International Review of Financial Analysis 47 (2016) 70–85

Contents lists available at ScienceDirect

International Review of Financial Analysis

Validation of default probability models: A stress testing approach☆

Fábio Yasuhiro Tsukahara [a], Herbert Kimura [b], Vinicius Amorim Sobreiro [b,*], Juan Carlos Arismendi Zambrano [c,d]

[a] Midway Finance, 500 Leão XIII, São Paulo, São Paulo, 02526-000, Brazil
[b] University of Brasília, Department of Management, Campus Darcy Ribeiro, Brasília, Federal District, 70910-900, Brazil
[c] Department of Economics, Federal University of Bahia, Rua Barão de Jeremoabo, 668-1154, Salvador, Brazil
[d] ICMA Centre, Henley Business School, University of Reading, Whiteknights, Reading RG6 6BA, United Kingdom

ARTICLE INFO

Article history:
Received 15 July 2015
Received in revised form 31 May 2016
Accepted 28 June 2016
Available online 8 July 2016

Keywords:
Portfolio
Credit risk
Banking
Default probability
Validation techniques

ABSTRACT

This study aims to evaluate the techniques used for the validation of default probability (*DP*) models. By generating simulated stress data, we build ideal conditions to assess the adequacy of the metrics in different stress scenarios. In addition, we empirically analyze the evaluation metrics using the information on 30,686 delisted US public companies as a proxy of default. Using simulated data, we find that entropy based metrics such as measure *M* are more sensitive to changes in the characteristics of distributions of credit scores. The empirical sub-samples stress test data show that *AUROC* is the metric most sensitive to changes in market conditions, being followed by measure *M*. Our results can help risk managers to make rapid decisions regarding the validation of risk models in different scenarios.

1. Introduction

In summary, the Basel II Accord allows banks to develop internal models for measuring risk (BCBS, 2006; Kiefer, 2009) and the Basel III Accord aims to enhance the stability of the financial system by strengthening risk coverage and highlighting the importance of on- and off-balance sheet risks, including derivatives exposure (BCBS, 2011). In addition, the Accords also require validation of risk models to determine,[1] qualitatively and quantitatively, the models' performance and adherence to the institution's goals. In this context, Stein (2007) states that the validation process is of great importance, since it allows the benefits generated by the use of risk models to be fully obtained. However, effectively validating risk models is still a great challenge, because this is a recent aspect of banking regulation and the primary methods are still under development. In particular, credit model validation has major impediments, i.e., the small number of observations to accurately evaluate model performance (Lopez & Saidenberg, 2000).[2]

Many validation techniques of models for bank risk management have been proposed or submitted in recent years, for market risk (Alexander & Sheedy, 2008; Boucher, Daníelsson, Kouontchou, & Maillet, 2014), credit risk (Lopez & Saidenberg, 2000; Agarwal & Taffler, 2008), and model risk (Kerkhof & Melenberg, 2004; Alexander & Leontsinis, 2011; Alexander & Sarabia, 2012; Colletaz, Hurlin, & Pérignon, 2013). Blöchlinger (2012) presents a methodology where the validation of default probability (*DP*) is produced over credit rating methodologies. Medema, Koning, and Lensink (2009) proposes a practical methodology for validation of statistical models of *DP* for portfolio of individual loans where no credit rating can be associated. However, there are no studies that attempt to identify or guide managers regarding which model is most appropriate for a given situation. With regard to the methods for estimating credit risk parameters, *DP* models are, according to BCBS (2005a), those that have the most developed validation

☆ This document is a collaborative effort.
* Corresponding author.
E-mail addresses: fabioyat@gmail.com (F.Y. Tsukahara), herbert.kimura@gmail.com (H. Kimura), sobreiro@unb.br (V.A. Sobreiro), j.arismendi@icmacentre.ac.uk (J.C.A. Zambrano).

[1] For financial institutions to be able to use their internal models to calculate capital requirements, the Basel II Accord requires that the models be validated by an independent internal team. For this internal validation process it is necessary to develop techniques to consistently assess the performance of the models used. In this study, we present some validation techniques that are widely used in the financial market for assessment of models, such as *KS* and *AR*, and other less traditional measures, such as *CIER* and measure *M*.

[2] The growth of credit activity is an important aspect of economic development, because credit is a major source of funds for private and public organizations (Hagedoorn, 1996). However, increases in credit supply bring more exposure to credit risk and, in extreme cases, overreliance on credit can compromise the stability of the financial system (Abou-El-Sood, 2015; Arnold, Borio, Ellis, & Moshirian, 2012). Economic crises, such as the one in 2008, indicate a need for greater control and regulation of financial institutions by supervisors and for the development of risk management models. In this context, the Basel I, II, and III Accords are examples of how regulatory agencies are concerned with securing a solid international financial system; they are dynamically adjusting their requirements due to an ever-changing economic environment.

http://dx.doi.org/10.1016/j.irfa.2016.06.007
1057-5219/

methodology. Tasche (2006) separates the performance validation process for these models into two parts, discriminative ability and calibration.

Our contribution to the literature is twofold. First we evaluate the stress test the adequacy of the primary models for risk management and thereby support the decision-making of managers regarding the model selection process. More specifically, we present the characteristics and main properties of different techniques that allow a manager to choose among classic validation models, such as the Kolmogorov–Smirnov (*KS*) statistic, Accuracy Ratio (*AR*), and Brier Score, and newer validation models, such as the Conditional Information Entropy Ratio (*CIER*) and Measure *M*. The stress test simulation[3] is carried out in two phases: (i) an assessment of the performance of models to separate good and bad borrowers among the risk groups is performed, (ii) the accuracy of the probabilities estimated by each model is evaluated.[4] The models were applied to credit portfolios, which were compiled using Monte Carlo simulations, to identify good and bad borrowers and how the characteristics (e.g., dependencies or moments) of these portfolios impacted the results of the models. According to Zott (2003), when there are significant limitations on gathering empirical data and variables have complex interrelationships, simulation may be useful and can actually lead to superior insights into the phenomenon.[5] The objective of this study is not to exhaustively explore the subject but rather to enable managers to quickly identify a small number of optimal models.

Second, we analyze the default probability validation metrics using controlled sub-samples of market data. Our empirical stress analysis includes financial data of from 30,686 public US firms from 1950 and 2014, using delisting information as a proxy for default. We develop a methodology that aggregates different groups of years by high–low mean, variance, and correlation related to the financial explanatory variables. Although using empirical data does not allow as total control as using simulated data, the method gives some control over the distribution of credit scores and dependence among variables. Therefore, we can also analyze the behavior of *DP* evaluation metrics on empirical sub-sample data.

In the case of controlled stress simulations, for independent explanatory variables, we found that (i) the measure *M* was the only metric able to detect changes in the mean of the explanatory variables,[6] while there was no metric sensitive to changes in the variances; (ii) all metrics were very sensitive to the number of observations; therefore, the study can help in the validation of models for the retail and large corporations segment. In the case of controlled stress simulations, for dependent explanatory variables, we found that (i) the only metric that captured a performance decrease for both increases and decreases in the correlation parameter was measure *M*, all other measures exhibited an increase in performance as the strength of the correlation was decreased; (ii) modeling using the *T* copula and Gaussian copula provided no difference in the sensitivity results of the metrics.

The remainder of this study is structured as follows: in Section 2, a literature review of credit is presented; Section 3 and 4 address the aspects used to compare the models and their results; Section 5 presents a empirical application; in Section 6, the primary conclusions are presented and discussed.

[3] Simulated portfolios to study credit risk have been explored in the literature. For instance, Kalkbrener, Lotter, and Overbeck (2004) develops an importance sampling Monte Carlo technique to study capital allocation for credit portfolios and Jobst and Zenios (2005) use simulation to analyze the sensitiveness of credit portfolio values to default probability, recovery rates, and migration of ratings. In addition, Hlawatsch and Ostrowski (2011) study loss given default based on simulated datasets to analyze the synthesized loan portfolios.

[4] Since there are many classification techniques used for credit scoring (Baesens et al., 2003), performance measurement is necessary to assess model adequacy (Verbraken et al., 2014).

[5] Davis, Eisenhardt, and Bingham (2007) presents a reference to the theory developed using simulation methods.

[6] Explanatory variables are any variables that can lead to a causal explanation of the relationships in default, such as the ones included in the *Z-score* of Altman (1968).

2. Literature review

The Basel II Accord aims to improve the awareness of the financial institutions regarding their credit risk (Hakenes & Schnabel, 2011). The Basell II Accord first pillar aims to guide the calculation of minimum capital requirements, i.e., it reviews the main ideas presented in the Basel I Accord. The minimum capital requirement is calculated based on the Internal Rating Based (*IRB*) method, which is generally estimated internally by a bank based on the following parameters: (i) *DP*; (ii) Exposure at Default (*EAD*); (iii) Loss Given Default (*LGD*); and (iv) Maturity (*M*). It is worth noting that in the simplified version of the *IRB*, it is only necessary to calculate the *DP* value because the other parameters are defined by regulatory bodies. From this point of view, the calculation of *DP* becomes crucial.

2.1. Validation tests for default probability models

Two of the most used validation tests are the Cumulative Accuracy Profile (*CAP*) curve and *AR* developed by Sobehart, Keenan, and Stein (2000a). Their calculation is performed by ranking all parties based on the scores estimated by the model. Once ranked, for a certain cutoff score, it is possible to identify the fraction of defaults and non-defaults with scores that are less than the cutoff score. The *CAP* curve is obtained by calculating these fractions for all possible cutoff points, as shown in Fig. 1.

According to Engelmann, Hayden, and Tasche (2003), the *AR* can be defined by:

$$AR = \frac{a_R}{a_P}, \tag{1}$$

where a_R,a_P are the areas defined in Fig. 1. The closer the *AR* is to one, the greater the discriminative ability of the model.

The Receiver Operating Characteristic (*ROC*) curve and the area under the *ROC* curve are other widely used validation measures developed by Tasche (2006). The *ROC* curve is obtained by plotting $HR(C)$ versus $FAR(C)$, where $HR(C)$ is the hit rate and $FAR(C)$ the false alarm rate at score *C*. According to Engelmann et al. (2003), the higher the area under the *ROC* curve of the model, the better the performance. Considering the ideal situation, i.e., an *ROC* area equal to 1, the area may be calculated using Eq. (2):

$$AUROC = \int_0^1 HR(FAR)d(FAR). \tag{2}$$

The Pietra Index developed by Pietra (1915)[7] is a widely used index, whose geometric interpretation corresponds to half of the shortest distance between the *ROC* curve and the diagonal. This index can be calculated as:

$$PI = \frac{\sqrt{2}}{4}\max_c|HR(C) - FAR(C)| \tag{3}$$

Sobehart et al. (2000a) defined the *CIER* measure according to:

$$CIER = \frac{H_0(P) - H_1}{H_0(P)}, \tag{4}$$

where $H_0(P)$,H_1 are entropy functions developed by Jaynes (1957) and related to Kullback-Leibler (*KL*) distance, with the purpose of finding a function with conditions of continuity, monotonicity, and composition law, that represents the uncertainty of a probability distribution. Keenan and Sobehart (1999) defined the measure $H_0(p)$ as the entropy of a binary event for which *p* is the default rate of the sample.

[7] See Eliazar and Sokolov (2010) for a recent economic application.