QUANTITATIVE METHODS IN ECONOMICS
Vol. XVI, No. 1, 2015, pp. 108 – 115

APPLICATION OF MIXED MODELS AND FAMILIES OF CLASSIFIERS TO ESTIMATION OF FINANCIAL RISK PARAMETERS

Urszula Grzybowska, Marek Karwański
Department of Informatics
Warsaw University of Life Sciences – SGGW in Warsaw
e-mail: urszula_grzybowska@sggw.pl, marek_karwanski@sggw.pl

Abstract: The essential role in credit risk modeling is Loss Given Default (LGD) estimation. LGD is treated as a random variable with bimodal distribution. For LGD estimation advanced statistical models such as beta regression can be applied. Unfortunately, the parametric methods require amendments of the "inflation" type that lead to mixed modeling approach. Contrary to classical statistical methods based on probability distribution, the families of classifiers such as gradient boosting or random forests operate with information and allow for more flexible model adjustment. The problem encountered is comparison of obtained results. The aim of the paper is to present and compare results of LGD modeling using statistical methods and data mining approach. Calculations were done on real life data sourced from one of Polish large banks.

Keywords: LGD, mixed models, random forests, gradient boosting

INTRODUCTION

New Basel Accords introduced a possibility of applying IRB systems in banks for the need of risk parameters estimation. Within that approach three key risk parameters should be estimated: PD (Probability of Default), LGD (Loss Given Default) i. e., the percentage of total exposure at the time of default that cannot be recovered and EAD (Exposure at Default). Contrary to PD, LGD estimation has received much less attention so far. If fact it has been a subject of more intense scientific research for hardly five years now. The reasons are, among others, lack of unified definitions of default or economic loss as well as the scarcity of LGD data. Moreover, LGD can exhibit difficult behavior. Its values are fractions, often with high concentration at 0 (full recovery) and/or at 1 (total loss). LGD is treated as a random variable, frequently with a bimodal distribution.

In the first section of the paper we briefly describe models utilized so far in LGD modeling. Our idea was to apply families of classifiers to modeling LGD. While parametric models are easier to explain, ensemble methods are able to better cope with often bimodal or highly skewed distribution of LGD. Therefore in the next section we present two ensemble models: gradient boosting and random forests that have to our knowledge not been applied yet in LGD modeling. In the final section we present the results of our research. We compare classical methods applied in LGD estimation with ensemble methods. We compare both approaches using graphical methods, among others the REC curve.

PARAMETRIC MODELS FOR LGD ESTIMATION

Industrial approach to LGD modeling

Capital loss in credit risk is represented by a random variable L. Its expected value can be calculated as:

$$E(L) = EAD \cdot LGD \cdot PD \tag{1}$$

In the paper we consider only LGD for individual transactions.

Models of LGD

LDG is expressed in percentage therefore it is a fractional target variable. LGD is usually modeled by regression methods, like fractional regression or beta regression [Loterman et al. 2012]. Fractional regression was introduced by Papke and Wooldridge in 1996 [Papke et al. 1996]. Fractional regression is a Generalized Linear Model (GLM) with logit link function.

$$G(\beta x) = \frac{1}{1+\exp(-\beta x)} \tag{2}$$

Beta regression was first applied to model proportions in 2004 by Ferrari and Cribari-Neto [Frerrari et al. 2004]. The distribution function for $p > 0,\ q > 0$ and $y\epsilon(0,1)$ is given by:

$$f(y; p, q) = \frac{y^{p-1}(1-y)^{q-1}}{B(p,q)} = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1} \tag{3}$$

where $B(\cdot,\cdot)$ is a Beta function and $\Gamma(\cdot)$ denotes Gamma function. Ferrari and Cribari-Neto proposed a transformation of p and q into a location/mean parameter μ and dispersion/precision parameter ϕ. We have:

$$E(Y) = \frac{p}{p+q} = \mu \tag{4}$$

$$Var(Y) = \frac{pq}{(p+q)^2(p+q+1)} = \frac{\mu(1-\mu)}{\phi+1}, \text{ where } \phi = p + q \tag{5}$$

Ferrari and Cribari-Neto estimated parameters using maximum likelihood function and their approach was similar to the maximum likelihood method applied in GLM. Beta function is well suited to describe highly skewed data [Karwański et al. 2015]. Unfortunately, the model proposed by Ferrari and Cribari-Neto is restricted to the open interval (0,1). The observations on boundaries are neglected. Real data indicate that observations at boundaries appear with high frequencies. Only in the recent three years extensions of beta regression models covering the whole range [0,1] have been discussed. The first generalization of beta regression for the boundaries of the (0,1) interval was proposed by Ospina and Ferrari [Ospina et al. 2011]. The authors proposed a zero or one inflated beta regression for modeling fractional outcomes. In 2012 Calabrese proposed a mixed model

$$B_{inf01}(y,\pi_0,\pi_1,\mu,\phi)=\begin{cases}\pi_0 & for\ y=0\\ \pi_1 & for\ y=1\\ \pi_0+[1-\pi_0-\pi_1]f(y;\mu,\phi) & for\ y\epsilon(0,1)\end{cases} \tag{6}$$

where $f(y;\mu,\phi)$ is a beta distribution [Calabrese 2012]. The model is a mixture of Bernoulli distribution and Beta distribution.

A zero and one beta inflated model was first introduced in 2014 by Xiao Yao [Xiao Yao et al. 2014]. It can be written in the following way:

$$B_{inf01}(y,\pi,\psi,\mu,\phi)=\begin{cases}\pi(1-\psi) & for\ y=0\\ \pi\psi & for\,y=1\\ (1-\pi)f(y;\mu,\phi) & for\ y\epsilon(0,1\,)\end{cases} \tag{7}$$

$$E(y)=\pi\psi+(1-\pi)\mu \tag{8}$$

In our calculations we have applied three different models based on beta distribution. Two of them belong to the family of mixed models. The parameters of our models were estimated by maximum likelihood method.

FAMILIES OF CLASSIFIERS

Random forests and gradient boosting are extensions of regression trees, that is simply the partition of the space X, which consists of predictors of target variable y, into disjoint regions R$_j$. We will describe briefly both methods.

Random forests

Random forests were introduced in 2001 by L. Breiman as a method of classification [Breiman 2011]. In this approach a large number of simple trees is constructed with a random sample of predictors taken before each node is split. The object is classified based on an average vote of the set of de-correlated trees [Berk 2008]. The random forest algorithm can be described as follows [Hastie et al. 2009]:

```
Create N bootstrap samples {S1,…,SN} out of a data set S as follows:
Si: random drawings of |N| elements from S with replacement
For each i=1,…,N  select random set of attributes {X*}   h*i=Learn(Si ; X*)
```

Output H=[{h*1,…,h*N}, majority Vote].

Gradient boosting

Gradient boosting was introduced by J. Fridman in 1999. In gradient boosting, similarly as in random forests, a family of trees is grown. Each tree is constructed based on a previous one in such a way that one minimizes a given loss function in the gradient direction. Gradient boosting can be described in the following way [Berk 2008]:

Let Y=learn(h(X)) and $b_m(x)$ be a set of predictors of Y

$$h(x) = \sum_{m=1}^{M} b_m(x) = \sum_{m=1}^{M} \beta_m b(x; \theta_m) \tag{8}$$

Put: $h_0(x) \leftarrow 0$. For m=1,…,M and the loss function L( ):

$$r(x,y) \leftarrow -\frac{d}{dh} L(h_{m-1}(x), y) \tag{9}$$

$$b_m \leftarrow \arg\min_b \sum_{(x,y)} \big(b(x) - r(x,y)\big)^2 \tag{10}$$

$$h_m(x) \leftarrow h_{m-1}(x) + v \cdot b_m(x) \tag{11}$$

In gradient boosting one applies a property that for small ν:

$$\sum_{(x,y)} L(h_{m-1}(x) + b_m(x), y) \approx \sum_{(x,y)} L(h_{m-1}(x), y) + \sum_{(x,y)} \frac{d}{dh} L(h_{m-1}(x), y)\, b_m(x) \tag{12}$$

DATA

Our calculations were made based on data covering small and medium enterprises sourced from one of large Polish banks. Data was collected in a few operational data bases built for the needs of various bank departments. It comprised both information about the client as well as about products offered. The data was collected over the period of 3 years with the defaults registered until October 2007 and the recovery process observed till October 2008. In the calculations 12000 observations and 12 variables were used. The data was normalized to make the comparison possible. The selected variables used in calculations are described in Table 1.

Table 1. Explanatory variables used in the analysis

X1	Average monthly withdrawal in the last 3 months
X2	Average overdraft balance in the last 6 months
X3	Average credit balance in the last 3 months
X4	Average balance in the last 3 months
X5	Trend for average balances in the last 12 months
X6	Ratio of total balance to average balance in the last 3 months
X7	Total interest delinquency ratios at the time of analysis (PIT[*] approach)
X8	Total capital delinquency ratios at the time of analysis (PIT approach)
X9	Average increas in capital arrears in the last 6 months
X10	Total amount of monthly payments done by the client
X11	Coefficient of debt/loan repayment
X12	Status of the first loan account

Source: own preparation

RESULTS AND CONCLUSIONS

In our research we have applied three regression models. The models were selected in accordance with GLM model selection. The first model was a beta regression model proposed by Ferrari and Cribari-Neto (we denote it by beta model). The second one was a zero-one inflated beta regression with constant parameters π, ψ and ϕ (we denote it by inflated beta model 1) and the last one was a zero-one inflated beta regression with parameters following logistic distribution (we denote it by inflated beta regression 2). The parameters of our regression models were estimated by maximum likelihood method. The calculations were done in SAS 9.4.

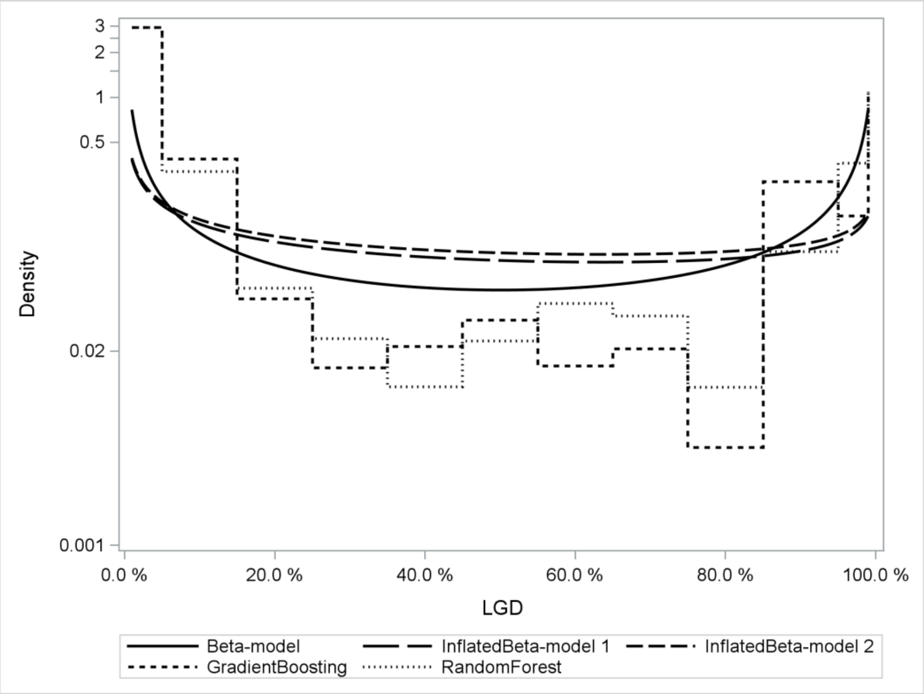
We have also applied two ensemble methods: random forest and gradient boosting. The calculations were done in SAS Enterprise Miner ver. 13.2. To make the model comparison possible, the target variable LGD in ensemble modeling was categorized into ten classes.

The research that is based on two different approaches encounters the problem of comparison between obtained results. No sufficient theory of comparing statistical and data mining models has been developed so far. Namely, regression models are based on minimization of residuals, while data mining methods are based on information maximization. In the latter Gini coefficient or entropy measure are commonly used. In fact, there are no popular measures that compare simultaneously both approaches.

[*] PIT (point in time), a methodology of evaluating risk parameters opposite to TTC

In order to compare the results of our research we have used an underrated measure called REC (Regression Error Characteristic) [Bi et al. 2003]. REC curve is a powerful tool for visualizing and comparing model results. REC curves plot the error tolerance (loss function) versus the percentage of points predicted within the tolerance (cumulative distribution). The REC curve visually presents commonly-used statistics. The area-over-the-curve (AOC) is a biased estimate of the expected error. The R^2 statistic can be estimated using the ratio of the AOC for a given model to the AOC for the null model. Moreover, the shapes of these curves give some additional information about model goodness-of-fit.

Figure 1. Density functions $f(LGD|x's\ equal\ their\ averages)$ estimated by various models

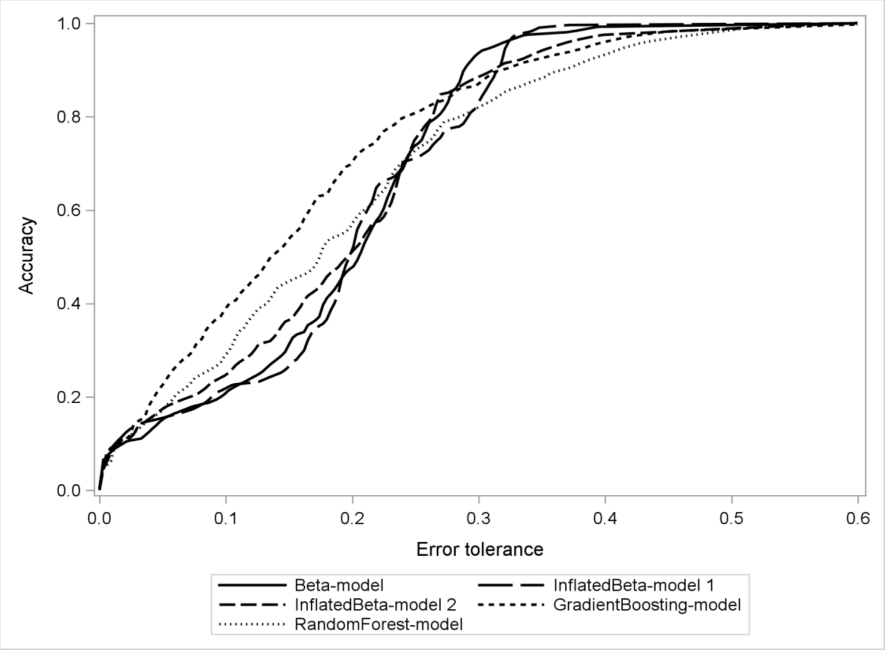


Source: own calculations

The plots of density functions for LGD shown on Figure 1 indicate that all regression models and families of classifiers give similar results. The figure was plotted for average values of all covariates.

The REC curves in Figure 2 show that gradient boosting outperforms other methods for the majority of residual values. Beta regression models exhibit similar behavior.

Figure 2. Comparison of REC curves for all beta regression models, random forests and gradient boosting



Source: own calculations

Table 2. Area Over the REC Curve as the measure of goodness of fit

	Area Over the REC Curve (AOC)
Beta model	0.1831
Inflated beta model 1	0.1868
Inflated beta model 2	0.1812
Gradient boosting	0.1525
Random forest	0.1813

Source: own preparation

The gradient boosting model clearly dominates the others over the greater part of range of possible errors. On the contrary, the performances of beta model and both beta inflated regression models are harder to compare. For smaller errors inflated beta model 2 dominates, but for larger errors beta model overcomes. The decision which of these two models is preferable may be domain dependent. Random forests behave similarly as the area over the curve (i.e., the expected error) is almost equal to that of inflated beta regression 2.

The aim of our research was to show that ensemble methods can be applied in LGD estimation. The results revealed that families of classifiers not only can successfully be applied in LGD modeling but also gradient boosting outperforms all considered beta regression models.

REFERENCES

Berk R. A. (2008) Statistical learning from a regression perspective, Springer, New York.

Bi J., Bennett K. (2003) Regression error characteristic curves, Proceedings of the 20th International Conference on Machine Learning.

Breiman L. (2001) Random Forests, Machine Learning, Vol. 45.

Calabrese R. (2012) Regression model for proportions with probability masses at zero and one. Working Paper, http://www.ucd.ie/geary/static/publications/workingpapers/gearywp201209.pdf.

Ferrari S. L. P., Cribari-Neto F. (2004) Beta Regression for Modeling Rates and Proportions, Journal of Applied Statistics, No. 31.

Hastie T., Tibshirani R., Friedman J. (2009) The elements of statistical learning. Data Mining, Inference and Prediction. Second Edition, Springer, New York.

Karwański M., Gostkowski M., Jałowiecki P. (2015) LGD Modeling: an application to data from a polish bank, On-line Risk Journals available on http://www.risk.net/.

Loterman G., Brown I., Martens D., Mues Ch., Baesens B. (2012) Benchmarking regression algorithms for loss given default modeling, International Journal of Forecasting, No. 28.

Ospina R., Ferrari S. L. P. (2012) A General Class of Zero-or-one inflated beta Regression Models, Computational Statistics and Data Analysis, No. 56.

Papke L, Wooldridge J. (1996) Econometric Methods for Fractional Response Variables with an Application to 401(K) Plan Participation Rate, Journal of Applied Econometrics, Vol. 11.