Cumulative Overdue Model of Employees’ Housing Accumulation Fund Loans Based on Decision Tree

Li-ping AN\textsuperscript{a,1,*}, Ying-lu ZHOU\textsuperscript{b,2}, Song ZHANG\textsuperscript{a} and Yue-yang XIE\textsuperscript{a}

\textsuperscript{a}Business School, Nankai University, Tianjin 300071, China
\textsuperscript{b}Huisen Middle School, Tianjin 300150, China

1anliping@nankai.edu.cn
2missluluau@126.com
*Corresponding author

Keywords: Housing accumulation fund, Overdue loan, Decision tree, Data mining.

Abstract. To predict and control the risks that the housing accumulation fund loans taken out by employees will be cumulatively overdue, we selected the cumulative overdue rate as the target variable and 31 independent variables based on the data on housing accumulation fund loans of City T over the years. We did some data-mining using the decision tree and found 62 rules that apply to a “low risk” situation and 12 rules that apply to a “high risk” situation. The classification accuracy of the rule set on the testing data set is 81.72%. Therefore when considering an loan application, we can use this rule set to predict the risk that the loan will be overdue and be better informed in credit evaluating.

Introduction

In recent years, with the rapid development of housing accumulation fund service and the improvement of information system, the business is growing at an incredible speed. Though the housing accumulation fund system of City T has accumulated a lot of data, these data are somehow dispersed and independent. Previous analysis on these data offer nothing beyond a simple calculation or collection of numbers. None of them goes deep enough to reveal the hidden information. In a word, the decision-making support provided by the existing report system cannot give satisfactory answers to some frequently asked questions posed by decision-makers such as “What are the factors that will limit the market share of housing accumulation fund service in different periods?” “What kinds of employees will prefer accumulation fund loans?” “What kinds of employees are more likely to have overdue loans?”

The introduction of data-mining has brought decision-making support to a new level. There are only a handful of Chinese researchers who sought to apply data-mining to solve problems in the use and management of accumulation funds. Zhao Liang\cite{1} built a simple linear regression model and a seasonal trend ratio model to analyze the data on the housing accumulation funds of Chongqing. Dai Jiagang\cite{2} built an ARIMA prediction model and a neural network prediction model using the data collected and extracted from accumulation funds since 2000 and gave a comparative analysis of the results. Zhuang Yunhao\cite{3} applied association data-mining to figure out the relations between age, purchasing price and floor area, age and the settlement/withdrawl of housing accumulation funds, as well as working years and the amount of loans. Chen Gang\cite{4} applied data-mining to analyze the fund deposit and loan origination records in the online audit database using cluster analysis and time series analysis and got some practical results. Wu Jianxin\cite{5} built a decision tree to analyze the repaying behaviors of customers based on data from individual housing loans, arrived at a number of classification rules, further analyzed the rules themselves, and gave some information as an reference for the banks. As for other problems in accumulation fund management, some researchers also commented on the data-mining schemes\cite{6,7}. Because of the difference between Chinese and foreign housing systems and data-mining technologies, we have little foreign literature to refer to.
Our aim is to analyze loan data using data-mining methods, identify the characteristics and behavior patterns of employees that are likely to have cumulative overdue loans, and test them with the data on all employees on loans. The results can be applied to the real-time prediction of the possibility that an applicant will have cumulative overdue loans. Based on the prediction, we can have targeted control over loan risks.

Variable Selection and Preprocessing

In this paper, we built a decision tree based on employees’ loan records in the housing accumulation fund administration center of City T to identify the characteristics and behavior patterns of employees likely to have cumulative overdue loans. Based on these patterns, when considering an application, we can predict the risk level of the loan applied for by the employee.

We selected the cumulative overdue rate (which is also our greatest concern) as the target variable. According to our understanding of the business and our statistical analysis, the possibility that an employee will have cumulative overdue loans depends mainly on personal finances, social conventions, social environment, age, gender, working years as well as education and profession backgrounds. So we chose 31 variables as independent variables which determine the target variable, the cumulative overdue rate of loans, and are used to establish a cumulative overdue loan forecasting model based on the decision tree algorithm.

Calculation and Discretization of Target Variable

We used two formulas when calculating the cumulative overdue rate of the target variables:

Cumulative Overdue Rate When the Loans Are Still Being Granted= \( \frac{\text{cumulative overdue times of the borrower}}{\text{how many months of loans that the borrower is supposed to repay till now}} \times 100 \)

Cumulative Overdue Rate When the Loans Have Been Paid Off= \( \frac{\text{cumulative overdue times of the borrower}}{\text{how many months of loans that the borrower has taken out}} \times 100 \)

To build a model, we should first discretize the target variables based on the concept of overdue and the actual condition of the data. We defined loan risk from the perspective of overdue as follows: an accumulation fund loan that has never been overdue or has been overdue less than once every 12 months is a “low risk” loan; an accumulation fund loan that has been overdue once or more every 12 months is a “high risk” loan.

Now in the housing accumulation fund administration center of City T, we have 249617 records of active employees who have taken out loans for over two months. The overdue rate for once every 12 months is \( \frac{1}{12} \times 100\% = 8.333\% \). Therefore we discretize the variable of the 249617 accumulation funds—the cumulative overdue rate as shown in Table 1.

<table>
<thead>
<tr>
<th>Value</th>
<th>Lower Limit</th>
<th>Upper Limit</th>
<th>Sum of Records</th>
<th>Initial Percentage(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Risk</td>
<td>0</td>
<td>8.333</td>
<td>225781</td>
<td>90.451</td>
</tr>
<tr>
<td>High Risk</td>
<td>8.333</td>
<td>-</td>
<td>23836</td>
<td>9.549</td>
</tr>
</tbody>
</table>

Selection and Valuation of Independent Variables

We chose the following 31 variables as the independent variable, namely, (1) ratio of housing accumulation fund loan to social average income, (2) sum of repayment in the first month, (3) use of spouse’s fund, (4) gender of the borrower, (5) ratio of the fund to be contributed by the employer to the fund to be contributed by the employee, (6) fund to be contributed each month (when the application form was submitted, by the borrower), (7) base of the contribution and deposit (when the application form was submitted by the borrower), (8) sum to be repaid each month (now), (9) ratio of the sum to be repaid each month to the total sum, (10) social wage level at the date of borrowing, (11) remaining sum of contribution and deposit, i.e. the remaining sum of fund when the
application form was submitted by the borrower, (12) ratio of the remaining sum of contribution and deposit to average income, (13) whether the borrower has supplementary accumulation fund now, (14) education background, (15) occupation, (16) age at the date of borrowing (by years), (17) age at the date of opening an account for the first time(by years), (18) working years at the date of borrowing, (19) use of the loan, (20) district/county, (21) floor area of the house, (22) selling price of the house, (23) price per square meter of the house, (24) sum to be raised by the borrower himself, (25) proportion of the sum to be raised by the borrower himself, (26) born at City T or not, (27) economic type, (28) employer type, (29) times of job-hopping, (30) way of repayment, (31) months of borrowing by the borrower.

Model Construction and Evaluation

To build a model and test its effectiveness, we needed to divide the 249617 records into two parts---the training set and the testing set. We used the former to summarize the regularities and rules and the latter to test their accuracy and reliability. Based on our experiment and relevant data-mining practicies, we put 70% of the data into the training set to build the model and 30% into the testing set to evaluate the model.

Model Construction Based on Decision Tree Algorithm

We built a model based on the decision tree in SPSS Modeler and set up a rule for every path from the roots to the leaves. The rule set was formed based on the classification rule of “if..., then...”. A conjunctive term of the “if” part was formed for every attribute-value pair on an given path, and the “then” part by the class predictions contained in the leaf nodes. All the rules were saved in the rule base. We obtained 61 rules that apply to a “low risk” situation and 12 rules that apply to a “high risk” situation. However, we were more concerned about the latter. Please see as below.

Rule 1  IF remaining sum of accumulation fund contribution and deposit <= 5665.820 AND months of borrowing <= 239 AND floor area <= 78.670, THEN high risk.

Rule 2  IF remaining sum of accumulation fund contribution and deposit <= 5665.820 AND months of borrowing> 239 AND months of borrowing<= 240 AND house price <= 430,065 AND education background IN ["master and above" “vocational or technical secondary school” “technical school” “university”], THEN high risk.

Rule 3 IF remaining sum of accumulation fund contribution and deposit <= 5665.820 AND months of borrowing> 239 AND months of borrowing <= 240 AND house price <= 430,065 AND education background IN ["elementary school” “college” “high school” “middle school”], THEN high risk.

Rule 4 IF remaining sum of accumulation fund contribution and deposit <= 5665.820 AND months of borrowing > 240 AND months of borrowing <= 343, THEN high risk.

Rule 5 IF remaining sum of accumulation fund contribution and deposit <= 5665.820 AND months of borrowing > 240 AND months of borrowing > 343, THEN high risk.

Rule 6 IF remaining sum of accumulation fund contribution and deposit > 5665.820 AND remaining sum of accumulation fund contribution and deposit <= 9394.260 AND sum to be raised by the borrower himself <= 47,000 AND age at the date of opening an account for the first time <= 27.583, THEN high risk.

Rule 7 IF remaining sum of accumulation fund contribution and deposit> 5665.820 AND remaining sum of accumulation fund contribution and deposit <= 9394.260 AND sum to be raised by the borrower himself > 47,000 AND sum to be raised by the borrower himself <= 103,485 AND district/county IN [“JH” “XQ” “BC” “HP” “HG” “JX” “DG” “HX”], THEN high risk.

Rule 8 IF remaining sum of accumulation fund contribution and deposit> 9394.260 AND remaining sum of accumulation fund contribution and deposit <= 13200.680 AND sum to be raised by the borrower himself > 47,000 AND sum to be raised by the borrower himself <= 79,978, THEN high risk.

Rule 9 IF remaining sum of accumulation fund contribution and deposit > 9394.260 AND remaining sum of accumulation fund contribution and deposit <= 13200.680 AND sum to be raised
by the borrower himself $> 79,978$ AND sum to be raised by the borrower himself $\leq 131,000$ AND
district/county IN [“KF” “BC” “HP” “HQ” “HB” “DG” “JN” “HX”], THEN high risk.

Rule 10 IF remaining sum of accumulation fund contribution and deposit $> 13200.680$ AND
remaining sum of accumulation fund contribution and deposit $\leq 17431.810$ AND months of
borrowing $> 230$ AND months of borrowing $\leq 240$ AND ratio of the fund to be contributed by the
employer to the fund to be contributed by the employee $\leq 20$, THEN high risk.

Rule 11 IF remaining sum of accumulation fund contribution and deposit $> 13200.680$ AND
months of borrowing $> 240$ AND education background IN [“vocational or technical secondary
school” “technical school” “elementary school” “university” “college”], THEN high risk.

Rule 12 IF remaining sum of accumulation fund contribution and deposit $> 13200.680$ AND
months of borrowing $> 240$ AND district/county IN [“HD” “JH” “HP” “NH” “HG” “DG” “JN” “HX”],
THEN high risk.

Model Accuracy Evaluation

The executing results of the decision tree indicated the factors that will influence the cumulative
overdue rate of loans taken out by employees. To test their accuracy and reliability, we must first
analyze and evaluate the decision tree model. After building a decision tree model from the training
data set, we tested it with the testing data set. The classification accuracy of the rule set to the
testing set is 81.72%.

Conclusion

These results indicate that this model is practical and effective in analyzing and predicting the risks
that the accumulation housing fund loans taken out by clients will be cumulatively overdue. It can
help managers better understand the factors that will influence the possibility that a loan will be
overdue so that in their work, they can set targeted policies for high cumulative overdue rate clients,
control asset risks and ensure the security of the accumulation funds contributed and deposited by
employees in City T. This is crucial for improving the competitiveness and management of the
center as well its relationship with clients.

Acknowledgement

This work is supported by Tianjin Social Science Planning Project (No. TJGL15-010)

References


