

Validation of Rating Models

Bernd Appasamy, Stefan Hengstmann, Georg Stapper, Egbert Schark
d-fine, Frankfurt, Germany, Email for correspondence
bernd.appasamy@d-fine.de

Introduction

In the course of the upcoming new capital accord, Basel II, the different methodologies for the construction of internal rating models were in the center of interest within the last few years. Special attention was often put to the optimization of the discriminatory power and on statistical stability, frequently only obtainable by joining several portfolios. However, in practice it turned out that the discriminatory power itself is not a suitable measure for the quality of a rating system. Recently it became evident, that attention must be paid to the comprehensive validation of the systems just besides mature rating methodologies (Bundesbank 2003).

Rating methodologies serve for the assessment of the creditworthiness and finally the expected default probability (PD) of obligors. Available credit risk factors are assessed by means of different methods and combined to a single indicator (score) specifying the creditworthiness. However, this assessment is flawed with prediction uncertainties due to different error sources inherent in the system. The most important error sources are:

- statistical uncertainties
(e.g. uncertainties due to optimizations during calibration)
- missing, uncertain, respectively incorrect input data
(e.g. undisclosed reserves in a balance sheet)
- idealized assumptions of a rating methodology
(e.g. assumption of normally distributed financial ratios)

The design of a robust rating model with high discriminatory power has to reflect these error sources. However, it is a central task of the validation to prove that the probability of default predicted by the rating model corresponds to the observed default probability. This article presents a quantitative and flexible method which takes into account the error sources mentioned above and quantifies the prediction uncertainty of the probability of default and the discriminatory power of the rating model. The corporate rating for non-profit organizations of the German bank *Bank für Sozialwirtschaft* is chosen to illustrate this validation approach. This example is chosen with the awareness that the portfolio is small and shows a low average probability of default ($\overline{PD} < 1\%$), two conditions which are typically considered to be difficult for the development of a robust and reliable rating model.

Prediction Uncertainty of a Rating Model

Model supported rating systems are generally combinations of statistical methods and expert systems. While statistical methods are calculating one risk sensitive score for the obligor by means of a score function whose parameters are calculated with the help of an optimization method, the estimation of the creditworthiness of an expert systems is carried out by means of evaluations of soft-facts for which the accumulated experience of credit experts was formalized. Both, the model supported and the expert rating system must be examined with regard to discriminatory power, prediction uncertainty and robustness. This

examination is an essential part of the rating validation and usually affects the development of a robust and reliable rating model. A corporate rating is usually a hybrid model and takes into account both a statistical model for the analysis of financial statements and an expert system for soft-facts.

In the further course of this article we start out exemplarily from a hybrid rating model for corporate that incorporates both categorical risk factors (for example the corresponding soft-fact for the evaluation of the development of the company's line of business) and metric risk factors (for example the financial ratio: *equity-to-fixed-assets*). The result of the rating model should be directly the corresponding probability of default of the rating class the obligor is assigned to. We assume, for simplicity, that the default probability is averaged over the score interval of the corresponding rating class which reflects the creditworthiness of the obligor.

The prediction uncertainty of a rating model can be verified directly by means of counting defaults: For this, forecasts are made for the probability of default at time t_0 for all obligors using the rating model. The observed defaults of obligors of a rating class within the time $[t_0|t_0 + 1 \text{ year}]$ are used for calculating the observed probability of defaults of the rating class. These observed probability of defaults are compared to the predicted values. Observation and prediction are compatible if they are in agreement taking into account their statistical uncertainty.

The error of the number of defaults within a specific rating class is determined by the binomial distribution, since the average default probability of the portfolio, the number of defaults and the total number of obligors are known. It should be stressed that it can even be calculated to what extent an observation of no default within a rating class is compatible with the model prediction. The forecast uncertainty of the model is estimated using the method of bootstrapping (Efron and Tibshirani 1993), (Venables and Ripley 1999) which is outlined in the following.

Random samples of indicators such as financial ratios can generally not be described assuming simple distribution models like a normal distribution. Estimating the empiric distribution on the basis of observed samples it is usually necessary to accumulate a sufficient amount of data over a long period of time. Especially in the credit business this is not a practical approach for the distribution of defaulted obligors. An alternative approach is to generate several random samples by drawing from a single observed distribution. All simulated samples produced this way show characteristics as the observed distribution. This technique is known as bootstrapping (Efron and Tibshirani 1993), (Venables and Ripley 1999). To demonstrate the bootstrapping for estimating the uncertainty range of the probability of default curve the following steps have to be performed:

- Assume a portfolio of N obligors for which the score was calculated at time t_0 by the rating model using all available information at time t_0 . The number of observed defaults within a time period of one year is assumed to be N_D (N_D non-performing obligors) while the number of performing obligors is $N_A = N - N_D$. The non-performing obligors

are labeled with "D" for *defaulted* obligors while the survivors are labeled with "A" for *active* obligors. These two samples (so-called reference samples) represent the score distributions for performing and non-performing obligors.

- Since the number of non-performing obligors may be different in the next year, as a first step, the number of obligors n_D which may default during one year are simulated by means of a binomial distribution using the average probability of default and the number of obligors within the portfolio as input parameters.
- Next, a sample of n_D non-performing obligors is drawn from the non-performing reference sample and $N - n_D$ obligors are drawn from the performing reference sample. Attention must be paid that each obligor is drawn from the complete reference sample because the whole reference sample serves as an approximation for the distribution (sampling with replacement). This sampling process may be repeated several times (typically 250 times) obtaining several possible random samples, each representing a simulated portfolio of obligors with n_D defaults.
- The default probability curve $PD_i(\text{score})$ of each simulated portfolio may then be calculated in the usual way using Bayes's rule (Duffie and Singleton 2003):

$$PD_i(\text{score}) = \frac{\varphi_D(\text{score}) \cdot n_D}{\varphi_A(\text{score}) \cdot n_A + \varphi_D(\text{score}) \cdot n_D}$$

$\varphi_D(\text{score})$ denotes the conditional distribution at time t_0 of the score for obligors which are defaulted within the period $[t_0|t_0 + 1 \text{ year}]$ while $\varphi_A(\text{score})$ denotes the conditional distribution at time t_0 of the score for performing obligors.

The distribution of the $PD_i(\text{score})$ curves may then be used to estimate the expected probability of default curve and its uncertainty by calculating the corresponding confidence intervals.

The bootstrapping method can directly analyze the impact of typical problems like incomplete or biased information (two examples are: opening balance sheets for start ups and hidden reserves), a small number of observed defaults (Peduzzi et al. 1996) and the problem of over-parameterization (Efron and Tibshirani 1993), (Venables and Ripley 1999). This is an important advantage of the method allowing an efficient development of robust rating models with high discriminatory power. Furthermore the bootstrapping provides a powerful tool for the completion of missing data, known as imputation (Venables and Ripley 1999). The bootstrapping technique is also helpful when analyzing the effect of incorrect model assumptions on the predictive power of the rating system. An example could be the assumption of normally distributed indicators when applying discriminant analysis. Bootstrapping may also be used to investigate minimum required number of defaults to build a stable rating system. Our practical experience has shown that about 30 defaults may be sufficient for a robust and sensitive rating tool.

Example: non-profit organizations

The *BfS* is a bank which is, among others, specialized on financing corporates, non-profit organizations, and friendly societies operating in the social economy. The corresponding portfolio of the *BfS* shows a small average probability of default and is relatively small but homogeneous.

In the context of the new capital accord, Basel II, an internal rating model has been developed to determine the creditworthiness of counterparties operating in the social economy. The probability of default is estimated by two different systems, a modified expert system and a statistical rating system. These systems are monitoring the performance of each other. The expert system, in use since 1996, has been extended to calculate the default probability by the help of the numerical derivation of the receiver-operating-characteristic curve (ROC-curve) (Fritz, Popken and Wagner 2002). Both rating systems use beside current and application information also predictions about future developments in the social economy. The internal rating model is designed as a hybrid rating model. The model consists of a financial module and a scoring module for the soft-facts. The financial module analyzes the liquidity (and debt service), the capital structure, the profitability and the cash-flow using a logistic regression model (Venables and Ripley 1999), (King and Zeng 2000) with four financial ratios. The soft-facts module contains four soft-facts analyzing management quality, the line of business, the corporate's market position and the relationship of the corporate to the bank. In addition a group logic and an override feature is implemented. Overrides modify the final rating according to an override catalogue.

The internal rating model has been designed in 2001. It accounts for both quantitative and qualitative input, that is financials and soft-facts. The following discussion focuses on the validation of the predicted default probability and the discriminatory power on the basis of the financial ratios. Using the soft-facts the discriminatory power can be increased without affecting the quality of the rating system. For the validation we analyze the PD-curve and the ROC-curve of the internal rating model in detail:

Figure 1 shows PD (blue line) and the confidence interval (green lines) as a function of the score of the rating model. The results are presented using a half-logarithmic scale for the y-axis. The 95%-confidence interval reflecting the uncertainty of the PD-curve depending on the score has been determined by means of a multivariate bootstrapping. In contrast to the simple bootstrapping algorithm outlined before, for the multivariate bootstrapping (Efron and Tibshirani 1993) algorithm the individual risk factor values are drawn taking their correlation into account. Since the risk factors are not normally distributed the reference sample was initially transformed using a QQ-plot (Venables and Ripley 1999). The observations together with their uncertainties are depicted as squares with corresponding error bars. It can be seen that observations and expectations of default probabilities are in good agreement for all values of the score. An indication for a robust rating system is the relatively small confidence interval over the full range of the score. Since the portfolio of the *BfS* contains only very few counterparties with score values

smaller than -3.5 and bigger than 1.0 the uncertainty increases in these regions.

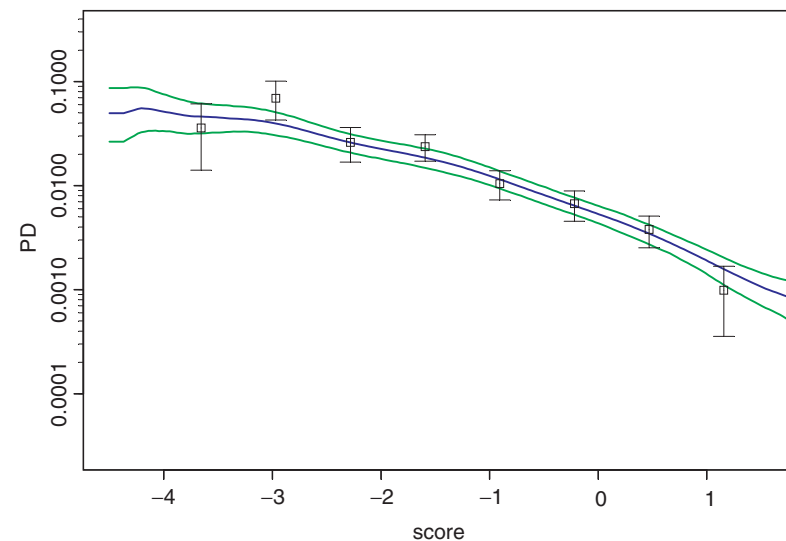


Figure 1: PD of the financial module as a function of the score (blue line) with the corresponding 95% confidence levels (green lines) and the observed probability of default (squares) with error bars

The ROC-curve of the financial rating module is presented in Figure 2a. A CoC-value of 86.3% is achieved corresponding to a Gini-coefficient of 72.6%. In addition to the observed ROC-curve (black line) the expected ROC-curve (blue line) and the 95%-confidence interval (green lines) is plotted. The small confidence interval guarantees the robustness of the rating model. Obligors of poor creditworthiness are well identified, reflected by the steep rise of the ROC-curve. The horizontal part of the ROC-curve shows that the counterparties of high creditworthiness are also identified by the financial rating module. Figure 2b shows the corresponding distribution of CoC-values. The large CoC-value shows a statistical uncertainty of plus 4% / minus 4.5% manifesting the quality of the implemented rating method.

Figure 3 shows the PD-curve of the complete rating model. The CoC-value of the complete rating model is 95.4%. The complete model incorporates both financial ratios and soft-facts. In addition a group logic and overrides according to an override catalogue were implemented. The validation proves the rating system to be robust and credit risk sensitive. A system of poor robustness would show large prediction uncertainties. Observation uncertainties being smaller than the expectation uncertainties of the PD-curve would be reflected by the confidence interval of the ROC-curve and a large spread of CoC-values.

Those positive features of the investigated rating model are independent of the absolute size of the underlying portfolio. Increasing the

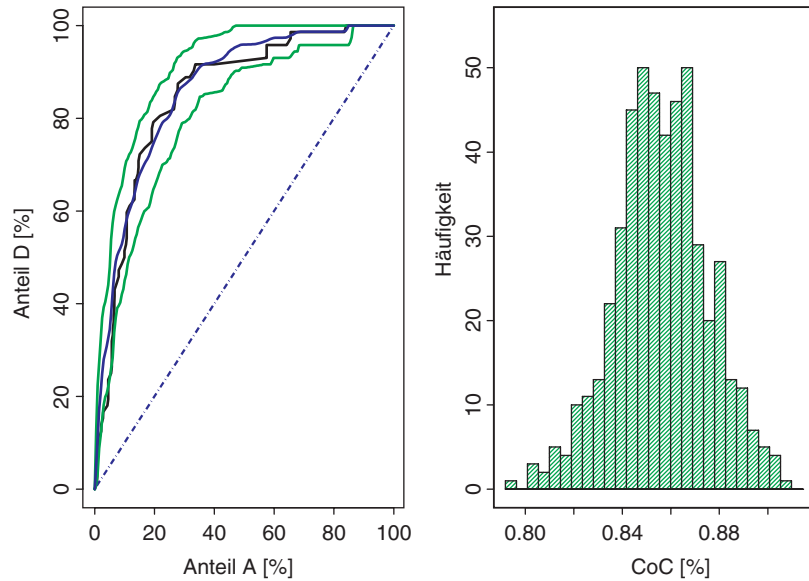


Figure 2: a) Current ROC-Curve for the financial module of the corporate rating model of the *Bank für Sozialwirtschaft* (black line), corresponding expected ROC-curve (blue line), 95% confidence interval (green lines) b) frequency distribution of CoC-values corresponding to the uncertainty range of the ROC-curve

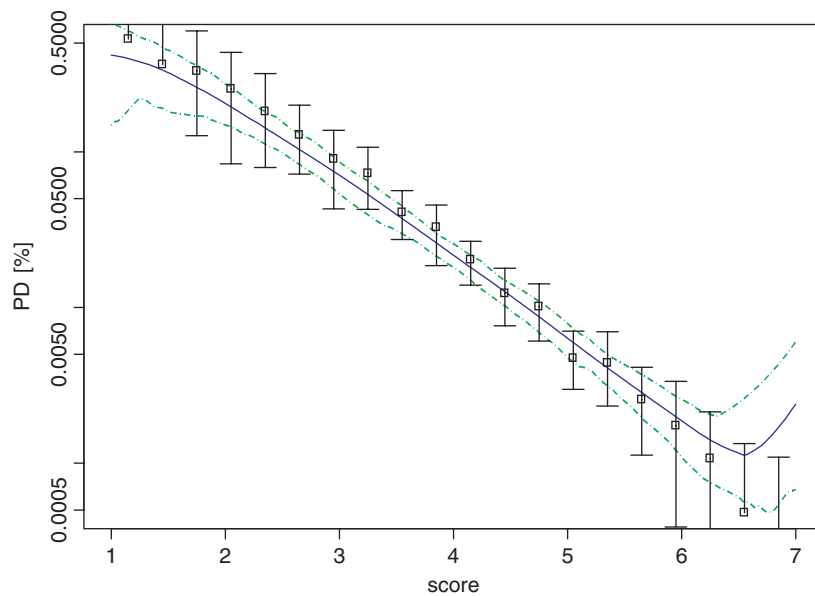


Figure 3: PD of the complete rating model as a function of the score (blue line) with the corresponding 95% confidence levels (green lines) and the observed probability of default (squares) with error bars

amount of data by, for example, merging several portfolios would just decrease the observation uncertainty without changing the prediction uncertainty.

The explained bootstrapping method accounts for the typical error sources in the PD estimation process and allows statements about the quality of the rating model. The approach is demonstrated on a productively used rating model for corporates of the *Bank für Sozialwirtschaft*. This example was chosen consciously because the corresponding portfolio shows difficult prerequisites for the construction of a reliable rating model, since the specific portfolio has got a low average default probability on a low population.

REFERENCES

- *Validierungsansätze für interne Ratingsysteme*; Monatsbericht der Deutschen Bundesbank; September 2003.
- B. Efron and R. Tibshirani; *An Introduction to the Bootstrap*; Chapman and Hall; New York 1993.
- W.N. Venables and B.D. Ripley; *Modern Applied Statistics with S-PLUS*; Springer Verlag, New York, third edition 1999.
- P. Peduzzi, J. Concato, E. Kemper, T.R. Holford, and A.R. Feinstein; *A simulation study of the number of events per variable in logistic regression analysis*; *Journal of Clinical Epidemiology*; **49** 1373 pp (1996).
- D. Duffie, K.J. Singleton; *Credit Risk* Princeton Series in Finance; Princeton and Oxford 2003.
- S.G. Fritz and L. Popken, and C. Wagner; *Scoring and Validating Techniques for Credit Risk Rating Systems*; Risk Books/Credit Ratings; 2002.
- G. King and L. Zeng; *Logistic Regression in Rare Events Data*; Political Analysis; **9.2** 1pp (2000).