Generalisierte Additive Modelle im Credit Rating: Eine Fallstudie zum Vergleich verschiedener Verfahren

Marlene Müller
Beuth Hochschule für Technik Berlin, Fachbereich II
Luxemburger Str. 10, D–13353 Berlin
E-mail: marlene.mueller@beuth-hochschule.de

16. Mai 2013

1 Einleitung

Im Kredit Scoring soll ein Ratingscore nicht nur ein optimales Klassifkationsergebnis darstellen, sondern ist gleichzeitig eine Komponente eines (im Regelfall recht komplexen) Ratingsystems. Wir betrachten im folgenden die Schätzung von Kredit-Ratingscores mit Hilfe semiparametrischer Logit-Modelle. Im klassischen parametrischen Logit-Modell ist der Score eine gewichtete Summe der Ratingfaktoren (erklärende Variablen). Auf diese Weise kann der Score einfach interpretiert werden und auch von Nichtstatistikern nachvollzogen werden.

Die Anpassung eines Modells für den Ratingscore erfolgt typischerweise in mehreren Einzelschritten. Der erste Schritt ist dabei die Auswahl der geeigneten Ratingfaktoren gefolgt von einer passenden, möglicherweise nichtlinearen Transformation. Diese Transformation der Rohdaten sorgt dafür, dass der Score später als lineare Funktion der transformierten Ratingfaktoren in der finalen Modellanpassung geschätzt wird. Als Alternative zu diesem zweistufigen Verfahren können generalisierte additive Modelle (GAM) verstanden werden. Sie erlauben die Transformation der Rohdaten und die Modellschätzung in einem Schritt durchzuführen.

Die Untersuchung soll hier verschiedene Schätzverfahren für generalisierte additive Modelle (GAM) vergleichen. Als Grundlage dienen verschiedene Kreditdatensätze, die verschiedene Eigenschaften solcher Daten widerspiegeln: Kleine Ausfallraten, Mischung kategorieller, diskreter und stetiger Variablentypen sowie möglicherweise nichtlineare Abhängigkeiten zwischen den Ratingfaktoren.

2 Kredit Rating

Die statistischen Aspekte des Kredit Scoring haben durch die Implementation der international vereinbarten Regeln zur Eigenkapitalunterlegung, die kurz als Basel II und III bezeichnet werden (Basel Committee on Banking Supervision; 2004), neue Bedeutung gewonnen. Banken und Finanzinstitutionen haben dadurch die Möglichkeit, ihre Eigenkapitalunterlegung an ihr eigenes Kreditportfolio anzupassen. Dies erfordert die eigene Schätzung der relevanten Komponenten auf Basis der Daten in den eigenen Kreditportfolien.

Die wichtigsten Komponenten des Schätzverfahrens sind die Bestimmung der individuellen Ratingscores und die Zuordnung der Ausfallwahrscheinlichkeiten. Beide Terme werden typischerweise als Funktionen der erklärenden Variablen (Ratingfaktoren) modelliert. Hier werden in der Praxis meist Modelle vom Logit- bzw. Probit-Typ verwendet, die in der Lage sind Scores (als lineare Prädiktoren) und Ausfallwahrscheinlichkeiten (als angepasste Bernoulliwahrscheinlichkeiten) simultan zu schätzen.

Statistisch gesehen haben wir ein Klassifikationsproblem mit zwei Gruppen, wofür Methoden zur binären Regression eingesetzt werden. Darüberhinaus sind verschiedene Aspekte des Risikomanagements zu betrachten:

- Das Kreditrisiko ist nur ein Teil des Gesamtrisikos einer Bank, d.h. die Kreditrisikoschätzung wird später mit anderen Risiken aggregiert.
- Die Schätzungen basierend auf historischen Daten dienen später als Grundlage für Stresstests zur Simulation zukünftiger Extremsituationen, der einfachen Anpassung des Ratingssystems an zukünftige Änderungen des Kreditportfolios und der Extrapolation in Ratingsegmente ohne oder mit nur wenigen Ausfallbeobachtungen.

Die Schätzung der individuellen Ratingscores und der zugehörigen Ausfallwahrscheinlichkeiten besteht meist aus den folgenden Schritten: Startpunkt sind die Rohdaten, d.h. die Bebachtungen der Ratingfaktoren X_j (der erklärenden Variablen). Ein erster Schritt ist dann eine (nichtlineare) Transformation $X_j \to \widetilde{X}_j = m_j(X_j)$, die gleichzeitig die Behandlung von Ausreißern (Extremwerten) und die Modellierung nichtlinearer Effekte der Ratingfaktoren auf die Risikoschätzung ermöglicht. Der Score ist also gegeben durch:

$$S = w_1\widetilde{X}_1 + \ldots + w_d\widetilde{X}_d\,.$$

Die Ausfallwahrscheinlichkeit (PD, Probability of Default) wird dann durch eine binäre Regression geschätzt, d.h. durch Verwendung des Modells

$$PD = P(Y = 1|\boldsymbol{X}) = G(w_1\widetilde{X}_1 + \ldots + w_d\widetilde{X}_d)$$

wobei G z.B. die logistische oder standardnormale Verteilungsfunktion (Logit- oder Probit-Modell) bezeichnet.

Ziel dieser Analyse ist es eine Fallstudie mit verschiedenen Ratingsdatensätzen (Querschnittsdaten) durchzuführen, die verschiedene Ansätze zur Schätzung generalisierter additiver Modelle (GAM) vergleicht. Insbesondere sollen dabei Modelle verwendet werden, die auch qualitative Variablen verwenden können (partiell lineare Prädiktoren). Das Interesse liegt dabei sowohl bei der simultanen Schätzung der Transformationen der Ratingfaktoren, der Ratingscores und der Ausfallwahrscheinlichkeiten.

3 Generalisierte Additive Modelle

Binäre Regressionsmodelle, insbesondere Logit- und Probit-Modelle, sind Spezialfälle des generalisierten linearen Modells (GLM):

$$E(Y|\boldsymbol{X}) = G\left(\boldsymbol{X}^\top \boldsymbol{\beta}\right). \tag{1}$$

Das klassische generalisierte additive Modell modifiziert dieses Modell derart, dass die linearen additiven Komponenten zu nichtparametrisch geschätzten Funktionen verallgemeinert werden:

$$E(Y|\boldsymbol{X}) = G\left\{c + \sum_{j=1}^{p} m_j(X_j)\right\}, \quad m_j \text{ nichtparametrisch}. \tag{2}$$

Im folgenden betrachten wir eine Erweiterung, das generalisierte additive partiell lineare Modell, das oft auch als semiparametrisches GAM bezeichnet wird. Dieses Modell erlaubt zusätzliche lineare Komponenten:

$$E(Y|\boldsymbol{X}_1, \boldsymbol{X}_2) = G\left\{c + \boldsymbol{X}_1^\top \boldsymbol{\beta} + \sum_{j=1}^{p} m_j(X_{2j})\right\}, \quad m_j \text{ nonparametric}. \tag{3}$$

Der zusätzliche lineare Teil des Prädiktors macht es möglich, für qualitative Ratingfaktoren (kategorielle erklärende Variablen) zu kontrollieren.

Die statistische Programmumgebung R (R Core Team; 2013) enthält zwei Standardpakte zur Schätzung von generalisierten additiven Modellen: Die Function `gam` aus dem Paket `gam` (kurz: `gam::gam`) implementiert den Backfitting-Algorithmus zusammen mit dem Local Scoring (Hastie and Tibshirani; 1990) und die Funktion `gam` aus dem Paket `mgcv` implementiert penalisierte Regressionssplines (Wood; 2006). Die folgende Fallstudie vergleicht diese zwei Schätzverfahren.

4 Fallstudie

Insgesamt werden folgende Schätzungen in den Vergleich einbezogen: Mit **logit** bezeichnen wir die Schätzung des Logit-Modells, d.h. die Schätzung von (1) mit $G(u) = 1/\{1+\exp(-u)\}$ als (inverser) Linkfunktion. Diese Logit-Schätzung wird ergänzt durch **logit2** und **logit3**, die Logit-Schätzungen mit zusätzlichen polynomialen Termen zweiten und dritten Grades für die stetigen Ratingfaktoren darstellen. Zum weiteren Vergleich betrachten wir auch eine Logit-Schätzung, bei der die stetigen Ratingfaktoren in 4–5 Faktorstufen kategorisiert werden. Diese Schätzung bezeichnen wir mit **logitc**. Die Bezeichnungen **gam** and **mgcv** werden für die binären GAM-Schätzungen zum Modell (3) mit der (inversen) Logit-Linkfunktion aus den R-Paketen verwendet, d.h. für `gam::gam` bzw. `mgcv::gam` mit Splinetermen für die stetigen Ratingfaktoren. Zusätzlich betrachten wir die Schätzung `gam::gam` mit einer ergänzenden Glättungsparameteroptimierung, diese wird mit **gamo** bezeichnet. Hier wird der der optimale Glättungsparameter durch Optimierung des AIC mit Hilfe der R-Funktion `optim` bestimmt.

Übersicht der Schätzverfahren

Abkürzung	Schätzung
logit	Logit-Schätzung, d.h. binäres GLM mit $G(u) = 1/\{1 + \exp(-u)\}$
logit2, **logit3**	Logit-Schätzung mit polynomialen Termen 2. und 3. Grades für die stetigen Variablen
logitc	Logit-Schätzung mit 4–5 Faktorstufen für die stetigen Variablen
gam	binäres GAM aus R-Funktion gam::gam mit `s()` Termen für die stetigen Variablen
gamo	binäres GAM aus R-Funktion `gam::gam` mit `s()` Termen für die stetigen Variablen, der Glättungsparamter `df` wird optimiert bzgl. AIC
mgcv	binäres GAM aus R-Funktion `mgcv::gam` mit `s()` Termen für die stetigen Variablen

Diese Fallstudie ergänzt die in Müller (2012) dargestellten Ergebnisse um die Schätzung mittels `gam::gam` mit Glättungsoptimierung. Sie beinhaltet außerdem detailliertere Ergebnisse zu den folgenden vier Kreditdatensätzen (siehe Abschnitte 6 und 7):

Übersicht der Datensätze

			Ratingfaktoren		
Datensatz	Stichprobenumfang	Kreditausfälle	stetig	diskret	kategoriell
German	1000	30.00%	3	–	17
Australian	678	55.90%	3	1	8
French	8178	5.86%	5	3	15
UC2005	5058	23.92%	12	3	21

5 Vergleich der Modelle

Die Validierung eines Ratingsystems umfasst zwei Aspekte: Zum einen soll das Ratingsystem trennscharf sein, d.h. der Ratingscore soll ermöglichen die Kreditausfälle und die Nichtausfälle zu trennen. Zum anderen soll das Ratingssystem gut kalibiriert sein, damit bezeichnet man eine gute Anpassungsgüte der geschätzen Ausfallwahrscheinlichkeiten.

Die Trennschärfe wird typischerweise durch die sogenannte CAP- (Cumulated Accuracy Profile) oder Lorenz-Kurve visualisiert. Diese Kurve kann als Variante der ROC-Kurve angesehen werden. Bei der ROC-Kurve würde man die Nichtausfall- gegen die Ausfallverteilungsfunktion der Scores dargestellen ($\widehat{F}_0$ vs. $\widehat{F}_1$, jeweils geschätzt durch die empirischen Verteilungsfunktionen der Scores). Im Unterschied dazu stellt die CAP-Kurve die Verteilungsfunktion aller Scores gegen die der Ausfälle dar, d.h. $\widehat{F}$ gegen $\widehat{F}_1$ (wobei $\widehat{F}$ die empirische Verteilungsfunktion aller Scorewerte bezeichnet). In Analogie zur Fläche unter der Kurve (AUC, Area under Curve) der ROC-Kurve fasst man auch hier die Trennschärfe in einem Zahlenwert, dem Accuracy Ratio (AR) zusammen. Abbildung 1 veranschaulicht die Konstruktion der CAP-Kurve. Der Accuracy Ratio AR als Trennschärfemaß ist eine lineare

Funktion der Fläche unter der Kurve AUC der ROC-Kurve:

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

Der AR wird berechnet als die Fläche zwischen der CAP-Kurve und der Diagonalen (die keine Trennschärfe zeigt) im Verhältnis zur optimalen CAP-Kurve (volle Trennschärfe bei perfekter Trennung). In der Praxis variiert der AR zwischen 0 und 1 (bzw. 0% und 100%), sofern es eine monotone wachsende Beziehung zwischen Score und Ausfallrisiko gibt.

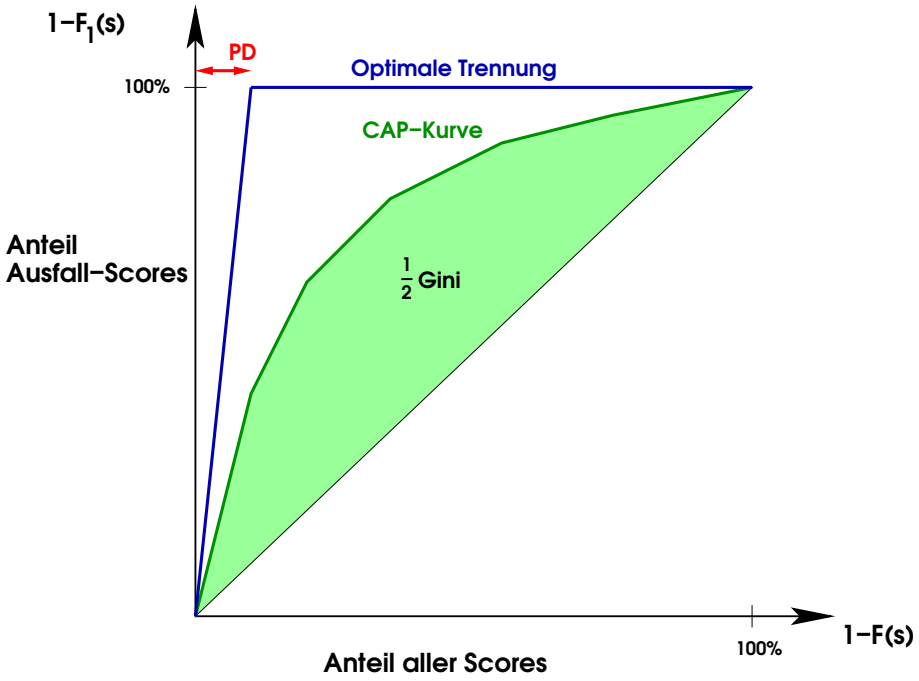

Abbildung 1: Schematische Darstellung der CAP-Kurve (Cumulated Accuracy Profile)

Die im folgenden dargestellten AR-Werte wurden durch eine Out-of-Sample-Validierung bestimmt. Wir benutzen einen blockweisen Kreuzvalidierungsansatz bei dem jeweils x% der Daten aus der Schätzung ausgelassen und zur Berechnung des AR verwendet wurden. Der Anteil x% wird für die verschiedenen Datensätze in Abhängigkeit der Ausfallrate verschieden hoch gewählt.

Zur Einschätzung der Kalibrierung der Ausfallwahrscheinlichkeiten verwenden wir zusätzlich die Out-of-Sample-Devianz der Modellschätzung:

$$\text{Devianz} = -2 \sum_{i=1}^{n} \left\{ y_i \log(\widehat{PD}_i) + (1 - y_i) \log(1 - \widehat{PD}_i) \right\} .$$

Die Devianzwerte werden ebenfalls durch den oben beschriebenen blockweisen Kreuzvalidierungsansatz berechnet.

6 Deutsche Kredit-Daten

Der hier betrachtete Datensatz, den wir im folgenden mit **German** bezeichnen ist einer der wenigen frei verfügbaren Datensätze zu Ratingdaten, bei dem auch eine detaillierte Variablenbeschreibung verfügbar ist. Der Datensatz ist eine geschichtete Stichprobe, in der der

Ausfallanteil höher ist als in der Realität (30% Ausfallrate in der Stichprobe, die wahre Ausfallrate dürfte bei ca. 5% liegen). Drei der erklärenden Variablen können als stetige Variablen angesehen werden (Alter des Kreditnehmers, Betrag und Laufzeit des Kredits).

			Ratingfaktoren		
Datensatz	Stichprobenumfang	Kreditausfälle	stetig	diskret	kategoriell
German	1000	30.00%	3	–	17

Quelle: `http://www.stat.uni-muenchen.de/service/datenarchiv/kredit/kredit.html`

Abbildung 2 zeigt die geschätzten additiven Komponentenfunktionen für die stetigen Variablen Alter (age), Kreditbetrag (amount) und Laufzeit (duration), jeweils durch `gam::gam` (blau), `gam::gam` mit Glättungsoptimierung (dunkelblau) und `mgcv::gam` (schwarz) berechnet. Hier wurden für beide Schätzungen die Standardwerte der R-Funktionen benutzt. Die zusätzlich dargestellten Konfidenzbänder (gestrichelte Linien) sind die durch `mgcv::gam` berechneten.

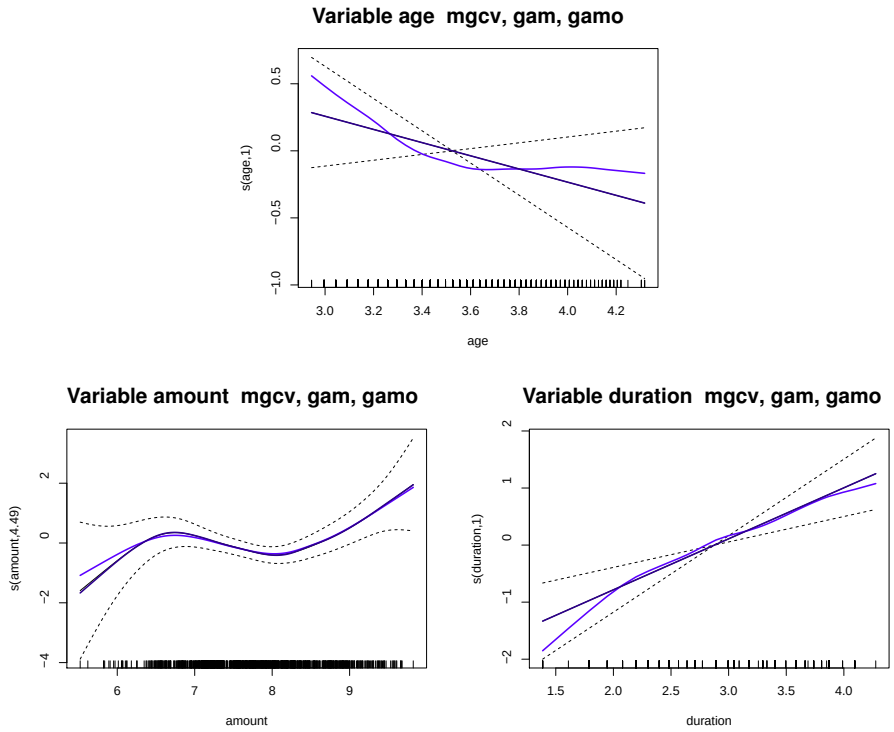

Abbildung 2: German Credit Data: Additive component functions for continuous regressors

Die nachfolgende Abbildung 3 zeigt den Out-of-sample-Vergleich (blockweise Kreuzvalidierung mit 5 Blöcken) für die verschiedenen Schätzer: Boxplots und Profillinien der AR-Werte (obere Panels) und der Devianzwerte (untere Panels). In Bezug auf die AR-Werte zeigt sich, dass sich `mgcv::gam` und `gam::gam` (mit Glättungsparameteroptimierung) sehr ähnlich zueinander und am besten im Vergleich zu den anderen Schätzern verhalten. Bei den Devianzen erkennt man einen Ausreißer in der vierten Out-of-Sample-Validierung zu `mgcv::gam`. Dieser kommt durch einen extrem kleinen Wert der Ausfallwahrscheinlichkeitsschätzung bei der Extrapolation auf die Out-of-Sample-Daten zustande. Zusammenfassend kann man schließen, dass die Trennschärfe bei `mgcv::gam` und `gam::gam` mit Glättungsparameteroptimierung

vergleichbar ist, in der Kalibrierung der Ausfallwahrscheinlichkeiten jedoch bei `mgcv::gam` in einem Fall deutlich schlechter abschneidet.

Wenn wir nur die drei stetigen erklärenden Variablen in der Schätzung verwenden tritt das Problem in der Kalibrierung nicht auf (vgl. Abbildung 4).

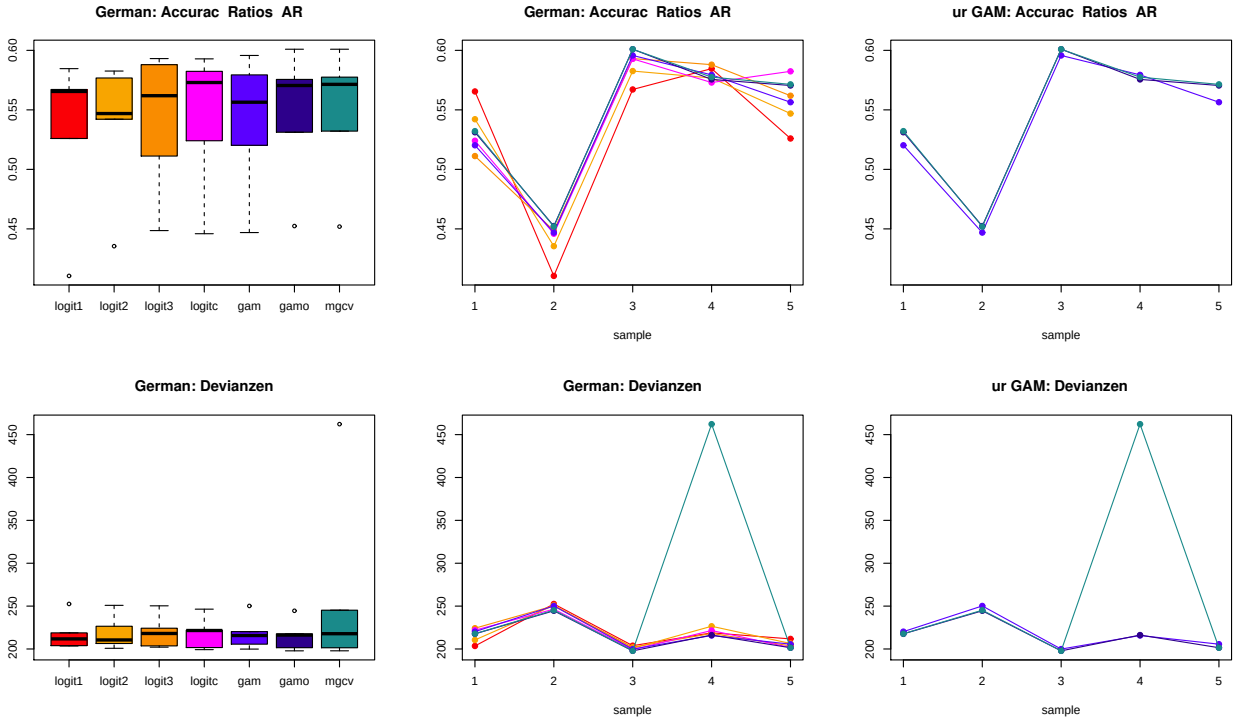

Abbildung 3: Vergleich der Modelle für den Datensatz German

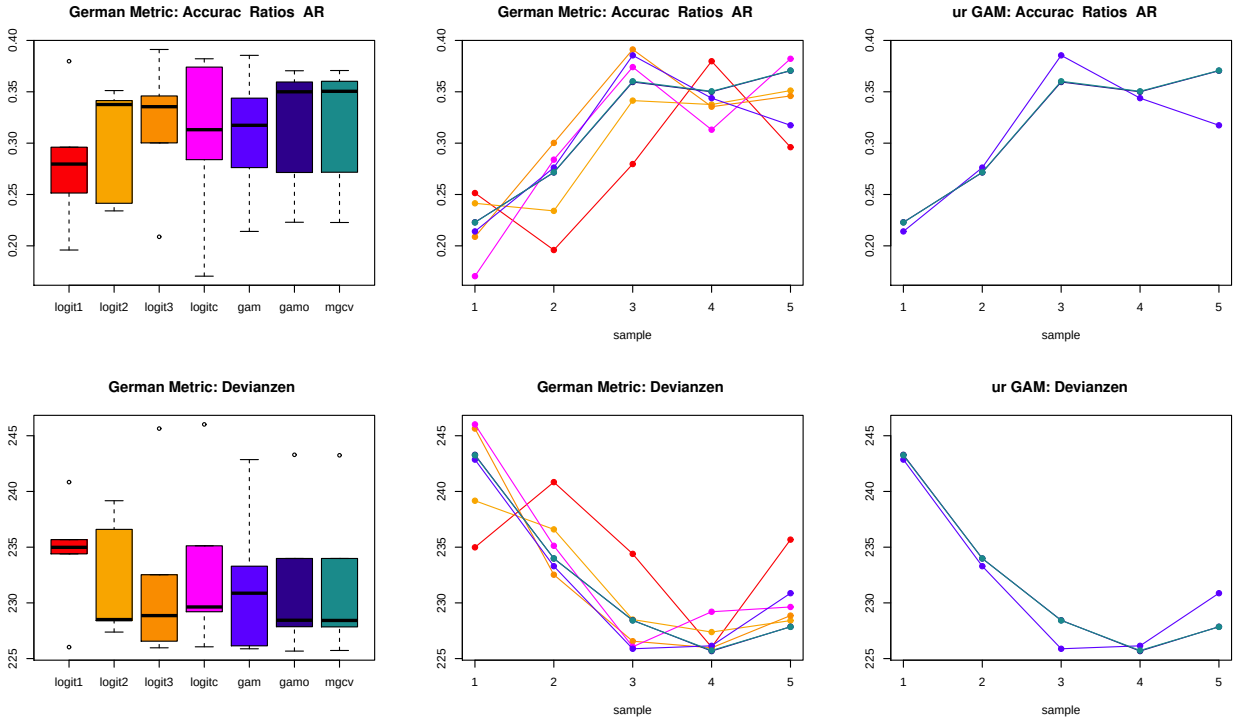

Abbildung 4: Vergleich der Modelle für den Datensatz German, hier Modelle mit nur stetigen Variablen

7 Weitere Kredit-Datensätze

In diesem Abschnitt sollen die weiteren drei Fallbeispiele vorgestellt werden. Alle drei Datensätze bestehen aus anonymisierten Variablen, d.h. die Resultate sind inhaltlich nicht interpretierbar. Die Datensätze werden im folgenden mit **Australian**, **French** und **UC2005** abgekürzt.

Übersicht der Datensätze

Datensatz	Stichprobenumfang	Kreditausfälle	Ratingfaktoren stetig	diskret	kategoriell
Australian[1]	678	55.90%	3	1	8
French[2]	8178	5.86%	5	3	15
UC2005[3]	5058	23.92%	12	3	21

[1] Quelle: http://archive.ics.uci.edu/ml/datasets/Statlog+(Australian+Credit+Approval)
[2] Gleicher Datensatz wie in Müller and Härdle (2003)
[3] Quelle: UC 2005 Klassifikationswettbewerb

Für den Datensatz Australian erkennt man, dass die Schätzungen durch `gam:gam` tendenziell am besten abschneiden (vgl. Abbildung 5).

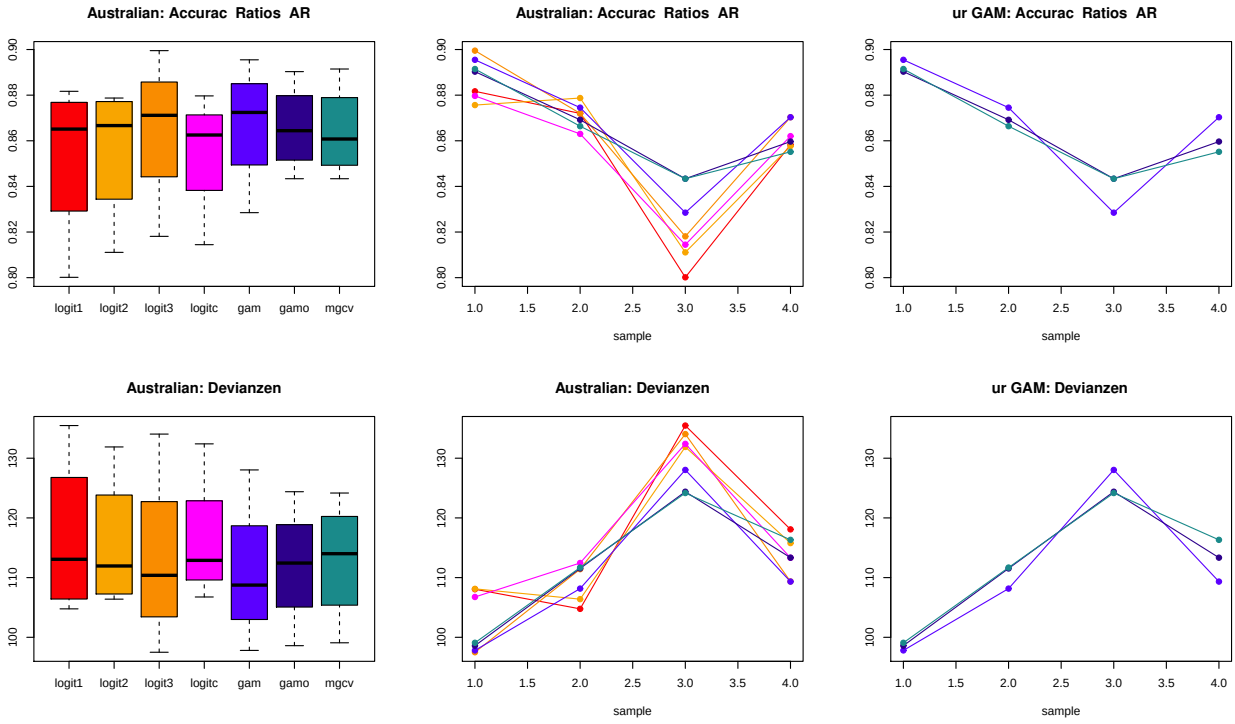

Abbildung 5: Vergleich der Modelle für den Datensatz Australian

Am Beispiel der Datensätze French (vgl. Abbildung 6) und UC2005 (vgl. Abbildung 7) lassen sich jeweils wieder eine einzelne schlechte Kalibrierung der Ausfallwahrscheinlichkeiten erkennen. In der Tendenz sieht man jedoch, dass alle semiparametrischen GAM-Schätzer vergleichsweise gleich gut abschneiden.

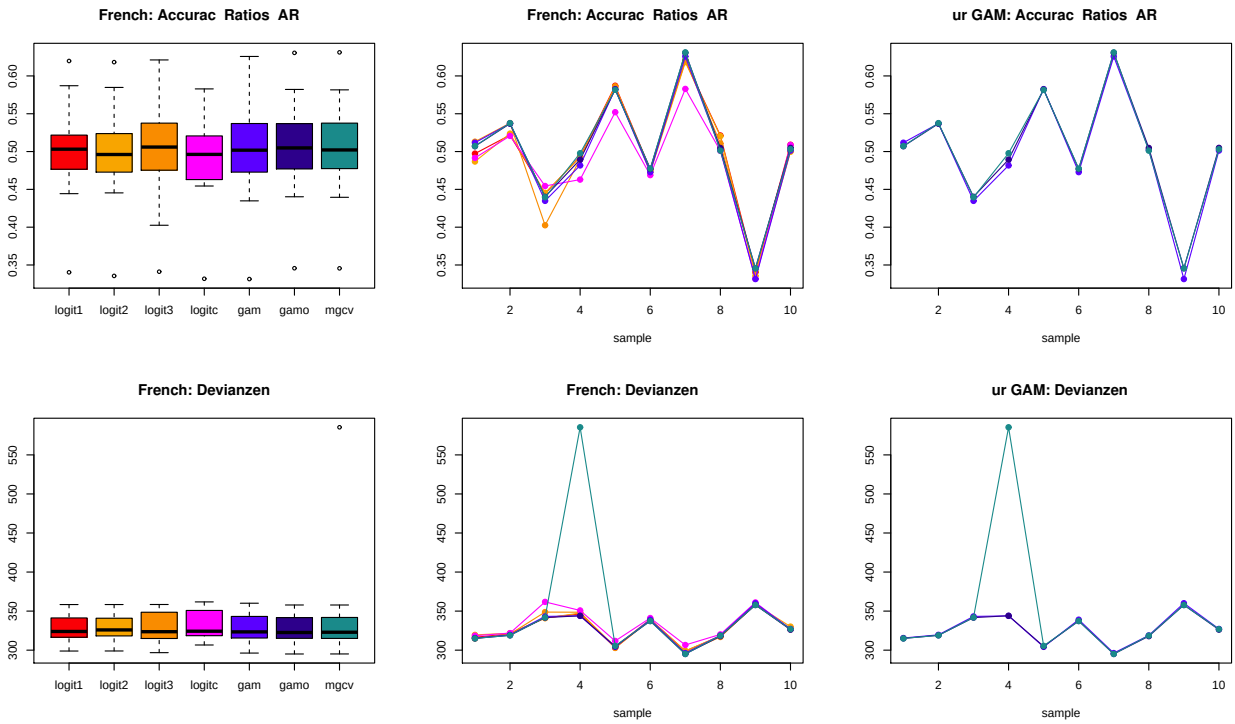

Abbildung 6: Vergleich der Modelle für den Datensatz French

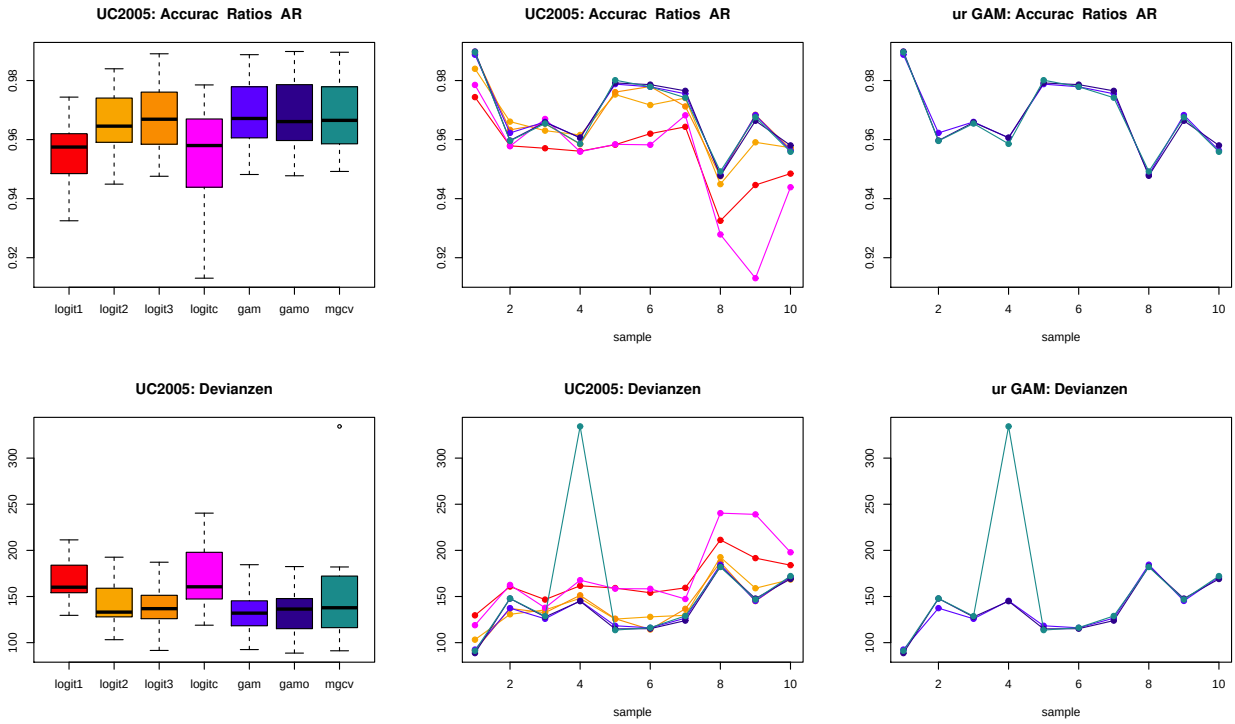

Abbildung 7: Vergleich der Modelle für den Datensatz UC2005

8 Fazit

Die Analyse konzentriert sich auf semiparametrische GAM-Schätzungen für Ratingscores und Ausfallwahrscheinlichkeiten. Da Ratingdatensätze üblicherweise eine Mischung kategorieller, diskreter und stetiger Variablentypen enthalten, ist es sinnvoll die folgenden in R

verfügbaren Schätzmethoden zu vergleichen: Das klassische Backfitting mit Local Scoring (in R: `gam::gam`) stellt schnell numerisch stabile Ergebnisse zur Verfügung, diese lassen sich jedoch meist noch mit einer Glättungsparameteroptimierung im Hinblick auf Trennschärfe und Kalibrierung verbessern. Diese Optimierung nimmt jedoch gerade für große Datensätze deutlich Zeit in Anspruch. Die von (Wood; 2006) implementierten penalisierten Regressionssplines (in R: `mgcv::gam`) liefern vergleichbar gute Ergebnisse wie `gam::gam` mit Glättungsparameteroptimierung, da hier eine Optimierung schon automatisch durch die R-Funktion durchgeführt wird. Hier gibt es jedoch in Einzelfällen Probleme bei der Extrapolation auf Out-of-Sample-Daten.

Literatur

Basel Committee on Banking Supervision (2004). *Basel II: International Convergence of Capital Measurement and Capital Standards: a Revised Framework*, Bank for International Settlements (BIS), Basel, Switzerland.
URL: *http://www.bis.org*

Härdle, W., Müller, M., Sperlich, S. and Werwatz, A. (2004). *Nonparametric and Semiparametric Modeling: An Introduction*, Springer, New York.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*, Vol. 43 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, London.

Müller, M. (2012). A case study on using generalized additive models to fit credit rating scores, *Bulletin of the International Statistical Institute Proceedings of the 58th World Statistics Congress 2011, Dublin*, International Statistical Institute, The Hague, The Netherlands.
URL: *http://2011.isiproceedings.org/*

Müller, M. and Härdle, W. (2003). Exploring credit data, *in* G. Bol, G. Nakhaeizadeh, S. Rachev, T. Ridder and K.-H. Vollmer (eds), *Credit Risk - Measurement, Evaluation and Management*, Physica-Verlag.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *http://www.R-project.org/*

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*, Texts in Statistical Science, Chapman and Hall, London.