International Journal of Forecasting 29 (2013) 548–562

Contents lists available at ScienceDirect

International Journal of Forecasting

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

A zero-adjusted gamma model for mortgage loan loss given default

Edward N.C. Tong *, Christophe Mues, Lyn Thomas

University of Southampton - Southampton Management School, Southampton, United Kingdom

ARTICLE INFO

Keywords:
Regression
Finance
Credit risk modelling
Mixture models
LGD
Basel II

ABSTRACT

The Internal Ratings Based (IRB) approach introduced in the Basel II Accord requires financial institutions to estimate not just the probability of default, but also the Loss Given Default (LGD), i.e., the proportion of the outstanding loan that will be lost in the event of a default. However, modelling LGD poses substantial challenges. One of the key problems in building regression models for estimating the loan-level LGD in retail portfolios such as mortgage loans relates to the difficulty of modelling their distributions, as they typically contain extensive numbers of zeroes. In this paper, an alternative approach is proposed where a mixed discrete-continuous model for the total loss amount incurred on a defaulted loan is developed. The model accommodates the probability of a zero loss and the loss amount given that a loss occurs simultaneously. The approach is applied to a large dataset of defaulted home mortgages from a UK bank and compared to two well-known industry approaches. Our zero-adjusted gamma model is shown to present an alternative and competitive approach to LGD modelling.

1. Introduction

The advanced Internal Ratings Based (IRB) approach outlined in the Basel II and Basel III Accords allows banks to calculate their own regulatory capital requirements based on internal credit risk model estimates (Basel Committee on Banking Supervision, 2005).

It requires banks to develop suitable methods for estimating three key parameters for each segment of their loan portfolios: PD (probability of default in the next 12 months), LGD (loss given default, i.e., the proportion of the outstanding loan that will be lost in the event of a default) and EAD (exposure at default).

For consumer credit, probability of default modelling has been a main objective of credit scoring for several decades. However, the additional IRB requirement of having to model LGD has posed substantial challenges, partly because of the properties of its distribution. Datasets of defaulted loan observations for residential mortgage portfolios or other retail portfolios usually exhibit a large probability mass at zero where no losses have been incurred, either because the account has cured and returned to performing status, or, in the case of mortgage loans, because the property has subsequently been repossessed and the sale price covered the loan balance at default adequately (Leow & Mues, 2012; Loterman, Brown, Martens, Mues, & Baesens, 2012; Thomas, Matuszyk, & Moore, 2012). Also, whereas the actual LGD observations of some individual loan defaults may fall outside the (0, 1) range, as LGD is supposed to include all economic costs (e.g., additional collection costs) and recoveries (e.g., penalties paid), the model estimates themselves are expected to be constrained to this interval.

Although the LGD research literature has traditionally focused more on corporate loan portfolios, LGD modelling for residential mortgages is a growing research area, given the impact of the new Accords on consumer lending, and the importance of mortgage loss estimation in the current financial context. One published approach for mortgages involved modelling LGD directly using ordinary

* Corresponding author.
E-mail address: E.Tong@soton.ac.uk (E.N.C. Tong).

0169-2070/$ – see front matter
http://dx.doi.org/10.1016/j.ijforecast.2013.03.003

least squares regressions (Qi & Yang, 2009). Their approach was developed using data from private mortgage insurance companies for a set of high loan-to-value loans, and although the ordinary least squares model was used, its use could be criticized because of the non-normal distribution of LGD.

Alternatively, a two-stage approach has been introduced in industry by Lucas (2006) and further investigated in the academic literature by Leow and Mues (2012). The method incorporates a probability of repossession (foreclosure) model developed with logistic regression and a haircut model using an OLS regression. The haircut represents the discount factor to be applied to the estimated sale price of the property, given that repossession occurs. The two models are then combined to produce an expected loss percentage given default. They showed the two-stage approach to perform better than the single-stage approach with a standard OLS regression.

There has also been interest recently in using quantile regressions or quantiles of model estimates to obtain LGD predictions (Somers & Whittaker, 2007; Zhang & Thomas, 2012). Somers and Whittaker (2007) have argued that using the low tail of the property value is more predictive of probable losses than the average value estimates in those settings where no loss is incurred in most accounts.

Where censored approaches to LGD modelling are concerned, Tobit regressions have been suggested as one of the methods to be used for modelling the restricted range of the LGD distribution (Bellotti & Crook, 2012). In a Tobit regression, the observed range of the response variable has a Gaussian distribution; however, Sigrist and Stahel (2012) introduced a censored model which allows the response to be Gamma distributed. For LGD data fitted with the Tobit model, the Gaussian assumption may not be suitable, and therefore the censored Gamma regression was developed to overcome the skewed nature of this interval. They also proposed a zero-inflated Gamma model for excess zeroes which were dealt with in probit regression.

Most of the existing literature on LGD modelling in the consumer and corporate credit risk domains has focused on modelling the LGD distribution directly. However, this distribution is known to be challenging to model accurately, due partly to its strongly unimodal or sometimes bimodal nature and lack of predictive characteristics. Therefore, in this paper, rather than modelling LGD (i.e. the loss as a *proportion* of the exposure) directly, we propose to model the incurred financial loss *amount*. Once an estimate of the amount has been obtained, one can then simply infer the LGD parameter by dividing the predicted loss by the loan balance or exposure.

In our proposed approach, the loss amount is modelled as a continuous response variable using a semi-parametric discrete-continuous mixture model approach with the zero-adjusted gamma distribution. Firstly, since the non-zero or positive loss amount exhibits heavy right-skewness, it is modelled using the gamma distribution. Both the mean and the dispersion of the positive loss amount are modelled explicitly as functions of explanatory variables. Secondly, the probability of the (non-)occurrence of a zero loss amount is modelled using a logistic-additive model. All of the mixture model components, i.e., the logistic-additive component for the probability of a zero loss and the log-additive components for the mean and dispersion of the loss amount conditional on there being a positive loss, are estimated using loan-level application and behavioural characteristics and house price index (HPI) covariates. The LGD parameter is then estimated by dividing the predicted loss amount by the loan balance.

When modelling the relationship between the response variable and continuous covariates, past credit risk research has focused on categorizing such covariates using binning methods. Such techniques can be arbitrary and result in a loss of information and precision for the estimated coefficients (Harrell, 2001; Royston, Altman, & Sauerbrei, 2006). Categorization also assumes that the relationship between the response and the covariate is flat within intervals, which may be unreasonable. Another common method would be to assume that continuous covariates are related to the response variable linearly, which would be incorrect for non-linear relationships. For example, such a method would not allow either the magnitude or the sign of coefficients to vary according to the range of covariate values. Our approach adopts a semi-parametric route by allowing non-linear relationships with the loss amount response variable through the use of regression splines (Eilers & Marx, 1996). Exploiting such non-linear relationships will reduce the bias in the estimates, improve the predictive performance of the model, and offer additional insights into the effects of covariates, while still retaining a fair level of model interpretability (Harrell, 2001; Hastie, Tibshirani, & Friedman, 2009).

Although the proposed approach has not yet been attempted in the context of consumer lending (to the best of our knowledge), the concept of estimating the expected loss amount for the exposures in a portfolio has been proposed previously in insurance modelling for policy claim amounts. Heller, Stasinopoulos, Rigby, and De Jong (2007) developed a discrete-continuous mixture model for estimating the total claim amount at a policy level from a portfolio of motor insurance policies. They used two components—the negative binomial distribution for modelling the number of claims for individual policies, and the inverse Gaussian for the claim amount given that a claim occurred. With the risk factors for prospective policy holders, the expected total claim size is then obtained from the product of the expected number of claims and the expected claim size for an individual claim.

Other discrete-continuous mixture models which use a mixture of Bernoulli and beta random variables have also been developed for the recovery rate modelling of corporate loans by Calabrese (2010) and Calabrese and Zenga (2010). They propose two logistic regression models for the recovery rates at the 0 and 1 end-points. For the (0, 1) interval, a joint beta regression model is developed to accommodate skewness and heteroscedastic errors by modelling the mean and dispersion of the response variable jointly. However, note that these methods can only be used to model LGD directly, not to model the loss amount itself.

To validate our approach empirically, the zero-adjusted gamma model is applied to a large dataset of defaulted