International Journal of Forecasting 28 (2012) 161–170

Contents lists available at SciVerse ScienceDirect

International Journal of Forecasting

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

Benchmarking regression algorithms for loss given default modeling

Gert Loterman [a], Iain Brown [b,*], David Martens [c], Christophe Mues [b], Bart Baesens [b,d]

[a] Department of Business Administration and Public Management, University College Ghent, Ghent University, Voskenslaan 270, B-9000 Ghent, Belgium
[b] School of Management, University of Southampton, Southampton, S017 1BJ, United Kingdom
[c] Faculty of Applied Economics, University of Antwerp, Prinsstraat 13, B-2000 Antwerpen, Belgium
[d] Department of Decision Sciences & Information Management, Catholic University of Leuven, Naamsestraat 69, B-3000 Leuven, Belgium

ARTICLE INFO

Keywords:
Basel II
Credit risk
LGD
Data mining
Prediction

ABSTRACT

The introduction of the Basel II Accord has had a huge impact on financial institutions, allowing them to build credit risk models for three key risk parameters: PD (probability of default), LGD (loss given default) and EAD (exposure at default). Until recently, credit risk research has focused largely on the estimation and validation of the PD parameter, and much less on LGD modeling. In this first large-scale LGD benchmarking study, various regression techniques for modeling and predicting LGD are investigated. These include one-stage models, such as those built by ordinary least squares regression, beta regression, robust regression, ridge regression, regression splines, neural networks, support vector machines and regression trees, as well as two-stage models which combine multiple techniques. A total of 24 techniques are compared using six real-life loss datasets from major international banks. It is found that much of the variance in LGD remains unexplained, as the average prediction performance of the models in terms of R^2 ranges from 4% to 43%. Nonetheless, there is a clear trend that non-linear techniques, and in particular support vector machines and neural networks, perform significantly better than more traditional linear techniques. Also, two-stage models built by a combination of linear and non-linear techniques are shown to have a similarly good predictive power, with the added advantage of having a comprehensible linear model component.

1. Introduction

With the recent turmoil in credit markets, the topic of credit risk modeling has arguably become more important than ever before. Also, to comply with the Basel II Accord introduced at around the same time, financial institutions have had to invest heavily in the development of improved credit risk models. The Basel II Capital Accord sets out a framework that regulates the minimum amount of capital that financial institutions are required to hold as a safety cushion against unexpected credit-, market- and/or operational losses. More specifically, the accord allows institutions to build credit risk models for three key risk parameters: probability of default (PD), loss given default (LGD) and exposure at default (EAD). From these, the regulatory capital is then derived.

So far, credit risk research has largely focused on the estimation and validation of the PD parameter. On the other hand, the LGD parameter measures the economic loss, expressed as a percentage of the exposure, in case of default. In other words, LGD is the proportion of the remaining loan amount that the bank would not be able to recover. This parameter is a crucial input to the Basel II regulatory capital calculations, as it enters the capital requirement formulas in a linear way (unlike PD, which therefore has less of a direct effect on minimum capital). Hence, any changes in the LGD estimates produced by

* Corresponding author.
E-mail addresses: gert.loterman@hogent.be (G. Loterman), i.brown@soton.ac.uk (I. Brown), david.martens@ua.ac.be (D. Martens), c.mues@soton.ac.uk (C. Mues), b.m.m.baesens@soton.ac.uk, bart.baesens@econ.kuleuven.be (B. Baesens).

0169-2070/$ – see front matter
doi:10.1016/j.ijforecast.2011.01.006

Table 1
Overview of dataset characteristics.

Dataset	Type	Inputs	Total size	Training size	Test size
BANK1	Personal loans	44	47 853	31 905	15 948
BANK2	Mortgage loans	18	119 211	79 479	39 732
BANK3	Mortgage loans	14	3 351	2 232	1 119
BANK4	Revolving credit	12	7 889	5 260	2 629
BANK5	Mortgage loans	35	4 097	2 733	1 364
BANK6	Corporate loans	21	4 276	2 851	1 425

models have a strong bearing on the capital of a financial institution, and thus its long-term strategy as well. It is therefore crucial to have models that estimate LGD as accurately as possible. This is not straightforward, however, as industry models typically show low R^2 values, particularly for consumer lending portfolios. Such models are often built using ordinary least squares regression or regression trees (Bastos, 2009; Bellotti & Crook, 2007; Caselli & Querci, 2009; Gupton & Stein, 2002). Using a set of six real-life default loss datasets, this first large-scale LGD benchmarking study investigates whether or not other approaches can improve the prediction performances of these LGD models.

The remainder of this paper is organized as follows. Section 2 gives an overview of the regression techniques examined, the performance metrics used to evaluate and compare the models, the available datasets and the experimental set-up used in this study. Next, Section 3 reports and discusses the experimental results obtained, and Section 4 concludes the paper.

2. Datasets, techniques and methodology

This section describes the datasets, regression techniques and performance metrics used in this study. We subsequently outline the experimental benchmarking framework used to assess the performances of the various models built on these datasets.

2.1. Dataset characteristics

Table 1 shows the characteristics of six real-life LGD datasets obtained from a series of financial institutions, each of which contains loan-level data about defaulted loans and their resulting losses. The number of dataset entries varies from a few thousands to just under 120,000 observations. The number of available input variables ranges from 12 to 44. The types of loan portfolios included are personal loans, corporate loans, revolving credit and mortgage loans. The empirical distribution of LGD values observed in each of the datasets is displayed in Fig. 1. Note that the LGD distribution in consumer lending often contains spikes around LGD $=$ 0 (in which case there was a full recovery) and/or LGD $=$ 1 (no recovery). Also, some of the datasets include some LGD values that are either negative (e.g., because of penalties paid, gains in collateral sales, etc.) or larger than 1 (e.g., due to additional collection costs incurred), while in other datasets, values outside the unit interval were truncated to 0 or 1 by the banks themselves. Importantly, LGD does not appear to be normally distributed in any of these datasets.

2.2. Regression techniques

The experiments comprise a selection of one-stage and two-stage techniques, listed in Table 2. The one-stage techniques can be divided into linear and non-linear techniques. The linear techniques included in our study model the (original or transformed) dependent variable as a linear function of the independent variables, whereas the non-linear techniques fit a non-linear model to the dataset. Two-stage models are a combination of the aforementioned one-stage models. These either combine the comprehensibility of an OLS model with the added predictive power of a non-linear technique, or use one model to first discriminate between zero- and higher LGDs and a second model to estimate LGD for the subpopulation of non-zero LGDs. Further details are provided in Table 2.

2.3. Performance metrics

Several different performance metrics can be used to evaluate the extent to which the models produce accurate predictions for the dependent variable. Each of the metrics listed in Table 3 has its own method of quantifying the model performance, as specified in the second column of the table. The next two columns show the metric values for the worst (null model) and best possible performances, respectively.[1] The final column indicates whether the metric evaluates calibration or discrimination (Van Gestel & Baesens, 2009). Calibration indicates how close the predicted values are to the observed values, whereas discrimination refers to the ability to provide an ordinal ranking of the dependent variable. A good ranking does not necessarily imply a good calibration. On the other hand, a good calibration always implies good discrimination.

2.4. Experimental set-up

After data preprocessing, the models are built on the training sets and performance metrics are reported for the test sets. Several of the techniques included require setting or tuning parameters, and/or benefit from variable selection; further details of both are provided below, along with the procedure used to assess whether the observed performance differences are statistically significant or not.

[1] Note that the R^2 measure defined here could possibly lie outside the [0,1] interval when applied to non-OLS models. Although alternative generalized goodness-of-fit measures have been put forward for evaluating various non-linear models (see for example the R^2 of Nagelkerke, 1991), the measure defined in Table 3 has the advantage that it is widely used and can be calculated for all techniques.