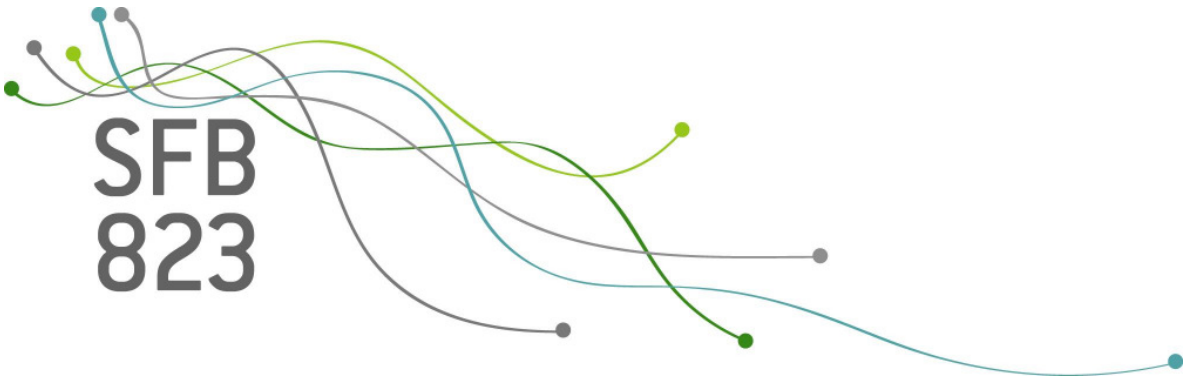
SFB 823

Statistischer Qualitätsvergleich von Kreditausfallprognosen

Walter Krämer, Michael Bücker

Nr. 30/2009

Discussion Paper



Statistischer Qualitätsvergleich von Kreditausfallprognosen

Prof. Dr. Walter Krämer[1]

Technische Universität Dortmund, Fakultät Statistik, 44221 Dortmund, Deutschland
walterk@statistik.tu-dortmund.de

Michael Bücker

Technische Universität Dortmund, Fakultät Statistik, 44221 Dortmund, Deutschland
buecker@statistik.tu-dortmund.de

Zusammenfassung

Die statistische Qualität von Kreditausfallprognosen läßt sich auf unterschiedliche Art und Weise messen und vergleichen. Die vorliegende Arbeit faßt die in der Literatur gemachten Vorschläge zusammen und diskutiert deren Eignung für das Alltagsgeschäft von Kreditausfallprognosen im Privatkundenbereich. Es zeigt sich, daß nicht alle Qualitätskriterien hier sinnvoll anzuwenden sind. Insbesondere scheinen die etwa in der Meteorologie beliebten Brier Scores und verwandte Kriterien für diese Anwendungen nur schlecht geeignet.

[1] Diese Arbeit entstand im Rahmen des DFG-Sonderforschungsbereiches 823: „Statistik nichtlinearer dynamischer Prozesse in Wirtschaft und Technik“

1. Das Problem

Die vorliegende Arbeit befaßt sich mit Aussagen der Gestalt: „Die Wahrscheinlichkeit, dass Kunde A (Firma B, Land C) binnen eines – zukünftigen – Zeitraums z ihren oder seinen Kreditverbindlichkeiten nicht in vollem Maße nachkommt, beträgt p Prozent." Dergleichen Wahrscheinlichkeitsprognosen sind aus offensichtlichen Gründen von höchstem praktisch-ökonomischem Interesse – ob ein Unternehmen oder eine Gebietskörperschaft von einer der etablierten Rating-Agenturen mit einem A oder mit einem AA oder gar mit einem der begehrten AAAs bewertet wird, hat maßgeblichen Einfluß auf die Kosten der Verschuldung. Oder wie es der bekannte New York Times Journalist Thomas L. Friedman (1996) einmal formulierte:

> *„There are two superpowers in the world today in my opinion. There's the United States and there's Moody's Bond Rating Service. The United States can destroy you by dropping bombs, and Moody's can destroy you by downgrading your bonds. And believe me, it's not clear sometimes who's more powerful."*

Und auch für Privatkunden haben dergleichen Wahrscheinlichkeitsprognose – in der Praxis meist verkleidet als der erreichte „Score" in einem Scorekartensystem – erhebliche praktische Bedeutung; sie entscheiden oft darüber, ob er oder sie überhaupt einen Kredit erhält. Der Konkretheit halber werden sich die folgenden Ausführungen vor allem auf solche Kreditausfallprognosen für Privatkunden und die dazu verwendeten sogenannten „Scorekarten" beziehen.[2]

Hier gibt es ein erstes, oft unerkanntes oder mißverstandenes Problem, das auch einen Großteil der einschlägigen Literatur durchzieht und stört. Nämlich, wie eine Aussage wie die obige überhaupt zu verstehen ist. Das ist alles andere als offensichtlich. Oft z. B. wird die prognostizierte Ausfallwahrscheinlichkeit als eine Eigenschaft des Kunden angesehen. Dabei steht ein (hypothetisches) Zufallsexperiment im Hintergrund, bei dem jeder Kunde aus einer kundenspezifischen Schublade von hypothetischen Kunden gezogen wird, in der p% ausfallen (siehe etwa Hamerle u. a. 2003). Sonst weiß man nichts. Und die Aufgabe des Statistikers ist es dann, dieses p zu schätzen.

Diese Sichtweise ist für die Praxis aber nicht die relevante. In der Praxis werden am Ende der Prognoseperiode alle Karten aufgedeckt, d.h. ein Kunde fällt entweder aus oder nicht. Oder anders ausgedrückt: Konkrete Kunden sind hier nicht Träger von Wahrscheinlichkeiten! Wahrscheinlichkeiten gibt es nur für zufällige Ereignisse in Zufallsexperimenten, etwa dem: „Wähle zufällig aus allen Kunden mit den Attributen männlich, geschieden, über 30, einen aus." Dann ist die Aussage sinnvoll: „Die Wahrscheinlichkeit, dass dieser Kunde ausfällt, beträgt 10 Prozent." Und diese Aussage ist darüber hinaus nicht nur sinnvoll, sondern auch wahr, wenn tatsächlich in dieser Kundengruppe jeder zehnte Kunde ausfällt.

---

[2] Zu Ausfallprognosen bei Unternehmen siehe auch Krämer und Güttler (2008).

Formal gesehen handelt es sich hier um bedingte Wahrscheinlichkeiten, gegeben die tatsächlichen Ausfälle am Ende der Periode. Interpretiert man Ausfallwahrscheinlichkeiten auf diese Art und Weise, werden viele auf den ersten Blick verwirrende Phänomene im Kreditbewerten sofort leichter zu verstehen. Etwa folgendes: Prognostiker A behauptet: Kunde X fällt aus mit Wahrscheinlichkeit 2%, Prognostiker B behauptet, Kunde X fällt aus mit Wahrscheinlichkeit 4%. Wer hat Recht?

Interpretiert man Wahrscheinlichkeiten als Eigenschaften eines Kunden: höchstens einer. Auf keinen Fall aber beide zusammen. Interpretiert man Wahrscheinlichkeiten dagegen als Eigenschaften eines Kollektivs: Möglicherweise beide. Das hängt allein davon ab, in welche Schublade die beiden Prognostiker den konkreten Kunden einsortieren. Und da ist es problemlos möglich, dass Prognostiker A den Kunden in eine Schublade sortiert, in der 2% aller Kredite ausfallen, und Prognostiker B sortiert den Kunden in eine Schublade ein, in der 4% aller Kunden ausfallen. Oder anders ausgedrückt: Aus der Sicht eines konkreten Kreditnehmers ist die Ausfallwahrscheinlichkeit eine Funktion der anderen Kunden, mit denen er oder sie vom Kreditbewerter zusammengeworfen wird. Nach dem Motto „mitgefangen, mitgehangen“ ist die Ausfallwahrscheinlichkeit allein eine Eigenschaft eines real existierenden *Kollektivs*. Und je nachdem, wie sich das Kollektiv zusammensetzt, ist sie mal größer und mal kleiner.

Der Traum eines jeden Kreditbewerters ist dann natürlich eine Prognose mit nur zwei Schubladen – in die eine kommen alle Kreditbewerber hinein, die später ausfallen, in die andere alle die, die ihren Verbindlichkeiten nachkommen. Das ist in der Praxis aber kaum jemals zu erreichen. Die vorliegende Arbeit widmet sich nun der Frage, wie man entscheidet, welcher Prognostiker bzw. Kreditbewerter bzw. welches Scorekartensystem diesem Ideal am nächsten kommt.

Aus der Sicht der mathematischen Statistik werfen dergleichen Wahrscheinlichkeitsprognosen natürlich noch eine ganze Reihe weiterer Fragen auf. Etwa wie man sie überhaupt produziert. Hier konkurrieren logistische Regressionsmodelle mit neuronalen Netzen, Stützvektormethoden, Baumverfahren oder Diskriminanzanalysen, ohne dass das letzte Wort bereits gesprochen ist. Einen ersten und immer noch aktuellen Überblick über diese inzwischen riesige Literatur geben etwa Rosenberg und Gleit (1994); die damit zusammenhängenden Fragen bleiben aber im Weiteren außer Betracht. Ebenfalls außer Betracht bleibt im Weiteren das für die Praxis geradezu entscheidende Problem, ob und wie sich das Ausfallverhalten der Population, die der Konstruktion der Prognose zugrundelag („Lernstichprobe“) von dem Ausfallverhalten der Anwendungspopulation unterscheidet. Wir konzentrieren uns hier vielmehr auf die Frage, wie gut eine Wahrscheinlichkeitsprognose die Ausfälle von den Nichtausfällen trennt, unabhängig davon, wie die Prognose zustandgekommen ist, und welche von mehreren konkurrierenden Wahrscheinlichkeitsprognosen in einem geeigneten, noch zu diskutierenden Sinn die beste ist.

Dabei ist immer zu bedenken, dass dergleichen Wahrscheinlichkeitsprognosen in aller Regel kein Selbstzweck, sondern nur Mittel zu einer darauf aufbauenden Entscheidungsfindung sind, etwa: bekommt der Kunde einen Kredit oder nicht? Und die letztendlich hier interessierende Frage ist: welche Wahrscheinlichkeitsprognose kann

diese Entscheidung am besten unterstützen? Auch darauf geht die vorliegende Untersuchung abschließend etwas ein.

2. Qualitätskriterien für Wahrscheinlichkeitsprognosen

Qualitätskriterien für Wahrscheinlichkeitsprognosen kann man in drei Gruppen unterteilen:

(i) Halbordnungen von Wahrscheinlichkeitsprognosen. Beispiel: Prognose A ist „besser“ als Prognose B, wenn ihre Lorenzkurve (dazu weiter unten mehr) überall oberhalb der von B verläuft (oder zumindest niemals unterhalb). Dann kann man sich in der Regel alle weiteren Vergleiche sparen.

(ii) Skalarwertige Qualitätskriterien. Die sind insbesondere immer dann gefragt, wenn sich zwei Prognosen im Sinn der in (i) diskutierten Halbordnungen nicht vergleichen lassen, etwa wenn sich ihre Lorenzkurven schneiden. Die in der Praxis am weitesten verbreiteten Beispiele sind hier der Gini-Koeffizient, die Fläche unter der ROC-Kurve oder der Brier-Score. Auch dazu weiter unten mehr.

(iii) Abstandsmaße. Wie weit liegen die vorhergesagten Ausfallwahrscheinlichkeiten der Ausfälle und der Nichtausfälle auseinander? Hier kommt es weniger auf korrekte Wahrscheinlichkeitsprognosen an, sondern es wird nur gefragt: haben die Ausfälle eine höhere vorhergesagte Ausfallwahrscheinlichkeit (im Kontext von Konsumentenkrediten: einen schlechteren „Score“ in dem benutzen Scorekartensystem) als die nicht ausgefallenen? Und an dieser Frage sieht man auch schon, dass sich die Ziele „Gute Wahrscheinlichkeitsprognose“ und „Hohe Trennschärfe“ nicht notwendig decken, zumindest nicht auf den ersten Blick. Diese Problematik soll daher am Anfang der vorliegenden Untersuchung stehen, unter Wiederholung einiger Argumente aus Krämer (2003).

Trennschärfe versus Kalibrierung

Angenommen, 10% aller Kredite eines bestimmten Portfolios fallen binnen eines Jahres aus (wobei hier nicht zu interessieren braucht, was das genau bedeutet). Eine Prognose A versieht jeden davon mit dem Etikett "Ausfallwahrscheinlichkeit 10%". Diese Prognose ist "kalibriert"[3] (synonym auch "valide" = valid oder "zuverlässig" = reliable, siehe Sanders 1963 oder Murphy 1973). Kalibriert bedeutet: Unter allen Krediten mit dem Etikett "Ausfallwahrscheinlichkeit x%" fallen x% tatsächlich aus. Trotzdem ist diese Bewertung wertlos – sie liefert keine neuen Informationen, das alles hat man vorher schon gewußt. Oder anders ausgedrückt: Kalibrierung ist allenfalls eine notwendige, aber keine hinreichende Bedingung für eine "gute" Wahrscheinlichkeitsprognose.

[3] Der Ausdruck „kalibrieren“ wird in der Literatur in mehrfacher Bedeutung verwendet. Zuweilen meint man damit auch nur „Schätzen von Ausfallwahrscheinlichkeiten für bestimmte Klassen von Forderungen“.

Prognose B teilt das Portfolio in zwei Gruppen auf: die erste mit Ausfallwahrscheinlichkeit 5%, die zweite mit Ausfallwahrscheinlichkeit 15%. Auch diese Bewertung sei kalibriert: In der ersten Gruppe fallen tatsächlich 5%, in der zweiten 15% der Kredite aus. Dann ist Prognose B ganz offensichtlich "besser" als Prognose A. In der Literatur wird das auch "trennschärfer" genannt (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%. Die trennschärfste Wahrscheinlichkeitsprognose läßt nur zwei Aus-sagen 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; sie sagt jeden Kreditausfall mit Sicherheit exakt voraus.

Halbordnungen zwischen Wahrscheinlichkeitsprognosen

Auch bei kalibrierten, aber nicht maximal trennscharfen Prognosen ist es sinnvoll, nachzufragen: Welche von mehreren kalibrierten Prognosen kommt dem Ideal einer maximal trennscharfen Prognose am nächsten? In obigem Beispiel ist Prognose B trennschärfer als A. Und nochmals trennschärfer sind zwei Prognosen C und D, welche die Kredite in die Ausfallklassen 2,5%, 7,5% und 22,5% bzw. 2,5%, 5% und 15% aufteilen. Und am trennschärfsten ist natürlich eine Prognose E, die alle Ausfälle exakt voraussagt. Tabelle 1 (angelehnt an Krämer 2003) zeigt eine mit Kalibrierung verträgliche Verteilung der Kredite auf die verschiedenen Ausfallklassen in diesen fünf Prognosesystemen.

Tabelle 1. Prognostizierte Ausfallwahrscheinlichkeiten und ihre Verteilung auf die Gesamtzahl der Kredite

Ausfallwahrscheinlichkeit	Verteilung der Kreditbewerber auf die verschiedenen Ausfallklassen (Anteile)				
	A	**B**	**C**	**D**	**E**
0%	0	0	0	0	0,9
2,5%	0	0	0,25	0,2	0
5%	0	0,5	0	0,25	0
7,5%	0	0	0,5	0	0
10%	1	0	0	0	0
15%	0	0,5	0	0,55	0
22,5%	0	0	0,25	0	0
100%	0	0	0	0	0,1

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 10%. Aber auch die B-Prognose läßt

sich ihrerseits aus der C-Prognose ableiten: Alle Kredite mit der C-Prognose 2,5% und eine zufällig ausgewählte Hälfte aller Kredite mit der C-Prognose 7,5% erhalten das Etikett 5%, die übrigen das Etikett 15%. 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 2,5% und 5% sowie ein zufällig ausgewähltes Elftel der D-Prognosen 15% erhalten das Etikett 5%, die übrigen das Etikett 15%. Das Ergebnis ist wieder eine kalibrierte Prognose mit der gleichen Trennschärfe wie B.

Die Prognosen C und D lassen sich jedoch 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.

Unabhängig von Trennschärfe und Kalibrierung ist es sinnvoll, beim Vergleich zweier Prognosen A und B zu fragen: "Welche von beiden gibt den ausgefallenen Krediten die höheren a-priori-Ausfallwahrscheinlichkeiten?" Diese Frage führt zum Begriff der "Ausfalldominanz" (Vardeman und Meeden 1983): Eine Prognose A ist besser als eine Prognose B im Sinne der Ausfalldominanz, falls A die ausgefallenen Kredite systematisch schlechter einstuft als B.

Formal: Sei $q_A(p_i|1)$ der Anteil der ausgefallenen Kredite, die von Prognose A die prognostizierte Ausfallwahrscheinlichkeit p_i (i=1,...,k) erhalten. Analog $q_B(p_i|1)$ usw. Dann ist A besser als B im Sinne der Ausfalldominanz, falls

$$\sum_{i=1}^{j} q_A(p_i \mid 1) \le \sum_{i=1}^{j} q_B(p_i \mid 1) \quad \text{für alle } j = 1,...,k.$$

Für kalibrierte Scorekarten errechnen sich die $q_A(p_i| 1)$ durch

$$q_A(p_i \mid 1) = \frac{p_i \times q_A(p_i)}{p}$$

mit p als Gesamtausfallwahrscheinlichkeit für alle Kredite insgesamt.

Analog läßt sich auch in Bezug auf die *nicht* ausgefallenen Kredite fragen, ob eine von zwei zu vergleichenden Scorekarten diese systematisch mit niedrigeren a-priori-Ausfallwahrscheinlichkeiten versieht. Sei dazu $q_A(p_i \mid 0)$ der Anteil der *nicht* ausgefallenen Kredite, die von Prognose A in die verschiedenen Scoreklassen $p_i (i = 1,..., K)$ eingeordnet worden sind. Analog $q_B(p_i \mid 1)$. Dann ist A besser als B im Sinn der Nichtausfall-Dominanz, falls

$$\sum_{i=1}^{j} q_A(p_i \mid 0) \ge \sum_{i=1}^{j} q_B(p_i \mid 0) \text{ für alle } j = 1,...,k.$$

Für kalibrierte Wahrscheinlichkeitsprognosen errechnen sich die $q_A(p_i \mid 0)$ als

$$q_A(p_i \mid 0) = \frac{(1-p_i)\times q_A(p_i)}{1-p} \ .$$

In der Sprache der Mathematik handelt es sich hier um einen Vergleich von Wahrscheinlichkeitsverteilungen über Ausfallwahrscheinlichkeiten. Prognose A ist in dieser Sprache besser als Prognose B im Sinne der Ausfalldominanz, wenn die bedingte Verteilung von A, gegeben Ausfall, diejenige von B stochastisch dominiert. Und A ist besser als B im Sinne der Nichtausfall-Dominanz, wenn die bedingte Verteilung von B, gegeben kein Ausfall, diejenige von A stochastisch dominiert.

Ein weiteres, von Kalibrierung unabhängiges und in der Praxis sehr populäres Qualitätskriterium (siehe z. B. Sobehard und Keenan 2001) gründet sich auf den Polygonzug durch die Punkte

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

Diese Kurve heißt in der angelsächsischen Literatur auch *power curve, cumulated accuracy profile* oder *Gini curve* und im deutschen Sprachgebiet meist "Lorenzkurve"[4]. Eine Wahrscheinlichkeitsprognose A ist dann besser als eine Wahrscheinlichkeitsprognose B in diesem Sinne, wenn A's Lorenzkurve immer oberhalb (präziser: niemals unterhalb) der von B verläuft.

Eine Prognose mit gleichen prozentualen Ausfallanteilen für alle vorhergesagten Ausfallwahrscheinlichkeiten hätte als Lorenzkurve die Diagonale. Diese Prognose liefert keine Informationen und ist in diesem Sinne die schlechtest mögliche. Zuweilen nennt man die so erzeugten Wahrscheinlichkeitsprognosen auch „uninformativ".

Das folgende Zahlenbeispiel möge diese Zusammenhänge verdeutlichen: Angenommen, zur Tabelle 1 gehören insgesamt 800 Kredite, davon 10% = 80 schlecht. Betrachten wir einmal die Lorenzkurve von Karte D. Diese prognostiziert für 160 Kredite eine Ausfallwahrscheinlichkeit von 2,5% (=guter Score), für 200 eine Ausfallwahrscheinlichkeit von 5% (=mittlerer Score), und für 440 eine Ausfallwahrscheinlichkeit von 15% (schlechter Score. Karte D ist kalibriert, d.h. in der ersten Gruppe fallen im Mittel 4 Kredite (= 2,5% von 160) tatsächlich aus, in der zweiten Gruppe fallen 10 Kredite aus (=5% von 200), in der dritten Gruppe 66 (= 15% von 440). Insgesamt gibt es 80 Ausfälle (10% von 800). Gruppiert man die Kredite nach ihren Scores von schlecht nach gut, und stellt ihnen die kumulierten Anteile der Ausfälle an den Ausfällen insgesamt gegenüber, ergibt sich Tabelle 2:

[4] Nach dem amerikanischen Statistiker Max O. Lorenz (1880-1962). Ursprünglich hatte Lorenz seine Kurve zur Beschreibung von Einkommens- und Vermögensungleich-heiten vorgesehen. Diese Lorenzkurven fangen mit den ärmsten an („die 10% Ärmsten haben 1% des Gesamteinkommens") und sind anders als die hier interessierenden. Lorenzkurven nach unten gebeugt. Zuweilen findet man auch die Schreibweise Lorentzkurve, aber die ist falsch.

Tabelle 2: Ausfälle versus Gesamtzahl der Kredite

Score	Kumulierter Anteil an Gesamtzahl der bewerteten Kredite	Kumulierter Anteil der Ausfälle an der Gesamtzahl der Ausfälle
schlecht	55 %	66/80 = 82,5%
Mittel	80 %	76/80 = 95%
Gut	100 %	80/80 = 100%

Diese Punkte, in ein 2-dimensionales Koordinatensystem übertragen und durch Geraden verbunden, erzeugen die in Abbildung 1 wiedergegebene Lorenzkurve der Prognose D. Ebenfalls eingezeichnet ist die optimale Lorenzkurve einer Prognose, die alle 80 Ausfälle, und nur diese, in die schlechteste Bonitätsklasse aufgenommen hätte. Diese unterscheidet sich von der Lorenzkurve der Prognose C um die Fläche B.

Die Lorenzkurve ist invariant gegenüber monotonen Transformationen der Ausfallwahrscheinlichkeiten. Werden alle prognostizierten Ausfallwahrscheinlichkeiten verdoppelt, ist die Prognose zwar nicht mehr kalibriert, aber die Lorenzkurve bleibt die gleiche. Da die Lorenzkurve sich nur darum kümmert, welcher Prozentsatz der Ausfälle auf die so-und-soviel Prozent schlechtesten Scores entfällt, nicht aber darum, welche prognostizierten Ausfallwahrscheinlichkeiten zu diesen Scores gehören, ist Kalibrierung für die Lorenzkurve irrelevant.

Die Lorenzkurve entdeckt jedoch, ob die tatsächlichen Ausfallwahrscheinlichkeiten mit höheren prognostizierten Ausfallwahrscheinlichkeiten (=schlechteren Scores) zunehmen: In diesem Fall ist sie konvex, und die zugehörige Prognose heißt auch „semikalibriert“.

Ein enger Verwandter der Lorenzkurve ist die ROC-Kurve (für „Receiver Operating Characteristic“[5]). Die Lorenzkurve trägt – beginnend mit den schlechten Scores – den Anteil der Schlechten in der Gruppe an allen Schlechten insgesamt gegen den Anteil aller Kredite ab. Die ROC-Kurve trägt – ebenfalls beginnend mit den Schlechten – den Anteil der Schlechten gegen den Anteil der Guten ab. Das dehnt die Lorenzkurve sozusagen nach links: Die Ordinatenwerte für die Stützpunkte bleiben gleich, die Abszissenwerte ändern sich. Bei der Lorenzkurve der Karte D aus Abbildung 1 entfallen etwa auf die schlechtesten 55% der Scores 66/80 = 82,5% aller Ausfälle, und auf die schlechtesten 80% der Scores entfallen 76/80 = 95% der Ausfälle. Bei der ROC-Kurve überlegt man anders: In der schlechtesten Gruppe sind 374 von insgesamt 720 Guten, das sind 51,94%, und wie gehabt 82,5% aller Schlechten. Damit verschiebt sich der Abszissenwert des ersten Stützpunkts nach links, von 0,55 nach 0,5194. In den beiden schlechtesten Gruppen zusammen sind 530 von insgesamt 720 Guten, das sind 73.61%

[5] Der Name kommt aus der Nachrichtentechnik und wird dort seit den 50er Jahren verwendet, um die falsch positiven gegen die richtig positiven Signale eines Empfängers (receivers) abzutragen. Die Kurve, die man durch Verschieben des Schwellenwertes erhält, „charakterisiert“ den Empfänger, daher „receiver operating characteristic“.

und wie gehabt 95% aller Schlechten. Damit verschiebt sich der Abszissenwert des zweiten Stützpunkts ebenfalls nach links, und zwar von 0,8 nach 0,7361.

Abbildung 2 zeigt die resultierende ROC-Kurve. Eine Prognose A ist dann besser als eine Prognose B im Sinne dieses Kriteriums, wenn A's ROC-Kurve immer oberhalb der von B verläuft.

Skalarwertige Qualitätskriterien

Die bisher vorgestellten Methoden erlauben Aussagen der Art: „Prognose A ist besser als Prognose B". Das hat den Vorteil, dass man keine konkreten Qualitätskriterien berechnen muß. Es läßt sich sogar zeigen (siehe etwa Krämer 2006), dass eine Dominanz etwa bezüglich des Trennschärfekriteriums zur Folge hat, dass die in diesem Sinne besseren Prognosen auch besser sind, ganz gleich welches sonstige skalarwertige Qualitätskriterium aus einer großen Klasse von sinnvollen Kriterien man auch benutzt.

Skalarwertige Qualitätsmaße lassen sich unterteilen in solche, die von der ROC- oder Lorenzkurve, und solche, die von den vorhergesagten Wahrscheinlichkeiten abhängen. Das in der Kapitalmarktpraxis beliebteste Maß aus der ersten Klasse ist das Verhältnis der Fläche A zur Fläche A+B; es heißt auch "Trefferquote" ("accuracy ratio") oder „Gini-Koeffizient" (nach Corrado Gini (1884-1965), der sich große Verdienste in der Ungleichheitsmessung erworben hat). Je höher der Gini-Koeffizient, desto näher kommt eine Prognose an die in obigem Sinn optimale Prognose heran. Der maximale Wert ist 1; er wird nur erreicht für eine Fläche B=0 und bedeutet: Alle Schlechten, und nur diese, sind in der schlechtesten Scoreklasse vertreten, die Prognose liefert eine perfekte Trennung der nicht ausgefallenen (=guten) und der ausgefallenen (=schlechten) Kredite.

In medizinische Anwendungen, etwa bei der Prognose von Therapieerfolgen, verwendet man dagegen lieber die Fläche unterhalb der ROC-Kurve.[6] Dieses Kriterium heißt auch „Area under ROC" (AUROC). Sowohl die Lorenz- als auch die ROC-Kurve sind umso besser, je weiter sie sich von der Diagonalen nach oben wegbiegen. Der Gini-Koeffizient und die AUROC quantifizieren dieses Wegbiegen durch Flächen. Alternativ könnte man auch den maximalen vertikalen Abstand oder die Länge der Kurven nehmen. In der ökonometrischen Ungleichheitsmessung sind diese Koeffizienten durchaus üblich und auch mit eigenen Namen versehen, siehe Krämer (1998). Im Kontext von Kreditbewertungen kommt vor allem der maximale vertikale Abstand zwischen der ROC-Kurve und der Diagonalen vor, in der angelsächsischen Literatur auch „ROC-gap" genannt.

Unter den Kriterien, die auf den vorhergesagten Ausfallwahrscheinlichkeiten beruhen, ist der „Brier-Score" (nach G. W. Brier 1950) in der Praxis bei weitem der populärste. Sei p^j die vorhergesagte Ausfallwahrscheinlichkeit für Kredit Nr. j (aus insgesamt n zu

[6] Oder auch nur die Fläche zwischen ROC-Kurve und Diagonale; die Literatur ist sich hier nicht einig.

bewertenden Krediten), und sei $\theta^j = 1$ bei Ausfall und $\theta^j = 0$, wenn kein Ausfall eintritt. Dann ist dar Brier-Score definiert als[7]

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

Es wurde und wird bislang vor allem zum Qualitätsvergleich von Wettervorhersagen eingesetzt, ist aber grundsätzlich in allen Kontexten einsetzbar, in denen Wahrscheinlichkeitsprognosen zu vergleichen sind.

Der Brier-Score liegt immer zwischen 0 und 1. Je kleiner der Brier-Score, desto besser die Wahrscheinlichkeitsprognose. Der bestmögliche Wert von 0 ergibt sich für Prognosen von immer nur 0% oder 100% für Ausfall, bei denen stets das Vorhergesagte tatsächlich eintritt. Der schlechtestmögliche 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. Verwandte Maße sind die Mittlere Logarithmische Abweichung (Good 1952)

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

oder der sog. „Sphärische Score"

$$S = \frac{1}{n} \sum_{j=1}^{n} - \frac{\left| p^j - \theta^j - 1 \right|}{\sqrt{p^{j2} + \left(1 - p^j\right)^2}} \ .$$

Die Logik hinter diesen Maßen steht hier nicht zur Debatte; bei Winkler (1994, 1996) findet man eine ausführliche Übersicht der Hintergründe dieser und weiterer skalarwertiger, auf vorhergesagten Ausfallwahrscheinlichkeiten beruhender Qualitätskriterien.

Abstandsmaße

Abstandsmaße wollen quantifizieren, wie weit die Verteilungen der prognostizieren Ausfallwahrscheinlichkeiten der Guten und der Schlechten auseinander liegen. Das bekannteste derartige Abstandmaß ist die sogenannte „Divergenz“, definiert als

$$D = 2 \frac{(m_G - m_s)^2}{s_G^2 + s_s^2}.$$

Dabei ist m_g die mittlere prognostizierte Ausfallwahrscheinlichkeit der guten Kredite, m_s die mittlere prognostizierte Ausfallwahrscheinlichkeit der schlechten Kredite, und s_g^2 bzw. s_s^2 sind die entsprechenden empirischen Varianzen. Diese Größe ist invariant gegen Lineartransformationen der vorhergesagten Ausfallwahrscheinlichkeiten, d.h.

[7] Zuweilen wird in der Literatur auch das Negative dieses Ausdrucks als Brier-Score bezeichnet.

man kann diese auch o.B.d.A auch durch die in der Praxis meist benutzen Scores ersetzen, die in aller Regel durch geeignete Lineartransformationen der vorhergesagten Ausfallwahrscheinlichkeiten entstehen.

Alternativ kann man auch die kompletten Verteilungen (sei es die der Scores, sei es die der vorhergesagten Ausfallwahrscheinlichkeiten) betrachten, so wie in Abb. 3, und überprüfen, wie sehr sich diese überschneiden. Die schraffierte Fläche („Überschneidungsfläche“) wäre dann ein weiteres Maß für den Abstand der Verteilungen. Sie liegt immer zwischen 0 und 1; je größer diese Fläche, desto schlechter werden gute und schlechte Kredite getrennt, je kleiner diese Fläche, desto besser werden sie getrennt. Und schließlich kann man auch noch mit formalen statistischen Testverfahren überprüfen, ob die beiden Verteilungen identisch sind – je deutlicher der Test ablehnt (je „signifikanter“ die jeweilige Teststatistik), desto weiter sind dann die Verteilungen voneinander entfernt. Damit liefert ebendiese Teststatistik ein bequemes Maß für den Abstand der Verteilungen. In der Praxis findet man dafür vor allem den Zwei-Stichproben t-Test auf Gleichheit der Erwartungswerte, den nichtparametrischen Mann-Whitney U-Test auf Identität der kompletten Verteilungen, den nichtparametrischen Wilcoxon Rangsummentest auf Identität der kompletten Verteilungen und den nichtparametrischen Kolmogoroff-Smirnoff-Test auf Identität der kompletten Verteilungen.

Die Prüfgröße des Kolmogoroff-Smirnoff-Tests ist der maximale Abstand der empirischen Verteilungsfunktionen der Scores der Guten und der Schlechten, multipliziert mit einem Proportionalitätsfaktor, der von der Anzahl der Guten und der Schlechten in der Population abhängt. Die Prüfgröße des Wilcoxon Rangsummentests ist die Summe der Ränge der Schlechten (alternativ auch der Guten), die man erhält, wenn man alle Kredite, Gute wie Schlechte gleichermaßen, ihrem Score entsprechend aneinanderreiht. Die Prüfgröße des Mann-Whitney-U-Tests erhält man durch paarweisen Vergleich aller Schlechten mit allen Guten, und ist dann einfach der Anteil der Paare, in denen der Gute einen besseren Score aufweist als der Schlechte.[8] Bei perfekter Trennung ist dieser Anteil 100%.

3. Beziehungen zwischen den Qualitätskriterien

Die Fülle der hier vorgestellten Qualitätskriterien scheint in der Praxis ein Dilemma aufzuwerfen: Auf welches soll man nun vertrauen? Zum Glück liefern aber viele Kriterien die gleichen Informationen, nur anders verpackt. Dieser Abschnitt stellt die wichtigsten einschlägigen Resultate aus der Literatur zusammen. So ist etwa die Divergenz numerisch identisch zur quadrierten Prüfgröße des Zwei-Stichproben t-Tests auf identische Erwartungswerte bei normalverteilten Grundgesamtheiten (Hartung 1987, S. 510), und sind der Kolmogoroff-Smirnoff-Test, die Überschneidungsfläche und das ROC-Gap monotone Funktionen voneinander.

[8] Ein gleicher Score zählt zur Hälfte.

Abb. 4 möge dies illustrieren. Die Abbildung zeigt exemplarisch die Verteilungsfunktionen der Scores der Guten (=G(s)) und Schlechten (=S(s)) für eine ausgewählten Population von Kreditnehmern. Der Kolmogoroff-Smirnoff-Test gründet sich auf den maximalen Abstand dieser Verteilungsfunktionen (Teil a); die konkrete Prüfgröße ist dieser maximale Abstand, multipliziert mit dem Faktor $(Np(1-p))^{1/2}$ (siehe etwa Hartung 1987, S. 521). Dabei ist p der Schlechtanteil und N der Umfang der betrachteten Kreditnehmerpopulation. Der Abstand der Verteilungsfunktionen ist maximal, wenn die beiden Steigungen S'(s) und G'(s) übereinstimmen. Der Quotient dieser Steigungen ist aber gerade die Steigung der ROC-Kurve an der Stelle G(s).[9] Mit anderen Worten, für S'(s) = G'(s) hat die ROC-Kurve die Steigung 1, und genau an dieser Stelle ist der Abstand der ROC-Kurve zur Diagonalen maximal (Teil b).

Analog ist der Zusammenhang zwischen ROC-Gap und Überschneidungsfläche nachzuweisen (Teil c): Das ROC-Gap ist gerade S(s) – G(s), evaluiert an der Stelle S'(s) = G'(s) (:=s*). Aber S(s*) und G(s*) sind ja als Intergrale von S' und G' die Flächen unterhalb S' und G', von links bis zur Stelle s*, und je größer deren Differenz, desto kleiner die Überschneidungsfläche. Das alles kann man auch bei Hoadley und Oliver (1998) in mathematisch ausführlicherer Formulierung finden. Damit ist klar, dass der Kolmogoroff-Smirnoff-Test, die Überschneidungsfläche und das ROC-Gap im wesentlichen die gleichen Informationen liefern. Optimiert man eines der drei Kriterien, dann optimiert man auch die beiden anderen.

Genauso sind auch der Gini-Koeffizient, die AUROC und die Prüfgrößen des Mann-Whitney U-Tests und des Wilcoxon Rangsummentests auf Gleichheit der Verteilungen monotone Funktionen voreinander. Auch hier gilt also: optimiert man eines der Kriterien, dann optimiert man simultan auch alle anderen. Die Prüfgröße des U-Tests ist sogar numerisch identisch zur AUROC, und diese wiederum bestimmt den Gini-Koeffizienten durch die Beziehung

$$\text{Gini-Koeffizient} = 2 \times (\text{AUROC} - 0{,}5) .$$

Diese Zusammenhänge sind nicht ohne weiteres grafisch zu vermitteln; ausführliche Beweise sind etwa in Hanley und McNeil (1982) oder Engelmann et al. (2003) zu finden.

Ähnliche Zusammenhänge gibt es auch bei den Halbordnungen. Zum Beispiel ist es aus der Definition der Lorenzkurve und der ROC-Kurve klar, dass eine Scorekarte A im Sinn der Lorenzordnung genau dann besser ist als eine Scorekarte B, wenn sie auch im Sinn der ROC-Ordnung besser ist. Oder anders ausgedrückt: zwei ROC-Kurven schneiden sich genau dann, wenn sich auch die zugehörigen Lorenzkurven schneiden.[10] Für

---

[9] Das sieht man intuitiv wie folgt: Die Steigung der ROC-Kurve an der Stelle G ist der Grenzwert, für $\Delta G \to 0$, von $\Delta S/\Delta G$. Erweitere diesen Quotienten um Δs: $(\Delta S/\Delta s)\ /(\Delta G/\Delta s)$. Dann erhält man für $\Delta s \to 0$ gerade $S'(s)/G'(s)$. Hier habe ich, wie auch im weiteren, implizit die Differenzierbarkeit aller relevanten Funktionen unterstellt. Das ist bei empirischen ROC-und Verteilungskurven nicht überall der Fall. Dann approximiert man die empirischen Funktionen durch monotone differenzierbare Funktionen.

[10] Das folgt sofort daraus, dass ROC-Kurven nichts anderes als nach links verschobene Lorenzkurven sind.

semikalibrierte Scorekarten A und B gilt weiterhin: (i) Ist A besser als B im Sinn der Ausfalldominanz, dann ist A auch trennschärfer als B, und (ii) Ist A trennschärfer als B, so liegen sowohl die Lorenzkurve als auch die ROC-Kurve von A oberhalb der von B. Die Beweise für diese Zusammenhänge sind bei Vardemann und Meeden (1983) und Krämer (2005) nachzulesen.

Die Umkehrung dieser Implikationen gilt nicht. Mit anderen Worten, es kann vorkommen, dass eine Ausfallprognose A eine bessere Lorenzkurve hat als B, aber nicht trennschärfer ist als B. Für Einzelheiten siehe Krämer (2005).

4. Welches Kriterium für welchen Zweck?

Gewisse der hier vorgestellten Kriterien, etwa alle Halbordnungen von Wahrscheinlichkeitsprognosen, sind nur dann sinnvoll, wenn die zu vergleichenden Ausfallprognosen auf identische Grundgesamtheiten von Kreditbewerbern angewendet werden, und, noch viel wichtiger: wenn man *für alle* Schwellenwerte, nicht nur für den aktuell in der Praxis ausgewählten, die Guten und die Schlechten kennt (oder zumindest indirekt erschließen kann; diese „reject inference“ ist ein Thema für sich und würde den Rahmen dieser Arbeit sprengen). Dieser Gesichtspunkt ist vor allem für die allein auf historischen Ausfalldaten beruhende Bewertung existierender Scorekarten von zentraler Bedeutung („performance assessment“ oder „monitoring“, siehe Hand 2005). Hier hat man in aller Regel nur einen einzigen, gegebenen, festen Schwellenwert; Kunden darüber erhalten einen Kredit, Kunden darunter nicht. Für letztere ist also unbekannt, ob sie zu den Guten oder zu den Schlechten gehören. Gründet man jetzt den Vergleich zweier Scorekarten nur auf den Scoreverteilungen von Kunden mit Score über (dem für beide Prognosen gleichen) Schwellenwert, können sich Resultate ergeben wie etwa dieses: Prognose A hat einen besseren Gini-Koeffizienten als Prognose B, aber trotzdem ist für Prognose B der Schlechtanteil geringer. Das kann etwa so geschehen, dass A zwar mehr Schlechte über Schwellwert hat als B, dass sich aber bei A die Scores der Schlechten knapp über dem Schwellenwert konzentrieren, und das ist günstig für den Gini-Koeffizienten: ein hoher Schlechtanteil bei kleinen Scores.

Technisch gesehen handelt es ich hier um *bedingte* Scoreverteilungen der Guten und der Schlechten, gegeben der Score übersteigt einen bestimmten Schwellenwert. Im weiteren unterstelle ich dagegen die unbedingten Verteilungen, und da kommen dergleichen Unstimmigkeiten nicht vor.

Von Bedeutung , allerdings von großer Bedeutung, für die praktische Eignung der hier vorgestellten Qualitätskriterien ist dann nur noch deren Abhängigkeit vom Gut-Schlecht-Verhältnis in der Gesamtpopulation aller untersuchten Kreditnehmer. Hängt das fragliche Kriterium ceteris paribus (d.h. bei gegebener Scoreverteilung der Guten und der Schlechten) von diesem Verhältnis ab? Wenn ja, dann ist es in der Praxis nur mit großer Vorsicht zu verwenden.

Hier zeigt sich ein großer Nachteil aller auf Wahrscheinlichkeitsprognosen abstellender Qualitätskriterien; deren Werte werden – zumindest bei kalibrierten Prognosen - mit kleiner werdendem Schlechtanteil immer besser, unabhängig von der Qualität der Prognose. Betrachten wir einmal das Zahlenbeispiel aus Tabelle für den Fall von 800 Kreditnehmern bei einem Schlechtanteil von 10%. Dann erhalten wir für den Brier-Score der Scorekarten A, B und C:

$$B_A = \frac{1}{800}\left[80(0{,}1-1)^2 + 720(0{,}1-0)^2\right] = 0{,}090$$

$$B_B = \frac{1}{800}\left[20(0{,}05-1)^2 + 380(0{,}05-0)^2 + 60(0{,}15-1)^2 + 340(0{,}15-0)^2\right] = 0{,}088$$

$$B_C = \frac{1}{800}\left[5(0{,}025-1)^2 + 195(0{,}025-0)^2 + 30\,(0{,}075-1)^2 + 370(0{,}075-0)^2 + 45(0{,}225-1)^2 + 155\,(0{,}225-0)^2\right] = 0{,}084$$

Mit anderen Worten, die Werte unterscheiden sich kaum und sind darüber hinaus auch alle sehr gut (beim Brier Score bedeutet klein=gut). Selbst die Trivialprognose A von „Ausfallwahrscheinlichkeit von 2% für alle Kredite gleichermaßen“ liefert noch eine Score von 0,09.

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

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

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: Selbst Trivialprognosen sind in diesem Sinn“sehr gut“. Es empfiehlt sich daher in den Anwendungen auf jeden Fall, einen realisierten Brier-Score relativ zu dem Trivialscore (*) zu sehen. Der Vergleich von Scorekarten, die auf verschiedene Populationen von Kreditnehmern angewandt werden, bleibt aber auf jeden Fall problematisch.

Als kleiner Vorteil steht dagegen, dass ein Anwender seinen subjektiv erwarteten Brier-Score immer dann minimiert, wenn er als Prognose für die Ausfallwahrscheinlichkeit seine wahre subjektive Ausfallwahrscheinlichkeit einsetzt (De Groot und Fienberg 1983). Insofern belohnt der Brier-Score also „ehrliches“ Verhalten. Abweichungsmaße mit dieser Eigenschaft heißen in der angelsächsischen Literatur auch „proper scoring rules“ (Winkler 1996). Ein deutscher Ausdruck dafür wäre „anreizkompatible Abweichungsmaße“. Auch die Mittlere logarithmische Abweichung ist anreizkompatibel. 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. Da aber die in der Praxis benutzten Scorekarten genau dieses menschliche Ermessen ausschalten, fällt dieser Vorteil bei maschinell entscheidenden Scorekarten nicht ins Gewicht.

Der entscheidende Nachteil von Qualitätsmaßen, die wie der Brier-Score unmittelbar auf Ausfallwahrscheinlichkeiten aufbauen, ist aber der: Es kann vorkommen, dass der Brier-Score sagt: eine Prognose ist besser als eine andere, aber die ROC-Kurve der „besseren" Prognose liegt überall unter der der „schlechteren". Nehmen wir die Prognose C aus Tabelle 1, mit 800 Kreditnehmern wie in Tabelle 2. Die Prognose unterscheidet drei Klassen von Kreditbewerbern, mit Ausfallwahrscheinlichkeiten von 2,5%, 7,5% und 22,5%, das liefert für diese Population einen Brier-Score von B_c = 0.084 (siehe oben). Das ist kleiner = besser als der Brier Score B_b = 0.088 von Prognose B. Wie man leicht nachprüft, liegt deren ROC-Kurve auch überall – außer natürlich an den Endpunkten - unter der von C, B ist also auch in diesem Sinne schlechter als C.

Jetzt definieren wir eine neue Prognose C* und transformieren die vorhergesagten Ausfallwahrscheinlichkeiten monoton: aus 2,5% wird 10%, aus 7,5% wird 15%, und aus 22,5% wird 30%. Diese Prognose alias Prognose ist natürlich nicht mehr kalibriert, aber die ROC Kurve von C* ist die gleiche wie die ROC-Kurve von C. Insbesondere ist C* in diesem Sinne also besser als B, die ROC-Kurve von C* liegt überall über der von B. Aber der Brier-Score von C* ist schlechter als der Brier-Score von B! Einfaches Ausrechnen ergibt B_{c^*} = 0.9, und das ist größer = schlechter als B_b = 0, 88. Die unmittelbar auf prognostizierten Ausfallwahrscheinlichkeiten für jeden Kreditnehmer basierenden Qualitätskriterien sind also für einen Qualitätsvergleich von Scorekarten nicht geeignet.

Wie steht es nun um die Abhängigkeit der auf der Lorenz- bzw ROC-Kurve basierenden Qualitätskriterien vom Gut-Schlecht-Verhältnis in der Population?

Es folgt unmittelbar aus der Definition, dass die ROC-Kurve nicht vom Gut-Schlecht Verhältnis in der Population abhängt. Damit sind auch alle auf der ROC-Kurve basierenden skalarwertigen Qualitätsmaße wie etwa AUROC oder ROC-gap ceteris paribus (d.h. bei gegebener Scoreverteilung für die Guten und die Schlechten) unabhängig vom Gut-Schlecht Verhältnis in der Population.

Wegen der Gleichung

$$\text{Gini-Koeffizient} = 2 \times (\text{AUROC} - 0{,}5)$$

gilt das auch für den Gini-Koeffizienten**.** Es gilt aber *nicht* für die Lorenzkurve. Deren Form hängt vielmehr bei gegebenen Score-Verteilungen für die Guten und die Schlechten auch noch vom Gut-Schlecht-Verhältnis in der Grundgesamtheit ab.

5. Berücksichtigung der Kosten

Die bisherigen Betrachtungen waren in gewisser Weise *global*: Sie hatten das ganze Spektrum der Kreditnehmer und Kreditnehmer-Scores zum Gegenstand, von den ganz guten bis zu den ganz schlechten, und berücksichtigen nicht, dass in konkreten Anwendungen immer bei einem ganz bestimmten Score s abgeschnitten wird: Kreditbewerber mit einem besseren Score bekommen einen Kredit, die anderen nicht. Was links oder rechts von diesem Schnittpunkt geschieht, ist weniger relevant, von Interesse sind allein die Konsequenzen, vor allem die Konsequenzen für Gewinne und Verluste, an einer bestimmten Stelle der Scoreverteilung, also was geschieht *lokal*?

Hier sind zwei Aspekte zu unterscheiden: Die Konsequenzen eines Fehlers 1. Art (eine Schlechter bekommt einen Kredit) und die Konsequenzen eines Fehlers 2. Art (ein Guter bekommt keinen Kredit).[11] Bei einem Fehler 1. Art fallen Verluste an, bei einem Fehler 2. Art entgehen Gewinne. Bei welchem Schwellenwert ist die Summe aus beiden minimal?

Eine korrekte Antwort ist natürlich nur möglich bei Kenntnis ebendieser Kosten und Gewinne. Die sind nicht für alle Kunden gleich, und in aller Regel auch a priori nict exakt bekannt, so dass je nach den Annahmen, die man in diese Analyse hineinsteckt, andere Antworten resultieren.

Vergleichweise einfach ist das ganze bei einem gegebenen Schwellenwert s. Abbildung 5 zeigt eine typische ROC-Kurve, die zu einer gegebenen Scorekarte gehört. Hohe Scores sind gut, G(s) ist wie gehabt die Verteilungsfunktion der Scores der Guten, S(s) die Verteilungsfunktion der Scores der Schlechten. Dann ist die ROC-Kurve gerade der Graph von S(s) auf der senkrechten Achse gegen G(s) auf der waagerechten Achse, für alle denkbaren Schwellenwerte s. Bei gegebenem Schwellenwert s und einem dadurch implizit definierten Anteil G(s) von abgelehnten Guten liegen ceteris paribus die Gewinne aus vergebenen Krediten fest. Das einzige, was dann je nach Sorekarte noch variiert, ist der Anteil 1-S(s) akzeptierter Schlechter. Daraus folgt sofort, dass bei einem gegebenem Schwellenwert s ist diejenige Scorekarte am besten ist, deren ROC-Kurve für den durch diesen Schwellenwert festgelegten Anteil G(s) abgelehnter Guter den größten Wert annimmt. Liegt also die ROC-Kurve einer Prognose A überall über der einer Prognose B, ist Prognose A auf jeden Fall profitabler als Prognose B, unabhängig von dem gewählten Schwellenwert. Und da die ROC-Kurve von A genau dann über der ROC-Kurve von B liegt, wenn die Lorenzkurve von A über der von B liegt, gilt gleiches auch für die Lorenzkurve.

Zur Wahl des Schwellenwertes gibt es in der Literatur eine Reihe von Vorschlägen, siehe etwa Hoadley und Oliver (1998), Stein (2005) oder Blöchlinger und Leippold (2006). Alle bauen auf gewissen Annahmen zu Gewinnen und Verlusten auf. Der einfachste Fall ist der:

---

[11] Diese Notation ist hier nicht immer einheitlich. Mit dem Fehler 1. Art ist in der Statistik üblicherweise die fälschliche Ablehnung einer korrekten Nullhypothese gemeint. Das wäre dann hier: Der Kunde ist schlecht. Ob das aber die richtige Nullhypothese ist, mit der man Geschäftspartnern begegnet, sei einmal dahingestellt.

V := durchschnittlicher Nettoverlust pro Schlechter ist bekannt und für alle Schwellenwerte gleich.

E := durchschnittliche Nettoeinnahme pro Guter ist bekannt und für alle Schwellenwerte gleich.

Sei nun wie gehabt p der Schlechtanteil in der ganzen Population. Diese habe den Umfang N. Dann gilt für einen gegebenen Schwellenwert s:

Gesamte Einnahmen durch Gute = $E \times (1-G(s)) \times (1-p) \times N$.

Gesamte Verluste durch Schlechte = $V \times (1-S(s)) \times p \times N$.

Bei kleinstmöglichem Schwellenwert – jeder bekommt einen Kredit - sind sowohl die gesamte Einnahmen durch Gute wie die gesamten Verluste durch Schlechte maximal. Mit wachsendem Schwellenwert s nehmen dann die vergebenen Kredite zusammen mit den Einnahmen und mit den Verlusten ab, jedoch fallen die Verluste zunächst schneller, der Gewinn nimmt zu. Der maximale Gesamtgewinn entsteht also für dejenigen Schwellenwert s, an dem die Verluste genauso schnell abnehmen wie die Gewinne, oder formal, wo gilt:

$$G'(s) \times E \times (1-p) = S'(s) \times V \times p$$

bzw.

$$S'(s)/G'(s) = (E \times(1-p)) / (V \times p).$$

Die linke Seite dieser letzten Gleichung entspricht aber genau der Steigung der ROC-Kurve an der Stelle s (siehe etwa Fußnote 8 oder Stein 2005). Daraus folgt sofort, dass sich bei schwellenwertunabhängigen durchschnittlichen Gewinnen und Verlusten pro Kredit eine Anhebung des Schwellenwertes (d.h. eine Einschränkung der Kreditvergabe) solange lohnt, wie die Steigung der ROC-Kurve den Quotienten $(E \times(1-p)) / (V \times p)$ übersteigt. Oder mit anderen Worten: Der optimale Schwellenwert s und der dazu gehörende Anteil G(s) abgelehnter Guter liegt umso weiter rechts, je kleiner der Quotient $(E \times(1-p)) / (V \times p)$. Daraus folgt sofort, dass der optimale Schwellenwert ceteris paribus umso höher ist (die Kriterien an vergebene Kredite sind umso strenger), je größer der Schlechtanteil in der Population und je höher die Verluste pro schlechtgewordenem Kredit. Oder nochmals anders ausgedrückt: Man ist bei der Kreditvergabe umso vorsichtiger, je mehr Schlechte es gibt und je höher die Verluste pro Schlechte ausfallen. Das weiß jeder Kreditbearbeiter zwar auch so, aber es ist doch immer wieder beruhigend, wenn die Wissenschaft den gesunden Menschenverstand bestätigt.

Nehmen wir als Zahlenbeispiel die Prognose D aus Tabelle 1, d.h. eine Scorekarte mit den Verteilungen der Scores der Guten und Schlechten wie in Abb. 3 und einer ROC-Kurve wie in Abb. 5. Angenommen, der Schlechtanteil beträgt 10%, der Verlust für jeden faulen Kredit im Durchschnitt 10.000 Euro, der Ertrag für jeden guten Kredit im Durchschnitt 1000 Euro. Dann gilt:

$$(E \times (1-p)) / (V \times p) / = 0{,}9$$

und es lohnt sich, die Kreditvergabe solange einzuschränken, bis eine Gerade mit dieser Steigung die ROC-Kurve tangiert, so wie in Abbildung 6. Es würden also nur Kreditbewerber aus der schlechtesten Ratingklasse abgelehnt. Und zwar solange, wie die Steigung der Geraden sich oberhalb eine Wertes von

$$(0{,}95 - 0{,}825) / (0{,}7361 - 0{,}5194) = 0{,}58.$$

befindet. Steigt nun der Schlechtanteil von 10% auf 15%, so sinkt auch die Steigung der Geraden, und zwar von 0,9 auf 0,57. Das ist kleiner als 0,58, also vergrößern sich jetzt die Gewinne, wenn auch Kreditnehmer aus der zweitschlechtesten Klasse abgewiesen werden.

Diese Überlegungen beruhen auf durchschnittlichen Verlusten und Erträgen, die nicht von den Scores der Kreditbewerber abhängen. Nun könnte man sich in der Praxis sehr gut Situationen vorstellen, wo hohe Scores auch hohe Erträge versprechen, und in dem Umfang, wie sich dann auch das *Verhältnis* von Erträgen und Verlusten ändert, wären die obigen Ableitungen dann entsprechend anzupassen. Das ist aber ein Thema für sich, und muß künftigen Untersuchungen vorbehalten bleiben.

Literatur

Blöchlinger, A. und Leippold, M. (2006): “Economic benefit of powerful credit scoring”, *Journal of Banking & Finance 30*, 851-873.

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

DeGroot, M. und Fienberg, S.E. (1983): "The comparison and evaluation of probability 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.

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

Friedman, Thomas L. (1996): Interview in der “News Hour with Jim Lehrer”, PBS-Televison, 13. Februar.

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

Hamerle, A., Rauhmeier, R., und Roesch, D. (2003): “Uses and misuses of measures for credit rating accuracy”, Unveröffentlichtes Arbeitspapier, Universität Regensburg.

Hand, D. J. (2005). “Good practice in retail credit scorecard assessment.” *Journal of the Operational Research Society* 56, 1109-1117.

Hanley, A. und McNeil, B. J. (1982): “The meaning and use of the area under a receiver operating characteristics (ROC) curve”, *Diagnostic Radiology 143*, 29-36.

Hartung, J. (1987): *Statistik*, 6. Auflage, München (Oldenbourg).

Hoadley, B. und Oliver, R. M. (1998): “Business measures of scorecard benefit”, *IMA Journal of Mathematics Applied in Business & Industry* 9, 55-64.

Krämer, W. (1998): "Measurement of inequality." in A. Ullah/D. Giles (Hrsg.): *Handbook of Applied Economic Statistics*, New York, 39-62.

Krämer, W. (2003): "Die Bewertung und der Vergleich von Kreditausfallprognosen." *Kredit & Kapital* 36, 395 – 410.

Krämer, W. (2005): "On the ordering of probability forecasts." *Sankhya – The Indian Journal of Statistics* 67, 662 - 669.

Krämer, W. (2006): “Evaluating probability forecasts in terms of refinement and strictly proper scoring rules,” *Journal of Forecasting* 25, 223 - 226.

Krämer, W. und Güttler, A. (2008): „On comparing the accuracy of default predictions in the rating industry.“ *Empirical Economics* 34, 343-356.

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

Rosenberg, E. und Gleit, A. (1994): „Quantitative methods in credit management: A Survey.“ *Operations Research* 42, 589-613.

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

Sobehart, J. und Keenan, S. (2001): “Measuring Default Accurately”, *Risk*, March, 31-33.

Stein, R. M. (2005): “The relationship between default prediction and lending profits: Integrating ROC analysis and loan pricing”, *Journal of Banking & Finance 29*, 1213-1236.

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. (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.

Abb. 1. Eine beispielhafte Lorenzkurve von Kreditausfällen

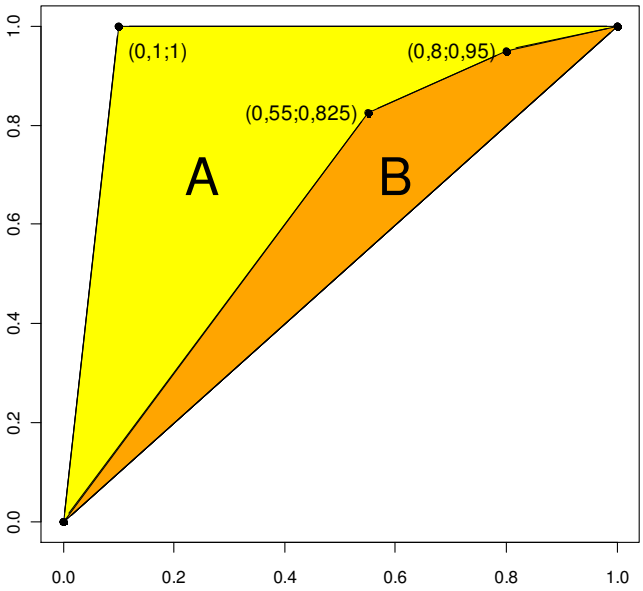


Abb. 2: Eine beispielhafte ROC-Kurve von Kreditausfällen

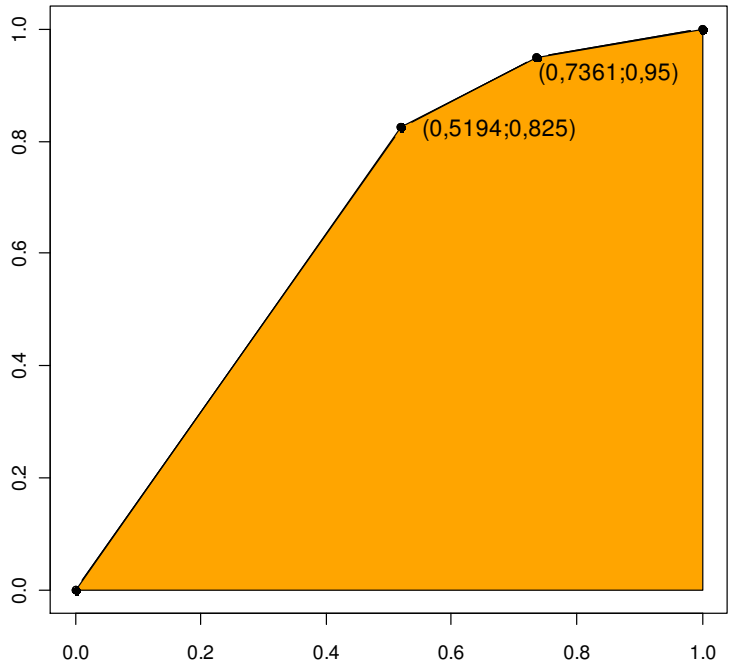


Abb. 3: Verteilungsvergleich der Guten und der Schlechten

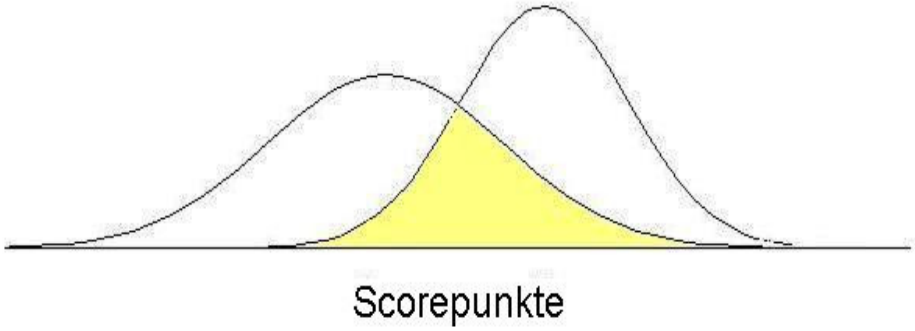
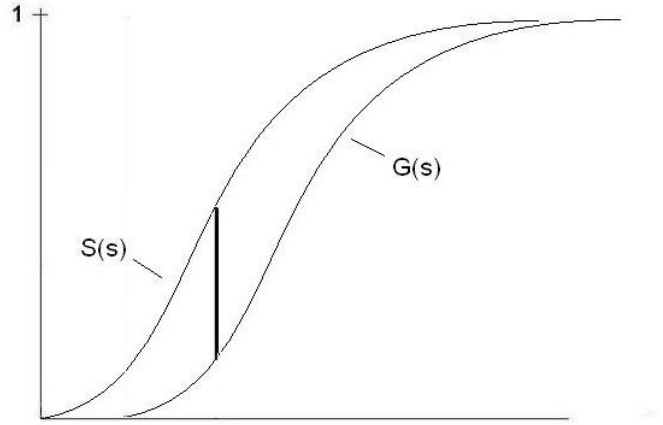
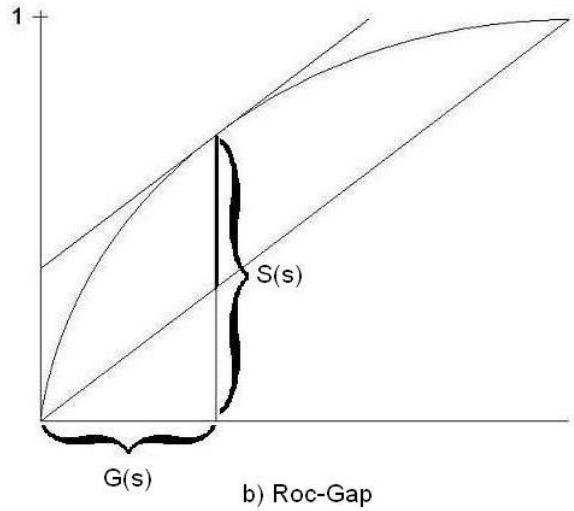


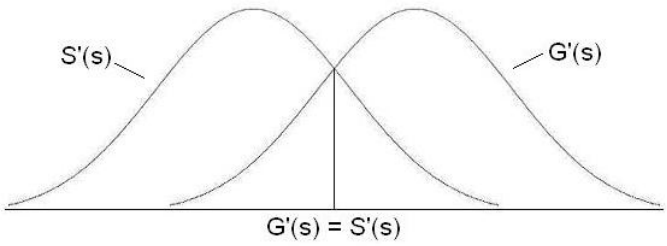
Abb. 4. Zusammenhang zwischen Verteilungsfunktionen, ROC-Gap und Überschneidungsfläche.



a) Kolmogoroff-Smirnoff



b) Roc-Gap



c) Überschneidungsflächen

Abb. 5: Typische ROC-Kurve für eine gegebene Scorekarte

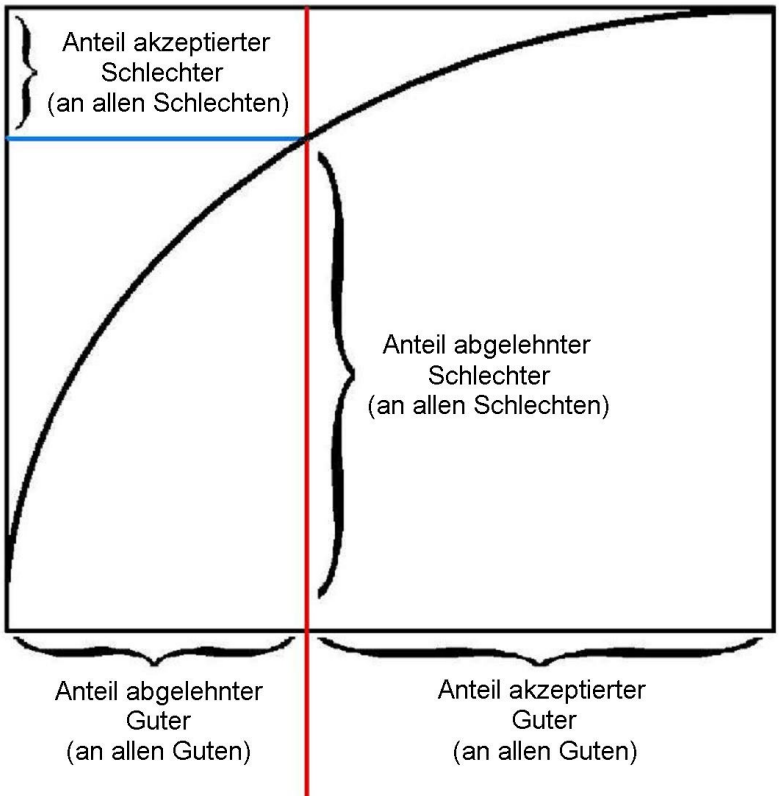


Abb. 6. Bestimmung des optimalen Schwellenwertes