European Journal of Operational Research 241 (2015) 202–211

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Innovative Applications of O.R.

Identifying future defaulters: A hierarchical Bayesian method

Fan Liu [a,b,*], Zhongsheng Hua [a], Andrew Lim [b]

[a] The School of Management, University of Science and Technology of China, Hefei 230026, Anhui, China
[b] Department of Management Sciences, City University of Hong Kong, Tat Chee Ave, Kowloon Tong, Hong Kong

ARTICLE INFO

Article history:
Received 26 November 2013
Accepted 4 August 2014
Available online 13 August 2014

Keywords:
Risk management
Credit scoring
Mixture cure model
Bayesian analysis

ABSTRACT

Traditional methods of applying classification models into the area of credit scoring may ignore the effect from censoring. Survival analysis has been introduced with its ability to deal with censored data. The mixture cure model, one important branch of survival models, is also applied in the context of credit scoring, assuming that the study population is a mixture of never-default and will-default customers.

We extend the standard mixture cure model through: (1) relaxing the independence assumption of the probability and the time of default; (2) treating the missing defaulting labels as latent variables and applying an augmentation technique; and (3) introducing a discrete truncated exponential distribution to model the time of default. Our full model is written in a hierarchical form so that the Markov chain Monte Carlo method is applied to estimate corresponding parameters.

Through an empirical analysis, we show that both mixture models, the standard mixture cure model and the hierarchical mixture cure model (HMCM), are more advanced in identifying future defaulters while compared with logistic regression. It is also concluded that our hierarchical Bayesian extension increases the model's predictability and provides meaningful output for risk management.

1. Introduction

Under Basel II, banks are allowed to build their own credit scoring models to risk assess their customers and calculate the capital in need for the portfolio. It makes it more important to discover suitable scoring models to better understand customers payment behavior. Traditionally, credit scoring is modeled as a classification problem – loan applicant is either classified as credit worthy or credit worthless. Dozens of techniques have been proposed by numerous researchers, for details we refer the readers to excellent surveys (Crook, Edelman, & Thomas, 2007; Hand & Henley, 1997; Rosenberg & Gleit, 1994). Most of these approaches require a development sample that contains properly labeled customers who have been assigned to default group or non-default group based on their repayment history (Hand & Kelly, 2001) over an observation period. Such assignment is not always available, especially for long term products, which limits the applicability of such methods in practice.

Let us take a real industrial application as an example. One of the largest commercial banks in China wanted to build a scorecard before the end of 2008 for their mortgage business. Samples were selected from existing accounts where the average term of loan was over 10 years. To obtain accurate label for each customer, we were forced to use data that were at least 10 years old. In the rapid developing country like China, there is a dramatic change in economic environments, data that are too old cannot accurately reflect the current situation. Observation period cannot be too short either. Mortgage customers rarely default immediately after taking the loan. It will take some time before we can observe sufficient number of defaults to train any meaningful model. The time window is always determined by some ad hoc rules, such as vintage analysis. In this example, it shows in Fig. 1 that historically 70% of the defaults have occurred within 24 months. Therefore, 24 months was a good time window for selecting training data. The defaulting labels for customers were then determined according to their behavior before 31st December 2008 rather than the loan lifetime: that is, the data was right censored. However, the non-default group included customers that may default after the censoring time. Directly apply classification based method to this data set was inappropriate. Intuitively, ignoring the censoring effect may make the model unable to identify future defaulters since they were labeled as non-default in model development procedure.

Survival analysis, with its ability to deal with the censored data, has been introduced in the context of credit scoring. Exploratory works include Banasik, Crook, and Thomas (1999), Hand and Kelly (2001) and Stepanova and Thomas (2002), in which various

* Corresponding author at: The School of Management, University of Science and Technology of China, Hefei 230026, Anhui, China.
E-mail address: liuf1989@mail.ustc.edu.cn (F. Liu).

http://dx.doi.org/10.1016/j.ejor.2014.08.008
0377-2217/

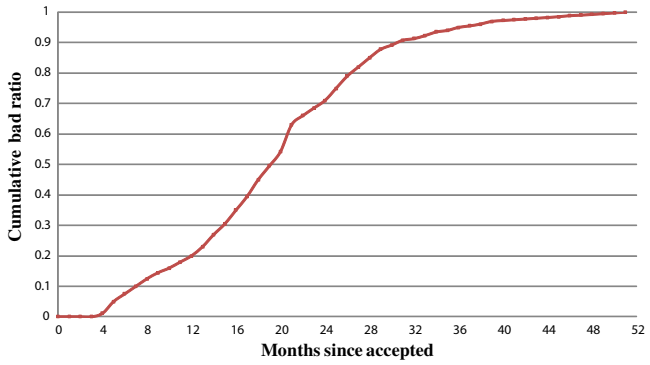

Fig. 1. Vintage analysis.

standard survival models, both parametric and non-parametric, are applied and compared. These works are followed by Bellotti and Crook (2009) and Im, Apley, Qi, and Shan (2012) which both incorporate the time-dependent variables into the proportional hazards model, with (Bellotti & Crook, 2009) including seven macroeconomic variables as the time-varying covariates and (Im et al., 2012) combining all the external factors into one coefficient, which accelerates the occurrence of defaulting. Baesens, Gestel, Stepanova, and Vanthienen (2005) proposes several neural-network based survival models that allow the covariates to be incorporated in a nonlinear mode. Andreeva, Ansell, and Crook (2007) introduces a combination score method to capture the relationship between the customers' profitability and the time of default. More recently, the mixture cure models are proposed in Yildirim (2008) and Tong, Mues, and Thomas (2012) on the consideration that the customers are latently divided into two groups. One group includes customers who will never default, and the other includes those who will. They use mixture approaches to model the event of default by two parts, an incidence part, which describes the probability of the occurrence of an event, and a latency part, describing the time of the occurrence given it would happen. These papers concluded that advantages of survival analysis are that, it can be used to better discriminate applicants, estimate defaulting probability as a function of time, assist banks to make decisions on credit management.

We retain the desirable properties of the survival analysis and complement it by extending the mixture cure model to a hierarchical model, the HMCM. First, we establish a joint multivariate normal distribution of the defaulting predictor and the logarithm of the hazard rate, so that the assumption of the independence of the probability and the time to default is relaxed. The covariance matrix estimated using our method can be used to test whether the dependency exists. It is also meaningful for banks, because a customer with higher defaulting probability may contribute more profit if he defaults later (Stepanova & Thomas, 2002). Second, a discrete truncated exponential distribution model is developed to accommodate two characters of the defaulting time for a fixed-term loan. That is, the reported defaulting time is discrete, usually monthly, and finite, as it is restricted by the loan term. At last, we incorporate the unobserved default labels into the model as latent variables, and augment them according to their posterior distributions, which makes our model flexible and easy to implement. Peng and Dear (2000), Corbire and Joly (2007) and Tong et al. (2012) have introduced such latent variable. However, they do not augment it but compute its posterior expectation which is required by the expectation–maximization algorithm. Since the likelihood under our model has no analytical form, the parameters are estimated using Bayesian method. Markov chain Monte Carlo (MCMC) procedure is one of the approaches to deal with hierarchical models by generating the parameters with a Markov chain.

We compare and contrast the performance of the HMCM with the mixture cure model proposed in Yildirim (2008) and a logistic regression model. Via an application on a real data, we conclude that ignoring the effect of censoring will make the model lose its ability to identify the customers who default later than the censoring time. Our extension highly improves the performance of the mixture model in terms of identifying the future defaulters, even some of them default much later than the censoring time. From the correlation coefficient derived in our data, the riskier the customer, the longer he may survive. It reveals the reason why logistic regression underperforms the mixture models.

Our method is similar to that in Abe (2009) which used a hierarchical Bayesian extension of the Prato/NBD model to forecast customers future purchases. Through an empirical analysis on three datasets, they conclude that the customer purchasing behavior is better tracked and, most importantly, the correlation estimated provides significant insights in customer relationship management. Unlike the consumer behavior discussed in Abe (2009), the event of default on a fixed-term loan cannot occur repeatedly, so the Poisson assumption of customer behavior is replaced by a Bernoulli distribution.

The remainder of this paper is organized as follows. The mathematic notations, the statistical models and how our extensions are made are described in Section 2. In Section 3, we estimate the parameters using MCMC algorithm and list the main results. An empirical analysis is conducted in Section 4 to compare our model with the logistic regression model and the mixture cure models proposed in Tong et al. (2012). Section 5 concludes our paper with several directions of future research.

2. Model specification

2.1. Mathematic notation

Suppose the observation is taken on N customers, and the ith customer is accepted for a loan with term L_i. Let T_i denote the defaulting time since the loan is approved. T_i may or may not be observed during the observation period C_i. Although we used the same time window to select customers, due to the difference in the starting time of the loans, the observation period for each customer may differ. The ith customer record takes the form $(t_i, \delta_i, \boldsymbol{x_i})$. Where, $\delta_i = 1$ indicates the customer defaults within the observation period, that is, $T_i \leqslant C_i$, and $\delta_i = 0$ otherwise; $t_i = \min\{T_i, C_i\}$ is the reported time; $\boldsymbol{x_i}$ is a $1 \times p$ vector of the explanatory variables. y_i is defined as the unobserved defaulting label, indicating whether a customer will eventually default. If the ith customer defaults before C_i, i.e. $T_i \leqslant L_i$ and $\delta_i = 1$, then $y_i = 1$, else y_i could be either 0 or 1. For simplicity of presentation we omitted the subscript for customer for all symbols in the following discussion.

2.2. Standard mixture cure model

Let $q = P(y = 1)$ denote the probability that the customer will eventually default. Then the probability that a customer survives after t is given by:

$$S(t) = p(T > t) = 1 - q + qS(t|y = 1) \tag{1}$$

where $S(t|y = 1)$ is the probability of the customer survives after t conditioning on that the customer will eventually default. To incorporate explanatory variables into this model, we specify q using logistic regression:

$$logit(q) = \beta_{q0} + \sum_{j=1}^{p} \beta_{qj} x_j \tag{2}$$