European Journal of Operational Research 240 (2015) 528–538

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Innovative Applications of O.R.

Support vector regression for loss given default modelling

Xiao Yao [*], Jonathan Crook, Galina Andreeva

Credit Research Centre, The University of Edinburgh Business School, 29 Buccleuch Place, Edinburgh EH8 9JS, UK

ARTICLE INFO

Article history:
Received 18 July 2013
Accepted 27 June 2014
Available online 5 July 2014

Keywords:
Support vector regression
Loss given default
Recovery rate
Credit risk modelling

ABSTRACT

Loss given default modelling has become crucially important for banks due to the requirement that they comply with the Basel Accords and to their internal computations of economic capital. In this paper, support vector regression (SVR) techniques are applied to predict loss given default of corporate bonds, where improvements are proposed to increase prediction accuracy by modifying the SVR algorithm to account for heterogeneity of bond seniorities. We compare the predictions from SVR techniques with thirteen other algorithms. Our paper has three important results. First, at an aggregated level, the proposed improved versions of support vector regression techniques outperform other methods significantly. Second, at a segmented level, by bond seniority, least square support vector regression demonstrates significantly better predictive abilities compared with the other statistical models. Third, standard transformations of loss given default do not improve prediction accuracy. Overall our empirical results show that support vector regression techniques are a promising technique for banks to use to predict loss given default.

1. Introduction

The introduction of the Basel II and Basel III Accords (BIS, 2005a, 2005b, 2011) requires that banks in the G20 countries hold specified amounts of capital to reduce the chance of their insolvency. The amount of capital required under the Internal Rating Based (IRB) advanced approach is based on the calculation of the proportions of defaulted loans that the bank will never recover, termed Loss Given Default (LGD). Similarly the proportion has been recovered can be defined as recovery rate (RR) equals to one minus LGD. Yet compared with the extensive research on modelling the probability of default, there is relatively little research on LGD, and that which has been published shows very poor predictive accuracy. In this paper we present improved support vector regression (SVR) models that give substantial increases in predictive accuracy compared with previously published methods.

Two types of predictive models have been applied in the empirical literature: parametric and non-parametric. Among the parametric models the most popular are linear regression models that have shown robustness and effectiveness in LGD prediction and explanation. Acharya, Bharath, and Srinivasan (2007) conclude from including the industry distress dummies into a linear regression model that industry distress conditions have negative effects on the RR of defaulted firms' debts. Their results suggest RR falls during distress periods due to both the downward trend in asset values and liquidity constraints. Qi and Yang (2009) in a study of LGD of residential mortgages demonstrate that LGD can be explained by linear regression that includes debt characteristics, with loan-to-value playing the single most important role. These results are confirmed by Khieu, Mullineaux, and Yi (2012) who estimate RR of bank loans with loan characteristics, borrower characteristics and macroeconomic conditions. They suggest loan characteristics are more significant determinants of RR than the other factors. Leow, Mues, and Thomas (2013) investigate the role of macroeconomic variables in two retail loans data sets. They find that the inclusion of macroeconomic variables can improve the prediction of residential mortgage LGD but bring little improvements for personal loan LGD.

Empirical LGD distributions are often bi-modal and usually bounded between [0, 1], suggesting that a linear regression model might fit poorly. Therefore, in order to improve the fit and predictive accuracy of the model, various transformations of LGD have been tried prior to the modelling stage. Gupton and Stein (2002) propose to transform the distribution of LGD into a normal distribution by a beta distribution function and then to model the transformed target with nine factors. They conduct extensive validation studies showing that such beta transformed linear regression gives better predictions than historical average methods. Another attractive alternative to linear regression is a generalized linear model such as a fractional response model. Dermine and Neto De Carvalho (2006) employ a complementary log–log model to predict

[*] Corresponding author.
E-mail address: X.YAO-2@sms.ed.ac.uk (X. Yao).

http://dx.doi.org/10.1016/j.ejor.2014.06.043
0377-2217/© 2014 Elsevier B.V. All rights reserved.

the cumulative RR of corporate loans from a Portuguese bank and report the R^2 as 0.13 for the 12-month prediction. Jacobs and Karagozoglu (2011) propose a beta-link generalized linear model to estimate LGD at firm and instrument levels jointly and report a significant improvement in terms of both in-sample and out-of-sample performances. Leow and Mues (2011) investigate a two-stage model to predict the LGD of UK residential mortgage loans with a combination of a probability of repossession model and a haircut model (a model that predicts a proportion of lost value for a repossessed property). This study suggests that such a two-stage modelling approach works better than a single-stage model. Calabrese (2010) applies an inflated beta regression model to predict RR of loans from The Bank of Italy where the dependent variable is assumed as a mixture of a continuous beta distribution on $(0,1)$ and a discrete Bernoulli distribution to model the probability mass at the boundaries 0 and 1. This study shows that the out-of-sample prediction of the inflated beta regression model outperforms fractional response regression models in terms of both MSE and MAE. Bellotti and Crook (2012) benchmark a number of different transformations and algorithms to predict the LGD for a credit cards data set. Surprisingly, they find that linear regression (OLS) with no variable transformations gives greater predictive accuracy.

Although parametric models are simple to implement and easy to explain, past research reports rather poor predictions of LGD, and generalized linear regression models do not achieve significant improvements compared with linear regression. Zhang and Thomas (2010) compare both linear regression and survival regression for modelling RR of personal loans from a UK bank, and report the out-of-sample R^2 as low as 0.0904 for linear regression, and the parametric survival models exhibit even poorer predictions. It is also surprisingly interesting to see that given the versatility of the distribution allowed in the Cox approach, the predictive accuracies can still not be improved compared with linear regression model. Similar evidence provided by Bellotti and Crook (2012) show the model fit of simple linear regression to be rather weak with R^2 of 0.1428, and still the predictions of this model outperform the other ones including logit and probit models.

In contrast, non-parametric methods provide much more flexibility in modelling LGD, although literature on this topic is not as extensive as for parametric models. One of the major advantages of non-parametric methods is that they do not assume a specific distribution for LGD. Unlike parametric models which imply a specific form of the LGD distribution, non-parametric methods do not make any prior assumptions when fitting a regression model. This often leads to a better performance compared with parametric techniques, as reported by previous research. For example, Bastos (2010) compares parametric fractional response regression and a non-parametric regression tree model to forecast bank loans RR and finds that the latter is superior. More strong evidence comes from Qi and Zhao (2011) who compare six modelling methods including four parametric statistical models and two data mining techniques (decision trees and neural networks) for a mixed portfolio of bonds and loans. They find non-parametric methods perform significantly better than other parametric methods in terms of both model fit and prediction accuracy. Tong, Mues, and Thomas (2013) develop a zero adjusted gamma model to predict LGD of a UK bank where the non-parametric smoothing splines are incorporated into the predictor of a mixture gamma distribution. The findings show that such a semi-parametric formulation gives favorable out-of-sample predictions compared with the traditional linear regression.

This study focuses on another promising non-parametric data mining technique: support vector machines (SVM) and application to LGD modelling. SVM was first studied by Vapnik (1995, 1998) and are widely applied in engineering, bioinformatics and decision sciences. Previous research has revealed that SVM can not only handle non-linear problems well, but also avoid the over-fitting problem that is common in neural networks based on the principle of structural risk minimization. SVM models have been widely applied in credit risk modelling as a tool to solve classification problems such as in credit scoring, i.e. to classify credit applicants into 'Good' or 'Bad' risks. On the other hand, support vector regression (SVR) adapted to regression problems has been developed and effectively applied to non-linear regression and to time series prediction problems. However, until now only one published paper, by Loterman, Brown, Martens, Mues, and Baesens (2011), has investigated the application of SVR to LGD modelling. They conduct a comprehensive benchmarking study on six retail loan data sets with 24 techniques, some of which are two-stage models including both linear and non-linear techniques, and they find that non-linear techniques including neural networks and SVR models consistently outperform other traditional linear methods. But they do not make any further improvements on SVR models.

Our paper makes three distinct contributions based on the analysis of the RR of corporate bonds. First, the predictive performance of RR is modelled by using different intercepts or dummy variables to explain the unobservable heterogeneity of different bond seniorities. Second, SVR models are applied to losses from corporate bonds for the first time. In addition, the dataset comprises a longer time series of observations than previous studies and uses a more comprehensive set of predictor variables, including the debt characteristic, the accounting ratios from obligors' financial statements. Macroeconomic factors are also included to allow for any possible systematic differences in LGD over time. Third, the paper investigates whether transforming LGD values using a logistic or beta transformation prior to analysis can improve SVR model fitting and prediction accuracy. The results show that all SVR models substantially outperform other statistical models in terms of both model fit and out-of-sample predictive accuracy, and we find that the robustness of SVR models is comparable to that of statistical models. However, a logistic or beta transformation prior to modelling does not provide any improvement in prediction.

The rest of the paper is organized as follows. Section 2 presents the models, the data used in this research is described in Section 3. Section 4 discusses the results and conclusions are drawn in Section 5.

2. Models

In this section both parametric regression and SVR models are presented and the proposed SVR models are elaborated in more detail. Note that in line with literature and our data the target variable is RR instead of LGD.

2.1. Linear regression

Previous empirical research shows that linear regression models appear to be of comparable predictive accuracy as other more complicated statistical models (Bellotti & Crook, 2012; Qi & Zhao, 2011) even though they have the potential risk to make predictions out of the range between 0 and 1. Consider a dataset $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ with the covariates $\mathbf{x}_i \in R^m$ which is m-dimensional and the related dependent variable is $y_i \in R$, and $\boldsymbol{\beta}$ denotes a vector of population parameters. The linear regression model is given as

$$\begin{aligned} y_i &= \boldsymbol{\beta}^T \mathbf{x}_i + \varepsilon_i \\ \varepsilon_i &\sim N(0, \sigma^2), \end{aligned} \tag{1}$$

Maximum likelihood methods can be applied to estimate the parameters.