Data Mining for Automated Assessment of Home Loan Approval

Wanyok Atisattapong(✉), Chollatun Samaimai, Salinla Kaewdulduk, and Ronnagrit Duangdum

Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Bangkok 12120, Thailand
awanyok@tu.ac.th

Abstract. Banks receive large numbers of home loan applications from their own customers and others each day. In this study, we investigated the use of data mining to decide whether or not to extend credit, based on two analytical approaches: Naïve Bayes and decision tree. Four independent factors were considered: the loan period, the net income of the applicant, the size of the loan, and other relevant characteristics of the potential borrower. Models were constructed that produce three outcomes: approval, conditional approval, and rejection. The predictive accuracy of the models was compared, to evaluate the effectiveness of the classifiers and a Kappa statistic was applied, to evaluate the degree of accuracy with which the models predicted the final outcome. The decision tree model performed better on both accuracy and the Kappa statistic. This model had an accuracy of 90% and a Kappa of 0.8140, whereas the Naïve Bayes had an accuracy of 65% and Kappa of 0.3694. We therefore recommend the use of decision tree-based models for home loan ranking. Data mining of the applicants history can support the decision-making of financial organizations, and can also help applicants realistically evaluate their own chances of securing a loan.

Keywords: Data mining · Naïve Bayes · Decision trees · Home loans

1 Introduction

Homebuyers need money to purchase a house and pay for decorating and other expenses. For most individuals, this involves taking out a loan. Banks and financial institutions are willing to offer loans for qualified borrowers. However, the credit risk assessment process takes at least two weeks to process customer data and approve the loan. Since the number of bank customers has significantly increased, the efficiency of credit granting methods must be improved for the benefit of both customers and the banking system.

Many techniques are used to support automatic decision making [1–3]. Approaches include techniques such as fuzzy logic [4,5], logistic regression [6,7], and artificial neural networks [8,9].

S. O. Kuznetsov et al. (Eds.): RCAI 2018, CCIS 934, pp. 11–21, 2018.
https://doi.org/10.1007/978-3-030-00617-4_2

Levy et al. [4] proposed the application of fuzzy logic to commercial loan analysis using discriminant analysis. Mammadli [5] used a fuzzy logic model for retail loan evaluation. The model comprised five input variables: income, credit history, employment, character, and collateral. The single output rated the credit standing as low, medium, or high. Dong et al. [6] proposed a logistic regression model with random coefficients for building credit scorecards. Majeske and Lauer [7] formulated bank loan approval as a Bayesian decision problem. The loan approval criteria were computed in terms of the probability of the customer repaying the loan.

Data mining is an emerging technology in the field of data analysis, and has a significant impact on the classification scheme. Alsultanny [10] proposed Naïve Bayes classifiers, decision tree, and decision rule techniques, for predicting labor market needs. The comparison between these three techniques showed the decision tree to have the highest accuracy. Hamid [11] proposed a model from data mining to classify information from the banking sector and to predict the status of loans. Recently, Bach et al. [12] compared the accuracy of credit default classification of banking clients through the analysis of different entrepreneurial variables, when using decision tree algorithms.

In this work, two data mining approaches, Naïve Bayes and decision tree, were investigated for evaluation of applications for a home loan. Instead of classifying only into two classes, loan approval and rejection, we added a third class in the middle, loan approval with conditions. These conditions are things like search for syndicate partners, adding of collateral assets, and anything else that the borrower needs to clear for loan approval. When the conditions are met, the bank will extend credit.

The remainder of the paper is structured as follows. In Sect. 2, the processes for constructing the models from Naïve Bayes and decision tree are described. In Sect. 3, the performance of both predictive models measured in terms of the accuracy and a Kappa statistic is presented. Finally, the conclusions are outlined in Sect. 4.

2 Proposed Models and Implementation

The data mining process can be divided into four steps, as follows.

1. Prepare the training set, using records that already have a known class label.
2. Build the model by applying the learning algorithm to the training set.
3. Apply the model to a test set containing unclassified data.
4. Evaluate the accuracy of the model.

The flow chart is shown as Fig. 1.

Many variables affect the customer evaluation. Obtaining comprehensive set of actual data from the banks was difficult as such information is considered confidential and thus should be hidden from unauthorized entities. To choose the appropriate variables, bank's lending policies, application forms, and loan review systems were considered.

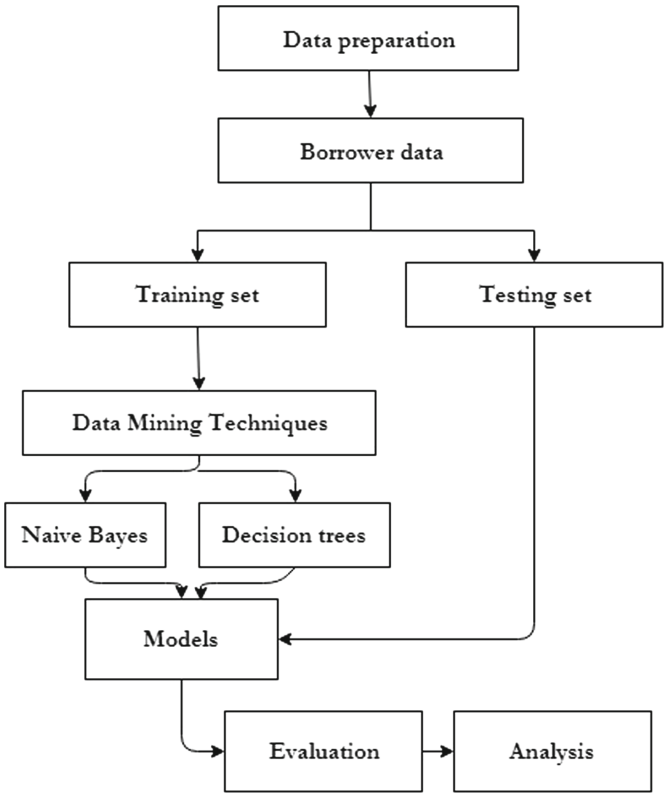

Fig. 1. Data mining process

In this work, the major variables were classified as independent or dependent. There were four independent variables: the loan period, net income, size of loan, and characteristics of the borrower. These are shown in Table 1.

For the period attribute, the timing was limited to a maximum of 35 to reflect bank policies. The 65 year age limit is based on the fact that the customer will usually cease earning at retirement. The net income represents the amount of money remaining after all expenses, interest, and taxes. From the bank's requirement, the net income per month must be greater than or equal to 15,000 Baht. The amount loaned ranged from one to ten million Baht. The relevant characteristics of the borrower were ranked into three levels: 'A', 'B', and 'C', with 'A' being the highest score and 'C' being the lowest. This was evaluated by scoring the customer's loan questionnaire, which collects data on education level, employment, any life insurance policies held, assets, and credit history.

The dependent variables were as follows: 'AP' indicated that the loan was approved. 'AC' indicated that the loan was conditionally approved. 'DN' indicated that the loan was rejected. The target class of the training set is based on the bank policy, which requires that a borrower taking a loan of one million Baht is able to make payments of not less than 7,000 Baht per month.

Table 1. Attribute description

No.	Attribute	Description	Unit	Data type
1	Period	Timing of disbursements (Calculated by 65 − the customer age ≤ 35)	Year	Numeric
2	Income	Excess of revenues over expenses	Baht	Numeric
3	Size of loan	Amount of loan required	Million Baht	Numeric
4	Character	Relevant characteristics of the borrower	Grade	Nominal

We first simulated an original dataset of 200 customers. This was divided into two sets, a training set comprising 80% of all data, and a testing set comprising 20%. Table 2 shows the original data set, which was designed to simulate the customer conditions. Each customer's data could be updated. As an example of a single data point, customer number 1 has a loan period equal to five years, a net income per month of 18,000 Baht, would like to take a loan for two million Baht, and has a good credit rating. The decision is that the loan is rejected.

Table 2. Dataset

Customer no.	Period	Income	Size of loan	Character	Class
1	5	18000	2	A	DN
2	20	50000	2.5	B	AP
3	26	15000	1	C	AC
4	16	25000	1.6	B	AP
5	10	20000	2.3	B	DN
6	32	45000	3	C	AP
7	33	55000	3.6	A	AP
8	35	65000	4.3	B	AP
9	35	75000	5	C	AP
10	35	85000	5.6	A	AP
...	...	...	...	...	...
200	24	27000	8.1	A	DN

Two models, Naïve Bayes and decision tree, were implemented using Weka Data Mining software version 3.9.2 [13] as shown in Fig. 2.

To investigate the appropriate size for the training set, we randomly chose data sets of ten different sizes, as shown in Table 3. Next, ten Naïve Bayes models were constructed and their accuracy computed automatically. After choosing the size that yielded the best accuracy, we used the Weka software to build a decision tree model of the same size to observe the nodes of the tree.

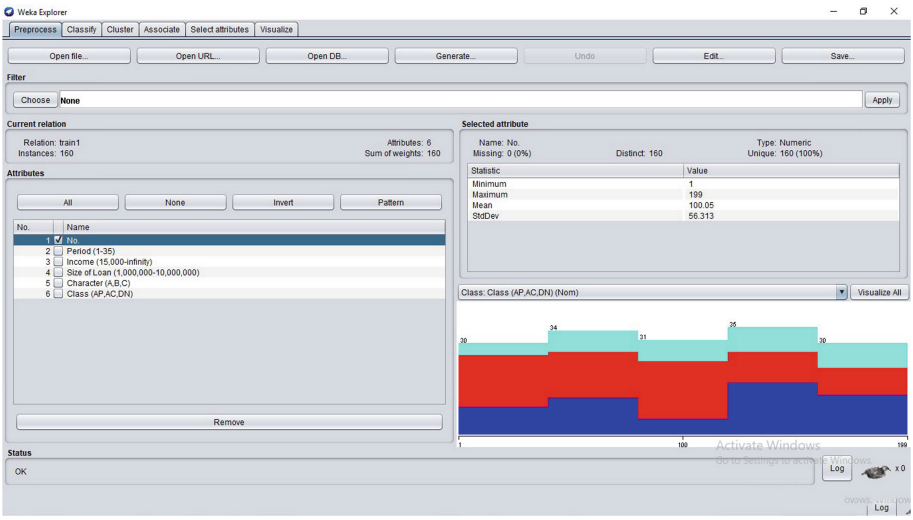

Fig. 2. Training set imported to Weka software

Table 3. Sizes of training sets

No.	Number of cases
1	16
2	32
3	48
4	64
5	80
6	96
7	112
8	128
9	144
10	160

3 Results

Classification accuracy is a standard metric used to evaluate classifiers [14]. Accuracy is simply the percentage of instances in which the method predicts the true outcome. The Kappa statistic [15] is used to measure the agreement of a prediction with the actual outcome. A Kappa of 1 indicates perfect agreement, whereas a kappa of 0 indicates agreement equivalent to chance. The scale of the Kappa is shown in Table 4.

Table 5 shows that the training set of 80 cases provided the best accuracy. The selected training dataset used to construct the decision tree is shown in

Table 4. Interpretation of Kappa

Kappa	Agreement
<0	Less than chance agreement
0.01–0.20	Slight agreement
0.21–0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–0.99	Almost perfect agreement

Table 5. Accuracy and training set size

No.	Number of cases	Accuracy
1	16	87.50
2	32	87.50
3	48	89.60
4	64	89.10
5	80	95.00
6	96	90.63
7	112	90.18
8	128	87.50
9	144	86.80
10	160	89.38

Tables 6 and 7 shows the accuracy and Kappa statistic of the Naïve Bayes and the decision tree models. The results showed the decision tree model to provide more precise results.

Figure 3 shows the structure of the tree obtained. As can be observed, the initial node of the tree was the net income of the applicant. This showed the first factor affecting the decision to be income. The other attributes were ranked as shown. Rectangular boxes show the final classification. The numbers in the box count the correctly and incorrectly categorized cases. In total, only four cases were incorrectly classified.

To find the appropriate class for a given customer, we start with the value at the root of the tree and keep following the branches until a leaf is reached. For example, the attribute values of customer number 1 are as follows: Income = 18,000, Size of Loan = 2, Character = A, and Period = 5. To classify this case, we start at the root of the tree in Fig. 3, which is labeled 'Income', and follow the true branch (Income $\leq$ 50,000) to the node 'Size of loan'. As a loan of 2 million Baht is less than the 3.3 million Baht threshold, Character is considered. Character 'A' leads finally to a leaf labeled 'AP', indicating that the bank can proceed with a loan approval.

Table 6. Example training set

No.	Customer no.	Period	Income	Size of loan	Character	Class
1	3	26	15000	1	C	AC
2	8	35	65000	4.3	B	AP
3	10	35	85000	5.6	A	AP
4	11	35	95000	6.3	B	AP
5	14	19	120000	8	B	AP
6	19	24	75000	6	A	AP
7	22	5	31000	5.2	A	DN
8	23	16	25000	7.3	A	DN
9	25	20	19000	1.9	B	AC
10	26	25	48000	3.2	C	AP
11	29	5	320000	9.8	B	AC
12	32	10	62000	1.9	C	AP
13	38	26	16000	3	C	DN
14	39	22	27000	2.5	A	AP
15	40	15	39000	6	A	DN
...	...	...	...	...	...	
71	175	16	55000	7	C	DN
72	176	24	65000	4.5	B	AP
73	185	20	23000	3	B	AC
74	187	21	64000	5.6	C	AP
75	188	19	29000	2.2	C	AC
76	189	30	15000	3.5	A	DN
77	192	20	46000	7.1	B	DN
78	194	18	25000	3	B	AC
79	195	19	35000	4.2	B	AC
80	196	19	25000	8	C	DN

Table 7. Accuracy and Kappa of two models from the 80 case training set

Technique	Accuracy	Kappa
Naïve Bayes	67.5%	0.4669
Decision tree	95.0%	0.9191

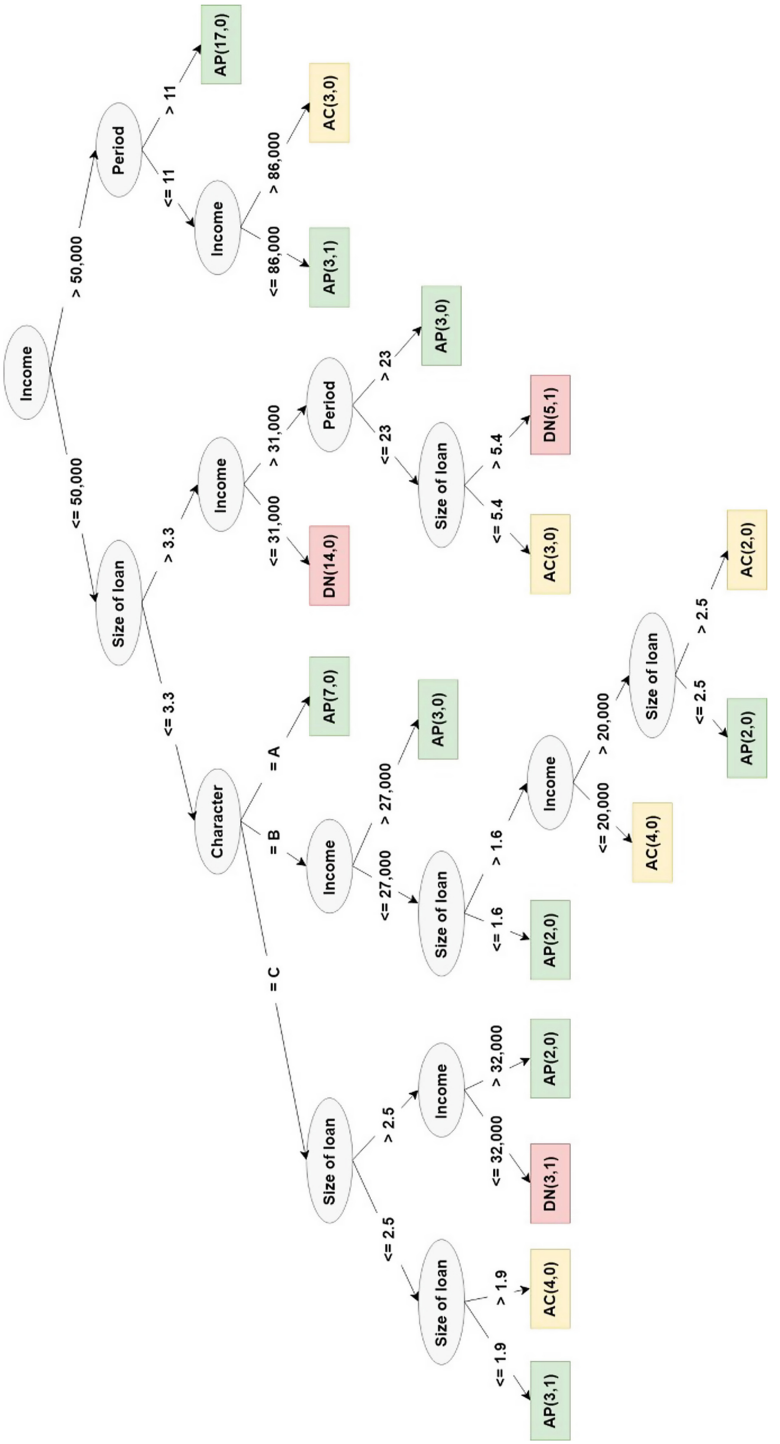

Fig. 3. Decision tree for the training set

Table 8. Test set

No.	Customer no.	Period	Income	Size of loan	Character	Target class
1	1	5	18000	2	A	AP
2	13	16	75000	8.8	A	AP
3	24	28	28000	6.4	A	DN
4	27	27	15000	1	C	AP
5	48	30	34000	3	A	AP
6	95	13	16000	3	A	AP
7	105	33	16000	1.4	A	AP
8	121	20	18000	4.6	C	DN
9	124	25	18000	2	C	AC
10	128	10	190000	8.7	B	AP
11	131	20	24000	3.3	A	AP
12	146	20	23000	5.2	C	DN
13	159	33	15000	1.5	C	AP
14	165	30	37000	4	C	AP
15	169	20	16000	2	B	AC
16	171	20	61000	9.2	C	AP
17	174	25	31000	3.3	B	AP
18	180	22	20000	4	C	DN
19	182	20	79000	4	B	AP
20	197	20	23000	3.3	C	AC

Table 9. Accuracy and Kappa of two models from the 20 case testing set

Technique	Accuracy	Kappa
Naïve Bayes	65%	0.8140
Decision tree	90%	0.3694

Next, the two models were applied to a testing set. Since the full dataset was split 80:20 into training and testing sets, and 80 case was determined to be the optimal size of the training set, 20 case was chosen for the testing set (Table 8). The accuracy and Kappa statistic from the two techniques are shown in Table 9. The decision tree model again yielded better accuracy.

4 Conclusions

In this study, two algorithms, Naïve Bayes and decision tree, were used to build predictive models and to classify applications for home loans. After adding a median approval class, the models were implemented in Weka and used to classify

applicants base on simulated data. The results showed the decision tree model had both a higher accuracy and Kappa statistic than the Naïve Bayes model. The decision tree technique was efficient in predicting the values of instances that were not in the training set. It was able to deal with missing attribute values, because in some case the timing of disbursement is not critical. By applying this technique, banks can improve their predictions of which clients will have a higher chance of getting their loan application approved. This will expand access to home loans.

Future work should investigate the use of a wider range of data mining techniques. As the performance of an algorithm is known to be dependent on the domain and type of the dataset, it would be interesting to explore other classification algorithm such as those used in machine learning.

Acknowledgment. We would like to thank Mr. John Winward for comments and suggestions on the manuscript.

References

1. Chambers, M., Garriga, C., Schlagenhauf, D.: The loan structure and housing tenure decisions in an equilibrium model of mortgage choice. Rev. Econ. Dyn. **12**, 444–468 (2009)
2. Trönnberg, C.-C., Hemlin, S.: Lending decision making in banks: a critical incident study of loan officers. Eur. Manag. J. **147**, 362–372 (2014)
3. Lee, C.C., Ho, Y.M., Chiu, H.Y.: Role of personal conditions, housing properties, private loans, and housing tenure choice. Habitat Int. **53**, 301–311 (2016)
4. Levy, J., Mallach, E., Duchessi, P.: A fuzzy logic evaluation system for commercial loan analysis. OMEGA. Int. J. Manag. Sci. **19**, 651–669 (1991)
5. Mammadli, S.: Fuzzy logic based loan evaluation system. Proc. Comput. Sci. **102**, 495–499 (2016)
6. Dong, G., Lai, K.K., Yen, J.: Credit scorecard based on logistic regression with random coeggicients. Proc. Comput. Sci. **1**, 2463–2468 (2012)
7. Majeske, K.D., Lauer, T.W.: The bank loan approval decision from multiple perspectives. Expert Syst. Appl. **40**, 1591–1598 (2013)
8. Malhotra, R., Malhotra, D.K.: Evaluating consumer loans using neural networks. OMEGA. Int. J. Manag. Sci. **31**, 83–96 (2003)
9. Doori, M.A., Beyrouti, B.: Credit scoring model based on back propagation neural network using various activation and error function. Int. J. Comput. Sci. Netw. Secur. **14**, 16–24 (2014)
10. Alsultanny, Y.: Labor market forecasting by using data mining. Proc. Comput. Sci. **18**, 1700–1709 (2013)
11. Hamid, A.J., Ahmed, T.M.: Developing prediction model of loan risk in banking using data minding. Mach. Learn. Appl.: Int. J. **3**, 1–9 (2016)
12. Bach, M.P., Zoroja, J., Jakovi, B., and Sarlija, N.: Selection of variables for credit risk data mining models: preliminary research. In: 40th International Convention on Information and Communication Technology, Electronics and Microelectronics, pp. 1367–1372 (2017)
13. Witten, I.H., Frank, E., Hall, M., Pal, C.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, Burlington (2016)

14. Perlich, C., Provost, F., Simonoff, J.: Tree induction vs. logistic regression: a learning-curve analysis. J. Mach. Learn. Res. **4**, 211–255 (2003)
15. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**, 159–174 (1977)