DEUTSCHE
BUNDESBANK

Measuring the Discriminative Power of Rating Systems

Bernd Engelmann
(Deutsche Bundesbank)

Evelyn Hayden
(University of Vienna)

Dirk Tasche
(Deutsche Bundesbank)

Discussion paper
Series 2: Banking and Financial Supervision
No 01/2003

The discussion papers published in this series represent the authors' personal opinions and do not necessarily reflect the views of the Deutsche Bundesbank or its staff.

Deutsche Bundesbank, Wilhelm-Epstein-Strasse 14, 60431 Frankfurt am Main,
P.O.B. 10 06 02, 60006 Frankfurt am Main

Tel +49 69 9566-1
Telex within Germany 41227, telex from abroad 414431, fax +49 69 5601071

Please address all orders in writing to: Deutsche Bundesbank,
Press and Public Relations Division, at the above address or via fax No. +49 69 9566-3077

ISBN 3–935821–67–0

Measuring the Discriminative Power of Rating Systems

Abstract: Assessing the discriminative power of rating systems is an important question to banks and to regulators. In this article we analyze the Cumulative Accuracy Profile (CAP) and the Receiver Operating Characteristic (ROC) which are both commonly used in practice. We give a test-theoretic interpretation for the concavity of the CAP and the ROC curve and demonstrate how this observation can be used for more efficiently exploiting the informational contents of accounting ratios. Furthermore, we show that two popular summary statistics of these concepts, namely the Accuracy Ratio and the area under the ROC curve, contain the same information and we analyse the statistical properties of these measures. We show in detail how to identify accounting ratios with high discriminative power, how to calculate confidence intervals for the area below the ROC curve, and how to test if two rating models validated on the same data set are different. All concepts are illustrated by applications to real data.

Keywords: Validation, Rating Models, Credit Analysis
JEL Classification: C 52, G 10

Messung der Trennschärfe
von Ratingverfahren

Zusammenfassung: Die Beurteilung der Trennschärfe von Ratingverfahren ist sowohl für Banken als auch für die Bankenaufsicht von großer Bedeutung. In dieser Arbeit untersuchen wir das Cumulative Accuracy Profile (CAP) und die Receiver Operating Characteristic (ROC), welche beide in der Praxis häufig verwendet werden. Wir interpretieren die Konkavität dieser beiden Kurven anhand testtheoretischer Überlegungen und zeigen, wie diese Beobachtung zu einer effizienteren Ausnutzung des Informationsgehaltes von Bilanzkennzahlen verwendet werden kann. Darüber hinaus beweisen wir die Äquivalenz des Accuracy Ratio und der Fläche unter der ROC-Kurve und analysieren deren statistische Eigenschaften. Wir erläutern im Detail, wie Bilanzkennzahlen mit hoher Trennschärfe identifiziert, wie auf einfache Weise Konfidenzintervalle für die Fläche unter der ROC-Kurve berechnet und wie zwei Ratingverfahren, die auf demselben Datensample validiert werden, auf gleiche Trennschärfe getestet werden können. Sämtliche Konzepte werden auf reale Daten angewendet.

Schlagwörter: Validierung, Ratingverfahren, Kreditrisikoanalyse
JEL Klassifizierung: C 52, G 10

Contents

I. Introduction

A variety of rating methodologies and credit risk modelling approaches has been developed in the last three decades. Therefore, the question arises which of these methods are preferable to others. The need to judge the quality of rating systems has become increasingly important in recent years after the Basel Committee on Banking Supervision (2001) has published the second consultative document of the new capital adequacy framework where it has announced that an internal ratings-based approach could form the basis for setting capital charges with respect to credit risk in the nearest future. This is forcing banks and supervisors to develop statistical tools to evaluate the quality of internal rating systems. The importance of sound validation techniques for rating systems stems from the fact that rating models of poor quality could lead to sub-optimal capital allocation. Therefore, the Basel Committee on Banking Supervision (2000) has emphasized that the field of model validation will be one of the major challenges for financial institutions and supervisors in the foreseeable future.

In this article we focus on the evaluation of the discriminative power of rating systems. The most popular validation technique currently used in practice is the Cumulative Accuracy Profile (CAP) and its summary statistic, the Accuracy Ratio. A detailed explanation of this method can be found in Sobehart, Keenan, and Stein (2000). A concept similar to the CAP is the Receiver Operating Characteristic (ROC) and its summary statistic, the area below the ROC curve. This method has its origin in signal detection theory, psychology and especially in medicine (e.g. Hanley and McNeil (1982)).[1] ROC curves are used to evaluate the quality of medical diagnosis for many years. There exists a large body of literature that analyses the properties of ROC curves. Sobehart and Keenan (2001) explain how to use this concept for validating internal rating models. In their article, they concentrate on the fundamental features of ROC curves like their calculation and their interpretation. However, both the articles by Sobehart, Keenan, and Stein (2000) and Sobehart and Keenan (2001) do not analyse the measures presented in these articles from a statistical point of view.

In this article our focus will be on the statistical properties of the CAP and the ROC. In our analysis we will concentrate on the ROC curve for two reasons. First, concentrating on the ROC allows us to use the results given in the medical literature and second, the properties of ROC curves are much more intuitive than the results for the CAP. We will show how the area below the ROC curve can be interpreted in terms of a probability, how confidence intervals for the area below the ROC curve can be calculated, and how the areas below the ROC curves of two different rating methods can be compared statistically. We will demonstrate how these techniques have to be modified that they are applicable also for the CAP.

[1] An interesting overview of the variety of possible applications of ROC curves is given in Swets (1988).

The rest of the article is organised as follows. In part II, to keep this article self-contained, we briefly review the concepts of the CAP and the ROC. For both concepts it is possible to summarize the information about the quality of a rating system with a single number, namely with the Accuracy Ratio and the area below the ROC curve. In part III we will analyse the statistical properties of both the ROC and the CAP. We will start with a detailed description of the properties of the ROC and show how these properties have to be modified to be applicable to the CAP. In part IV, we will apply the techniques presented in the second part to real data and discuss their reliability. The final section concludes.

Throughout this article we will assume rating systems that produce a finite number of rating scores. This is the situation that is mainly found in practice. However, it is straightforward to apply all methods presented in this article to rating systems that deliver continuous scores.

II. The Cumulative Accuracy Profile and the Receiver Operating Characteristic

Consider a rating model which assigns to each debtor a score s out of k possible values $\{s_1, \ldots, s_k\}$ with $s_1 < \ldots < s_k$. A high rating score indicates a low default probability. It is our aim to evaluate the quality of this rating model. We can do this by assigning scores to debtors from a data sample that is used for the validation, and checking if the debtors will default over the next period or remain solvent. In this context, we introduce three random variables, S_T, S_D, and S_{ND}. The random variable S_T describes the score distribution of all debtors, S_D and S_{ND} model the score distributions of the defaulters and the non-defaulters, respectively. The probability that a defaulter has a score value s_i is denoted by p_D^i, $p_D^i \geq 0$, $\sum_{i=1}^k p_D^i = 1$. The probability that a non-defaulter has a score value s_i is called p_{ND}^i. Given the a-priori default probability π of all debtors, we find for the probability p_T^i that an arbitrary debtor has a score value s_i

$$p_T^i = \pi p_D^i + (1 - \pi) p_{ND}^i.$$

We define the cumulative probabilities

$$CD_D^i = \sum_{j=1}^{i} p_D^j,\ i = 1, \ldots, k, \tag{1a}$$

$$CD_{ND}^i = \sum_{j=1}^{i} p_{ND}^j,\ i = 1, \ldots, k, \tag{1b}$$

$$CD_T^i = \sum_{j=1}^{i} p_T^j,\ i = 1, \ldots, k, \tag{1c}$$

where CD_D, CD_{ND}, and CD_T denote the distribution function of the score values of the defaulters, the non-defaulters, and the total sample of debtors, respectively. For instance, CD_D^i denotes the probability that a defaulter's score is not greater than i. Additionally, we define $CD_D^0 = CD_{ND}^0 = CD_T^0 = 0$.

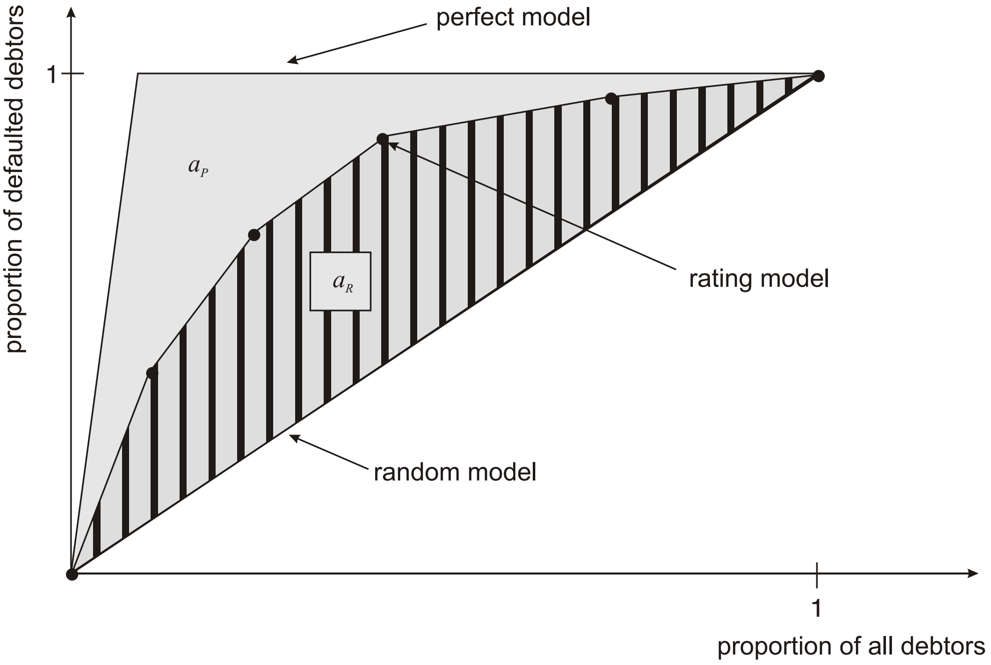
II.1. Cumulative Accuracy Profile

The Cumulative Accuracy Profile is defined as the graph of all points $(CD_T^i, CD_D^i)_{i=0,\ldots,k}$ where the points are connected by straight lines (linear interpolation). This is illustrated in Figure 1.

A perfect rating model would assign the lowest scores to the defaulters. In this case the CAP is increasing linearly and then staying at one. For a random model without any discriminative power the fraction x of all debtors with the lowest rating scores will contain x percent of all defaulters,

Figure 1. Cumulative Accuracy Profile

This figure illustrates the concept of a CAP. The polygon shows the performance of the model being evaluated in depicting the percentage of defaults captured by the model at different percentages of the data set, while the straight line below represents the naive case of zero information or random assignment of rating scores. The third line represents the case of perfect information where all defaults are assigned to the lowest rating scores. The Accuracy Ratio is the ratio of the performance improvement of the model being evaluated over the naive model (a_R) to the performance improvement of the perfect model over the naive model (a_P).



i.e. in this case we will have $CD_D^i = CD_T^i$, $i = 0, \ldots, k$. Real rating systems will be somewhere in between these two extremes. The quality of a rating system can be summarized by a single number, the Accuracy Ratio AR. It is defined as the ratio of the area a_R between the CAP of the rating model being validated and the CAP of the random model, and the area a_P between the CAP of the perfect rating model and the CAP of the random model, i.e.

$$AR = \frac{a_R}{a_P}. \tag{2}$$

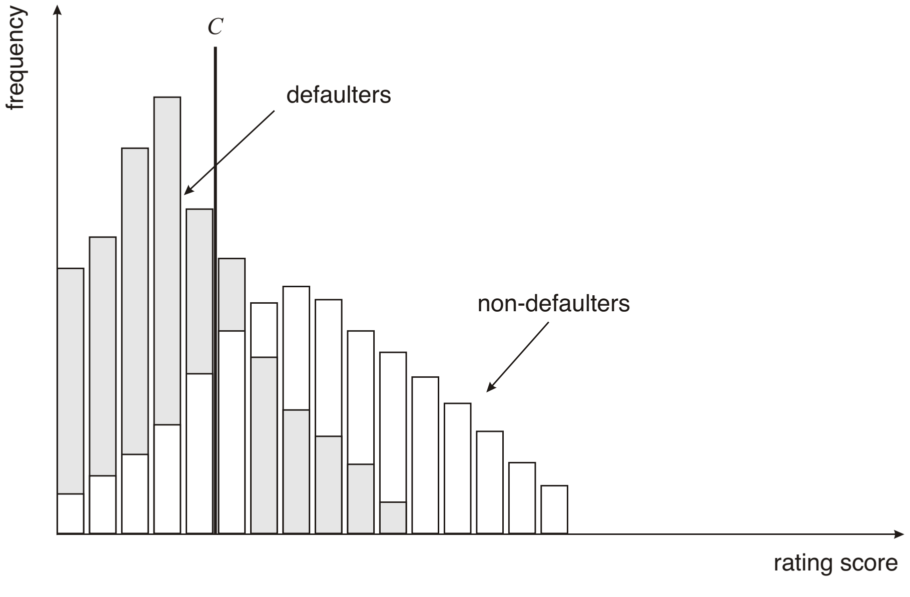
Thus, the rating method is the better the closer AR is to one.

II.2. Receiver Operating Characteristic

In this part, we explain the ROC and its associated summary statistic, the area under the ROC curve. The construction of a ROC curve is illustrated in Figure 2 which sketches possible distributions of rating scores for defaulting and non-defaulting debtors. For a perfect rating model the left distribution and the right distribution in Figure 2 would be separate. For real rating systems perfect discrimination in general is not possible. Both distributions will overlap as illustrated in Figure 2.

Figure 2. Distribution of rating scores for defaulting and non-defaulting debtors

This figure depicts possible distributions of rating scores for defaulting and non-defaulting obligors. For a perfect rating model the distributions would be separate. For real rating systems, however, perfect discrimination in general is not possible and the two distributions overlap.



Assume someone has to use the rating scores to decide which debtors will survive during the next period and which debtors will default. One possibility for the decision-maker would be to introduce a cut-off value C as in Figure 2, and to classify each debtor with a rating score lower than C as a potential defaulter and each debtor with a rating score higher than C as a non-defaulter. Then four decision results would be possible. They are summarized in Table 1.

Table 1
Decisions results given the cut-off value C

This table summarizes the possible consequences for a decision-maker using the cut-off value C.

		default	no default
rating score	below C	correct prediction (hit)	wrong prediction (false alarm)
	above C	wrong prediction (miss)	correct prediction (correct rejection)

If the rating score is below the cut-off value C and the debtor defaults subsequently, the decision was correct. Otherwise the decision-maker wrongly classified a non-defaulter as a defaulter (type I error). If the rating score is above the cut-off value and the debtor does not default, the classification was correct. Otherwise a defaulter was incorrectly assigned to the non-defaulters group (type II error). Using the notation of Sobehart and Keenan (2001), we define the hit rate $HR(C)$ (equal to the grey area on the left hand side of the cut-off value C in Figure 2) as

$$HR(C) = P(S_D \leq C). \tag{3}$$

The false alarm rate $FAR(C)$ (equal to the white area on the left hand side of the cut-off value C in Figure 2) is defined as

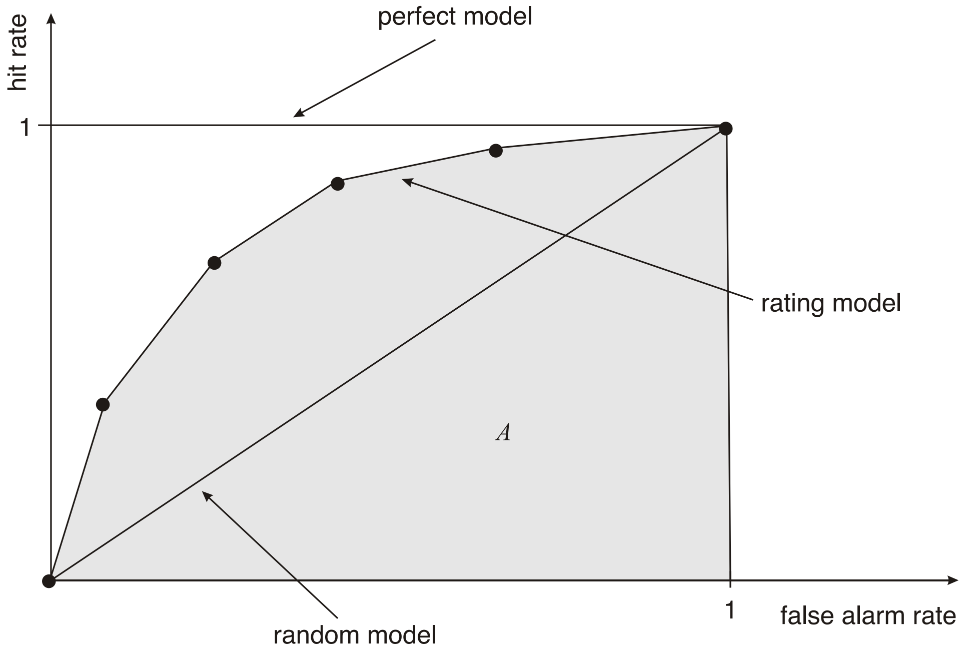
$$FAR(C) = P(S_{ND} \leq C). \tag{4}$$

The ROC curve is constructed as follows. For all cut-off values C that are contained in the range of the rating scores the quantities $HR(C)$ and $FAR(C)$ are computed. The ROC curve is a plot of $HR(C)$ versus $FAR(C)$ for all values of C. In our setting, the ROC curve consists of all points $(CD_{ND}^i, CD_D^i)_{i=0,\ldots,k}$. As in the case of the CAP these points are connected by linear interpolation. This is illustrated in Figure 3.

A rating model's performance is the better the steeper the ROC curve is at the left end and the closer the ROC curve's position is to the point (0,1). Similarly, the model is the better the larger the area under the ROC curve is. We denote this area by AUC (area under curve). It can be interpreted as the average power of the tests on default / non-default corresponding to all possible cut-off values C. The area AUC is 0.5 for a random model without discriminative power and is 1.0 for a perfect model. It is between 0.5 and 1.0 for any reasonable rating model in practice.

Figure 3. Receiver Operating Characteristic Curves

This figure shows a ROC curve. For all possible cut-off values C the fraction of defaulters predicted correctly ($HR(C)$) and the fraction of false alarms ($FAR(C)$) are computed. The ROC curve is a plot of $HR(C)$ versus $FAR(C)$.



III. Properties of the ROC and the CAP

In this section we analyse some statistical properties of both the CAP and the ROC. We will start with the ROC because it offers more intuitive results than the CAP. For this reason, there exists a large body of literature on the ROC curve in medicine and psychology. We will mainly refer to the results provided in this literature in the first part of this section. In the second part of this section, we will show how the results for the ROC can be transferred to the CAP.

III.1. Properties of the ROC

Most of the results we present here are well known in the medical literature. The probabilistic interpretation of the ROC curve and an efficient way to calculate confidence intervals using asymptotic normality are based on an article of Bamber (1975). The test to compare the areas under the ROC curves of two different rating systems that are validated on the same data is based on DeLong, DeLong, and Clarke-Pearson (1988).

III.1.1. Shape of the ROC curve

From its definition, it is obvious that the ROC curve is non-decreasing. It is also well known (Bamber 1975) that the ROC curve is concave if and only if the Likelihood Ratio

$$LR_i = \frac{p_D^i}{p_{ND}^i}, \quad i = 1, \ldots, k, \tag{5}$$

is non-increasing in i. This property is quite intuitive since the probability of receiving a high score should be large for a non-defaulting debtor but small for a defaulting debtor. It is also easy to see that concavity of the CAP is equivalent to the monotonicity of the likelihood ratio.

Actually, concavity of the ROC curve has also a decision-theoretic interpretation. Besides the cut-off decision rules as described in Section II.2 above, a lot of other rules are conceivable. For instance, there might be rating systems such that very high or very low scores indicate default. However, it can be shown (Tasche 2002) that monotonicity of the Likelihood Ratio is equivalent to the optimality of the cut-off rules in the following sense: For any fixed cut-off value, there is no decision rule with both lower type I and type II errors. In the case of rating systems with finitely many categories the monotonicity can always be reached by reordering. This is current practice in the medical sciences (Lee 1999).

III.1.2. Probabilistic Interpretation

We continue by providing a probabilistic interpretation of AUC. Consider the following experiment. Two debtors are drawn at random, the first one from the distribution of defaulters, the second one from the distribution of non-defaulters. The scores of the defaulter and the non-defaulter determined this way can be interpreted as realizations of the two independent random variables S_D and S_{ND} we have introduced at the beginning of Section II. Assume someone has to decide which of the debtors is the defaulter. A rational decision-maker might suppose that the defaulter is the debtor with the lower rating score. If both debtors had the same score she would toss a coin. Therefore, the probability that her decision is correct is equal to $P(S_D < S_{ND}) + \frac{1}{2}\,P(S_D = S_{ND})$. A simple calculation shows that this probability is exactly equal to the area AUC below the ROC curve.

$$\begin{aligned}
AUC &= \sum_{i=1}^{k} \frac{1}{2}\,(CD_D^i + CD_D^{i-1})\,(CD_{ND}^i - CD_{ND}^{i-1}) \\
&= \sum_{i=1}^{k} \frac{1}{2}\,\big(P(S_D \le s_i) + P(S_D \le s_{i-1})\big)\,P(S_{ND} = s_i) \\
&= \sum_{i=1}^{k} \big(P(S_D \le s_{i-1}) + \frac{1}{2}\,P(S_D = s_i)\big)\,P(S_{ND} = s_i) \\
&= \sum_{i=1}^{k} P(S_D \le s_{i-1})\,P(S_{ND} = s_i) + \frac{1}{2}\,\sum_{i=1}^{k} P(S_D = s_i)\,P(S_{ND} = s_i) \\
&= P(S_D < S_{ND}) + \frac{1}{2}\,P(S_D = S_{ND}) \qquad (6)
\end{aligned}$$

III.1.3. Calculation of Confidence Intervals for AUC

In this part of the article we discuss a simple method of calculating confidence intervals for AUC, the area below the ROC curve. The interpretation of AUC as a probability relates to the test statistic of the U-test of Mann and Whitney (1947). If we draw a defaulter with score s_D from S_D and a non-defaulter with score s_{ND} from S_{ND} and define $u_{D,ND}$ as

$$u_{D,ND} = \begin{cases} 1, \text{if}\, s_D < s_{ND} \\ \frac{1}{2}, \text{if}\, s_D = s_{ND} \\ 0, \text{if}\, s_D > s_{ND} \end{cases}, \qquad (7)$$

then the test statistic $\hat{U}$ according to Mann-Whitney is defined as

$$\hat{U} = \frac{1}{N_D\, N_{ND}} \sum_{(D,ND)} u_{D,ND}, \tag{8}$$

where the sum is over all pairs of defaulters and non-defaulters (D, ND) in the sample. The numbers of defaulters and non-defaulters in the validation sample are denoted by N_D and N_{ND} respectively. Observe that $\hat{U}$ is an unbiased estimator of $P(S_D < S_{ND}) + \frac{1}{2}\, P(S_D = S_{ND})$, i.e.

$$AUC = E(\hat{U}) = P(S_D < S_{ND}) + \frac{1}{2}\, P(S_D = S_{ND}). \tag{9}$$

Furthermore, we find that the area $\widehat{AUC}$ below the ROC curve calculated from the empirical data is equal to $\hat{U}$. For the variance $\sigma^2_{\hat{U}}$ of $\hat{U}$ we find the unbiased estimator $\hat{\sigma}^2_{\hat{U}}$ as

$$\begin{aligned}\hat{\sigma}^2_{\hat{U}} = \frac{1}{4\,(N_D - 1)\,(N_{ND} - 1)}\, &[\hat{P}_{D \neq ND} + (N_D - 1)\, \hat{P}_{D,D,ND} \\ &+ (N_{ND} - 1)\, \hat{P}_{ND,ND,D} - 4\,(N_D + N_{ND} - 1)\, (\hat{U} - \frac{1}{2})^2]\end{aligned} \tag{10}$$

where $\hat{P}_{D \neq ND}$ is an estimator for $P(S_D \neq S_{ND})$ and $\hat{P}_{D,D,ND}$ and $\hat{P}_{ND,ND,D}$ are estimators for the expressions $P_{D,D,ND}$ and $P_{ND,ND,D}$ which are defined as

$$\begin{aligned} P_{D,D,ND} = {} & P(S_{D,1}, S_{D,2} < S_{ND}) + P(S_{ND} < S_{D,1}, S_{D,2}) \\ & - P(S_{D,1} < S_{ND} < S_{D,2}) - P(S_{D,2} < S_{ND} < S_{D,1}), \end{aligned} \tag{11a}$$

$$\begin{aligned} P_{ND,ND,D} = {} & P(S_{ND,1}, S_{ND,2} < S_D) + P(S_D < S_{ND,1}, S_{ND,2}) \\ & - P(S_{ND,1} < S_D < S_{ND,2}) - P(S_{ND,2} < S_D < S_{ND,1}). \end{aligned} \tag{11b}$$

In (11a) and (11b), the quantities $S_{D,1}$, $S_{D,2}$ are independent observations randomly sampled from S_D and $S_{ND,1}$, $S_{ND,2}$ are independent observations randomly sampled from S_{ND}. This unbiased estimator $\hat{\sigma}^2_{\hat{U}}$ is implemented in many standard statistical software packages.

For $N_D, N_{ND} \to \infty$ it is known that $(AUC - \hat{U})/\hat{\sigma}_{\hat{U}}$ is asymptotically normally distributed with mean zero and standard deviation one. This allows the calculation of approximate confidence intervals at level α for AUC by

$$\Big[\hat{U} - \hat{\sigma}_{\hat{U}} \Phi^{-1}\big(\frac{1+\alpha}{2}\big), \hat{U} + \hat{\sigma}_{\hat{U}} \Phi^{-1}\big(\frac{1+\alpha}{2}\big)\Big], \tag{12}$$

where Φ denotes the cumulative distribution function of the standard normal distribution. Our numerical explorations in Section IV indicate that the number of defaults should be at least 50 in order to guarantee that (12) is a good approximation. We note that there is no clear rule for which values of $\hat{U}$ the asymptotic normality of $\hat{U}$ is a valid approximation, because $\hat{U}$ can solely take values in the interval [0,1]. If $\hat{U}$ is only a few standard deviations away from one it is clear that the normal approximation might be inaccurate[2]. However, as illustrated in our examples below, even in this situation the normal approximation can lead to reasonable results.

III.1.4. Testing for Discriminative Power

The confidence intervals for AUC can be used to test if a rating system has any discriminative power at all. In this case, the null hypothesis would be $AUC = 0.5$ or, equivalently, $S_D = S_{ND}$ in distribution. Under the null hypothesis (10) simplifies considerably. One obtains

$$\sigma^2_{\hat{U}} = \frac{N_D + N_{ND} + 1}{12\, N_D\, N_{ND}}. \tag{13}$$

Given a confidence level α asymptotic normality can be applied to test if the rating system has enough discriminative power to reject the null hypothesis of no discriminative power.

III.1.5. Comparing two Areas under the ROC Curve

One major application of both the CAP and the ROC is the comparison of different methods on the same data. We consider the case of comparing two rating systems 1 and 2 with areas below the ROC curve AUC_1 and AUC_2. Just comparing the single numbers only is insufficient since they are not very meaningful from a statistical point of view. Comparing confidence intervals could also be misleading because a potential correlation of both rating methods is neglected in this case. To construct a rigorous test on the difference of AUC_1 and AUC_2 it is necessary to calculate the variances $\hat{\sigma}^2_{\hat{U}_i}$ for the estimators $\hat{U}_i$ of AUC_i, $i = 1, 2$. In addition, we need the covariance $\hat{\sigma}_{\hat{U}_1,\hat{U}_2}$ between the estimators $\hat{U}_1$ and $\hat{U}_2$ of AUC_1 and AUC_2. We find for the covariance

$$\begin{aligned} \hat{\sigma}_{\hat{U}_1,\hat{U}_2} = {} & \frac{1}{4\,(N_D - 1)\,(N_{ND} - 1)}\,[\tilde{P}^{12}_{D,D,ND,ND} + (N_D - 1)\,\tilde{P}^{12}_{D,D,ND} \\ & + (N_{ND} - 1)\,\tilde{P}^{12}_{ND,ND,D} - 4\,(N_D + N_{ND} - 1)\,(\hat{U}_1 - \frac{1}{2})\,(\hat{U}_2 - \frac{1}{2})], \end{aligned} \tag{14}$$

[2] Several methods for the computation of confidence intervals without relying on the assumption of asymptotic normality are known which lead in general to very conservative confidence intervals. An overview of these methods is given in Bamber (1975). One could rely on these methods if the normal approximation is questionable as in the case of very few defaults in the validation sample.

where $\tilde{P}^{12}_{D,D,ND,ND}$, $\tilde{P}^{12}_{D,D,ND}$ and $\tilde{P}^{12}_{ND,ND,D}$ are estimators for $P^{12}_{D,D,ND,ND}$, $P^{12}_{D,D,ND}$ and $P^{12}_{ND,ND,D}$ which are defined as[3]

$$\begin{aligned} P^{12}_{D,D,ND,ND} &= P(S^1_D > S^1_{ND}, S^2_D > S^2_{ND}) + P(S^1_D < S^1_{ND}, S^2_D < S^2_{ND}) \\ &\quad - P(S^1_D > S^1_{ND}, S^2_D < S^2_{ND}) - P(S^1_D < S^1_{ND}, S^2_D > S^2_{ND}), \end{aligned} \tag{15a}$$

$$\begin{aligned} P^{12}_{D,D,ND} &= P(S^1_{D,1} > S^1_{ND}, S^2_{D,2} > S^2_{ND}) + P(S^1_{D,1} < S^1_{ND}, S^2_{D,2} < S^2_{ND}) \\ &\quad - P(S^1_{D,1} > S^1_{ND}, S^2_{D,2} < S^2_{ND}) - P(S^1_{D,1} < S^1_{ND}, S^2_{D,2} > S^2_{ND}), \end{aligned} \tag{15b}$$

$$\begin{aligned} P^{12}_{ND,ND,D} &= P(S^1_D > S^1_{ND,1}, S^2_D > S^2_{ND,2}) + P(S^1_D < S^1_{ND,1}, S^2_D < S^2_{ND,2}) \\ &\quad - P(S^1_D > S^1_{ND,1}, S^2_D < S^2_{ND,2}) - P(S^1_D < S^1_{ND,1}, S^2_D > S^2_{ND,2}). \end{aligned} \tag{15c}$$

The quantities S^i_D, $S^i_{D,1}$, and $S^i_{D,2}$ are independent draws from the sample of defaulters. The upper index i indicates whether a score of the rating model 1 or a score of the rating model 2 has to be taken. The meaning of S^i_{ND}, $S^i_{ND,1}$, and $S^i_{ND,2}$ is analogous.

To carry out the test on the difference between the two rating methods (where the null hypothesis is equality of both areas below the ROC curve), we have to evaluate the test statistic T which is defined as

$$T = \frac{(\hat{U}_1 - \hat{U}_2)^2}{\sigma^2_{\hat{U}_1} + \sigma^2_{\hat{U}_2} - 2\sigma_{\hat{U}_1,\hat{U}_2}}. \tag{16}$$

This test statistic is asymptotically χ^2-distributed with one degree of freedom. Given a confidence level α, we can calculate critical values from the $\chi^2(1)$-distribution for the test statistic T.

III.2. Properties of the CAP

All concepts we have presented in Section III.1 are also applicable to the CAP and its summary statistic AR. The key to transfer the statistical results for the ROC and AUC to the CAP and AR is the relation

$$AR = 2\,AUC - 1. \tag{17}$$

A proof of (17) is given in Appendix A. Using (17), we get an estimator for the Accuracy Ratio by the Mann-Whitney test statistic

$$\hat{V} = \frac{1}{N_D\,N_{ND}} \sum_{(D,ND)} v_{D,ND}, \tag{18}$$

[3] The expressions given in DeLong, DeLong, and Clarke-Pearson (1988) look different from the expressions here. However, it can be shown that both are equivalent. We used this notation to be consistent with the notation of Section III.1.3.

with $v_{D,ND}$ defined as

$$v_{D,ND} = \begin{cases} 1, & \text{if } s_D < s_{ND} \\ 0, & \text{if } s_D = s_{ND} \\ -1, & \text{if } s_D > s_{ND} \end{cases}, \tag{19}$$

where s_D and s_{ND} are the scores of a randomly chosen defaulter and a randomly chosen non-defaulter, respectively[4].

For the variance $\hat{\sigma}^2_{\hat{V}}$ of $\hat{V}$ we find

$$\begin{aligned} \hat{\sigma}^2_{\hat{V}} = \frac{1}{(N_D-1)\,(N_{ND}-1)}\,[&\hat{P}_{D\neq ND} + (N_D-1)\,\hat{P}_{D,D,ND} \\ &+ (N_{ND}-1)\,\hat{P}_{ND,ND,D} - (N_D+N_{ND}-1)\,\hat{V}^2], \end{aligned} \tag{20}$$

where $\hat{P}_{D\neq ND}$, $\hat{P}_{D,D,ND}$, and $\hat{P}_{ND,ND,D}$ are defined exactly as in Section III.1.3. For the covariance $\hat{\sigma}_{\hat{V}_1,\hat{V}_2}$ between two Accuracy Ratios V_1 and V_2 we find

$$\begin{aligned} \hat{\sigma}_{\hat{V}_1,\hat{V}_2} = \frac{1}{(N_D-1)\,(N_{ND}-1)}\,[&\tilde{P}^{12}_{D,D,ND,ND} + (N_D-1)\,\tilde{P}^{12}_{D,D,ND} \\ &+ (N_{ND}-1)\,\tilde{P}^{12}_{ND,ND,D} - (N_D+N_{ND}-1)\hat{V}_1\,\hat{V}_2], \end{aligned} \tag{21}$$

where $\tilde{P}^{12}_{D,D,ND,ND}$, $\tilde{P}^{12}_{D,D,ND}$ and $\tilde{P}^{12}_{ND,ND,D}$ are defined as in Section III.1.5.

Taking all this together allows the calculation of confidence intervals for AR and the comparison of different rating systems by their Accuracy Ratios.

[4] This implies a probabilistic interpretation of AR, namely that $AR = P(S_{ND} > S_D) - P(S_D > S_{ND})$.

IV. Applications

In this section, we apply the concepts presented in Section III to real data. We use a database of the Deutsche Bundesbank which contains balance sheets of small and medium companies that are not listed on exchanges for the years 1987 – 1999. It contains about 300,000 balance sheets and about 3,000 defaults where default was defined as insolvency. In the first part of this section, we will show how to use the concept presented in Section III.1.1 to identify accounting ratios with high discriminative power that could be included into rating systems. In the second part, we will calculate confidence intervals for AUC using the normal approximation of Section III.1.3 and compare the results to bootstrapping in order to get a feeling for the reliability of this approximation. In the final part of this section, we illustrate the test on the difference of two rating models presented in Section III.1.5 by real examples. We carry out all applications using AUC as a quality measure. All this could also be done using AR as outlined in Section III.2.

In all the examples of this section we assume a rating system with 20 rating categories. The obligors are distributed to the rating categories in such a way that the categories are approximately of equal size. To be more precise, after we estimated a rating model, the debtors are ordered from the lowest score to the highest score. In the next step the debtors are distributed to the rating categories. All debtors in one rating category get the number i, $1 \leq i \leq 20$, of the category as their rating score.

IV.1. Identification of Accounting Ratios with Discriminative Power

In this section we apply the technique presented in Section III.1.1. When designing a rating system it is crucial to identify accounting ratios with high discriminative power. The calculation of AUC for the accounting ratios could be misleading in some situations. This is illustrated in Figure 4 where we see a score distribution of the defaulters which is partly on the left and partly on the right of the distribution of the non-defaulters. Such a score distribution clearly has discriminative power. A straightforward calculation of AUC, however, results in a value close to 0.5, the same value a score function without discriminative power would result in.

Instead of calculating the AUC for rating scores of the defaulters and the non-defaulters, it is more reasonable to calculate AUC using likelihood ratios as a score. This ensures that accounting ratios or models with high discriminative power can be identified by AUC. We illustrate this in Figures 5, 6, and 7.

Figure 5 shows the analysis for the accounting ratio “Ordinary Business Income/Total Assets”. We see that in this case the rating score is almost perfectly correlated to the likelihood ratio.

Figure 4. Score function where AUC would be misleading

In this figure, the score distribution of the defaulters is partly on the left and partly on the right of the distribution of the non-defaulters. A straightforward calculation of AUC would result in a value close to 0.5 which could lead to the wrong conclusion that the score function has no discriminative power.

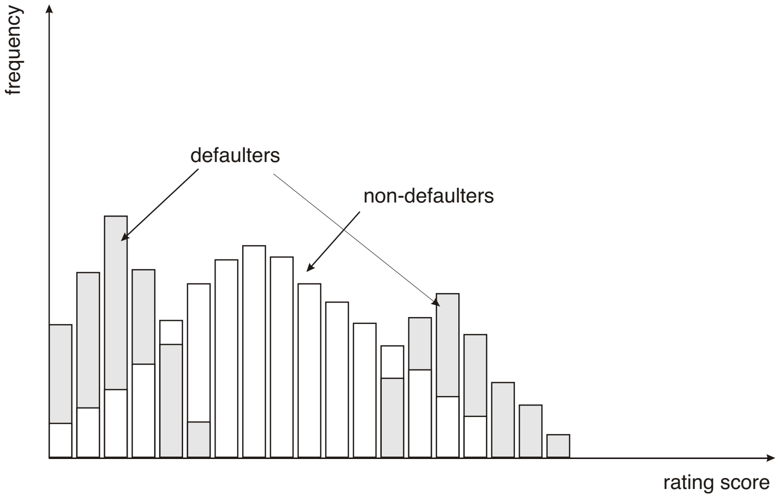
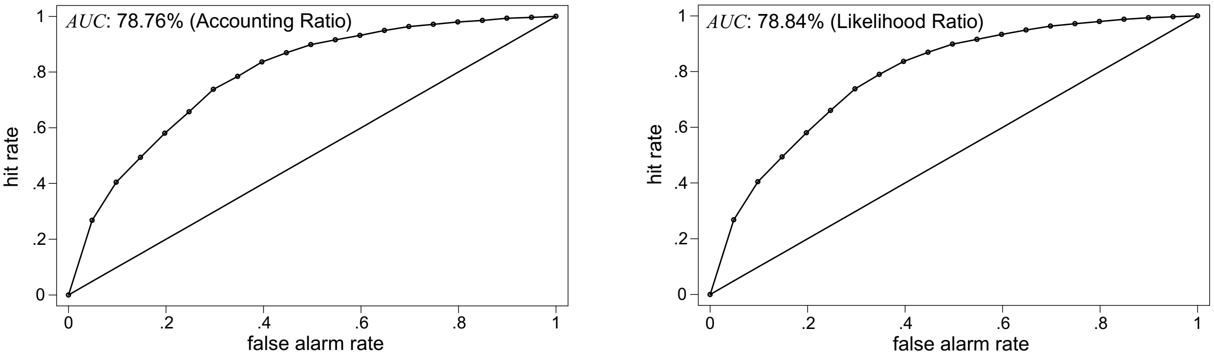


Figure 5. Ordinary Business Income / Total Assets

The figure on the left shows the ROC (and the corresponding AUC) when the debtors are sorted by the accounting ratio "Ordinary Business Income/Total Assets". The figure on the right shows the resulting ROC when debtors are sorted by their corresponding likelihood ratios.



The situation is different for the accounting ratio “Change in (Net Sales/Total Assets)” as illustrated in Figure 6 below. The ROC curve is not concave in this situation. Therefore, using the likelihood ratio is necessary to gain the full information on the discriminative power of this accounting ratio. The same is true for “Current Assets/Total Assets” in Figure 7 although this accounting ratio does not contain much discriminative power.

Figure 6. Change in (Net Sales/Total Assets)

The figure on the left shows the ROC (and the corresponding AUC) when the debtors are sorted by the accounting ratio “Change in (Net Sales / Total Assets)”. The figure on the right shows the resulting ROC when debtors are sorted by their corresponding likelihood ratios.

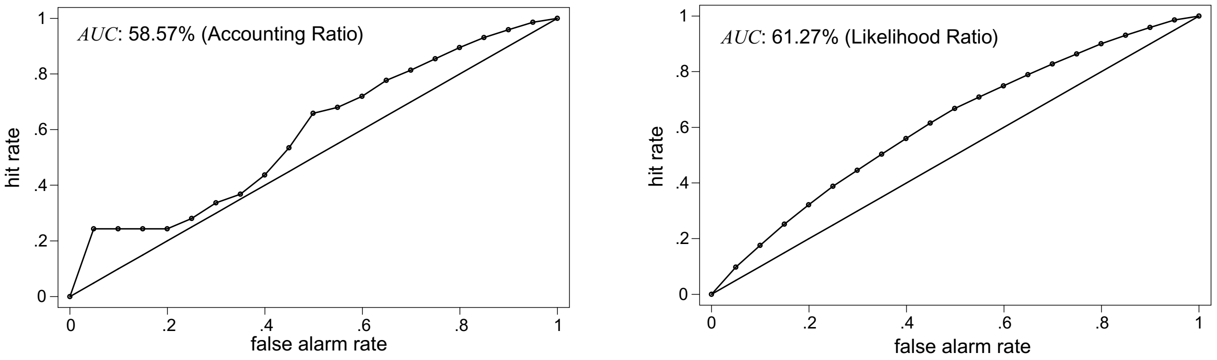
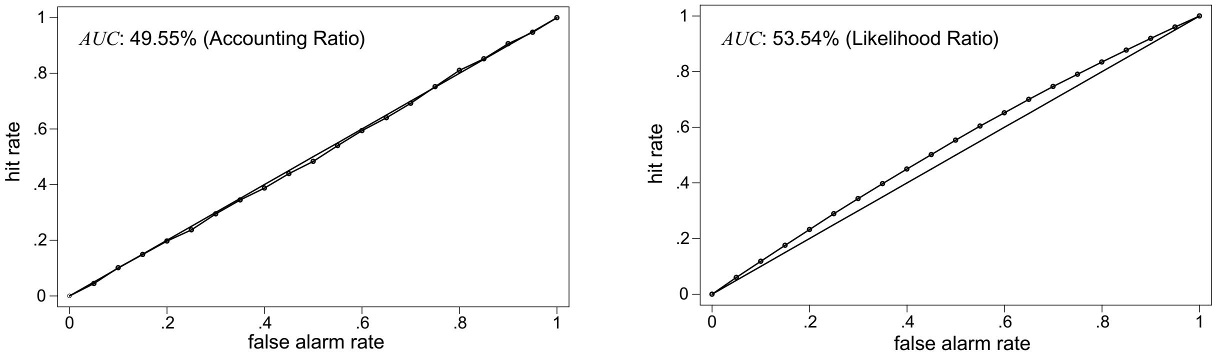


Figure 7. Current Assets/Total Assets

The figure on the left shows the ROC (and the corresponding AUC) when the debtors are sorted by the accounting ratio “Current Assets / Total Assets”. The figure on the right shows the resulting ROC when debtors are sorted by their corresponding likelihood ratios.



We see that likelihood ratios are valuable in detecting accounting ratios with high discriminative power. Their use is optimal from a theoretical point of view as explained in Section III.1.1. Therefore, they should be used as inputs for the estimation of a rating model instead of the pure accounting ratios or any other transformation.

IV.2. Calculation of Confidence Intervals for AUC

In this part of the article we analyze the calculation of confidence intervals for AUC based on formula (12). Since this formula is based on an asymptotic result it is not clear for which values of N_D and N_{ND} it is a reasonable approximation. As a benchmark we compute confidence intervals based on bootstrapping. A good overview on bootstrapping is given in Efron and Tibshirani (1998).

We construct three logit-models using four accounting ratios for each model to carry out the validation exercises. We estimate the models using the balance sheets of the years 1987 – 1993 from the database we described above. The logit-scores of the three models are given in (22a), (22b), and (22c).

$$\begin{aligned}\text{Model 1} = \quad & -7.74 + 2.85 \cdot \text{Liabilities/Total Assets} \\ & -0.40 \cdot \text{Net Sales/Total Assets} \\ & -12.18 \cdot \text{Ordinary Business Income/Total Assets} \\ & +1.93 \cdot \text{Current Liabilities/Total Assets}\end{aligned} \tag{22a}$$

$$\begin{aligned}\text{Model 2} = \quad & -4.01 - 1.53 \cdot \text{Equity/Total Assets} \\ & -5.43 \cdot \text{EBIT/Interest Expenses} \\ & -5.04 \cdot \text{Ordinary Business Income/Total Assets} \\ & +0.97 \cdot \text{Bank Debt/Liabilities}\end{aligned} \tag{22b}$$

$$\begin{aligned}\text{Model 3} = \quad & -5.25 - 1.10 \cdot \text{Equity/Total Assets} \\ & -0.40 \cdot \text{Net Sales/Total Assets} \\ & -12.08 \cdot \text{Ordinary Business Income/Total Assets} \\ & +2.18 \cdot \text{Current Liabilities/Total Assets}\end{aligned} \tag{22c}$$

In our first exercise, we validate Model 1, Model 2, and Model 3 on the the whole data set of the years 1994 – 1999 from the database described above. This sample of the database contained about 200,000 balance sheets and about 825 defaults. We calculate $\hat{U}$ and $\hat{\sigma}_{\hat{U}}$ as in (8) and (10) for all three models. Furthermore, we compute 95% confidence intervals and 99% confidence intervals for $\hat{U}$ with (12) which is based on asymptotic normality. To evaluate the quality of the normal approximation, we additionally calculate confidence intervals by bootstrapping[5]. Not surprisingly, for this large data sample we find almost perfect agreement between the confidence intervals based on asymptotic normality and the confidence intervals computed by bootstrapping. The results are summarized in Table 2.

In a second validation experiment we want to evaluate the accuracy of (12) for small values of N_D. We randomly draw four portfolios of 500 obligors. The first portfolio contains 100 defaulters, the second portfolio 50 defaulters, the third portfolio 20 defaulters, and the fourth portfolio 10 defaulters. For each portfolio we compute $\hat{U}$, $\hat{\sigma}_{\hat{U}}$, 95% confidence intervals for $\hat{U}$, and 99% confidence intervals for $\hat{U}$ for the three rating models (22a), (22b), and (22c). The results are given in Table 3. We see that for the portfolio with 100 defaults and the portfolio with 50 defaults the confidence intervals based on asymptotic normality agree almost perfectly with the confidence intervals calculated by bootstrapping. For the portfolio with 20 defaults and especially for the portfolio with 10 defaults we would expect that the normal approximation is rather inaccurate. In fact, the confidence intervals based on bootstrapping are no longer symmetric. However, the results using the normal approximation are still close to the bootstrapping results. Therefore, we conclude that the normal approximation is applicable to practically all rating systems we could observe in practice. The main advantage of using the normal approximation for the calculation of confidence intervals is the considerably lower computational time for obtaining them. Bootstrapping can take several hours especially if the portfolio is large.

Table 2
Confidence intervals for Model 1, Model 2, and Model 3 for the total portfolio

This table shows the results for $\hat{U}$, $\sigma_{\hat{U}}$, 95%, and 99% confidence intervals (derived by asymptotic normality and bootstrapping) for Model 1, Model 2, and Model 3 on the total portfolio.

	$\hat{U}$	$\hat{\sigma}_{\hat{U}}$	95% conf. int. (analytical)	95% conf. int. (bootstrap)
Model 1	0.8119	0.0063	[0.7996,0.8242]	[0.7999,0.8248]
Model 2	0.7791	0.0070	[0.7654,0.7928]	[0.7662,0.7933]
Model 3	0.8081	0.0063	[0.7958,0.8205]	[0.7958,0.8208]
	$\hat{U}$	$\hat{\sigma}_{\hat{U}}$	99% conf. int. (analytical)	99% conf. int. (bootstrap)
Model 1	0.8119	0.0063	[0.7959,0.8281]	[0.7962,0.8280]
Model 2	0.7791	0.0070	[0.7611,0.7969]	[0.7614,0.7974]
Model 3	0.8081	0.0063	[0.7919,0.8241]	[0.7917,0.8244]

[5] All bootstrapping results in this article were obtained by carrying out 5,000 simulation runs.

Table 3
Confidence intervals for Model 1, Model 2, and Model 3 for four subportfolios

This table shows the results for $\hat{U}$, $\sigma_{\hat{U}}$, 95%, and 99% confidence intervals (derived by asymptotic normality and bootstrapping) for Model 1, Model 2, and Model 3 on the four subportfolios with 100 defaulters and 400 non-defaulters, with 50 defaulters and 450 non-defaulters, 20 defaulters and 480 non-defaulters, and 10 defaulters and 490 non-defaulters.

a) 400-100	$\hat{U}$	$\hat{\sigma}_{\hat{U}}$	95% conf. int. (analytical)	95% conf. int. (bootstrap)
Model 1	0.8375	0.0204	[0.7976,0.8774]	[0.7977,0.8754]
Model 2	0.8206	0.0214	[0.7787,0.8626]	[0.7772,0.8620]
Model 3	0.8381	0.0203	[0.7984,0.8778]	[0.7963,0.8763]
	$\hat{U}$	$\hat{\sigma}_{\hat{U}}$	99% conf. int. (analytical)	99% conf. int. (bootstrap)
Model 1	0.8375	0.0204	[0.7850,0.8900]	[0.7826,0.8865]
Model 2	0.8206	0.0214	[0.7655,0.8757]	[0.7620,0.8737]
Model 3	0.8381	0.0203	[0.7859,0.8903]	[0.7800,0.8880]

b) 450-50	$\hat{U}$	$\hat{\sigma}_{\hat{U}}$	95% conf. int. (analytical)	95% conf. int. (bootstrap)
Model 1	0.8133	0.0227	[0.7689,0.8578]	[0.7660,0.8562]
Model 2	0.7800	0.0282	[0.7247,0.8353]	[0.7231,0.8325]
Model 3	0.8133	0.0227	[0.7689,0.8578]	[0.7681,0.8557]
	$\hat{U}$	$\hat{\sigma}_{\hat{U}}$	99% conf. int. (analytical)	99% conf. int. (bootstrap)
Model 1	0.8133	0.0227	[0.7549,0.8718]	[0.7522,0.8698]
Model 2	0.7800	0.0282	[0.7073,0.8527]	[0.7062,0.8503]
Model 3	0.8133	0.0227	[0.7550,0.8717]	[0.7516,0.8703]

c) 480-20	$\hat{U}$	$\hat{\sigma}_{\hat{U}}$	95% conf. int. (analytical)	95% conf. int. (bootstrap)
Model 1	0.8594	0.0377	[0.7855,0.9333]	[0.7804,0.9229]
Model 2	0.8281	0.0456	[0.7388,0.9175]	[0.7334,0.9066]
Model 3	0.8516	0.0382	[0.7766,0.9265]	[0.7742,0.9155]
	$\hat{U}$	$\hat{\sigma}_{\hat{U}}$	99% conf. int. (analytical)	99% conf. int. (bootstrap)
Model 1	0.8594	0.0377	[0.7623,0.9565]	[0.7455,0.9340]
Model 2	0.8281	0.0456	[0.7107,0.9456]	[0.7049,0.9285]
Model 3	0.8516	0.0382	[0.7531,0.9501]	[0.7389,0.9313]

d) 490-10	$\hat{U}$	$\hat{\sigma}_{\hat{U}}$	95% conf. int. (analytical)	95% conf. int. (bootstrap)
Model 1	0.8724	0.0620	[0.7510,0.9939]	[0.7377,0.9666]
Model 2	0.8673	0.0534	[0.7626,0.9721]	[0.7550,0.9500]
Model 3	0.8677	0.0616	[0.7466,0.9881]	[0.7395,0.9616]
	$\hat{U}$	$\hat{\sigma}_{\hat{U}}$	99% conf. int. (analytical)	99% conf. int. (bootstrap)
Model 1	0.8724	0.0620	[0.7128,1.0000]	[0.6944,0.9749]
Model 2	0.8673	0.0534	[0.7297,1.0000]	[0.7112,0.9652]
Model 3	0.8677	0.0616	[0.7086,1.0000]	[0.6863,0.9738]

IV.3. Comparison of AUC for two Different Rating Systems

In this part of the article we apply the test from Section III.1.5 on the difference of the area below the ROC curve to two rating models. We carry out pairwise comparisons of our three rating models (22a), (22b), and (22c) on the total validation sample from 1994 – 1999. The rating models (22a) and (22c) differ only by one accounting ratio. From Table 2 we see that their AUC has almost the same value and that the confidence intervals for AUC are overlapping on a very large range. On a first glance one might conclude that both rating models are of similar quality. In Table 4 we report the value of the test statistic (16), the corresponding p-value, and the correlation coefficient between the areas below the ROC curve for all pairwise comparisons of the three rating models.

Table 4
Results of the test of the difference of the areas below the ROC curve for pairwise comparison of Model 1, Model 2, and Model 3, validated on the total portfolio

In this table, we report the results of the test of Section III.1.5 for the total portfolio. We report the value of the test statistic T, the p-value, and the correlation ρ of the Mann-Whitney test statistics of the two rating models that are compared. We find that the differences between all rating methods are highly significant.

Models	T	p-value	ρ
1 & 2	55.93	$<$0.0001	0.79
1 & 3	11.58	0.0007	0.98
2 & 3	39.98	$<$0.0001	0.39

We find that Model 1 and Model 2 are different with high significance. The same is true for Model 2 and Model 3. Surprisingly the p-value of the test on the difference of Model 1 and Model 3 is only 0.0007. Therefore, both models are also different with high significance. The reason is the high correlation of 0.98. We give an intuitive explanation of this result. If we carried out bootstrapping both models would yield similar values for AUC in all simulations. However, due to the high correlation, the value of Model 1 would be in almost all cases higher than the value of Model 3. Therefore, Model 1 is superior to Model 3 with high significance.

If we carry out the same analysis for the sample portfolio with 500 obligors that contains 100 defaulters the picture is different. None of the pairwise comparisons of the three rating models leads to a significant difference. The detailed results are given in Table 5.

Table 5

Results of the test for the difference of the areas below the ROC curve for pairwise comparison of Model 1, Model 2, and Model 3 validated on the 100-400 portfolio

In this table, we report the results of the test of Section III.1.5 for the portfolio with 500 obligors that contains 100 defaults. We report the value of the test statistic T, the p-value, and the correlation ρ of the Mann-Whitney test statistics of the two rating models that are compared. We find on these small validation samples that no difference of any pair of rating models is statistically significant.

Models	T	p-value	ρ
1 & 2	1.40	0.2367	0.77
1 & 3	0.03	0.8648	0.98
2 & 3	1.44	0.2296	0.76

V. Conclusion

We have introduced a method to improve the discriminative power of accounting ratios by replacing them with their corresponding likelihood ratios. Furthermore, we have analysed statistical properties of the CAP and the ROC. By demonstrating the correspondence of the area AUC below the ROC curve and the Accuracy Ratio, we have shown that these summary statistics of the CAP and the ROC are equivalent. Furthermore this result enables us to use a simple analytical method, based on Bamber (1975), to obtain confidence intervals for these statistics. Additionally, by means of a methodology introduced by DeLong, DeLong, and Clarke-Pearson (1988), we are able to compare these summary statistics for two different rating methods being validated on the same data set. Examples with real data demonstrated that these methods are reliable even for small portfolios.

A. Proof of $\mathbf{AR = 2\,AUC - 1}$

Using our notation, we find for the area below the ROC curve AUC and the area below the CAP $a_R + 0.5$

$$\begin{aligned} AUC &= \sum_{i=1}^{k} \frac{1}{2}\,(CD_D^i + CD_D^{i-1})\,(CD_{ND}^i - CD_{ND}^{i-1}), && (23) \\ a_R + 0.5 &= \sum_{i=1}^{k} \frac{1}{2}\,(CD_D^i + CD_D^{i-1})\,(CD_T^i - CD_T^{i-1}). && (24) \end{aligned}$$

For a_P a simple calculation yields

$$a_P = \frac{1}{2}\,(1-\pi), \tag{25}$$

where π is the a-priori default probability of all debtors. To proof the desired relation, we start with (24).

$$\begin{aligned} a_R + 0.5 &= \sum_{i=1}^{k} \frac{1}{2}\,(CD_D^i + CD_D^{i-1})\,(CD_T^i - CD_T^{i-1}) \\ &= \sum_{i=1}^{k} \frac{1}{2}\,(CD_D^i + CD_D^{i-1})\,\big(\pi\,(CD_D^i - CD_D^{i-1}) + (1-\pi)\,(CD_{ND}^i - CD_{ND}^{i-1})\big) \\ &= (1-\pi)\sum_{i=1}^{k} \frac{1}{2}\,(CD_D^i + CD_D^{i-1})\,(CD_{ND}^i - CD_{ND}^{i-1}) \\ &+ \pi\sum_{i=1}^{k} \frac{1}{2}\,(CD_D^i + CD_D^{i-1})\,(CD_D^i - CD_D^{i-1}) \\ &= (1-\pi)\,AUC + \frac{1}{2}\,\pi\sum_{i=1}^{k}\Big(\big(CD_D^i\big)^2 - \big(CD_{ND}^{i-1}\big)^2\Big) \\ &= (1-\pi)\,AUC + \frac{1}{2}\,\pi && (26) \end{aligned}$$

Taking (2), (25), and (26) together, we obtain

$$AR = \frac{a_R}{a_P} = \frac{(1-\pi)\,(AUC - \frac{1}{2})}{\frac{1}{2}\,(1-\pi)} = 2\,AUC - 1. \tag{27}$$

References

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, 2000, Supervisory Risk Assessment and Early Warning Systems, Bank for International Settlements.

Basel Committee on Banking Supervision, 2001, The Internal Ratings-Based Approach, Bank for International Settlements.

DeLong, E., D. DeLong, and D. Clarke-Pearson, 1988, Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach, *Biometrics* 44, 837–845.

Efron, B., and R. J. Tibshirani, 1998, *An Introduction to the Bootstrap*. (Chapman & Hall Boca Raton).

Hanley, A., and B. McNeil, 1982, The Meaning and Use of the Area Under a Receiver Operating Characteristics (ROC) Curve, *Diagnostic Radiology* 143, 29–36.

Lee, W. C., 1999, Probabilistic Analysis of Global Performances of Diagnostic Tests: Interpreting the Lorenz Curve-based Summary Measures, *Statistics in Medicine* 18, 455–471.

Mann, H., and D. Whitney, 1947, On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other, *Annals of Mathematical Statistics* 18, 50–60.

Sobehart, J., and S. Keenan, 2001, Measuring Default Accurately, *Credit Risk Special Report, Risk* 14, 31–33.

Sobehart, J., S. Keenan, and R. Stein, 2000, Benchmarking Quantitative Default Risk Models: A Validation Methodology, Moodys Investors Service.

Swets, J., 1988, Measuring the Accuracy of Diagnostic Systems, *Science* 240, 1285–1293.

Tasche, D., 2002, Remarks on the Monotonicity of Default Probabilities, Working Paper.