Kredit und Kapital, 36. Jahrgang, Heft 3
Seiten 395–410

Die Bewertung und der Vergleich von Kreditausfall-Prognosen

Von Walter Krämer*, Dortmund

I. 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 II.),
- einer Gegenüberstellung der Ausfälle in den „guten" Ratingklassen (Abschnitt III.),
- dem Ausmaß der Konzentration der Ausfälle in den „schlechten" Ratingklassen (Abschnitt IV.) und

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

- einem direkten Vergleich von Wahrscheinlichkeitsprognosen mit tatsächlich eingetretenen Ereignissen (Abschnitt V.).

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.

II. Trennschärfe versus Kalibrierung

Angenommen, 2% aller Kredite eines größeren Portfolios fallen erfahrungsgemäß binnen eines festen Zeitraums, etwa eines Jahres, 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) oder *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 wohl bekannt (siehe etwa *Oehler* und *Unser* (2001), Kapitel III.2) und müssen deshalb hier nicht weiter 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.

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

Tabelle 1

Prognostizierte Ausfallwahrscheinlichkeiten und ihre Verteilung auf die 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

Ergebnis ist wieder eine kalibrierte 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 V.

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

III. 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 $p_i (i = 0, \ldots, K)$ 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 II.

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.

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* aus-

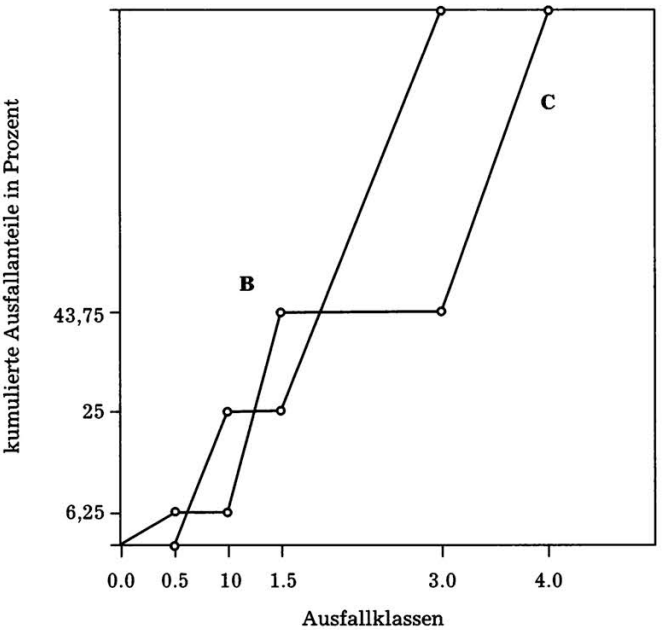

Abbildung 1: Kumulierte Verteilung der ausgefallenen Kredite auf die Ratingklassen

gefallenen Kredite, die von System A in die verschiedenen Ratingklassen $p_i (i = 0, \ldots, \mathrm{K})$ einsortiert worden sind. Analog $\tilde{q}_\mathrm{B}(p_i)$ usw. Dann ist B besser als A im Sinn der Nichtausfall-Dominanz, falls

$$\sum_{i=0}^{j} \tilde{q}_\mathrm{A}(p_i) \geq \sum_{i=0}^{j} \tilde{q}_\mathrm{B}(p_i) \qquad \text{für alle } j = 0, \ldots, \mathrm{K}.$$

In kalibirierten 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 II.

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.

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 %

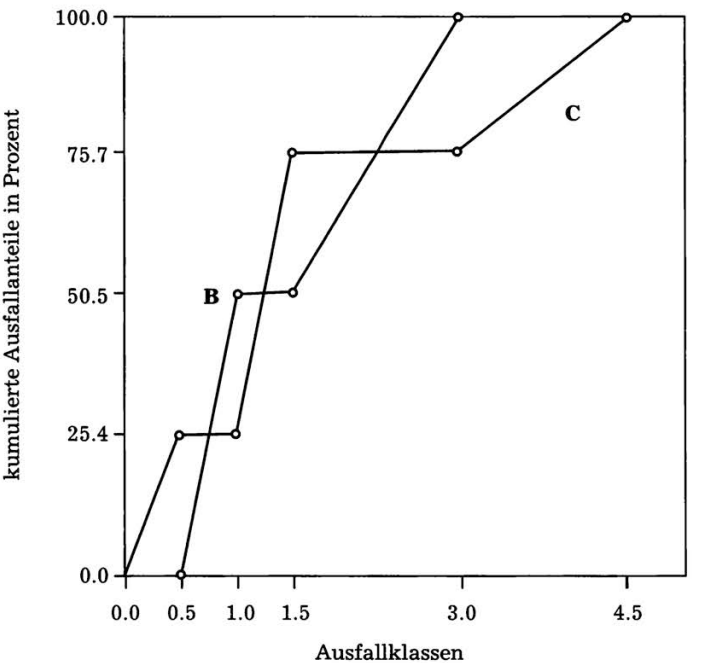

*Abbildung 2: Kumulierte Verteilung der **nicht** ausgefallenen Kredite auf die Ratingklassen*

Bei kalibrierten Systemen ist das der Normalfall. Auch die entsprechenden Kurven von B und D sowie von 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 C 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 II. in die Sprache der stochastischen Dominanz übertragen (*DeGroot* und *Eriksson* (1985)): Ein kalibriertes 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.

IV. Die Lorenzkurve der Kreditausfälle

Angenommen, im Beispiel aus Abschnitt II. 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 Weiteren 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.

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".

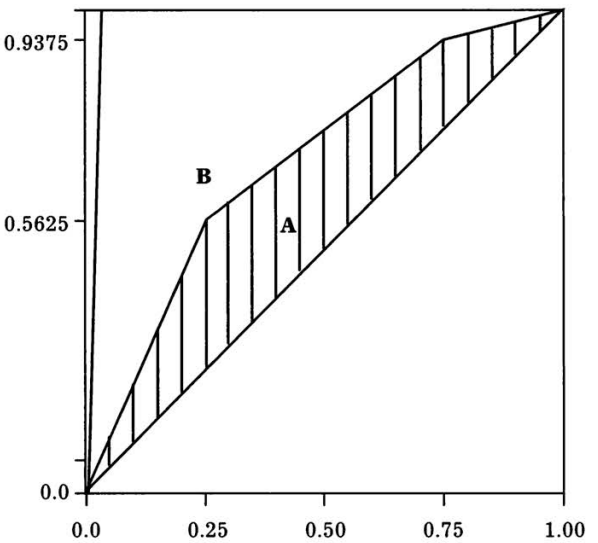

Abbildung 3: Eine beispielhafte Lorenzkurve von Kreditausfällen

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 zu Kredit-Rationierungszwecken, sind Systeme mit einer hohen Trefferquote ideal.

Alternativ zur Lorenzkurve werden häufig auch die ROC-Kurve („Receiver Operating Characteristic-Kurve") und die Fläche unter dieser Kurve eingesetzt. Die ROC-Kurve trägt die kumulierten Ausfallanteile, statt gegen die kumulierten Anteile an *allen* bewerteten Krediten, gegen die kumulierten Anteile an den *nicht* ausgefallenen Krediten ab. Sie wird häufig in der Medizin zur Bewertung konkurrierender Diagnose-

methoden eingesetzt (siehe etwa *Zweig* und *Campbell* (1993)), liefert aber keine neuen Informationen. Es ist leicht zu sehen (*Krämer* (2002), Theorem 3), dass zwei ROC-Kurven sich genau dann schneiden, wenn die entsprechenden Lorenzkurven sich schneiden, und auch die Fläche zwischen ROC-Kurve und Winkelhalbierender liefert keine zusätzlichen Unterscheidungsmöglichkeiten; sie ist nummerisch identisch mit der weiter oben definierten Trefferqoute (siehe etwa *Engelmann* et al. (2003); dieser Sachverhalt wurde unabhängig von verschiedenen anderen Autoren ebenfalls bemerkt).

V. Abweichungsmaße

Eine alternative Möglichkeit zur Beurteilung der Qualität von Wahrscheinlichkeitsprognosen ist der direkte Vergleich von Prognose 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

$$B = \frac{1}{n} \sum_{j=1}^{n} (p^j - \theta^j)^2. \tag{1}$$

Der Brier-Score ist das bekannteste Maß zur Bewertung von Wahrscheinlichkeitsprognosen. Er wurde und wird bislang vor allem zum Qualitätsvergleich von Wettervorhersagen 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 III. liefert für die Ratingsysteme A, B und C (unter der Annahme, dass die tatsächlichen Ausfälle mit den erwarteten Ausfällen zusammenfallen):

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

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

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

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: $B_A = 0{,}0196$).

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

$$\bar{B} = p(1-p)^2 + (1-p)p^2. \tag{2}$$

Dieser Ausdruck strebt für $p \longrightarrow 0$ ebenfalls gegen 0 (dito für $p \longrightarrow 1$). 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. Derart adaptierte Abweichungsmaße werden auch „skill-scores" genannt (*Winkler* (1986)).

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))

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

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 offen zu legen. Untertreibungen oder Übertreibungen der subjektiv für richtig gehaltenen Ausfallwahrscheinlichkeiten verschlechtern den subjektiven Erwartungswert des Abweichungsmaßes und werden insofern bestraft.

Unabhängig von der Art des verwendeten Abweichungsmaßes stellt sich ferner das Problem seiner stochastischen Eigenschaften. Ist ein Prognosesystem „systematisch" besser als eine Zufallsprognose (d.h., ist die Trefferquote „signifikant" größer als Null)? Ist eine Rating-Agentur tatsächlich „besser" als die Konkurrenz, oder geht ein Vorsprung, etwa gemessen durch die Trefferquote oder den Brier-Score, nur auf zufällige Abweichungen der Stichprobe von den „wahren" Populationsparametern zurück? Hier gibt es erste Ansätze (siehe etwa *Blochwitz* et al. (2000) oder *Engelmann* et al. (2003), die einen Signifikanztest für die Trefferquote entwickeln), aber im Großen und Ganzen steht eine Antwort auf diese Fragen im Augenblick noch aus.

VI. Ausblick

Der Einfachheit halber wurden in den obigen Beispielen relative Häufigkeiten und Wahrscheinlichkeiten gleichgesetzt. Neben der Problematik der stochastischen Eigenschaften von Trefferquoten und Abweichungsmaßen wurde damit auch die empirische Überprüfung der Kalibrierungseigenschaft ausgeblendet, denn in der Praxis werden auch bei kalibrierten Systemen die realisierten relativen Ausfallhäufigkeiten zufällig von den prognostizierten Ausfallwahrscheinlichkeiten abweichen. Darauf gehen etwa *Höse* und *Huschens* (2003) näher ein. Auch die für die Praxis zentralen Probleme der Definition von „Kreditausfall" und der Festlegung des Prognosehorizontes, wie auch die Konstruktion von Ratingsystemen im Allgemeinen, wurden hier nicht weiter diskutiert (siehe dazu etwa *Krahnen* und *Weber* (2001)).

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 $p_0 < \ldots < p_K$ (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 kalibrierten Ratingsystem A, und analog $v_B(p_i)$ die vorhergesagten Ausfallwahrscheinlichkeiten unter einem kalibrierten Ratingsystem 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
>
> $$\sum_{i=0}^{j-1} (p_j - p_i)(v_A(p_i) - v_B(p_i)) \geq 0 \qquad \text{für alle } j = 1, \ldots, K-1.$$

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

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

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

Blochwitz, S./*Liebig,* Th./*Nyberg,* M. (2000): „Benchmarking Deutsche Bundesbank's, default risk model, the KMV private firm model and common financial ratios for German cooperations." Bundesbank-Diskussionspapier. – *Brier,* G. W. (1950): „Verification of forecasts expressed in terms of probability." *Monthly Weather Review* 78, 1–3. – *DeGroot,* M./*Fienberg,* S. (1983): „The comparison and evaluation of forecasters." *The Statistician* 32, 12–23. – *DeGroot,* M./*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. – *Engelmann,* B./*Hayden,* E./*Tasche,* D. (2003): „Testing rating accuracy." *Risk* 16, 82–86. – *Falkenstein,* E./*Boral,* A./*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. – *Höse,* S./*Huschens,* S. (2003): „Sind interne Ratingsysteme im Rahmen von Basel II evaluierbar? – Zur Schätzung von Ausfallwahrscheinlichkeiten durch Ausfallquoten." *Zeitschrift für Betriebswirtschaft* 73, 139–168. – *Krahnen,* J. P./*Weber,* M. (2001): „Generally accepted rating principles: A primer." *Journal of Banking and Finance* 25, 2–24. – *Krämer,* W. (2002): „On the ordering of probability forecasts." SFB 475 Diskussionspapier 50/02, Dortmund. – *Murphy,* A. H. (1973): „A new vector partition of the probability score." *Journal of Applied Meteorology* 12, 595–600. – *Oehler,* A./*Unser,* M. (2001): *Finanzwirtschaftliches Risikomanagement,* Berlin (Springer). – *Sanders,* F. (1963): „On subjective probability forecasting." *Journal of Applied Meteorology* 2, 191–201. – *Vardeman,* S./*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.

Zusammenfassung

Die Bewertung und der Vergleich von Kreditausfall-Prognosen

Der Aufsatz zeigt, dass Kreditausfallprognosen im Sinne von Wahrscheinlichkeitsprognosen in vielfacher Hinsicht in eine Qualitätsreihenfolge überbracht werden können. Insbesondere macht er deutlich, dass die Übereinstimmung von vorhergesagter Ausfallwahrscheinlichkeit und der relativen Häufigkeit der tatsächlich ausgefallenen Kredite für sich alleine noch keine Qualität verbürgt. Letztere tritt erst durch die „Spreizung" der vorhergesagten Ausfallwahrscheinlichkeiten in Richtung 0% und 100% ein. Darüber hinaus werden auch verschiedene skalarwertige Maße für Prognosequalitäten diskutiert. (JEL C11, C52, G33)

Summary

Evaluation and Comparison of Default Forecasts in the Rating Industry

This article shows that probability forecasts may in many respects be ranked in terms of quality. It makes clear in particular that correspondence of the predicted default probability and the relative frequency of the actual defaults is no quality guarantee in itself. The latter would only be the case when the predicted default probability is near 0% and near 100%. In addition, this article discusses various scalar measures permitting a ranking of the forecasting quality.

Résumé

L'évaluation et la comparaison des prévisions sur les pertes de crédit

Dans cet article, l'auteur montre que les prévisions sur les pertes de crédit, dans le sens de prévisions de probabilités, peuvent être mises à plusieurs points de vue dans une série de qualité. Il explique surtout clairement que la concordance de la probabilité prévue de pertes et la fréquence relative des crédits réellement perdus n'est pas en soi une garantie de qualité. La qualité n'est assurée que par l' »écartement« des probabilités prévues de pertes entre 0% et 100%. En plus, l'auteur discute de différentes mesures d'échelles pour les qualités des prévisions.