Approaches to the validation of internal rating systems

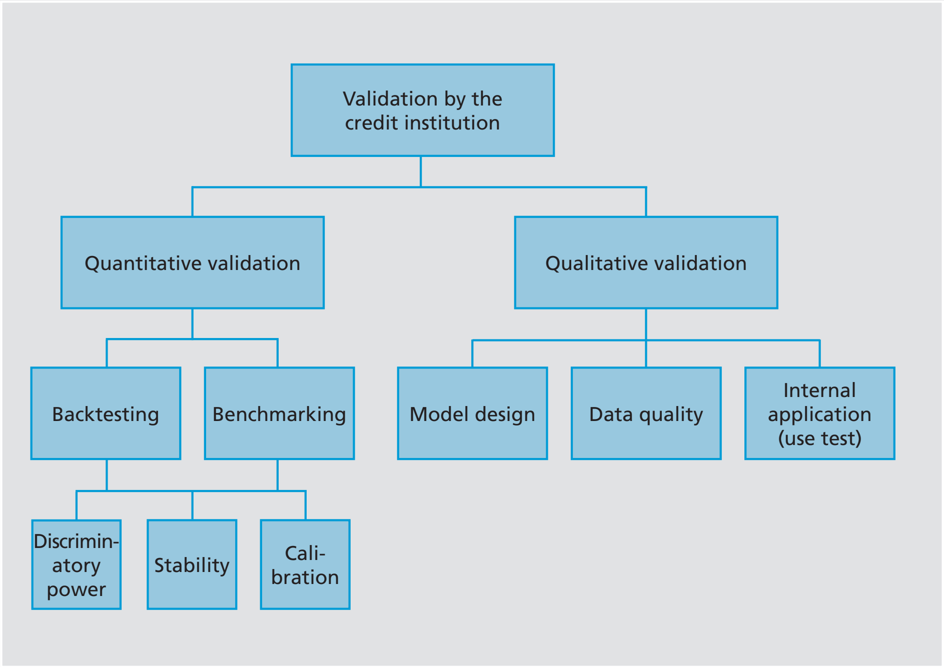
The new international capital standard for credit institutions (Basel II) permits banks to use internal rating systems for determining the risk weights relevant for calculating the capital charge. In return the banks are obliged to regularly review their rating systems (validation). Regulatory standards for validation are designed to ensure a uniform framework for the prudential certification and ongoing monitoring of the internal rating systems used.

Validation represents a major challenge for both banks and supervisors. It is true that the statistical methods used for quantitative validation are useful indicators of possible undesirable developments. As a rule, however, it is not possible to deduce from them a stringent criterion for assessing the suitability of a rating system. For this reason qualitative criteria will play an important role in validation.

It is likely that the methods described in this article will be further developed and refined in the coming years, not least owing to the increasing availability of reliable data. In particular, the future discussions generated both by research and banking practice will provide additional insights into the methods used for estimating the risk parameters.

Rating systems serve to determine the credit risk of individual borrowers. Using various

Aspects of validation



Deutsche Bundesbank

methods, rating scores are assigned to individual borrowers to indicate their degree of creditworthiness.

In view of the envisaged prudential recognition of banks' internal rating systems under the two IRB (Internal Ratings-Based) approaches, the problems associated with their quantitative and qualitative validation are currently the subject of much discussion. The term validation denotes the entire process of assessing an internal rating system, from validating its discriminatory power to process-oriented validation ("use test"). The chart on this page gives an overview of the main components of the validation process for rating systems.

The task of validating rating systems is closely connected with the validation of additional risk parameters that are derived from the rating assessments and which, under the IRB approaches of the new Basel minimum requirements (Basel II), largely determine the amount of capital which a bank needs to maintain. This article examines the problems associated with validation without expressing any prudential choice for or against particular methods. It reflects some of the best practices as ascertained from a survey of German banks carried out in the spring of 2003.

Quantitative aspects of validation

The precise nature of both quantitative and qualitative validation greatly depends on the character of the rating system in use. A basic distinction is drawn between model-based systems and systems based on expert judgement.

Model-based rating systems

Model-based systems, such as discriminant analysis or various kinds of regression analysis, are typically developed on the basis of historical default data. If such data are not available on a sufficient scale, many practitioners resort to a "shadow rating" which adopts the credit assessment of external rating agencies. A feature shared by all model-based systems is that – using statistical methods – they capture a number of risk factors (eg total exposure, equity capital or sector/profession) in a risk ratio (rating score).

Expert judgement

If little statistically significant information is available or if the credit operations are of material importance or complex, the bank will normally rely instead on expert judgement. In such a rating system, too, a standardised procedure is normally applied for assigning the ratings. The main difference between this and model-based methods is that there is no statistical modelling of the rating score.

Hybrid systems

In practice, the most common methods used are hybrid forms combining elements of both types of rating system. In such hybrid systems the responsible credit expert can correct the model-based rating if he has information of which the model-based rating system takes no or insufficient account.

Criteria for quantitative validation

All rating systems – whether model-based or based on expert judgement – can essentially be validated by quantitative means. However, a quantitative validation requires a sufficient number of loan defaults. This requirement is typically met in the case of retail business, ie loans to small and medium-sized enterprises or to individuals. The principal criteria for the quantitative validation of a rating system are its discriminatory power, its stability and its calibration.

Discriminatory power and stability

Discriminatory power of a rating system

The discriminatory power of a rating system denotes its ability to discriminate *ex ante* between defaulting and non-defaulting borrowers. The discriminatory power can be assessed using a number of statistical measures of discrimination, some of which are described in detail in the Annex to this article. However, the absolute measure of the discriminatory power of a rating system is only of limited meaningfulness. A direct comparison of different rating systems, for example, can only be performed if statistical "noise" is taken into account. Such a comparison must be based on the same dataset.

Moreover, the discriminatory power should be tested not only in the development dataset but also in an independent dataset (out-of-sample validation). Otherwise there is a danger that the discriminatory power may be overstated by over-fitting to the development dataset. In this case the rating system will then frequently exhibit a relatively low discriminatory power on datasets that are independent of but structurally similar to the development dataset. Hence the rating system would have a low stability.

Criteria for assessing the quality of rating systems

Discriminatory power:

The discriminatory power of rating systems denotes their *ex ante* capability to identify borrowers who are in danger of defaulting. Thus a rating system with maximum discriminatory power would be able to precisely identify in advance all borrowers who subsequently default. In practice, however, such perfect rating systems do not exist. A rating system is said to have a high discriminatory power if the "good" grades subsequently turn out to contain only a small percentage of defaulters and a large percentage of non-defaulters, with the converse applying to the "poor" grades.

Stability:

A characteristic feature of a stable rating system is that it adequately models the cause-effect relationship between the risk factors and creditworthiness. It avoids spurious dependencies based on empirical correlations. In contrast to stable systems, unstable systems frequently show a sharply declining level of forecasting quality over time.

Accuracy of calibration:

Calibration normally denotes the mapping of the Probabilities of Default (PD) to the rating grades. A rating system is well calibrated if the estimated PDs deviate only marginally from the actual default rates. Used in a broader sense, the calibration of the rating system also includes the mapping of additional risk parameters, such as the Loss Given Default (LGD) and the Exposure At Default (EAD).

Deutsche Bundesbank

Stability of a rating system

One way of assessing the stability of a model-based rating system is to measure the statistical significance of the risk factors applied. In addition, a test should be made to ascertain any correlation effects. High or unstable correlations may adversely affect the stability of the rating system.

Calibration

The risk parameters under Basel II

Under both IRB approaches of Basel II, a bank's capital requirements are determined by internal estimations of the risk parameters for each exposure. These are derived in turn from the bank's internal rating scores. These notably include the borrower's Probability of Default (PD) and, for the advanced IRB approach, the expected Loss Given Default (LGD) as well as the Exposure At Default (EAD). In this connection one also speaks of the calibration of the rating system. As the risk parameters can be determined by the bank itself, the quality of the calibration is a decisive prudential criterion for assessing rating systems.

Not only the PD but also the LGD and the EAD are random variables as they are not fully known to the bank when rating borrowers' creditworthiness. They depend, in particular, on the intrinsic value of the collateral and on the amount of credit drawn down by the time of the default. Unlike the PD, however, these parameters have to be estimated by the bank itself only under the advanced IRB approach, whereas under the IRB foundation approach they are laid down by the supervisors.

Methods of calculating PD

There are several tried and tested statistical methods for deriving the PDs (Probabilities of Default) from a rating system. Firstly, a distinction needs to be drawn between direct and indirect methods. In the case of the direct methods, such as Logit, Probit and Hazard Rate models, the rating score itself can be taken as the borrower's PD. The PD of a given rating grade is then normally calculated as the mean of the PDs of the individual borrowers assigned to each grade.

Where the rating score cannot be taken as the PD (as in the case of discriminant analysis), one may resort to indirect methods. One simple method consists of estimating the PD for each rating grade from historical default rates. Another method is the estimation of the score distributions of defaulting borrowers, on the one hand, and non-defaulting borrowers, on the other. A specific PD can subsequently be assigned to each borrower using Bayes' Formula.

In practice a bank's PD estimates will differ from the default rates actually observed subsequently. The key question is whether the deviations are purely random or whether they occur systematically. A systematic underestimation of PDs merits a critical assessment – from the point of view of supervisors and bankers alike – since in this case the bank's computed capital requirement would not be adequate to the risk it has incurred.

Various statistical methods of assessing the estimation quality of PDs are discussed in the academic literature. Most of these methods are based on backtesting. However, these methods display shortcomings in practice which argue against their mechanical application. These can be illustrated by means of the binomial test, the technical details of which are described in the Annex.

Binomial test

The binomial test was first incorporated into prudential practice in connection with the backtesting of market risk models. For the assessment of PDs, too, it is possible to construct a statistical test (using simplified assumptions) based on the binomial distribution. It is assumed that the defaults per rating grade are statistically independent. Under the hypothesis that the estimated PDs of the rating grades are correct, the actually observable number of defaults per rating grade after one year would then be binomially distributed. If major differences are evident between the default rate and the estimated PD of the rating grade, the hypothesis of a correct estimation must be rejected. The rating model would thus be poorly calibrated.

One problem associated with this test is the assumption that the defaults of the borrowers constitute independent events. In reality, however, the defaults are more or less strongly correlated owing to cyclical influences. Theoretically, a solution to this problem would be conceivable if the default correlations were known. But determining the default correlations is difficult. Hence even a modified binomial test is suitable at most as an indicator of a good or poor calibration.

Another approach to the statistical validation of PDs is the use of benchmark portfolios. In banking practice, for example, using external

Use of benchmark portfolios and external data sources

data from rating agencies and other commercial providers as a benchmark is widespread. Systematic deviations of the bank's internal estimates from the estimates in the benchmark portfolio would have to be checked. Benchmarking can serve as a useful complement to the validation process. However, the usefulness of this approach depends very much on the choice of a suitable benchmark portfolio. The choice of a benchmark rating is likewise generally not an easy task.

Measuring the LGDs

Besides an estimation of the PD, the IRB advanced approach under Basel II will also permit banks to themselves estimate the LGD (Loss Given Default) and the EAD (Exposure At Default). A quantitative validation of the LGDs consists in verifying the bank's internal estimates. The LGD of bank loans is determined mainly by the realisation of the loan collateral. If a loan is not repaid, the credit institution does not know how high the actual loss is until the liquidation period is terminated. The liquidation period may vary greatly, depending on the precise features of the loan and, in particular, on the collateral. As a rule it is between 18 months and three years, but in exceptional cases it may even exceed ten years.

In order to calculate the actual loss it is necessary to take account of all payment streams that flow during the liquidation process and, where appropriate, to assign them to individual collateral items. The payment streams comprise payments made to the bank and payments which the bank itself has to make. The former consist primarily of partial payments made by the borrower or of proceeds from realising collateral. The latter consist, for example, of lawyer's costs, court costs plus cumulative interest charges and refinancing costs during the liquidation process. Given the duration of the liquidation process, the payment streams have to be discounted before the actual economic LGD can be calculated.

A number of statistical studies already exist which can be used to determine the LGDs of exchange-traded corporate bonds. By contrast, standardised databases concerning losses from unsecured loans are still at a rudimentary stage of development. But in the case of unsecured loans, too, it is likely that the LGDs are very much sector-specific and are strongly correlated with the default rates. The LGD database must capture the losses completely and must also contain those defaulted loans where the unsecured shares have not led to losses. The exclusive inclusion of loans that have actually led to losses would lead to an overstating of the LGD. It is also common for several loans to be secured by one and the same collateral item (eg a global land charge). As a rule a credit institution will try to estimate a separate realisation rate for each category of collateral. In the case of global collateral the collateral must be distributed across the individual loans.

EAD

Like the LGD, the validation of the EAD (Exposure At Default) is based on the verification of the bank's internal estimates. For balance sheet assets the Basel minimum requirements envisage that the estimated values must not be less than the currently drawn credit amount (though netting effects may be taken into account). For derivative transactions the credit equivalent amount is calculated from the replacement cost plus an add-on for future potential liabilities. The supplementary prudential requirements in respect of the bank's internal estimates of the EAD are thus concentrated on off-balance-sheet transactions. A central problem is determining the drawn share of credit line amounts at the time of default. Studies indicate that there are significant correlations between the EAD and the residual maturity of the loan, and between the EAD and the borrower's credit rating. Additional utilisation of the credit line tends to increase the EAD in accordance with the length of the residual maturity of the loan. This is plausible, since the longer the residual maturity of a loan, the greater is the probability that the borrower's credit rating will deteriorate and his potential access to alternative financing sources will diminish. Other study findings indicate that the degree of utilisation of the credit line by the time of default tends to decrease in accordance with the quality of the borrower's credit rating at the time the credit line was granted. The argument put forward to explain this is that, faced with a borrower with a poor credit rating, a bank will insert clauses into the credit agreement that hamper his utilisation of the approved credit line in the event of a further deterioration in his rating.

The estimates can be greatly simplified if dependencies on the creditworthiness and the residual maturity do not have to be taken into account. However, this harbours the risk that neglecting these dependencies may systematically distort the estimates for the credit utilisation.

Qualitative aspects of validation

Qualitative validation as a complement to quantitative validation

The quantitative validation methods have to be complemented by qualitative – ie non-statistical – methods. Qualitative validation serves not least to safeguard the applicability of quantitative methods. In these cases the qualitative validation will have to be performed before the quantitative validation. The qualitative analyses primarily test three aspects: the design of the rating models, the quality of the data for the rating development and deployment as well as the internal use of the rating system in the credit-granting process ("use test").

Model design

Testing the model design plays a major role in the case of model-based systems, in particular, but not just for these. This is especially true whenever a quantitative validation is subject to limitations owing to the dataset. In any case the process of assigning the rating must be transparent and well documented. The influence of the risk factors should be discretely disaggregated and economically plausible. In addition, the demonstration of statistical foundation is crucial in the case of model-based systems.

Data quality and availability

A bank should, as a general rule, pay close attention to the integrity of its data and their consistent collection. Only a sound database with a sufficiently large data history makes possible the development of a high-quality rating system and reliable estimates of the prudentially stipulated risk parameters. If the credit institution itself has only a small database of default information, it may resort if necessary – as mentioned above – to external data sources.

Use test

A second major criterion for the qualitative validation of internal rating systems is the actual use of rating results in banks' internal risk management and reporting. This kind of qualitative validation tests the design of the internal bank processes and is therefore referred to as "process-oriented validation". Examples of credit risk management using rating systems include ratings-based credit decisions and credit-granting competencies, a credit risk strategy geared to rating grades and correspondingly structured limit systems. In all of these applications a credit institution bases important business policy decisions on the risk assessment generated by internal ratings.

From a prudential point of view the way in which the bank uses its rating system for internal decision-making processes reflects the confidence it has in its own system. Wherever banks' own rating systems are not used internally or are used only for individual, isolated purposes, this can be interpreted as an internal assessment of the (deficient) quality of the rating systems. A rating system that is not sufficiently integrated into the bank's internal credit processes will therefore not receive prudential approval.

The quantification of risk, expressed in PDs and realisation or LGD rates, should likewise be used for the bank's internal purposes. The most important example of this is the calculation of the standard risk costs as part of contribution margin costing. The calculation of risk provisions based on standard risk costs is another conceivable indicator of the internal use of rating systems.

Independence

In addition, the Basel minimum requirements for internal rating systems stipulate that rating decisions must not be influenced by other business divisions that profit from the credit decision either directly or indirectly. A particularly important requirement is the independent assignment of ratings when using expert judgements. In these cases the final rating competence must lie with the back office staff and not the front office staff. This applies even more so if the sales staff are remunerated according to the volume of transactions concluded. One of the qualitative criteria is therefore that the initial rating proposal, which may potentially be made by the customer account staff, must be reviewed and confirmed by an independent third party.

Other factors

Other key points of the validation process are appropriate training for the staff and the acceptability of the rating systems to their users. These must have a good understanding of the rating system and actually apply the rating system in everyday business.

Prospect of using a central credit register for validation purposes

If an institution seeks approval for internal rating systems under Basel II, it must demonstrate that its system has been adequately validated. The task of the supervisors is to certify the rating systems and to continuously monitor compliance with these minimum requirements by the bank. In the context of this process the bank's internal validation procedures also have to be assessed. In this connection central credit registers can play an important role. The Basel Committee on Banking Supervision is therefore currently considering their possible application for this purpose.

Required database

The main precondition for using a central credit register for prudential purposes is the availability of information about loan defaults, banks' internal rating grades and the collateralisation of the loans. This information is already available in part on the central credit registers of certain countries. A central credit register has the advantage over the alternative of individual enquiries, by virtue of standardised borrower IDs, of being able to compare the rating of different banks for one and the same borrower (benchmarking). The sample selected for the comparison could be defined flexibly. Moreover, the reporting system would cover the entire range of banks.

Use for backtesting

Another field of application of credit registers in connection with validation questions is backtesting. As explained above, backtesting involves comparing the estimated PDs with the actually observed defaults. In principle, this would make it possible to test banks' internal quantitative validation.

Thus central credit registers could, in principle, play a supporting role in the prudential certification of rating systems and their oversight. Depending on the scale of the investigations which this would necessitate, this would require making modifications to the central credit registers in their current form. This should be decided on the basis of careful cost/benefit considerations. The primary application of central credit registers will probably be the benchmarking of estimates of different credit institutions. By contrast, the use of credit registers for backtesting comes up against limits owing to the high degree of detail required for the relevant credit information.

Outlook

Banks and banking supervisors are currently preparing intensively for the validation of rating systems. With a view to further developing the approaches to validation, a working group for validation issues has been set up under the stewardship of the Research Task Force of the Basel Committee on Banking Supervision. The steep increase in the number of publications on this subject in recent years shows that academics are also considering this question. However, the suitability of individual methods is still disputed. One thing that is certain is that the assessment of internal rating systems cannot be based on a single validation method but instead will emerge as the synthesis of various quantitative and qualitative methods. The current discussion will lead to the further refinement of validation methods. In addition, the quality and quantity of the available data will improve substantially in the coming years. The resulting insights will be incorporated into the prudential validation standards.

Annex

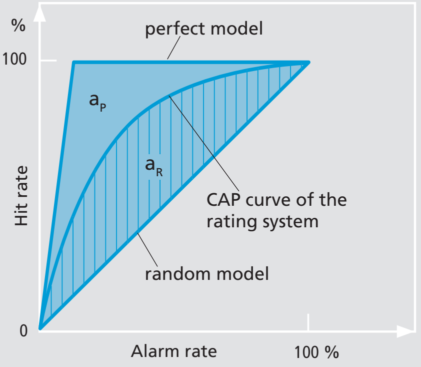
Statistical measures of discriminatory power

Cumulative Accuracy Profile (CAP)

The CAP curve provides a graphical illustration of the discriminatory power of a rating process. For this purpose, the creditworthiness indicator (score) of every borrower is established for the dataset to be used to examine the rating model's discriminatory power. This score can be continuous, for instance the result of a discriminant analysis or a Logit regression, or it may be an integer which represents the rating grade to which the borrower has been assigned. In the following analysis, it is assumed that a high score is a reflection of a good rating. In a first step the borrowers are arranged in an ascending order of scores. The CAP curve is then determined by plotting the cumulative percentage of all borrowers ("alarm rate") on the horizontal axis and the cumulative percentage of all defaulters ("hit rate") on the vertical axis. This is shown in the adjacent chart. If, for example, those 30% of all debtors with the lowest rating scores include 70% of all defaulters, the point (0.3;0.7) lies on the CAP curve. The steeper the CAP curve at the beginning, the more accurate the rating process. Ideally, the rating process would give all defaulters the lowest scores. The CAP curve would then rise linearly at the beginning before becoming horizontal. The other extreme would be a purely random rating classification. Such a rating process would not have any discriminatory power. The expected CAP curve would, in this case, be identical to the diagonal. In reality, rating classifications are neither perfect nor random. The corresponding CAP curve therefore runs between these two extremes. Using the CAP curve, the discriminatory power of a rating process can be aggregated into a single figure, the so-called "Gini coefficient"[1] (GC). In the above chart, the area between the perfect rating and the random rating is denoted by a_p and the area between the actual rating and the random rating is denoted by a_R. The Gini coefficient is defined as the ratio of a_R to a_p, which means

Gini coefficient (GC)

Cumulative Accuracy Profile (CAP)* and Gini coefficient**



* For each rating score the alarm rate measures the fraction of borrowers with a lower-than-specified score within all borrowers. The hit rate gives, for each rating score, the fraction of defaulters with a lower-than-specified score within all defaulters. Processing all possible scores yields the points of the CAP curve. — ** The Gini coefficient is the quotient of the area a_R, between the CAP curve and the diagonal, and the area a_P of the shaded triangle. The larger the quotient, the more accurate the rating process.

Deutsche Bundesbank

$$GC = \frac{a_R}{a_P}.$$

The Gini coefficient is always between minus one and one. A rating system is the more accurate the closer it is to one.

Receiver Operating Characteristic (ROC)

The ROC curve is a concept related to the CAP curve. In order to plot this curve, the empirical score distribution for defaulters, on the one hand, and for non-defaulters, on the other, is deter-

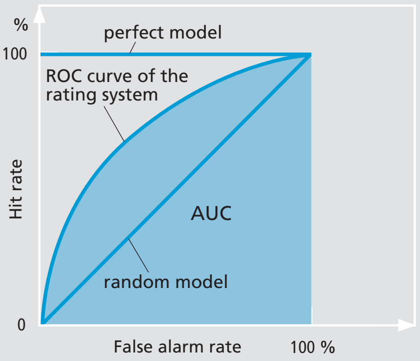
[1] The Gini coefficient is often termed the "accuracy ratio".

mined. The result could be similar to that shown in the chart on page 70. Next, a score C is set. Using this score C, it is possible to define a simple decision-making rule for identifying potential defaulters. All borrowers with a score greater than C are deemed to be creditworthy and those with a lower score are deemed to be not creditworthy. One of the features of a good rating system is that it has as high a hit rate as possible (correct classification of a borrower as a potential defaulter) and at the same time as low a false alarm rate as possible (incorrect classification of a creditworthy borrower as a potential defaulter). In order to analyse the discriminatory power of a rating system irrespective of the chosen cut-off value C, both the false alarm rate and the hit rate are calculated for every C between the maximum and the minimum score. The points determined in this way yield the ROC curve (see adjacent chart). The steeper the ROC curve at the beginning, the more accurate the rating system. In a perfect rating system, the ROC curve would be plotted solely on the line defined by the points (0;0), (0;1) and (1;1). In a purely random rating system, the ROC curve would be plotted exactly along the diagonal in the adjacent chart.

Area under the Curve (AUC)

As for the CAP curve, an aggregated ratio can also be given for the ROC curve. This ratio results from the area under the ROC curve and is called the AUC. The AUC ratio is always between zero and one. The closer the AUC is to one, the more accurate the rating system. The connection between the AUC and the GC as well as the statistical properties of the AUC and the GC are dealt with in the next section. The equivalence of the AUC and the GC is a key result. It is possible to convert one ratio into the other through a simple linear transformation.

Receiver Operating Characteristic (ROC)* and Area under the Curve (AUC)**



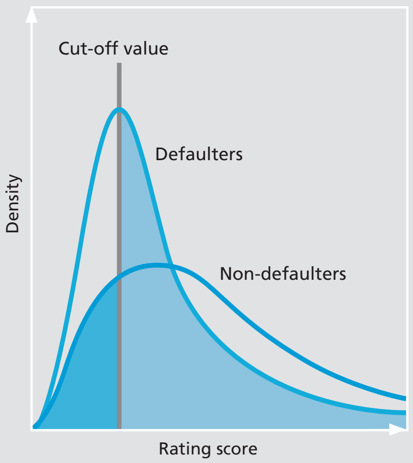
* For each rating score the false alarm rate measures the fraction of non-defaulters with a lower-than-specified score within all non-defaulters. The hit rate gives, for each rating score, the fraction of defaulters with a lower-than-specified score within all defaulters. Processing all possible scores yields the points of the ROC curve. — ** The AUC (shaded area under the ROC curve) is the average hit rate formed by computing the arithmetic mean of the hit rates belonging to all possible false alarm rates.

Deutsche Bundesbank

Minimum classification error rate

Another measure of discriminatory power widely applied in practice is the minimum classification error rate, the calculation of which is illustrated in the chart on page 70. The classification error rate is the term used to describe the mean of the relative frequencies for defaulters and non-defaulters who were incorrectly classified with a cut-off value of C. The fraction of defaulters who were deemed to be creditworthy in view of the cut-off value C corresponds to the area to the right of C under the defaulters’ score distribution curve. Similarly, the fraction of non-defaulters who were incorrectly classified as not creditworthy corresponds to the area to the left of C under the non-defaulters’ score distribution curve. The classification error rate is obtained by halving the total content of these two areas. The minimum classification error rate is obtained by calculating the classification error rate for every C value between the minimum

Probability densities of the rating scores* and classification error rate**



* For the distributions on the populations of the defaulters and non-defaulters, respectively. — ** With the given cut-off value, the classification error rate is obtained by halving the total content of the two shaded areas.

Deutsche Bundesbank

and the maximum score and determining the minimum level. The more accurate the rating system, the lower the minimum classification error rate. Alternatively, the minimum classification error rate can be determined using the Kolmogoroff-Smirnoff statistic, which measures the maximum difference between the two score distribution functions.

Statistical properties of the GC and the AUC

There is a simple linear relationship between the Gini coefficient (GC) and the area under the ROC curve (AUC) as two measures of discriminatory power, ie.

$$GC = 2 \cdot AUC - 1.$$

In the following, the statistical properties of mainly the AUC will be described as these can be interpreted more illustratively. The equivalent properties can be obtained for the GC using the preceding equation.

If all pair combinations of one defaulter and one non-defaulter are formed, the Mann-Whitney statistic can be defined as

$$U(a,b,c) = \frac{1}{N_D \cdot N_{ND}} \sum_{(D,ND)} u_{D,ND},$$

where N_D is the number of defaulters and N_{ND} is the number of solvent debtors. The expression $u_{D,ND}$ is defined as

$$u_{D,ND} = \begin{cases} a, \text{ if } S_D < S_{ND} \\ b, \text{ if } S_D = S_{ND} \\ c, \text{ if } S_D > S_{ND} \end{cases}.$$

Here, S_D is the defaulter's rating score and S_{ND} is the solvent borrower's rating score. The relationship

$$AUC = U(1, 0.5, 0)$$

can be proven for the AUC as a measure of discriminatory power. If the definition of U is taken into account, one obtains

$$AUC = P(S_D < S_{ND}) + 0.5\,P(S_D = S_{ND}).$$

This equation can be explained in illustrative terms. If one debtor is randomly chosen from all of the defaulters and one debtor is randomly chosen from all of the solvent borrowers, one would assume that the borrower with the higher rating score is the solvent borrower. If both borrowers have the same rating score, then lots are drawn. The probability that the solvent borrower can be identified using this decision-making rule turns out to be $P(S_D<S_{ND}) + 0.5\ P(S_D=S_{ND})$. This probability is identical to the area under the ROC curve.

Confidence intervals and tests for the AUC and the GC

The connection between the area under the ROC curve and the Mann-Whitney statistic can be used to calculate confidence intervals for the AUC in a relatively simple manner. Moreover, it also makes it possible to test for differences between the AUC values of two rating systems which are validated on the same dataset. In both cases, advantage is taken of the fact that the Mann-Whitney statistic or the normed difference between two Mann-Whitney statistics is subject to asymptotically normal distribution. The associated variances can be easily calculated using the empirical data.[2]

Mathematical description of the binomial test

The following is a description of how the binomial test works. The binomial test can be used on an individual rating grade. In doing so, it is assumed that all K debtors in a rating grade have the same Probability of Default PD. The binomial distribution turns out to be the distribution of default events within the rating grade if it is assumed that the default events are statistically independent. Each debtor is assigned an indicator variable I_i, where I_i is given the value one if the debtor defaults, otherwise it is equal to zero. The number of default events D_K is obtained as follows

$$D_K = \sum_{i=1}^{K} I_i.$$

The null hypothesis that the actual Probability of Default at most has a value PD can now be rejected at a confidence level α if the actual default rate exceeds a critical value $d_{K,\alpha}$, which is determined by

$$P\left[D_K \geq d_{K,\alpha}\right] \leq \alpha.$$

Using the density of the binomial distribution, $d_{K,\alpha}$ is calculated as

$$d_{K,\alpha} = \min\left\{ d : \sum_{i=d}^{K} \binom{K}{i} PD^i (1 - PD)^{K-i} \leq \alpha \right\}.$$

Therefore, the probability that the critical value $d_{K,\alpha}$ is exceeded under the assumption of binomial distribution is at most α. In determining $d_{K,\alpha}$, it is assumed that all of the default events in a rating grade are independent. This is not the case in reality as default rates fluctuate in the business cycle and thus default events are correlated with one another. As a consequence, the binomial test generally underestimates $d_{K,\alpha}$. The binomial test is therefore a conservative indicator of the quality of calibration of a rating grade's Probability of Default.

[2] The relevant formulas are deliberately not given in full here. They are very complex. However, this is not a constraint for the users of these methods as the methods have been integrated into the commonly-used statistical software packages.