Die Bewertung und der Vergleich von Kreditausfall-Prognosen

Professor Dr. Walter Krämer
Fachbereich Statistik, Universität Dortmund

1. Das Problem

Die finanzwirtschaftliche Bedeutung von Kreditausfällen und Kreditausfall-Prognosen bedarf keiner weiteren Begründung. Dieser Bedeutung angemessen, gibt es inzwischen eine Vielzahl von Modellen und Verfahren, die Ausfallwahrscheinlichkeiten von Bankkrediten oder Industrieanleihen zu schätzen; immer mehr Firmen und Institute bieten solche Schätzungen an, und immer mehr Anleger und Banken beziehen solche geschätzten Ausfallwahrscheinlichkeiten in ihre Entscheidungen mit ein.

Gegeben diese reichhaltige und stetig wachsende Palette von Methoden zur Prognose von Ausfallwahrscheinlichkeiten, erhebt sich fast von selbst die Frage: Welche Ausfallprognose ist "die beste"? Oder allgemeiner: wann ist eine Rating-Agentur oder ein Rating-Verfahren A "besser" als eine Rating-Agentur oder ein Rating-Verfahren B? Oder noch allgemeiner: Wie soll man überhaupt die Qualität einer Wahrscheinlichkeitsvorhersage bestimmen?

Die Anführungszeichen um "besser" und "die beste" deuten schon auf die fehlende Eindeutigkeit einschlägiger Qualitätsmaßstäbe hin. Die folgenden Seiten stellen die wichtigsten Kriterien vor. Diese beruhen im wesentlichen auf

- der „Spreizung“ der Wahrscheinlichkeitsprognosen in Richtung 0 und 100% (Abschnitt 2)
- einer Gegenüberstellung der Ausfälle in den „guten“ Ratingklassen (Abschnitt 3)
- dem Ausmaß der Konzentration der Ausfälle in den „schlechten“ Ratingklassen (Abschnitt 4) und
- einem direkten Vergleich von Wahrscheinlichkeitsprognosen mit tatsächlich eingetretenen Ereignissen (Abschnitt 5).

Die darauf aufbauenden Qualitätskriterien wurden vielfach zunächst im Kontext anderer Sachzusammenhänge wie Wetterprognosen in der Meteorologie der Krankheitsdiagnosen in der Medizin entwickelt, lassen sich aber unmittelbar auf die Prognose von Kreditausfällen übertragen. Um die Diskussion nicht mit sachfremden Problemen zu belasten, sei dabei unterstellt, daß über die Definition von "Kreditausfall" Konsens besteht und daß Bonitätsurteile sich eindeutig in Ausfallwahrscheinlichkeiten übersetzen lassen.

2. Trennschärfe versus Kalibrierung

Angenommen, 2% aller Kredite eines größeren Portfolios fallen erfahrungsgemäß binnen eines festen Zeitraums aus. Eine Rating-Agentur A, um eine Bewertung der Kredite dieses Portfolios gebeten, versieht jeden davon mit dem Etikett "Ausfallwahrscheinlichkeit 2%".

Diese Prognose ist "kalibriert" (synonym auch "valide" = valid oder "zuverlässig" = reliable, siehe Sanders 1963 oder Murphy 1973). Kalibriert bedeutet: Unter allen Krediten mit dem Etikett "Ausfallwahrscheinlichkeit x%" fallen langfristig x% tatsächlich aus.

Trotzdem ist dieses Rating wertlos – es liefert keine neuen Informationen, das alles hat man vorher schon gewußt. Oder anders ausgedrückt: Kalibrierung ist eine notwendige, aber keine hinreichende Bedingung für eine "gute" Wahrscheinlichkeitsprognose.

Agentur B teilt das Portfolio in zwei Gruppen auf: die erste mit Ausfallwahrscheinlichkeit 1%, die zweite mit Ausfallwahrscheinlichkeit 3%. Auch diese Bewertung sei kalibriert: In der ersten Gruppe fallen tatsächlich 1%, in der zweiten 3% der Kredite aus. Dann ist Agentur B ganz offensichtlich "besser" als Agentur A.

Das Rating von B heißt auch "trennschärfer" als das von A (synonym auch "sharper" oder "more refined", siehe Sanders 1963 und DeGroot und Fienberg 1983). Trennschärfe ist ein Maß für das "Spreizen" der Wahrscheinlichkeitsprognosen in Richtung 0 bzw. 100 Prozent. Die trennschärfste Wahrscheinlichkeitsprognose läßt nur zwei Aussagen zu: "Ein Kredit fällt sicher aus" (Prognose 100%), oder "ein Kredit fällt sicher nicht aus" (Prognose 0%). Ist eine solche extrem trennscharfe Prognose außerdem noch kalibriert, dann ist sie absolut perfekt: Das Rating sagt jeden Kreditausfall mit Sicherheit exakt voraus.

Eine solche Perfektion ist in der Praxis natürlich nie erreichbar. Maximal trennscharfe Systeme, etwa auf der Diskriminanzanalyse aufbauende Verfahren, die nur die Prognosen „Ausfall“ oder „Kein Ausfall“ zulassen, sind notwendigerweise niemals kalibriert. Sie müssen vielmehr mit zwei Arten von Fehlern leben: Bei einer Ausfallprognose von 0% tritt dennoch ein Ausfall ein – der Alpha-Fehler – oder bei einer Ausfallprognose von 100% tritt kein Ausfall ein – der Beta-Fehler. Je nach Bewertung und Wahrscheinlichkeit von Alpha- und Beta-Fehler lassen sich dann maximal trennscharfe Systeme hinsichtlich ihrer Prognosequalität vergleichen. Die einschlägigen Methoden sind seit langem wohlbekannt (siehe etwa Oehler und Unser 2001, Kapitel III.2) und müssen deshalb hier nicht ausführlich erörtert werden. Die folgende Diskussion beschränkt sich vielmehr auf kalibrierte, aber nicht maximal trennscharfe Ausfallprognosen, so wie sie für moderne Rating-Agenturen typisch sind, auf die das Konzept des Alpha- und Beta-Fehlers nicht direkt übertragbar ist.

Auch bei diesen kalibrierten, aber nicht maximal trennscharfen Prognosen ist es sinnvoll, nachzufragen: Welches von mehreren kalibrierten Rating-Systemen kommt dem Ideal einer maximal trennscharfen Prognose am nächsten? In obigem Beispiel ist System B trennschärfer als A. Und nochmals trennschärfer sind zwei Systeme C und D, welche die Kredite in die Ausfallklassen 0,5%, 1,5% und 4,5% bzw. 0,5%, 1% und 3% aufteilen.

Tabelle 1 zeigt eine mit Kalibrierung verträgliche Verteilung der Kredite auf die verschiedenen Ausfallklassen in den vier Prognosesystemen.

Tabelle 1:
Prognostizierte Ausfallwahrscheinlichkeiten
und ihre Verteilung auf Gesamtzahl der Kredite

Prognostizierte Ausfallwahrscheinlichkeit	Verteilung der Kredite auf die prognostizierten Ausfallwahrscheinlichkeiten			
	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

Mathematisch ist "trennschärfer" bei kalibrierten Prognosen dadurch definiert, daß sich die trennschwächere Prognose in gewissem Sinn aus der trennschärferen ableiten läßt. Das ist bei einem Vergleich von A und B ganz offenbar der Fall: unabhängig vom B-Etikett erhalten alle Kredite unter A die Prognose 2%. Aber auch die B-Prognose läßt sich ihrerseits aus der C-Prognose ableiten: Alle Kredite mit der C-Prognose 0,5% und die zufällig ausgewählte Hälfte aller Kredite mit der C-Prognose 1,5% erhalten das Etikett 1%, die übrigen das Etikett 3%. Das Ergebnis ist eine kalibrierte Prognose mit der gleichen Trennschärfe wie B.

Die B-Prognose läßt sich aber auch aus der D-Prognose ableiten: Alle D-Prognosen 0,5% und 1% sowie ein zufällig ausgewähltes Elftel der D-Prognosen 3% erhalten das Etikett 1%, die übrigen das Etikett 3%. Das Ergebnis ist wieder eine kalibirierte Prognose mit der gleichen Trennschärfe wie B.

Die Prognosen C und D lassen sich allerdings in diesem Sinne nicht vergleichen: Weder ist D trennschärfer als C, noch C trennschärfer als D. Die Trennschärfe erzeugt also keine vollständige Ordnung, sondern nur eine Halbordnung unter allen kalibrierten Wahrscheinlichkeitsprognosen; es gibt kalibrierte Wahrscheinlichkeitsprognosen, die nach dem Kriterium der Trennschärfe nicht vergleichbar sind. In solchen Fällen empfiehlt sich ein Rückgriff auf die weiter unten vorgestellten Qualitätsmaße aus Abschnitt 5.

Im Anhang findet sich ferner eine allgemeine Formel, die bei beliebigen kalibrierten Ratingsystemen entscheidet, ob die Systeme im Sinn der Trennschärfe vergleichbar sind.

3. Das Konzept der Ausfalldominanz

Unabhängig von Trennschärfe und Kalibrierung ist es sinnvoll, beim Vergleich zweier Ratingsysteme A und B zu fragen: "Welches der beiden Systeme hat die ausgefallenen Kredite am schlechtesten bewertet?" Diese Frage führt zum Begriff der "Ausfalldominanz" (Vardeman und Meeden 1983): Ein Ratingsystem B ist besser als ein Ratingsystem A im Sinne der Ausfalldominanz, falls B die ausgefallenen Kredite systematisch schlechter einstuft als A.

Formal: Sei $q_A(p_i)$ der Anteil der ausgefallenen Kredite, die von System A in die durch die prognostizierte Ausfallwahrscheinlichkeit p_i (i=0,...,K) definierte Ratingklasse einsortiert worden sind. Analog $q_B(p_i)$ usw. Dann ist B besser als A im Sinne der Ausfalldominanz, falls

$$\sum_{i=0}^{j} q_A(p_i) \leq \sum_{i=0}^{j} q_B(p_i) \qquad \text{für alle } j = 0,\ldots,K.$$

In kalibrierten Ratingsystemen errechnen sich die $q_A(p_i)$ durch

$$q_A(p_i) = \frac{p_i \times v_A(p_i)}{p} \;.$$

Dabei ist p die Gesamtausfallwahrscheinlichkeit und $v_A(p_i)$ der Anteil der von System A in Klasse p_i einsortierten Kredite. Analog $q_B(p_i)$, $v_B(p_i)$ usw.

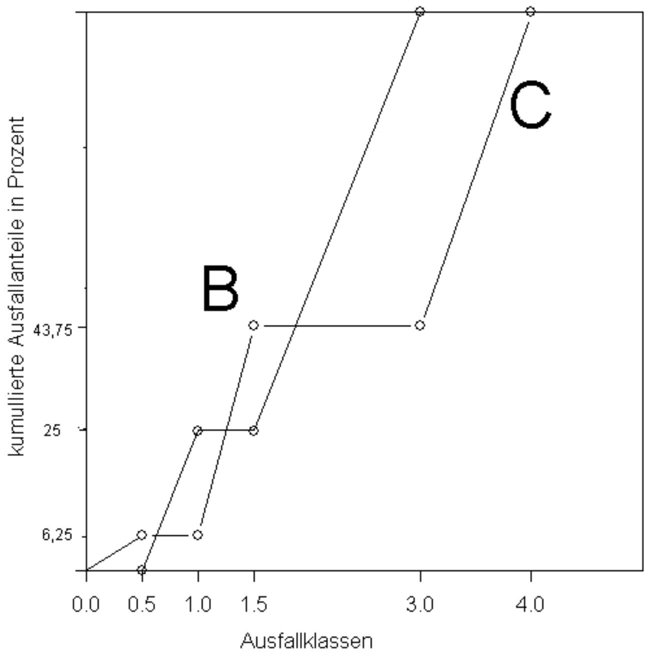
Tabelle 2 zeigt die so berechneten Anteile der ausgefallenen Kredite in den verschiedenen Ratingklassen der Systeme B und C aus Abschnitt 2.

Tabelle 2:
Verteilung der ausgefallenen Kredite auf die Ratingklassen

Ratingklasse	$q_B(p_i)$	$q_C(p_i)$
0,5%	0%	6,25%
1%	25%	0%
1,5%	0%	37,5%
3%	75%	0%
4,5%	0%	56,25%

Abbildung 1 stellt die kumulierten Summen der Ausfallanteile der Systeme B und C aus diesem Beispiel auch grafisch gegenüber. Es zeigt sich, daß keines der beiden Systeme das andere im Sinne der Ausfallordnung dominiert.

Abbildung 1:
Kumulierte Verteilung der ausgefallenen Kredite auf die Ratingklassen



Analog läßt sich auch in Bezug auf die nicht ausgefallenen Kredite fragen, ob eines von zwei zu vergleichenden Ratingsystemen diese systematisch besser bewertet. Sei dazu $\tilde{q}_A(p_i)$ und der Anteile der nicht ausgefallenen Kredite, die von System A in die verschiedenen Ratingklassen $p_i (i = 0, \ldots, K)$ einsortiert worden sind. Analog $\tilde{q}_B(p_i)$ usw. Dann ist B besser als A im Sinn der Nichtausfall-Dominanz, falls

$$\sum_{i=0}^{j} \tilde{q}_A(p_i) \geq \sum_{i=0}^{j} \tilde{q}_B(p_i) \qquad \text{für alle } j = 0,...,K.$$

In kalibrierten Ratingsystemen errechnen sich die $\tilde{q}_A(p_i)$ als

$$\tilde{q}_A(p_i) = \frac{(1 - p_i) \times v_A(p_i)}{1 - p} \, .$$

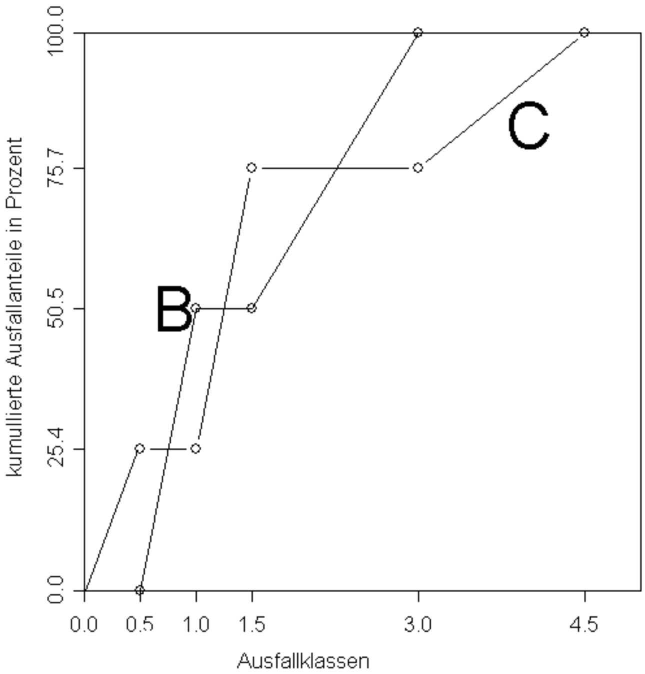
Analog $\tilde{q}_B(p_i)$ usw. Tabelle 3 zeigt die so berechneten Anteile der nicht ausgefallenen Kredite in den verschiedenen Ratingklassen der Systeme B und C aus Abschnitt 2.

Tabelle 3:
Verteilung der nicht ausgefallenen Kredite auf die Ratingklassen

Ratingklasse	$\tilde{q}_B(p_i)$	$\tilde{q}_C(p_i)$
0,5%	0%	25,4%
1%	50,5%	0%
1,5%	0%	50,3%
3%	49%	0%
4,5%	0%	24,3%

Das folgende Diagramm stellt auch diese Verteilungen grafisch dar. Wie aus der Abbildung zu sehen, ist also auch bezüglich der "Nichtausfallordnung" keines der beiden Systeme besser als das andere.

Abbildung 2:
Kumulierte Verteilung der nicht ausgefallenen Kredite auf die Ratingklassen



Bei kalibrierten Systemen ist das der Normalfall. Auch die entsprechenden Kurven von B und D sowie von und C und D schneiden sich. Insofern hilft das Konzept der Ausfalldominanz in vielen Anwendungen nicht weiter. Wenn es aber greift, d.h. wenn eine Prognose, die auch noch kalibriert ist, eine andere in diesem Sinne dominiert, dann ist der Verlierer wirklich schlecht. Daher empfiehlt sich das Konzept der Ausfalldominanz vor allem zum Aussortieren von Substandard-Systemen.

In der Sprache der Mathematik handelt es sich hier um einen Vergleich von Wahrscheinlichkeitsverteilungen über Ratingklassen. System B ist in dieser Sprache besser als System B im Sinn der Ausfalldominanz, wenn die in Tabelle 2 wiedergegebene bedingte Verteilung von B, gegeben Ausfall, diejenige von C stochastisch dominiert. Und B ist besser als C im Sinn der Nichtausfall-Dominanz, wenn die bedingte Verteilung von C, gegeben kein Ausfall, diejenige von B stochastisch dominiert.

Analog läßt sich auch der Trennschärfe-Vergleich aus Abschnitt 2 in die Sprache der stochastischen Dominanz übertragen (DeGroot und Eriksson 1985): :Ein kalibiriertes System A ist genau dann trennschärfer als ein kalibriertes System B, wenn die unbedingte Verteilung der Kredite auf die Ratingklassen unter A diejenige von B stochastisch in 2. Ordnung dominiert.

4. Die Lorenzkurve der Kreditausfälle

Angenommen, im Beispiel aus Abschnitt 2 sind insgesamt 800 Kredite zu bewerten. Agentur C prognostiziert für 200 davon eine Ausfallwahrscheinlichkeit von 0,5%, für 400 eine Ausfallwahrscheinlichkeit von 1,5%, und für 200 eine Ausfallwahrscheinlichkeit von 4,5%. Agentur C ist kalibriert, d.h. in der ersten Gruppe fällt im Mittel 1 Kredit (= 0,5% von 200) tatsächlich aus, in der zweiten Gruppe fallen 6 Kredite aus (= 1,5% von 400), in der dritten Gruppe 9 (= 4,5% von 200). Insgesamt gibt es im Mittel 16 Ausfälle (2% von 800). Im weiten sei der Einfachheit halber unterstellt, daß die erwarteten Ausfälle mit den tatsächlichen Ausfällen übereinstimmen. Gruppiert man die Kredite von schlecht nach gut, und stellt ihnen die kumulierten Anteile an den Ausfällen gegenüber, ergibt sich folgende Tabelle:

Tabelle 4:
Bonität vs. Ausfallanteile

Anteil an Gesamtzahl der bewerteten Kredite	Anteile an der Gesamtzahl der Ausfälle
0	0/16
0,25	9/16
0,75	15/16
1	16/16

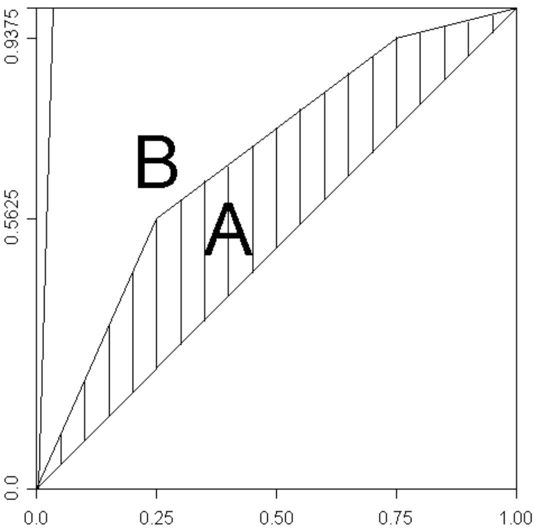
Diese Punkte, in ein 2-dimensionales Koordinatensystem übertragen und durch Geraden verbunden, erzeugen die Lorenzkurve – in der angelsächsischen Literatur auch "power curve" – der Kreditausfälle. Sie wird etwa von Moody's (siehe Falkenstein et al. 2000) zum Vergleich und zur Bewertung von Rating-Systemen eingesetzt.

Die aus der Statistik-Grundausbildung bekannte Definition der Lorenzkurve lautet etwas anders. Man sortiert die zu untersuchenden Objekte von klein nach groß, und die resultierende Lorenzkurve ist nach unten gebogen. Die Lorenzkurve der Kreditausfälle dagegen sortiert die zu untersuchenden Objeke von groß (=Hohe Ausfallwahrscheinlichkeit) nach klein, und ist in der Regel nach oben gebogen.

Ein System, das in jeder Rating-Klasse die gleichen prozentualen Ausfallanteile hätte, hat als Lorenzkurve die Diagonale. Dieses System liefert keine Informationen und ist in diesem Sinne das schlechtest mögliche.

Abbildung 3 zeigt die Lorenzkurve der Prognose C. Ebenfalls eingezeichnet ist die optimale Lorenzkurve eines Ratingsystems, das alle 16 Ausfälle, und nur diese, in die schlechteste Bonitätsklasse aufgenommen hätte. Diese begrenzt zusammen mit der Winkelhalbierenden die Fläche B.

Abbildung 3:
Eine beispielhafte Lorenzkurve von Kreditausfällen



Das Verhältnis der Fläche A zur Fläche B heißt „Trefferquote“ („accuracy ratio“). Je höher die Trefferquote, desto näher kommt ein Rating-System an die in obigem Sinn optimale Prognose heran.

Die Lorenzkurve der C-Prognosen ist konkav. Das bedeutet: In einer schlechteren Ratingklasse fallen prozentual mehr Kredite aus als in einer besseren. Ratingsysteme mit dieser Eigenschaft heißen auch "semi-kalibriert".

Lorenzkurven von Ausfallwahrscheinlichkeiten sind invariant gegenüber monotonen Transformationen der vorhergesagten Ausfallwahrscheinlichkeiten. Hätte das System C statt 0,5%, 1,25% und 4,5% die Ausfallwahrscheinlichkeiten 10%, 20% und 50% vorhergesagt, bliebe die Lorenzkurve der Kreditausfälle unverändert. Lorenzkurven von Kreditausfällen messen also nur, inwieweit die Rangfolge der Bonitätsurteile mit den tatsächlichen Ausfallanteilen korrespondiert; über die Treffersicherheit der in diesen Bonitätsurteilen enthaltenen Wahrscheinlichkeitsprognosen (also über die Kalibrierung des Systems) sagen sie nichts.

Im Sinn der Lorenzkurve „gute“ Ratingsysteme sind also nicht ohne weiteres zur Preisfindung bei Krediten einzusetzen; hier kommt es auf die genaue Prognose der Ausfallwahrscheinlichkeiten an. Sind aber nur die x% schlechtesten Kredite auszufiltern, etwa zur Kredit-Rationierungszwecken, sind Systeme mit einer hohen Trefferquote ideal

5. Abweichungsmaße

Eine alternative Möglichkeit zur Beurteilung der Qualität von Wahrscheinlichkeitsprognosen ist der direkte Vergleich von Prognosen und tatsächlich eingetretenem Ereignis. Seien $p_0, p_1, \ldots, p_K$ mit $p_0 = 0$, $p_K = 1$ die möglichen, zur Prognose zugelassenen Ausfallwahrscheinlichkeiten. (Zur Erinnerung: hier ist unterstellt, daß sich Ratingklassen in eindeutige Ausfallwahrscheinlichkeiten übersetzen lassen). Insgesamt gebe es n zu bewertende Kredite. Sei p^j die Prognose für Kredit j, und sei $\theta^j = 1$ bei Ausfall und $\theta^j = 0$ (kein Ausfall). Dann ist das Brier-Maß ("Brier-Score", nach G.W.Brier 1950) definiert als

$$(1) \quad B = \frac{1}{n} \sum\nolimits_{j=1}^{n} (p^j - \theta^j)^2 .$$

Der Brier-Score ist das bekannteste Maß zur Bewertung von Wahrscheinlichkeitsprognosen. Er wurde und wird bislang vor allem zum Qualitätsvergleich von Wettervorhersagern eingesetzt, ist aber grundsätzlich in allen Kontexten einsetzbar, in denen Wahrscheinlichkeitsprognosen zu vergleichen sind.

Je größer der Brier-Score, desto schlechter die Wahrscheinlichkeitsprognose. Der schlechtest mögliche Wert von B = 1 ergibt sich für eine Prognose von immer nur 0 oder 100% Wahrscheinlichkeit für Ausfall, bei der stets das Gegenteil des Vorhergesagten eintritt. Der bestmögliche Wert von 0 ergibt sich für eine Prognose von immer nur 0% oder 100% für Ausfall, bei der stets das Vorhergesagte tatsächlich eintritt.

Das Zahlenbeispiel aus Abschnitt 3 liefert für die Ratingssysteme A, B, und C (unter der Annahme, daß die tatsächlichen Ausfälle mit den erwarteten Ausfällen zusammenfallen):

$$B_A = \frac{1}{800}\left[16\,(0{,}02-1)^2 + 784\,(0{,}02-0)^2\right] = 0{,}0196$$

$$\mathrm{B_B} = \frac{1}{800}\Big[4\ (0{,}01-1)^2 + 396\,(0{,}01-0)^2 \\ + 12\,(0{,}03-1)^2 + 388\,(0{,}03-0)^2\Big] = 0{,}0195$$

$$B_C = \frac{1}{800}\Big[1\,(0{,}005-1)^2 + 199\,(0{,}005-0)^2 \\ + 6\ (0{,}015 - 1)^2 + 394\,(0{,}015-0)^2 \\ + 9\,(0{,}045 - 1)^2 + 191\ (0{,}045-0)^2\Big] = 0{,}0194$$

In System D sind die erwarteten Ausfälle nicht ganzzahlig. Damit können bei 800 zu bewertenden Krediten die erwarteten und tatsächlichen Ausfälle nicht übereinstimmen, und es unterbleibt hier eine numerische Auswertung.

Die obigen Brier-Scores weichen kaum voneinander ab. Außerdem sind sie alle sehr klein (d.h. sehr gut). Das ist ein gravierender Mangel des Standard-Brier-Scores: Ist die Gesamtausfallwahrscheinlichkeit sehr klein, wie etwa 2% in obigem Zahlenbeispiel, so liefert schon die Trivialprognose von 2% Ausfallwahrscheinlichkeit für alle Kredite einen guten Brier-Score (in obigem Beispiel: $\mathrm{B_A} = 0{,}0196$).

Bei einem Gesamtausfall-Anteil p hat die Trivialprognose "Ausfallwahrscheinlichkeit von p für jeden Kredit" den (erwarteten) Brier-Score

$$(2) \qquad \overline{\mathrm{B}} = \mathrm{p}(1-\mathrm{p})^2 + (1-\mathrm{p})\mathrm{p}^2\ .$$

Dieser Ausdruck strebt für $p \to 0$ ebenfalls gegen 0 (dito für $p \to 1$). Das ist bei Anwendungen wie Kreditausfallprognosen, mit sehr kleinen Wahrscheinlichkeiten für das fragliche Ereignis, ein Problem. Es empfiehlt sich daher in den Anwendungen auf jeden Fall, einen realisierten Brier-Score relativ zu dem Trivialscore (2) zu sehen.

Es ist leicht zu überprüfen (De Groot und Fienberg 1983), daß ein Anwender seinen erwarteten Brier-Score immer dann minimiert, wenn er als Prognose für die Ausfallwahrscheinlichkeit seine wahre subjektive Ausfallwahrscheinlichkeit einsetzt. Insofern belohnt der Brier-Score „ehrliches“ Verhalten. Abweichungsmaße mit dieser Eigenschaft heißen in der angelsächsischen Literatur auch „proper scoring rules“ (Winkler 1969).

Ein deutscher Ausdruck dafür wäre „anreizkompatible Abweichungsmaße“. Ein weiteres anreizkompatibles Abweichungsmaß ist die Mittlere logarithmische Abweichung (Good 1952)

$$(3) \qquad L = \frac{1}{n} \sum_{j=1}^{n} -\log\left(\left|p^j + \theta^j - 1\right|\right).$$

Anreizkompatible Abweichungsmaße wie der Brier-Score oder die Mittlere logarithmische Abweichung bieten sich als Entlohnungskriterium für Kreditsachbearbeiter an: Es lohnt sich, die wahren subjektiven Ausfallwahrscheinlichkeiten offenzulegen. Untertreibungen oder Übertreibungen der subjektiv für richtig gehaltenen Ausfallwahrscheinlichkeiten verschlechtern den subjektiven Erwartungswert des Abweichungsmaßes und werden insofern bestraft.

6. Würdigung

Der Einfachheit halber wurden in den obigen Beispielen relative Häufigkeiten und Wahrscheinlichkeiten gleichgesetzt. In der Praxis ergibt sich noch das zusätzliche Problem, daß auch bei kalibrierten Systemen die realisierten relativen Ausfallhäufigkeiten zufällig von den prognostizieren Ausfallwahrscheinlichkeiten abweichen können (und zuweilen sogar müssen – siehe System D bei 800 Krediten insgesamt). Um festzustellen, ob solche zufälligen Abweichungen wirklich nur zufällig oder aber systematisch und damit ein Indikator für fehlende Kalibrierung sind, gibt es formale statistische Tests, die hier nicht weiter interessieren sollen. Auch die für die Praxis zentralen Problem der Definition von „Kreditausfall“ und der Festlegung des Prognosehorizontes wurden hier nicht weiter diskutiert.

Die Notwendigkeit, konkurrierende Ratingsysteme gegeneinander abzuwägen, bleibt davon unberührt. Auch bei nicht ganz perfekten Daten helfen die oben vorgestellten Kriterien bei praktischen Entscheidungen. Sie erzwingen die Anerkennung des für Wahrscheinlichkeitsprognosen zentralen Sachverhaltes, daß die Übereinstimmung von prognostizierten und realisierten relativen Häufigkeiten für sich allein noch keine gute Prognose darstellt, und stellen die Vorteile der „Spreizung“ der prognostizierten Wahrscheinlichkeiten in den Mittelpunkt. Je „gespreizter“ eine Wahrscheinlichkeitsprognose, desto besser schneidet sie ceteris paribus bei allen oben aufgeführten Qualitätskriterien ab, und desto nützlicher ist sie ganz offensichtlich für die Praxis (denn desto näher kommt sie an die sichere Vorhersage heran). Insofern geht also die oben dargestellte abstrakte Theorie mit den Erfordernissen der Praxis Hand in Hand.

Mathematischer Anhang: Trennschärfevergleich von kalibrierten Ratingsystemen

Seien $p_0<...<p_K$ (mit $p_0 = 0$, $p_K = 1$) die zur Auswahl stehenden Ausfallwahrscheinlichkeiten. Sei $v_A(p_i)$ sei der Anteil der Kredite mit der vorhergesagten Ausfallwahrscheinlichkeit p_i unter einem kalibirierten Ratingsystem A, und analog $v_B(p_i)$ die vorhergesagten Ausfallwahrscheinlichkeiten unter eine, kalibrierten Ratingsstem B. Dann gilt folgendes allgemeine Resultat (De Groot und Fienberg 1983, Theorem 1):

Ein kalibriertes Ratingsystem A ist genau dann trennschärfer als ein kalibriertes Ratingsystem B, wenn

$$\sum_{i=0}^{j-1} \left(p_j - p_i\right)\left(v_A(p_i) - v_B(p_i)\right) \geq 0$$

für alle j = 1,...,K-1.

Diese Formel zeigt sofort, daß die Systeme C und D aus Abschnitt 2 nicht zu vergleichen sind. Es gilt:

Für j=2: $(p_2 - p_1)(v_C(p_1) - v_D(p_1) = (1 – 0{,}5)(0{,}25 – 0{,}2) = 0{,}5 \,.\, 0{,}5 = 0{,}25 > 0$.

Für j=3: $(p_3 - p_1)(v_C(p_1) - v_D(p_1) + (p_3 – p_2)(v_C(p_2 - v_D(p_2) =$
$(1{,}5 – 0{,}5)(0{,}25 - 0{,}2) + (1{,}5 – 1)(0 – 0{,}25) =$
$0{,}05 – 0{,}125 = - 0{,}075 < 0$.

Literatur:

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

DeGroot, M. und Fienberg, S. (1983): "The comparison and evaluation of forecasters." *The Statistician* 32, 12 – 23.

DeGroot, M. und Eriksson, E.A. (1985): „Probability forecasting, stochastic dominance, and the Lorenz curve,“ in: S. S. Gupta und J. O. Berger (Hrsg): *Statistical decision theory and related topics III*, Vol 1, New York (Academic Press), S. 291-314.

Falkenstein, E., Boral, A. und Kocagil, A.E. (2000): „RiskCalc for private companies II: More results and the Australian Model.“ *Moody's Investor Services*, Report No. 62265.

Good, I.J. (1952): "Rational decisions." *Journal of the Royal Statistical Society B* 14, 107 – 114.

Murphy, A.H. (1973): "A new vector partition of the probability score." *Journal of Applied Meteorology* 12, 595 – 600.

Oehler, A. und Unser, M. (2001): *Finanzwirtschaftliches Risikomanagement*, Berlin (Springer) .

Sanders, F. (1963): "On subjective probability forecasting." *Journal of Applied Meteorology* 2, 191 – 201.

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

Winkler, R.L. (1969): "Scoring rules and the evaluation of probability assessors." *Journal of the American Statistical Association* 64, 1073 – 1078.

Winkler, R.L. (1986): "On good probability appraisers." In: P. Goel und A. Zellner: *Bayesian Inference and Decision Techniques*, Amsterdam (Elsevier), S. 265 – 278.

Die Arbeit entstand im Rahmen des Sonderforschungsbereiches 475 "Komplexitätsreduktion in multivariaten Datenstrukturen", Teilprojekt B1: "Kapitalmarktpreise". Ich danke Martin Weber für Kommentare und konstruktive Kritik.