Expert Systems with Applications 37 (2010) 993–997

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Application of data mining to the spatial heterogeneity of foreclosed mortgages

Tsung-Hao Chen [a,*], Cheng-Wu Chen [b]

[a] Departments of Business Administration, Shu-Te University, Yen Chau, Kaohsiung 82445, Taiwan, ROC
[b] Departments of Logistics Management, Shu-Te University, Yen Chau, Kaohsiung 82445, Taiwan, ROC

ARTICLE INFO

Keywords:
LGD
Data mining
Heterogeneity
Residential mortgage loans
Foreclosure

ABSTRACT

The loss given a default (LGD) is a key component when calculating the credit risk associated with an asset portfolio. However, the issue of default probability has not often been addressed in past mortgage loan data mining studies. The LGD has rarely been used to assess the comprehensive credit risk for a portfolio of mortgage loans. The location of a mortgaged property is strongly correlated with the price of that property as well as providing social, demographic, and economic information which inherently characterizes the mortgage loan population. Thus, to make an accurate assessment of the credit risk associated with the loan portfolio, one requires a specific data mining technique capable of determining the heterogeneity of the portfolio across regions. The sample utilized in this study consists of data on two thousand foreclosed mortgages in Kaohsiung City. We first test the homogeneity between the different city districts; second, we estimate the magnitude of the heterogeneity, including the spatial heterogeneity; third, a prior distribution for the heterogeneity is formulated using data mining methods; finally, the overall LGD, showing the credit risk for a given default probability is calculated.

1. Introduction

Data mining is a technique for extracting knowledge from information. It can be applied to determine the various social, medical, demographic, financial and economic factors, and collect information (Bayam, Liebowitz, & Agresti, 2005; Cho & Kim, 2004; Chun & Kim, 2004; Chun & Park, 2005; Chun & Park, 2006; Coussement & Van den Poel, 2008; Ha & Park, 1998; Hsu & Chen, 2007; Hwang, Chang, & Chen, 2008; Kuo & Chen, 2001; Prinzie & Van den Poel, 2005). There are many factors that could have an influence on the default behavior of residential mortgages, such as the present value of the mortgage payments, characteristics of the family, loan to value (LTV) ratio, home equity, unemployment rate, and divorce rate (Ciochetti, Lee, Shilling, & Yao, 2001; Deng, Quigley, & Van Order, 1996; Deng, Quigley, & Van Order, 2000; Deng, 1997; Lambrecht, Perraudin, & Satchell, 2003; Marrison, 2002). In their study of the effects of counseling on mortgage default behavior, Hartarska and Gonzalez-Vega (2005, 2006) concluded that counseled borrowers were less likely to default on their mortgage than non-counseled borrowers, and that this also affected the optimal exercise.

Ambrose, Capone, and Deng (2001) decomposed the boundary conditions for optimal default exercise to look at the economic dynamics leading to optimal default timing for mortgage foreclosure. Ong, Neo, and Tu (2007) showed that price expectations, volatility and equity losses are influential factors for individual households, with past price movement being the most important of these. However, there has been little research on how these factors influence mortgages prior to foreclosure or how they are correlated with location.

Thus, in this study, we test homogeneity in different city districts. The remaining part of this paper is organized as follows. In Section 2 we discuss our motivation for adopting a nonlinear mixed model. In Section 3 we present an empirical analysis of random effects and in Section 4 some conclusions are offered.

2. Research framework

2.1. Fixed effect model

Relationships between $\pi_i(X) = E(Y_i|x)$ and x are usually nonlinear rather than linear. Logistic regression analysis is a statistical modeling approach for analyzing dichotomous response data which can accommodate the adjustment of one or more explanatory variables (Walker, 2002). For a binary response (Y), where the explanatory variables are denoted as a vector X, let $\pi_i(X)$ represent success probability. To simplify the notation, we use $\pi_i(X) = E(Y_i|X)$ to represent the conditional mean of Y given X when a logistic distribution is specified, giving us

$$\ln\left(\frac{E(Y_i)}{1-E(Y_i)}\right) = \log it(\pi_i) = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots\cdots + \beta_k X_{ik}. \quad (1)$$

* Corresponding author. Tel.: +886 952492037.
E-mail addresses: thchen@mail.stu.edu.tw (T.-H. Chen), cwchen@mail.stu.edu.tw (C.-W. Chen).

0957-4174/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2009.05.076

2.2. Overdispersion and diagnostics

Several problems can cause overdispersion in a fitted model, such as a large residual deviance relative to the number of degrees of freedom or a lack of important explanatory variables (Collect, 2003). Allison (1999) suggested that lack-of-fit could be associated with the dependence of the observations. Collect (2003) suggested that overdispersion might be accommodated by including a random effect in the model.

There are two conventional goodness-of-fit tests, the Pearson χ^2 and the likelihood ratio χ^2 tests, also known as deviance χ^2 (Allison, 1999; Lawal, 2003; Myers, Montgomery, & Vining, 2002). Pregibon (1981) carried out a theoretical analysis that extended linear regression diagnostics to a logistic regression.

2.2.1. Pearson χ^2 test

The Pearson χ^2 test uses the statistic

$$\sum_j \frac{(O_j - E_j)^2}{E_j}, \tag{2}$$

where O_j indicates the observed frequency, and E_j is the expected frequency in cell j.

Table 1
Variable explanations.

Variables	Explanations
Y – with a bidder versus without a bidder	With a bidder = 1; without a bidder = 0
X_1 – upset price (reserve bid)	Lowest price requirements
X_2 – price of average square footage	Price of average square footage
X_3 – floor area	Building floor area
X_4 – square meters of ground	Square meters of ground surrounding building
X_5 – number of bids	Number of bids
X_6 – rented or not	Dummy variables; rent = 1, otherwise = 0
X_7 – handed over	Dummy variables; yes = 1, no = 0
X_8 – floor	Floor on which residence is on
X_9 – total stories	Total number of floors in building
X_{10} – width of road	Width of a road bordering the house
X_{11} – mass transit station distance	Distance from the mass transit station to the house
X_{12} – school distance	Distance from a school to the house
X_{13} – economic growth rate	Economic growth rate
X_{14} – unemployment rate	Unemployment rate

2.2.2. Likelihood ratio χ^2 test

The likelihood ratio χ^2 test uses the statistic

$$2\sum_j O_j \log\left(\frac{O_j}{E_j}\right), \tag{3}$$

where O_j indicates the observed frequency, and E_j is the expected frequency in cell j.

2.3. Random effects: nonlinear mixed model

Sample fractions may give only a poor estimation of πi for some cells with few observations, and produce large standard errors (Agresti, 2007). Collect (2003) proposed that overdispersion could be accommodated for by including a random effect in the model.

Davidian and Gallant (1993) used Gaussian quadrature methods for nonlinear mixed models. Davidian and Giltinan (1995) and Vonesh and Chinchilli (1996) provided an overview and a discussion of the general theoretical development of nonlinear mixed models.

The nonlinear mixed model can be written as

$$\log it(\pi_i) = \log\left[\frac{p_{ij}}{1-p_{ij}}\right] = u_i + \alpha + \beta_1 x_{ij}, \tag{4}$$

where u_i are independent $N(0,\sigma^2)$, and n_{ij} is the total number of observations.

3. An illustration with a portfolio of foreclosed mortgages

3.1. Sample description

The original data set collected for this study includes data on individual residential foreclosed mortgages from 1987 to 2007. Data were collected from the Taiwan Kaohsiung District Court in the Cianjin District, Kaohsiung City, Taiwan. The data set included 891 cases without a bidder and 989 cases with a bidder. The censoring time was at the end of 2007.

Table 1 shows the variables used in this study. They include one dependent variable and fourteen independent variables. When there is bidder for the foreclosed mortgage, the case is classified as a "bid". The bids for foreclosed mortgages are associated with the following factors: upset price (reserve bid) (X_1), average price per square foot (X_2), square footage (X_3), square meters of ground (X_4), number of bids (X_5), rented or not (X_6), handed over or not

Table 2
Maximum likelihood estimates.

	Original mode		Reduced model	
Parameter	Coefficients	P-value	Coefficients	P-value
Intercept	2.6273	<.0001***	3.3644	<.0001***
X_1 – upset price	0.000362	0.0366*	0.000419	0.0070**
X_2 – price of average square footage	−0.2561	<.0001***	−0.2530	<.0001***
X_3 – square footage	−0.00597	0.0007***	−0.00468	0.0002***
X_4 – square meters of ground	0.0197	0.2086		Deleted
X_5 – number of bids	−0.1370	0.0003***	−0.1233	0.0008***
X_6 – rented or not	−0.00509	0.9195		Deleted
X_7 – handed over or not	0.1205	0.4639		Deleted
X_8 – floor	−0.00581	0.6044		Deleted
X_9 – total stories	0.0573	<.0001***	0.0500	<.0001*
X_{10} – width of road	−0.00422	0.3279	Deleted	
X_{11} – MRT station distance	0.000580	0.0027**	0.000674	0.0004***
X_{12} – school distance	0.000563	0.0562	Deleted	
X_{13} – economic growth rate	0.0332	0.1083	Deleted	
X_{14} – unemployment rate	−0.4783	<.0001***	−0.5714	<.0001

* $P < .05$.
** $P < .01$.
*** $P < .0.001$.