Personal bankruptcy prediction using decision tree model

Sharifah Heryati Syed Nor
Department of Economics and Financial Studies, Universiti Teknologi Mara, Bandar Puncak Alam, Malaysia

Shafinar Ismail
Department of Finance, Universiti Teknologi Mara Melaka, Melaka, Malaysia, and

Bee Wah Yap
Advanced Analytics Engineering Centre, FSKM, Universiti Teknologi Mara, Shah Alam, Malaysia

Received 10 August 2018
Revised 10 August 2018
Accepted 26 September 2018

Abstract

Purpose – Personal bankruptcy is on the rise in Malaysia. The Insolvency Department of Malaysia reported that personal bankruptcy has increased since 2007, and the total accumulated personal bankruptcy cases stood at 131,282 in 2014. This is indeed an alarming issue because the increasing number of personal bankruptcy cases will have a negative impact on the Malaysian economy, as well as on the society. From the aspect of individual's personal economy, bankruptcy minimizes their chances of securing a job. Apart from that, their account will be frozen, lost control on their assets and properties and not allowed to start any business nor be a part of any company's management. Bankrupts also will be denied from any loan application, restricted from travelling overseas and cannot act as a guarantor. This paper aims to investigate this problem by developing the personal bankruptcy prediction model using the decision tree technique.

Design/methodology/approach – In this paper, bankrupt is defined as terminated members who failed to settle their loans. The sample comprised of 24,546 cases with 17 per cent settled cases and 83 per cent terminated cases. The data included a dependent variable, i.e. bankruptcy status (Y = 1(bankrupt), Y = 0 (non-bankrupt)) and 12 predictors. SAS Enterprise Miner 14.1 software was used to develop the decision tree model.

Findings – Upon completion, this study succeeds to come out with the profiles of bankrupts, reliable personal bankruptcy scoring model and significant variables of personal bankruptcy.

Practical implications – This decision tree model is possible for patent and income generation. Financial institutions are able to use this model for potential borrowers to predict their tendency toward personal bankruptcy.

Social implications – Create awareness to society on significant variables of personal bankruptcy so that they can avoid being a bankrupt.

The authors would like to express their gratitude to Agensi Kaunseling dan Pengurusan Kredit (AKPK), Kuala Lumpur, for their continuous support and cooperation.

Journal of Economics, Finance and Administrative Science
Vol. 24 No. 47, 2019
pp. 157-170
Emerald Publishing Limited
2077-1886
DOI 10.1108/JEFAS-08-2018-0076

Originality/value – This decision tree model is able to facilitate and assist financial institutions in evaluating and assessing their potential borrower. It helps to identify potential defaulting borrowers. It also can assist financial institutions in implementing the right strategies to avoid defaulting borrowers.

Keywords Data mining, Credit scoring, Decision tree model, Personal bankruptcy, Random undersampling

Paper type Research paper

1. Introduction

Personal bankruptcy cases in Malaysia have been on an upward trend since 2007. In Malaysia, a debtor is declared a bankrupt, pursuant to an adjudication order made by the High Court against the debtor if he/she is unable to pay his/her debts of at least RM30,000 (Malaysia Department of Insolvency, 2017). The Credit Counseling and Debt Management Agency (Agensi Kaunseling dan Pengurusan Kredit, AKPK) reportedly claimed in Utusan Online (Zainon, 2016) that the reasons for financial difficulty include:

- improper financial planning (49.7 per cent);
- business failure (15.2 per cent);
- cost of living after retirement (11.7 per cent);
- higher medical cost (11 per cent);
- unemployment (9.5 per cent); and
- other miscellaneous reasons (2 per cent).

Meanwhile, Datuk Seri Azalina Othman Said, a minister at the Prime Minister's Department, stated that personal bankruptcy cases often occur due to hire purchase loans, personal loans and housing loans. She also added that a total number of 22,581 personal bankruptcy cases, as recorded by the Insolvency Department between 2012 and September 2016, had involved individuals aged between 25 and 34 years old (Bernama, 2016).

In 2014, the Insolvency Department of Malaysia reported that personal bankruptcy cases have increased, from 13,238 cases in 2007 to 22,351 cases in 2014. This showed an increase of 68.8 per cent, with the total accumulated personal bankruptcy cases of 131,282 in 2014. This is alarming because, if the number of personal bankruptcy cases continues to increase, it will have a negative impact on the Malaysian economy and on the society. From the aspect of individuals' personal economy, bankruptcy minimizes their chances of securing a job.

As one of the efforts taken to curb the increasing household debt which mainly leads to personal bankruptcy, Bank Negara Malaysia has set up a debt management agency. This agency is an avenue for potential individual borrowers and distressed borrowers to acquire assistance and seek advice in managing their debts and finances. Thus, this paper illustrates the application of data mining techniques to determine the conditional probability of a borrower belonging to a class (bankrupt or non-bankrupt) using the decision tree model. The findings from this study are useful for various parties to make decisions and management agencies, hire-purchase companies and credit companies. These actions are important to avoid or to prevent default payment, bad debts and personal bankruptcy. Therefore, the objectives of this paper are to identify the significant predictors and to determine the conditional probability of a borrower belonging to a class (bankrupt or non-bankrupt) using the decision tree model.

This paper is organized as follows: Section 2 provides a review on personal bankruptcy studies and the methods used in the analysis. The methodology for the decision tree model is covered in Section 3. The results are discussed in Section 4. Finally, Section 5 concludes the paper.

2. Literature review

Personal bankruptcy issues have been a pressing concern to the governments, bankers, creditors and financial researchers in recent years. Offering important body of information to financial institutions helps them evaluate the risk of their credit portfolio in a timely manner, as well as help them to formulate their respective risk management strategies (Min and Lee, 2008). Erroneous decision-making by financial institutions will likely lead to financial crises and distress. Therefore, bankruptcy prediction and credit scoring are very important when making financial decisions (Tsai, 2014). Moreover, the efforts to reduce the probability of a customer defaulting, which predicts customer risk, support and help maximize the expected profit from that customer, especially for banks and credit companies (Abdou and Pointon, 2011).

Fisher (2005) found that household heads who are older, white, less educated and in poor health are significantly more likely to file for bankruptcy. On the one hand, Agarwal *et al.* (2011) indicated that mobility, rural residency, homeownership, marital status and age are the significant predictors for personal bankruptcy. In addition, Jullamon (2012) revealed that the bankrupts are usually those in their fifties, low-income earners, unemployed, without property ownership and defaulted on the loan agreement.

Apart from household characteristics, technologies and regulations have also impacted the level of bankruptcy. Livshits *et al.* (2010) reported that credit market innovations or changes that reduce the cost of bankruptcy (such as filing fees), together with the cost of borrowing (such as interest rates) play an essential role toward the rise in bankruptcy cases. The banking deregulation and technology changes play an important role in the increase of consumer bankruptcy. Debts, defaults and state laws have also contributed to bankruptcy filing (Dick and Lehnert, 2010). Bland *et al.* (2007) found that the causes of bankruptcy include:

- overspending on credit cards;
- poor financial management;
- problems with house payment or property taxes; and
- loss of jobs.

Azaizeh (2010) also pointed out that those with higher credit card debts, older household heads and have bad payment history are more likely to file for bankruptcy. Correspondingly, Zhu (2013) reported that households that file for bankruptcy have spent beyond their means by extending their credit facilities. Meanwhile, Dawsey (2014) indicated that borrowers' number of loans significantly increase the probability of bankruptcy. Additionally, anti-harassment, garnishment and exemption law impact the borrowers' choice of informal bankruptcy, formal bankruptcy and repayment.

Despite the continual growth of personal bankruptcy cases since 2007, this issue remains an under-investigated research area. To the best of the researcher's knowledge, only six studies were conducted on personal bankruptcy issues in Malaysia. Selvanathan *et al.* (2016) explored the factors affecting personal bankruptcy cases based on a sample of Klang Valley residents using Pearson correlation coefficient and multiple regression analysis. They found that there were positive relationships between money management, financial literacy and non-performing loan with personal bankruptcy. Nair *et al.* (2016) applied logistic regression to identify the determinants of civil servants' bankruptcy probability in Malaysia. They found that asset ownership, attitude toward debts and financial management practices are significant predictors of personal bankruptcy. Noordin *et al.* (2012) investigated the relationship between knowledge about credit card and knowledge about bankruptcy, and the relationship between lifestyle and bankruptcy. The method they deployed was descriptive statistics, namely, correlation. They found that there is a negative relationship between knowledge about credit card and bankruptcy due to credit card debts, as well as there is no relationship between lifestyle and bankruptcy due to credit card debts.

Eaw *et al.* (2014) focused on the causality factors of personal bankruptcy, and later, Eaw *et al.* (2015) examined the moderating effects of psychographic factors on the association between financial numeracy and financial management outcome using structural equation modeling. They found that good financial numeracy leads to a better financial management outcome, and less likely to cause financial stress and bankruptcy. In their 2015 research, they found that there was a positive relationship between financial numeracy and financial management outcome. Individuals with low materialistic value were also found to be more likely to avoid high borrowing when they have high level of financial numeracy. Othman *et al.* (2015) studied the profiles of bankrupts, sources of bankruptcy, the loan types leading to bankruptcy and financial status before bankruptcy. They analyzed their data using descriptive statistics and independent samples *t*-test. Their findings revealed that poor financial management, overspending and failure in business are the reasons for bankruptcy.

Personal bankruptcy prediction has been an increasing concern, both to the industry and for the purpose of academic investigation, as it often results in significant losses to the creditors (Xiong *et al.*, 2013). Financial distress and crises deeply affect the shareholders, managers, workers, lenders, suppliers, clients, communities and governments. Therefore, it is very important to develop financial distress or personal bankruptcy prediction model (Tsai, 2014). Apart from the development of the bankruptcy prediction model, the accuracy of personal bankruptcy prediction is a major issue to the shareholders, creditors, policy makers and business managers (Olson *et al.*, 2012). According to Daskalaki *et al.* (2003), insolvency or bankruptcy prediction makes sense in business terms if it is applied early enough to be of any use for the company.

Credit scoring has been regarded as a core appraisal tool by different institutions for the last few years and has been widely investigated in different areas, such as finance and accounting (Abdou and Pointon, 2011). The credit risk model provides important information to help financial institutions formulate good risk-management strategies (Min and Lee, 2008). The credit risk model evaluates the risk in lending to a particular client as the model estimates the probability that an applicant, with any given credit score, will be “good” or “bad” (Řezáč and Řezáč, 2011). It also quantifies the risks associated with credit requests by evaluating the social, demographic, financial and other data collected at the time of the application (Paleologo *et al.*, 2010). A broad scope of statistical techniques are used in building credit scoring models. Techniques, such as weight-of-evidence measure, discriminant analysis, regression analysis, probit analysis, logistic regression, linear programming, Cox’s proportional hazard model, support vector machines, neural networks, decision trees, K-nearest neighbor (K-NN), genetic algorithms and genetic programming are all widely used in building credit scoring models by statisticians, credit analysts, researchers, lenders and computer software developers (Abdou and Pointon, 2011).

Decision tree (DT) is also commonly used in data mining. It is frequently used in the segmentation of population or predictive models. It is also a white box model that indicates the rules in a simple logic. Because of the ease of interpretation, it is extremely popular in assisting users to understand various aspects of their data (Choy and Flom, 2010). DTs are produced by algorithms that identify various ways of splitting a data set into branch-like segments. It has a set of rules for dividing a large collection of observations into smaller homogeneous groups with respect to a particular target variable. The target variable is usually categorical, and the DT model is used either to calculate the probability that a given record belongs to each of the target category or to classify the record by assigning it to the most likely category (Ville, 2006).

Several studies have shown that DT models can be applied to predict financial distress and bankruptcy. For example, Chen (2011) proposed a model of financial distress prediction that compares DT classification to logistic regression (LR) technique using samples of 100 Taiwan firms listed on the Taiwan Stock Exchange Corporation. The DT classification approach had better prediction accuracy than the LR approach.

Irimia-Dieguez *et al.* (2015) developed a bankruptcy prediction model by deploying LR and DT technique on a data set provided by a credit agency. They then compared both models and confirmed that the performance of the DT prediction had outperformed LR prediction. Gepp and Kumar (2015) indicated that financial distress and the consequent failure of a business are usually extremely costly and disruptive event. Therefore, they developed a financial distress prediction model by using the Cox survival technique, DT, discriminant analysis and LR. The results showed that DT is the most accurate in financial distress prediction. Mirzei *et al.* (2016) also believed that the study of corporate default prediction provides an early warning signal and identify areas of weaknesses. Accurate corporate default prediction usually leads to numerous benefits, such as cost reduction in credit analysis, better monitoring and an increased debt collection rate. Hence, they used DT and LR technique to develop a corporate default prediction model. The results from the DT were found to best suit the predicted corporate default cases for different industries.

3. Research method

This study involved a data set obtained from an authorized debt management agency. The data consisted of settled members and terminated members. Settled members were those who managed to settle their loans, while terminated were those who were unable to pay their loans. There were 4,174 settled members and 20,372 terminated members. The total sample size was 24,546 with 17 per cent (4,174) settled and 82.99 per cent (20,372) terminated cases. It is noted here that the negative instances belong to the majority class (terminated) and the positive instances belong to the minority class (settled); imbalanced data set. According to Akosa (2017), the most commonly used classification algorithms data set (e.g. scorecard, LR and DT) do not work well for imbalanced data set. This is because the classifiers tend to be biased toward the majority class, and therefore perform poorly on the minority class. He added, to improve the performance of the classifiers or model, downsampling or upsampling techniques can be used. This study deployed the random undersampling technique. The random undersampling technique is considered as a basic sampling technique in handling imbalanced data sets (Yap *et al.*, 2016). Random undersampling (RUS), also known as downsampling, excludes the observations from the majority class to balance with the number of available observations in the minority class. The RUS was used by randomly selecting 4,174 cases from the 20,372 terminated cases. This RUS process was done using IBM Statistical package for the Social Science (SPSS) software. Therefore, the total sample size was 8,348 with 50 per cent (4,174) representing settled cases and 50 per cent (4,174) representing terminated cases for the balanced data set. This study used both sample sizes for further analysis to see the differences in the results of the statistical analyses of this study.

The data covered the period from January 01, 2010 to October 31, 2015, which were received in Excel files. Data cleaning was the first step to remove outliers and redundant data. Once the data cleaning process was completed, the Excel data file was converted into a SAS file using SAS 9.4 software. The LR, scorecard and DT models were run using the SAS Enterprise Miner 14.1 software.

A DT model consists of a set of rules for dividing a large heterogeneous population into smaller, more homogeneous groups with respect to a particular target variable. The target variable is usually categorical, and the DT model is used either to calculate the probability that a given record belongs to each of the categories or to classify the records by assigning it to the most likely class (Linoff and Berry, 2011).

According to Ville (2006), the Gini index is used as a measure for node impurity. Linoff and Berry (2011) mentioned that purity measures for evaluating splits for categorical target variables include the Gini index. Sarma (2017) added that, when the target variable is binary, the impurity reduction achieved by the split is measured by Gini index. Hence, this study used Gini index as the splitting criteria. The Gini index compares impurity reduction on the splits and selects the one that achieves the greatest impurity reduction as the best split (Sarma, 2017). Gini is one of the popular splitting criteria in selection of attributes (or variables) in building the DT. The variables are ranked based on their Gini values. The Gini splitting criteria was used to develop the DT model.

For a binary split (a split with two nodes) for variable X, the Gini coefficient for each variable is calculated as follows (Linoff and Berry, 2011):

$$G_L = Gini_{X(left\,node)} = (p_{good})^2 + (p_{bad})^2$$

$$G_R = Gini_{X(right\ node)} = (p_{good}\ \ )^2 + (p_{bad})^2$$

where p is the proportion of good and bad cases in the respective node. Then,

$$Gini_X = \left(\frac{n_L}{n}\right) * G_L + = \left(\frac{n_R}{n}\right) * G_R \ \text{ where } n_L + n_R = n$$

An example of the Gini coefficient calculation based on Nodes 2 and 3 (refer to Figure 2) is presented below:

$$G_L = \ (0.7198)^2 + (0.2802)^2\, 0.5966$$

$$G_R = (0.9297)^2 + (0.0703)^2 = 0.8692$$

$$Gini_{Outstandingloan} = \left(\frac{3498}{3498 + 3867}\right) * 0.5966 + \left(\frac{3867}{3867 + 3498}\right) * 0.8692 = 0.73977 \approx 0.74$$

*outstanding loan = LoanOutstandingRange (variable name in the SAS system)

Gini is a measure of purity and ranges from 0 to 1. A Gini close to 1 indicates that the split is pure (that is the variable can split the good and bad cases based on that variable). The variable with the highest Gini score will be used in the first split. Successive splits will be based on Gini score too.

4. Results

Based on Figure 1, these are the decision rules of the DT model (imbalanced data):

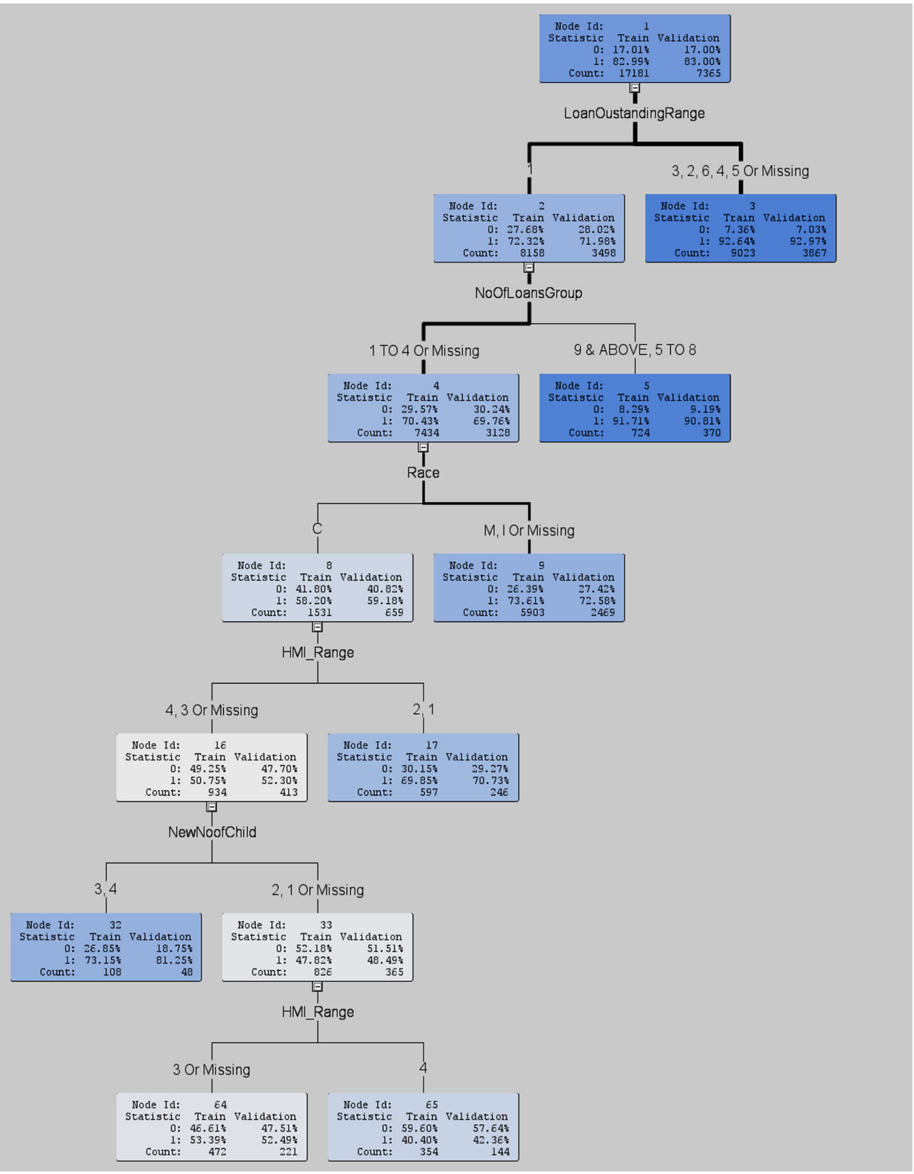

Figure 1.
The decision tree model for imbalanced data and number of decision rules

- If the borrowers have outstanding loan RM30k and above, then their status is bankrupt (Node 3).
- If the borrowers have number of loans between five and eight or nine and above, outstanding loan is RM29,999 and below, then their status is bankrupt (Node 5).

- If the borrowers are Malay or Indian, number of loans are between one and four, outstanding loan is RM29,999 and below, then their status is bankrupt (Node 9).
- If the borrowers are Chinese, number of loans are between one and four, outstanding loan is RM29,999 and below, household monthly income is RM1,000 and below or between RM1,001 and RM2,000, then their status is bankrupt (Node 17).
- If the borrowers are Chinese, number of loans are between one and four, number of children are between four and six or seven and above, outstanding loan is RM29,999 and below, household monthly income is between RM2,001 and RM3,000 or RM3,001 and above, then their status is bankrupt (Node 32).
- If the borrowers are Chinese, number of loans are between one and four, number of children are between zero and four, outstanding loan is RM29,999 and below, household monthly income is between RM2,001 and RM3,000, then their status is bankrupt (Node 64).
- If the borrowers are Chinese, number of loans are between one and four, number of children are between zero and four, outstanding loan is RM29,999 and below, household monthly income is above RM3,000, then their status is non-bankrupt (Node 65).

Figure 1 displays the DT model for imbalanced data and number of decision rules.

From the above decision rules, the profiles of a bankrupt from imbalanced data set are:

- a borrower who has outstanding loan of RM30k and above;
- a borrower who has the number of loans of five and above;
- a borrower who is a Malay or Indian with number of loans between one and four; and
- a borrower who is a Chinese with household monthly income of RM2,000 and below.

Figure 2 displays the DT model for balanced data and number of decision rules.

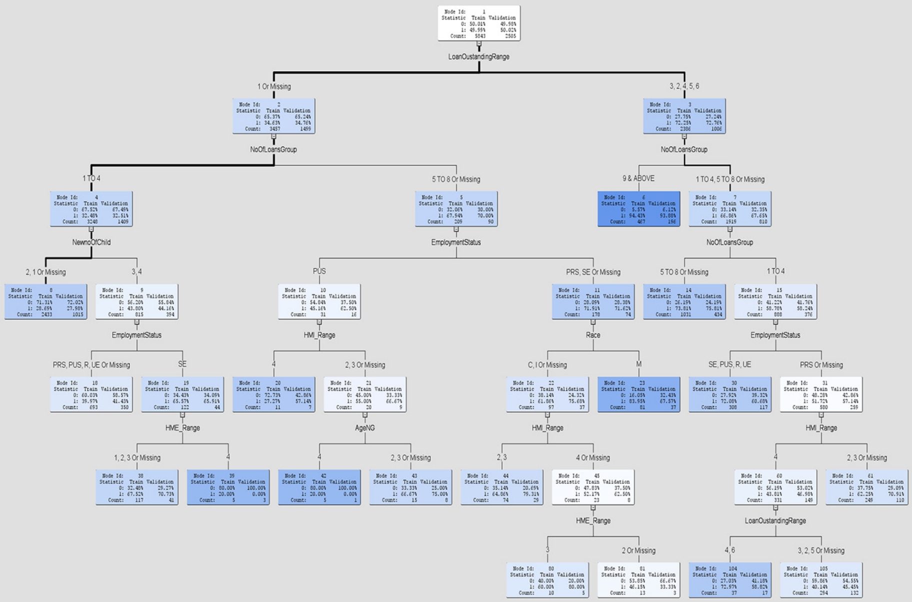
Node Id: 1 Statistic Train Validation 0: 50.01% 49.98% 1: 49.99% 50.02% Count: 5843 2505
LoanOustandingRange
1 Or Missing
3, 2, 4, 5, 6
Node Id: 2 Statistic Train Validation 0: 65.37% 65.24% 1: 34.63% 34.76% Count: 3457 1499
Node Id: 3 Statistic Train Validation 0: 27.75% 27.24% 1: 72.25% 72.76% Count: 2386 1006
NoOfLoansGroup
NoOfLoansGroup
1 TO 4
5 TO 8 Or Missing
9 & ABOVE
1 TO 4, 5 TO 8 Or Missing
Node Id: 4 Statistic Train Validation 0: 67.52% 67.49% 1: 32.48% 32.51% Count: 3248 1409
Node Id: 5 Statistic Train Validation 0: 32.06% 30.00% 1: 67.94% 70.00% Count: 209 90
Node Id: 6 Statistic Train Validation 0: 5.57% 6.12% 1: 94.43% 93.88% Count: 467 196
Node Id: 7 Statistic Train Validation 0: 33.14% 32.35% 1: 66.86% 67.65% Count: 1919 810
NewnoOfChild
EmploymentStatus
NoOfLoansGroup
2, 1 Or Missing
3, 4
PUS
PRS, SE Or Missing
5 TO 8 Or Missing
1 TO 4
Node Id: 8 Statistic Train Validation 0: 71.31% 72.02% 1: 28.69% 27.98% Count: 2433 1015
Node Id: 9 Statistic Train Validation 0: 56.20% 55.84% 1: 43.80% 44.16% Count: 815 394
Node Id: 10 Statistic Train Validation 0: 54.84% 37.50% 1: 45.16% 62.50% Count: 31 16
Node Id: 11 Statistic Train Validation 0: 28.09% 28.38% 1: 71.91% 71.62% Count: 178 74
Node Id: 14 Statistic Train Validation 0: 26.19% 24.19% 1: 73.81% 75.81% Count: 1031 434
Node Id: 15 Statistic Train Validation 0: 41.22% 41.76% 1: 58.78% 58.24% Count: 888 376
EmploymentStatus
HMI_Range
Race
EmploymentStatus
PRS, PUS, R, UE Or Missing
SE
4
2, 3 Or Missing
C, I Or Missing
M
SE, PUS, R, UE
PRS Or Missing
Node Id: 18 Statistic Train Validation 0: 60.03% 58.57% 1: 39.97% 41.43% Count: 693 350
Node Id: 19 Statistic Train Validation 0: 34.43% 34.09% 1: 65.57% 65.91% Count: 122 44
Node Id: 20 Statistic Train Validation 0: 72.73% 42.86% 1: 27.27% 57.14% Count: 11 7
Node Id: 21 Statistic Train Validation 0: 45.00% 33.33% 1: 55.00% 66.67% Count: 20 9
Node Id: 22 Statistic Train Validation 0: 38.14% 24.32% 1: 61.86% 75.68% Count: 97 37
Node Id: 23 Statistic Train Validation 0: 16.05% 32.43% 1: 83.95% 67.57% Count: 81 37
Node Id: 30 Statistic Train Validation 0: 27.92% 39.32% 1: 72.08% 60.68% Count: 308 117
Node Id: 31 Statistic Train Validation 0: 48.28% 42.86% 1: 51.72% 57.14% Count: 580 259
HME_Range
AgeNG
HMI_Range
HMI_Range
1, 2, 3 Or Missing
4
4
2, 3 Or Missing
2, 3
4 Or Missing
4
2, 3 Or Missing
Node Id: 38 Statistic Train Validation 0: 32.48% 29.27% 1: 67.52% 70.73% Count: 117 41
Node Id: 39 Statistic Train Validation 0: 80.00% 100.00% 1: 20.00% 0.00% Count: 5 3
Node Id: 42 Statistic Train Validation 0: 80.00% 100.00% 1: 20.00% 0.00% Count: 5 1
Node Id: 43 Statistic Train Validation 0: 33.33% 25.00% 1: 66.67% 75.00% Count: 15 8
Node Id: 44 Statistic Train Validation 0: 35.14% 20.69% 1: 64.86% 79.31% Count: 74 29
Node Id: 45 Statistic Train Validation 0: 47.83% 37.50% 1: 52.17% 62.50% Count: 23 8
Node Id: 60 Statistic Train Validation 0: 56.19% 53.02% 1: 43.81% 46.98% Count: 331 149
Node Id: 61 Statistic Train Validation 0: 37.75% 29.09% 1: 62.25% 70.91% Count: 249 110
HME_Range
LoanOustandingRange
3
2 Or Missing
4, 6
3, 2, 5 Or Missing
Node Id: 80 Statistic Train Validation 0: 40.00% 20.00% 1: 60.00% 80.00% Count: 10 5
Node Id: 81 Statistic Train Validation 0: 53.85% 66.67% 1: 46.15% 33.33% Count: 13 3
Node Id: 104 Statistic Train Validation 0: 27.03% 41.18% 1: 72.97% 58.82% Count: 37 17
Node Id: 105 Statistic Train Validation 0: 59.86% 54.55% 1: 40.14% 45.45% Count: 294 132

Figure 2.
The decision tree model for balanced data and number of decision rules

Based on Figure 2, these are the decision rules of the DT model (balanced data):

- If the borrowers have nine loans and above, outstanding loan is RM30k and above, then their status is bankrupt (Node 6).
- If the borrowers have number of loans between one and four, number of children is nil or between one and three outstanding loan is RM29,999 and below, then their status is non-bankrupt (Node 8).
- If the borrowers have number of loans between five and eight, outstanding loan is RM30k and above, then their status is bankrupt (Node 14).
- If the borrowers have number of loans between one and four, number of children are between four to six or seven and above, outstanding loan is RM29,999 and below, employment status is private sector or public sector or retired or unemployed, then their status is non-bankrupt (Node 18).
- If the borrowers have number of loans between five and eight, outstanding loan is RM29,999 and below, household monthly income is more than RM3k, employment status is public sector, then their status is non-bankrupt (Node 20).
- If the borrowers are Malay, number of loans are between five and eight, outstanding loan is between RM29,999 and below, employment status is private sector or self-employed, then their status is bankrupt (Node 23).
- If the borrowers have number of loans between one and four, outstanding loan is RM30k and above, employment status is self-employed or public sector or retired or unemployed, then their status is bankrupt (Node 30).
- If the borrowers have number of loans between one and four, number of children are between four and six or seven and above, outstanding loan is RM29,999 and below, household monthly expenses below RM1,000 or between RM1,001 and RM2,000 or RM2,001 and RM3,000, employment status is self-employed, then their status is bankrupt (Node 38).
- If the borrowers have number of loans between one and four, number of children are between four and six or seven and above, outstanding loan is RM29,999 and below, household monthly expenses is RM3,001 and above, employment status is self-employed, then their status is non-bankrupt (Node 39).
- If the borrowers have number of loans between five and eight, outstanding loan of RM29,999 and below, household monthly income is between RM1,001 and RM2,000 or RM2,001 and RM3k, employment status is public sector, age is between 50-59 years, then their status is non-bankrupt (Node 42).
- If the borrowers have number of loan between five and eight, outstanding loan is RM29,999 and below, household monthly income is between RM1,001 and RM2,000 or RM2,001 and RM3k, employment status is public sector, aged between 30-39 and 40-49 years, then their status is bankrupt (Node 43).
- If the borrowers are Chinese or Indian, number of loans are between five and eight, outstanding loan is RM29,999 and below, household monthly income is between RM1,001 and RM2,000 or RM2,001 and RM3k, employment status is private sector or self-employed, then their status is bankrupt (Node 44).
- If the borrowers have number of loans between one and four, outstanding loan RM30k and above, household monthly income is between RM1,001 and RM2,000 or

RM2,001 and RM3k, employment status is private sector, then their status is bankrupt (Node 61).

- If the borrowers are Chinese or Indian, number of loans are between five and eight, outstanding loan is RM29,999 and below, household monthly income is RM3k and above, household monthly expenses are between RM2,001 and RM3,000, employment status private sector or self-employed, then their status is bankrupt (Node 80).
- If the borrowers are Chinese or Indian, number of loans are between five and eight, outstanding loan is RM29,999 and below, household monthly income is more than RM3k, household monthly expenses are between RM1,001 and RM2,000, employment status private sector or self-employed, then their status is non-bankrupt (Node 81).
- If the borrowers have number of loans between one and four, outstanding loan is between RM90k and RM119.9k or RM150k and above, household monthly income is more than RM3k, employment status is private sector, then their status is bankrupt (Node 104).
- If the borrowers have number of loans between one and four, outstanding loan is between RM30k and RM59.9k or RM60k and RM89.9k or RM120k and RM149.9k, household monthly income is more than RM3k, employment status is private sector, then their status is non-bankrupt (Node 105).

From the above decision rules, the profiles of a bankrupt based on balanced data set are:

- borrowers who have number of loans of more than five and high outstanding loan (RM30k and above);
- borrowers who are Malay and self-employed or work in the private sector;
- borrowers who are self-employed, work in the public sector or retired or unemployed with high outstanding loan (RM30k and above);
- borrowers who are self-employed, number of children are more than four and household monthly expenses is RM3,000 and below;
- borrowers are between 30 and 39 years or 40 and 49 years old and work in the public sector;
- borrowers who are Chinese or Indian with household monthly income of RM3,000 and below;
- borrowers who work in the private sector with high outstanding loan (RM30k and above) and household income of RM3,000 and below;
- borrowers who are Chinese or Indian, work in the private sector or self-employed with household expenses between RM2,001 and RM3,000 per month; and
- borrowers who work in the private sector with outstanding loan between RM90k and RM119.9k or RM150k and above.

Table I displays the significant variables of a DT model for the balanced and imbalanced data sets. For the imbalanced data set, there are five significant variables and eight for the balanced data set.

Performance measure consists of misclassification, accuracy, sensitivity and specificity rate. Misclassification is the probability of the model has wrongly predicted bankrupt as non-bankrupt and non-bankrupt as bankrupt. Accuracy means the probability of the model

Model	Significant variables Imbalanced data (*n* = 24,546)	Balanced data (*n* = 8,348)
DT	Outstanding loan	Outstanding loan
	Number of loans	Number of loans
	Number of children	Number of children
	Household monthly income	Household monthly income
	Race	Household monthly expenses
		Employment status
		Race
		Age

Source: Output from SAS Enterprise Miner 14.1 software

Table I. Results (significant/ important variables)

correctly predicted bankrupt and non-bankrupt. Sensitivity is the probability that the model can correctly predict bankrupt, and specificity means the probability of the model can correctly predict non-bankrupt. A good model consists of lower misclassification rate and higher accuracy and sensitivity rate (Akosa, 2017; Brown, 2014).

Table II displays the performance measure results. The results showed that the validation classification accuracy and specificity for the imbalanced data set is 83.29 per cent and 6.62 per cent, respectively. The specificity rate is low, less than 10 per cent, and sensitivity rate is 99 per cent (high rate, almost perfect) which indicated that the DT model was affected by the imbalanced data. Then, undersampling was performed by randomly selecting 4,174 cases from the 20,372 distressed cases, and re-evaluating the model using the balanced sample of 8,348 cases. The validation classification accuracy decreased slightly to 70.9 per cent, but the sensitivity rate increased to 81.23 per cent.

5. Conclusion

This paper discussed the improvements in the classification of personal bankruptcy using random undersampling to correct the imbalanced data. The application of DT in this study showed that the specificity rate had increased after the random undersampling strategy was applied. In practical applications, classification methods which are easy to understand such as DTs are more appealing to users (Yap *et al.*, 2011). In conclusion, the predictive performance of a DT model based on a balanced data set is more reasonable compared to an imbalanced data set. In future research, we intend to consider the LR model, support vector machine and naive Bayes model.

Model	Imbalanced data (*n* = 24,546) Training (%)	Validation (%)	Balanced data (*n* = 8,348) Training (%)	Validation (%)
DT				
Accuracy	83.38	83.29	71.32	70.90
Misclassification	16.61	16.70	28.68	29.10
Specificity	7.22	*6.62*	80.42	*81.23*
Sensitivity	98.99	99.00	62.18	60.57

Source: Output from SAS Enterprise Miner 14.1 software

Table II. Performance measure results (accuracy, precision, specificity and sensitivity)

References

Abdou, H. and Pointon, J. (2011), “Credit scoring, statistical techniques and evaluation criteria: a review of the literature”, *Intelligent Systems in Accounting, Finance and Management*, Vol. 18 Nos 2/3, pp. 59-88, doi: 10.1002/isaf.325.

Agarwal, S., Chomsisengphet, S. and Liu, C. (2011), “Consumer bankruptcy and default: the role of individual social capital”, *Journal of Economic Psychology*, Vol. 32 No. 4, pp. 632-650, doi: 10.1016/j.joep.2010.11.007.

Akosa, J.S. (2017), “Predictive accuracy: a misleading performance measure for highly imbalanced data classified negative”, SAS Global Forum, pp. 1-12, available at: http://support.sas.com/resources/papers/proceedings17/0942-2017.pdf

Azaizeh, S.Y. (2010), “Essays on household demand for credit cards, bankruptcy and overspending”, Available from ProQuest Dissertations and Theses Global, 763225189, available at: http://search.proquest.com.ezaccess.library.uitm.edu.my/docview/763225189?accountid=42518

Bernama (2016), “Alarming 22,581 M’sians aged 25 to 34 declared bankrupt in last four years”, *New Strait Times*, available at: www.nst.com.my/news/2016/11/189631/alarming-22581-msians-aged-25-34-declared-bankrupt-last-four-years

Bland, E.M., Christi, C. and Stokes, P.P. (2007), “A comparison of perceptions: students and bankruptcy filers”, *Journal of Business and Economics Research*, Vol. 5 No. 7, pp. 53-64.

Brown, I. (2014), *Developing Credit Risk Models Using SAS Enterprise Miner and SAS/STAT*, SAS Institute, Cary, NC.

Chen, M.Y. (2011), “Predicting corporate financial distress based on integration of decision tree classification and logistic regression”, *Expert Systems with Applications*, Vol. 38 No. 9, pp. 11261-11272.

Choy, M. and Flom, P. (2010), “Building decision trees from decision stumps”, SAS Global Forum 2010, (094), pp. 1-7.

Daskalaki, S., Kopanas, I., Goudara, M. and Avouris, N. (2003), “Data mining for decision support on customer insolvency in telecommunications business”, *European Journal of Operational Research*, Vol. 145 No. 2, pp. 239-255, doi: 10.1016/S0377-2217(02)00532-5.

Dawsey, A. (2014), “Externalities among creditors and personal bankruptcy”, *Journal of Financial Economic Policy*, Vol. 6 No. 1, pp. 2-24, doi: 10.1108/JFEP-09-2013-0037.

Dick, A. and Lehnert, A. (2010), “Personal bankruptcy and credit market competition”, *The Journal of Finance*, Vol. 65 No. 2, available at: http://onlinelibrary.wiley.com/doi/10.1111/j.1540-6261.2009.01547.x/full

Eaw, H.C., Khong, K.W., Rajagopalan, U. and A. Hamid, B. (2014), “Causality of personal bankruptcy in Malaysia”, Emerging Trends in Scientific Research, April 2006, pp. 363-367.

Eaw, H.C., Rajagopalan, U., Hamid, B.A., Khong, K.W. and Ahmad, R. (2015), “Psychographic factors: does it influence personal bankruptcy in Malaysia?”, *Journal of Business and Economic Studies*, Vol. 1 No. 2.

Fisher, J.D. (2005), “The effect of unemployment benefits, welfare benefits, and other income on personal bankruptcy”, *Contemporary Economic Policy*, Vol. 23 No. 4, pp. 483-492, doi: 10.1093/cep/.

Gepp, A. and Kumar, K. (2015), “Predicting financial distress: a comparison of survival analysis and decision tree techniques”, *Procedia Computer Science*, Vol. 54, pp. 396-404, available at: https://doi.org/10.1016/j.procs.20

Irimia-Dieguez, A.I. and Blanco-Oliver Vazquez-Cueto, M.J. (2015), “A comparison of classification/regression trees and logistic regression in failure models”, *Procedia Economics and Finance*, Vol. 23, pp. 9-14, available at: https://doi. org/10.1016/S2212-5671(15)00493-1

Jullamon, K. (2012), "Consumer bankruptcy in Thailand", Available from ProQuest Dissertations and Theses Global, 562209086, available at: http://search.proquest.com.ezaccess.library.uitm.edu.my/docview/1562209086?accountid=42518

Linoff, G.S. and Berry, M.J.A. (2011), *Data Mining Techniques: for Marketing, sales, and Customer Relationship Management*, 3rd ed., John Wiley and Sons, Hoboken, NJ.

Livshits, I., MacGee, J. and Tertilt, M. (2010), "Accounting for the rise in consumer bankruptcies", *American Economic Journal: Macroeconomics*, Vol. 2 No. 2, pp. 165-193, doi: 10.1257/mac.2.2.165.

Malaysia Department of Insolvency (2017), "Annual report", Bahagian Hal Ehwal Undang-Undang, Jabatan Perdana Menteri, available at: www.bheuu.gov.my/portal/pdf

Min, J.H. and Lee, Y.C. (2008), "A practical approach to credit scoring", *Expert Systems with Applications*, Vol. 35 No. 4, pp. 1762-1770, doi: 10.1016/j.eswa.2007.08.070.

Mirzei, M., Ramakrishnan, S. and Bekri, M. (2016), "Corporate default prediction with industry effects: evidence from emerging markets", *International Journal of Economics and Financial Issues*, Vol. 6 No. 2001, pp. 161-169.

Nair, Y., Paim, L., Sabri, M.F. and Rahim, H.A. (2016), "Predictors of bankruptcy probability among Malaysian civil servants: examining the subjective measurement", *Journal of Emerging Economies and Islamic Research*, Vol. 2011 No. 2011, pp. 1-12.

Noordin, N., Zakaria, Z., Zool, M., Mohamed, H. and Ngah, K. (2012), "Bankruptcy among young executives in Malaysia", *International Conference on Economics, Marketing and Management*, Vol. 28, pp. 132-136.

Olson, D.L., Delen, D. and Meng, Y. (2012), "Comparative analysis of data mining methods for bankruptcy prediction", *Decision Support Systems*, Vol. 52 No. 2, pp. 464-473, doi: 10.1016/j.dss.2011.10.007.

Othman, M.A., Abdul Rahim, H. and Sabri, M.F. (2015), "Differences in financial information and proceedings of the Australasian conference on business and social sciences 2015", pp. 525-531, available at: www.aabss.org.au/system/files/published/000959-published-acbss-2015-sydney.pdf

Paleologo, G., Elisseeff, A. and Antonini, G. (2010), "Subagging for credit scoring models", *European Journal of Operational Research*, Vol. 201 No. 2, pp. 490-499, doi: 10.1016/j.ejor.2009.03.008.

Řezáč, M. and Řezáč, F. (2011), "How to measure the quality of credit scoring models. Finance a uver", *Czech Journal of Economics and Finance*, Vol. 61 No. 5, pp. 486-507.

Sarma, K.S. (2017), *Predictive Modeling with SAS Enterprise Miner: practical Solutions for Business Applications*, SAS Institute, Cary, NC.

Selvanathan, M., Krisnan, U.D. and Wen, W.C. (2016), "Factors effecting towards personal bankruptcy among residents: case study in Klang valley, Malaysia", *International Journal of Human Resource Studies*, Vol. 6 No. 3, p. 98, doi: 10.5296/ijhrs.v6i3.10092.

Tsai, C.F. (2014), "Combining cluster analysis with classifier ensembles to predict financial distress", *Information Fusion*, Vol. 16 No. 1, pp. 46-58, doi: 10.1016/j.inffus.2011.12.001.

Ville, B.D. (2006), "Decision trees – what are they?", *Decision Trees for Business Intelligence and Data Mining*, SAS Institute, Cary, NC, pp. 1-4, available at: http://support.sas.com/publishing/pubcat/chaps/57587.pdf

Xiong, T., Wang, S., Mayers, A. and Mong, E. (2013), "Personal bankruptcy prediction by mining credit card data", *Expert Systems with Applications*, Vol. 40 No. 2, pp. 665-676, doi: 10.1016/j.eswa.2012.07.072.

Yap, B.W., Rahman, H.A.A., He, H. and Bulgiba, A. (2016), "Handling imbalanced dataset using SVM and k-NN approach", in *AIP Conference Proceedings*, *AIP Publishing*, Vol. 1750 No. 1, p. 20023.

Yap, B.W., Ong, S.H. and Husain, N.H.M. (2011), "Using data mining to improve assessment of credit worthiness via credit scoring models", *Expert Systems with Applications*, Vol. 38 No. 10, pp. 13274-13283, available at: https://doi.org/10.1016/j.eswa.2011.04.147

Zainon, Z. (2016), "AKPK selesai masalah kewangan RM476.6j: Utusan online", available at: www.utusan.com.my/bisnes/korporat/akpk-selesai-masalah-kewangan-rm476-6j-1.417133#nav-allsections

Zhu, N. (2013), "Household consumption and personal bankruptcy", *Journal of Legal Studies*, Vol. 40 No. 1, pp. 1-37.

Corresponding author
Sharifah Heryati Syed Nor can be contacted at: sharifahheryati_syednor@yahoo.com