THE DISCRIMINATIVE POWER OF RATING FUNCTIONS

Claudia Beinert[*], Peter Reichling[**], Bodo Vogt[***]

Abstract

We discuss different measures of the discriminative power of rating functions to examine the forecast strength of ratings. Besides the measures discussed in the literature, we also consider stochastic dominance to evaluate rating functions. We apply these criteria to compare empirically the ratings of Standard & Poor's and Moody's Investors Service.

Key Words: Credit risk, rating accuracy, discriminative power.
JEL classification: C52, G21, G33.

1. Introduction

According to the Basel II accord, a bank can choose between the implementation of an internal ratings-based (IRB) approach and the standard approach based on ratings provided by external agencies in order to determine the minimum capital requirement for its credit risk. A goal of every rating is to estimate the specific default probability of a debtor. Within rating systems debtors with similar estimated default probabilities are assigned to the same rating class. Assessing creditworthiness with a reliable forecast represents a central success factor in a bank's credit business.

The validation of rating systems has become a prevailing topic since the financial supervisory authorities began to accept applications from banks for the IRB approach approval. The qualifying examinations require the validation of rating systems before approval. Recently the adequacy of measures of the discriminative power to validate rating functions has been discussed in the literature[1]. These measures refer to methods which were developed 30 years ago and are used in medicine, weather forecasting, and signal detection theory. In addition to the existing literature on measures of the discriminate power, we show that stochastic dominance can also be used to evaluate rating functions.

Our paper is organized as follows: Section 2 presents the measures of discriminative power. The comparison of rating functions using stochastic dominance will be a main point here. In Section 3 we empirically analyze the ratings of Standard & Poor's and Moody's Investors Service. We conclude with a short summary in Section 4.

2. Measures of Discriminative Power

If a rating function exhibits a high discriminative power, it can differentiate between debtors with high and low creditworthiness. The measures to evaluate the discriminative power discussed in this section are: contingency table, receiver operating characteristic, cumulative accuracy profile, stochastic tendency, and stochastic dominance.

[*] Otto-von-Guericke-University Magdeburg, Germany.
[**] Otto-von-Guericke-University Magdeburg, Germany.
[***] Otto-von-Guericke-University Magdeburg, Germany.

[1] See, e.g., Sobehart, Keenen, Stein (2000), Sobehart, Keenan (2001), Hayden (2002, pp. 78-95), Engelmann, Hayden, Tasche (2003) and Hamerle, Rauhmeier, Roesch (2003).

2.1. Overview of Common Measures

At first, we consider a rating model with the two possible ratings 'default' and 'non-default' to evaluate the rating function. Here the originator of the ratings assigns one of the two possible outcomes to every debtor in point in time $t = 0$ and then observes the realizations in $t = 1$ (see Table 1).

Table 1

Contingency Table

		Observation in $t = 1$	
		Default	Non-default
Forecast in $t = 0$	Default	*A*	*B*
	Non-default	*C*	*D*

There is a series of ratios to evaluate the discriminative power of the underlying rating function. Ratios commonly used here are the hit rate HR and the false alarm rate FAR:

$$\text{HR} \equiv \frac{A}{A+C} \quad \text{and} \quad \text{FAR} \equiv \frac{B}{B+D}. \tag{1}$$

The contingency table is not only applicable to models which simply allocate the ratings 'default' or 'non-default'. It can also be extended to any number of rating classes. For models with more than two possible ratings it is necessary to define a cut-off point (CoP). A cut-off point represents the limit beyond which a debtor is rated as at risk of default. In a multi-rating class model the cumulative hit rate HR_s for a cut-off point CoP_s results as the sum of the single hit rates of the worst rating classes 1,...,*s*. The cumulative false alarm rate FAR_s is calculated analogously.

The contingency table only allows assessing the ability of a rating function to separate with regard to the chosen cut-off point. The total discriminative power of a rating function between debtors with strong and weak creditworthiness, aggregated over all cut-off points, is shown by the receiver operating characteristic.

2.1.1. Receiver Operating Characteristic and Area under Curve

The receiver operating characteristic (ROC) was developed in the fifties to test the strength of noise-affected radio signals. The ROC curve is generated by plotting the hit rates against the false alarm rates for all possible cut-off points. Figure 1 shows the ROC curve[1]. Cut-off points are indicated by dots there.

The area below the ROC curve (bold line in Figure 1) is called the area under curve (AUC). This area evaluates the discriminative power by comparison with the area below the perfect rating function (thin line). An AUC value of 50% corresponds to a random experiment to forecast insolvency (dotted line).

The ability of a rating function to assign the worst possible ratings to debtors becoming insolvent and the best ones to those remaining solvent is not evaluated by the area under curve. Within the ROC framework it is only important that debtors with weak creditworthiness receive worse ratings than those with strong creditworthiness.

[1] See Sobehart, Keenan (2001) and Engelmann, Hayden, Tasche (2003).

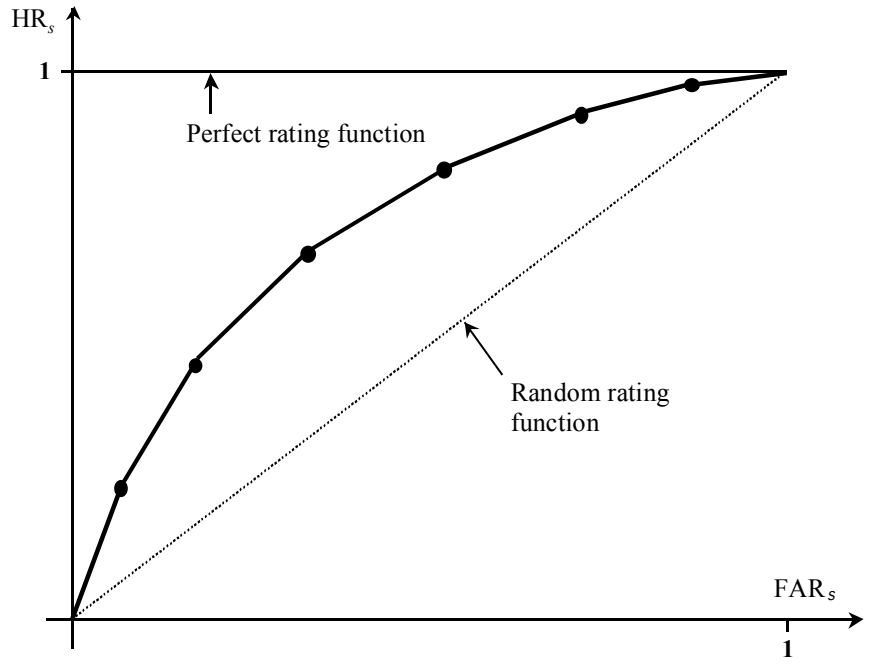

Fig. 1. Receiver Operating Characteristic

2.1.2. Cumulative Accuracy Profile and Accuracy Ratio

The cumulative accuracy profile (CAP) arises in accordance with the Lorenz curve. It measures along the lines of the receiver operating characteristic the ability of a rating function to position debtors with low creditworthiness in worse rating classes than those with high creditworthiness.

The CAP curve is generated by sorting all debtors according to their ratings. Subsequently, it is calculated which fraction of debtors with the worst ratings WR_s exhibits which hit rates of insolvencies. The perfect CAP curve (thin line in Figure 2) is determined by the default rate DR of the portfolio.

With the cumulative accuracy profile the accuracy ratio AR, which is calculated in accordance with the standardized Gini coefficient of the Lorenz curve, serves to evaluate the rating function. This ratio determines the relationship between two areas, the first being the area between the line of the observed rating function (bold line in Figure 2) and the line of the random rating function (dotted line) and the second being the area between the perfect rating function (thin line) and the random rating function[1].

[1] See Sobehart, Keenan, Stein (2000) and Engelmann, Hayden, Tasche (2003).

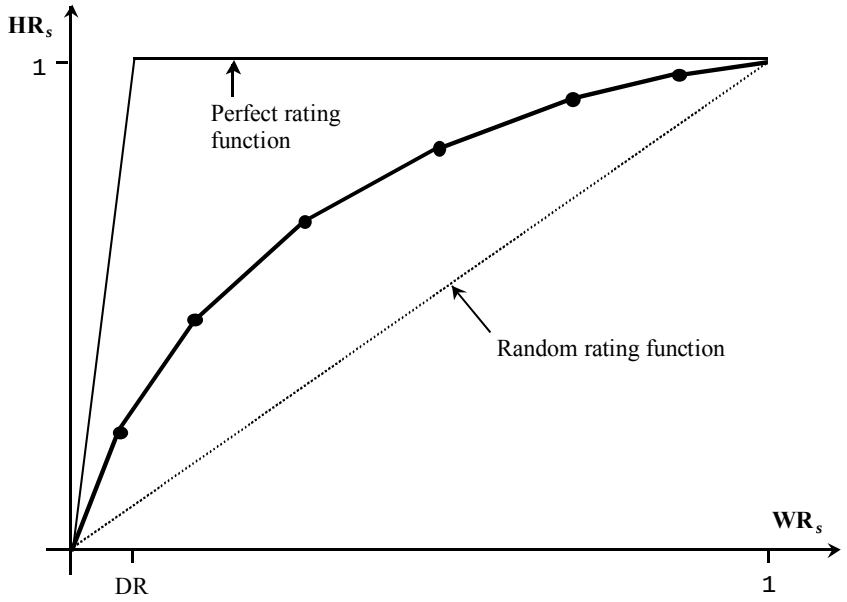

Fig. 2. Cumulative Accuracy Profile

In contrast to the Gini coefficient the accuracy ratio can become negative. Accuracy ratio and area under curve can be transferred into each other as follows[1]:

$$\mathrm{AR} = 2 \cdot \mathrm{AUC} - 1. \tag{2}$$

2.1.3. Stochastic Tendency

Stochastic tendency can also be used in order to assess whether the hit rate and false alarm rate distributions differ from each other. In our discrete framework the false alarm rate distribution tends to be stochastically larger than the hit rate distribution, if the following applies:

$$\mathrm{RE} \equiv \sum_{s} \frac{\mathrm{HR}_s + \mathrm{HR}_{s-1}}{2} \cdot \left(\mathrm{FAR}_s - \mathrm{FAR}_{s-1}\right) > \frac{1}{2} \quad \text{where} \quad \mathrm{HR}_0 = \mathrm{FAR}_0 \equiv 0. \tag{3}$$

With a relative effect RE above 0.5, companies remaining solvent in the tendency were placed into good rating classes and insolvent debtors tended to be positioned into bad rating classes. Relative effect and area under curve, both add up the products of false alarm rate per rating class and cumulative hit rate. Therefore, both concepts compute the area under the ROC curve.

Let S^{d} and S^{nd} denote the rating of debtors who have become insolvent and remained solvent, respectively. Then, in a probabilistic interpretation, the hit rates and the false alarm rates are:

$$\mathrm{HR}_s = \mathrm{Prob}\,(S^{\mathrm{d}} \le s) \quad \text{and} \quad \mathrm{FAR}_s = \mathrm{Prob}\,(S^{\mathrm{nd}} \le s)\,. \tag{4}$$

In case of independent ratings, this formulation gives the following interpretation of the relative effect and the area under curve, respectively[2]:

[1] See Engelmann, Hayden, Tasche (2003). Note, that these measures with given ratings depend, however, on the default rate of the credit portfolio. See Hamerle, Raumeier, Roesch (2003), and Sobehart, Keenan (2004).
[2] See (Bamber, 1975).

$$\mathrm{RE} = \mathrm{AUC} = \mathrm{Prob}(S^{\mathrm{d}} < S^{\mathrm{nd}}) + \frac{1}{2} \cdot \mathrm{Prob}(S^{\mathrm{d}} = S^{\mathrm{nd}}). \quad (5)$$

The area under curve corresponds to the probability that an insolvent company received a worse rating than a debtor who remained solvent (where the probability of equal ratings is weighted by a half). Note, that the relative effect does not react to a transformation that preserves order[1]. Therefore, relative effect, area under curve, and accuracy ratio are appropriate for ordinal rating scores. Irrespective of whether the rating function provides scorings or default probabilities, these measures come to a consistent result as long as the estimated default probability increases with an inferior rating score.

As an intermediate result, we get that the area under the ROC curve equals the relative effect of stochastic tendency. Both terms are in a linear relationship with the accuracy ratio. Therefore, the measures of discriminative power discussed so far produce identical results.

2.2. Stochastic Dominance

First-order stochastic dominance examines the cumulative probabilities of two random variables. In our framework stochastic dominance can be applied to the hit rate and false alarm rate distributions of one rating function or to the hit rate distributions of two rating functions.

In case of one rating function, we have first-order stochastic dominance of the false alarm rate distribution over the hit rate distribution if the following applies:

$$\mathrm{FAR}_s \leq \mathrm{HR}_s \ \forall_s \quad (6)$$

and this inequality is strictly fulfilled for at least one rating class s. Then the observed ROC curve lies above the diagonal in Figure 1, resulting in an area under curve above 50%.

For a comparison of two rating functions we consider the hit rate $\mathrm{HR}(s) = \mathrm{HR}_s$ and false alarm rate $\mathrm{FAR}(s) = \mathrm{FAR}_s$. Then the ROC curve is given by the vector $(\mathrm{HR}(s), \mathrm{FAR}(s))$. In the context of first-order stochastic dominance, rating function R dominates rating function T in case of identical false alarm rates if the following inequality holds and is strictly fulfilled for at least one rating class s:

$$\mathrm{HR}_T(s) \leq \mathrm{HR}_R(s) \ \forall s. \quad (7)$$

Since the false alarm rate is a monotone transformation of the rating score, inequality (7) holds for all false alarm rates, too. Therefore, in case of identical false alarm rates the ROC curve of rating function R runs above the curve of rating function T. Consequently, the AUC value of rating function R turns out to be higher than the value of rating function T.

If first-order stochastic dominance is not given, then it will be appropriate to try to apply the second-order stochastic dominance criterion. We have second order stochastic dominance of the false alarm rate distribution over the hit rate distribution, if the following inequality holds and is strictly fulfilled for at least one rating class i:

$$\sum_{s=1}^{i} \mathrm{FAR}(s) \leq \sum_{s=1}^{i} \mathrm{HR}(s) \ \forall i. \quad (8)$$

From this dominance it does not necessarily follow that the area under curve lies above 50%, as proved by the example in Table 2. The example is constructed in a way that although there is second-order stochastic dominance of the false alarm rate distribution over the hit rate distribution an AUC value of 47% is calculated.

[1] See (Bamber, 1975).

Table 2

Second-order Stochastic Dominance for Hit Rates

Score	Issuers	Defaults	HR_s	FAR_s	$\sum_{s=1}^{i} HR_s$	$\sum_{s=1}^{i} FAR_s$
1	160	100	33 %	20 %	0.33	0.20
2	40	30	43 %	23 %	0.77	0.43
3	200	30	53 %	80 %	1.30	1.23
4	200	140	100 %	100 %	2.30	2.23

From this example we see that the common measures of discriminative power focus on ratios of hit rates and false alarm rates. From an economic perspective this must not lead to the same result as analyzing the probability of default in every rating class. For example, without taking different credit amounts, opportunity costs, and credit spreads into consideration, using the rating function from Table 2 a credit rationing strategy would avoid the default cost of 100 and 130 debtors, if credit applications from clients with scores 1 and 2, respectively, are rejected. When applying a random rating function with identical false alarm rates, only 60 or rather 70 applications of clients becoming bankrupt are refused.

In the context of second-order stochastic dominance we again compare rating functions *R* and *T* with identical false alarm rates. Then rating function *R* dominates rating function *T* if the following inequality holds and is strictly fulfilled for at least one rating class *i*:

$$\sum_{s=1}^{i} \mathrm{HR}_T(s) \leq \sum_{s=1}^{i} \mathrm{HR}_R(s) \quad \forall i. \tag{9}$$

If we compare the rating function from Table 2 to a random rating function with the same false alarm rates, the random rating function will show a higher AUC value of 0.5. However, the rating function from Table 2 is preferred in the sense of second-order stochastic dominance. Our examples with identical false alarm rates show that the concepts of area under curve and accuracy ratio, respectively, may contradict with the criterion of second-order stochastic dominance.

3. Empirical Results

Our empirical examination aims to clarify two questions: Firstly, what discriminative power do the ratings of the agencies Standard & Poor's (S & P) and Moody's Investors Service (Moody's) exhibit? As these companies are the most well-known rating agencies worldwide, we expect that they possess high and significant AUC and AR values. Secondly, does the rating function of one agency dominate the function of the other? In order to answer these questions we analyze the period from 1982 to 2001. The number of issuers as well as the default rates for this period and the seven classes of both rating scales of S & P (AAA, AA ... CCC) and Moody's (Aaa, Aa ... C) were available. Reports from these agencies constituted the database[1].

The Moody's data includes issuers of long-term bonds from the sectors industry, transport, utilities, and financial institutions. Issuers of structured financial products and public issuers are not included. The S & P data also come from issuers of long-term bonds, mainly from the sectors industry, utilities, financial institutions, and insurance. Again issuers of structured financial products, public issuers, and issuers whose ratings were exclusively based on public information were ex-

[1] See Moody's Investors Service (2002) and Standard & Poor's (2002). Moody's numbers of issuers were provided by Moody's KMV.

cluded. Figure 3 demonstrates the AUC values for our sample. All calculated AUC values are significantly different from 0.5 using the Mann-Whitney statistics (all p-values are below 0.1%).

S & P's and Moody's rating functions show discriminative power for our sample. The average AUC value of the Moody's sample is 90.9%, with S & P this value is 90.5%. According to a sign test Moody's AUC values are significantly higher than those of S & P (p value below 6%). All the ascertained ROC curves run above the diagonal in Figure 1. In every case considered here, we find first-order stochastic dominance of the false alarm rate distribution over the hit rate distribution.

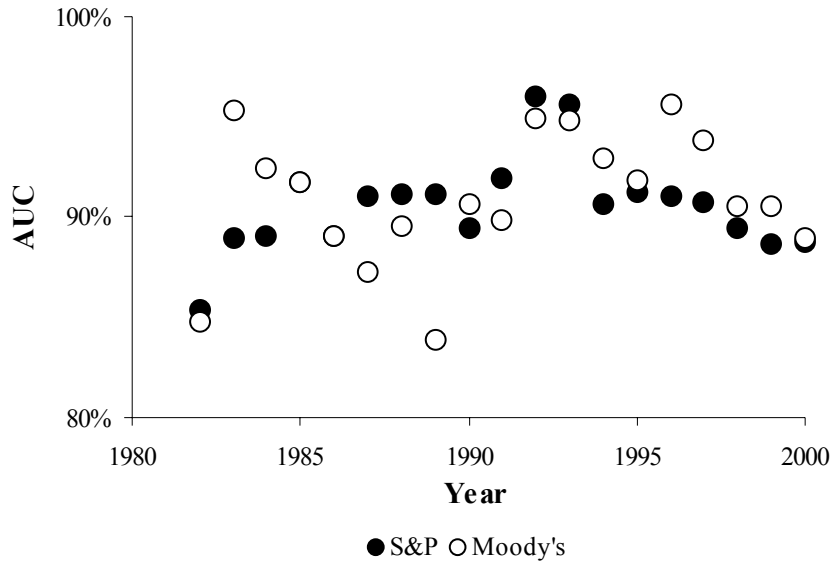

Fig. 3. AUC Values for S & P and Moody's Ratings (1982-2001)

A different result emerges when only the hit rate distributions are compared. In 17 of the 20 years examined a ranking of the rating functions is possible according to second-order stochastic dominance with no significant advantage of one rating agency. In the years 1985, 1986, and 1990 we observe different rankings between stochastic dominance of the hit rates and the AUC criterion. Since the differences in false alarm rates seem to be small, this supports our theoretical argument from section 2.2. Table 3 gives an overview.

Table 3

Second-order Stochastic Dominance and AUC Dominance

Year	2nd-order hit rate stochastic dominance	AUC dominance	FAR difference
1982	none	S & P	4.1 %
1983	Moody's	Moody's	1.5 %
1984	none	Moody's	2.7 %
1985	S & P	Moody's	2.1 %
1986	S & P	Moody's	2.9 %
1987	S & P	S & P	3.6 %
1988	S & P	S & P	3.3 %
1989	S & P	S & P	2.8 %
1990	S & P	Moody's	2.2 %

Table 3 (continuous)

Year	2nd-order hit rate stochastic dominance	AUC dominance	FAR difference
1991	S & P	S & P	1.8 %
1992	S & P	S & P	2.4 %
1993	S & P	S & P	2.4 %
1994	none	Moody's	2.5 %
1995	Moody's	Moody's	2.6 %
1996	Moody's	Moody's	2.4 %
1997	Moody's	Moody's	2.6 %
1998	Moody's	Moody's	2.1 %
1999	Moody's	Moody's	1.7 %
2000	Moody's	Moody's	2.7 %
2001	Moody's	Moody's	3.4 %

$$\text{FAR difference} \equiv \sqrt{\frac{1}{6}\sum_{s=1}^{6}(\text{FAR}_s^{\text{S\&P}} - \text{FAR}_s^{\text{Moody's}})^2}\ .$$

4. Conclusion

Besides the common measures to evaluate the discriminative power of rating functions, i.e. area under curve, accuracy ratio, and relative effect, we introduced stochastic dominance into this topic. Although first-order stochastic dominance leads to the same results as the common measures, second order stochastic dominance turns out to produce possibly different results.

Our empirical analysis provides the following result. The rating functions of Standard & Poor's and Moody's Investors Service possess discriminative power, as expected, for the years 1982 to 2001. However, a persistent dominance of one rating agency cannot be observed here.

References

1. Bamber, D. The Area above the Ordinal Dominance Graph and the Area below the Receiver Operating Characteristic Graph // Journal of Mathematical Psychology, 1975. – №12. – pp. 387-415.
2. Engelmann, B., E. Hayden, D. Tasche. Testing Rating Accuracy // Risk, 2003. – №16. – pp. 82-86.
3. Hamerle, A., R. Rauhmeier, D. Roesch. Uses and Misuses of Measures for Credit Rating Accuracy // Working Paper, University of Regensburg, 2003.
4. Hayden, E. Modeling an Accounting-Based Rating System for Austrian Firms // Dissertation Thesis, University of Vienna, 2002.
5. Moody's Investors Service. Default & Recovery Rates of Corporate Bond Issuers // Special Comment, 2002 – February.
6. Sobehart, J., S. Keenan. Measuring Default Accurately // Risk, 2001. – №14.– pp. 31-33.
7. Sobehart, J., S. Keenan. The Score for Credit // Risk, 2004. – №17.– pp. 54-58.
8. Sobehart, J., S. Keenan, R. Stein. Validation Methodologies for Default Risk Models // Credit, 2000. – №5. – pp. 51-56.
9. Standard & Poor's. Ratings Performance 2001 // Special Report, 2002 – February.