



Using data mining to improve assessment of credit worthiness via credit scoring models

Bee Wah Yap^{a,*}, Seng Huat Ong^{b,1}, Nor Huselina Mohamed Husain^{a,2}

^a Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia

^b Institute of Mathematical Sciences, University of Malaya, 50603 Kuala Lumpur, Malaysia

ARTICLE INFO

Keywords:

Data mining
Credit scoring
Logistic regression
Decision tree
Classification
Predictive modeling

ABSTRACT

Credit scoring model have been developed by banks and researchers to improve the process of assessing credit worthiness during the credit evaluation process. The objective of credit scoring models is to assign credit risk to either a “good risk” group that is likely to repay financial obligation or a “bad risk” group who has high possibility of defaulting on the financial obligation. Construction of credit scoring models requires data mining techniques. Using historical data on payments, demographic characteristics and statistical techniques, credit scoring models can help identify the important demographic characteristics related to credit risk and provide a score for each customer. This paper illustrates using data mining to improve assessment of credit worthiness using credit scoring models. Due to privacy concerns and unavailability of real financial data from banks this study applies the credit scoring techniques using data of payment history of members from a recreational club. The club has been facing a problem of rising number in defaulters in their monthly club subscription payments. The management would like to have a model which they can deploy to identify potential defaulters. The classification performance of credit scorecard model, logistic regression model and decision tree model were compared. The classification error rates for credit scorecard model, logistic regression and decision tree were 27.9%, 28.8% and 28.1%, respectively. Although no model outperforms the other, scorecards are relatively much easier to deploy in practical applications.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

Credit scoring models are very useful for many practical applications especially for banks and financial institutions. The decision-making process of accepting or rejecting a client's credit by banks is commonly executed via judgmental techniques and/or credit scoring models. Most banks and financial institutions use the judgmental approach which is based on the 3C's, 4C's or 5C's which are character, capital, collateral, capacity and condition. Credit scoring is a system creditors use to assign credit applicants to either a “good credit” one that is likely to repay financial obligation or a “bad credit” one who has a high possibility of defaulting on financial obligation. Generally, Linear Discriminant Analysis and logistic regression are two popular statistical tools to construct credit scoring models (Abdou, Pointon, & El-Masry, 2008; Desai, Crook, & Overstreet, 1996; Gao, Zhou, Gao, & Shi, 2006; Hand &

Henley, 1997; Thomas, 2000; Vojtek & Kocenda, 2006). However, with the advance in information and computer technology new techniques are appearing under the name of data mining. Data mining software such as SAS[®] Enterprise Miner and SPSS PASW[®] 13 modeler provide not only the classical methods but new novel predictive modeling and classification techniques such as decision tree, neural networks, support vector machine (SVM), and *k*-nearest neighbors.

Although credit scoring methods are widely used for loan applications in financial and banking institutions, it can be used for other type of organizations such as insurance, real estate, telecommunication and recreational clubs for predicting late payments. For example, Gschwind (2007) showed a data mining application in real estate for predicting late payments by tenant. Due to privacy concerns and unavailability of data from banks, for this paper, historical payment of monthly subscription from members of a local recreational club was used. Payment of the monthly subscription fee is an obligation of the club members besides paying the permanent membership fee. The management faces the problem in the rising number of defaulters. So far, there has been no significant effort to improve cash flow by proactively predicting non-payments using quantitative methods, and taking corrective actions before a late payment happened. Discussion with the

* Corresponding author. Tel.: +60 03 55435461; fax: +60 03 55435501.

E-mail addresses: beewah@tmsk.uitm.edu.my, yapbeewah@salam.uitm.edu.my (B.W. Yap), ongsh@um.edu.my (S.H. Ong), huselina_11@yahoo.com (N.H.M. Husain).

¹ Tel.: +60 03 79674306; fax: +60 03 79674143.

² Tel.: +60 03 55435461; fax: +60 03 55435501.

management of the club revealed that they use judgmental techniques to determine the defaulters or non-defaulters and whether to terminate the membership of defaulters. The main source of income for most recreational clubs is the membership monthly payments. A large number of defaulters will result in cash flow problem and loss of income for the club. This will affect the financial planning of the club activities and the management faces the problem of ensuring that the club does not go bankrupt. The objective of this paper is to illustrate the use of data mining in assessing credit worthiness using credit scoring models and for prediction of an event such as default in payment so that early intervention can be done to prevent financial loss.

This paper is organized as follows. Section 2 provides a review of the applications of data mining and credit scoring models. Then, the conceptual framework is presented. The methodology for constructing the credit scoring models is covered in Section 3. The results are discussed in Section 4. Finally, the limitations of the data mining approach to the construction of credit scoring models are highlighted in the concluding section.

2. Literature review

2.1. Data mining

Data mining refers to the extraction of useful patterns or rules from a large database. The data mining process involves identifying the business problem and data mining goal, retrieving the database needed, and using data mining techniques to analyze the data with the final aim of achieving important results for making strategic decisions (Berry & Linoff, 2004). Data mining is an integral part of knowledge discovery in databases (KDD). Data mining involves techniques such as anomaly detection, association analysis, clustering, and predictive modeling (Berry & Linoff, 2004; Han & Kamber, 2001; Tan, Steinbach, & Kumar, 2006). Anomaly detection involves algorithm that can discover real anomalies such as detection of fraud, unusual disease, unusual weather conditions and ecosystem disturbances. Association analysis enables the discovery of group of objects that occur frequently together such as items that are often bought together or genes that are similar. Clustering involves segmentation of objects in a large database into clusters or segments with similar characteristics. Such segmentation is especially useful for target marketing. Predictive modeling involves using statistical models, machine-learning technique such as decision tree algorithm or artificial intelligence model to predict the outcome of a dependent variable based on several attributes (or independent variables). Data mining has been applied in many fields such as banking, finance, telecommunication, manufacturing, healthcare, insurance, real estate, education, marketing, customer relationship management and weather study such as avalanche forecasting (Abdou et al., 2008; Ang, Chua, & Bowling, 1979; Chien & Chen, 2008; Chien, Hsiao, & Wang, 2004; Chien, Wang, & Chen, 2005; Cho & Ngai, 2003; Davis, Elder, Howlett, & Bouzaglou, 1999; Gschwind, 2007; Kurt, Ture, & Kurum, 2008; Lee, Chiu, Chou, & Lu, 2006; Rygielski, Wang, & Yen, 2002).

2.2. Credit scoring models

Credit scoring was first introduced in the 1940s and over the years had evolved and developed significantly. In the 1960s, with the creation of credit cards, banks and other credit card issuers realized the advantages of credit scoring in the credit granting process. In the 1980s, credit scoring was used for other purposes such as aiding decision in approving personal loan applications. In recent years, credit scoring has been used for home loans, small business loans and insurance applications and renewals (Koh, Tan, &

Goh, 2004; Thomas, 2000). A credit scoring model provides an estimate of a borrower's credit risk – i.e. the likelihood that the borrower will repay the loan as promised, based on a number of quantifiable borrower characteristics (Dinh & Kleimeier, 2007). Credit scoring is based on statistical or operational research methods. Historically, discriminant analysis and linear regression have been the most widely used techniques for building scorecards. Other techniques include logistic regression, probit analysis, non-parametric smoothing methods especially *k*-nearest neighbors, mathematical programming, Markov chain models, recursive partitioning, expert systems, genetic algorithms and neural networks (Hand & Henley, 1997).

Artificial neural networks (ANNs) have been criticized for its 'black box' approach and interpretative difficulties. Multivariate adaptive regression splines (MARS), classification and regression tree (CART), case based reasoning (CBR), and support vector machine (SVM) are some recently developed techniques for building credit scoring models. Huang, Chen, Hsu, Chen, and Wu (2004) investigated the performance of the SVM approach in credit rating prediction in comparison with back propagation neural networks (BNN). However, only slight improvement of SVM over BNN was observed. Huang, Chen, and Wang (2007) reported that compared with neural networks, genetic programming and decision tree classifiers, the SVM classifier achieved identical classification accuracy with relatively few input variables.

Lee et al., 2006 demonstrated the effectiveness of credit scoring using CART and MARS. Their results revealed that, CART and MARS outperform traditional discriminant analysis, logistic regression, neural networks, and support vector machine (SVM) approaches in terms of credit scoring accuracy. Recently, with the development of data mining software the process involved in building credit scoring model is made much easier for credit analysts. Despite the development of new novel techniques, for practical applications the popular techniques for banking and business enterprises are credit scorecards, logistic regression and decision trees as it is relatively easy to identify the important input variable, interpret the results and deploy the model.

2.3. Conceptual framework

In building a scoring model, or "scorecard", historical data on the performance of previously made loans and borrowers characteristics are required. A good scoring model should give a higher percentage of high scores to 'good borrowers' and a higher percentage of low scores to those who are 'bad borrowers'. Ang et al. (1979) investigated the profiles of late-paying consumer loan borrowers using variables such as gross amount of loan, age, sex, marital status, number of dependents, years lived at residence, monthly take home pay, monthly take home pay of spouse, own or rent residence, other monthly income, total monthly payments on all debts, type of bank accounts, number of credit references listed, years on job, total family monthly income per month, debt to income ratio, total number of payments on the loan, and annual percentage interest on the loan. Koh et al. (2004) used age, annual income, gender, marital status, number of children, number of other credit cards held and whether the applicant has an outstanding mortgage loan to construct a credit scoring model to predict credit risk of credit card applicants as bad loss, bad profit and good risk. Abdou et al. (2008) used twenty variables some of which were loan amount, loan duration, sex, marital status, age, monthly salary, additional income, house owned or rent, and education level for building credit scoring models to evaluate credit risk (paid or default) for personal loan. Gschwind (2007) concluded that mining basic tenant data, accounts receivable data, and government-published data can generate predictions of late payments of rental. Mavri, Angelis, and Loannou (2008) used variables such as gender,

متن کامل مقاله

دریافت فوری ←

ISIArticles

مرجع مقالات تخصصی ایران

- ✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
- ✓ امکان دانلود نسخه ترجمه شده مقالات
- ✓ پذیرش سفارش ترجمه تخصصی
- ✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
- ✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
- ✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
- ✓ دانلود فوری مقاله پس از پرداخت آنلاین
- ✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات