Comparing the accuracy of default predictions in the rating industry: The case of Moody's vs. S&P *

by

Walter Krämer
Fachbereich Statistik, Universität Dortmund
D-44221 Dortmund, Germany

and

André Güttler
Fakultät für Wirtschaftswissenschaften, Universität Frankfurt
D-60325 Frankfurt, Germany

Version September 2003

Abstract

We consider 1927 borrowers from 54 countries who had a credit rating by both Moody's and S&P as of the end of 1998, and their subsequent default history up to the end of 2002. Viewing bond ratings as predicted probabilities of default, we show that it is unlikely that both agencies are well calibrated, and that the ranking of the agencies depends crucially on the way in which probability predictions are compared.

Key words: credit rating, probability forecasts, calibration.

*Research supported by Deutsche Forschungsgemeinschaft (DFG) under SFB 475. We are grateful to our colleagues at KMW (Mark Wahrenburg, Jan-Pieter Krahnen, Martin Weber) and to Atanasios Mitropoulos for helpful criticism and comments.

1 Introduction

The evaluation of the quality, in whichever way defined, of default predictions in the credit industry has received considerable attention recently (Carey 2002, Tabakis and Vinci 2002, Engelmann et al. 2003 and many others). In particular, there is a growing interest in comparing the accuracy of competing rating agencies, or to rate the raters, so to speak. Also, in the wake of Basel II, there will be a rapidly increasing number of rating producers, in addition to the established rating agencies, and an increasing number of borrowers who are rated by at least two of them, so it is natural to ask: which rater rates best?

This question can be answered in a variety of ways. The most popular method is based on the accuracy ratio, i.e. on how successful a rating system is in concentrating the defaults in the "bad" grades, or its equivalent, the area under the ROC-curve (see section 4 or Sobehard and Keenan 2001 for a convenient introduction). Below, we also follow a different approach by viewing borrower ratings as predicted probabilities of default, and by comparing the accuracy of these predictions across rating agencies. In doing so, we borrow heavily from mathematical statistics, where the evaluation of probability forecasts has a long and distinguished history (see e.g. Dawid 1982, DeGroot and Fienberg 1983, Vardeman and Meeden 1983, DeGroot and Eriksson 1985). So far, this methodology has mostly been applied to weather forecasts (i.e. forecasts of the probability of rain) but it can easily be extended to default predictions in the rating industry.

Our analysis is based on 1927 borrowers, mostly industrial firms and financial institution from the US (68 % of all borrowers), who had a credit rating by both Moody's and $S\&P$ as of Dec. 31, 1998. We followed these firms up to the end of 2002 and recorded all defaults. The data were obtained from Bloomberg and are available from the authors on request.

Below we apply various orderings of probability forecasts to this data set. As these orderings are scattered in the statistics literature, we start by collecting and briefly reviewing them in section 2. Section 3 is concerned with mapping rating grades to probabilities of default, and section 4 compares the accuracy of the ratings, both in terms of partial orderings and in terms of various scalar measures of performance which have been suggested in the literature. Section 5 checks for statistical significance of the observed differences in performance, and a brief discussion of the shortcomings of our analysis in section 6 concludes.

2 Partial orderings of probability forecasts

Let $0 = a_1 < a_2 < \ldots < a_k = 1$ be k predicted probabilities of default. In practice, k varies from 6 to about 20. The US-based Loan Pricing Corporation has $k = 10$. The rating agencies which concern us in the present paper, Moody's and $S\&P$, both have scales with $k = 21$. For ease of comparability with these established agencies, most commercial banks also employ scales with $k = 20$ in their post Basel II internal rating systems.

Below we take the mechanism employed for the predictions as given. Producing the predictions is a separate problem which has engendered an enormous literature, but will not concern us here. Modern methods, as surveyed in e.g. Crouhy et al. (2001) or Arminger et al. (1997), include logit and probit analysis, neural networks, or classification trees. Rather, our point of departure is the discrete bivariate probability function $r(\theta, a_j)$; $\theta = 1, 2$; $j = 1, ..., k$, resulting from some such method, whichever it may be, with $\theta = 1$ indicating default and $\theta = 0$ indicating non-default.

The following additional notation will be used:

$p(1) := \sum_j r(1, a_j)$ = overall relative frequency of default.

$p(0) := \sum_j r(0, a_j)$ = overall relative frequency of no default.

$q(a_j)$:= relative frequency with which default probability forecast a_j is made.

$p(1|a_j) := \frac{r(1,a_j)}{q(a_j)}$ = conditional relative frequency of default given probability forecast a_j.

$p(0|a_j) := \frac{r(0,a_j)}{q(a_j)}$ = conditional relative frequency of no default given probability forecast a_j.

$q(a_j|1) := \frac{r(1,a_j)}{p(1)}$ = conditional relative frequency of predicted default probability a_j given default.

$q(a_j|0) := \frac{r(0,a_j)}{p(0)}$ = conditional relative frequency of predicted default probability a_j given no default.

The problem is: given two forecasters A and B, characterized by their respective bivariate probability functions $r^A(\theta, a_j)$ and $r^B(\theta, a_j)$, which one is "better"?

One sensible requirement is that among borrowers with predicted default probability a_j, the relative percentage of defaults will be roughly equal to a_j. Formally:

$$a_j \overset{!}{=} p(1|a_j) = \frac{r(1, a_j)}{q(a_j)}$$

whenever $q(a_j) > 0$. Such forecasters are called "well calibrated" (Dawid 1982).

However, calibration, though desirable, is not sufficient for a useful forecast. For instance, a probability forecaster attaching default probability $p(1)$ to all borrowers is well calibrated but otherwise quite useless.

Let $r^A(\theta, a_j)$ and $r^B(\theta, a_j)$ be the joint probability functions of forecasters A and B, respectively, with a nondegenerate marginal distribution $p(\theta)$. We assume that this marginal distribution is the same for both forecasters, i.e. that both agencies rate the same set of borrowers. First, we confine ourselves to forecasters which are both well calibrated. Following DeGroot and Fienberg (1983), we say that A is more refined than B, in symbols: $A \geq_R B$, if there exists a $k \times k$ Markov matrix M (i.e. a matrix with nonnegative entries whose columns sum to unity) such that

$$q^B(a_i) = \sum_{j=1}^{k} M_{ij} q^A(a_j), \quad \text{and} \tag{1}$$

$$a_i q^B(a_i) = \sum_{j=1}^{k} M_{ij} a_j q^A(a_j), \; i = 1, \dots, k. \tag{2}$$

Equation (1) means that, given A's forecast a_j, an additional independent randomisation is applied according to the conditional distribution M_{ij} ($j = 1, ..., k$) which produces forecasts with the same probability function as that of B. Condition (2) ensures that the resulting forecast is again well calibrated.

Table 1, from Krämer (2003a), provides an example. Forecaster A attaches a default probability of 2 % to all borrowers. If the overall default probability is indeed 2 %, he is obviously well calibrated. Forecaster B is more refined; he attaches default probabilities 1 % and 3 %, respectively, to 50 % of all borrowers. We assume that he, too, is well calibrated. Likewise forecasters C and D with distributions across predicted default probabilities as given in the table.

— table 1 about here —

Obviously, B, C and D are more refined than A. Also, C and D are more refined than B: If all borrowers who receive a 0,5 % rating from C, and a randomly selected 50 % of those who receive a rating 1,5 %, are given a rating of 1 %, the rest a rating of 3 %, we obtain a new, well calibrated forecast with the same probabilistic properties as B's.

The same can be done with D: All borrowers with ratings 0,5 % and 1 %, and a randomly selected one-eleventh of borrowers rated 3 %, are given a new rating of 1 %, the rest a new rating or 3 %. Again, this yields a new, well calibrated forecast with the same probabilistic properties as B's.

On the other hand, C and D cannot be compared according to the refinement ordering. DeGroot and Fienberg (1983, Theorem 1) show that, for well calibrated forecasters A and B,

$$A \geq_R B \iff \sum_{i=1}^{j-1}(a_j - a_i)[q^A(a_i) - q^B(a_i)] \quad \geq \quad 0, \quad \text{j=1,...,k-1.} \tag{3}$$

and this condition is violated for C and D in our example. It is equivalent to the fact that the distribution $q^A(a_i)$ second-order stochastically dominates the distribution $q^B(a_i)$ (DeGroot and Eriksson 1985).

Vardeman and Meeden (1983) suggest to alternatively order probability forecasters according to the concentration of defaults in the "bad" grades. This will here be called the VM-default order. Formally:

$$A \geq_{VM(d)} B \quad :\iff \quad \sum_{i=1}^{j} q^A(a_i|1) \quad \leq \quad \sum_{i=1}^{j} q^B(a_i|1), \quad \text{j=1,...,k.} \tag{4}$$

Or to put this differently: A dominates B in the Vardeman-Meeden default ordering if its conditional distribution, given default, first-order stochastically dominates that of B.

The same can be done for the non-defaults. A is better than B in the VM-non-default sense if non-defaults are more frequent in the "good" grades. Formally:

$$A \geq_{VM(nd)} B \iff \sum_{i=1}^{j} q^A(a_i|0) \quad \geq \quad \sum_{i=1}^{j} q^B(a_i|0), \quad \text{j=1,...,k.} \tag{5}$$

Finally, A dominates B in the Vardeman-Meeden sense (in symbols $A \geq_{VM} B$) if both $A \geq_{VM(d)} B$ and $A \geq_{VM(nd)} B$.

A final criterion which is favoured in the banking industry (see e.g. Falkenstein et al. 2000) is based on joining the points

$$(0,0), \left(\sum_{i=0}^{j-1} q(a_{k-i}), \sum_{i=0}^{j-1} q(a_{k-i}|1) \right), \quad j = 1, ..., k \tag{6}$$

by straight lines. The resulting plot is variously called the power curve, the Lorenz curve, the Gini curve, or the cumulative accuracy profile, and a forecaster A is considered better than a forecaster B in this - the Gini-default-sense (formally: $A \geq_{G(d)} B$) - if A's Gini curve is nowhere below that of B.

Alternatively, one might consider the receiver-operating-characteristic curve (ROC-curve) defined by the points

$$(0,0), \left(\sum_{i=0}^{j-1} q(a_{k-i}|0), \sum_{i=0}^{j-1} q(a_{k-i}|1) \right), \quad j = 1, ..., k. \tag{7}$$

It is however easily seen (see e.g. Krämer 2002, Theorem 3) that two Gini curves intersect if and only if the respective ROC-curves intersect, so these orderings are equivalent. Moreover, the area under the ROC-curve, which is often used as a scalar criterion of goodness, is numerically identical to the well known accuracy ratio derived from the Gini-curve, so the ROC-curve does not produce any independent information.

Both the power and the ROC curve are invariant to monotone transformations of the predicted probabilities of default. They only require that rating grades are ordered in terms of probability of default and can therefore be applied in a wider context than the one discussed above.

3 Mapping rating grades to default probabilities

Next, we apply the orderings described above to real world default predictions. This is done in two stages. First, we take empirical relative frequencies as substitutes for unknown true probability functions and check the partial orderings from section 2. Next, we explicitly recognize the randomness in the empirical relative frequencies and test for the statistical significance of the differences observed.

Table 2 summarizes our data base. For each rating grade, it shows the number of debtors carrying this rating as of Dec. 31, 1998, and the number of defaults up to the end of 2002. There are 17 grades, with all debtors rated worse than B- lumped together into grade C.

— table 2 about here —

Both the refinement and the VM-orderings require that rating grades are converted into default probabilities. We first proceed under the assumptions that (i) the true probabilities of default, given the rating grade, are the same for both agencies (where the correspondence between grades is as in table 2), and (ii) that the observed differences in empirical relative frequencies are due to random noise. This assumption will be later on relaxed. In addition, to obtain larger samples, we disregard the + and - subdivisions and estimate the grade specific default probabilities by averaging the empirical relative frequencies from Moody's and $S\&P$.

Column 3 in table 3 gives the results. Columns 4 and 5 give the historical 4-year default frequencies as reported by the agencies themselves.

— table 3 about here —

The table shows that the 4-year default rates in our sample are somewhat higher than the historical ones reported by the agencies themselves. The main reason is that our horizon covers the years 2001 and 2002, which saw an exceptionally large number of defaults: 70 of the 209 defaults in our sample occurred in 2001 and 58 occured in 2002. On the other hand the default rates reported by the agencies are averages of 18 (S&P) or 30 (Moody's) four-year default horizons, covering various phases of the business cycle.

In addition to averaging over default horizons, some additional smoothing is sometimes applied across rating grades to eliminate any remaining random noise (see e.g. Tabakis and Vinci 2002, pp. 13 – 15 or Blum et al., 2003, pp. 21 – 26). One can for instance fit a logit or a probit curve to the observed default frequencies in order to obtain a smooth and increasing sequence of default probabilities. Such issues will not be touched upon in this paper, as we are mainly concerned with systematic differences in forecasting ability, not with short-run effects induced by random deviations from a long run performance standard.

4 Comparing the accuracy of default predictions

We start by checking the partial orderings from Section 2. Figure 1 shows the Moody's and $S\&P$ power curves, as derived from table 2, with + and - subdivisions lumped together. It is seen that the power curves intersect, so the rating agencies cannot be compared according to this criterion.

— figure 1 here —

To obtain a similar result for the VM-criteria, table 4 lists the respective distributions of class frequencies, given default and given no default. It shows that Moody's dominates S&P with respect to VM(d) and that S&P dominates Moody's with respect to VM(nd). This comes as no surprise in view of theorem 1 in Krämer (2002), which states that the VM-ordering implies the Gini-ordering. As Moody's and S&P cannot be compared according to the Gini-ordering, they cannot be compared to the VM-ordering a forteriori. The most one could hope for is comparability according to either VM(d) or VM(nd), but not according to both (in the sense that one dominates the other according to both criteria). This is exactly what we find.

— table 4 about here —

As to the refinement ordering, we have to check calibration first. Here we have the problem that the data are not consistent with the fact that both agencies are well calibrated, at least if the distribution $q(a_j)$ of borrowers across rating grades a_j from table 2 can be viewed as typical for the agencies. A necessary condition for calibration is that the overall predicted relative frequency of default $p(1)$ be the same for both agencies. Plugging the default probabilities a_j from table 3 into the general formula

$$p(1) = \sum_j r(1, a_j) = \sum_j p(1|a_j)q(a_j) = \sum_j a_j q(a_j) \tag{8}$$

shows that we obtain different results for Moody's and for S&P. This is so no matter which column of table 3 is used for the predicted default probabilities a_j. For instance, taking our own estimates from column 3 gives $P^{S\&P}(1) = 9,89\%$ and $P^M(1) = 11,80\%$. For other columns, discrepancies are even larger.

One way out of this dilemma is to acknowledge that the equivalence of the rating grades established in table 3 is not quite correct, i.e. that a rating of BBB by S&P implies a (slightly) different predicted probability of default than a rating of Baa2 by Moody's. This in turn implies that we have $k = 14$ rather than $k = 7$ predicted probabilities of default (taking table 3 as our point of departure), with the probabilities themselves given for instance by columns 4 and 5. Plugging these probabilities into formula (8) gives $p(1) = 8,45\%$ for Moody's and $p(1) = 8,02\%$ for S&P, so we still have the result that calibration for both agencies is inconsistent with the data.

However, if we identify realized default frequencies with predicted ones, both agencies are well calibrated by definition. It then makes sense to check whether one is more refined than the other. We call this the empirical refinement ordering. Table 5 gives the results. It shows that the integrals of the distribution functions intersect, so none of the agencies is in this sense more refined than the other.

— table 5 about here —

The non-comparability of the default predictions in terms of the empirical refinement ordering implies that different scalar measures of performance will rank the predictions differently. Most popular among these is the Brier score (Brier 1950), defined as

$$B = \frac{1}{n} \sum_{i=1}^{n} (p_i - \theta_i)^2, \tag{9}$$

where p_i is the predicted probability of default, and $\theta_i = 1$ in case of default and $\theta_i = 0$ in case of no default. It takes its optimum value of $B = 0$ when the only predicted probabilities of default are 0 and 1, and when predictions are always correct (= perfect foresight). It takes its worst value of $B = 1$ when the only predicted probabilities of default are 0 and 1, and when always the opposite of what has been predicted occurs.

If we attach to each borrower the default probability from table 3, column 4 (Moody's) and column 5 (S & P), we obtain

$$B^M = 0,0684, \quad B^{S\&P} = 0,0735. \tag{10}$$

If we attach to each borrower the default probability from table 3, column 3, we obtain

$$B^M = 0,0662, \quad B^{S\&P} = 0,0689, \tag{11}$$

and if we attach to each borrower the observed default rate of the class these borrower has been sorted into, we obtain

$$B^M = 0,0660, \quad B^{S\&P} = 0,0686. \tag{12}$$

As small values of the Brier score are "good", Moody's outperforms S&P according to this criterion. It also outperforms S&P according to the logarithmic score, defined as

$$L = \frac{1}{n}\sum_{i=1}^{n} \ell n(|p_i + \theta_i - 1|). \tag{13}$$

The logarithmic score is always negative, with closeness to zero signalling a good performance. For our data set, it takes the following values if default probabilities from table 3, column 4 (Moody's) and column 5 (S & P) are used:

$$L^M = -0,2135, \quad L^{S\&P} = -0,2260. \tag{14}$$

If we attach to each borrower the default probabilities from table 3, column 3, we obtain

$$L^M = -0,2068, \quad L^{S\&P} = -0,2141. \tag{15}$$

And if we attach to each borrower the observed default rates of the class the borrower has been sorted into, we obtain

$$L^M = -0,2005, \quad L^{S\&P} = -0,2056. \tag{16}$$

As large values of the logarithmic score are "good", Moody's outperform $S\&P$ also according to this criterion. They also outperform $S\&P$ according to the spherical score, defined as

$$S = \frac{1}{n}\sum_{i=1}^{n} \frac{|p_i + \theta_i - 1|}{\sqrt{p_i^2 + (1-p_i)^2}}. \tag{17}$$

This gives

$$\begin{aligned}
S^M &= 0,9025, \quad S^{S\&P} = 0,8963 \quad \text{(column 4 and 5)} \\
S^M &= 0,9048, \quad S^{S\&P} = 0,9015 \quad \text{(column 3)} \\
S^M &= 0,9051, \quad S^{S\&P} = 0,9019 \quad \text{(column 1 and 2)}
\end{aligned}$$

As the spherical rule is always positive, with large values signalling superior performance, Moody's wins here as well.

However, it is easy to find scores such that this ranking is reversed. This reversal is made possible by the noncomparability of Moody's and S&P in terms of the empirical refinement ordering. It is well known (see e.g. DeGroot and Eriksson 1985 for a review of this literature) that second order stochastic dominance of a distribution $q^B(a_i)$ by a distribution $q^A(a_i)$ is equivalent to the fact that

$$\sum_i g(a_i) q^A(a_i) \geq \sum_i g(a_i) q^B(a_i) \tag{18}$$

for all continuous convex functions g on the unit interval. On the other hand, it is also well known (see e.g. Winkler 1996; the basic theorem is due to Savage 1971) that, for well calibrated forecasters, all proper scoring rules

$S(p_1, \ldots, p_n; \theta_1, \ldots, \theta_n)$ depend on the p_i and θ_i only via the a_j's and can be written as

$$S(p_1, \ldots, p_n; \theta_1, \ldots, \theta_n) = \sum_{j=1}^{K} g(a_j) q(a_j) \tag{19}$$

with some strictly convex function g. For the Brier score, for instance, we have

$$g(p) = p(1-p). \tag{20}$$

If second order stochastic dominance fails, one can therefore always find two convex functions f and g (corresponding to two proper scoring rules S_f and S_g) such that the ranking of two forecasters is reversed.

An example is the asymmetric version L^* of the logarithmic score suggested by Winkler (1994). Setting $c = 0,001$ (Winkler's notation) and equating observed default rates to predicted ones, we obtain values of $L^{*,M} = 0,2446$ and $L^{*,S\&P} = 0,2457$, so $S\&P$ is slightly better now. For details, see Krämer (2003 b).

5 Statistical Significance

Next we explicitly acknowledge the randomness in our data and briefly comment on the statistical significance of the differences in performance which we have found. For instance, the accuracy ratios derived from figure 1 are 0,833 for Moody's (AR_M) and 0,819 for S & P (AR_S), so there is a slight but insignificant advantage for Moody's here. Adapting result on areas under the ROC-curve from medical statistics, Engelmann et al. (2003) show that the statistic

$$T = \frac{(AR_M - AR_S)^2}{\sigma^2_{AR_M} + \sigma^2_{AR_S} - 2\sigma_{AR_M, AR_S}} \tag{21}$$

is asymptotically chi-squared with one degree of freedom. Approximating the variances $\sigma^2_{AR_S}$ and $\sigma^2_{AR_M}$ and the covariance σ_{AR_S,AR_M} of the respective accuracy ratios AR_S and AR_M by bootstrapping produces a p-value of 0,087. Estimating these for the areas under the ROC-curve by sample values produces a slightly lower p-value of 0,076. None of these indicates a systematic difference.

The asymptotic null distribution of the T-statistic (21) should be applied with caution, however. It is derived on the analogy with similar significance tests for ROC-curves in medical statistics (see e.g. Hajian-Tilaki et al. 2002) and assumes two independent simple random samples from the bivariate distributions $[q^M(a_j|1), q^S(q_j|1)]$ and $[q^M(a_j|0), q^S(a_j|0)]$, respectively, with sample sizes n_1 and n_2 fixed in advance. None of these requirements is met in the credit rating context. If we consider the 1927 ratings from the present paper as a random sample from a hypothetical universe of potential ratings, then the sample sizes n_1 and n_2 are not fixed but random and perfectly negatively correlated. And more importantly, a sample of n observations from the bivariate distribution $r(\theta, a)$ will in practice never be simple as the observed θ's are known to be positively correlated in practice. As the observed values of θ and a for a given borrower are also highly correlated, this can then be shown to translate into correlation among draws from the conditional distributions $q(a|1)$ and $q(a|0)$, which are therefore not a simple random sample. As H_0 in our case is not rejected anyway, we do not investigate this issue any further here.

The same argument applies even more forcefully when assessing the significance of the difference of other scalar measures of performance. For the Brier score, it is easily seen (see e.g. Redelmeier et al. 1991) that the statistic

$$Z = \frac{\sum_{i=1}^{n}(\theta_i - \pi_i)(p_i^S - p_i^M)}{\sqrt{\sum_{i=1}^{n}\pi_i(1-\pi_i)(p_i^S - p_i^M)^2}} \tag{22}$$

where $\pi_i = (p_i^S + p_i^M)/2$, is asymptotically standard normal when population Brier scores are identical. In our sample, Z takes the value 3,78, which at first sight is highly significant. However, as in the case of the Engelmann et al. T-statistic, its limiting null distribution obtains only for simple random samples, which in the rating context will almost never be observed in practice due to the positive correlation among the θ's.

6 Discussion

There are various shortcomings in our data. For instance, in order to obtain a reasonable data base, we had to collect all ratings as of Dec. 31, 1998 irrespective of the date the rating was produced or changed. This implies that the default probabilities from columns (1) - (3) of table 3 are for a horizon of slightly more than four years. However, as "no change" need not imply "no assessment of creditworthyness", and both Moody's and $S\&P$ are known for keeping a close track of their customers, the presumption is that the ratings observed in December 1998 closely mirror the then prevailing economic situation. Also, there are no large deviations in the age of the ratings between Moody's and S&P, so this "ragged edge" problem is unlikely to produce additional discrepancies.

The major question of course is whether or not the sample summarized in table 2 can be taken as typical for the performance of the agencies. It does not cover a full business cycle but rather the end and apex of an extraordinary upswing and the beginning of a downturn in 2001 and 2002, were the majority of the defaults occurred. Therefore, it is necessary keep on collecting data on other 4-years prediction periods to check the robustness of our results.

Table 1:
The refinement ordering
among well calibrated probability forecasters

forecast of default probability %	distribution of borrowers across predicted default probabilities A	B	C	D
0.5	0	0	0.25	0.2
1	0	0,5	0	0.25
1.5	0	0	0.5	0
2	1	0	0	0
3	0	0,5	0	0.55
4.5	0	0	0.25	0

Table 2:
Distribution of borrowers across rating grades

S&P			Moody's		
rating grade	frequency	number of defaults	rating grade	frequency	number of defaults
AAA	55	0	Aaa	42	0
AA+	33	0	Aa1	47	0
AA	80	0	Aa2	90	0
AA-	157	0	Aa3	142	0
A+	167	1	A1	160	0
A	201	0	A2	191	2
A-	171	2	A3	154	0
BBB+	170	3	Baa1	170	3
BBB	189	4	Baa2	180	1
BBB-	148	9	Baa3	165	9
BB+	77	9	Ba1	69	6
BB	77	11	Ba2	50	2
BB-	85	26	Ba3	90	24
B+	147	53	B1	76	19
B	106	49	B2	104	36
B-	43	25	B3	114	50
C	21	17	C	83	57
	1927	209		1927	209

Table 3:
Default Probabilities (%)

Grade	our sample			historical	
	Moody's	S & P	average	Moody's	S & P
	(1)	(2)	(3)	(4)	(5)
AAA / Aaa	0	0	0	0.04	0.07
AA / Aa2	0	0	0	0.16	0.17
A / A2	0.40	0.56	0.48	0.36	0.48
BBB / Baaa2	2.52	3.16	2.84	1.69	2.58
BB / Ba2	15.31	19.25	17.41	8.76	11.69
B / B2	35.71	42.91	39.32	27.04	27.83
C / C	68.68	80.95	71.15	55.05	51.25

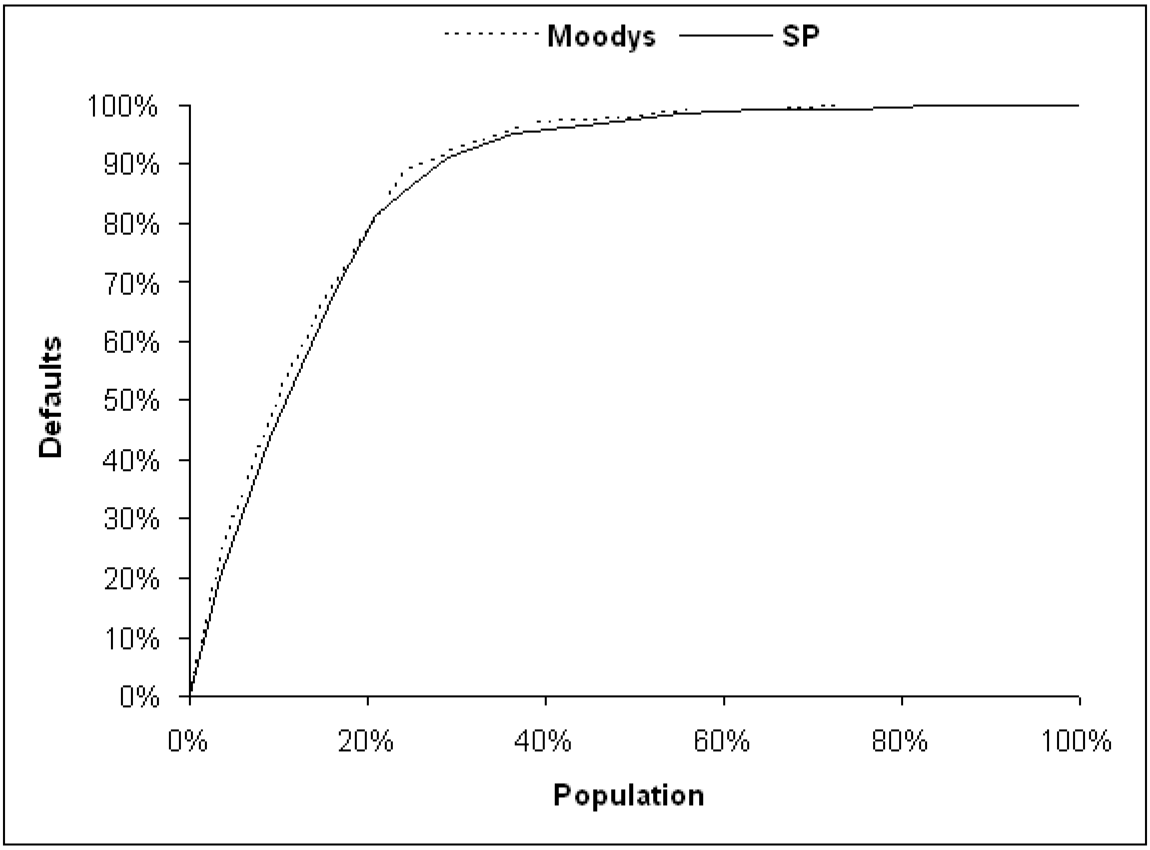
Table 4:
Conditional grade distributions given default and no given default, respectively

Grade	S&P		Moody's	
	$\sum q(a_j\|1) \times 209$	$\sum q(a_j\|0) \times 1718$	$\sum q(a_j\|1) \times 209$	$\sum q(a_j\|0) \times 1718$
AAA / Aaa	0	55	0	42
AA / Aa2	0	325	0	321
A / A2	3	861	2	824
BBB / Baaa2	19	1352	15	1326
BB / Ba2	65	1545	47	1503
B / B2	192	1712	152	1692
C / C	209	1718	209	1718

Table 5: Second order stochastic dominance of the distributions $q(a_i)$

a_i	Moody's		S&P	
(%)	$q(a_i)$	Integral	$q(a_i)$	Integral
0	16.66	0	16.86	0
0.40	26.21	0.065	0	0.067
0.56	0	0.135	27.97	0.094
2.52	26.72	0.978	0	0.976
3.16	0	1.418	26.31	1.259
15.31	10.85	9.977	0	9.906
19.25	0	13.043	12.40	12.706
35.71	15.26	26.289	0	26.463
42.91	0	33.171	15.27	32.471
68.68	4.31	57.832	0	57.933
80.95	0	70.111	1.19	70.065

Figure 1: Power Curves for Moody's and S&P



References

Arminger, G., Enache, D. and Bonne, T. (1997): "Analyzing credit risk data: A comparison of logistic discrimination, classification tree analysis, and feedforward networks." *Computational Statistics* 12, 293 – 310.

Blum, C., Overbeck, L. and Wagner, L. (2003): *An introduction to credit risk modelling*, Boca Raton (Chapman & Hall/CRC).

Brier, G.W. (1950): "Verification of forecasts expressed in terms of probability." *Monthly Weather Review* 78, 1 – 3.

Carey, M. (2002): "Some evidence on the consistency of banks' internal credit ratings." In: M. Ong (ed.) *Credit ratings: Methodologies, Rationale and Default Risk*, London (Risk Books).

Crouhy, M., Galai, D. and Mark, R. (2001): "Prototype risk rating systems." *Journal of Banking and Finance* 25, 47 – 95.

Dawid, A.P. (1982): "The well calibrated Bayesian." *Journal of the American Statistical Association* 77, 12 – 22.

DeGroot, M. and Fienberg, S.E. (1983): "The comparison and evaluation of probability forecasters", *The Statistican* 32, 12 – 22.

DeGroot, M. and Eriksson, E.A. (1985): "Probability forecasting, stochastic dominance, and the Lorenz curve." in J.M. Bernardo et al. (eds.): *Bayesian Statistics* 2, Amsterdam, 99 – 118.

Engelmann, B., Hayden, E. and Tasche, D. (2003): "Testing rating accuracy." *Risk* 16, 82 – 86.

Hajian-Tilaki, K. and Henley, J.A. (2002): "Comparison of three methods for estimating standard error of the area under the curve in ROC analysis of quantiative data." *Academic Radiology* 9, 1278 – 1285.

Krämer, W. (2002): "On the ordering of probability forecasts." *SFB 475 discussion paper* No.50, Dortmund.

Krämer, W. (2003a): "Die Bewertung und der Vergleich von Kreditausfallprognosen." to appear in *Kredit und Kapital*.

Krämer, W. (2003b): "Evaluating probability forecasts in terms of refinement and strictly proper scoring rules." SFB 475 discussion paper No. 24.

Redelmeier, D.A., Block, D.A. and Hickam, D.H. (1991): "Assessing predictive accuracy: How to compare Brier scores." *Journal of Clinical Epidemiology* 44, 1141 – 1146.

Savage, L.J. (1971): "Elicitation of personal probabilities and expectations." *Journal of the American Statistical Association* 66, 783 – 801.

Sobehart, J. and Keenan, S. (2001) : "Measuring default risk accurately", *Risk*, 31 – 33.

Tabakis, E. and Vinci, A. (2002): "Analysing and combining multiple credit assessments of financial institutions." *ECB working paper*, series No. 123.

Vardemann, S. and Meeden, G. (1983): "Calibration, sufficiency and domination: Considerations for Bayesian probability assessors." *Journal of the American Statistical Association* 78, 808 – 816.

Winkler, R.L. (1994): "Evaluating probabilities: Asymmetric scoring rules." *Management Science* 40, 1395 – 1405.

Winkler, R.L. (1996): "Scoring rules and the evaluation of probabilities." *Test* 5, 1 – 60.