Paper 159-2008

Identifying Potential Default Loan Applicants - A Case Study of Consumer Credit Decision for Chinese Commercial Bank[1]

Qiwei Gan, Binjie Luo
Southwestern University of Finance and Economics,
Chengdu, Sichuan, China
Zhangxi Lin
Texas Tech University
Lubbock, TX, USA

ABSTRACT

Consumer credit is a lucrative but risky business. In order to control risk and maximize profits, commercial banks around the world have made great efforts to develop various analytic models to identify potential default loan applicants. This is also critically important to China, as the non-performing individual loans of Chinese commercial banks have been fast growing because of the myopic business attitude and the overheated economic growth. To prevent the US sub-prime kind of crisis in China, Chinese commercial banks are adopting advanced analytic means to help make loan decisions. This paper reports a data mining application in the analysis of default loan applicants using a real dataset consisting of 641,988 observations obtained from a Chinese commercial bank, located in the southwest of China. An exploratory study of the dataset led to a number of interesting statistic figures that may characterize the applicants in the western region of China. In the analytic study, we constructed two types of models, unweighted and weighted, with SAS® Enterprise Miner. The models have revealed a number of useful findings that meet our expectation. They demonstrated good predictive power for loan decision making. The success of the study has consoled our concern in the data quality that is relevant to the efficiency of the data collection system rooted in the existing Chinese business system.

INTRODUCTION

Over the past decade, consumer credit[2] business of commercial banks has been expanding rapidly in China. In 1998, the balance of consumer loans of commercial banks in China was about US$ 10 billion. By the end of 2005, the balance reached US$ 293 billion (China State Statistics Bureau, 2006) – increasing more than 29 times in seven years. On the other hand, the fierce competition of consumer credit in China seems to harbor potential risks for commercial banks. The real estate prices in China have been skyrocketing over the years and bubbles are forming. Traditionally, consumer credit in China seems to have less risk, especially mortgage loans—because it just postpones risks not eliminates them and Chinese commercial banks began to provide mortgage loan about ten years ago and the maturity of most loans is more than fifteen years. As the real estate market is booming, the loans generate profits to banks and the risk of customer default is less, the business seems promising. If the prices of real estates fall down, the risk will become significant and the default on loans would incur great loss – this is very possible as the warning of the overheated Chinese economy is getting louder. However, the myopic attitude and the pressure of competition are driving many Chinese commercial banks for more customers without carefully screening. As getting a mortgage loan from a Chinese commercial bank is relatively easy, the non-performing consumer loans are increasing with real estate market fluctuation of recent years in China (http://www.buyusa.gov/china/en/bank. html). According to Shanghai Banking Supervision Bureau's statistics, the non-performing individual loans of Chinese commercial banks in Shanghai in the third season of 2006 totaled RMB 3.275 billion, that is, RMB 0.4 billion more than the last season and RMB 0.842 billion more than the beginning of the year. The non-performing rate of individual loans is 1.14%, an increase of about 10% than last season and about 20 % than the beginning of the year respectively. In particular, the non-performing consumer loans accounted for 90% of those non-performing individual loans (http://old.go24k.com/bank/news.asp?id=41470&anclassid=6).

With recent US sub-prime mortgage crisis, the consumer credit decision of Chinese commercial banks becomes critical. The US sub-prime mortgages are those cases that banks gave cash to people with poor or no credit history

[1] Upon the request from the bank, the bank's name is not mentioned here. The findings and conclusions reported in this paper are solely the opinions of the author(s) and do not necessarily represent the viewpoints of their institutions.
[2] In China, consumer credit includes both mortgage loans and others individual consumption loans, however, about 90% of consumer credit is mortgage.

and on low incomes who are in no position to pay back the loans, at higher than repayment levels. When the economic conditions get worse and the interest rate rises, many of them will default and incur much loss for the banks. In order to prevent the US sub-prime kind of crisis, from now on, Chinese commercial banks must be more serious in the consumer credit decisions. They must try to predict who will default before extending loans to them. While consumer credit is a lucrative business, the banks do not want to refuse those who will not default. Therefore, banks want to understand their existing customers and sort out the common features of defaulted customers and pinpoint potential default customers from loan applications.

In order to help commercial banks manage their risks and maintain financial stability, the People's Bank of China (China's central bank) set up a nation-wide distributed database system in 2005, the Personal Credit Record System (PCRS) (www.pbc.gov.cn), which has been keeping tracks of individual and business credit records. The credit transactions of anyone in China who had some credit business with financial institutions in China have been recorded in the database, such as getting a loan from a bank, using credit card, etc. When an applicant applies for a loan from a bank, the bank will require the applicant to authorize it to view his credit records from the database, if any. Then loan offices will evaluate these records with their experience, or, they will use a credit scoring model, if any, of their bank to score the applicants. This can help banks make better decisions.

This paper is to report a data mining application research effort in the analysis of default loan applicants from a real dataset obtained from a Chinese commercial bank. Our focuses at the first stage of the research project include: data cleansing and data quality evaluation, basic credit scoring modeling with SAS® Enterprise Miner. It is worth of noting that data profiling and quality evaluation is very important in analyzing the data collected in China. Without quality data, the analytical effort will result in inaccurate outcomes. The authors intends to show how the data mining technique can be applied in bank loan decision making in the context of China and how it is used to predict the potential valuable and default customers with real data from one of the Chinese state-owned commercial banks.

PROBLEM DEFINITION

We performed a case study on the consumer credit decisions of a Chinese commercial bank. The bank is a Sichuan Provincial Branch of bank, which is one of the "big four" state-owned commercial banks in China. Sichuan province is located in southwest of China with an area of 485,000 km² and a population of 87 million. In 2006, the GDP of the province is US$ 115 billion, with the per capita GDP of 1,132 US dollars. The bank maintains a vast database of consumer credit records, and intends to identify the potential default customers from their profiles. For historical reasons (policy-oriented loans to low income people), the customer default rate in Sichuan province is relatively high. However, the default customers usually account for only a small percentage among the total, as of 5 to 10 percent. This then challenges us with the imbalanced dataset problem.

The above situation characterizes our research issue in two aspects:

1) As we are to study the data mining problem with an imbalanced dataset, we articulated to have the same composition of the dataset from the previous database as that to be available in the current database. Many of those default applicants have actually been granted the loan benefited from government's policy but there is no indicator in the dataset to show which of these default applicants have previously been granted the loan.

2) Although Sichuan is one of the largest provinces in China, its economic development status does not rank high compared with the more developed coast areas in China, for example, Guangdong province. The composition of the loan applicants is significantly featured by the economic development status in the southwestern region of China.

The main objective in this paper is to develop a classification model that can provide optimum credit decision-making support for a Chinese commercial bank in the western region of China.

THE DATASET

The bank provided a database of actual clients' information for this research project. For confidential purpose, we eliminated the variables that can identify the clients. After pre-processing, the working database contains 14 variables and 641,988 observations. Each observation, the information of one client of the bank, includes sex, age, education, professional title, occupation, the years of continuing work, marriage status, household income, health, housing status, inhabitation period, number of dependant(s), and bank's classification of clients.

Table 1. Variable Details

Variable	Description	Measurement	Missing rate
Birth	Date of birth	Interval, in the format of yyyymmdd	59%
Age	Age in 2004	Interval	59%

Sex	Gender	Binary	0%
Edu	The highest education received	Ordinal, 10 values from elementary school to doctoral education	0%
Title	Professional titles	Ordinal, four values from junior to senior.	0%
Occu	Occupation	Nominal, 14 values	0%
Workyear	Years of continuing work	Interval	70%
Marriage	Marriage status	Nominal, 3 values: 1 - single, 2 - married, 3 - devoiced	52%
Income	Household annual income	Interval	46%
Health	Health conditions	Ordinal, 3 values: 1- good, 2 - average, 2 - bad	21%
House	Housing status	Binary, 1 - self-owned, 2 - rent	24%
Inhabit	Inhabitation	Interval, the number of years living in the same place	27%
Support	The number of dependents	Interval	48%
Class	The bank's classification on a client	Ordinal, four values: 1 - Excellent, 2 - Good, 3 - Limited, 4 - Default	0%

Class is the target variable in this study. Class 1 and 2 clients are typically good loan borrowers. Class 3 clients are "not good enough", but Chinese lenders are still providing loans to them, trying to take more market share in the competitive market. Figure 1 shows that Class 4 clients take about 8% of the population and are to be identified by the model.

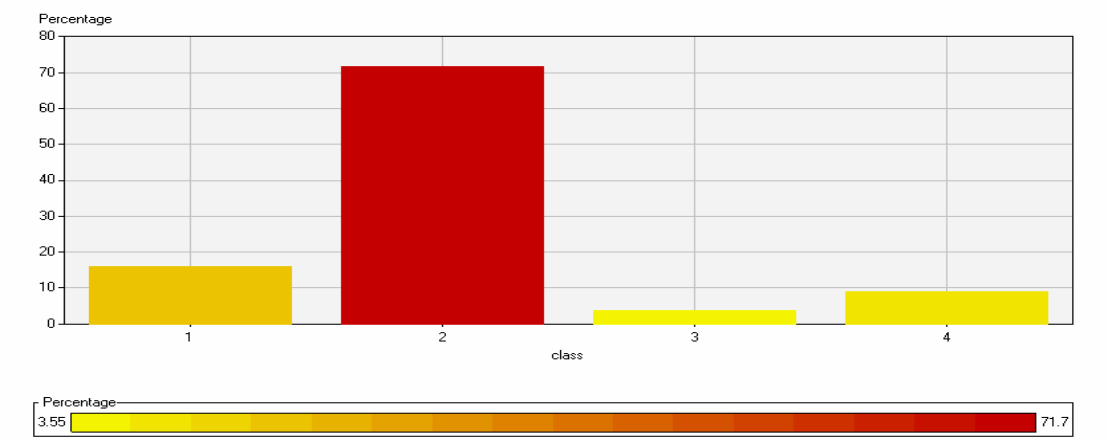

Figure 1a: The distribution of target variable – Class (graph)

Target Value	Prior Probability
4	0.0807260594
3	0.0386254529
2	0.7200110283
1	0.1606374594

Figure 1b: The distribution of target variable – Class (table)

The quality of the dataset is not very good, with many missing values and data errors. Therefore, data preprocessing is important to guarantee quality data mining outcomes. We mainly treated the age values of the data in steps. First, we checked the consistency between Birth and Age, and corrected apparent errors. We then explored the dataset and did primitive data mining. The results showed that the value of age has little effect on the model, which can be

explained as the effect of too many missing values. Then we simply set a derived variable from Age, “Age_Yes”, to indicate if the age value was present or not. We will show in the model section that Age_Yes does contribute to the model.

The following are some interesting findings from the exploratory study of the dataset, which may characterize the features of loan applicants in the western region of China:

1) Majority (about 90%) of loan applicants are male.
2) More than 40% loan applicants only studied at the senior high school.
3) Almost 80% loan applicants are peasants from the countryside. This characterizes the customer population in this specific bank.
4) Apart from more than 45% missing value, about 40% loan applicants fall into the group with the lowest income.
5) More than 65% of loan applicants have good health condition. Considering 20% applicants did not provide the data, the percentage of applicants with lower health condition is very low.
6) About 65% of the loan applicants have lived in their house for more than 10 years. More than three quarters of them already have their own house.
7) Less than 20% of the loan applicants claimed they have more than one dependent. Taking into account that the majority of these applicants are peasants from the countryside, we can realize the effect of one-family-one-child policy in China.
8) More than one third of loan applicants are at the age of 27 to 32.

These features can also be used to design specific financial products and services for the existing and future customers of the bank.

SAS MODELS

We analyzed the data in two steps. We first simply built up a classification model to identify those default loan applications for a binary decision problem. Then we extended the model by classifying the clients into all four classes with a carefully conceived profit matrix.

CLASSIFICATION MODEL FOR BINARY DECISION MAKING

We used SAS® Enterprise Miner 4.3 to try three classification models, decision tree, logistic regression, and neural network. We finally discarded the neural network model because its performance is not as good as other models. Figure 2 shows the final data mining modeling diagram being reported in this paper. The configurations of nodes are:

1) Data Input node: set Age, Birth, Class rejected, and use Target41 as the target variable, which is defined as 1 for Class 4 clients, 0 for others.
2) Transform Variables node: defines a new variable Age_Yes, 1 for the presence of age value, and 0 for the non-presence of age value.
3) Data Partition node: the proportion of training, validating, and testing dataset sizes is 40%:30%:30%. Data partitions are stratified with variable Target41.
4) Replacement node: uses tree imputation for interval variables, and distribution-based imputation for class variables.
5) Regression node: uses the Stepwise modeling method and Quasi-Newton optimization method.
6) Tree node (default): default
7) Tree node (Gini): uses Gini reduction for splitting criteria, and Total leaf impurity (Gini index) as the model assessment method
8) Assessment node: default
9) Score node: apply training data score code to score data set
10) Data input node: select scoring dataset—scoring.smpb25X6
11) Insight node: select dataset—DMDATA.SD738XH

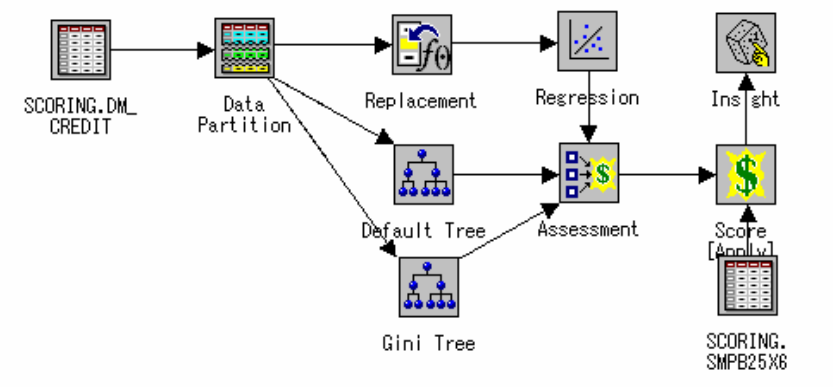

Figure 2: The SAS Data mining model

The results show that the classification performance of the Tree model is much better than the regression model, and the Tree model configured with Gini splitting criteria slightly outperformed the Tree model using default configuration (Figure 3). The results are satisfactory. For example, at the 40 percentile, the lift value of the model is about 2.0, and captures about 80% of interested observations.

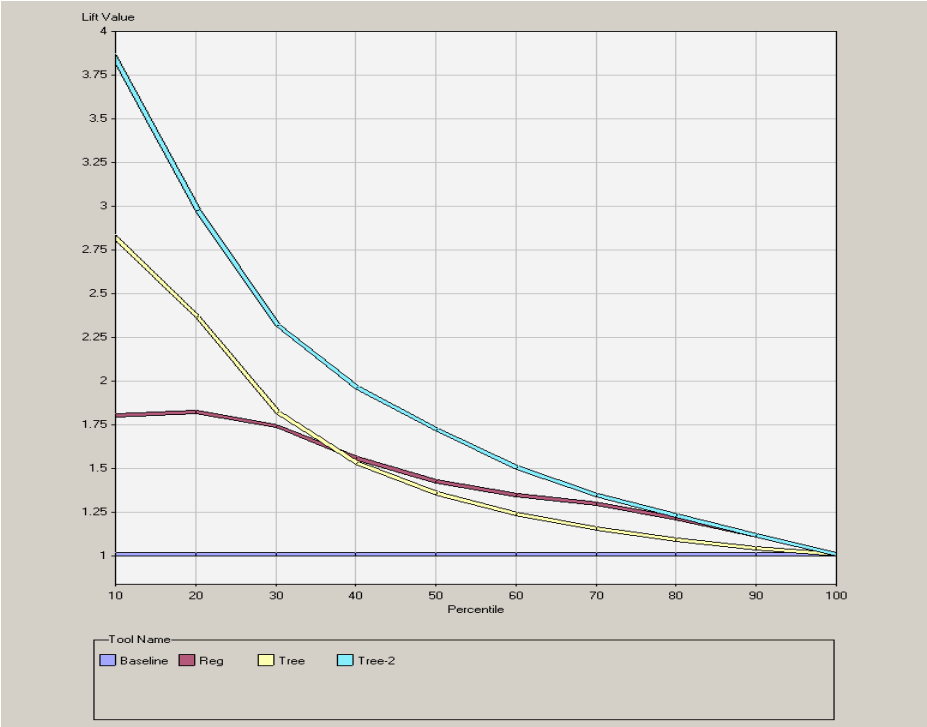

Figure 3a: Lift of the model (Tree-2 denotes the Tree model with Gini criteria)

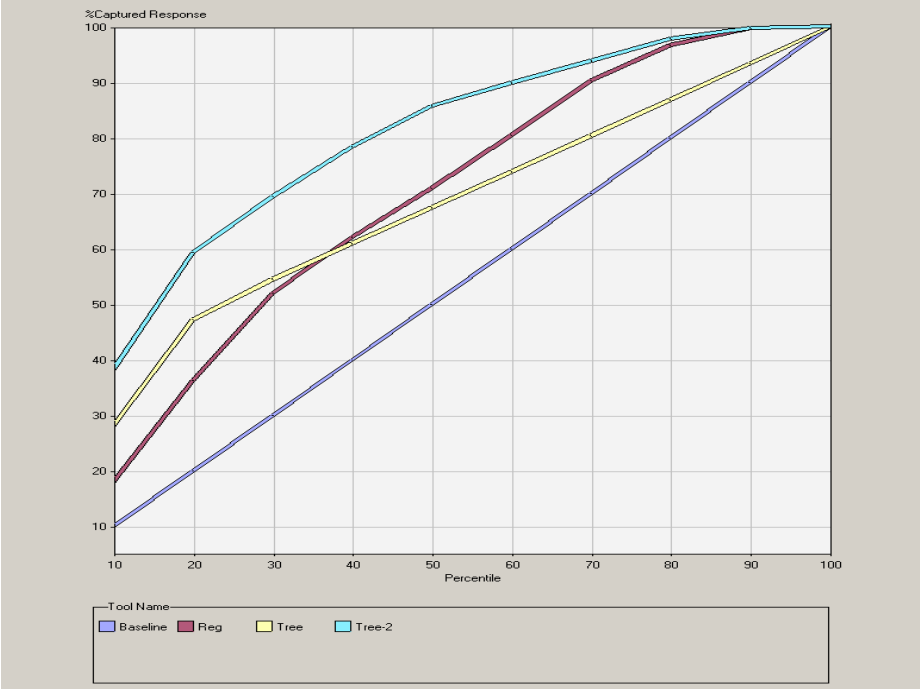

Figure 3b: Captured responses from the dataset.

The following findings from the Tree model are interesting (Figure 4):

1) The duration that the applicant stays at the same address is the number one variable to classify the loan applicants. It seems that applicants with a value of Inhabit are likely to have a better credit in average than others without Inhabit values.

2) In the branch of Inhabit value being missing, the applicants with Age value have a lower probability to be of Class 4. An applicant with a higher education level tends less likely to be of Class 4.

3) In the branch of Inhabit value being present, the applicants with higher incomes have lower possibility to be of Class 4. Those having fewer dependents seem to have better credit than those having more dependents.

The above findings sound reasonable and match the common sense of people.

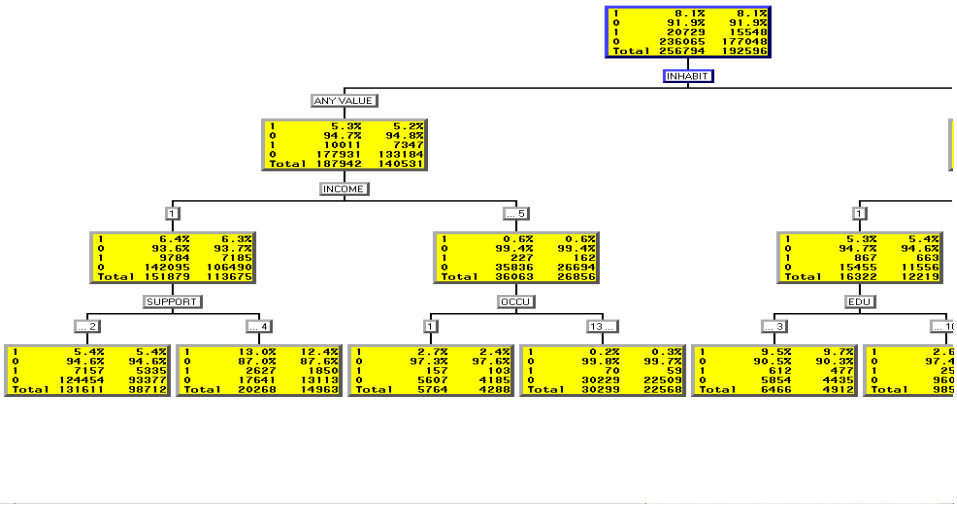

Figure 4a: The outcomes of the decision tree model (the left-hand part)

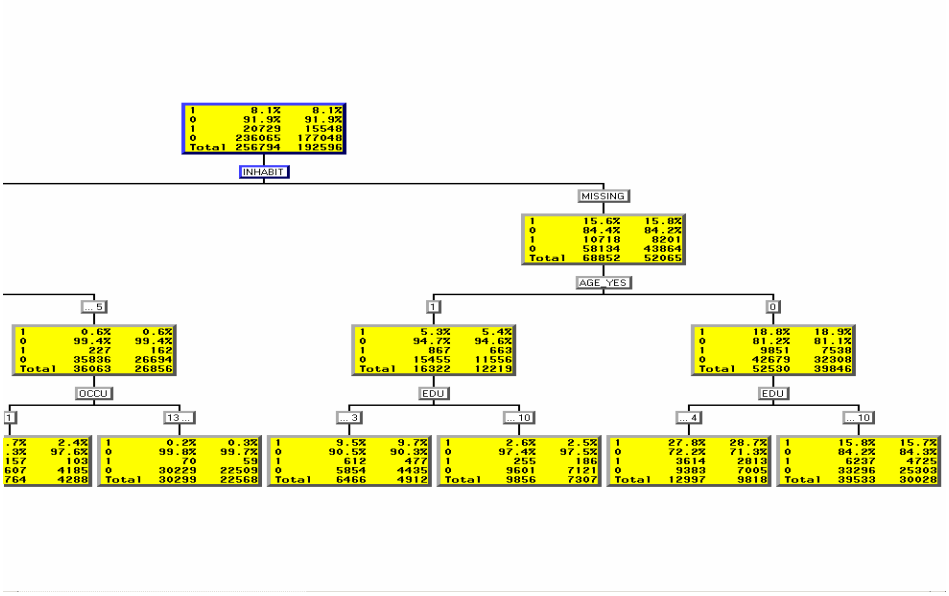

Figure 4b: The outcomes of the decision tree model (the right-hand part)

CLASSIFICATION FOR 4-CLASS DECISION MAKING

This classification model has a model diagram similar to the previous one but different in that it uses Class as the target variable that comes with a profit matrix (Figure 5). The criteria in designing the profit matrix include:

1) When the class of a client is correctly identified, the profit reaches the maximum among others.

2) The absolute value of negative profit by wrongly classifying a Class 4 client to other classes is higher than the profits earned from any other class of client assigned to the same class.

3) Once identified as of Class 4, the client will not be considered for loan.

LEVEL	4	3	2	1
4	0	-10	-30	-50
1	0	1	10	20
2	0	1	10	5
3	0	1	1	1

Figure 5: Profit matrix

Figure 6 demonstrates that the performance of the Tree model with the Gini splitting criteria is again the best among three models. The average profits from the Gini Tree model, the default Tree model, and the Regression model are 7.46, 7.37, and 5.30 respectively.

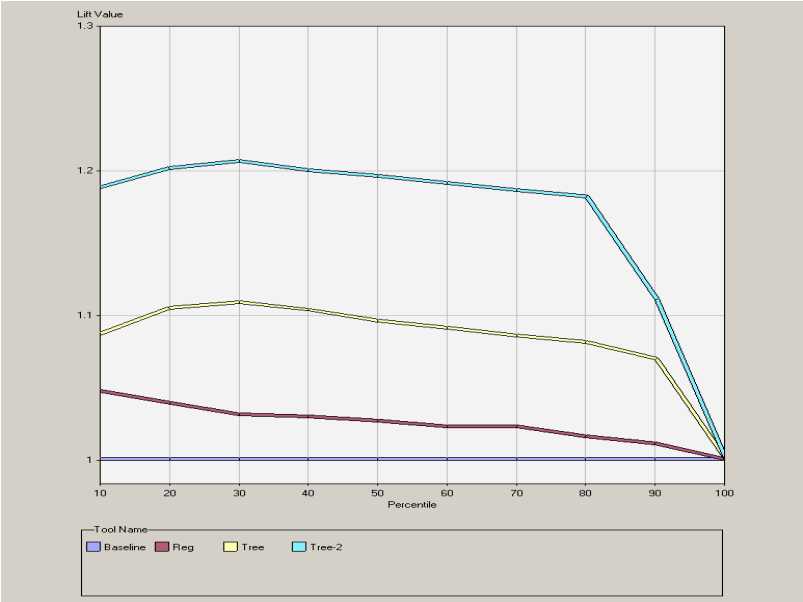

Figure 6a: Lift values

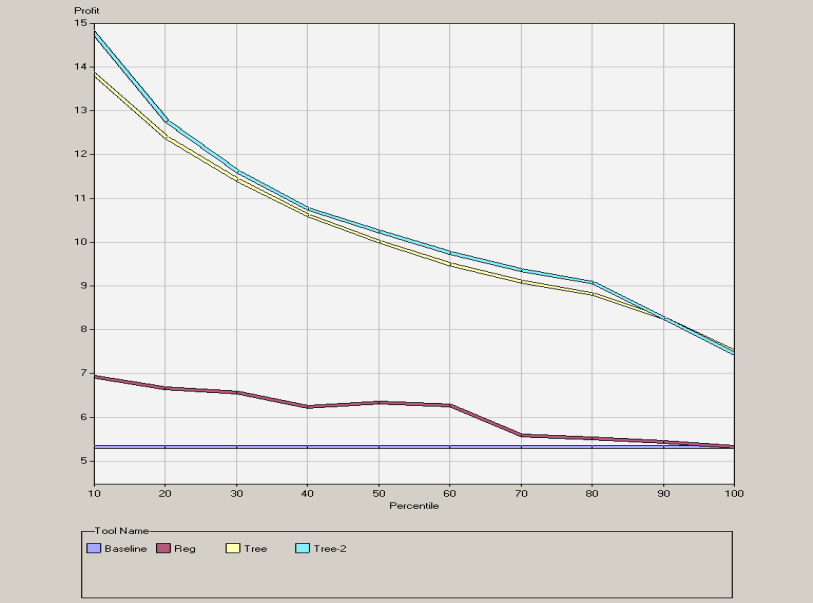

Figure 6a: Profit curves

The tree view of upper two levels of results from the Gini Tree model hints the following general rules for decision-making (Figure 8):

1) Inhabit is still the first criteria to classify applicants when the classification is weighted. The larger value of Inhabit implies better profitability. The intuition is that when one stays in the same address for a longer period of time, he may have a stable job and a stable salary; therefore, he may have good credit standing.

2) The diagram shows that most of the rules lead to Class 2 decision. The intuition is that more than 72% clients are of Class 2.

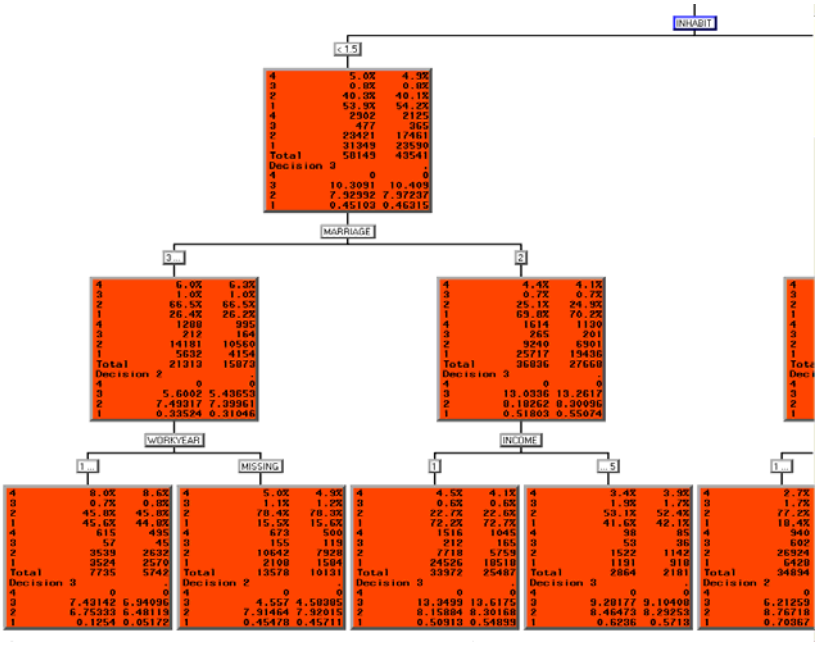

Figure 8a: The left-hand side of the Decision Tree

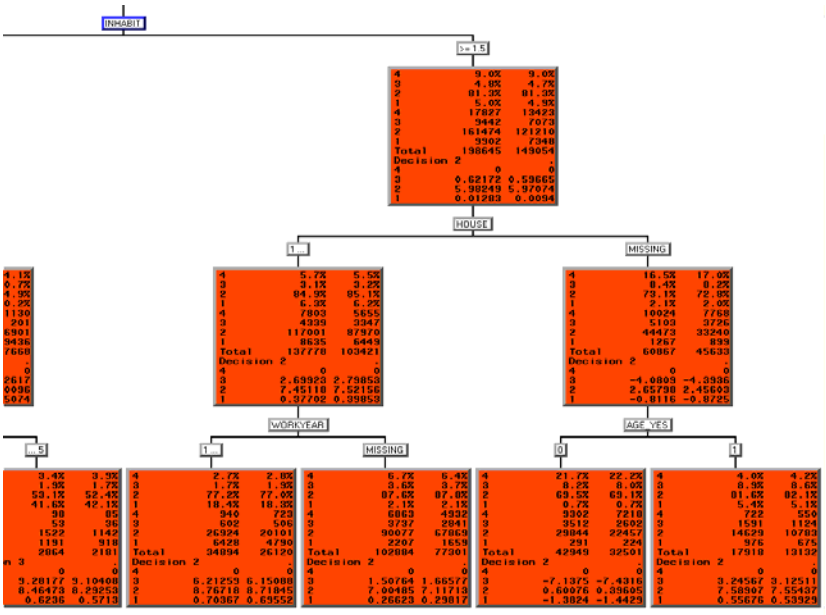

Figure 8b: The right-hand side of the Decision Tree

With a more thorough examination of the tree view to the deeper levels, we can find the following two Class 4 rules:

1) If Support is 3 or 4, Inhabit is between 1.5 and 3.5, Worksyear is missing, and House is not missing, there is a probability of 55.5% the applicant is of Class 4. This rule has identified 820 default applicants out of 1,477 observations.

2) If Occu is 1, 2, 3, 4, 6, 8, 10, 11, 13, or 14, Health = 1, Age_yes = 0 (missing), House is missing, and Inhabit > 1.5, there is a probability of 23.75% the applicant is of Class 4. This rule has identified 8,894 default applicants out of 36,181 observations.

The above outcomes are from a training dataset of 20,729 default applicant records, meaning more than half of Class 4 applicants are not covered by these rules.

THE RESULT OF SCORING

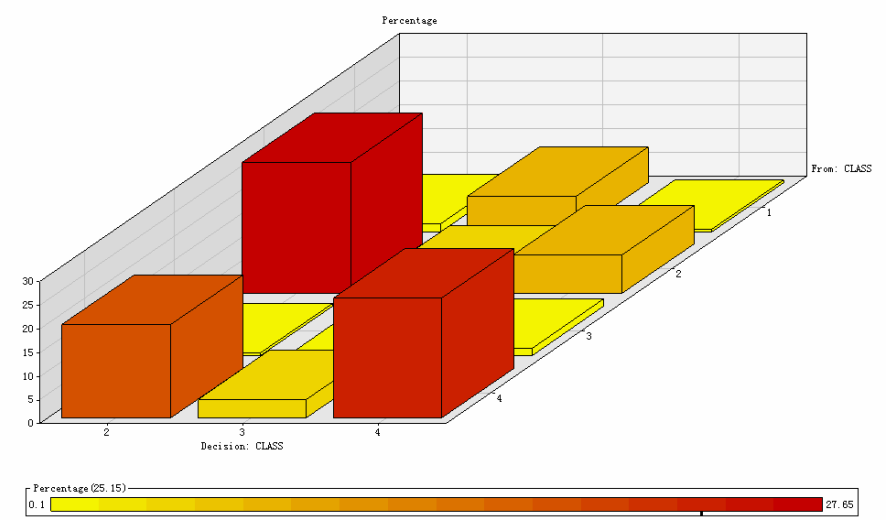

Figure 9a: Scoring Result

Decision / From	Class 1	Class 2	Class 3	Class 4
Class 1	0	17.01%	76.80%	6.19%
Class 2	0	69.65%	9.95%	20.40%
Class 3	0	25.58%	4.65%	69.77%
Class 4	0	40.25%	7.84%	51.91%

Figure 9b: Scoring result

Figure 9a and figure 9b report the scoring result of model graphically and quantitatively.

1) None of the scored customers, whether he is class 1 originally or not, is classified into class 1. Originally, class 1 customers account for only 9.7% of the total customers, and our target is to identify class 4 customers. This may cause the above-mentioned result. If we define the problem differently, say, to identify the best customers, the result will be different.

2) Most (69.65%) of original class 2 customers are classified into class 2. This implies that the model has good performance in identifying class 2 customers. Class 2 customers are normal customers of the bank and generate normal profits for the bank, so in the future research, we will focus on these customers and try to retain these customers for the bank.

3) Only 4.65% of original class 3 customers are classified into class 3, but 69.77% of original class 3 customers are classified into class 4. Both class 3 and class 4 are potential default customers and the bank would decline them. Our setting is to maximize profit, while class 3 and class 4 customers would incur loss to the bank, so the model is prudent and will decline most of class 3 and class 4 customers, as a matter of fact, many class 3 clients would shift to class 4 when the economy is getting worse.

4) 51.91% of original class 4 customers are classified into class 4, and 7.84% of original class 4 customers are classified into class 3. If we add them together, 59.75% of original class 4 customers would be refused by the bank. The result seems not bad. However, still there are 40% of class 4 customers are classified into class 2. In the future research, we will explore class 2 and class 4 customers and try to find out their specific characters in order to make the model have better performance.

DISCUSSIONS AND CONCLUSIONS

In this paper, we report the data mining outcomes from a real dataset of loan applicants obtained from a commercial bank in Sichuan province, China, applying basic SAS data mining models. The results from two different analytic

projects, unweighted and weighted, are encouraging. In particular, they resolved our concern about the usability of the data because of the data quality problem in the context of Chinese loan applicants and the specific data collection system in the Chinese commercial bank.

In the data preparation stage, from the preliminary explore of the data, we have a better understanding of the customers of the bank, and summarize their common characteristics. These characteristics are bank- and region-specific. They will help us to explain and interpret the results of data mining. Moreover, they are very useful for the future financial products design.

When it comes to missing value of some variables, we first delete those observations with missing values, but it turns out that the result is not satisfactory. Because many observations with missing values are falling in the category of class 1, when they are deleted, their characteristics cannot be represented in the model. We then replace some missing values and the result is not satisfactory either. At last, we transform some variables into binary variable which just indicates whether the value is missing or not. The result is much satisfactory than the previous two attempts. Our intuition tells us that this is reasonable to human behavior. When the bank is collecting customer information, the customers are not required to provide all the information. They tend to provide favorable information. This can be used in our model to make decisions.

In the modeling stage, when we divide the customers into only two classes: granting and turning down the application, we get a few very interesting findings. These findings may reflect the characteristics of the western region of China as the household income of most people in this dataset is below the national average level. Since about 40% loan applicants fall into the group with the lowest income apart from more than 45% missing value, income level is not significant in the model, which is different from what people with common intuition expect. Hence, we turn to other indicators to investigate the financial stability of customers, for example, their inhabitation status.

The second model is profit oriented and subject to changes with regard to the profit matrix. The bank providing the dataset has its own accounting of the profit matrix which is not available to us at this stage for the consideration of confidentiality. However, we define our profit matrix with rationality from our experience and the information from the Internet and other reports. We may further test the sensitivity of the model to the profit matrix. The relationship between the composition of the profit matrix and model performance should be an interesting research topic in the next step.

The quality of the dataset we have used is not very good with many missing values and errors. The model we constructed here may be improved if we get a better dataset. Taking the advantage of her close connection to China's banking system[3], Southwestern University of Finance and Economics helped us to access the present dataset. For further research, on-going cooperation between the university and banks in China will help us get more up-to-date datasets of better quality. Our future research will include improving the accuracy rate of identifying potential default customers, identifying the most valuable customers, training our model from imbalanced dataset, scoring the applicants, and profit matrix optimization.

As mentioned earlier, the default customers in the dataset account only for a small percentage and the imbalance dataset problem arises. This could be one of negative factors which have lowered the performance of our analytic models. So far, a lot research effort has been put in this aspect (Kubat and Matwin 1997; Chawla et al 2004; Guo and Viktor 2005). From the research angle of computer science, when using imbalanced training set, traditional machine learning algorithms may induce a classifier which will be biased towards the majority class and produce poor predictive accuracy over the minority class. But a biased classifier may be more suitable for certain problems, such as consumer credit assessment. When learning from imbalanced datasets, machine learning algorithms tend to produce high predictive accuracy over the majority class, but poor predictive accuracy over the minority class (Maloof 2003). We have done an extensive exploratory study on this problem with a series of experiments and obtained some interesting findings. Due to the limited scope of this paper, we leave this research to be reported later.

REFERENCES

N.V. Chawla, N. Japkowicz, A. Kolcz (2004), Special issue on learning from imbalanced data sets, ACM SIGKDD Explorations 6 (1). In W.-K. Chen, Linear Networks and Systems, Belmont, CA: Wadsworth, 1993, pp. 123–135.

China State Statistics Bureau (2006), China Statistical Yearbook 2006.

Jike Ge, et al (2006), "Application of Data Mining Technique to Personal Credit Evaluating Model," Computer Technology and Development, December: 172-174.

[3] Southwestern University of Finance and Economics (SWUFE) was previously affiliated to People's bank of China, and more than 30,000 alumni are currently working in China's banking system.

Hongyu Guo, Herna L. Viktor (2004), “Learning from Imbalanced Data Sets with Boosting and Data Generation: The DataBoost-IM Approach”, SIGKDD Explorations 6(1).

M. Kubat, and S. Matwin (1997), “Addressing the curse of imbalanced training sets: One-sided selection”, Proceedings of the 14th International Conference on Machine Learning, Morgan Kaufmann, pp. 179–186.

M.A. Maloof (2003), “Learning when data sets are Imbalanced and when costs are unequal and unknown,” ICML-2003 Workshop on Learning from Imbalanced Data Sets II.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Qiwei Gan
Southwestern University of Finance and Economics
55 Guanghuacun street, Chengdu, Sichuan 610074
P.R.China
Work Phone: +86 135 4005 4300
Fax:+86 28 8735 2040
E-mail: qwgan@swufe.edu.cn, qiweigan@gmail.com
Web: http://www.swufe.edu.cn

Binjie Luo
Southwestern University of Finance and Economics
55 Guanghuacun street, Chengdu, Sichuan 610074
P.R.China
Work Phone: +86 135 4005 4300
E-mail: luobinjie@gmail.com
Web: http://www.swufe.edu.cn

Zhangxi Lin
Texas Tech University
Lubbock, TX 79409-2101, USA
Work Phone: +1 (806) 742-1926
Fax: +1 (806) 742-3193
E-mail: zhangxi.lin@ttu.edu
Web: http://caabi.ba.ttu.edu