Knowledge-Based Systems 36 (2012) 245–252

Contents lists available at SciVerse ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

A Bayesian latent variable model with classification and regression tree approach for behavior and credit scoring

Ling-Jing Kao [a], Chih-Chou Chiu [a,*], Fon-Yu Chiu [b]

[a] Department of Business Management, National Taipei University of Technology, Taiwan
[b] Institute of Commerce Automation and Management, National Taipei University of Technology, Taiwan

ARTICLE INFO

Article history:
Received 2 April 2012
Received in revised form 5 July 2012
Accepted 10 July 2012
Available online 21 July 2012

Keywords:
Behavior scoring
Credit scoring
Bayesian
Latent variable model
Classification and regression tree

ABSTRACT

A Bayesian latent variable model with classification and regression tree approach is built to overcome three challenges encountered by a bank in credit-granting process. These three challenges include (1) the bank wants to predict the future performance of an applicant accurately; (2) given current information about cardholders' credit usage and repayment behavior, financial institutions would like to determine the optimal credit limit and APR for an applicant; and (3) the bank would like to improve its efficiency by automating the process of credit-granting decisions. Data from a leading bank in Taiwan is used to illustrate the combined approach. The data set consists of each credit card holder's credit usage and repayment data, demographic information, and credit report. Empirical study shows that the demographic variables used in most credit scoring models have little explanatory ability with regard to a cardholder's credit usage and repayment behavior. A cardholder's credit history provides the most important information in credit scoring. The continuous latent customer quality from the Bayesian latent variable model allows considerable latitude for producing finer rules for credit granting decisions. Compared to the performance of discriminant analysis, logistic regression, neural network, multivariate adaptive regression splines (MARS) and support vector machine (SVM), the proposed model has a 92.9% accuracy rate in predicting customer types, is less impacted by prior probabilities, and has a significantly low Type I errors in comparison with the other five approaches.

1. Introduction

Evaluating a customer's credit risk is crucial for financial institutions due to the high risks associated with inappropriate credit-granting decisions. It becomes an even more important task today after the recent financial crisis involving subprime loans. Three challenges are encountered by the bank while making credit-granting decisions. First, the bank would like to predict the future performance of an applicant accurately. Second, given information about current cardholders' credit usage and repayment behavior, financial institutions would like to determine the optimal level of product attributes, such as credit limit and annual percentage rate, for an applicant. Third, the bank would like to improve its efficiency by automating credit-granting decisions.

Financial institutions try to solve these three challenges by making heavy use of predictive scoring models. Depending on a financial institution's target activities and available customer information, the scoring models can be classified into credit scoring models or behavioral scoring models. A credit scoring model aims to make decisions about whether to grant new customers a particular financial product (e.g., credit), whereas a behavioral scoring model is designed to evaluate the credit of existing customers, i.e., how reliably customers keep up to date with payments.

Scoring models usually produce a binary credit rating which classifies customers into two groups (such as 'good/bad'; 1/0) according to classification rules and a set of explanatory variables. However, the choice of classification rules and explanatory variables is ad hoc and usually relies on managers' know-how and experience. Many statistical methods (e.g. logistic regression model; discriminant analysis) and artificial intelligence approaches (e.g. neural networks and rule-based approaches) have been used to construct scoring or fraud detection models [7,23,18,25,17,26]. These approaches have good credit or behavior scoring capabilities, but they are criticized for their black-box extraction of data feature vectors [19].

The objective of this paper is to provide a combined model that can overcome all three of these challenges involved in credit-granting decisions in the financial industry and that can overcome the drawbacks of traditional scoring approaches. It is called a "combined" model because the model we proposed is not only to

* Corresponding author.
E-mail addresses: lingjingkao@ntut.edu.tw (L.-J. Kao), chih3c@ntut.edu.tw (C.-C. Chiu), bdyucu@yahoo.com.tw (F.-Y. Chiu).

0950-7051/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2012.07.004

apply methods in sequence but also to establish the linkage between the credit scoring model and the behavior scoring model. It allows us to study the decision of credit granting given both credit terms (APR and etc.) and credit usage behavior.

The idea of integrating multiple approaches is not new in literature. For example, Chen [4] integrated feature selection and CPDA-based rough set approach to classify Asian banks' credit rating. The combined approach proposed in this paper consists of two steps. First, given customer types (e.g., good or bad) determined by credit analysts in a bank, a hierarchical Bayes model is developed to estimate heterogeneous customer quality scores. A cardholder's quality score is a continuous latent variable which summarizes this cardholder's repayment decisions and credit usage behaviors. Then, given posterior estimates of quality value and customers' credit reports as well as demographic information, we use the Classification and Regression Tree (CART) algorithm to deduce decision rules that can be used to determine whether to grant an applicant credit, and to determine the optimal levels of product attributes, such as credit limits and annual percentage rate.

Data provided by a bank in Taiwan is used to illustrate the combined approach proposed in this paper. This data set consists of each credit card holder's credit usage and repayment data, demographic information, and credit report. The empirical study shows that demographic variables used in most credit scoring models have little explanatory ability with regard to a cardholder's credit usage and repayment behavior. A cardholder's credit history provides the most important information in credit scoring. We demonstrate that the proposed model can result in a 92.9% accurate prediction and the lowest Type I error.

The remainder of the paper is organized as follows: In Section 2, the proposed approach is developed by combining a Bayesian behavior scoring model and a CART-based credit scoring model. An empirical study of a credit card and its results are presented in Section 3. Conclusions are offered in Section 4.

2. A combined approach

2.1. Bayesian behavior scoring model

Let y_{ij} denote the credit rating of the ith cardholder assigned by credit analysts in a bank. y_{ij} equals one if the ith credit card holder was evaluated as creditworthy during jth time period. y_{ij} equals zero if the ith credit card holder was evaluated as not creditworthy during jth time period. Let z_{ij} be the quality score of the ith cardholder during the time period j. A cardholder's quality (z_{ij}) is a continuous latent variable that summarizes this cardholder's credit usage and repayment behavior and is assumed to associate with a stochastic component ε_i and the vector of independent variables X_{ij}. Let $X_{ij} = (x_{1ij}, x_{2ij}, \ldots, x_{kij})'$ be a vector of variables used to measure the "quality" of the ith cardholder during the jth time period. The behavior scoring model can be written as follows:

$$y_{ij} = \begin{cases} 1 & z_{ij} \geqslant 0 \\ 0 & otherwise \end{cases} \qquad (1)$$
$$z_{ij} = X_{ij}\beta_i + \varepsilon_{ij} \quad \varepsilon_{ij} \sim N(0, \sigma_i^2), \forall i = 1, 2, \ldots, H; j = 1, 2, \ldots, T$$

Eq. (1) is a latent variable model with binary responses. We use Eq. (1) to describe a cardholder's quality (z_{ij}) and its relationship to the corresponding discrete binary outcome variable (y_{ij}). Threshold value is set at zero and σ^2 is set to one to overcome the location and scale identification problems respectively. The latent variable model has been an important tool in the social sciences literature [1]. The binary observations are often assumed to be driven by unobservable behavior mechanisms that are associated with a set of explanatory variables. For example, Bartholomew [2] uses the latent variable to obtain insights into the structure of the responses in the Workplace Industrial Relations Survey. Hand and Crowder [10] proposed a latent variable model which measures the underlying quality of a customer from a retail bank's perspective. Economists also use latent variable models to study labor force participation, the choice of occupation, consumer choice among alternatives, etc. [9].

Both the literature and industry practice suggest that inter-payment time and repayment ability are good indicators to determine a customer's value [14,15]. Therefore, we assume that the vector X_{ij} in Eq. (1) consists of two independent variables (x_{1ij}, x_{2ij}), in which x_{1ij} represents repayment ability and x_{2ij} represents the inter-payment time of the ith customer at time period j. A customer has a higher chance of being in default if a long inter-payment time is observed or a low percentage of credit debts is paid. A customer is defined as being in default if the minimum payment is not met after passing the due date for a certain period of time (e.g., 120 days, depending on bank rules).

A customer's repayment ability can also be evaluated by computing the ratio between the payment and the entire outstanding balance. This ratio also represents the customer's liquidity. A consumer who can pay most of his balance is considered to have good liquidity. Depending on banks' particular rules, the payment made by a customer can be allocated into either principal or interest. Each bank also has its own method (i.e., average daily balance method) to compute the incurred interest.

Let w_{ij} be the ith customer's payments to the principal, while p_{it} is the ith customer's total principal, Q_{ij} is the ith customer's payments to the interest incurred from the loan, and s_{ij} is the ith customer's total interest incurred from the loan during the time period j. A metric for evaluating the ith customer's repayment ability is developed as follows:

$$x_{1ij} = \rho\left(\frac{w_{ij}}{p_{ij}}\right) + (1-\rho)\left(\frac{Q_{ij}}{s_{ij}}\right) \qquad (2)$$

In Eq. (2), the first term (w_{ij}/p_{ij}) represents a customer's principal repayment ability, and the second term (Q_{ij}/s_{ij}) is used to measure a customer's interest repayment ability. ρ represents the weight given to the importance of repayment ability. For example, if the ith customer is a convenience user who uses a credit card as a convenient payment device and pays off the outstanding balance every period, the transaction fees collected from merchants are the only contribution he brings to the bank. Then, by Eq. (2), his payment ability at this period will be 0.5. However, those customers who use credit and have incurred interest over time and repay most of their credit debts, can contribute the most to the bank and are considered to have good repayment ability. In our research, we assume that principal repayment ability and interest repayment ability are equally important.

The Gibbs sampler and data augmentation [8,11,20,22] were used to estimate the customer quality model. Let $Z_i = (z_{i1}, z_{i2}, \ldots, z_{iT})'$ and $X_i = (X_{i1}, X_{i2}, \ldots, X_{iT})'$. Given conjugate priors

$$\beta_i \sim N(\bar{\beta}, V_\beta),$$
$$\bar{\beta} \sim N(\mu_0, A_0),$$
$$V_\beta \sim IW(f_0, G_0),$$

posterior estimates of parameters can be obtained by generating draws from the following full conditional distributions.

$$[z_{ij}|\cdot] \propto \prod [y_{ij}|z_{ij}][z_{ij}|X_{ij}\beta_i] \sim Truncated\,Normal(X_{ij}\beta_i, 1)$$

$$[\beta_i|\cdot] \propto [Z_i|X_i\beta_i][\beta_i|\bar{\beta}, V_\beta]$$
$$\sim N((X_i'X_i + V_\beta^{-1})^{-1}(X_i'Z_i + V_\beta^{-1}\bar{\beta}), (X_i'X_i + V_\beta^{-1})^{-1})$$