

In: [Wilmott Magazine](#), 2005, January, pp. 2-6

Myth and reality of discriminatory power for rating systems

Stefan Blochwitz, Alfred Hamerle, Stefan Hohl, Robert Rauhmeier, Daniel Rösch

This Draft: 2004/07/27

Disclaimer:

The views expressed herein are our own and do not necessarily reflect those of the Deutsche Bundesbank, the BIS, or the KfW.

-
- ¹ Dr. Stefan Blochwitz, Abteilung Bankgeschäftliche Prüfungen und Implementierung Basel II, Deutsche Bundesbank, Wilhelm-Epstein-Straße 14, D-60431 Frankfurt a.M., Germany
Phone: +49-69-9566-3040, Fax +49-69-9566-3040
Email: Stefan.Blochwitz@bundesbank.de
- ² Professor Alfred Hamerle, Department of Statistics, Faculty of Business Management, Economics, and Management Information Systems, University of Regensburg, 93040 Regensburg, Germany
Phone: +49-941-943-2588, Fax: +49-941-943-4936
Email: Alfred.Hamerle@wiwi.uni-regensburg.de
Internet: <http://www.wiwi.uni-regensburg.de/hamerle/>
- ³ Stefan Hohl, Bank for International Settlements, Representative Office for Asia and the Pacific, Hong Kong
Phone: +852-2878-7109
Email: Stefan.Hohl@bis.org
- ⁴ Dr. Robert Rauhmeier, KfW-Bankengruppe, Risk Management & Controlling, Methods and Modelling, Palmengartenstr. 5-9, 60325 Frankfurt a.M., Germany, Tel.: +49-69-7431-3212, Fax: +49-69-7431-3758
Email: Robert.Rauhmeier@kfw.de
- ⁵ Dr. Daniel Rösch, Assistant Professor, Department of Statistics, Faculty of Business Management, Economics, and Management Information Systems, University of Regensburg, 93040 Regensburg, Germany
Phone: +49-941-943-2752, Fax: +49-941-943-4936
Email: Daniel.Roesch@wiwi.uni-regensburg.de
Internet: <http://www.danielroesch.de>
-

Abstract

Under the Revised International Capital Framework, debtors' credit ratings and probabilities of default in credit portfolios have a key role in determining minimum regulatory capital requirements. To measure discriminatory power of rating systems, methods such as the area under a ROC curve, the Accuracy Ratio and PowerStat are often used both in academic studies and in practice. This paper shows that these measures need to be interpreted with caution since their values hinge crucially on the characteristics of the portfolio under consideration and *not* just on the quality of the rating system.

1 Introduction

For the prospective implementation of the Revised International Capital Framework (Basel II) the approach used to determine minimum regulatory capital requirements under its pillar 1 is based on the concept of measuring the credit risk by the probability of default (PD) of a borrower, see Basel Committee on Banking Supervision (2004). Probabilities of default are to be estimated by banks applying their internal credit ratings. Measuring the quality of internal rating systems is therefore of key importance for the capital ratio of a bank.

The power to discriminate credit risk (between “good” and “bad” debtors) is often highlighted as the main criterion for assessing the quality of a rating system. To measure this discriminatory power, the Accuracy Ratio (AR), the PowerStat, the Gini coefficient or the area under a receiver operating characteristic (ROC) have found widespread use both in academic studies and in practice.¹

This paper discusses the characteristics of these measures and shows that the quality of a rating system in terms of its discriminatory power (measured, for example, by AR) hinges crucially on the structure and quality of the portfolio (expressed by the true probabilities of default across all debtors). We show that the interpretation of these measures may produce misleading results in many cases with regard to the assessment of the quality of a rating system.

2 PDs, rating systems and measures of discriminatory power

The ‘default’ event occurs for debtor i among N_t debtors in period t with the probability π_{it} . This is known as the ‘probability of default’ (PD). Furthermore, it is assumed that the individual default events occur independently of each other over time and among borrowers.

Hence, PDs

- are individual, ie for each debtor i they are different ($i=1, \dots, N_t$) and
- vary over time t . In credit risk measurement usually one year PDs are considered.

¹ See, for example, Stein (2002), Sobehart/Keenan/Stein (2000), Sobehart/Keenan (2001).

Let us look first at the ideal case. Forecasting whether a customer or a loan will default in the next year implies predicting the future outcome of a random experiment. If this is compared to an analogous familiar random experiment, ie throwing the dice, it is easy to see that such a strategy is actually not possible. In the best case, it may be possible to correctly predict the *probability* of the possible outcomes of the random experiment. This is also the standpoint expressed by the Basel Committee since, in the new proposals, capital adequacy is to be determined by PDs. We therefore describe the knowledge (or correct prediction) of all probabilities of default of all debtors in a loan portfolio as the ‘ideal case’.

The correspondence between rating systems and probabilities of default can be established by means of the type of risk classification undertaken by a rating system: a rating is called a *perfect relative rating* if it can place the debtors in the correct order with regard to their individual PDs. It is not necessary to know the absolute level of the PDs to do this. If the rating predicts the correct (true) individual PD (π_{it}) for every debtor i for a given period t , the rating is called a *perfect absolute rating* for this period. If a rating system is perfect in the absolute sense, it is also perfect in the relative sense. The reverse is clearly not the case, however. A perfect absolute rating is obviously the ideal state that can be achieved in developing the rating if it is a question of forecasting the risk over a precisely defined time horizon (one year, for example).² In most cases, real ratings will depart from both the perfect absolute rating and from the perfect relative rating. The former is expected, for example, owing to estimation and forecasting uncertainties, whereas the latter occurs if the ranking produced by the rating differs from the (true) ranking of the PDs.

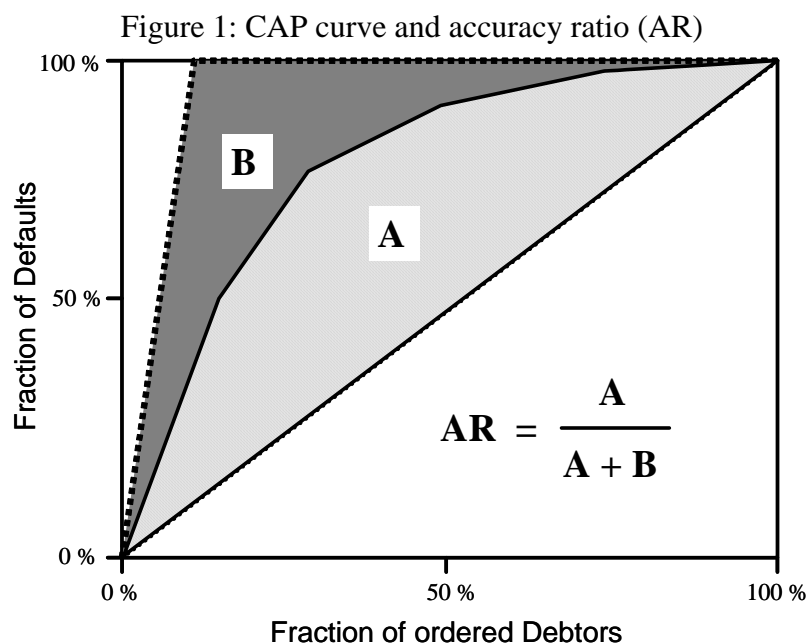
Now, how is the performance of rating systems measured in terms of discriminatory power? Let us assume that there are (ordinal or metric) risk classifications, based on a rating, for the next year for N debtors.³ After this year has elapsed, the random default events will be realised in line with the true debtor PDs.⁴ A ranking of the debtors is then established in line with

² What is important in this context is that the attribute “perfect” always refers to a precisely defined target criterion, namely, in this case, the one-year probability of default. A rating may also have as its aim the measurement of another risk, say, in a multi-year approach. It is assumed below that there is a precisely defined time horizon.

³ For the sake of simpler presentation, the t index is dropped below. It should be clearly borne in mind, however, that the probabilities of default are dependent on time.

⁴ As a matter of principle, the rating quality should be assessed out-of-sample since a rating with maximum alignment quality can always be constructed in-sample.

the assessment of their risk, starting with the riskiest debtor and ending with the debtor classified as being the least risky. In a system of coordinates, the cumulative share of the ranked debtors is then plotted on the abscissa and the share of default events accounted for by the relevant percentage of debtors relative to all defaults in the portfolio is marked off on the ordinate. This produces the power curve or CAP curve, which is illustrated in Figure 1.⁵



To measure the quality of rating systems, the area between the power curve and the diagonal (area A) is often placed in relation to the area between the “maximum” power curve and the diagonal (A + B). This gives the *accuracy ratio* (AR) or *PowerStat*. The idea behind this comparison is that, in an extreme case, all defaults should accumulate among the borrowers classified as the riskiest debtors and the power curve would therefore be identical to the hatched line.⁶ AR or PowerStat corresponds to Somer's D, which has been known for more than four decades.⁷ The accuracy ratio may also be calculated using the area under the receiver operating characteristic (area under a ROC, AUROC) or, as an equivalent, the Mann-Whitney U statistic, as⁸

⁵ This line of argument applies similarly to risk *classes*. In this case, more than one debtor has the same probability of default at a given point in time.

⁶ If the debtors were to be arranged in reverse sequence, a mirror image of the curves would be obtained on the diagonal with the same information value.

⁷ For the derivation, see Hamerle/Rauhmeier/Rösch (2003).

⁸ See Bamber (1975), Agresti (1984), Engelmann/Hayden/Tasche (2003).

$$(1) \quad AR = 2(AUROC - 0.5).$$

Now, it should be noted that the default events of this single year represent merely *one* realisation of a “random experiment”. The same is true for the CAP curve and the accuracy ratio from Figure 1. Another run of the “experiment” proceeding in accordance with the same given criteria (probabilities of default) would have resulted in a different realisation. These random deviations induce a “random error” for which the rating system is not responsible.⁹

To assess the random fluctuations, the probability distribution of AUROC or AR has to be determined. Of particular significance in this context are the expectation values $E(AUROC)$ and $E(AR)$, which may be interpreted as “averages” of the measured values of the discriminatory power given frequent repetitions of the random experiment. For the objectives focused on in this paper, it is sufficient to confine ourselves to the expectation values.

Let N debtors be ranked according to the existing score. Let the relevant score values be $S_{(1)}, \dots, S_{(N)}$, with $S_{(1)}$ denoting the score of the debtor rated “best” and $S_{(N)}$ denoting the score of the debtor rated as the riskiest. The following then holds¹⁰

$$(2) \quad \begin{aligned} E(AR) &= 2(E(AUROC) - 0.5) \\ &= \frac{1}{1 - \bar{\pi}} \left(\frac{2}{N^2 \cdot \bar{\pi}} \left(1 \cdot \pi_{S_{(1)}} + 2 \cdot \pi_{S_{(2)}} + 3 \cdot \pi_{S_{(3)}} + \dots + N \cdot \pi_{S_{(N)}} \right) - 1 - \frac{1}{N} \right) \end{aligned}$$

with $\pi_{S_{(1)}}$ denoting the (unknown) probability of default of the debtor *classified* as least risky by the rating system, etc. $\bar{\pi}$ is the average probability of default of the portfolio. It should be noted that, generally, $\pi_{S_{(1)}}$ is not necessarily the lowest probability of default, $\pi_{S_{(2)}}$ not the second-lowest, etc. If a perfect relative rating exists, however, the ranking of the rating system corresponds to the ranking according to the actual size of the PDs, ie in this case

$$\pi_{(1)} = \pi_{S_{(1)}}, \dots, \pi_{(N)} = \pi_{S_{(N)}}.$$

⁹ Therefore it is clear that a required minimum value of the realisation of the accuracy ratio can in all cases only be the lower limit of a confidence level, see also below (final section (3)).

¹⁰ For the derivation, see Hamerle/Rauhmeier/Rösch (2003).

The following statement may be derived from (2):

$E(AR)$ is maximal for a given portfolio of N debtors with a perfect relative rating.

In this case, the expression in the inner bracket is maximal because the highest probability of default is combined with the maximum rank N , the second-highest probability of default is combined with the second-highest rank $N - 1$ etc. Hence, $E(AR)$ is then also maximal.

3 Pitfalls and misinterpretations

The measures AR [AR is PowerStat, see above] and AUROC as well as all other measures of discriminatory power, based on the sequence in which the borrowers are ranked by a rating system, are interconvertible. They possess the same information content and the following statements apply equally to all measures.

Discriminatory power and randomness

The measured value of the AR is a random realisation. It is therefore by no means helpful when measuring discriminatory power to use a maximum value of one as a benchmark. That is due to the fact that a maximum value of one arises only if all the debtors are ranked correctly in relation to the random (and therefore actually unpredictable) default event. Obviously, this is not possible.

Also setting a minimum value smaller than one would be questionable. Even if a confidence interval was to be constructed around the minimum value with the sole constraint that every rating system should achieve at least an AR above the lower limit of this confidence interval, there are further compelling reasons for not specifying a minimum – as will be shown in the next section.

Discriminatory power and portfolio

The key fact for interpretation may be seen in formula (2). The value of the AR hinges crucially on the unknown probabilities of default of the debtors in the portfolio – the expression in the inner bracket.¹¹

The expected AR value achievable when analysing the portfolio depends not only on the quality of the rating system but also on the individual PDs of the debtors in the portfolio.

It follows directly from this that no assessment of a rating system's discriminatory power can be made on the basis of a given value of the discrimination measure *per se*. A simple example will illustrate this. Let us assume there are only three possible PDs. The portfolio of a bank A with 1,000 debtors is divided into

500 debtors with PD = 1%

500 debtors with PD = 5 %

Let us assume that the relative perfect rating system of this bank places the debtors without error into the two rating classes R_1 and R_2 , debtors with the small PD in R_1 , debtors with the larger PD in R_2 . We obtain $E(AR)=0.344$.

Consider now the portfolio of a second bank B divided into

500 debtors with PD = 1%

500 debtors with PD = 20 %

Let there again be a perfect rating with the same allocation to the rating classes as above. This produces $E(AR) = 0.505$. Both rating systems are perfect in the above sense and neither of the two banks can construct a better rating system because all the debtors have been ranked in the right order. Nevertheless, the two expected ARs differ considerably.

¹¹ Recently, Sobehart/Keenan (2004) showed that the measures dependent on the *default rate*. The fundamental difference should be noted: the measures depend not only on the realisation of the random variable *default rate* but also on each borrower's individual *probability of default* in the portfolio.

Discriminatory power and quality of the rating system

In practice, the quality of rating systems is very often demonstrated by means of their discriminatory power. However, as the above example has shown and, as is discussed in depth in the next section, such an assessment is shortsighted and arguably inadequate and under no circumstances provides a complete picture of a rating system's quality. Measures of discriminatory power can be meaningfully applied in situations where different rating systems can be compared on the same portfolio in the same period. This is usually the case when rating systems are in their development phase when the developer has to select one rating function among several.

Formula (2) and the remarks on the portfolio-dependency of the measures of discrimination already make it clear that the size of the absolute values of such measures by no means reflects how well the rating classifies risk. Large values do not inevitably indicate a good rating system. Low values do not necessarily mean that the rating is bad.

The false assessment becomes evident when analysing a homogeneous portfolio in which all the debtors have the same probability of default. A perfect absolute rating which assigns precisely this probability of default to the debtors has an expectation value for the AR measure of 0. It is widely held opinion, however, that a value of 0 or close to 0 for an AR measure is a "random rating". Here, it becomes evident that discriminatory power is merely one aspect of a rating system's quality which is relevant only if there is "something to discriminate".

Comparison of rating systems based on discriminatory power

(1) Different portfolios, same time period

Since the PDs of the debtors of portfolios generally differ, it follows from the remarks made in the last section that the discriminatory power of two rating systems cannot be compared on the basis of the AR values.

From this insight, it follows, in particular, that the widespread approach of making a split into a development and a validation sample is not necessarily helpful either if AR values (or the

aforementioned measures with the same information content) are used for performance measurement. An example will illustrate this.

Let us reconsider the portfolio of bank A. Of the 1,000 debtors, 700 are assigned to the development sample with the assumption (which the bank, of course, does not know) that 400 of them have a PD of 1% and 300 a PD of 5%. The remaining 300 debtors (100 with a PD of 1% and 200 with a PD of 5%) form the validation sample. Furthermore, it is assumed that the bank succeeds in constructing a perfect relative rating. For the perfect rating, expected AR measures of 0.371 and 0.252 are produced in the two samples. It is evidently wrong to conclude from this that the applied rating system has shortcomings.

(2) Same portfolio, different time periods

Non-comparability also holds if the ratings concern the same portfolio but rating A relates to point in time t_1 and rating B relates to point in time t_2 . This follows from the fact that the PDs of the debtors of a single portfolio generally change over time.

(3) Same portfolio, same time period

In this case, it is possible, in principle, to compare the discriminatory power of two rating systems on the basis of the AR values. When constructing a rating in practice, the debtors' PDs are not known and have to be predicted. Discrimination measures are validated and determined when default events have realised. But these are generated in line with the actual probabilities of default of the debtors in the portfolio. Now, if two (or more) different rating systems are analysed on the same portfolio, the probability distributions of the AR measures depend on the same probabilities of default of the debtors in the portfolio. In particular, it is the case that the "better" rating system (with the "more correct" ranking) possesses the higher expectation value. On this basis, a comparison of the expectation values can now be tested statistically. A test statistic and confidence intervals for this case are given, for example, in Engelmann/Hayden/Tasche (2003). However, it is only in this situation, ie on the same portfolio, that the test result can be interpreted in a meaningful way.

(4) Different portfolios, different time periods

This case best describes the reality at credit institutions and is also the most difficult. The non-comparability of various rating systems in this situation is due to both the differing PDs of the debtors in various portfolios and to the different points in time with changing PDs over time.

4 Conclusion

We have shown that describing the quality of a rating system by means of its discriminatory power is meaningful only in very limited situations which, in most cases, are not relevant in practice. An approach to verifying quality which goes far beyond the application of purely statistical methods therefore has to be chosen to ensure the quality of a rating system as the core of regulatory capital measurement in Basel II. Credit institutions should give careful consideration to the inclusion of all elements which might have an impact on the rating system – closely modelled on the validation puzzle picture proposed in Blochwitz/Hohl (2003). In particular, a formalised, structured and clearly documented process with clear competencies in the validation of rating systems is needed to make up for the shortcomings of the statistical procedures. Developing and implementing such a validation process is the responsibility of the credit institution.

References

- Agresti, A, 1984, *Analysis of Ordinal Categorical Data*, New York et al.
- Bamber, D, 1975, The Area Above the Ordinal Dominance Graph and the Area Below the Receiver Operating Graph, *Journal of Mathematical Psychology* 12, 387-415.
- Basel Committee on Banking Supervision, 2004, *The New Basel Capital Accord, Consultative Document*, June.
- Blochwitz, S, Hohl, S, 2003, Validierung bankinterner Ratingverfahren, *Handbuch MaK*, Schaeffer Pöschel Verlag, 261-283 (in German).
- Engelmann, B, Hayden, E, Tasche, D, 2003, Testing for Rating Accuracy, *Risk* 16, January, 82-86.
- Hamerle, A, Rauhmeier, R, Rösch, D, 2003, *Uses and Misuses of Measures for Credit Rating Accuracy*, Working Paper, University of Regensburg.
- Sobehart, JR, Keenan, SC, 2004, The Score for Credit, *Risk* 17, February, 54-58.

- Sobehart, JR, Keenan, SC, 2001, Measuring Default Accurately, Credit Risk Special Report, Risk, 14, March, 31-33.
- Sobehart, JR, Keenan, SC, Stein, RM, 2000, Benchmarking Quantitative Default Risk Models: A Validation Methodology, Moody's Rating Methodology.
- Somers, RH, 1962, A New Asymmetric Measure of Association for Ordinal Variables, American Sociological Review, 27, 799-811.
- Stein, RM, 2002, Benchmarking Default Prediction Models: Pitfalls, Moody's KMV, Technical Report #020305, New York.