Advances in Social Science, Education and Humanities Research, volume 543
Proceedings of the 2021 6th International Conference on Social Sciences and Economic Development (ICSSED 2021)

Research and Design of Screening Model of Target Customers for Internet Bank Financial Product

Naizhang Zhai, Eric McDermott

1325 North Lincoln Ave, apt2006, Urbana, IL, The United States, 61801
corresponding author：214 David Kinley Hall, 1407 W. Gregory Drive,1407 W Gregory Dr, M/C 707, Urbana, IL 61801

Corresponding author email: earwyn@qq.com

ABSTRACT
With the development of Internet finance, rapid and profound changes are taking place in the global technology and financial fields. The development of an innovative business model that combines technology and finance is booming in the field of Internet banking. In 1995, Security First Network, the first pure Internet bank in the United States, was established. Up to now, the US Internet bank has gone through more than two decades of development. According to a recent report, the total assets of Internet banking in the United States accounted for 5.1% of the total assets of the banking industry. Tencent's WeChat Bank and Alibaba's Zhejiang Internet Commercial Bank also borrow from the mature Internet banking model in Europe and the United States to occupy China Internet Banking market.

The booming of Internet Banking also bring great challenges to the managers. One the one hand, although the Internet bank has paid a lot of money, human resources for marketing, and developing more customers, the proportion of customers of specific banking business is still lower than traditional physical banks. On the other hand, some financial business has special requirement for costumers selection to ensure the smooth management and probability of such business. Credit business of Internet Bank among Small and Middle Size Enterprises (SMEs) is one of the typical examples. The faster development of SMEs has raised great opportunities for Internet Bank credit business to enlarge their costumers when competing with traditional physical banks. However, the disadvantages of SMEs request for special costumer screening of Internet Bank when offering such financial service. Marketing to invalid customers means wasting marketing resources. Therefore, how to use relevant resources to effectively identify target customers and reflect the value of marketing resources has become an urgent problem to be solved in the development of Internet Banking services. This article focused especially on Internet Banks' credit business for SMEs, trying to analyze the process of costumers screening on this field. This article attempts to use the Bayesian network to build a Internet bank SME credit access screening model to ensure the healthy and rapid development of Internet Bank's SME Credit business.

Keyword: *internet bank, finance, screening model*

1. THE NECESSITY OF COSTUMER SCREENING FOR INTERNET BANK'S SME CREDIT BUSINESS

1.1. The risks of SME credit business for Internet Bank

(1) False costumer information and fraud risk. Internet bank credit business analyzes and evaluates the credit status of borrowers with the help of Internet data and information. However, network virtualization makes the process of obtaining customer information more random. The authenticity, accuracy and timeliness of key information such as costumers' qualifications and debt paying abilities cannot be discerned by the bank on the spot. Some small sized enterprises who do not have access qualification may obtain credit resources by providing false information or fabricating false data. Some unscrupulous credit intermediaries, package the applicants and even defraud them through mastering the online loan application rules of various banks.

(2) Risk of deviation of the online loan customer warning model. Internet bank credit business can use information technology systems to build intelligent early warning model programs, through big data analysis and screening risks. However, due to incomplete information channels and the low utilization rate of internal and external data integration, some key risk information (such as the list of untrustworthy costumer, black clients involved in complaints, etc.) cannot be obtained in time. Some models have not been fully automated in the application process, and some information still needs to be collected and entered manually, which is prone to errors and even artificial data tampering.

(3) The characteristics of SMEs. Most small and middle sized enterprises don't have stable cash flow and profiting mode and are weak in paying debts in general. These require the bank managers to more carefully review the qualifications of each applications, especially the SMEs. Therefore, screening the costumers in specific business field is very important.

1.2. The benefits of SME credit business for Internet Bank

In recent years, small and medium-sized enterprises have developed very quickly, and have become an important part of the market economy. The growth of small enterprises is vast, the financial demand is rich, and the total loan demand is huge. As long as Internet banks change their minds, update their concepts, and follow proper methods, actively developing and innovating products and services that meet the needs of small enterprises, the business of SMEs credit are completely promising.

In addition, many large enterprises are built by small and medium-sized enterprises step by step, if Internet banks support a small enterprise, after the small company becomes a big one, it will continue to build a good relationship with the bank, which shall create a lot of comprehensive income for the bank. For example, HSBC in Hong Kong and the Li Ka-shing family business are such a cooperative relationship. Furthermore, many small and medium-sized enterprises are supporting services for large enterprises, and are upstream suppliers or downstream distributors of large enterprises. If the Internet banks develop credit business for these small and medium-sized enterprises, the corresponding large enterprises shall also stabilize and develop their business in the same banks.

2. METHOD FOR DESIGN OF COSTUMER SCREENING MODEL

2.1. Bayesian classification theory

The problem of classification is actually purposed to study the relationship between attribute variables and categorical variables to get the value of the unknown category. Bayesian network has a solid mathematical foundation, rich probabilistic expression ability, fusion ability, multi-source information expression and uncertainty problem processing ability. Therefore, Bayesian network is an effective tool for solving classification model problems. Due to the fact that Bayesian classification can not only represent the interdependence between variables, but also can integrate prior knowledge and sample data to train the classification performance, which effectively avoids the "black box" defects of artificial neural networks. It can also dynamically adjust the classification model during the classification process. On this basis, the Bayesian network classification model has high research value.

Assume $U=\{X, C\}$ as a finite set of random variables, where C is a categorical variable and the value range is $\{C_1, \dots \dots C_m\}$, $X=\{X_1, X_2 \dots \dots X_n\}$is the attribute variable set, According to Bayesian formula, the probability of sample $X_i =(X_1, X_2, \dots, \dots X_n)$ belonging to C_i can be expressed as follows:

$$P(C = c_j|X = x_i) = \frac{P(C_j)P(X_1,X_2 \dots,X_n|C_j)}{P(X_1,X_2,\dots,X_n)} = \alpha \cdot P(c_j) \cdot P(X_1, X_2 \dots, X_n|C_j) \quad (1)$$

Among the above formula, $P(X_1, X_2 \dots, X_n|C_j)$ is the likelihood of class C_j with respect to X_i, $P(c_j)$ is the prior probability of the class, α is the normalization factor, according to the chain rule of probability, the formula above could be expressed as:

$$P(c_j|x_1, x_2, \dots, x_n) = \alpha \cdot P(c_j) \cdot \prod_{i=1}^{n} P\left(x_i|c_j, \pi(x_i)\right) \quad (2)$$

Given a training sample set $D = \{u_1, u_2, \dots, u_n\}$, the purpose of classification task is to analyze the training sample D to confirm a mapping function $f: (x_i, x_2, \dots, x_n) \rightarrow C$, so that for any instance of unknown category $X_i =(X_1, X_2, \dots, \dots X_n)$, the class label could be calibrated. According to Bayesian Maximum Posterior Criterion, set an example of $X_i =(X_1, X_2, \dots, \dots X_n)$, Bayesian classification model shall choose the category which owns the maximum posterior probability $P(c_j|x_1, x_2, \dots, x_n)$ as the category label of this research.

Using Bayesian network as the classification tool is actually relying on formula (2). Since the Bayesian network expresses the full joint probability distribution of the variable set, as long as the Bayesian network structure of the variable set and the conditional probability distribution of the attribute variables are confirmed, $P\left(x_i|c_j, \pi(x_i)\right)$ can be obtained.

2.2. Tree Augmented Naïve Bayesian classifier (TAN)

According to different assumptions about the network structure, Bayesian networks can be divided into Naive Bayesian classification model, Tree Augmented Naïve Bayesian classification model and general Bayesian classification model. This article focuses on Tree Augmented Naïve Bayesian classification model which is developed on the basis of the other two models.

Naive Bayesian classifier faces directly the classification target, and has a concise and clear structure. However, the independence of its attribute variables is actually inconsistent with the real world condition. The Tree Augmented Naïve Bayesian classifier is improved on the basis. This classifier takes categorical variables as the parent nodes of all attribute nodes, and the attribute nodes form a tree structure. Categorical variables have no parent nodes. Other attribute nodes except C, have at least one other attribute as the parent node. The network structure (see the below Figure 1) shows that each attribute can have at most one related arc pointing to it except the pointing of categorical variable.

The working process of Tree Augmented Naïve Bayesian classifier has the following steps:

(1) Input training samples, initialize to a uniform format, define class variables and attribute variables, and

(2) If it is a classification task, then turn to (4), if it is a training task, then turn to (3), and

(3) Build TAN structure and Bayesian probability table and scan all training samples.

1) calculate the conditional mutual information entropy $I(x_i ; x_j|C)$ of each pair of attribute variables $i \neq j$

$$I(x_i ; x_j|C) = \sum_{x_i,x_j,c_k} \widehat{P_D}\ (x_i, x_j, c_k) \times \log \frac{\widehat{P_D}(x_i,x_j|c_k)}{\widehat{P_D}(x_i|c_k)\widehat{P_D}(x_j|c_k)} \quad (3)$$

Among them, P(.) is the empirical distribution of the frequency of event (.)

2) create a undirected weighted graph with its node of $X_1,X_2,\ldots,X_n$, the weight between X_i and X_j is $I(x_i ; x_j|C),\ \neq j$

3) build the maximum weight span tree of the undirected graph (first sort the edges according to the weight from large to small, and then follow the principle that the selected edge cannot form a loop, selecting the edge according to the weight of the edge from large to small. Then the largest weight span tree is formed)

4) specify an attribute node as the parent node, set the direction of all edges from the parent node to outwards, and convert the undirected graph into a directed graph

5) add a categorical node in the directed graph, and add an arc pointing from the categorical node to each attribute node to form a Tree Augmented naive Bayesian network structure

6) according to the Tree Augmented naive Bayesian network structure diagram obtained in the previous step, a Bayesian probability table is established

(4) Utilize the Bayesian probability table to get the classification result.

The model TAN established by the above process can maximize the network log-likelihood function of the given training data, and includes time complexity $O(n^2,N)$, where n is the number of attribute variables, N is the number of training samples. Experimental results show that under the same computational complexity and strength, the accuracy of TAN model is higher than that of the naive Bayesian classifier.

2.3. Feasibility analysis of constructing screening model through Bayesian network

The application of Bayesian network has achieved good results in many fields. With reference to the research and application of previous research, this article applies Bayesian network to construct a customer screening model for Internet Bank SMEs credit business. The feasibility of applying such model could be seen as follows:

(1) The Bayesian network organically combines directed acyclic graphs with probability theory. It not only has a strict probability theory foundation, but also has a more intuitive knowledge representation form. On the one hand, it can directly express the knowledge possessed by humans with directed graphs. On the other hand, it can also integrate statistical data into the model in the form of conditional probability, which clearly shows the correlation between variables. Many financial data used in designing the costumer screening model of Internet Bank have certain correlations, and Bayesian networks can use conditional probability to express their relationship.

(2) Bayesian network can learn the cause and effect relationship, not only can predict the results with data, but also find the cause of the problem when it occurs, and provide the decision maker with a most feasible solution under uncertain circumstances.

(3) Bayesian network can handle with the situation where there is missing data items. Many SMEs' financial data are incomplete. Bayesian network can solve this problem well.

(4) There is no definite input or output node in the Bayesian network, and the nodes affect each other. Obtaining the observation value of any node will affect other nodes, and the observation values of other nodes can by estimated and predicated through Bayesian network.

2.4. The process of applying Bayesian network in designing costumer screening model

The following process could be considered when applying Bayesian network:

(1) Design of indicator system. According to certain principles and referring to previous studies, the index system of the model is designed according to the actual needs of the model.

(2) Data selection and preprocessing. According to the designed index system, certain sample data of SMEs costumers are selected, and the samples are divided into two parts: the training set and the test set. The training set consisting 70% of the total simples is used to train the model and the remaining 30% test set is used to test the network. According to the mathematical principle of Bayesian network, the data needs to be preprocessed for discretization.

(3) Structure learning. If it is a Naive Bayesian network method, this part is not needed. If it is a Tree Augmented Bayesian network, the structure of the model should be learned. It is difficult to completely learn the structure of the Bayesian network so sometimes it can only sacrifice accuracy. The complete data is learned through heuristic search, such as K2 algorithms, greedy search CS algorithms, etc. For the learning of missing data, there are EM algorithms and MCMC algorithms. According to the actual situation, different algorithms are used to learn the structure of the model to finally obtain the topology of the Bayesian network model.

(4) Parameter learning. Learning the model parameters according to the obtained Bayesian network model topology structure, to obtain the conditional probability of each relevant node, so as to facilitate the next step of predicting the accuracy of the model.

(5) Test the accuracy of the model. Using the conditional probability obtained by parameter learning to calculate the classification of the test data, then compare the obtained results with the actual results to get the accuracy of the model.

3. CONSTRUCTION AND APPLICATION OF THE COSTUMER SCREENING MODEL IN INTERNET BANK CREDIT BUSINESS ON THE BASIS OF BAYESIAN NETWORK

3.1. variables selection, data selection and prepossessing

For the construction of the costumer screening model under the business of SMEs credit, the research in this article is based on the existing data of Internet banks, mainly based on the statistical data, statements and stable records submitted by the SMEs, as well as the annual reports disclosed by Internet Banks.

3.1.1. Variables selection

To confirm the relevant variables, this article considered both the important principle in selecting variables as well as the prior research results.

On the one hand, the following principles should be kept in mind when choosing the proper variables. (a) Systematic principles request for careful analysis of Internet Bank's business practices, so that the selected indicators can fully reflect the real operating conditions of the bank, and comprehensively reflect the credit conditions of SMEs from different angles and different levels. (b) The principle of importance. One of the main reasons for the failure of SMEs is the shortage of funds and the reduction of their solvency. These two reasons determine that the costumer screening model of SME credit business should pay special attention to relative liquidity indicators, including but not limited to current ratio, quick ratio, receivables turnover ratio and other indicators. (c) The principle of sensitivity. The selected indictors can sensitively reflect the credit status of SMEs, that is to say, once the indictor changes, it can be sensitively reflected in the screening model. Therefore, when designing the indicator system, we should give full consideration to the relevance between these indicators and whether the company breaches the contract, and try to select the indicators that have high correlation and strong leading capacity. (d) Dynamic principle. As the enterprises have different characteristics at different stages of development, the selected indictors system should change automatically according to the changes of SMEs. Replacing static indicators with active indictors to realistically reflect the status of the enterprise and ensure the accuracy of the model. (e) Operability. Ensure that the selected indictor system shall strive to obtain relevant and reliable information from a wide range of economic data. The data source should be reliable, the reasoning process is scientific, easy to quantify, suitable for operation, and the direct correlation among these indictors is preferably weak. (f) The indictor system should reflect the

company's profitability, solvency, operating ability, cash flow ability and other abilities.

On the other hand, when referring to the prior research results, In the article of "Financial Ratio as Predictors of Failure", Beaver selected six financial ratios from the all 29 financial ratios, and used the univariate analysis method to build an early warning model. The specific indicators in his research includes cash flow/total debt, which is the best indicator to predict the success or failure of the enterprise, net profit/Total assets, total liabilities/total assets, working capital/total assets, current ratio, (quick assets-current liabilities)/operating expenses.

This article selects the indicators from the four aspects of debt paying ability, profitability, operating ability and developing ability. Considering that SME loans are mainly short-term loans, more short-term indicators that reflect liquidity are selected, such as current ratio, Quick ratio, etc. In addition, considering that dynamic indicators can better reflect the short-term situation of the customer, this article uses dynamic indicators to replace static indicators, such as cash flow ratio, main business growth rate, etc.

3.1.2. data selection and preprocessing

The research in this article is based on the database of SMEs stored by Internet banks, and 640 enterprise samples are extracted from it, and the data samples are divided into 70% training set and 30% test set. Then the training set has 448 samples and the test set has 192 samples. This article collect and organize relevant indicators and data on the basis of the reports, steady operation records and statistical data that regularly submitted by SMEs.

Take the value of categorical variable C as 0 and l, where 0 and 1 represent default and non-default, respectively.

According to the mathematical principle of Bayesian network, the accuracy of the Bayesian network model established with discrete variables is higher than the Bayesian network model established with continuous variables. Therefore, before establishing the model, it is necessary to discretize the continuous variable financial indicators. According to the average distribution method proposed by Sumit Sarkar and Rma S Sriran, the financial ratios are graded at equal intervals.

3.2. The structure and parameter learning of the costumer screening model for Internet Bank credit business

3.2.1. The structure learning of the model

This article applies K2 algorithm to learn the structure of Bayesian network model. The key problem of using K2 algorithm is how to determine the input order of node variables, so as to reduce the search space, reduce the computational complexity, and learn the Bayesian network structure with high efficiency. According to the mutual information theory, conditional mutual information can indicate the degree of dependence among attributes. Therefore, in this article, the conditional mutual information between each attribute pair is calculated based on the known category C, and the input order of node under the K2 algorithm is determined according to the degree of correlation.

(1) Calculate conditional mutual information between attribute variables

In order to study the interdependence between attributes and categories and construct a directed graph, it is necessary to calculate the conditional mutual information between each attribute and category in the data set $I(x_i ; x_j|C)$. When the conditional mutual information of the two attribute variables x and y is very small, x and y are called conditional independence; otherwise, the probability dependence is stronger. That is, when the conditional mutual information between X and y is large, it means that y may be the parent node of x.

According to the conditional mutual information between the variables given in Table 4, arrange the conditional mutual information of attribute variables and the root node C from the sequence of large to small, and the node input order under the K2 algorithm can be derived as C, x_2, x_3, x_1, x_6, x_5, x_{10}, x_4, x_8, x_7, x_{11}, x_{13}, x_9, x_{14}, x_{12}, the specific values are shown in Table 1.

Table 1 K2 algorithm node input sequence and conditional mutual information between the attributes and root node C

root / node	x_2	x_3	x_1	x_6	x_5	x_{10}	x_4
C	0.45505	0.36916	0.35046	0.26213	0.23654	0.019241	0.017463
root / node	x_8	x_7	x_{11}	x_{13}	x_9	x_{14}	x_{12}
C	0.016156	0.015282	0.003045	0.001924	0.000199	0.000125	0.000316

(2) Use K2 algorithm to determine the topology of the model

This article uses the K2 learning algorithm function learn_struct_K2( ) defined in the Bayesian network toolbox in Matlab 7.0, to obtain the topological structure of the Bayesian network model.

3.2.2. parameters learning of the model

After determining the topology of the Internet Bank costumer screening model of credit business based on the Bayesian network, we need to learn the parameters of the model through the data, which is, the conditional probability distribution of the node of the Bayesian network model. The conditional distribution probability tables of some model nodes obtained through data learning.

3.3. Model empirical result analysis

After the structure learning and parameter learning of the model, the next step is to use the test set to check the accuracy of the model, and then compare the obtained results with the actual results of the test set to get the accuracy of the model. Input teh 192 test sets into the established model, and the classification prediction results are shown in the following table 2.

Table2: classification prediction results

Actual results / predicted results	Defaulting enterprises	Non-default enterprise
Defaulting enterprises	160	3
Non-default enterprises	16	13

From the above table 2, it can be concluded that the overall accuracy rate of the Internet bank costumer screening model of credit business based on the Bayesian network is 90.10%, that the accuracy rate of prediction for non-defaulting enterprises is 90.91% and the accuracy rate of prediction for defaulting enterprises is 81.25%, among which the accuracy rate of non-default companies prediction is 9.66% high than that of default companies prediction. It can be seen that the accuracy rate of the model is still quite high, reaching the expected goal, so the Bayesian network-based costumer screening model is effective and feasible.

4. CONCLUSION

The research results of this article have important practical guidance for the managers of Internet banks. To identify the characteristics of SME that can enter the bank's credit process through a screening model based on the Tree Augmented Naïve Bayesian Network. Therefore, this model can be used as a useful decision support tool to help auditors make decisions and improve audit efficiency and effectiveness. With the help of this decision support tool, bank auditors can identify credit business customers with greater risk of default at an early stage with reliable accuracy. Scholars can also use the semi-automatic decision-making system studied in this article as a reference to investigate the deviations of Internet bank auditors during the audit process, such as comparing the control experiment completed by the auditor with the prediction results of the probability model to better understand the audit deviation. At the same time, the research results of this articled can also be used as the basis for further exploration, and this probability model can be refined to provide better prediction performance.

There are still some deficiencies in the research of this article. For example, due to the difficulty of data collection, the sample of SMEs collected in this article comes from various industries, which ignores the differences in the credit access policies of Internet banks for SMEs in different industries. Therefore, the work that needs to be further supplemented is to study the credit costumer screening model for SMEs according to different industries.

REFERENCES

[1] B.David, A.V Thakor. “Bond Covenants and Delegated Monitoring”. Journal of Finance, 1957, 43:397—412.

[2] Stiglitz, Weiss A. “Credit Rationing in Markets with Imperfect Information”. The American Economic Review, 1981, 71:393—409.

[3] Y Shen, M Shen. “Bank Size and Small—and Medium-sized Enterprise(SME) Lending：Evidence from China”. World Development, 2009, 4(37): 800—811.

[4] M Agostino, F. Gagliardi, F.Trivieri. “Credit Market Structure and Bank Screening: An Indirect Test on Italian Data”. Review of Financial Economics, 2010 , 19:15 1-160.

[5] M Amano. “Credit Rationing of a Bayesian Bank with Simple Screening Technologies”. Japan and the World Economy, 1999, 11:545—556.

[6] George R., G. Clarke, R. Cull. “Foreign Bank Participation and Access to Credit Across Firms in Developing Countries”. Journal of Comparative Economics, 2006, 34: 774—795.

[7] D.Hebb. “Financial Predictors for Different Phases of the Failure Process”. Omega,1949,21(2):215—228.

[8] Pearl.A. “Bayesian Method for the induction of Probabilistic Networks from Data”. Machine

Learning, 1986, 12:309—347.

[9] Shafer. “Comparing Bayesian Network Classifiers”. In the 15th Conference on Uncertainty in Artificial Intelligence,1990.

[10] Zellner. “Bayesian Limited Information Analysis of Simultaneous Equations Mode”. Econometrica, 1962, 44(5):1045—1075.

[11] S. Kanungo, K. Jain A. “Evaluation of a Decision Support System for Credit Management Decisions. ”Decision Support Systems, 1978, 30:419—436.

[12] Williamson. Costly Monitoring, “Financial Intermediation and Equilibrium Credit Rationing” Journal of Monetary Economics, 1986, 29: 102 — l10.

[13] Steven, “A Sharpe . A Survey of Credit and Behavioral Scoring: Forecasting Financial Risk of Lending to Consumers”. International Journal of Forecasting, 1990, 23:42—51.