DAV

DEUTSCHE
AKTUARVEREINIGUNG e.V.

Ergebnisbericht des Ausschusses Schadenversicherung

Aktuarieller Umgang mit Big Data in der Schadenversicherung

Köln, 17. Mai 2019

Präambel

Der Ausschuss Schadenversicherung der Deutschen Aktuarvereinigung e. V. hat den vorliegenden Ergebnisbericht erstellt.[1]

Zusammenfassung

Der Ergebnisbericht behandelt Fragestellungen zum Themenbereich Big Data / Data Science und versucht hierbei einen ersten Überblick über folgende Themengebiete zu geben: Mögliche Anwendungen, Datenhaushalt (insbesondere in Bezug auf die Sicherstellung eines verlässlichen Kern-Datenhaushaltes), Kurzdarstellung üblicher Methoden und Möglichkeiten zur Bewertung verschiedener Modelle. Die konkreteste Anwendbarkeit dürfte für Aktuare gegeben sein, die an der Schnittstelle zwischen der Abbildung von Schadenversicherungsprodukten in den Verwaltungssystemen und deren statistischer Abbildung arbeiten. Er betrifft Aktuare insbesondere der Schaden-/Unfallversicherung, kann jedoch gegebenenfalls am Rande auch für Aktuare, die in anderen Spezialgebieten tätig sind, interessante Aspekte enthalten. Der Anwendungsbereich umfasst grundsätzlich alle Produkte und alle Teilbereiche des Versicherungsgeschäftes.

Der Ergebnisbericht ist an die Mitglieder und Gremien der DAV zur Information über den Stand der Diskussion und die erzielten Erkenntnisse gerichtet und stellt keine berufsständisch legitimierte Position der DAV dar.[2]

Verabschiedung

Der Ergebnisbericht ist durch den Ausschuss Schadenversicherung am 17. Mai 2019 verabschiedet worden. Er ersetzt die gleichnamige Vorversion vom 20. November 2017, die um das *Kapitel 5* zur Bewertung von Modellen erweitert wurde.

---

[1] Der Ausschuss dankt der Arbeitsgruppe *Tarifierungsmethodik* ausdrücklich für die geleistete Arbeit, namentlich Prof. Dr. Beate Bergter, René Billing, Dr. Klaus Dräger, Frank Ellgring, Thomas Franze, Florian Fuchsbrunner, Peter Gorontzy, Matthias Göttlich, Roderich Heim, Jochen Kneiphof, Daniel König, Dr. Olaf Kruse, Ina Kühnel, Christof Lisakowski, Andreas Löffler, Dr. Tina Marquardt, Dr. Gero Nießen, Dr. Christian Ott, Johannes Pohl-Grund, Frank Rastbichler, Prof. Dr. Viktor Sandor, Frank Schönfelder, Dr. Jörg Schult, Dr. Michael Schüte, Dr. Gerald Sussmann (Leitung), Karsten Vogel, Wiltrud Weidner, Dr. Stefan Wetzel und Axel Wolfstein.

[2] Die sachgemäße Anwendung des Ergebnisberichts erfordert aktuarielle Fachkenntnisse. Dieser Ergebnisbericht stellt deshalb keinen Ersatz für entsprechende professionelle aktuarielle Dienstleistungen dar. Aktuarielle Entscheidungen mit Auswirkungen auf persönliche Vorsorge und Absicherung, Kapitalanlage oder geschäftliche Aktivitäten sollten ausschließlich auf Basis der Beurteilung durch eine(n) qualifizierte(n) Aktuar DAV/Aktuarin DAV getroffen werden.

Abstract

This paper gives a short survey over some topics resulting out of Big Data / Data Science applications relevant to the actuary in P/C insurance. After a short introduction in Chapter 1, a variety of possible applications of Big Data / Data Science is discussed in Chapter 2. It is emphasized that this survey does not serve as a complete reference for all Big Data / Data Science related topics in P/C insurance, but should be considered as a little glimpse into a huge world of future ideas and inventions. Chapter 3 gives a guideline to the actuary, which supports to keep structured and understandable core data describing the insurance business in the context of growing flexibility of inventory systems. In Chapter 4, a short overview of mostly used Machine Learning Algorithms with complementing literature is given.

Inhalt

1. Einleitung

Dieses Dokument gibt einen Überblick über aktuarielle Anwendungen von Big Data bzw. Data Science in der Schadenversicherung.

Wir verwenden Big Data als Oberbegriff für den Themenkomplex rund um Data Science und die dazugehörigen Begrifflichkeiten Machine Learning, Data Mining und ähnliche. Diese Begriffe sind nicht eindeutig abzugrenzen und werden in Literatur und Fachpresse nicht einheitlich verwendet.

Konkret geht es beim Thema Big Data um Daten, Infrastruktur, Datenanalyse und schließlich die Anwendung und Operationalisierung in das Geschäftsmodell. Das erste Kapitel ist eine Einleitung zum Thema mit einem Überblick über die generellen Aspekte und Entwicklungen. Das zweite Kapitel zeigt potentielle Anwendungsbeispiele für Schadenversicherer auf. Im dritten Kapitel wird das Thema Daten und Datenhaltung aufgegriffen. Abschließend beschreibt das vierte Kapitel die relevanten Methoden zur Datenanalyse, die oftmals unter dem Begriff „Machine Learning" subsummiert werden.

Wie der Leser sicher bemerkt, spiegelt sich in der vorliegenden Notiz die Tatsache wider, dass es sich um kein fest abgegrenztes Themengebiet mit einer einheitlichen Definition handelt, in dem Problemstellungen, Ansätze und Methoden einem mehr oder weniger konkreten Standard unterliegen. Die im *Kapitel 2* angesprochenen Ansätze sind demzufolge eher als erste Ideen zu sehen, denn als konkrete Versuche mit bekanntem Ansatz oder Erfolg. Im Wissen, dass die Datenbasen künftiger Analysen genauso speziell und vielfältig sein werden, wie die noch neu zu entdeckenden Anwendungen selbst, haben wir uns in *Kapitel 3* weniger dahin gewandt, künftige Ideen „voraus-zu-erfinden", als vielmehr die Wichtigkeit eines stabilen versicherungstechnischen Kerns zu betonen. Wir hoffen, hier den Aktuar durch geeignete Abstraktion in die Lage zu versetzen, diesen Kern auch und insbesondere in einer sich schnell verändernden Datenwelt so zu beschreiben, dass er nicht verloren geht. Der Leser von *Kapitel 4* kann aus unserer Notiz keinesfalls die genaue Beschreibung der im Themengebiet angewandten Mathematik entnehmen. Dennoch wird er hoffentlich einen groben Überblick gewinnen über die Grundideen, auf denen die verschiedenen Machine-Learning-Verfahren basieren – und dabei vielleicht die Erkenntnis, dass ein gutes Verständnis der bekannten klassischen statistischen Ansätze auch in der neu anbrechenden Welt wertvoll bleibt!

Die Sicherstellung der Einhaltung des Datenschutzes sei dem Leser an dieser prominenten Stelle dringend ans Herz gelegt. In dieser Notiz können und werden wir aufgrund der Komplexität dieses Themas keinerlei konkrete Aussagen treffen.

1.1. Hintergründe, Historie

Im Zuge der fortschreitenden Digitalisierung und Ausbreitung des Internets seit 1990 gab und gibt es parallel verlaufende und sich teilweise gegenseitig beeinflus-

sende Entwicklungen, die zu einer erheblichen Ausweitung des prinzipiell verfügbaren Datenvolumens führen nebst Möglichkeiten, diese zu analysieren und zu diversen kommerziellen Zwecken zu nutzen. Beispiele für diese Entwicklungen sind:

- die Nicht-Beschränkung der Vernetzung auf „Rechner" im engeren Sinne („Internet der Dinge")
- die auch im privaten Bereich beliebte Nutzung von „Social Media"
- die Möglichkeiten des „Cloud Computing" (= die Bereitstellung von IT-Infrastruktur wie Speicherplatz, Rechnerleistung und Anwendungssoftware über das Internet → Wikipedia)
- Fortschritte bei der Entwicklung von (auch Open Source) Software, die speziell auf die effiziente Verarbeitung hoher Datenvolumina zugeschnitten ist, wie z. B. dem Framework „Hadoop"
- die Wiederentdeckung und die Weiterentwicklung von Analyse-Methoden, die unter „Maschinenlernen" subsummiert werden und die oft schon den 90-er Jahren oder noch früheren Zeiten entstammen, aber mittlerweile auch das technische „Biotop" vorfinden, das sie benötigen, um ihren Nutzen zu entfalten.
- Manchen Quellen zufolge (siehe z. B. [1]) ist derzeit ein exponentielles Wachstum des Datenvolumens mit einer Verdoppelung alle zwei Jahre zu verzeichnen. Ein Großteil der Daten stammt aus (teilweise privater) Kommunikation in Wort, Ton und Bild. Insbesondere die Verbreitung des Smartphones hat zu einer deutlichen Beschleunigung des Datenvolumens beigetragen. Ebenfalls zu nennen sind die datengestützte Steuerung und Überwachung technischer Systeme, beginnend bei Haus oder Auto.

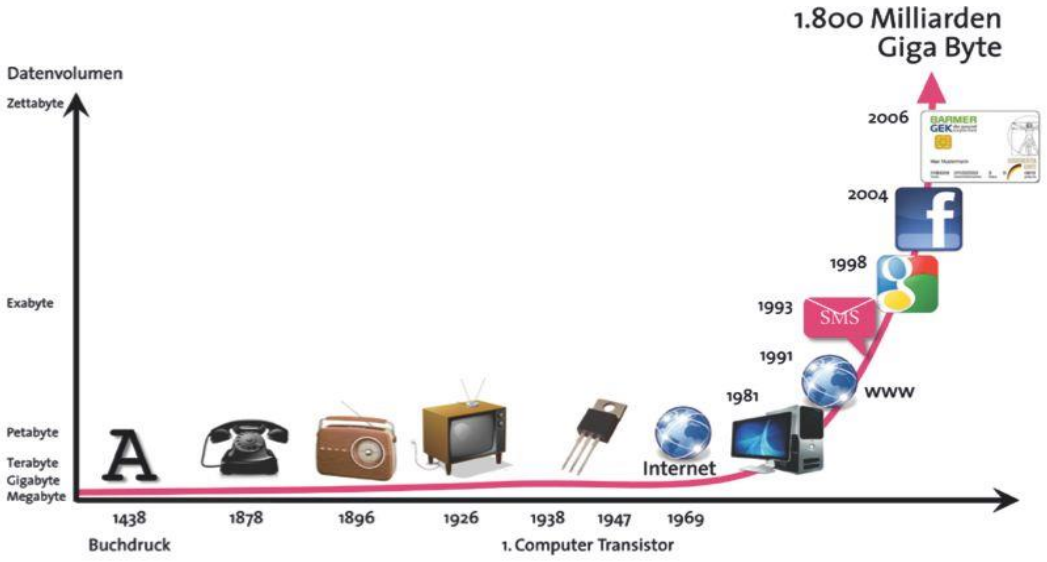


Abbildung 1: Wachstum der Datenmenge weltweit
[Quelle: BiTKOM [Hrsg.]: Big Data im Praxiseinsatz – Szenarien, Beispiele, Effekte; 2012]

Das exponentielle Wachstum der weltweit verfügbaren Datenmenge seit dem Mittelalter ist in *Abbildung 1* dargestellt.

1.2. Thesen zu Big Data in der deutschen Versicherungsbranche

Es fällt auf, dass viele prominente Beispiele von kommerziellen Big Data-Anwendungen der jüngeren Jahre nicht aus der Versicherungsbranche entstammen, in der Datensammlung und Erkenntnisgewinn durch Anwendung datenanalytischer Methoden seit jeher essenziell und existenziell mit dem Geschäftsmodell einhergehen. Möglicherweise neigte die Branche gerade deshalb dazu, sich auf wohlstrukturierten althergebrachten Datenmodellen „auszuruhen" und den Big Data-Trend etwas zu „verschlafen". Ob es sich bei diesem Trend vielleicht auch um Hype oder Hybris handelt, ist bereits diskutiert worden (siehe z. B. [3]). Ein gesundes Maß an Skepsis ist nie verkehrt; am Telematik-Beispiel werden in [4] Denkanstöße dargestellt, warum nicht jeder Digitalisierungs-/Big Data-Ansatz auch quasi von alleine einen wirtschaftlichen Mehrwert versprechen muss. Andererseits ist in zahlreichen Beispielen (siehe z. B. [5]) belegt worden, dass die „klassischen" Analyse-Methoden optimierbar sind. Gegenwärtig sind Umwälzungen und die Ausrichtung auf die neuen Möglichkeiten der Big Data-Welt in der deutschen Versicherungswirtschaft unverkennbar. Hierzu wollen wir mit diesem Papier, das sich insbesondere an Aktuare richtet, die bzgl. Big Data „Einsteiger" sind, einen Beitrag leisten.

1.3. Branchenübergreifende Beispiele von Big Data Konzepten

Die Anzahl erfolgreich implementierter Big Data & Analytics Use Cases wächst ständig. Auch wenn nicht alle Use Cases einen sofort messbaren positiven Business-Impact haben, gibt es dennoch etliche Erfolgsgeschichten aus unterschiedlichen Wirtschaftszweigen. Einige dieser Use Cases, die durch die Berichterstattung in der Tagespresse einen gewissen Bekanntheitsgrad erreicht haben, sind in der folgenden Aufstellung zusammengefasst.

- *Einführung neuer Kaffeesorte bei Starbucks*
  - Begleitende On-the-Fly-Analyse von Social-Media-Kommentaren
  - Ergebnis: Kaffee schmeckt OK, ist aber zu teuer => Preis wird ad hoc gesenkt
- *Maßgeschneiderte Werbung bei TARGET*
  - Personalisierte Werbung über Analyse des bisherigen Kaufverhaltens
  - „Ungerechtfertigte" Kunden-Beschwerde wegen Werbung für Baby-Kleidung
- *Optimierung von Öl-Bohrungen im Golf von Mexico*
  - Jede Fehlbohrung kostet 30 Mio. Dollar
  - Optimierung der Bohrungen über Analyse von Satellitendaten
- *Fuhrparksteuerung bei US EXPRESS*
  - Optimiertes Fahrzeug-/Flotten-Routing
  - Ad-hoc-Auswertung der Fahrzeug-/Motordaten

Das größte Transformationspotential durch Big-Data wird zurzeit in folgenden Bereichen gesehen:

- *Gesundheitswesen*
  - Smart Health / permanentes Monitoring / automatisierte Diagnose
- *Public Sector / Verwaltung*
  - Optimierung der Verwaltungsabläufe / Service-Angebote
  - Verkehrsmanagement / Bedarfsplanung der Versorger
- *Sicherheitswesen / Public Safety*
  - Cybersecurity - Counter-Terrorism – Kriminalitätsbekämpfung
- *Global Manufacturing (IoT / M_2M-Kommunikation)*
  - Prozessoptimierung / Produktionssteuerung / Qualitäts-Management
- *Vertriebssteuerung / Marketing*
  - Zielgenaues/individualisiertes Marketing (wem – was – wann – wo)
- *Personal Location Data*
  - Telematik / Google Car / Epidemiologie

1.4. Definition von Big Data

Der Begriff Big Data wird häufig (siehe z. B. [6]) über ursprünglich drei sogenannten „V"-Schlüsselwörter definiert, die mittlerweile um weitere „V"s ergänzt wurden:

- *Volume*
  - Exponentielles Wachstum des Datenvolumens
- *Velocity*
  - Geschwindigkeit der Daten-/Informationsentstehung
  - Geschwindigkeit des Daten-/Informationsflusses
- *Variety*
  - Unterschiedlichste Datenquellen und -formate
  - Strukturiert/Unstrukturiert / Text – Bilder – Video – Audio usw.
- *Veracity*
  - Daten-/Informationsqualität i. w. S.
  - Completeness - Consistency - Credibility - Accuracy - Objectivity - Unbiased –Truthful
- *Value*
  - Effizient aus Daten Informationen generieren

Ein bewusst anderer Definitionsansatz findet sich in [7]:

> "A pragmatic definition of big data must be actionable for *both* IT and business professionals.
>
> *The Definition of Big Data*
>
> *Big Data is the frontier of a firm's ability to store, process, and access (SPA) all the data it needs to operate effectively, make decisions, reduce risks, and serve customers.*
>
> To remember the pragmatic definition of big data, think SPA — the three questions of big data:
>
> - *Store.* Can you capture and store the data?
> - *Process.* Can you cleanse, enrich, and analyze the data?
> - *Access.* Can you retrieve, search, integrate, and visualize the data?"

Insgesamt muss wohl festgehalten werden, dass sich nicht bei jedem datenbasierten Analyseproblem eindeutig zuordnen lässt, ob es zu „Big Data" gehört; der Begriff bleibt also insofern „schwammig" (vgl. [6]).

1.5. Big Data als ganzheitlicher Transformationsprozess

In [1] wird Big Data im Versicherungsunternehmen als „Gesamt-Konzept" dargestellt, das keineswegs allein das Aktuariat betrifft, sondern Transformationspotenzial für das ganze Unternehmen hat - mit zahlreichen Facetten und Herausforderungen.

Symptome einer fehlenden bzw. nicht gesamthaft entwickelten gemeinsamen „Datensprache" sind:

- Probleme, die Daten zu verstehen bzw. auszuwerten/abzufragen
- Ständige Notwendigkeit einer Datenbereinigung bzw.-konsolidierung
- Kein koordinierter Zugriff auf alle Daten(-quellen)
- Notwendigkeit, auf Ad-hoc-Proxies zurückzugreifen
- Probleme, Analysen zu aktualisieren bzw. zu reproduzieren
- Probleme bei der Erklärung, Umsetzung bzw. Anwendung von Ergebnissen
- Widersprüchliche, inkonsistente Ergebnisse

Auf einem höheren Level können die entstehenden Probleme durch folgende Beispiele bzw. Symptome charakterisiert werden:

- Zusammenarbeit zwischen verschiedenen Datensilos ist schwierig bzw. nicht existent
- Vergleichbarkeit zwischen verschiedenen Business Units ist schwierig bzw. nicht existent
- Fehlende Transparenz bzw. Real-Time-Monitoring
- Detaillierte Übersicht
- „As-if"-Szenario-Analysen können nur ad-hoc erfolgen und sind schwer zu interpretieren

Als Maßnahmen, diese Symptome zu überwinden, werden empfohlen:

Wenn man sich auf eine gemeinsame Datensprache verständigt hat, sollte sie allgemeingültig fixiert werden. Dass kann durch ein Daten-Modell geschehen, welches folgende Eigenschaften haben sollte:

- Für alle Beteiligten verständlich strukturiert
- In einem koordinierten Prozess gemanagt/abgestimmt
- Die Daten sollten über die Zeit veränderbar sein.

Speziell die IT-seitigen Aspekte von Big Data-Herausforderungen werden in [2] mit einem sehr großen Angebot an weiterführender Literatur zusammengefasst. Auch wird Big Data als ganzheitlicher Ansatz verstanden, was sich z. B. in den sechs empfohlenen Prinzipien für das Design von Big Data-Systemen konkretisiert:

1. Gute Architekturen und Frameworks sind notwendig und sollen hohe Priorität bekommen.
2. Eine Vielzahl von Analysemethoden soll zur Verfügung stehen.
3. Es gibt keine Größe, die für alle Problemstellungen passt.
4. Die Analyse-Tools müssen (manchmal) zu den Daten gebracht werden (statt umgekehrt).
5. Verarbeitung muss im Arbeitsspeicher organisierbar sein.
6. Datenspeicherung muss (nach Partitionierung) im Arbeitsspeicher organisierbar sein.
7. Verarbeitungs- und Dateneinheiten müssen gut koordiniert werden.

1.6. Methodik

In [2] werden die Datenanalyse-Methoden im Zusammenhang mit Big Data wie in *Abbildung 2 d*argestellt zusammengefasst (Bei genauer Betrachtung, siehe insbesondere auch *Kapitel 4*, verschwimmen jedoch die Grenzen der hier dargestellten Teilgebiete, oder diese überlappen sich auch zum Teil):

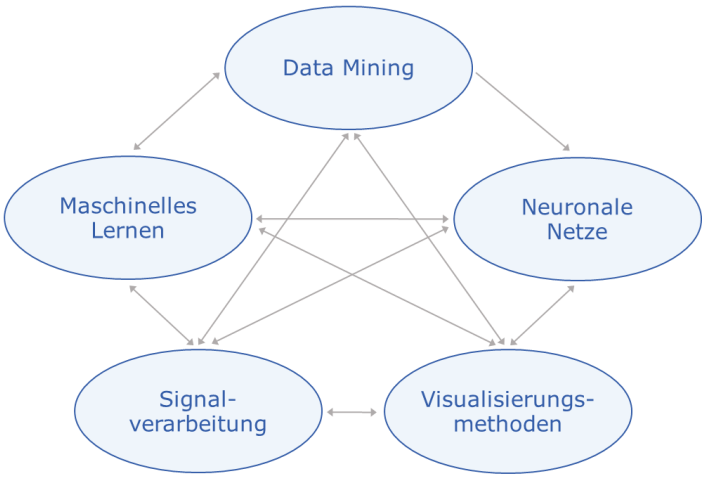


Abbildung 2: Big-Data-Datenanalyse-Methoden in Anlehnung an [2]

Dabei dürfte das Maschinenlernen am engsten verwandt sein mit der klassischen Tätigkeit eines Tarifierungs-Aktuars in der Schaden- und Unfallversicherung. Unterschieden wird dabei stets, ob es sich um *überwachtes* (englisch: *supervised*) oder *unüberwachtes* (englisch: *unsupervised*) Lernen handelt.

Bei ersterem liegt in den zu untersuchenden Daten eine Zielvariable vor (z. B. Schadenaufwand, Stornoprognose), deren Werte durch die übrigen Einflussvariablen möglichst exakt bestimmt werden sollen. Bei letzterem hingegen sollen allgemeine Aussagen über eventuelle Muster im Datensatz getroffen werden. Nimmt die im Umfeld des *Supervised Learnings* gegebene Zielvariable numerische Werte an, liegt ein Regressionsproblem vor (Höhe des Schadenaufwands). Enthält sie Ausprägungen einer Kategorie, handelt es sich um ein Klassifikationsproblem (z. B. Storno ja/nein).

In *Kapitel 4* werden einige dieser Methoden aufgegriffen und genauer erläutert.

2. Anwendungen

In diesem Kapitel geben wir einen Überblick über mögliche Anwendungsgebiete von Machine Learning in der Schadenversicherung. Ein Teil der beschriebenen Anwendungen wird bereits von Versicherern in der Praxis angewendet.

2.1. Vertrieb/Marketing

Für Vertrieb und Marketing lassen sich im Netz vorhandene Informationen für viele Zwecke heranziehen. Naheliegend ist ein *Chancenfrüherkennungssystem*, welches aus Foren z. B. Folgendes erkennt:

- Welcher Versicherungsnehmer bei anderen Versicherungsunternehmen ist gerade unzufrieden wegen Schadenregulierung, Beitragsrechnung, Vertragsverwaltung?
- Können wir das besser? Dann sollten wir Kontakt aufnehmen….
- Das lässt sich natürlich umdrehen zu:
  - Welcher Versicherungsnehmer unseres Versicherungsunternehmens ist gerade unzufrieden wegen Schadenregulierung, Beitragsrechnung, Vertragsverwaltung?
  - Können wir das besser? Dann sollten wir Kontakt aufnehmen….
- Bei wem zeichnen sich Veränderungen ab, die zu einem veränderten Versicherungsbedarf führen:
  - Welcher junge Mensch plant, demnächst auszuziehen?
  - Wer plant, eine Immobilie zu erwerben oder
  - sich ein Auto zu kaufen?
- Anhand der Informationen aus sozialen Medien können potentielle Kunden nach vielen Kriterien segmentiert werden (Sport/Hobbys, Urlaubsziel, …) und gezielt mit Bezug darauf beworben werden. Gleichzeitig lässt sich für jedes, auch das kleinste Teilsegment, grundsätzlich messen, wie erfolgreich die Kundenansprache war (*ROMI = Return On Marketing Investment*).

Die oben aufgeführten Segmentierungen sind neben der gezielten Kundenansprache im Vertrieb auch potentielle Merkmale in der Tarifierung. Bei jedem Segment können bei den entsprechenden Verträgen im Bestand auch das Risiko und die Reaktion auf Preisänderungen gemessen werden.

Für jeden Versicherungsnehmer lässt sich anhand der grundsätzlich im Netz verfügbaren aktuellen Wettbewerberpreise für seine konkreten Merkmale feststellen, wie weit die aktuelle Prämie davon abweicht. Daraus lassen sich Schlüsse auf die Wechselbereitschaft ziehen.

Die vorhandenen Informationen insbesondere zu weichen Merkmalen können mit frei zugänglichen Informationen abgeglichen werden (Fotos der Garage, tägliches weites Pendeln mit dem Auto aber Wenigfahrertarif, …).

2.2. Preisfindung

Im Rahmen der neuen Methoden ergeben sich auch neue Möglichkeiten bei der Preisfindung. In der klassischen Sicht der Preisfindung geht man vom Risikomodell über einen Kostenzuschlag zum technischen Preis, welcher über ein Tarifmodell zum Tarif entwickelt wird. Dies stellt einen statischen Blick auf die Produktionskosten dar und ignoriert den Bezug zum Wettbewerb und das Kundenverhalten.

Diese Sichtweise wurde und wird mit Hinblick auf den Wettbewerb erweitert. Hierbei verändert man den Tarif abweichend vom technischen Preis in Richtung einer marktgängigeren Prämie.

Meist geschieht dies im Hinblick auf Antragsfragen oder den Vergleich der Prämien bzw. Zuschläge in unterschiedlichen Detailgraden. Zusätzlich werden Modelle eingesetzt, welche das Kundenverhalten im Hinblick auf seine Nachfrage beschreiben.

Ein Ansatzpunkt von Big Data ist, diese Modelle durch immer mehr Daten zu verfeinern und somit die Vorhersage zu verbessern.

Dennoch sind diese Anpassungen meist eher statischer Natur und im Vergleich zu aktuell eingesetzten Methoden aus anderen Bereichen wie Onlineshopping, Vergleichsportalen oder Reiseanbietern wie Airlines recht einfach.

Alle folgenden Betrachtungen sollten jedoch immer die Natur des Versicherungsgeschäftes beachten. Eine verkaufte Police ist nicht gleich einem verkauften Produktionsgut, welches bereits hergestellt ist, oder einem Hotelzimmer, welches sonst leer steht. Es gibt im Versicherungsbetrieb selten Positionen, welche „sonst nur herumstehen" wie Waren im Lager oder „ungenutzt sind" wie Hotelzimmer und Flugsitze. Auch spielt im Massengeschäft mit Versicherungen Knappheit seltener eine Rolle als im Vergleich zu beliebten Reisezeiten oder ähnlichem.

Durch das Risikomodell und den Kostenaufschlag werden die Kosten aufgezeigt, welche in Zukunft verdient werden müssen. Die Preisfindung sollte immer den Preis und die zukünftigen Kosten gegenüberstellen. Hier sei jedoch auf die Tatsache hingewiesen, dass der Preis von Versicherungsprodukten über den Schadenerwartungswert und die Vertriebskosten in der Regel zum weit überwiegenden Teil durch reinrassig variable Kosten bestimmt ist. Dies stellt einen erheblichen Unterschied dar zu anderen Segmenten in Industrie, Handel und Dienstleistung, die wesentlich bestimmt sind durch Fixkosten, bei einem teilweise fast zu vernachlässigenden Teil variabler Kosten. Folge ist, dass die Bewirtschaftung von Preiselastizitäten und Kundenverhalten in der Versicherung einen deutlich weniger gewichtigen Einfluss auf den wirtschaftlichen Erfolg hat, als in derartigen Fixkosten-getriebenen Segmenten.

Dennoch gibt es auch im Versicherungsgeschäft bei der Preisfindung bedingt durch die moderne Technik mehr Möglichkeiten als in früherer Zeit. Diese sind beispielsweise

- Ad-hoc-Wettbewerberanalysen im Moment der Preisfindung
- Automatisierte Abbildung des Wettbewerbs für fortlaufende Tarifaktualisierungen
- *Self-Calibrating*-Tarifmaschinen, die ständig einen Abgleich von erwarteten und eingetretenen Abschlüssen durchführen und sich somit selbstständig kalibrieren
- Eliminierung des *Money-Left-on-the-Table*-Effektes; wenn ein Kunde zu einem Preis X abschließen würde, sollte man diesen nicht deutlich unterbieten. Hierzu ist eine genaue Kenntnis des Kundenverhaltens und des Marktumfeldes notwendig
- Echtzeit-Nutzung externer Daten
- Eine Berücksichtigung der Echtzeitdaten kann auch erkennen, ob sich der Kunde soeben intensiv mit dem Preis beschäftigt oder auf einen schnellen Abschluss drängt
- Webseiten, welche auf die individuellen Kundenbedürfnisse angepasst sind.

Ebenso durch neuere Techniken ermöglicht ist die Betrachtung des Kundenbedürfnisses, welches Produkt er wahrscheinlich nachfragen wird. Das Erkennen von sogenannten *Life-Events* hilft dabei eine Produktnachfrage vorherzusagen, aber auch ein Produktbedürfnis auszulösen.

Es wirken also viele Dinge gemeinsam; zum einen die Nutzung neuer Daten zur Modellverbesserung und -erweiterung, aber auch deren Anwendung in Echtzeit und zusätzlich die Verbesserung oder aber auch Automatisierung der Modellierung durch Machine Learning.

2.3. Tarifierung/Produktentwicklung

2.3.1. Risikobewertung

Machine Learning Verfahren können GLM-Prozessen vorgeschaltet werden, um Abhängigkeiten maschinell gestützt zu erkennen. Konkret bieten sich dazu *Tree Based Machine Learning (TBML)* Verfahren an. Neben den bereits bestehenden GLM wird mit einem *TBML* Verfahren ein separates Risikomodell errechnet. Beide Risikomodelle werden miteinander verglichen (z. B. Vergleich *GLM-Beta* mit *TBML-Variable Importance*). Bei Variablen, die im *TBML* einen wesentlichen höheren Erklärungsgrad besitzen als im GLM, wird der Grund genauer untersucht (z. B. mittels Mosaikplots). Die identifizierten Auffälligkeiten rühren i. d. R. aus Abhängigkeitsstrukturen her, die dann an den Analysten des GLMs übergeben werden können, der sie für den herkömmlichen GLM-Pricing- und Vertriebsprozess nutzt.

Ein weiterer Ansatz, Machine Learning im Pricing zu nutzen, ist das effiziente Aufbauen von Modellen der Wettbewerbspreise. Grundlage dafür sind Preise eines Wettberbers für wenige tausend bis zehntausend Nachfrager. Das so erhaltene Marktpreismodell der jeweiligen Wettbewerber kann dann z. B. Rückschlüsse für das eigene Risikomodell erlauben.

2.3.2. Reduktion der Antragsfragen / Ersetzen der Fragen durch „andere" Daten

Versicherer können sich nicht nur durch den Preis von ihren Konkurrenten absetzen. Betrachtet man die Anzahl der Fragen, die einem Kunden bei Abschluss gestellt werden, fällt einem deren erhebliche Menge auf.

- Manche Kunden schreckt die Anzahl ab
- Einzelne Fragen kann der Kunde nicht beantworten, was den Abschluss behindert
- Nicht alle Fragen möchte der Kunde beantworten

Die Anzahl der Fragen liegt zunächst auch daran, dass wir Aktuare das Risiko möglichst genau bewerten können wollen. Verdeutlicht man sich letzteren Punkt, ist nicht das einfache Weglassen von Informationen zielführend – man würde sich sofort dem Risiko der Antiselektion aussetzen. Es geht also nicht um das "Weglassen" von Merkmalen, sondern das Substituieren von beim Kunden erfragten Merkmalen durch andere Merkmale, welche aus alternativen Datenquellen abgeleitet werden können.

Viele Antragsfragen helfen uns, das Risiko möglichst gut zu beschreiben. Sie sind Näherungen. Die Fragen lassen sich oft nach weichen und harten Merkmalen kategorisieren.

Es gibt im Allgemeinen mehrere Möglichkeiten des Vorgehens:

- Alternative Merkmale, welches das vorhandene ersetzen
- Alternative Merkmale, welche das zu ersetzende beschreiben
- Zielgruppen identifizieren, bei denen das Merkmal nicht benötigt wird und somit eine Dynamisierung des Angebotsprozesses, so dass nicht immer alle Fragen gestellt werden
- Vorschlag von erwarteten Antworten
    - (Wieder-)Erkennung des Kunden via Login, ID oder Hash
    - Clevere Defaults, schon mal eingegeben für gewissen Hash
    - Kunden mit diesen Merkmalen hatten häufig auch …

Diese vorangeschriebenen Positionen bedingen das Ausnutzen der neuen *Data Lakes* im Zusammenspiel mit Machine Learning und deren teilweisen Anwendung in real-time. Es ist wenig zielführend, nach fünf Minuten festzustellen, dass ein Kunde schon einmal angefragt hat. Man braucht diese Information sofort im Moment der Anfrage.

Die Lösungen auf diese Frage werden selten für alle Vertriebswege gleich sein.

- Beispielsweise kennt ein Agent / Makler seinen Kunden.
- Auch eine Website mit Login erkennt seinen Kunden.
- Ein mobiles Endgerät erkennt den Standort und kann diesen somit vorschlagen.
- Eingebettete bzw. angehängte Lösungen von sozialen Netzwerken können auch diesem bekannten Werte zunächst als Default vorschlagen.
- Durch direktes Tracking oder via Hash können Kunden Defaults quer durch das Netz vorgeschlagen werden.

Ein Merkmal wird äußerst selten direkt durch ein anderes ersetzt werden können. Hierbei gilt es meist aus einer Vielzahl weiterer Merkmale durch Klassifizierungen eine möglichst gute Approximation abzuleiten.

2.3.3. Telematik

Telematikdaten werden zurzeit im deutschen Markt in der Regel mit heuristischen Algorithmen verwertet, die nicht auf statistischen Zusammenhängen beruhen (z. B. „Wenn schneller als 130 Km/h auf der Autobahn, dann schlechter Telematik-Score"). Bei weiter wachsenden Telematik-Beständen können Machine-Learning-Techniken in der Bewertung der versicherungstechnischen Risiken maßgeblich werden.

Daneben bieten die Telematikdaten eine Information, zu welcher Wochen- und Tageszeit auf welchen Straßen gefahren wurde. Grundsätzlich sind relativ einfach Muster der Fahrwege ermittelbar (überwiegende Verwendung auf weitgehend identischer Route bei Pendlern …). Muster des Fahrverhaltens selbst können aus den Daten ebenfalls abgeleitet werden, allerdings erfordert dies sicher tiefer gehende Analysen. Derartige objektiv gewonnene Informationen können ebenso wie die oben angesprochenen eher heuristischen Scores ähnlich wie klassische Risikomerkmale verwendet werden. Allen Daten gemeinsam ist, dass sie erst im Laufe der Versicherungsperiode zur Verfügung stehen.

Telematikdaten bieten auch die Möglichkeit, z.B. über die spielerische Anbindung von Apps und Social Media, eine schadenmindernde Wirkung auf ihre Nutzer zu entfalten („Ich habe einen noch besseren Score als Du").

Weitere Anwendungen sind die indirekte Werkstattsteuerung, die automatische Unfallmeldung, Serviceleitungen (z. B. Fahrtenbuch) und die Vermarktung weiterer Produkte (z. B. Ölwechsel, Reifen, Wartung); oder außerhalb der Kraftfahrzeugversicherung z. B. Smart Home.

2.3.4. Modelldiagnose

Es sei hier auf den hohen Einfluss der Treffgenauigkeit des Risikomodells auf den wirtschaftlichen Erfolg des Tarifs hingewiesen, wobei diese Treffgenauigkeit in ihrer Relation zu den sonstigen marktüblichen Tarifstrukturen entscheidend ist. Ein Misserfolg kann auch schon drohen, wenn ein eigener Tarif insgesamt einen sehr guten Fit an die Realität aufweist, jedoch in einem Teilsegment deutlich abweicht. Diese Problematik kann speziell dann relevant sein, wenn der neue Tarif auf Basis eines im Vergleich zu den üblichen Vorgehensweisen (GLMs, oft auf Basis von Faktoren, die aufgrund marktweiter Statistiken ermittelt sind) abweicht. Eine Beurteilung eines Tarifes, der auf traditioneller Basis durch Erweiterungen um zusätzliche Daten oder Wechselwirkungen entstanden ist, fällt dem Aktuar sicher nicht allzu schwer. Die Beurteilung eines Tarifes, der rein auf Basis von Machine Learning Verfahren entstanden ist, muss sich damit auseinander setzen, dass diese Verfahren sicher im Detail eine hohe Erklärungskraft besitzen. Die traditionellen Verfahren können jedoch durch ihre explizite Voraussetzung funktionaler Zusammenhänge – dort wo diese die Wirklichkeit gut beschreiben – zusätzlich von diesen Voraussetzungen im Hinblick auf ihre Treffgenauigkeit profitieren und stabilere Ergebnisse liefern. Nicht zu unterschätzen ist, dass die bei der Modellbildung traditioneller Tarife angesetzten funktionalen Zusammenhänge von der ganzen Branche und über Jahrzehnte fortentwickelt wurden. Hier steht der Aktuar bei Einführung eines auf Machine Learning basierenden Tarifs vor einer wirtschaftlich weitreichenden Qualitätsbeurteilung.

2.4. Service

2.4.1. Eingabe Partnerdaten

Bei Eingabe der Partnerdaten (Name, Anschrift, Risikodaten) erfolgt ein sofortiger Abgleich mit eigenen und fremden Datenbeständen. Die eingegebene Postleitzahl wird mit einem Postleitzahlverzeichnis abgeglichen und der zugehörige Ort übernommen, was Rechtschreibfehler vermeidet. Die eingegebene Straße wird mit einem Straßenverzeichnis abgeglichen.

Der eingegebene Name wird mit den bereits vorhandenen eigenen Partnerdaten abgeglichen. Hierdurch werden Dubletten im Partnerbestand vermieden und es ist unter Einbeziehung der Adresse möglich, familiäre Beziehungen zu identifizieren ("Neukundin X ist die Tochter von Y"). Bei diesem Abgleichen sollten Verfahren der Textanalyse (Phonetik, Abstandsmaße, Levenshtein-Distanz) und Häufigkeitsverteilungen von Vor- und Nachnamen berücksichtigt werden. So sind zwei Personen mit identischer Anschrift und dem Namen "A. Spriedelzurst" sehr wahrscheinlich verwandt oder sogar identisch. Zwei Personen mit Namen "M. Müller" mit identischer Anschrift in einem Hochhaus mit über 100 Bewohnern sind wahrscheinlich nicht identisch. Hier können also auch Mikroregionaldaten über Einwohnerzahlen je Anschrift einfließen.

2.4.2. Stornoprognose und Storno-Prophylaxe

Ziel der Stornoprognose ist es, für jeden individuellen Kunden zu prognostizieren, welche Stornoneigung bei ihm besteht. Da es sich hierbei um Bestandskunden handelt, liegen sehr viele interne Daten über den Kunden und den bisherigen Vertragsverlauf vor. Diese müssen zur Beurteilung der Stornoneigung konsequent ausgewertet werden. Beispielsweise hat sich gezeigt, dass neben dem Kundenalter insbesondere die Anzahl der insgesamt versicherten ("angebündelten") Sparten einen wesentlichen Einfluss auf die Stornoneigung hat.

Aber auch die Frage, ob der Vertrag mit einer mehrjährigen Vertragsdauer abgeschlossen wurde, hat einen Einfluss auf die Storno-Prognose: Verträge, welche z.B. drei Jahre lang nicht kündbar waren, weisen nach Ablauf des dritten und des vierten Jahres eine deutlich überdurchschnittliche Stornoquote ("Storno durch Versicherungsnehmer zum Ablauf") auf. Die Erhöhung nach dem vierten Jahr ist z. B. damit zu erklären, dass der Kunde die Frist nach dem dritten Jahr versäumt hat und nun "endlich" aus dem Vertrag aussteigen möchte.

Bei genauer Prognose der Stornoneigung kann das Unternehmen kundenindividuell und rechtzeitig z. B. durch speziell zugeschnittene Anpassungsangebote auf den Kunden zugehen.

2.4.3. Inkasso

Zahlungsausfälle bei der Vereinnahmung der Prämie erzeugen zunächst erheblichen Aufwand in der Verwaltung des Vertrages. Bei Lastschrift-Zahlern fallen Rückläufer-Kosten an, beim Maklerinkasso kommt es zu teils empfindlichen Störungen im Dreiecksverhältnis Kunde-Makler-Versicherer. In jedem Falle entstehen zusätzliche interne und externe Kosten durch das Mahnverfahren und die eventuelle Eintreibung von Außenständen durch externe Inkassounternehmen.

Der gesamte Mahnprozess sollte also möglichst effizient gestaltet werden. Hierzu ist die kundenindividuelle Zahlungshistorie auszuwerten, d. h. über alle mitversicherten Sparten des Kunden sollten die bisherigen Zahlungsausfälle (Erinnerungen, Mahnungen, Kündigungen wegen Nichtzahlung) zusammengetragen und zur Prognose der zukünftigen Gefahr eines Zahlungsausfalls analysiert werden. Kunden, bei denen in der Vergangenheit beispielsweise jede vierteljährliche Zahlung erst nach qualifizierter Mahnung erfolgt ist, können z. B. auf jährliche Zahlung per Lastschrift umgestellt werden. Auch könnte bei "notorischen Nichtzahlern" der Mahnprozess beschleunigt werden, indem auf ein freundliches Hinweisschreiben verzichtet wird und sofort mit gesetzlicher Frist die qualifizierte Mahnung ausgesprochen wird.

Zudem stellt das bisherige Zahlungsverhalten auch eine Bonitätsbewertung des Kunden dar und deren Einfluss auf das versicherungstechnische Risiko kann z. B. in die Bestandssteuerung einfließen.

Die Hinzunahme externer Bonitätsdaten mag im Neugeschäft sehr hilfreich sein, bei der Bewertung der eigenen Bestände sollte in jedem Fall zunächst der vorhandene Datenstand ausgeschöpft werden.

2.4.4. Beschwerdemanagement

Das aufsichtsrechtlich geforderte Beschwerdemanagement kann durch den Einsatz von Text- und Spracherkennungsverfahren unterstützt werden. Die eingehende Korrespondenz (E-Mail, Fax, Brief etc.) von Kunden oder Maklern kann mittels Textanalyseverfahren untersucht werden. Dabei wird nicht nur nach Schlüsselwörtern ("Unverschämtheit", "Frechheit", "unmöglich", "erneut" etc.) gesucht, sondern die Dokumente können daraufhin klassifiziert werden, ob der Sprachstil von Aggression, Enttäuschung oder Androhung geprägt ist. Indem die bisherige Korrespondenz mit dem Kunden hinzugezogen wird, kann zudem eine Einschätzung der Eskalationsgeschwindigkeit erfolgen.

Telefonanrufe (z. B. im Call-Center) können in Echtzeit auf die Gesprächsdynamik des Anrufers hin bewertet werden und es kann ein Hinweis an den Call-Center-Mitarbeiter erfolgen, dass das Gespräch in eine spezialisierte Abteilung weitergegeben werden sollte. Dabei kann zum einen der gewählte Telefonzugang ausgewertet werden ("Ist die Telefonnummer des Anrufers bekannt?"), zum anderen kann im Hintergrund die gesamte Kundenhistorie bewertet werden und bei Premiumkunden wird die Weiterleitung zu den Spezialisten evtl. früher eingeleitet als bei Kunden mit nur einer einzigen, niedrigpreisigen Police.

2.4.5. Steuerung Marketingmaßnahmen

Bei Versicherungsunternehmen mit Programmen "Kunden werben Kunden" ist es sinnvoll, nur bestimmte Kunden auf dieses Programm aufmerksam zu machen, z. B. in Werbebriefen oder im Rahmen der Jahresrechnung.

Kunden, deren Verträge selber schadenbelastet sind oder die in der Vergangenheit durch Zahlungsschwierigkeiten auffällig geworden sind, geben eventuell Empfehlungen in ein ähnlich gelagertes Umfeld ab. Hier könnte der Hinweis in der Jahresrechnung entfallen. Auch die Darstellung der Website könnte gesteuert werden, d.h. wenn ein solcher Kunde z. B. via Cookie erkannt wird, entfällt ein ansonsten eingeblendeter Hinweis auf eine aktuelle "Werberaktion".

2.4.6. Interne Ressourcenplanung

In einigen Versicherungssparten gibt es jahreszeitliche Häufungen in der Vertrags- und Schadenbearbeitung, z. B. Kraftfahrt Ende November wegen Hauptfälligkeit, Wohngebäude wegen Frost im Winter etc. Diese Peak-Phasen können bei der Personalplanung berücksichtigt werden, um etablierte Qualitätsstandards einzuhalten. Bei eventuell eingebundenen Dienstleisternetzwerken können zusätzliche Kapazitäten reserviert werden. Diese Ergebnisse basieren im Wesentlichen auf bereits im

Versicherungsunternehmen vorhandenen Daten, welche aber meist nicht in auswertbarer Form vorliegen.

Ergänzend dazu kann z. B. durch Auswertung von Nachrichtenfeeds das plötzliche Aufkommen von Kumulereignissen (Sturm, Hochwasser, Hagel) erkannt werden. In der Folge könnten kurzfristig Leistungsangebote für die Kunden eingerichtet werden, z. B. Notfall-Callcenter, Schaden-Schnellregulierungszentren etc.

2.5. Schaden / Betrug

Die Einsatzgebiete von Data Science bezüglich Schadenbearbeitung und Betrugsabwehr scheinen aus heutiger Sicht vielfältig. Im Folgenden haben wir mögliche Anwendungsbeispiele aufgezählt.

2.5.1. Schadensvermeidung

- Frühwarnsysteme, z. B. durch automatische Auswertung und Verarbeitung von *lokalen* Wetterdaten bzw. anderen lokalen Ereignissen (z. B. Einbruchsserien in Wohngebieten) → Vermeidung von Hagelschäden an Autos, Einbrüchen …

- Soziale Medien nutzen (Hinweise zur Schadensvermeidung) und auswerten (Kommunikation in sozialen Netzwerken z. B. über lokale Ereignisse wie aufziehende Gewitter etc.)
    - personalisierte Textnachrichten an Versicherungsnehmer schicken, da diese besser wahrgenommen werden als eine allgemeine Wetterwarnung.
    - *Beispiel:* „Wussten Sie, dass es momentan schüttet, aber das Unwetter in 18 Minuten vorbei sein wird?" → Der Angestellte, der gerade nach Hause gehen wollte, bleibt länger im Büro und wartet das Unwetter ab.

- Real-Time-Monitoring von besonders exponierten Risiken (Überwachungssysteme) → Daten auswerten und künftige Schäden vermeiden (*Predictive Maintenance*)
    - *Beispiel:* Real-time-Messdaten von Maschinen zusammen mit tatsächlichen Schadensdaten als Input für ein Machine Learning Modell verwenden. Dies kann dann zur Vorhersage von Defekten genutzt werden, d. h. man hat die Möglichkeit Bestandteile der Maschine auszutauschen so lange sie noch funktionieren und reduziert somit die tatsächliche Schadenszahl.

- Automatische Erkennung von Bedrohungen
  - Was könnte das nächste Katastrophenszenario sein?
  - *Beispiel 1:* Welche chemischen Substanzen haben das Potenzial, eine riesige Klagewelle wegen Gesundheitsschäden zu erzeugen?
    Mit Hilfe von Data Mining wissenschaftliche Veröffentlichungen untersuchen und Korrelationen zwischen bestimmten chem. Substanzen und Krankheitsbildern herstellen
  - *Beispiel 2:* Cyber-Versicherung. Die Plattform *Cyence* liefert in Echtzeit Informationen pro Unternehmen zu dessen Attraktivität und Verwundbarkeit gegenüber Hacker-Angriffen. Berechnung der Scores u. a. durch Sentiment Analysis in sozialen Netzwerken und automatisierte Scans von Hacker-Foren z.B. im Dark Net.

- Suchen und Erkennen von erklärenden Faktoren für Schäden, z. B. Einbrüche (Gibt es hier einen Zusammenhang mit der Lage des Hauses, Sackgassen etc.?)

- Kfz-Versicherung: Beobachtung/Analyse von Fahrerverhalten → Fahrer „in Spur" bringen
  - *Beispiel:* Wildunfälle vermeiden durch Verschneidung von Fahrtstrecken mit Waldbeständen

- Schutzmaßnahmen zur Vertragsbedingung machen
  - *Beispiel 1:* Gebäude in Überschwemmungsgebieten können nur versichert werden, wenn die vereinbarten Schutzmaßnahmen ergriffen sind.
  - *Beispiel 2:* bestimmte Gebäude können nur versichert werden, wenn eine Kamera mit Gesichtserkennung installiert ist

- Überprüfung von Schutzmaßnahmen mit Hilfe von Drohnen (insbesondere in der Industrieversicherung)

2.5.2. Schadenfrüherkennung

Schadenfrüherkennung speziell in der (fakultativen) Rückversicherung:

- Durch automatisiertes Durchforsten von Internet-Quellen erhält der Rückversicherer deutlich früher, einfacher und billiger Kenntnis über Schäden, die möglicherweise versicherungsrelevant sein können.
- Dies erlaubt zunächst Transparenz über Schadengeschehen, unabhängig von Beteiligungsfragen.
- Ein Abgleich mit internen Bestandssystemen klärt dann die Frage einer möglichen Rückversicherungs-Haftung.

- Das ermöglicht Schadenbearbeitern ein effizientes Handeln im *Opportunitäts-fenster* – dem Zeitraum, in dem Entscheidungen bezüglich Reparatur/Wiederherstellung oder Anschaffung von Ersatz gefällt werden (ohne diese Möglichkeit bleibt dem Rückversicherer nur, den getroffenen Entscheidungen zu folgen und zu bezahlen, was vereinbart wurde)
- Im Idealfall kann der Rückversicherer sogar den Erstversicherer über einen Schaden in dessen Portfolio informieren.

Schadenfrüherkennung in der (technischen) Erstversicherung, Produkt-Haftpflicht:

- Schadenfrüherkennung in technischen Versicherungen: Sensoren an Maschinenteilen können im Idealfall bereits dann auf Schäden hinweisen, wenn diese noch im Entstehen sind (z. B. Schwingungs-Sensoren in Generatoren oder Waggon-Achsen). Dies ermöglicht (im Idealfall) das vorbeugende Austauschen von Teilen (→ *Predictive Maintenance*, siehe Schadenvermeidung); zumindest aber lassen sich Schäden minimieren und sofort erkennen
- Automatisierte Qualitätsprüfung von Produkten: Mittels optischer, akustischer oder sonstiger geeigneter Sensoren kann automatisiert geprüft werden, ob Erzeugnisse den Qualitätsansprüchen genügen. Damit lassen sich ggf. Produkthaftungs-Ansprüche vermeiden oder zumindest erkennen, so lange sie noch klein sind
- In beiden Fällen kann das Versicherungsunternehmen dem Versicherungsnehmer nach Installation der Sensoren und Überwachungs-Systeme billigere Deckung anbieten

2.5.3. Betrugserkennung

Schätzungen zufolge beläuft sich der jährliche Schaden durch Versicherungsbetrug in Deutschland auf etwa 4 – 5 Mrd. Euro. Diese Schätzung signalisiert, dass in der Betrugsabwehr enorme Potenziale für jede Versicherungsgesellschaft liegen. Die Bandbreite reicht dabei von vorsätzlicher Brandstiftung über organisierten Kfz-Betrug (z. B. *Berliner Modell*) bis hin zu so genannten Bagatellschäden wie der vom Versicherungsnehmer selbst beschädigten Brille, die bei der Haftpflichtversicherung des Nachbarn gemeldet wird. Dies bedeutet, dass die Betrugserkennung geeignet sein muss, sowohl Gelegenheitstäter als auch organisierte Gruppen aufzudecken.

Die Anwendungsmöglichkeiten zur Betrugserkennung beschränken sich hierbei nicht nur auf den Bereich der Schadenmeldung, ebenso sind die Erkenntnisse auch im Bereich des Underwritings nutzbar.

Die Effektivität der Betrugsabwehr kann durch die Erweiterung der Betrugsbewertung auf die Risikoprüfung und Bestandsverwaltung deutlich verbessert werden. Dies kann erreicht werden, indem nach der Risikoprüfung auffällige Policen gekennzeichnet werden, so dass bei einer Schadenmeldung eine entsprechend sorgfältige Prüfung durchgeführt wird.

Im Rahmen der Analysen zur Betrugserkennung erscheint es sinnvoll, auf viele unterschiedliche Datenquellen (wie z. B. Soziale Netzwerke, internes Datensystem) zurückzugreifen. Als Ergänzung zu Expertenregeln können auch datenbasierte Regeln für die Betrugserkennung verwendet werden. Die Datenregeln ermöglichen die Aufdeckung neuer, auch den Experten vormals unbekannter Betrugsformen.

Eine Vielzahl an unterschiedlichen Methoden bietet sich für die Anwendung der Betrugserkennung an, z. B. können mit Hilfe von Textmining oder Spracherkennung Unfallbeschreibungen hinsichtlich bestimmter Auffälligkeiten analysiert werden. Zusätzlich können weitere Datenquellen wie Unfallbilder, Satellitenbilder, Street View oder ähnliches automatisiert analysiert werden. Während ein Mensch meist nur dreidimensional denkt, schafft es eine Maschine Gemeinsamkeiten und somit Auffälligkeiten in viel höheren Dimensionen zu erkennen.

Ebenso ist die Nutzung weiterer externer Datenquellen zur Betrugserkennung im Hinblick auf Falschdeklaration vorstellbar.

Bekannte Betrugsfälle können genutzt werden um Prognosemodelle abzuleiten.

Im Rahmen der Schadenbearbeitung bietet eine effiziente Betrugserkennung, neben der Möglichkeit möglichst viele unberechtigte Forderungen zu erkennen und abzuweisen, ebenso die Möglichkeit berechtigte Forderungen schnell abzuwickeln. Hierdurch lassen sich zum einen die Kosten der Schadenbearbeitung reduzieren und zum anderen die Kundenzufriedenheit steigern.

Zur Optimierung der Betrugsbekämpfung ist ein Prozess zur Re-Evaluierung von bestehenden Modellen und Kalibrierung von eingesetzten Regeln durch die Fachanwender einzuführen.

2.5.4. Schadenregulierung

Die operativen Aspekte der Schadenregulierung können einen deutlichen Einfluss auf aktuarielle Analysen wie etwa bezüglich. der Reserveposition haben.

Dabei geht es um Abwicklungsgeschwindigkeit, Qualitätskontrollen (automatisiert und manuell inklusive Betrugserkennung), Qualifizierung und Weiterentwicklung der Schadenbearbeiter, Personalkapazität mit der Implikation auf Arbeitsdruck, technische Ausstattung, Organisationsstruktur, etwaige externe Unterstützung z. B. durch Reha-Management, Zufriedenheit der Mitarbeiter und Arbeitsatmosphäre, Zufriedenheit der Geschädigten/Kunden, Kulanz, Kosten der Schadenregulierung. Des Weiteren können die Aufwendungen für die einzelnen Schäden i. d. R. nach einzelnen Leistungskennziffern herunter gebrochen werden.

Es ist klar, dass nicht alle dieser Dimensionen gleichzeitig optimiert werden können. So könnte eine Intensivierung der Qualitätskontrollen zu Lasten der Abwicklungsgeschwindigkeit gehen und umgekehrt eine Erhöhung der Geschwindigkeit zu erhöhten Schadendurchschnitten aufgrund vernachlässigter Kontrollen führen.

Ähnliches dürfte für Kulanz im Widerspruch zur Zufriedenheit der Geschädigten/Kunden gelten.

All diese Dimensionen mit all ihren Ausprägungen und ihrer durchaus verschiedenen Messbarkeit können im Sinne einer mindestens zweidimensionalen Optimierung analysiert werden: Geschwindigkeit/Qualität, Kulanz/Zufriedenheit etc.

Dabei können insbesondere durch eine hochfrequente Erhebung (z. B. Arbeitsbelastung auf Tagesbasis insgesamt sowie für Gruppen von Mitarbeitern) schnell recht große Datenmengen entstehen.

Darüber hinaus ist denkbar, für Validierungen von Schilderungen des Schadenhergangs z. B. eine Real-Time-Sicht auf den Schadenort oder auch Äußerungen von Beteiligten im Internet heranzuziehen (sofern dies datenschutzrechtlich zulässig ist).

2.5.5. Effektivität der Schadenregulierung

Bisher wird üblicherweise in der Schadenregulierung nur die Schnelligkeit der Regulierung und gegebenenfalls die Kundenzufriedenheit gemessen (z. B. *Net Promoter Score*). Durch die Ergänzung einer verlässlichen Messgröße für die Effektivität/Qualität der Schadenregulierung kann das Steuerungsdreieck vervollständigt werden.

Dabei wird auf Grundlage der Informationen einer Vielzahl von Schäden eines Portfolios gearbeitet. Aus den geschlossenen Schäden der Vor-Vorperiode wird ein Modell für den erwarteten Schadenbedarf der in der Vorperiode geschlossenen Schäden ermittelt. Anschließend wird der real benötigte Schadenbedarf der Schäden der Vorperiode (Realisation) mit dem modellierten Schadenbedarf der gleichen Schäden verglichen. Auf den resultierenden Vergleichssätzen erfolgt eine Analyse nach Auffälligkeiten. (Z. B. „Gibt es bestimmten Werkstätten, die in Kombination mit gewissen Gutachtern in der Realisation besonders von der Erwartung abweichen?“)

2.5.6. Belegprüfung

Bei der Schadenbearbeitung werden die eingereichten Unterlagen (Belege) in der Regel durch den Schadensachbearbeiter auf Schlüssigkeit und Angemessenheit hin geprüft. Dies sind in der Krankenversicherung z. B. verordnete Medikamente und Therapien, durchgeführte Operationen und Heilbehandlungen sowie sonstige erbrachte Leistungen.

In der Krankenversicherung sind jedoch einige Belege nicht erstattungsfähig. So sind Akkupunkturleistungen nur bei bestimmten Diagnosen oder nur in bestimmten versicherten Tarifen mitversichert, ähnliches gilt z. B. für Nahrungsergänzungsmittel und Vitamine. Oder die Positionen sind zwar grundsätzlich erstattungsfähig, die berechneten Kosten sind aber zu hoch, z. B. Chefarztleistung ohne entsprechenden Versicherungsschutz oder 3,5-facher Satz bei einfacher Fallkonstellation.

Hier ist es sinnvoll, dass der Schadensachbearbeiter nicht nach Bauchgefühl oder anhand von länglichen Prüflisten vorgehen muss, sondern der eingereichte Beleg in einer automatisierten Vorprüfung analysiert wird und Hinweise auf eventuell intensiver zu prüfende Belege ausgesteuert werden. Hochqualifizierte Schadensachbearbeiter sind ein knappes Gut und es erscheint daher sinnvoll, Vorgänge mit hohem Prüfbedarf bevorzugt bearbeiten zu lassen.

Im Bereich der Kraftfahrt- und Sachversicherung gibt es hier bereits weiter fortgeschrittene professionelle Ansätze. So werden im Bereich der Prüfung von Kraftfahrt-Schäden die eingereichten Rechnungen oder Kostenvoranschläge eingescannt und die enthaltenen Positionen per *optical character recognition (OCR)* möglichst vollautomatisiert extrahiert.

Anschließend erfolgt ein Abgleich u. a. unter folgenden Aspekten:

- Regional differenzierte Stundenverrechnungssätze (je Leistungsart)
- Ersatzteilpreise
- Herstellervorgaben zur Reparatur (z. B. Anzahl Klebeleisten bei Frontscheibe)
- Bisheriges Abrechnungsverhalten der Werkstatt

In einem mehrdimensionalen Modell wird dann die "Prüfbedürftigkeit" des einzelnen Vorgangs bewertet; die Bearbeitung der Vorgänge erfolgt in dieser Reihenfolge.

2.6. Customer Journey

Customer Journey definiert sich durch die verschiedenen Prozesse und Phasen, die ein (potentieller) Kunde durchläuft.

Bevor jemand zum Kunden wird, befasst diese Person sich als Interessent für ein bestimmtes Produkt mit der Beschaffung von Informationen wie etwa Preis, Deckungsumfang etc. Hierzu stehen verschiedene Kanäle und Medien zur Verfügung.

Für unsere Zwecke stehen vor allem datengenerierende Prozesse im Fokus – hier vor allem internetbasierte Zugangswege. Diese umfassen Websites einzelner Versicherer, Vergleichsportale, Informationsaustausch (z. B. *MotorTalk*) oder weitere, die im direkten Zusammenhang mit Versicherungsprodukten stehen wie etwa Gebrauchtwagenportale, aber auch generelle Bewegungen/Verhaltensweisen im Internet.

All diese Daten können im Grundsatz herangezogen werden, um Analysen/Charakterisierungen zu generieren. Dies umfasst Fragestellungen wie etwa, welcher Kundentyp welches Produkt kauft oder Preissensitivitäten mit Blick auf Abschlusswahrscheinlichkeiten oder Modellierung von Kommunikation (Wer reagiert positiv/negativ auf Mailerinnerungen? Wer braucht als Kaufanreiz ein Rabattangebot und wer nicht?).

Nach Abschluss einer Versicherung durchläuft der Kunde weitere Phasen wie etwa Service (z. B. Adressänderungen, Anpassung des Deckungsumfangs), Kundenkommunikation (z. B. Newsletter), Upselling (falls der Kunde als dafür prädestiniert modelliert wurde), Schadenregulierung, Erneuerung (beachte Preiselastizitäten, die zu Kündigung, Tarifwechsel o. ä. inklusive zugehöriger Kommunikation via Mail, Brief, Telefon führen können).

Zu nahezu all diesen Vorgängen können auf Basis der direkt zum Versicherungsvertrag gehörigen Daten, aber auch weitere z. B. internetbasierter Daten Modellierungen erfolgen. Damit kann etwa über Art/Inhalt der Kommunikation, etwaiger Preisanpassung, möglicher Upselling-Maßnahmen entschieden werden.

Dabei sind die Grenzen zwischen klassischer Modellierung wie etwa der Tarifierung (z. B. ein Tarif pro Jahr) und internetbasierten Prognosen von Kundenverhalten fließend. So wäre z. B. eine Neugeschäftstarifierung analog der für Benzinpreise (nicht nur täglich, sondern tageszeitabhängig) eine Zwischenstufe. Grundsätzlich ist es natürlich möglich, verschiedene Herangehensweisen zu verbinden wie etwa Modifizierung des klassischen Tarifs durch internetbasierte Prognosen auf Basis von Kundencharakterisierungen.

Dabei ist zu beachten, dass die Relevanz bzw. Aussagekraft der Daten von der Verwendung abhängen kann. So könnten zusätzliche Informationen zum Internetverhalten für Kommunikation entscheidender sein als für die Beschreibung des Risikos.

2.7. Controlling und Management Information

Im Rahmen der Unternehmenssteuerung nimmt das Controlling Planungs-, Koordinations- und Kontrollaufgaben wahr, um die Unternehmensführung mit den notwendigen Instrumenten und Informationen zu versorgen. Unter dem Begriff *Business Intelligence (BI)* werden Verfahren und Prozesse zur systematischen Analyse (Sammlung, Auswertung und Darstellung) von Daten in elektronischer Form verstanden. Es werden hierbei wichtige *Key Performance Indikatoren (KPIs)* analysiert, die dann die Grundlage für die Unternehmenssteuerung bilden. Neben den klassischen Themen wie Reporting und Planung ist durch Solvency II ein weiterer Schwerpunkt in Richtung Risikomanagement gesetzt worden. Das Risikomanagement erfordert umfangreiche Simulations- und Szenarien-Analysen.

Big Data bietet hierbei eine Reihe von Anwendungsmöglichkeiten:

- Mit Hilfe von Big Data können zukünftig bisher unberücksichtigte externe Einflüsse in die Unternehmenssteuerung mit aufgenommen werden. Da die Daten in unterschiedlichen Strukturen vorliegen, ist es notwendig, sich mit den Auswertungsmöglichkeiten auseinanderzusetzen. So besteht mit Hilfe von Text Mining die Möglichkeit der analytischen Erschließung von Texten. Dies bedeutet aber auch, dass eine Vielzahl an unstrukturierten Textdaten in kurzer Zeit zu verarbeiten ist. Hierdurch versetzt sich das Unternehmen in die

Lage, gesellschaftliche, politische oder wirtschaftliche Entwicklungen zu erfassen und diese externen Erkenntnisse wiederum in Kennzahlen zu übersetzen. Eine Verknüpfung dieser Kennzahlen mit internen Daten kann genutzt werden um wertvolle Schlüsse für die Planung abzuleiten.

- Durch erhöhte Rechenleistung kann eine Vielzahl von Einflussfaktoren und deren Wechselwirkungen in verschiedenen Szenarien abgebildet werden. Somit ist man in der Lage, zur Beurteilung der Risikosituation unterschiedliche Marktentwicklungen simulieren zu können.
- Im Reporting können Zusammenhänge zwischen unstrukturierten Massendaten in aussagekräftigen Grafiken visualisiert werden. Mittels statistischer Auswertungen können beispielsweise die Gründe für Plan-Ist-Abweichungen durch die Analyse der zugrundeliegenden Daten identifiziert und unter Verwendung einer automatischen Texterstellungssoftware als Erläuterungen in den Report integriert werden.

Diese Anwendungsmöglichkeiten verdeutlichen, dass sich das Anforderungsprofil der Mitarbeiter im Controlling verändern wird. Die Komplexität von Big Data macht es erforderlich, Zusammenhänge und Beziehungen zwischen den unterschiedlichsten Einflussfaktoren zu erkennen und zu beurteilen. Nur so ist der Controlling-Mitarbeiter in der Lage die richtigen Schlüsse aus den Daten ziehen und basierend hierauf fundierte Handlungsempfehlungen zur Unternehmenssteuerung auszusprechen. Dies bedeutet, dass der Mitarbeiter über notwendige Statistikkenntnisse verfügen sollte. Zusätzlich spielt die IT-Kompetenz eine wichtige Rolle.

2.8. Risikomanagement

Im Risikomanagement bieten sich vor allem in der Kumulkontrolle Anwendungsbereiche, wie beispielsweise

- Automatische Entdeckung von Hot Spots mit Hilfe von komplexen Algorithmen (z. B. dynamische Ringanalyse)
- Warnsysteme / automatische Hinweise, wenn die Zeichnung bestimmter Risiken dazu führt, dass ein Kumul entsteht
- Kumulkontrolle nicht nur im Hinblick auf Naturgefahren, sondern auch im Hinblick auf Feuer (Kreise mit kleinen Radien → sehr große Datenmengen, die überprüft werden müssen)
- Industrieversicherung: Analyse von Lieferbeziehungen (Lieferketten)
    - Kann der Ausfall eines Lieferanten mehrere Produzenten oder sogar Branchen treffen?
    - Die Beziehungen zwischen Produzenten und Lieferanten über mehrere Ebenen (*Tier 1*, *Tier 2* etc.) kann über unstrukturierte Daten (z. B. mittels *Web-Crawling*) ermittelt werden.

    - Datenscan öffentlicher Nennungen im Internet lässt erkennen und im zweiten Schritt visualisieren welche Unternehmen miteinander Lieferbeziehungen unterhalten.
- Höchstschadensschätzung: *Was-wäre-wenn-* (bzw. *Worst-Case-*)Szenarien generieren und analysieren
    - Was könnte das nächste Katastrophenszenario sein?

3. Daten

Die folgenden Ausführungen benennen wichtige Voraussetzungen, damit einerseits eine sachgerechte Statistikarbeit zu Kalkulation und Nachkalkulation möglich ist, andererseits Strukturen bereitstehen, die ermöglichen, dass der Aktuar seine Kalkulationsergebnisse umsetzen kann.

Die Beschaffung oder sinnvolle Verwendung von externen Daten (wie z. B. OpenStreetMaps, Social Media) oder von sehr großen Datenmengen (wie z. B. Telematik-Daten) ist nicht Gegenstand dieses Kapitels.

3.1. Einführung

Die Verantwortung des Aktuars in der Tarifgestaltung in der Schaden- und Unfallversicherung ist in [1] und [8] allgemein beschrieben. Die Wahrnehmung dieser Verantwortung beruht in entscheidenden Teilen darauf, dass die relevanten fachlichen und technischen Strukturen geeignet konzipiert sind und sachgerechte Prozesse und Rollendefinitionen vorliegen. Es muss sichergestellt sein, dass eine Kalkulation und Nachkalkulation auf der vorhandenen Datenbasis möglich ist und Strukturen für eine Umsetzung der aktuariellen Kalkulationsergebnisse in den Tarifen vorhanden sind.

Grundsätzlich sollte der Aktuar sich in einem eng gefassten Rollenverständnis darauf verlassen können, dass unter Federführung anderer hierfür zuständiger Organisationseinheiten (beispielsweise Fachbereiche, Betriebsorganisation oder IT) die Prozesse, Rollen und sowohl fachliche als auch technische Strukturen definiert sind, die die relevanten aktuariellen Anforderungen sicherstellen. Da jedoch die Praxis zeigt, dass derzeit hier in vielen Unternehmen Regelungslücken bestehen, versuchen wir an aus unserer Sicht entscheidenden Punkten dem Aktuar strukturierte Hilfe bei der Formulierung seiner Anforderungen zu geben. Bei genauem Lesen wird man gewisse Querverbindungen zu Art. 19 bzw. 272 der Delegierten Verordnung erkennen.

Hier soll dem Aktuar sowohl ein Überblick über die Probleme und Lösungsansätze als auch eine Argumentationsbasis für die Vertretung seines Standpunktes gegeben werden.

Gegenstand der Ausarbeitung ist ein sachgerechter Umgang mit den üblichen internen Versicherungsdaten. Deren sachgerechte Bereitstellung ist auch eine Voraussetzung für die Nutzung von Big Data.

3.2. Problemstellung

3.2.1. Veränderungen in der Produktgestaltung

In der klassischen Versicherung wird jedes versicherte Risiko in einem Vertrag festgehalten. Für jedes Risiko existiert im Prinzip aufgrund der festen Struktur ein Datensatz (ggf. normalisiert in mehreren Tabellen abgelegt). In diesem Datensatz sind alle für Tarifberechnung und Statistik notwendigen Informationen gespeichert.

Die moderne Produktgestaltung kennt demgegenüber einerseits die Granularisierung der Leistungen in zu- und abwählbare Teil- und Zusatzleistungen und andererseits die Zusammenfassung vielfältiger Leistungsinhalte in einer Police.

Als Beispiele für die Problemstellungen seien folgende Konstellationen angegeben:

Beispiel 1: As-you-like-it-Versicherung

> Die bekannte Gebäudeversicherung wird im Interesse einer vom Kunden nach Bedarf zusammenstellbaren Leistung in Bausteine zerlegt, die die Versicherung der Mauern, des Verputzes oder des Daches zum Gegenstand haben und jeweils in den Varianten *Ersatz des Zeitwertes* bzw. *Wiederherstellung zum aktuellen Stand der Wissenschaft* gewählt werden können.

Beispiel 2: All-inclusive-Versicherung

> Für Gewerbebetriebe sind pauschal alle Sachen gegen alle Arten von Beschädigungen versichert.

Beispiel 3: Standard-Police

> Klassische Sachversicherung von Gewerbebetrieben; jeder Bereich (jedes eigene Gebäude) ist mit seinen spezifischen Daten zu erfassen; die Underwriter können jedoch systembedingt die Betriebsart und die Details der Deckungen wie Selbstbehalt und Höchstentschädigung nur einmal eingeben. Die Produktstruktur entspricht hierbei der Meldeanleitung der Risikostatistik Sach des GDV.

Zusätzlich hat sich die Tarifberechnung von der klassischen (Papier-)Tariftabelle über das Produkt von Risikofaktoren zu multivariaten nichtlinearen Funktionen der Eingabeparameter weiterentwickelt.

Im Zuge von Big Data könnten die Eingabeparameter von Tarifrechnern selbst aus großen komplexen Datenmengen abgeleitet werden. Die Verwendung der vom GDV unverbindlich bekannt gegebenen *ZÜRS-Zone* in der Sachversicherung ist hierfür ein aktuelles Beispiel: Die Zone wird online bei Abfrage aus den gespeicherten Daten der Landkarte mit den definierten *ZÜRS-Zonen* und der Koordinate des interessierenden Objektes mit Hilfe geo-informatischer Berechnungen hergeleitet.

3.2.2. Anforderungen des Aktuars

In seiner Verantwortung in der Tarifgestaltung (siehe [8]) ist der Aktuar an Folgendem interessiert:

- Er kann die notwendige Statistik betreiben, um die ihm obliegende Aussage zur Auskömmlichkeit zu treffen. Die notwendigen Daten sind in der geeigneten Granularität strukturell verfügbar und erfasst.
- Die Ergebnisse seiner Berechnungen sind auch umsetzbar – im Tarifbuch und am besten auch im Risikomodell. Die fachlichen und technischen Strukturen stehen in einer Tiefe zur Verfügung, die der kalkulierten Granularität entspricht.

Um diese Verantwortung wahrnehmen zu können, muss sich der Aktuar darauf verlassen, dass Deckungen sachgerecht konzipiert werden und dass insbesondere

- ihre sachgerechte Regulierung gesichert ist
- die notwendigen Daten vorhanden sind
- die sachgerechte Statistik gesichert bzw. eine Abweichung verantwortet ist
- die notwendige Berechnung ermöglicht ist.

Damit diese im Wesentlichen auf die Ebene der Produktbausteine gerichtete Forderung des Aktuars erfüllbar ist, müssen die genannten Themen in den Ebenen oberhalb der Produktbausteine sachgerecht abgebildet werden.

Hierzu sollten alle Zusammenfassungen mit ihrer Datenstruktur und Funktionalität exakt in einem Produktstrukturmodell niedergelegt sein. Die Auswirkungen auf das Zeitenmodell und die Datenflüsse Produktbaustein-übergreifender Daten in allen Geschäftsvorfällen sollten definiert und die Wechselwirkung mit der Bedingungsgestaltung bekannt sein.

In diesem Kontext sollte der Aktuar auch seine im Zuge der Produktgestaltung zukünftig entstehenden Anforderungen an die Produktbaustein-übergreifende Datennutzung in das bestehende Produktstrukturmodell einbringen können.

Die oben genannten Tendenzen werfen in der Produktgestaltung mannigfache Probleme auf, deren Lösung Voraussetzung für die sachgerechte Arbeit des Aktuars ist.

Genau genommen handelt es sich hier um die Aufgabe der Facharchitekten. Ein großer Teil der auftretenden Probleme steht jedoch im Zusammenhang mit der Arbeit und den Anforderungen des Aktuars. Daher ist in der Praxis dieser gefordert, im Dialog mit Fachlichkeit und IT meist mehr als weniger eigeninitiativ darauf hinzuwirken, die Verfügbarkeit und Vergleichbarkeit von Daten im zeitlichen Verlauf als wesentliche Grundlage des Versicherungsprodukts sicherzustellen. Gegebenenfalls muss er auch unter Hinweis auf seine Verantwortung eine geeignete Umsetzung durchsetzen.

3.3. Umsetzungen

3.3.1. Anforderungen an die Datenhaltung

Die Aussage des Aktuars zur Auskömmlichkeit lebt in den beiden Dimensionen

- Leistungsinhalt („was") und
- Risikomerkmale („wer").

Eine Zwischenstellung nimmt die Entschädigungsfunktion („wieviel") ein, die von der zugesagten Leistung die tatsächlich zu zahlende Entschädigung (eingeschränkt durch Selbstbehalt oder Höchstentschädigung) ableitet.

Der Aktuar stellt allen gleichförmigen Leistungszusagen die daraus erwachsenden Entschädigungen gegenüber und differenziert diese nach den Risikomerkmalen. In Anlehnung an diese Logik sehen moderne Bestands- und Schadensysteme das Strukturelement *Produktbaustein* oder *Deckung* für jeden eindeutig definierten (atomaren) Typ von Leistungszusagen vor. Je Produktbaustein ist die relevante Liste der *Risiko-* bzw. *Tarifmerkmale* zu definieren und die zugehörigen Modelle zu berechnen. Im gewählten Tarifmodell können die für die Tarifberechnung nötigen Eingabeparameter andere als die des Risikomodells sein – daher die Unterscheidung zwischen Risiko- und Tarifmerkmalen. Während erstere ganz vordergründig für die konkrete Tarifberechnung nötig sind, kann eine Nachkalkulation des Risikomodells nur geschehen, wenn die ggf. abweichenden Parameter des Risikomodells ebenfalls erfasst werden.

Als Voraussetzung für die Nachkalkulierbarkeit muss jedem Produktbaustein in einer äquivalenten Struktur im Schadensystem die zugehörige Leistung gegenübergestellt werden. Dies setzt voraus, dass die Leistung eindeutig dem auslösenden Produktbaustein zugeordnet werden kann und dies in der konkreten Regulierung auch lückenlos durchgeführt wird. Zur Sicherstellung der Nachkalkulierbarkeit sollte der Aktuar in seiner Verantwortung darauf hinweisen, dass hierfür ein verbindlicher Prozess mit den entsprechenden Verantwortlichkeiten installiert wird.

Bei einer Neukalkulation von Produkten weichen häufig die neuen Produktbausteine von den bestehenden leicht ab (z. B. Erhöhung einzelner Höchstentschädigungsgrenze, Verzicht auf den Selbstbehalt für einzelne Leistungspositionen). In diesen Fällen ist es hilfreich, wenn eine Leistung nicht nur dem gesamten Produktbaustein, sondern gemäß dem Verursachungsprinzip der Leistungsversprechen (z. B. Beratungsgespräch bei Arbeitsrechtsschutz) zugeordnet ist. Auch dies setzt voraus, dass eine entsprechende Struktur im Schadensystem vorhanden ist.

Es sei angemerkt, dass vor dem Hintergrund einer verkaufstechnisch sinnvollen Produktdifferenzierung die praktische Umsetzung einer entsprechenden Granularisierung in der Schadenregulierung erhebliche Schwierigkeiten aufweisen kann. So kann die Aufteilung der Leistungen im Schadenfall (bspw. Einbruchdiebstahl) zwischen den einzelnen Deckungen (Grunddeckung, Deckung von Barmitteln, Wertsachen etc.) in der Praxis zumindest aufwändiger als eine Buchung in Gänze (bspw.

auf die Grunddeckung) sein. Eventuell ist diese aufgrund fehlender Daten zur Differenzierung gar unmöglich. Weiter sei darauf hingewiesen, dass eine unsaubere Umsetzung in der Schadenregulierung der Leistungsinhalte der einzelnen Produktbausteine nicht nur die Nachkalkulierbarkeit zerstört, sondern auch den Gleichbehandlungsgrundsatz der Kunden verletzt. Eine unsaubere Abbildung kann auch eine exakte Schadenregulierung konterkarieren – mit der Folge sowohl von nicht gerechtfertigter Überzahlung als auch Unzufriedenheit der Kunden.

In unserem *Beispiel 1* muss der Schadenregulierer den konkreten Schaden zwischen den gemäß Bausteindefinition relevanten Gewerken unterscheiden und je Gewerk gemäß der vereinbarten Modalität regulieren. Eine unsaubere Regulierung könnte ggf. die Kunden dazu verleiten, nicht alle notwendigen Bausteine zu versichern und stattdessen darauf zu vertrauen, dass durch geeignete Rechnungsstellung im Schadenfall der tatsächlich versicherte Baustein ausreicht.

Wird bei der Risikoerfassung und bei der Schadenbearbeitung keine saubere Trennung zwischen der Differenzierung von Leistungsinhalten und Risikomerkmalen durchgeführt, sollte der Aktuar auf eine sehr genaue und restriktive Bewirtschaftung dringen. Hier sei beispielsweise auf die oftmals als Risikomerkmal abgebildete Leistungserweiterung *Erhöhung Entschädigungsgrenze für Wertsachen* in der Verbundenen Hausratversicherung und das Paradoxon hingewiesen, dass Verträge mit Erhöhung der Entschädigungsgrenze (also erweiterter Deckung) weniger Schäden verursachen.

In Bezug auf die Kürzung der Entschädigung – üblicherweise ein Selbstbehalt auf den bedingungsgemäß ersatzpflichtigen Schaden – ist eine saubere Nachkalkulierbarkeit nur durch den getrennten Ausweis des ersatzpflichtigen Schadens neben der geleisteten Entschädigung gewährleistet. Eine derartige Speicherung ermöglicht nicht nur eine Kontrolle der sachgerechten Regulierung, sondern stellt auch einen klaren Ausweis der Leistung dem Kunden gegenüber dar.

Wird die Originalschadenhöhe nicht im System gespeichert, kommt es vor allem bei gemeldeten Schäden unterhalb der Selbstbehaltshöhe oder bei Stop-Loss-Selbstbehalten zu deutlichen Informationsverlusten, die korrekte Aussagen zur Entlastung verhindern. Diese Informationsverluste auf der Schadenseite können im Retail-Segment mit Kompromissen durchaus wirksam bearbeitet werden, indem bei der Risikomodellierung der Selbstbehalt als Merkmal abgebildet wird. Dies funktioniert aber nur bei einer kleinen Anzahl an Selbstbehaltsvarianten. Im gewerblichen Segment mit vielen individuellen Selbstbehalten ist diese Herangehensweise nicht zielführend.

Wird auf eine saubere Abbildung der Selbstbehalte im Schaden verzichtet, ist jede alternative Abbildung mit Kompromissen verbunden. Der Aktuar ist in seiner Verantwortung gut beraten, an dieser Stelle saubere Abbildungen und Prozesse sowie die Definition der Verantwortung hierfür einzufordern.

Unser *Beispiel 1* zeigt eine weitere Problematik: Während die traditionelle Versicherung je Ereignis über den gesamten Vertrag den vereinbarten Selbstbehalt abzieht,

stellt sich in der neuen Produktwelt die Frage, ob der Selbstbehalt je Baustein oder über die Summe der Leistungen aus den Bausteinen zu berechnen ist. Hier entsteht die Folgefrage, wie zu verfahren ist, wenn Kunden nur einen Teil der möglichen Bausteine (mit Rückwirkungen auf den Wert des Selbstbehaltes) versichert haben und wie die Umsetzung sichergestellt ist, wenn Abzüge in Ansatz gebracht werden, die sich auf die Summe mehrerer Leistungen beziehen? Hier seien die Jahreshöchstentschädigungen genannt.

Für jede Art von Produktbausteinen benötigt der Aktuar alle bekannten risiko- und tarifrelevanten Merkmale. Nur so kann er adäquate Risikomodelle entwickeln und somit zu guten Risikoeinschätzungen für jeden Baustein kommen. Auf dieser Basis können Quersubventionierungen aufgedeckt und in der Folge Aussagen zur Bestandsqualität getroffen werden.

3.3.2. Kalkulationsstatistik

Sind durch geeignete Modularität für jeden Produktbaustein die Leistungen exakt zuordenbar und alle Risiko- und Tarifparameter dispositiv verfügbar, kann der Aktuar grundsätzlich jeden Baustein sachgerecht (nach)kalkulieren, insoweit statistisch signifikante Bemessungsgrundlagen für die Risiken und Schäden vorliegen. Hierbei wird davon ausgegangen, dass das Datenmodell und dessen praktische Implementierung korrekte Aussagen über die Ausprägung aller gespeicherten Risiko-, Tarif- und Leistungsparameter für jeden beliebigen Zeitpunkt oder Zeitraum in der Vergangenheit ermöglichen.

Für die Kalkulation der Bausteine von *Beispiel 1* wird der Aktuar in einer ersten Phase wohl zunächst auf die Verbands-Statistiken zurückgreifen und das Risikomodell der Teilbausteine durch plausible Aufteilung erstellen; mit zunehmendem Datenvolumen wird er validere Kalkulationen des Risikomodells auf eigenen Daten durchführen können. Für volatile Gefahren wird er eventuell über längere Zeit eine Kombination von internen und externen Daten erwägen.

Für die Kalkulation von *Beispiel 2* wird der Aktuar in naheliegender Weise die Risikomodelle der zugrunde liegenden versicherten Sachen und Gefahren aus Verbandskalkulationen und internen Kalkulationen additiv zusammensetzen. Erfolgt die Regulierung und damit die Datenerfassung pauschal – also ohne Speicherung der Informationen über Schadenursache, Schadenart, Schadenort u. ä. –, so ist eine Nachkalkulation dieses Risikomodells nicht möglich.

Die genannten Anforderungen implizieren einerseits eine Grenze für den Umfang von Produktbausteinen und andererseits Zusammenfassungen, für die geeignete Schlüssel und Strukturen in der IT implementiert und in der Statistik abgebildet werden müssen. Es handelt sich zwar um keine rechtlich zwingenden Anforderungen; jedoch erschwert eine Abweichung die dem Aktuar obliegende Beurteilung der Genauigkeit von Kalkulation und Nachkalkulation. Der Aktuar sollte deshalb eine entsprechende Verantwortlichkeit bei der Produktgestaltung einfordern.

Es sei angemerkt, dass statistische Zeitreihen auf einer Vertragsgestaltung beruhen, die einen Vertrag *je Risiko* und einen Schaden *je Ereignis und Risiko* zugrunde legen. Beispiele für Deckungsveränderungen sind Verträge, die das traditionelle Risiko *Gebäude* z. B. auf die Gesamtwerte je Betriebsgrundstück erweitern oder auf (getrennt versicherbare) Gebäudeteile und Teile des üblichen Schadens granularisieren.

3.3.3. Verbandsstatistik und Unternehmensstatistiken

Grundsätzlich kann der Aktuar auf den internen Daten je Produktbaustein kalkulieren. Sollte eine Heranziehung von Verbandskalkulationen notwendig sein, ist in der Kalkulation die Überleitung zwischen Verbandskalkulation und geplanten Deckungen wie auch eine Anpassung auf das eigene Portfolio zu leisten. Im Gegenzug sollten die Daten in einem der Statistik entsprechenden Umfang an die Risikostatistik des GDV gemeldet werden. Für beide Aufgaben ist eine Zusammenfassung, Reduzierung bzw. Granularisierung auf einen der Risikostatistik näherungsweise entsprechenden Leistungsumfang notwendig. Gerade im Kraftfahrtbereich (Kasko) werden viele Leistungen angeboten, die nicht Gegenstand der unverbindlichen Musterbedingungen des Verbandes sind (z. B. Schlüsselersatz bei einfachem Diebstahl), die dann auch nicht in die Risikostatistik gemeldet werden sollten.

Die aus den Versicherungsdaten gewonnene Statistik sollte im Rahmen von üblichen Zeitreihen bzw. Peer-Vergleichen im Rechnungswesen, in der Verbandsstatistik des GDV oder anderen Datenpools abbildbar und aussagefähig sein. Eine Granularisierung oder Zusammenlegung üblicher Deckungen kann hier problematisch sein, falls Deckungen konstruiert werden, die mehrere Versicherungszweige bzw. -arten (Bilanz) sowie Branchenschlüsselzahlen (Verbandsstatistik) überdecken. Gegebenenfalls müssen solche Deckungen unter *Sonstige* erfasst werden, was eine Vergleichbarkeit in der Statistik erschwert.

Granulare Deckungen sowie die zugehörigen Leistungen sollten zur Herstellung der Vergleichbarkeit konsistent zu den sonst üblichen zusammengefasst, neuere Zusatz-Deckungen eventuell ausselektiert werden. Diese Konsistenz im zeitlichen Verlauf ist wichtig für

- die Meldung zur Verbandsstatistik des GDV
- betriebswirtschaftliche Benchmark-Vergleiche (z. B. Beitragseinnahmen, Schadenquote, Schadendurchschnitt des eigenen Unternehmens im Vergleich zum Markt)
- die Überprüfung ermittelter Risikoprofile und der Rückversicherung
- die Aussagekraft von Schadendreiecken, deren Bedeutung in den letzten Jahren gerade im Hinblick auf Solvency II deutlich zugenommen hat.

Beispiel 1 legt nahe, dass eine Erfassung des Schadens je einzelnem der drei Bausteine im Extremfall zu einer Verdreifachung der Schadenanzahl und entsprechenden Reduzierung des Schadendurchschnittes führt. Mindestens für die Meldung zur

Monats- und Risikostatistik des GDV sollten die Leistungen aus den drei materiellen Bausteinen zusammengefasst gemeldet werden.

Das Produkt aus *Beispiel 2* wiederum ermöglicht keine Abbildung in den Versicherungszweigen bzw. Branchenschlüsselzahlen der Sachversicherung.

3.3.4. Verantwortung

Für Entscheidungen dieser Tragweite sollte ein explizit Verantwortlicher definiert werden; der Aktuar sollte im Rahmen seiner Berufspflichten eine derartige Data Governance einfordern.

3.4. Hinweis auf mögliche Probleme in der Umsetzung

Geht es nur um die „akademische" programmiertechnische Umsetzung an sich, treten zunächst keine Probleme auf. Diese entstehen an den Schnittstellen, wenn reale Daten mit dem Tarifrechner verknüpft werden sollen. Wird hier nicht sachgerecht konzipiert, erhält der Aktuar ggf. keine korrekten Daten an die Schnittstelle des Tarifrechners geliefert. Als Konsequenz ist der Aktuar sehr wohl Stakeholder der hier auftretenden Fragestellungen:

3.4.1. Problem der Informationen auf verschiedenen Ebenen

Tarifrechner arbeiten üblicherweise auf Ebene der einzelnen Deckung mit den auf dieser Ebene relevanten Informationen. Grundsätzlich ermöglicht eine geeignete Modularisierung zusammen mit einer angemessenen Datenspeicherung die technische Abbildung und Umsetzung aller Kalkulationsergebnisse. Probleme ergeben sich vor allem dann, wenn wesentliche Informationen nicht der einzelnen Deckung als Datenfeld zur Verfügung stehen, sondern sich sachlogisch aus dem Kontext mehrerer Deckungen und des Vertrages ergeben:

Gewisse erforderliche Daten sind nur aus der übergreifenden Konstellation der abgeschlossenen Deckung ableitbar. Als Beispiele wären ein Bündelnachlass oder die Schadenerfahrung aus der bisher abgeschlossenen Gesamt-Deckung zu nennen. Gut sichtbar ist das Problem auch an einer möglichen Tarifierung *Zweitfahrzeug*: Hier muss in der Datenbank

- nach einem anderen Fahrzeug gesucht werden (des Kunden, seiner Ehefrau, …),
- aufgrund gewisser Informationen zwischen Erst- und Zweitfahrzeug unterschieden werden (Laufleistung, Motorleistung, …) und
- diese Information über die verschiedenen Prozess-Varianten von Ersatz, Stilllegung etc. verwaltet werden.

Auch der Wechsel der Deklaration des Fahrzeugs von Erstwagen zu Zweitwagen und umgekehrt ist hier zu erwähnen.

Diese Informationen sind aus der originalen abzuleiten und folgende Punkte zu beachten:

- Es sind übergreifende Strukturen zu definieren, die diese Ableitungen umsetzen. Die Daten stehen dann systematisch einmal für mehrere Produktbausteine zur Verfügung (1:n).
- Die Daten können in der genannten Struktur genau einmal (normalisiert) gespeichert werden, wobei die einzelnen Produktbausteine auf die gemeinsame Datenbasis zugreifen. Alternativ können die Daten in die Datenstruktur der einzelnen Produktbausteine denormalisiert werden, wobei dies allerdings der Datenbanktheorie zuwiderläuft.
- Bewirken die übergreifenden Daten einen Zwischenschritt der Tarifberechnung, der in den Rechenalgorithmen aller Produktbausteine ausgeklammert werden kann (z. B. gemeinsamer multiplikativer Bündelrabatt), so kann in der übergreifenden Struktur der entsprechende Wert gespeichert und direkt den Produktbausteinen übergeben werden.
- In allen Bausteinen sollten die gleichen Daten auch mit der gleichen Genauigkeit vorliegen.

Jede Änderung einer an der übergreifenden Konstellation beteiligten Deckung hat somit Einfluss auf die Preise aller anderen beteiligten Deckungen.

Eine Veranschaulichung und ein Problem zeigt *Beispiel 3*: Die dort beschriebene Struktur wurde gewählt, da eine sachgerechte Kalkulation nur für einzelne Häuser auf Basis der für jedes einzelne Haus relevanten Daten sinnvoll und möglich ist. Dies setzt die Erfassung (für die Durchführung der Tarifberechnung) und Speicherung (für spätere Nachkalkulationen) der Daten jedes einzelnen Hauses voraus. Die Schäden müssen auch für jedes einzelne Haus erfasst werden. Dies steht ggf. in Konflikt mit einer einmaligen Erfassung: Während die nur einmalige und übergreifende Erfassung der Korrespondenzanschrift und Kontonummer des Kunden sinnvoll ist, führt die einmalige Erfassung der Betriebsart dann zu Problemen, wenn die Betriebsarten der einzelnen Gebäude unterschiedlich sind (viele Kunden der Betriebsart *Präzisionsmetallbearbeitung* versichern u. a. auch Gebäude der Betriebsarten *Lager* oder *Bürogebäude*).

Dies ist entweder in allen beteiligten Prozessen abzubilden oder es ist eine Zeitenlogik zu definieren und in geeigneten Strukturen abzubilden, die die übergreifenden Informationen zu jedem beliebigen Zeitpunkt neben der tatsächlichen Entwicklung speichert.

Beispiel: Ein von den Beiträgen der erfassten Produkte abhängiger Bündelnachlass ändert sich grundsätzlich bei jeder Änderung eines der darunter liegenden Produkte. Wird im Gegensatz der Bündelnachlass einmalig (z. B. bei einer gemeinsamen jährlichen Hauptfälligkeit) berechnet und für das folgende Versicherungsjahr festgehalten, folgt der Bündelnachlass einer eigenen, von den darunter liegenden Produkten gegebenenfalls abweichenden Zeitenlogik.

Eine weitere Konsequenz ist, dass jede Tarifberechnung für einen einzelnen Produktbaustein übergreifende Daten aus der Gesamtheit der relevanten Gruppe von Produktbausteinen benötigt. Eine individuelle Behandlung einzelner Produktbausteine ohne Zugriff auf zentrale Informationen ist deshalb nicht mehr möglich. In der Konsequenz erfordert dies einen Echtzeit-Zugriff auf Bestands- und Schadendaten bei jedem Tarifberechnungsvorgang.

Konsequenzen ergeben sich auch auf die Vertragsgestaltung: Produktbausteine, deren Ausprägungen die Preise anderer Produktbausteine beeinflussen, sollten dem gleichen Vertrag angehören, oder mindestens über eine Vertragsstruktur verbunden sein, die diese Beeinflussung rechtlich und betreffend der Datenstrukturen ermöglicht.

3.4.2. Problem der abgeleiteten Merkmale

Gewisse Daten können nicht direkt in der Tarifberechnung verwendet werden, sondern müssen vorher aufbereitet werden, z. B. Klassifizierungen oder Zonierungen. Diese Ableitungen können sowohl relativ einfache Tabellenabfragen als auch Berechnungen auf großen Datenmengen interner oder externer Dienste mit komplizierten Algorithmen umfassen. Es liegt nahe, die Ergebnisse dieser Ableitungen in der Bestandsführung und der Statistik zu speichern. Jedenfalls ist eine Reproduzierbarkeit der notwendigen Abbildungen (auf die Klassen, Zonen etc.) zu sichern. Der konkrete Umgang mit den Daten muss unter dem Gesichtspunkt der Vertragsgestaltung, der statistischen Verwendung und der Praktikabilität festgelegt werden.

Ein Beispiel aus der aktuellen Versicherungswelt ist die Ableitung der Überschwemmungszone durch geo-informatische Verschneidung der konkreten Koordinaten mit umfangreichen aufbereiteten Landkarten-Daten in Echtzeit in dem Dienst *ZÜRS Geo*; in naher Zukunft könnte die Ableitung eines Fahrer-Scores aus Fahrzeug-, Landkarten- und Wetterdaten auf dem Server eines Dienstleisters möglich sein.

3.4.3. Problem der Kommunikation von (Teilen der) Preisermittlung sowie Beitragsveränderungen

Es besteht die Frage, ob es sinnvoll oder notwendig ist, die konkrete Prämienberechnungsformel oder Teile davon dem Kunden zu kommunizieren. Prominentes Beispiel ist der Schadenfreiheitsrabatt, der in seiner Höhe als multiplikativer Faktor – unabhängig von der sonstigen Tarifberechnung – oft dem Kunden mitgeteilt wird.

Sollte dies aus rechtlichen oder sonstigen Gründen unverzichtbar sein, sind die (relevanten Teile) der Algorithmen dem Kunden gegenüber zu beschreiben. Dies umfasst auch die Eingangsgrößen des jeweiligen Algorithmus – gerade auch dann, wenn Änderungen durch den Kunden zu melden sind und daraufhin eine Anpassung des Beitrags erfolgt.

Wenn dem Kunden Informationen kommuniziert werden, so müssen die Beitragsermittlung und die Datenhaltung hierzu (auch historisch) konsistent sein. Dies kann

bei sonst unproblematischen Änderungen in den Algorithmen (bei Kappungen, Wechselwirkungen, Rundungen etc.) einen signifikanten (Konzeptions-)Aufwand bei der Berechnung, Datenhaltung und Prozessen verursachen. Dies ist a priori – auch bei der Gestaltung der Formulierung dem Kunden gegenüber – zu berücksichtigen.

Darüber hinaus führt die Anwendung mancher Verfahren zu einem Bruch der vom Kunden erwartet Organik im Tarif. Verringert z. B. ein Versicherungsnehmer die vertragliche Kilometer-Laufleistung, würde ihn eine daraus resultierende Erhöhung der Prämie überraschen und die Wahrscheinlichkeit eines Stornos erhöhen.

3.4.4. Probleme bei den Statistiken

Die Statistik ist in jedem Fall auf Ebene der Produktbausteine zu erstellen. Hierzu ist eine ggf. normalisierte Datenhaltung im Statistik-System (mindestens logisch und temporär für diesen Zweck) zu denormalisieren. Im Gegenzug bedingt eine sachgerechte Statistik oft auch, dass Daten nach unterschiedlichen Aspekten zusammengefasst werden müssen. Dazu sind mindestens die relevanten Schlüssel nötig, um die Zusammenfassung logisch durchführen zu können. Zur Vereinfachung kann im Statistik-System eine Zusammenfassung implementiert werden. Das Statistik-System enthält hierbei die Schäden mit allen relevanten Detailinformationen zum Schadenaufwand in der Meldejahressicht (zwecks Vergleichbarkeit mit den Verbandsstatistiken) sowie für einzelne Abwicklungsjahre (bei mehrjährigen Statistiken).

Ein Beispiel für die Problematik ist die gewerbliche Sachversicherung: Die Deckung könnte im Einklang mit der unverbindlichen Empfehlung des GDV je Gebäude und Gefahr modelliert werden. Die Kalkulationsstatistik (inklusive der Kalkulation des GDV) wird auf dieser Ebene arbeiten. Im Unterschied dazu wird eine auf Kumule ausgerichtete Statistik (z. B. für die Rückversicherung) auf die Risikoorte abstellen (also auf die Zusammenfassung aller Gebäude je Risikoort), und eine vertrieblich ausgerichtete Statistik auf den ganzen Vertrag (der gelegentlich auch mehrere Risikoorte umfassen kann). Derartige Statistiken sind natürlich rein logisch neben einander problemlos möglich. Die praktische Konzeption erfordert jedoch oft einen nennenswerten Aufwand.

3.4.5. Problem sonstiger Zusammenfassungen

Bei der Bearbeitung und Kommunikation sind unterschiedliche Arten der Zusammenfassungen von Produktbausteinen sinnvoll (normalisierte Ein- bzw. Ausgabe von Gebäudedaten für die Vielzahl der Produktbausteine über die Gefahren; summierte Ausgabe der Prämien je Gefahr für eine Mehrzahl versicherter Objekte etc.).

Wiederum greifen wir die gewerbliche Sachversicherung auf: Für die Gebäudeversicherung kann es sinnvoll sein, die Adresse eines einzigen Betriebsgrundstückes nur einmal einzugeben und auf alle Deckungen der auf dem Betriebsgrundstück befindlichen Gebäude auszudehnen. Die Inhaltsversicherung und die BU können in

der Praxis dagegen oft nicht einzelnen Gebäuden oder Betriebsgrundstücken zugeordnet werden. Als Konsequenz muss die Adresse auf die Deckungen für Gebäude, Inhalt bzw. BU eventuell. in differenzierter Granularität angewendet werden. Für verschiedene Gefahren wird möglicherweise eine unterschiedliche Zusammenfassung verwendet. Die Kommunikation dem Kunden gegenüber geschieht oft auf Ebene der Gefahr, übergreifend über Gebäude, Inhalt und BU. Für jede der Zusammenfassungen ist eine eigene Datenbankstruktur (Schlüssel) vorzusehen.

Naheliegende Anforderungen an Tarifstrukturen wirken sich mitunter erheblich auf oberhalb der Produktbausteine liegenden Strukturen aus. Unter anderem können die Bedingungsgestaltung sowie die Prozesse inklusive. der globalen Infrastruktur des Datentransportes betroffen sein. Zudem sind Zusammenhänge mit strukturell abzubildenden Zusammenfassungen sonstiger Zwecke zu berücksichtigen.

4. Methoden

Das Kapitel Methoden beschreibt Verfahren, die bei der Anwendung von Maschinellem Lernen in der Schadenversicherung genutzt werden können. Dabei werden nur Verfahren dargestellt, die nicht in [22] beschrieben sind. Zusammenfassend ist vorauszuschicken, dass ein Großteil der Verfahren in ihrem Kern auf den Ideen der bekannten Verfahren wie insbesondere der Regression beruhen, auch wenn die numerischen Verfahren zur Parameterfindung oft andere, auf die großen Datenmengen und Dimensionszahlen abgestimmte, sind.

Der entscheidende Unterschied zwischen klassischer Statistik und der Herangehensweise des Machine Learnings scheint unseres Erachtens insbesondere in der Konkretheit der Modellverwendung und der Modelldiagnose zu liegen. Klassische Statistik unterstellt ein Modell (Verteilungsannahme und funktionale Zusammenhänge zwischen erklärenden Variablen und Zielvariable(n)). Die Parameter der funktionalen Zusammenhänge werden durch Maximierung der Likelihood auf Basis des gesamten Datenmaterials gesucht. Die Güte der Anpassung wird durch die Likelihood bestimmt. Eine klare Trennung zwischen Überdispersion und Modell-Missfit gelingt nicht. Machine Learning sucht dagegen funktionale Zusammenhänge und Parameter (mindestens zum Teil) durch Probieren, wobei der Fit nur auf einem Teil der Daten (Trainingsdaten) durchgeführt wird. Die letztendliche Auswahl von funktionalen Zusammenhängen und Parametern geschieht gemäß der Erklärungskraft auf dem disjunkten Rest der Daten (Testdaten). Damit gewinnt man eine starke Unabhängigkeit von einem von Anfang an unterstellten Modell und reduziert so die Gefahr eines Missfits. Im Gegenzug schwindet das „Verständnis" der Daten anhand des Modells, das häufig selbst Resultat einer fachlichen Datenkenntnis ist.

4.1. Generalized Additive Models (GAM)

Bei einem *Generalized Additive Model (GAM)* wird ein lineares Modell erweitert um Funktionen $f_i, i = 1, \dots, n$, welche auf die Regressoren vor der Regression angewandt werden:

$$g\big(E(Y)\big) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_n(x_n)$$

Die Funktionen müssen nicht weiter spezifiziert werden. Daher stellen die *GAM* die größtmögliche theoretische Abstraktion Linearer Modelle dar. Ist die Zufallsgröße Y weiter aus der Klasse der Exponentialfamilie, so ist das *GAM* bis auf die Funktionen $f_i, i = 1, \dots, n$ identisch mit einem *Generalized Linear Model*.

Im Folgenden seien die Funktionen $f_i, i = 1, \dots, n$ zweimal stetig differenzierbar. In der Regel werden die *GAM* in der Statistik genau unter dieser Restriktion betrachtet.

Je nach gewählter Linkfunktion g geben die Funktionen $f_i,\ i = 1, \dots, n$ dem Modell größere Freiheitsgrade oder schränken sie ein:

- Ist die Linkfunktion z. B. die Identität, dann handelt es sich um die klassische lineare Regression.

- Ist die Linkfunktion z. B. die Log-Funktion, dann handelt es sich um eine kategoriale Regression, es werden die Score-Werte geglättet; beim Merkmal *Fahrleistung* kann Organik der Score-Werte sichergestellt werden.

Auch GLM können um Funktionen $f_i, i = 1, \dots, n$ erweitert werden. Sie müssen aber vor Anwendung manuell z. B. als Polynome und Spline-Funktionen vorgegeben werden. Der entscheidende Unterschied bei den *GAM* ist, dass sich das *GAM*-Modell selber geeignete Funktionen $f_i, i = 1, \dots, n$ sucht.

Hierzu wird eine *penalisierte* Log-Likelihood-Funktion eingeführt:

Seien $2\, l\big(\beta_0, f_1, f_1(x_1), f_2(x_2), \dots., f_n(x_n)\big)$ die Standard-Log-Likelihood-Funktion und *penalty* ein allgemeiner Strafterm. Die Schätzer des *GAM* ist dann, die um den *Penalty*-Term erweiterte Maximum-Likelihood-Funktion:

$$2\, l\big(\beta_0, f_1(x_1), f_2(x_2), \dots., f_n(x_n)\big) - \ penalty$$

Als *Penalty*-Term wird in der Regel die zweite Ableitung der Funktionen f_i, $i = 1, \dots, n$ benutzt:

$$penalty: = \ \sum_{j=1}^{n} \lambda_j \int \left(f''_j(x_j)\right)^2 \ dx \quad \text{mit } \ \lambda_i \geq 0.$$

Da die zweite Ableitung größer wird, je stärker gekrümmt eine Funktion ist, bestraft der *Penalty*-Term also zum Beispiel Polynome mit höherer Ordnung. Durch den *Penalty*-Term müssen also zum Beispiel Polynome höherer Ordnung ihre Fähigkeit sich an die Daten anzupassen (und dieselben dadurch möglicherweise zu überfitten), dadurch ausgleichen, dass sie die Maximum-Likelihood-Funktion entsprechend stärker maximieren. Der *Penalty*-Term würde eine lineare Funktion vorziehen, da dann die zweite Ableitung null ist. Durch den *Penalty*-Term wird also das Gleichgewicht zwischen bestmöglicher Erklärung der Variablen und Komplexität des Modells gewahrt.

Mit den Parametern λ_j kann der Anwender des *GAM* die Bestrafung des *Penalty*-Terms abschwächen oder erhöhen. Je höher sie gewählt werden, desto mehr würde das *GAM* lineare Funktionen vorziehen. Mit den λ_j kann der Anwender also vorgeben, wie stark die einzelnen Variablen geglättet werden sollen.

Ohlsson und Johansson zeigten, dass man sich bei der Suche nach zweimal stetig differenzierbaren Funktionen auf kubische Splines beschränken kann.

Die λ_j können auch aus dem Datenmaterial z. B. mit *Generalized Cross Validation Criteria* geschätzt werden. Eine Diskussion dieser Verfahren würde aber den Rahmen dieser kurzen Einführung sprengen.

Nachteilig bei einem *GAM* wirkt sich in der Praxis aus, dass die Modellergebnisse kaum noch intuitiv interpretierbar sind: die Verbesserung der Schätzung durch

GAMs gegenüber GLM*s* geschieht auf Kosten der Interpretationsfähigkeit der Koeffizienten.

Literatur:

[9] Hastie, Tibshirani (1986): Generalized Additive Models

[22] Ohlsson, Johansson (2010): Non-Life Insurance Pricing with Generalized Linear Models, Kapitel 5

4.2. Shrinkage Methods

Auf Grundlage generalisierter linearer Modelle reduzieren *Shrinkage*-Methoden den Einfluss einzelner Kovariablen auf das final gewählte Modell. Dadurch soll ein Overfitting vermieden werden. Im Gegensatz zu anderen Verfahren der Variablen-Auswahl (z. B.: *Forward-*, *Backward-* oder *Stepwise*-Selektion) erfolgt die Selektion dabei kontinuierlich. Unter bestimmten Bedingungen wird allerdings auch bei den *Shrinkage*-Verfahren ein vollständiger Ausschluss einzelner Variablen aus dem Modell erreicht. Das ist z. B. bei *Lasso* und *Elastic Net* der Fall.

Gesucht wird bei allen *Shrinkage*-Verfahren:

$$\hat{\beta}(\lambda) = \arg\min_{\beta}[-\log L(\boldsymbol{y};\beta) + \lambda\, P(\beta)]$$

Dabei ist $L(\boldsymbol{y};\beta)$ die *Likelihood*. Der Strafterm $P(\beta)$ ist je nach Verfahrenstyp unterschiedlich. Optimiert wird also eine penalisierte Likelihood. Je größer der Parameter λ gewählt wird, desto geringer ist der Beitrag der Kovariablen.

$\boldsymbol{P(\beta)}$	**Parametrisierung**	**Verfahrensbezeichnung**	**Variablenausschluss**
$\sum_{\substack{alle \\ Kovariablen\ i}} \lvert\beta_i\rvert^q$	$q = 1$	Lasso	Ja
	$q = 2$	Ridge	Nein
	$0 < q \leq 1$	-	Ja
	$q > 1$	-	Nein
$\sum_{\substack{alle \\ Kovariablen\ i}} \left(\alpha\, {\beta_i}^2 + (1-\alpha)\lvert\beta_i\rvert\right)$	$0 \leq \alpha \leq 1$	Elastic Net	Ja, für $\alpha \neq 1$

Bei *Elastic Net* werden die Strafterme von *Ridge* und *Lasso* also linear gemischt.

Für die sinnvolle Anwendung müssen die Variablen vorab standardisiert bzw. als Dummy-Variablen kodiert werden, da ansonsten der Einfluss der jeweiligen β auf den Strafterm $P(\beta)$ verzerrt wäre.

Mittels *Shrinkage*-Verfahren sind sowohl Klassifikation als auch Regression möglich.

Literatur:

[10] Hastie, Tibshirani, Friedman (2009): The Elements of Statistical Learning, Abschnitt 3.4

[12] James et al. (2014): An Introduction to Statistical Learning, Abschnitt 6.2

4.3. Baumverfahren

Die Praxis unterscheidet verschiedene Algorithmen zur sukzessiven Unterteilung einer Datenmenge *M* von möglichen Werten eines Vektors $Y = (Y^{(1)}, \dots, Y^{(m)})$ in disjunkte Teilmengen $(M_1, \dots, M_n)$ von *M*, welche durch ein Baumdiagramm veranschaulicht werden. Im Folgenden wird das am häufigsten verwendete Baumverfahren *CART (Classification and Regression Trees)* beschrieben und anschließend ein Vergleich mit *C5.0* vorgenommen. Zusätzlich werden *Random Forests* aufgegriffen.

4.3.1. Classification and Regression Trees (CART)

CART ist eine nicht-parametrische Entscheidungsbaummethode, die entweder Klassifizierung oder Regression vornimmt. Entscheidungsbäume basieren stets auf einer Menge von Regeln, anhand derer einzelne Datensätze in Klassen eingeteilt werden, wobei gängige Kriterien Minimierung der Varianz innerhalb der erzeugten Klassen oder maximale Diversität zwischen den Klassen umfassen.

Im ersten Schritt wird diejenige Unterteilung der Lernmenge *M* bezüglich eines Merkmals in zwei disjunkte Teilmengen (Knoten) M_1 und M_2 bestimmt, bei welcher die Fehlerquadratsumme

$$RSS = \sum_{M_1} (y - \overline{y_{M_1}})^2 + \sum_{M_2} (y - \overline{y_{M_2}})^2$$

der beiden Knoten M_1 und M_2 am kleinsten ist. Anschließend werden beide Knoten M_1 und M_2 unabhängig voneinander analog unterteilt. Das Verfahren ist folglich rekursiv und wird so lange wiederholt, bis entweder eine zuvor bestimmte Maximaltiefe erreicht ist oder nach jedem einzelnen vorhandenen Merkmal, welches einen Zuwachs an Genauigkeit bieten kann, gesplittet wurde.

Man erhält so einen Baum, der in der Regel sehr groß ist. Die Vorhersage des Baumes wird durch die Mittelwerte der Endknoten bestimmt. Anschließend wird der so gebildete Baum auf einen Teilbaum zurückgeschnitten. Hierbei wird zunächst für jeden Parameter *a* derjenige Teilbaum T_a bestimmt, der die Mischung

$$R_\alpha(T) = R(T) + \alpha \cdot |T|$$

aus dem quadratischen Fehler $R(T)$ in den Endknoten und der Anzahl $|T|$ der Endknoten minimiert und anschließend derjenige Teilbaum T_a ausgewählt, der die beste Vorhersage auf einer zur Trainingsstichprobe unabhängigen Teststichprobe liefert.

Ein großer Vorteil von Baumverfahren ist die schnelle Überprüfbarkeit der Klassifikation einzelner Datensätze und die Verteilungsfreiheit. Darüber hinaus ist im Gegensatz zu nicht inhärent-deskriptiven Verfahren wie neuronalen Netzen die semantische Nachvollziehbarkeit sehr hoch. Die Erklärbarkeit der Klassen folgt streng den zugrundeliegenden Split-Kriterien, die direkt in der Eingabemenge ablesbar und ggf. darstellbar sind: Wo also die Reproduzierbarkeit der maschinell getroffenen Entscheidungen (bspw. um Tarifierungsfragen regulierungskonform quantifizieren zu können) wichtig ist, liegt hier ein zentraler Vorteil von Baum- und Regressionsverfahren. Die Vorteile der Baumverfahren gegenüber klassischen Regressionsverfahren wie GLM*s* sind die automatische Merkmalsauswahl, die Erkennung von Interaktionen und die einfache Interpretierbarkeit bei kleineren Bäumen. Ein Nachteil ist die Instabilität; bei geringen Änderungen in den Daten entsteht ein komplett anderer Baum. Bezüglich der Vorhersagekraft sind *Random Forests* zu überlegen.

R-Package:

`rpart`

Literatur:

[13] Breiman et al. (1984): Classification and Regression Trees

[10] Hastie, Tibshirani, Friedman (2009): The Elements of Statistical Learning, Abschnitt 9.2

[12] James et al. (2014): An Introduction to Statistical Learning, Kapitel 8

[15] Wu et al. (2008): Top 10 algorithms in data mining

4.3.2. C5.0

Der Algorithmus *C5.0* stellt einen populären Sonderfall der Entscheidungsbäume dar, dessen Kriterien explizit auf dem Maß der Informationsentropie basieren, also zur Entscheidung die Frage evaluieren, welcher Split den größten Informationsgewinn zur Folge hat. Die Trainingsstichprobe muss dazu aus bereits klassifizierten Samples bestehen, also Labels der Zielfunktion $\{s_1, \dots, s_n\} \subset \{1, \dots, L\}$ enthalten. Dann ist $H(S) \coloneqq -\sum_{l=1}^{L} p(l)\ \log_2 p(l)$ die Entropie von *S*, wobei $p(l)$ die relative Häufigkeit von l in $\{x_1, \dots, x_n\}$ ist. Wird *S* mit Hilfe eines Merkmals *A* in *K* Teilmengen $S_1, \dots, S_K$ zerlegt, so ist $IG(S, A) \coloneqq H(S) - \sum_{k=1}^{K} q(k) H(S_k)$ der Informationsgewinn von *A*. Dabei ist $q(k) = \frac{|S_k|}{n}$ der Anteil der Beobachtungen von *S* in S_k. Für den Split wird das Merkmal gewählt, das den größten Informationsgewinn ergibt.

Der Vorteil der Methode liegt im größtmöglichen Differenzierungspotential, wobei allerdings anzumerken ist, dass auch die Anfälligkeit für *Overfitting* insbesondere bei unbalancierten Zielfunktionsverteilungen überdurchschnittlich ausgeprägt ist. Die Vorgehensweise von *C5.0* ist im Sinne der Differenzierung optimal, hält allerdings die Herausforderung bereit, dass benachbarte Klassen keine Homogenität im

Sinne der informationstheoretischen Topologie des Parameterraumes haben. In solchen Situationen ist es oftmals ratsam, konvexe Verfahren wie *Support Vector Machines* zu verwenden, die verallgemeinerten Kontinuitätskriterien gerecht werden können.

R-Package:

`C50`

Literatur:

[14] Quinlan (1993): C4.5: Programs for Machine Learning

[15] Wu et al. (2008): Top 10 algorithms in data mining

4.3.3. Random Forests

Um die Vorhersagekraft von Baumverfahren zu erhöhen, wurden Verfahren entwickelt, welche nicht nur einen, sondern viele Bäume aus den Daten erstellen. Die Vorhersage ist dann bei Regressionsmodellen der Mittelwert der Vorhersagen, bei Klassifikationsaufgaben folgt man hingegen der Mehrheitsentscheidung der erzeugten Bäume.

Das bekannteste dieser Verfahren ist *Random Forests* (Breiman 2001). Hierbei wird aus den Modelldaten viele Male ein neuer Datensatz per Bootstrap erzeugt und dort das *CART*-Verfahren ohne Rückschnitt durchgeführt, wobei in jedem Knoten nur deutlich kleinere und zufällige gewählte Teilmengen der Merkmale für den Schnitt verwendet werden.

Durch dieses Verfahren werden oft die Varianz des Schätzers und die Überanpassung an die Daten verringert. Starke Verbesserungen ergeben sich insbesondere in unstabilen Situationen, in denen eine kleine Änderung der Lernmenge zu großen Änderungen der Schätzer führt.

Die Vorhersagekraft der *Random Forests* ist deutlich höher als diejenige des einfachen *CART*-Verfahrens. Dieses Verfahren wird in vielen verschiedenen Situationen eingesetzt und schneidet in Vergleichsrechnungen bezüglich der Vorhersagekraft meist sehr gut ab. Dafür geht die leichte Interpretierbarkeit bereits bei kleineren Bäumen verloren.

R-Package:

`randomForest`

Literatur:

[10] Hastie, Tibshirani, Friedman (2009): The Elements of Statistical Learning, Kapitel 15

[12] James et al. (2014): An Introduction to Statistical Learning, Abschnitt 8.2.2

4.4. Ensemble Learning

Ensemble-Methoden nutzen eine endliche Menge von verschiedenen Lernalgorithmen, um bessere Ergebnisse zu erhalten als mit einem einzelnen Lernalgorithmus. Aus vielen einfachen Modellen ein kombiniertes zu erzeugen, führt in vielen Fällen zu einer höheren Prädiktionsgüte. Dabei können auch unterschiedliche Methodenarten zur Anwendung kommen. Dies passiert allerdings auf Kosten der Interpretierbarkeit, z. B. sind die Zusammenhänge bei einem kombinierten Modell aus 100 verschiedenen Bäumen schwerer als bei einem einzelnen Baum zu verstehen. Ein weiterer Nachteil sind die erhöhten Rechenzeiten von *Ensemble Learners*.

Es gibt viele verschiedene Arten von *Ensemble Learnern*, u. a. *Bagging* und *Boosting*.

4.4.1. Bagging

Die Idee von *Bagging* ist es, dieselbe Methodik auf unterschiedliche Daten anzuwenden. Das Wort *Bagging* setzt sich zusammen aus *Bootstrap* und *Aggregation* und genau so funktioniert das Prinzip: Aus den vorhandenen Daten werden zunächst durch Ziehen mit Zurücklegen n Samples generiert *(Bootstrap).* Auf allen Samples wird die gewählte Methode angewandt und die n resultierenden Modelle kombiniert *(Aggregation)*. Dabei kann die *Aggregation* für Regressionen durch ein gewichtetes Mittel der Modelle ermittelt werden.

Random Forest ist z. B. eine *Ensemble-Learning*-Methode für Klassifizierungs- und Regressionsprobleme, die *Bagging* verwendet: Es werden mehrere individuelle Entscheidungsbäume durch *Bagging* erstellt und anschließend kombiniert.

4.4.2. Boosting

Boosting ist eine Verfahrensklasse, bei der ein Modell iterativ auf dem vorherigen Modell aufbaut, beispielsweise auf dem noch nicht erklärten Teil *(Residuen)*. Es gibt viele verschiedene Varianten von *Boosting*-Verfahren.

Beispielsweise wenden *Gradient-Tree-Boosting*-Modelle das *Boosting*-Prinzip an. Im *n*-ten Iterationsschritt wird $F_n(x)$ auf den Residuen $y - \sum_{i=1}^{n-1} F_i(x)$ trainiert. Das Modell $F_n(x)$ wird mit der vorab gewählten „Lernrate" *(Shrinkage-Parameter)* μ zum aktuellen Modell addiert: es ergibt sich somit das finale Modell durch $F(x) = \mu \sum_{n=1}^{N} F_n(x)$.

4.4.3. Gradient Boosting Machine (GBM)

Gradient Boosting Machine (Friedmann), auch *Generelized Boosting model* (in R), auch *Multiple Additive Regression Trees (MART)*

Idee:
Es wird schrittweise aus linearen additiven Komponenten eine Funktion *F(x): X → R* aufgebaut, mit der die Responsevariable y gelernt werden soll. Das *GBM* ist kein eigenständiger statistischer Ansatz, sondern als *Ensemble Learning* eine Kombination von mehreren, weitgehend schon beschriebenen Methoden des maschinellen Lernens:

1. *Boosting:* Anstatt direkt ein Modell zu schätzen wird eine Serie schwacher Modelle aufgebaut, die gegen ein Zielmodell konvergieren.

2. *Regression Trees:* Jedes der Teilmodelle ist ein Baum, dessen Tiefe bewusst suboptimal begrenzt wird.
   Die gemeinsame Tiefe der Bäume steuert die Komplexität des Modelles. Beschränkung auf die Tiefe 1 (engl. *Stumps*) führt auf ein Modell vergleichbar einem GLM ohne Interaktionen. Gleichzeitig wird über einen *Penalisierungs*-Parameter die Komplexität des Baumes, also die Anzahl der Knoten oder Blätter bewertet.

3. *Addition von Modellen:* Die Funktion $F(m+1)$ der Stufe $m+1$ wird i. W. berechnet als $F(m) + T(m+1)$,
   letzterer ist der *Regression Tree* der $(m+1)$-ten Stufe.

4. *Residuenanalyse* Jedes neue Teilmodell baut auf den Residuen der (kumulierten) vorhergehenden Modelle auf.

5. *Shrinkage* Ein Parameter beschränkt gleichmäßig die Wirkung jedes Modellschrittes. Es gilt in Präzisierung von *3.*:
   $F(m+1) = F(m) + \mu\, T(m+1)$ mit $0 < \mu \leq 1$.

6. *Kreuz-Validierung:* Nach jeder Iteration wird das auf der Trainingsmenge gewonnene Modell auf der/den Testmenge(n) validiert.

7. *Bagging* Die Trainingsmenge wird in jeder Iteration als definierter Anteil an der Gesamttrainingsmenge neu ausgelost

R-Package:

gbm

Literatur:

[10] Hastie, Tibshirani, Friedman (2009): The Elements of Statistical Learning, Abschnitt 10.2

[11] R-package gbm: Generalized Boosted Regression Models, https://cran.r-project.org/web/packages/gbm/gbm.pdf

4.5. Bayes'sche Netze (BN)

BN dienen zur Klassifikation und Entscheidungsfindung unter Unsicherheit. Sie sind probabilistische Netze, die graphisch die Faktorisierung der gemeinsamen Wahrscheinlichkeitsverteilung gegeben bedingter Unabhängigkeiten darstellen.

Als Grundlage zur Definition von *BN* wird eine Kombination aus Graphentheorie und Wahrscheinlichkeitstheorie benötigt. Ausgangspunkt ist ein gerichteter azyklischer Graph – *Directed Acyclic Graph (DAG)*, wobei die Knoten die Zufallsvariablen (ZV) und die Kanten die bedingten Abhängigkeiten repräsentieren. Zum Modell gehören ebenso die bedingten Wahrscheinlichkeitsverteilungen pro ZV. Ein *BN* ist als *DAG* definiert, der die Markov-Bedingung erfüllt: ein Knoten X ist bedingt unabhängig von den Nicht-Nachfahren gegeben die Elternknoten.

Gemäß der Graphenstruktur treten spezielle Verbindungen (d-Separationen: *serielle*, *divergente* oder *konvergente*) in *BN* auf. Die Markov-Blanket eines Knotens liefert ein Mengenkriterium für die bedingte Unabhängigkeit von allen anderen Knoten. Bei der Klassifikation lässt sich so die Anzahl der Merkmale reduzieren. Zur Modellierung eines *BN* gehören die Schritte:

- Bestimmung der Variablen = Knoten.
- Identifizieren der Zusammenhänge = Kanten
- Gemeinsame Wahrscheinlichkeitsverteilung der Knoten schätzen $P(X_1,..,X_n) = \prod_{i=1}^{n} P(X_i = x_i \mid \text{Menge der Elternknoten})$

Die Konstruktion des *BN* aus der Information der Daten wird als *Lernen Bayes'scher Netze* bezeichnet. Man unterscheidet zwischen *Parameterlernen* (Schätzen der Verteilungsparameter mittels ML, MAP oder bei missings EM-Algorithmus) und *Strukturlernen*. Für Letzteres existieren score-basierte Methoden (z. B. *Hill-Climbing*) oder constraint-basierte Methoden (z. B. mittels bedingter Unabhängigkeitstests). Als Score-Funktionen kommen *AIC*, *BIC*, *Score über Log-Likelihood-Funktion* oder *A-Posteriori-Verteilung* in Frage. Als Verteilungsannahmen eignen sich für kategoriale Variablen eine Multinomial-Verteilung für die Daten mit Dirichlet-Verteilung als A-Priori-Verteilung des Parameters, somit ist die A-Posteriori-Verteilung ebenso Dirichlet.

Verschiedene *BN*-Klassifikatoren finden Verwendung: Der naive Bayes-Klassifikator basiert auf dem Maximum der A-Posteriori-Verteilung. Allerdings ist die Annahme der Unabhängigkeit der Merkmale gegeben der Klasse notwendig *(sternförmiges BN)*. Abhilfe schafft der Baum-erweiterte naive Bayes-Klassifikator – *Tree-Augmented Naive Bayes (TAN)*.

R-Package:

`Blearn`

Literatur:

[16] Nagarajan, Scutari, Lèbre (2013): Bayesian Networks in R with Applications in Systems Biology, Kapitel 2

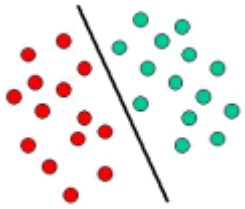
[17] Scutari, Denis (2014): Bayesian Networks with Examples in R, Kapitel 1 und 4

4.6. Support Vector Machines (SVM)

Ziel ist es, eine Klassifikationsaufgabe in K=2 Klassen zu bewältigen. Die Trainingsdaten $(x_1, y_1), \dots, (x_n, y_n)$ bestehen aus metrischen Kovariablen $x_i \in R^p$ und einer kategorialen Zielvariablen $y_i \in \{-1,1\}$. Aus den Trainingsdaten wird eine Funktion $f: R^p \longrightarrow R$ angepasst mit der Testdaten $x \in \mathbb{R}^p$ nach dem Vorzeichen von *f(x)* klassifiziert werden, also

$$Klasse\ von\ x = \begin{cases} 1, & falls\ f(x) > 0 \\ -1, & falls\ f(x) < 0 \end{cases}$$

4.6.1. Maximum Margin Classifiers



Gibt es eine Hyperebene, die Trainingsdaten perfekt trennt, dann ist *f* affin linear, also gibt es $b_0 \in R$ und $b \in R^p$ mit $f(x) = b_0 + \langle b, x \rangle$. Die Hyperebene wird so gewählt, dass der Abstand zu den Trainingsdaten maximal ist. Dieser Abstand heißt margin. Um *f* zu bestimmen, wird folgende Maximierungsaufgabe gelöst: Bestimme $M > 0$ unter der Nebenbedingung $\|b\| = 1$, so dass

$$\forall i = 1, \dots, n: \ \ y_i(b_0 + \langle x_i, b \rangle) \geq M \qquad (1)$$

gilt.

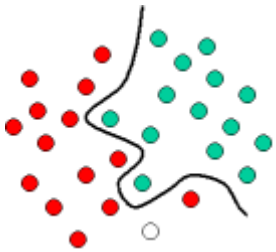
4.6.2. Support Vector Classifiers

Im Allgemeinen gibt es keine Hyperebene, die die Trainingsdaten trennt. Man weicht die Forderung der trennenden Hyperebene dahingehend auf, dass auch Fehlklassifikationen der Trainingsdaten zugelassen werden. Die Funktion *f* ist weiter affin linear. In der Maximierungsaufgabe in *4.6.1* wird dann (1) ersetzt durch

$$\forall i = 1, \dots, n: \;\; y_i(b_0 + \langle x_i, b\rangle) \geq M(1 - \varepsilon_i) \;\; \text{mit } \varepsilon_i \geq 0, \sum_{i=1}^{n} \varepsilon_i \leq C$$

Ist $\varepsilon_{i_0} > 1$ dann wird x_{i_0} falsch klassifiziert. Bei *C* handelt es sich um einen Tuningparameter, über den man die Anzahl der Fehlklassifikationen steuert. Es stellt sich heraus, dass nur die falsch klassifizierten Vektoren die Hyperebene bestimmen, diese heißen Support Vektoren.

4.6.3. Support Vector Machines



Es gibt Situationen, in denen die Forderung nach einer linearen Trennung durch eine Hyperebene der Struktur der Daten zuwider läuft. Dies führt zur Erweiterung der Methoden zu *SVM*. Es werden nun nicht-lineare Funktionen *f* betrachtet. Die Trainingsstichprobe wird nicht perfekt klassifiziert, d. h. die Menge $I \coloneqq \{i \in \{1, \dots, n\} \mid \text{sign}(x_i) \neq y_i\}$ ist nicht leer. Die Funktion *f* besitzt die Form $f(x) = \beta_0 + \sum_{i \in I} \alpha_i K(x, x_i)$ mit $K: R^p \times R^p \longrightarrow R$. Die Summe wird über die falsch klassifizierten Lerndaten gebildet, diese heißen wie in 1.5.2. *Support-Vektoren*. Die Funktion *K* heißt *Kern-Funktion* (englisch: *kernel*). In der Anwendung sind populär

- polynomiale Kerne: $K(x, x') = (1 + \langle x, x'\rangle)^d$ mit $d \in \mathbb{N}$
- radiale Kerne: $K(x, x') = \exp(-\gamma \|x - x'\|_2^2)$ mit $\gamma > 0$
- sigmoide Kerne: $\tanh(\kappa_1 \langle x, x'\rangle + \kappa_2)$

Der Raum R^p wird durch die Fläche $\{x \in \mathbb{R} \mid f(x) = 0\}$ getrennt, die Funktion *f* ergibt sich nun als Lösung einer quadratischen Optimierungsaufgabe. Im Fall von mehr als zwei Klassen *K*>2 werden mehrere Support Vector Machines konstruiert und mit Hilfe paarweiser Vergleiche auf den Fall *K*=2 zurückgeführt.

R-Package:

e1071

Literatur:

[10] Hastie, Tibshirani, Friedman (2009): The Elements of Statistical Learning, Abschnitt 12.3

[12] James et al. (2014): An Introduction to Statistical Learning, Kapitel 9

Abbildungen

dieses Abschnitts entnommen aus http://www.statsoft.com/textbook/support-vector-machines.

4.7. Neuronale Netze

Ein (künstliches) neuronales Netz ist eine z. T. heuristische, auf der Funktionsweise von biologischen Zerebralprozessen basierende Methode, die darauf abzielt, Probleme der Klassifizierung oder des Clusterings auf ein Netzwerk von sogenannten Neuronen und deren Verbindungen abzubilden.

4.7.1. Restricted Boltzmann Machine (RBM)

Eine konkrete Ausprägung neuronaler Netze stellen die sogenannten *Restricted Boltzmann Machines (RBMs)* dar. Deren Grundidee besteht darin, ein stochastisches Element in die neuronale Wirkungsweise einzufügen: Während die (binäre) Aktivierung der Neuronen im klassischen Fall deterministisch ist, fügen *RBMs* dem ein stochastisches Element hinzu. Dies geschieht, indem die Menge der Neuronen in drei Unterkategorien eingeteilt wird, die *normalen Neuronen* einerseits*, versteckte Klassen*, deren Gesamtheit die gesuchte Zielfunktion darstellt und eine *bias unit*, die die stochastische Modellierung übernimmt.

RBMs (und neuronale Netze im Allgemeinen) arbeiten durch das Aktualisieren der internen Zustände der Neuronen in Abhängigkeit von anderen Neuronen. In *RBMs* sind die Zustände der bekannten Neuronen, die die Eingabemenge modellieren, konstant, während die versteckten Einheiten mutabel sind. Ein Arbeitsschritt kann wie folgt abstrahiert werden:

1. Alle Nachbarn des Neurons senden ein Signal, das gewichtet wird. Dies erfolgt durch die Berechnung der Aktivierungsenergie $a_i = \sum_j w_{ij} x_j$, wobei die Summe über alle Verbindungen geht, die die Einheit i aufweist, w_{ij} das Gewicht dazwischen darstellt und $x_j \in \{0,1\}$ der Zustand des Neurons j ist.
2. Es wird $p_i = \sigma(a_i)$ gesetzt, wobei $\sigma(x) = \frac{1}{1+\exp(-x)}$ die logistische Funktion ist. Das Signal wird also zwischen den Binärzuständen interpoliert.

3. Die Wahrscheinlichkeit dafür, dass das Neuron i aktiviert wird, ist damit gerade durch p_i gegeben.

Die iterative Durchführung der Prozedur führt folglich dazu, dass gleichartige Werte gleichartige Reaktionen provozieren und die Wahrscheinlichkeit, dass stark unterschiedliche Eingaben die gleichen Neuronen aktivieren, exponentiell unterdrückt werden. Hierbei ist zu beachten, dass *RBMs* inhärent binäre Zielfunktionen bevorzugen und nicht originär zur direkten Tarifierung verwendbar scheinen, sondern eher um binäre Scores zu modellieren.

Ein wesentlicher Nachteil der Vorgehensweise bei den meisten Ausprägungen von neuronalen Netzen besteht darin, dass die internen Zustände der Neuronen entweder keinerlei Abbildung auf die Ein- oder Ausgangssemantik zulassen oder überhaupt nicht zugänglich sind, wenn es sich um eine gänzlich zustandsfreie Ad-hoc-Modellierung handelt. Der große Vorteil der Flexibilität der Netze wird hier gewissermaßen damit erkauft, dass eine Portierung auf andere Probleme, bei der nur die Eingabeformate oder Zielfunktionen geändert werden, üblicherweise nicht möglich ist.

R-Package:

`darch`

Literatur:

[18] Salakhutdinov, Mnih, Hinton (2007): Restricted Boltzmann machines for collaborative filtering

[19] Hinton (2010): A Practical Guide to Training Restricted Boltzmann Machines

4.7.2. Deep Learning

Deep Learning, auf Deutsch etwa *tiefgehendes Lernen*, bezeichnet eine Klasse von Optimierungsmethoden von künstlichen neuronalen Netzen, welche zahlreiche verborgene Schichten (englisch: *hidden layers*) zwischen Eingabeschicht und Ausgabeschicht haben und dadurch eine umfangreiche innere Struktur aufweisen. Das Adjektiv *deep* meint also eher die Tiefe der Netzstruktur als die Tiefe der gewonnenen Einsichten zu Problemen, die sich mit traditionellen Methoden nicht erschließen lassen. Dies ist insbesondere Mustererkennung in Bildern, Videosequenzen, gesprochener Sprache, Sentimentanalyse von Texten (Beispiele: Google *Deep Mind*, *Deep Face*).

Durch die Schichtung der Zwischenebenen werden die unstrukturierten Eingabemerkmale schrittweise in analysierbare Muster umgewandelt. Für das Training werden allerdings nur die Eingabedaten und die korrekten Label zur Klassifizierung verwendet. Grundsätzlich kann das gesamte Netz als ein Optimierungsproblem trainiert werden, wobei die inneren Ebenen unüberwacht vortrainiert werden können. Das

Training geschieht im Wechsel von Forward und Backward Propagation. Forward: Der Output jeder Schicht Neuronen wird in der nächsten Schicht über eine lineare Verknüpfung und eine sigmoide Funktion (z. B. $\frac{1}{1+e^{-x}}$, logistische Funktion) an die übernächste Schicht weitergereicht. Backward: Optimierung einer Fehlerfunktion, i. A. Minimierung der Fehlerquadrate, durch Gradientenabstieg.

Durch die Komplexität des Netzes besteht die Gefahr von Overfitting in den Zwischenebenen. Dieser kann begegnet werden durch Regularisierung, z. B.

- *Penalisierung* der Komplexität in der Fehlerfunktion der Optimierung,
- *Dropout*, ausblenden eines Teils der Einheiten in einzelnen Trainingsschritten
- *Stochastische Binäreinheiten*, welche mit einer bestimmten Wahrscheinlichkeit den Wert 1 senden, sonst 0
- *Autoencoding*, zur internen Komplexitätsreduktion mit einer verdeckten Mittelschicht aus erheblich weniger Einheiten als Ein- und Ausgabeschichten, u. U. mit Regularisierungsparameter. Die Wirkung eines linearen Autoencoders ist hier vergleichbar mit einer PCA.
- *Faltung (Convolution)* bei Bilderkennung: Schärfung durch Mittelung von Pixeln über eine Kernelmatrix

Literatur:

Der erste Satz ist zitiert nach

[20] Wikipedia: https://en.wikipedia.org/wiki/Deep_learning

4.8. Unsupervised Nearest Neighbor (UNN)

Das Verfahren *Unsupervised Nearest Neighbor (UNN)* ist eine unüberwachte Regression zur Dimensionsreduktion $F: y \rightarrow x$ für Muster $y \epsilon Y \subset \mathbb{R}^d$ und latente Punkte $x \epsilon X \subset \mathbb{R}^q$ mit $q < d$, d. h. der latente Raum X wird durch Umkehr eines Regressionsmodells f ermittelt. Dabei wird versucht, den Rekonstruktions-Fehler $E(Y,X) = \frac{1}{N}||Y - f(X)||_F^2$ mit der Methode der k-*Nearest-Neighbor (KNN)* als Regressionsmodell $f_{UNN}(X) = \frac{1}{K}\sum_{i \in N_K(X)} y_i$ und der Frobeniusnorm $||A||_F^2 = \sqrt{\sum_{i=1}^{d}\sum_{j=1}^{N}|a_{ij}|^2}$ zu minimieren, wobei $N_K(X)$ die Menge der k-nächsten Nachbarn darstellt.

Das *KNN*-Verfahren ist bei Daten mit vielen Dimensionen problematisch, da homogene Metriken des Ausgangsraumes zu wahlweise schlechter Genauigkeit (Homogenität) oder mangelhaftem Clustering führen, wenngleich eine Regularisierung beispielsweise via *PCA* dem entgegenwirken kann. Deshalb ist bei Einsatz von *UNN* oftmals die Verwendung einer vorgeschalteten hierarchischen Klassifizierungsmethode empfehlenswert.

Literatur:

[21] Kramer (2011): Dimensionality Reduction by Unsupervised K-Nearest Neighbor Regression

[10] Hastie, Tibshirani, Friedman (2009): The Elements of Statistical Learning, Kapitel 13.3

4.9. Principal Component Analysis (PCA)

Es werden *n* Beobachtungen erhoben von jeweils *p* Merkmalen. Bezeichne $X := (x_1, \dots, x_p)$ mit den *n*-dimensionalen Beobachtungsvariablen $x_1, \dots, x_p \in R^n$ die $n \times p$ Datenmatrix. Die *PCA* berechnet eine Orthonormalbasis $B = (b_1, \dots, b_p)$ des *p*-dimensionalen Vektorraums der Merkmale, sodass die Projektion auf einen $q < p$ – dimensionalen Unterraum, aufgespannt durch die ersten *q* Vektoren von *B*, einen möglichst großen Anteil der Information der Daten enthält. Die Vektoren von *B* heißen *Hauptkomponenten (principal components)*.

Das Ziel ist eine Komplexitätsreduktion durch Verringerung der Anzahl der Merkmale und zugleich eine Analyse der Struktur der Merkmale in den Daten durch Interpretation der Komponenten der Vektoren von *B*. Das Verfahren gehört zu den nicht überwachten Verfahren, d. h. es ist nicht auf eine Zielvariable *y* bezogen, sondern liefert allgemeine Strukturinformationen.

Beispiele für Anwendungen:

- Für eine Typklassifikation von Fahrzeugen sollen umfangreiche technische Daten analysiert werden.
- Für eine Regionalklassenanalyse sollen externe mikrogeographische Daten analysiert werden.

Die Hauptkomponenten sind die *p* Eigenvektoren der Kovarianzmatrix *S*=Cov(*X*,*X*) von *X* zu den Eigenwerten $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$. Die Spur $\sum_{i=1}^{p} \lambda_i$ von *S* entspricht der Varianz der Daten und $\frac{\lambda_i}{\text{Spur}(S)}$ ist der Anteil der *i*-ten Hauptkomponente an der Gesamtvarianz.

Bemerkungen:

1. Die *PCA* ist sensibel gegenüber der Skalierung einzelner Merkmalsvariablen. Wenn man nicht ausdrücklich unterschiedliche Dimensionierung in der Analyse zulassen möchte, müssen die Daten standardisiert werden.
2. Die *PCA* liefert aufsteigend dimensionierte Untervektorräume des R^p, die jeweils am nächsten zu den Beobachtungen *X* liegen, in dem Sinne, dass die Summe der euklidischen Abstände der x_i zu ihren Projektionen minimal ist.

3. Es gibt kein kanonisches Abbruchkriterium des Verfahrens. Man betrachtet den kumulierten erklärten Anteil der Gesamtvarianz $P(k) := \frac{\sum_{j=1}^{k} \lambda_i}{\text{Spur}(S)}$ und bricht z.B. bei *P(k)*>95 % ab.

R-Package:

```
stats, Funktionen: prcomp(), princomp()
```

R-Blogs:

https://www.r-bloggers.com/principal-component-analysis-using-r/
http://www.sthda.com/english/wiki/principal-component-analysis-in-r-prcomp-vs-princomp-r-software-and-data-mining.

Literatur:

[12] James et al. (2014): An Introduction to Statistical Learning, Abschnitt 10.2

4.10. Lineare Diskriminanzanalyse

Die lineare Diskriminanzanalyse ist ein Verfahren zur Klassifikation einer diskreten Zielvariablen *Y* mit Werten in {1, …, *K*} in Abhängigkeit von $x \in \mathbb{R}^p$, den dazugehörigen Werten von *p* Kovariablen. Unter der Annahme, dass P(*Y=k*)=π_k, *k=1,…,K* gilt und dass *X* gegeben *Y=k* die bedingte Dichte *f(x|Y=k) =:f*$_k$*(x)* besitzt, erhält man mit dem Satz von Bayes die bedingte Zähldichte

$$p_k(x) := \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l}. \qquad (2)$$

Einem $x \in \mathbb{R}^p$ wird diejenige Klasse k_0 zugeordnet, für die $p_k(x)$ maximal wird. Diese Maximierungsaufgabe erhält lineare Gestalt unter der zusätzlichen Annahme, dass *X* gegeben *Y=k* multivariat normalverteilt ist mit Erwartungswertvektor $\mu_k \in \mathbb{R}^p$ und Kovarianzmatrix $\Sigma \in \mathbb{R}^p \times \mathbb{R}^p$ (für alle *k* dieselbe Kovarianzmatrix). Eine leichte Umformung von (2) mit der Dichte $f_k(x)$ zeigt, dass die Maximierung von (2) in *k* äquivalent ist zur Maximierung von

$$\delta_k(x) := x^T \cdot \Sigma^{-1}\mu_k - \frac{1}{2}{\mu_k}^T \cdot \Sigma^{-1}\mu_k + \ln \pi_k$$

Aus der Lernstichprobe $(x_1,y_1), \ldots, (x_n,y_n)$ werden Schätzer $\hat{\pi}_k, \hat{\mu}_k$ und $\hat{\Sigma}$ für μ_k , π_k und Σ bestimmt. Der Schätzer $\hat{\pi}_k$ ist die relative Häufigkeit des Auftretens von *k* in den $y_1, \ldots, y_n$, die Parameter der multivariaten Normalverteilung werden wie üblich geschätzt. Einem Testdatum $x \in \mathbb{R}^p$ wird also die Klasse k_0 zugewiesen, für die

$$\delta_k(x) := x^T \cdot \hat{\Sigma}^{-1}\hat{\mu}_k - \frac{1}{2}{\hat{\mu}_k}^T \cdot \Sigma^{-1}\hat{\mu}_k + \ln \hat{\pi}_k$$

maximal ist.

R-package:

`MASS`

Literatur:

[10] Hastie, Tibshirani, Friedman (2009): The Elements of Statistical Learning, Abschnitt 4.3

[12] James et al. (2014): An Introduction to Statistical Learning, Abschnitt 4.4

5. Bewertung von Modellen

Die Vielzahl und Unterschiedlichkeit der vorgestellten Methoden gibt bereits einen Hinweis darauf, dass die Auswahl einer geeigneten Methode sehr wichtig ist. Hinzu kommt, dass bei unterschiedlicher Anwendung ein und derselben Methode Modelle unterschiedlicher Qualität entstehen. Die Entscheidung für das richtige oder das beste Modell ist dabei keinesfalls einfach. Zum einen steht neben der Vielzahl an möglichen Methoden eine ebenso unübersichtliche Anzahl an Möglichkeiten zum Vergleich von Modellen zur Verfügung. Zum anderen ist es bei Weitem nicht klar, was „richtig" oder „besser" konkret heißt. Diese Problematik erhält bei Nutzung der in *Kapitel 4* vorgestellten Verfahren des maschinellen Lernens eine besondere Bedeutung, da diese die Auswahl des konkreten Modells und insbesondere der benutzten Variablen im Allgemeinen automatisch übernehmen. Damit ist der Vergleich von Modellen bereits (implizit) in die Methoden eingebaut. Bei einem Vergleich der Verfahren kann damit die Wahl des Gütemaßes ein Präjudiz für das Abschneiden verschiedener Verfahren sein. Es besteht die Gefahr einer selbsterfüllenden Prophezeiung. Bei der Auswahl konkreter Modelle für den aktuariellen Einsatz ist deshalb eine intensive Beschäftigung mit den Grundlagen der Bewertung von Modellen anzuraten.

In diesem Kapitel werden verschiedene Ansätze zur Bewertung und Auswahl von Modellen vorgestellt und diskutiert. Die Abschnitte *5.1* und *5.2* gehen dabei auf die allgemeinen Grundlagen, sowie die Visualisierung und Interpretation der Ergebnisse ein. In den Abschnitten *5.3* bis *5.6* werden einzelne Gütemaße und Vergleichsmethoden vorgestellt.

5.1. Einleitung

Aktuarielle Modelle dienen in der Regel dazu, Erwartungen über die Zukunft zu prognostizieren. In der täglichen Arbeit reicht dabei der Prognosehorizont, also der Zeitraum, für den die Prognose getätigt wird, über Jahre. Die Vorhersagefähigkeit der verwendeten Modelle muss damit über einen deutlich längeren Zeitraum gewährleistet sein, als es für die meisten anderen Anwendungsbereiche statistischer Lernverfahren notwendig ist. Der Aktuar muss sich die Frage stellen, inwieweit die verwendete Methode eine solch langfristige Stabilität gewährleisten kann. Dafür genügt es im Allgemeinen nicht, sich auf die Messung von Korrelationen in historischen Daten zu stützen. Vielmehr stellt sich die Frage, ob ein Modell die Ursachen für eine bestimmte gemessene Abhängigkeit geeignet abbildet. Nur damit wäre es möglich, die zukünftige Wirkung korrekt oder zumindest geeignet vorherzusagen.

In der täglichen Arbeit muss man aber oft feststellen, dass ein formaler Ursache-Wirkung-Zusammenhang auf Basis der verwendeten Daten und mit Hilfe der zur Verfügung stehenden Methoden gar nicht dargestellt werden kann. Ein „physikalisches" Modell von dem, was ist, erscheint eher eine Illusion als tatsächlich realisierbar. Umso wichtiger ist es, Aussagen zur Eignung von Modellen treffen zu können.

Methoden zum Vergleich und zur Bewertung von Modellen sollen eine Objektivierung der Auswahl des richtigen oder des am besten geeigneten Modells aus der Bandbreite zur Verfügung stehender Modelle ermöglichen. Die Wahl eines konkreten Modells kann dabei sowohl quantitative als auch qualitative Aspekte berücksichtigen. Während wir zum Beispiel bei statistischen Tests eine Ablehnung der Nullhypothese auf Basis einer realisierten Teststatistik erwarten, setzen graphische Verfahren eine Interpretation durch den Anwender voraus. Für die aktuarielle Anwendung ist dabei die modellierte Zielgröße im Regelfall eine Zufallsvariable und keine durch einfaches Anschauen eindeutig bestimmbare Eigenschaft der untersuchten Objekte. Dieser Umstand wird in vielen Fällen ein anderes Herangehen an den Modellvergleich erfordern, als bei einigen anderen prominenten Anwendungen des maschinellen Lernens.

Bevor allerdings überhaupt ein Modell gewählt werden kann, müssen geeignete Methoden zum Vergleich der Modelle ausgewählt werden. Dabei kann die Wahl der Methode bereits die Wahl des finalen Modells determinieren. Es drängt sich damit die Frage auf: Nach welchen Kriterien sollten Methoden zum Vergleich von Modellen ausgewählt werden?

Um einen Ansatzpunkt zur Beantwortung der Frage zu erhalten, führen wir uns die Natur von Modellvergleichen vor Augen. Grundsätzlich sind mathematisch folgende zwei Sichtweisen möglich, um die Güte von Modellen zu messen:

- Zum einen könnte das Modell bevorzugt werden, das die Daten am besten beschreibt. Dabei wäre zu messen, wie nah die zu vergleichenden Modelle die Daten fitten. Um ein geeignetes Maß dafür zu erhalten, ist es also notwendig Aussagen zur Verteilung des Fehlers der Zielgröße zu tätigen, wobei einige Verfahren auf die explizite Angabe verzichten. Für diesen Ansatz sucht man eine Darstellung für $P(D|M)$, also die Wahrscheinlichkeit der Daten bei gegebenem Modell.
- Zum anderen könnte man das Modell suchen, aus welchem die Daten am besten hervorgegangen sein könnten. Das ist für aktuarielle Fragestellungen, wie schon erläutert, häufiger das eigentliche Anliegen. Dieser Ansatz erfordert aber neben dem Wissen über den Zufallseffekt in den Daten eine (explizite oder implizite) Annahme über die Verteilung des Modell- bzw. Parametrisierungsfehlers. Bei diesem Ansatz bewertet man $P(M|D)$, also die Wahrscheinlichkeit des Modells bei gegebenen Daten.

Die zweite Sichtweise bietet tatsächlich noch einen weiteren Erkenntnisvorteil. Mit dem Satz von Bayes sieht man, dass es sich um eine Erweiterung des ersten Ansatzes handelt (siehe *5.5*). Die berücksichtigte Unsicherheit des Modells wirkt dabei als Korrektiv gegen eine zu große Modellkomplexität und damit einem Overfitting entgegen. Tatsächlich lassen sich die Tuningparameter von Verfahren des maschinellen Lernens in diesem Sinn als Hyperparameter der zugrunde liegenden A-priori-Verteilung des unbekannten Modellparameters interpretieren, die man zum Teil sogar explizit darstellen kann.

Aus beiden Ansätzen ist aber klar: es existieren keine „verteilungsfreien" Vergleichsmöglichkeiten. Bestenfalls wird auf die Angabe der zugrundeliegenden Verteilung verzichtet. Man kann sich aber die Frage stellen, ob das Wissen von der und die Kontrolle über die Verteilungsannahme nicht einen bedeutenden Mehrwert für die aktuarielle Arbeit bieten.

Neben den diskutierten Überlegungen spielen bei der Wahl der Vergleichsmethoden folgende Fragen eine Rolle:

- Welche Aussage soll der Vergleich liefern? Wird vom Ergebnis eine konkrete Entscheidung erwartet oder wird ein qualitatives Ergebnis bevorzugt?
- Stehen die notwendigen Basisinformationen in den zugrundliegenden Modellen vollständig zur Verfügung?
- Welcher Verteilungstyp liegt den Daten zugrunde oder ist plausibel? Auf welchem Träger bewegen sich die Zieldaten? Sind weitere Eigenschaften wie z. B. die Schiefe prinzipiell bekannt?
- Entstammen die zu vergleichenden Modelle demselben Verfahren? Sollen genestete oder auch nicht genestete Modelle miteinander verglichen werden?

In der Praxis ist zudem zu prüfen, ob die notwendigen Daten und Werkzeuge zur Verfügung stehen und ob die Auswertungen performant genug ausgeführt werden können.

Bei allen Einschränkungen, die üblicherweise in der Praxis bestehen, ist unbestritten, dass die Wahl einer ungeeigneten Methode zur Auswahl von Modellen ungeeignete Ergebnisse hervorbringen wird. In der Regel kann durch die Wahl der Methode bereits das Ergebnis vorbestimmt werden und ein Bias zu einer Modellklasse entstehen. Letztlich können dadurch, wie schon erwähnt, selbst-erfüllende Prophezeiungen nicht ausgeschlossen werden. Um diese Gefahr zu reduzieren, ist es dringend anzuraten, für die Modellwahl auch bei klassischen Methoden einen anderen Datensatz heranzuziehen, als für die Anpassung des Modells.

Für Verfahren des maschinellen Lernens werden die verfügbaren Daten sogar in drei Datensätze zerlegt: den Trainings-, Validierungs- und Testdatensatz. Der Trainingsdatensatz wird verwendet, um die Modelle anzupassen. Die Wahl der Hyperparameter, die die implizite Modellwahl bestimmen, wird auf Basis der Validierungsdaten ermittelt. Der Testdatensatz (auch *holdout sample* genannt) wird zur Einschätzung des Prognosefehlers verwendet. Um einen Bias durch die Einteilung zu verhindern, sollte zudem nach dem Ausscheiden der Testdaten die Zuordnung zu Trainings- bzw. Validierungsdaten stetig neu festgelegt werden. Im Gegensatz dazu entfällt die Notwendigkeit zur Aufteilung von Trainings- und Validierungsdaten bei klassischen statistischen Verfahren.

Eine der wesentlichen Aufgaben beim Erstellen und Vergleich von Modellen ist die Vermeidung von *Over-*, *Under-* und *Missfitting*. Diese Begriffe bezeichnen jeweils ein fehlerhaftes Anpassungsverhalten, also eine nicht korrekte Trennung von systematischen und zufälligen Effekten. *Overfitting* bezeichnet die Anpassung zufälliger

Effekte im systematischen Teil, *Underfitting* das Gegenteil. Während das maschinelle Lernen oft zu *Overfitting* neigt, sind klassische Methoden anfällig für *Underfitting*, da Erklärungsgrößen nicht gefunden werden. Auf der anderen Seite können bei diesen Methoden extrinsische Vorgaben die Qualität und Stabilität erhöhen, da die Modellkomplexität durch sie von Beginn an eingeschränkt wird. Diese Vorgaben basieren in der Regel auf Erfahrungswissen oder gesundem Menschenverstand – beides ist in den Daten nicht oder nur unzureichend abgebildet. Wenn allerdings a priori falsche Annahmen für das zu erzielende Modell getroffen werden, droht *Missfitting*. Das verwendete a priori Wissen dominiert die Modellwahl dann unangemessen. Unter einem *Missfitting* versteht man im Allgemeinen die Anpassung nicht vorhandener Effekte durch die spezifische Wahl eines Modells. Die Möglichkeit der Berücksichtigung von a priori Annahmen erfolgt in der Regel bei maschinellem Lernen nicht, was als Vor- und Nachteil gesehen werden kann.

Alle drei Varianten von Anpassungsfehlern benötigen spezifische Maßnahmen, um sie beheben zu können. In der Regel zeigen Gütemaße für den Modellvergleich aber nur, wie gut oder schlecht ein Modell gegen ein anderes bei gegebenen Daten abschneidet. Ob in einem Fall *Missfitting* und im anderen *Overfitting* relevant ist, lässt sich gewöhnlich nicht sagen. Um das festzustellen, sind oftmals mehrere Vergleichsmethoden nötig und vor allem auch graphische Aufbereitungen hilfreich.

Die Überlegungen zielen bisher darauf, ein konkretes Modell hinsichtlich seiner Erklärungskraft zu bewerten. Dieser für die eigentliche Modellbildung fundamentalen Aufgabenstellung steht die Frage nach einer Bewertung der wirtschaftlichen Eignung des gewählten Modells gegenüber. Diese kann und will sich oft nicht mit Details der konkreten Modelle oder möglichen Alternativmodelle befassen, sondern sucht eine Antwort auf die Frage, wie das gewählte Modell seine Aufgabe im gegebenen wirtschaftlichen Umfeld erfüllt.

Zur Beantwortung derartiger Fragestellungen stehen die Werkzeuge des traditionellen aktuariellen Controllings zur Verfügung. Hier wird man ganz bewusst ohne Bezug auf Details von Tarif oder Modell auf eine Dokumentation der aus dem Tarif im Ergebnis zu erwartenden Schadenhäufigkeit und Schadenhöhenverteilung zurückgreifen, die ohnehin im Rahmen der aktuariellen Tarifentwicklung entsteht. Auf Basis dieser Dokumentation liefert die einfache Gegenüberstellung der unter dem kalkulierten Tarif erzielten Volumina zusammen mit der sich realisierenden Schadenerfahrung die Aussage, ob das zugrunde liegende Modell seine Aufgabe geeignet erfüllt. Eine derartige Herangehensweise ist im Rahmen des Control Cycle eine Selbstverständlichkeit und sollte nicht nur auf aggregierte Bestände angewandt werden, sondern insbesondere auch auf Teilbestände, die mit Hilfe komplizierter statistischer Modelle und Verfahren bepreist wurden. Sie liefert dann das notwendige Sicherheitsnetz, wenn noch weniger erprobte Tarifierungsansätze in überschaubaren Teilbeständen angewandt wurden, und eine mögliche Fehltarifierung im Rahmen der Risikobewirtschaftung geeignet berücksichtigt ist. Dadurch erlaubt ein sachgerechter Controlling-Prozess aus Risikogesichtspunkten durchaus ein Experimentieren mit innovativeren Methoden.

Der Nutzen von Data-Science-Methoden ist neben der Effektivität und Präzision zusätzlich anhand der Erklärbarkeit der Modellergebnisse zu bewerten. Hierbei sind zwei Aspekte zu unterscheiden: Einerseits betrachtet man die formal-mathematische Reproduzierbarkeit. Diese ist bei linearen Regressionen oder Baumverfahren sehr gut und nimmt bis zu tiefen neuronalen Netzen ab. Andererseits ist die heuristisch-didaktische Perspektive zu beachten, also: wie einfach ist eine Klassifikation oder Regression einem Fachfremden zu erläutern? Hierbei zeigt sich, dass unterschiedliche Methoden verschiedene Stärken haben. Grundsätzlich jedoch gilt, dass parametrische Lösungen gegenüber Blackbox-Ergebnissen dahingehend im Vorteil sind, als dass man generell bei kleiner Veränderung des Eingabedatensatzes auch nur eine kleine Veränderung des Modellergebnisses erhält.

Die Rolle des Aktuars ist hinsichtlich der Modellauswahl und -anpassung also nicht nur eine gestaltende, sondern auch eine moderierende. Er muss stets die Anforderungen der Erklärbarkeit seiner Ergebnisse im Auge behalten. Das gilt insbesondere in Bezug auf die verbraucherschutzrechtlich geregelte Auskunft gegenüber dem Kunden oder das Audit des Regulierers. Hierbei sei angemerkt, dass der Versicherer natürlich nicht verpflichtet ist, dem Kunden Tarifierungsdetails von der Art eines Geschäftsgeheimnisses preiszugeben. Wie auch schon in den vorangegangenen Kapiteln diskutiert ist jeweils zu klären, wie die Anforderungen an Nachvollziehbarkeit und Transparenz im Einzelnen zu erfüllen sind. Darüber hinaus ist es problematisch, das Management oder den Regulierer von einer Auskömmlichkeit des Tarifs nur anhand der Erwartungswerte eines Modells überzeugen zu wollen, wenn die zugrunde liegende Systematik nicht erschöpfend erklärt werden kann. Es versteht sich im Rahmen von Solvency II dabei von selbst, dass relevante Überlegungen und quantitative Tests zur Stabilität der verwendeten Lösung anzustellen sind. Wenn nun die verwendete Regressionsmethode beispielsweise keine Konfidenzintervalle zu berechnen vermag, so ist ihre Eignung zwar nicht grundsätzlich in Frage zu stellen. Man sollte aber mindestens von zusätzlichem Dokumentations- und Testaufwand ausgehen.

Die meisten Ansätze zur Bewertung der Modellgüte beziehen sich auf überwachte Lernverfahren. Während das überwachte Lernen eine klare Zielfunktion aufweist, die quantitative Performancemessungen ermöglicht, können für unüberwachte Modelle in der Regel keine allgemeingültigen Kennzahlen angegeben werden, da die zugrunde zu legende Bewertungsmethode ebenfalls Teil der Problemstellung ist. Anders gesagt: Ein unüberwachter Lernerfolg ist ohne seinen konkreten Kontext wertlos. Wird etwa eine unüberwachte Mustererkennung auf einem semi- oder unstrukturierten Datensatz mit der Absicht durchgeführt, neue Merkmale für ein Tarifierungsmodell zu extrahieren, so kann nicht der Output der Mustererkennungsmethode direkt bewertet werden, sondern nur die indirekte Veränderung der Tariffunktion unter Einbeziehung der extrahierten Merkmale. In diesem Falle sind daher die Kenngrößen der anschließenden Regression heranzuziehen, obwohl die durchgeführte Untersuchung selbst keine Regression darstellt. Demgegenüber könnte der gleiche Output der Mustererkennung möglicherweise ebenso für eine Klassifikation dienen, etwa bestimmter Kundensegmente bezüglich ihrer Merkmale. In diesem Fall

sind Kenngrößen der folgenden Klassifikation zu verwenden, um die relative Performance des Lernerfolgs zu bewerten. So decken unüberwachte Methoden üblicherweise Arbeitsschritte der Exploration und Vorverarbeitung ab und sind stets gemeinsam mit der Gesamtzielsetzung zu betrachten. Spezielle Verfahren zur Bewertung von unüberwachtem maschinellem Lernen betrachten wir deshalb hier nicht.

5.2. Visualisierung und Interpretation

Gerade in der praktischen Anwendung reicht es oftmals nicht aus, abstrakte Hyperparameter auf einem Validierungsdatensatz statistisch zu optimieren, um einem Modell „vertrauen" zu können. Neben der rein statistischen Bewertung der Modellgüte ist es bei vielen Modellierungsaufgaben ein Hauptanliegen, die wesentlichen Treiber und Effekte eines Modells zu erkennen und zu verstehen. Diese Aufgabe gestaltet sich per Definition besonders für die Verfahren schwierig, die der Kategorie „Black Box" zugeordnet werden. Um die Wirkungsweise eines Modells besser zu verstehen, eignen sich grafische Analysen, die eine Vielzahl von Informationen in einer für den Menschen schnell und leichter zu interpretierenden Art bereitstellen.

Im Folgenden werden ausgewählte Visualisierungen beschrieben. Diese sollen dem Leser einen ersten Eindruck vermitteln – es gibt zahlreiche weitere nützliche grafische Analysen, auch teilweise abhängig von dem verwendeten Modellierungsverfahren und der modellierten Größe.

5.2.1. Globale vs. lokale Interpretierbarkeit

Zur Modellanalyse kann zwischen globaler und lokaler Interpretierbarkeit unterschieden werden. Bei einer globalen Interpretation werden die Vorhersagen in Gänze betrachtet und durch Aggregation stark approximiert (z. B. „mit steigender Fahrleistung erhöhen sich die durchschnittlichen Prädiktionen"). Untersucht man das Modellverhalten lokal, schränkt man sich auf ein kleineres Teilsegment ein (z. B. „Mit steigender Fahrleistung erhöhen sich die Prädiktionen für Benziner deutlich stärker als für Diesel-Fahrzeuge").

Mathematisch betrachtet korrespondieren lokale Interpretationen mit bedingten Verteilungen auf Segmenten. Aktuariell betrachtet lassen sich bedingte Verteilungen zweier Segmente, die sich unterscheiden, auf Korrelations- und/oder Interaktionseffekte zurückführen. Die im Modell enthaltenen Interaktionseffekte sind bei klassischen linearen Verfahren bekannt (weil sie explizit in das Modell aufgenommen wurden), während bei vielen (nicht linearen) Modellen des maschinellen Lernens (ML-Modelle) diese Effekte implizit zahlreich und nicht offensichtlich vorhanden sind. Das bestimmt maßgeblich die Aufgabenstellung des Modellierers. Während bei klassischen linearen Verfahren der Modellierer auf der Suche nach Interaktionseffekten

ist, die er manuell in das Modell aufnimmt, um es zu verbessern („*Underfitting* vermeiden")[3], versucht er bei (nicht linearen) ML-Modellen die im Modell zahlreich vorhandenen Interaktionen im Wesentlichen zu verstehen und zu plausibilisieren („*Overfitting* vermeiden").

5.2.2. Randverteilungen "Modell vs. Beobachtung"

Eine einfache und effektive Art und Weise, die Güte eines Modells zu bestimmen, ergibt sich durch den direkten Vergleich der durchschnittlichen Beobachtungen und Prädiktionen auf verschiedenen Segmenten bzw. Bestandsquerschnitten. Querschnitte sind dazu beispielsweise über die zur Verfügung stehenden (untersuchten) Merkmale zu bilden („Randverteilungen"). Der Abgleich kann sowohl auf den Trainings- als auch auf den Testdaten erfolgen, um *Over*-, *Under*- und *Missfittings* festzustellen. Jede Abweichung des Modells von der Beobachtung wird prinzipiell als zufälliger Effekt interpretiert – oder aber das Modell erklärt einen systematischen Effekt nicht korrekt und ist dann folgerichtig anzupassen.

In *Abbildung 3* wurde am Beispiel des in der Kfz-Risikomodellierung gängigen, vorab gruppierten Merkmals „Fahrleistung" für jede Ausprägung sowohl die durchschnittliche Modellvorhersage als auch die Beobachtung visualisiert. Es ist zu erkennen, dass die Vorhersagen des Modells die Beobachtungen gut beschreiben. Mit diesem Prinzip kann der Datenbestand weiter segmentiert werden, z.B. um den Vergleich separat für Benzin- und Diesel-Fahrzeuge anzustellen.

[3] Das bedeutet im Umkehrschluss nicht, dass klassische Modelle frei von *Overfitting* sind.

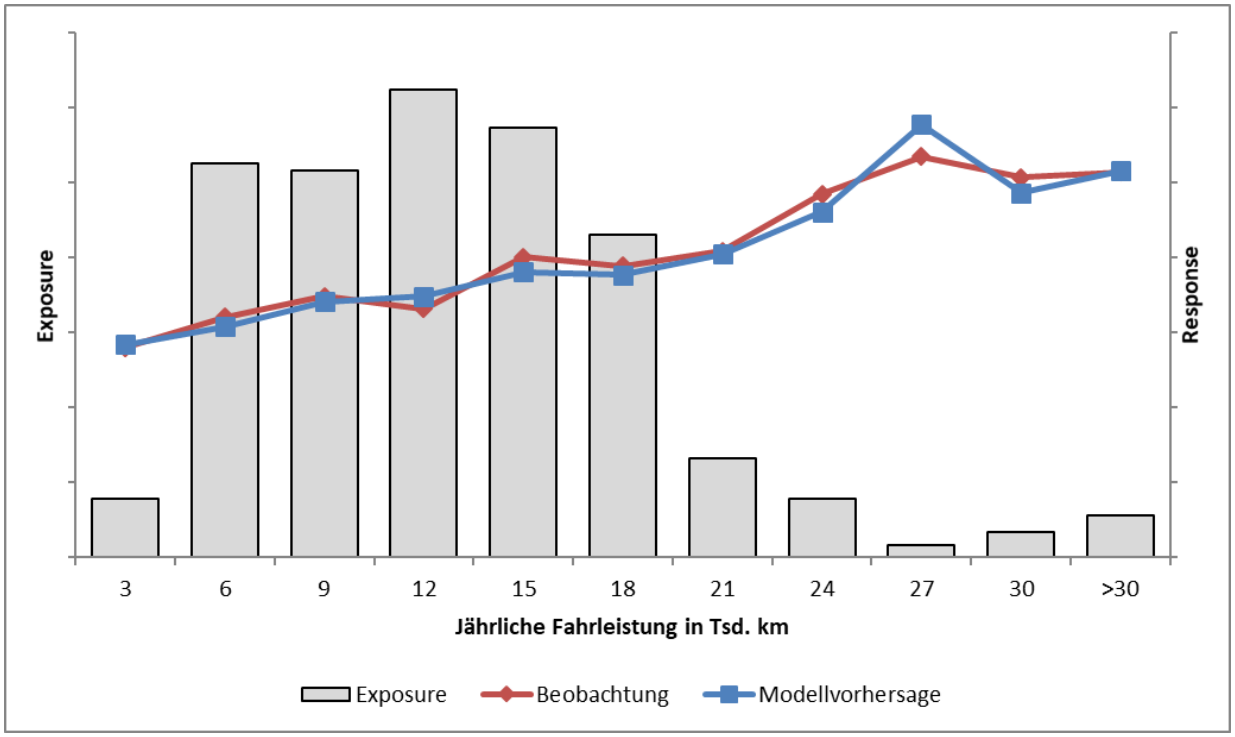


Abbildung 3: Beispielhafte Darstellung Randverteilung „Modell vs. Beobachtung“

Randverteilungen geben dem Modellierer ein tieferes Verständnis darüber, in welchen Segmenten die Stärken und Schwächen des Modells liegen. Die Bewertung mehrerer eindimensionaler Randverteilungen ist in der Regel zeitintensiv. Je höherdimensional die Randverteilung, desto größer die Gesamtzahl aller möglichen Kombinationen und somit die Komplexität der Bewertung einer Randverteilung. Daher können verschachtelte Querschnitte („die jungen Fahrer aus Süddeutschland mit altem Fahrzeug, aber hoher kW …“) mit dieser Technik in der Regel nicht systematisch erkannt bzw. nur auf Verdacht geprüft werden.

Der Vergleich ein- bzw. zweidimensionaler Randverteilungen erinnert an die Analyse von Verallgemeinerten Linearen Modellen (GLM*s*), kann aber zur Evaluierung für alle Arten von Modellen angewandt werden; an dieser Stelle sei jedoch darauf hingewiesen, dass insbesondere bei nicht parametrischen Verfahren zu prüfen ist, ob Marginaluntersuchungen überhaupt anwendbar sind und Einblick in die verwendete Heuristik adäquat zulassen. Die Visualisierung einer Randverteilung liefert kein „hartes“ Gütekriterium, sondern benötigt eine Interpretation und Bewertung. Denkbar sind auch darauf basierende Definitionen von statistischen Kennzahlen, wie z.B. eine mittlere gewichtete absolute Abweichung.

5.2.3. Partial Dependence Plots

Mit Hilfe von *Partial Dependence Plots (PDP)* können die Modelltreiber – auch von „Black Box“ Modellen – besser verstanden werden. *PDP* zeigen den „marginalen Effekt“ eines Merkmals auf die Vorhersagen eines Modells auf. Die Fragestellung erinnert an die GLMs und lautet: Wie ändern sich die Modellvorhersagen, wenn man die Ausprägungen *eines* Merkmals variiert?

Für GLMs lässt sich diese Frage für jedes Risiko eindeutig beantworten, solange keine Interaktionen mit diesem Merkmal vorliegen. Die Marginaleffekte entsprechen den Relativitäten (d. h. den multiplikativen Faktoren in multiplikativen Modellen) und lassen den Modellierer die Treiber des Modells quantifizieren. Das Prinzip kann darüber hinaus auch für komplexere nicht lineare Modelle genutzt werden. Anstatt für ein ausgewähltes Basisrisiko sind hier die Ausprägungen eines Merkmals für mehrere ausgewählte Risiken zu variieren und die dazugehörigen Modellvorhersagen zu berechnen. Über diesen repräsentativen Risikosatz lassen sich die Modellvorhersagen mitteln und der Einfluss des Merkmals einschätzen. Bei Erweiterung der Betrachtungsweise auf Abweichungen und Durchschnitte von Teilsegmenten oder ausgewählten Einzelrisiken über den Mittelwert hinaus lassen sich Aussagen zu Monotonie, Nichtlinearität und im Modell vorhandene „komplexe" Interaktionen ableiten. So gelangt man von einer globalen zu einer lokalen Interpretation.

Abbildung 4 zeigt eine beispielhafte Visualisierung. Es ist zu erkennen, dass sich mit steigender Fahrleistung die Prädiktionen im Schnitt erhöhen. Die Linie „min Spread" („max Spread") basiert auf dem Datensatz des repräsentativen Risikosatzes, der am wenigsten (meisten) spreizt[4]. Dadurch lässt sich aufzeigen, dass im Modell anscheinend komplexe Interaktionseffekte mit dem Merkmal Fahrleistung vorliegen. Während es Risiken gibt, bei denen die Erhöhung der Fahrleistung die Modellvorhersage stärker nach oben treibt, deutet die „min Spread"-Linie auf Datensätze hin, bei denen die Fahrleistung sogar einem gegenläufigen Trend folgt. Damit ist die Monotonie für das Merkmal Fahrleistung nicht gegeben, was in der Praxis – beispielsweise bei einem Tarifmodell – ein Problem darstellen kann. In diesem Fall empfiehlt es sich, das Modell zu überarbeiten und gegebenenfalls die Monotonie zu erzwingen.

[4] Spreizung geeignet definiert

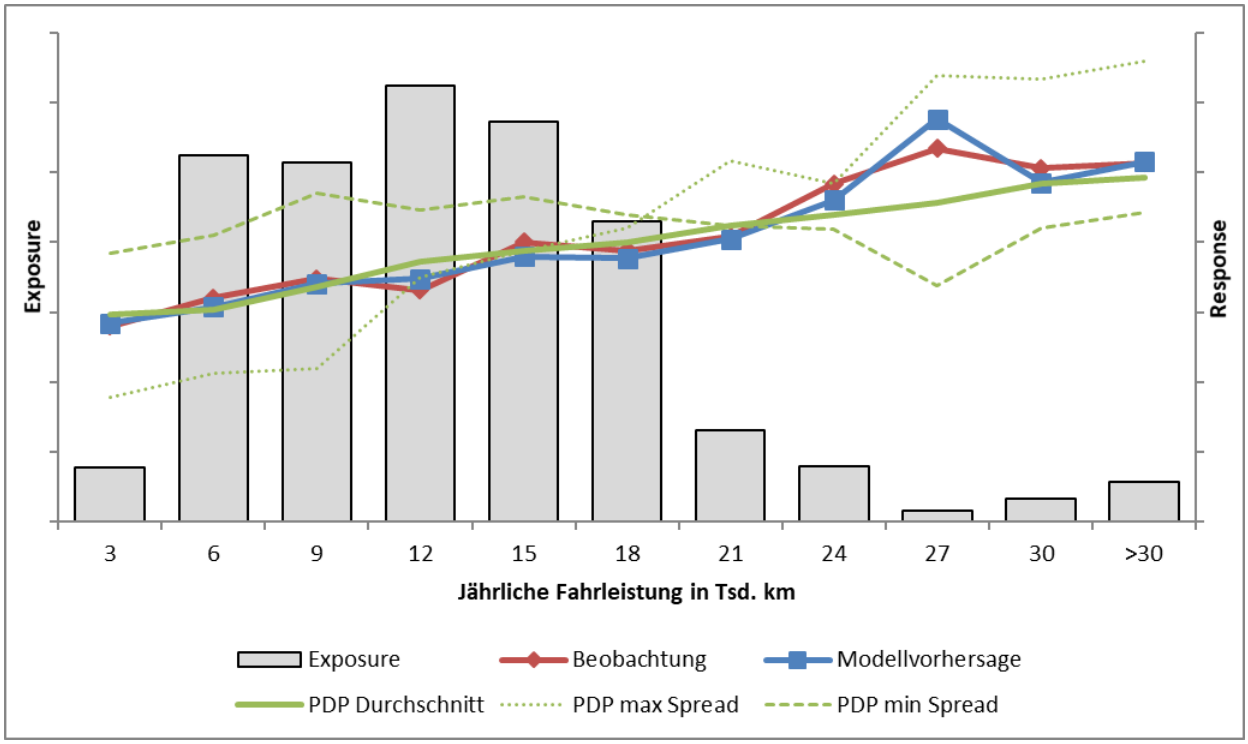


Abbildung 4: Beispielhafte Darstellung Partial Dependence Plot

5.2.4. Liftplot

Mit Hilfe eines *Liftplots* kann grafisch geprüft werden, ob für Risiken mit geringen bzw. hohen Modellvorhersagen tatsächlich auch geringe bzw. hohe Werte beobachtet wurden. Es wird also die Modellgüte in den Randbereichen überprüft. Die Idee ist einfach: Auf einem Datensatz (i. d. R. Testdaten) werden die Daten nach Modellvorhersage (z. B. Schadenhäufigkeit oder -durchschnitt) sortiert, über feste Abstände oder Perzentile gruppiert und die durchschnittlichen Beobachtungen gegen die Modellvorhersagen aufgetragen.

Ein *Liftplot* in 2 %-Perzentilschritten ist in *Abbildung 5* aufgetragen. Die Beobachtungen folgen tendenziell der per Konstruktion ansteigenden Modellvorhersage und bestätigen – abgesehen von Zufallsschwankungen – die „Treffsicherheit" des Modells auch an den extremen Rändern.

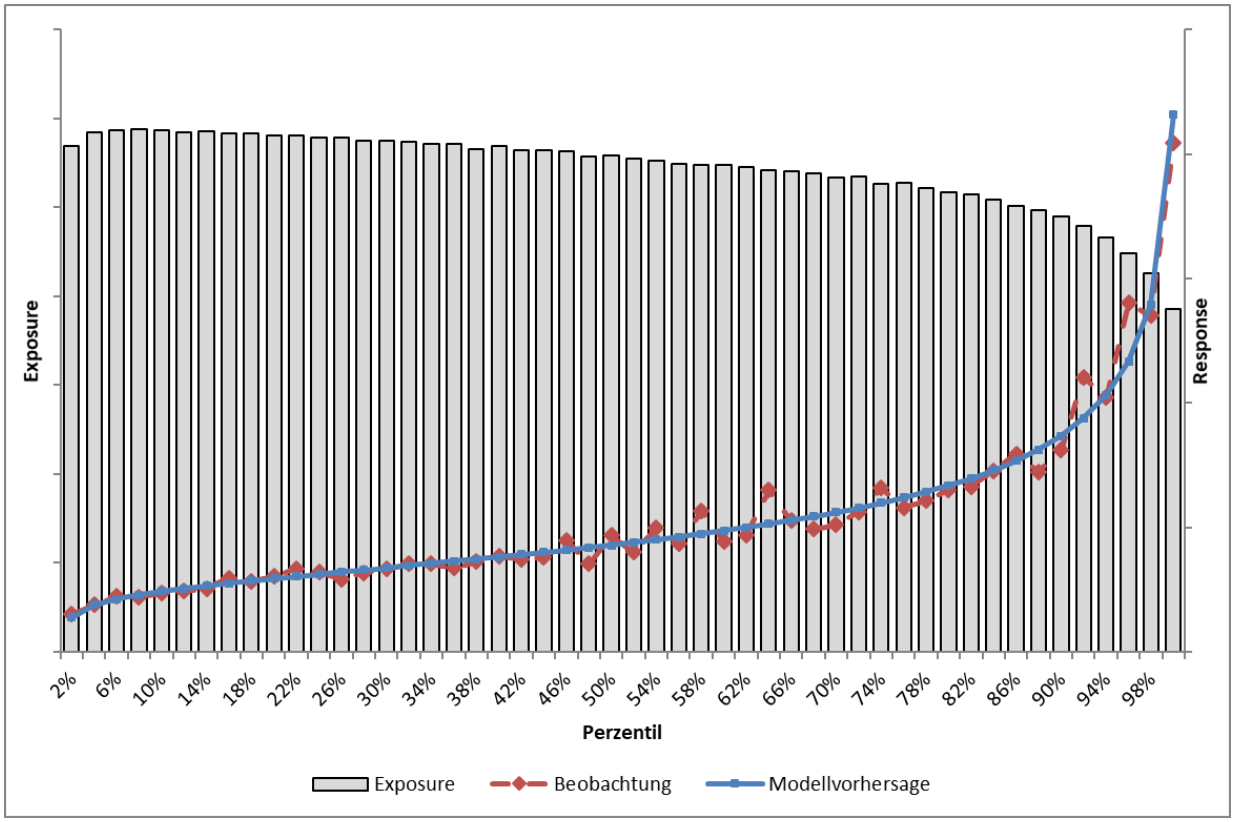


Abbildung 5: Beispielhafte Darstellung Einfacher Lift Plot

5.2.5. Double Lift Charts

Um zwei unterschiedliche Modelle zu vergleichen, eignet sich ein *Double Lift Chart*. Bei diesem werden die (Test-)Daten, sortiert nach dem Verhältnis der beiden Modellvorhersagen, gruppiert und die Gruppenmittel bestimmt. In der gewählten Segmentierung – beispielsweise in 10 %-Schritten oder in Grenzen mit gleichem Exposureanteil pro Segment – wird damit der visuelle Vergleich ermöglicht, welche Modellergebnisse näher an den Beobachtungen liegen.

Im Beispiel, vgl. *Abbildung 6,* beschreibt das grüne Modell die Beobachtungen deutlich besser als das blaue Modell – auch an den Rändern, in denen die beiden Modelle stark unterschiedliche Prädiktionen liefern. Der *Double-Lift-Chart* liefert in diesem Fall ein eindeutiges Ergebnis, welches Modell die Beobachtungen im Testdatensatz besser beschreibt.

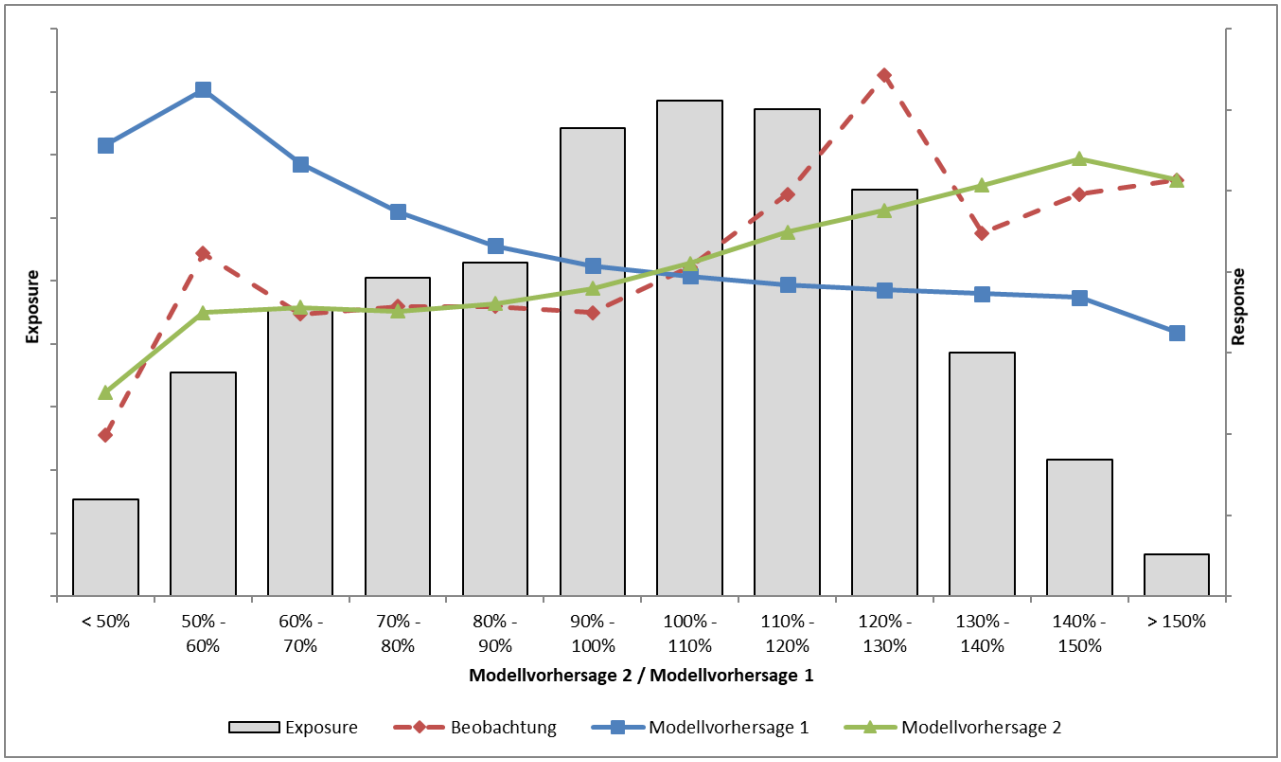


Abbildung 6: Beispielhafte Darstellung Double Lift Chart

5.2.6. Residuenanalyse

Residuen stellen die Abweichung der Modellvorhersage zur Beobachtung für jedes einzelne Risiko dar. Verschiedene Definitionen je nach Problemstellung sind gängig: absolute oder relative Abweichungen werden teilweise standardisiert und skaliert. Allgemein gesagt sollten die Residuen eines „guten" Modells zufällig verteilt sein; alle systematischen Effekte können durch das Modell beschrieben werden, die Residuen spiegeln die verbleibenden zufälligen Effekte wider.

Entscheidend für die Interpretation ist die Definition der Residuen, die von der Fragestellung und der modellierten Größe abhängt. Die Bestimmung der Residuen erfolgt auf Einzeldatensatzebene oder leicht aggregiert. Beispielsweise können bei 0/1-Modellen Datensätze zufällig in kleine Gruppen eingeteilt werden, um auf Gruppenebene sinnvolle residuale Größen bestimmen zu können. Diverse Visualisierungen der Residuen sind üblich. Eine gängige Methode ist es, die Residuen gegen die Modellvorhersagen darzustellen. Fallen dabei bestimmte Muster in den Residuenverteilungen auf, ist dies ein Indiz dafür, das Modell und das dahinter liegende Verfahren kritisch zu prüfen. Des Weiteren können extreme Ausreißer festgestellt werden; bei einem zu hohen Einfluss auf das Modell können diese aus den Daten für einen erneuten Modellfit ausgeschlossen werden. Beispielsweise können die Residuen auch gegen die beschreibenden Merkmale aufgetragen werden. Eine komprimierte Darstellung der Residuenverteilung zum Beispiel über Box-Whisker-Plots kann helfen, unterschiedliche Residuenverteilungen einfacher zu entdecken.

Der Residuenplot in *Abbildung 7* deutet an, dass die Residuenstreuung mit zunehmender Modellvorhersage für Modell 1 abnimmt, während die Streuung für Modell

2 strukturell gleich verläuft. Hier wäre für Modell 1 zu untersuchen, ob dies an einem *Overfitting* bei hohen oder einem *Underfitting* bei niedrigen Prädiktionen liegt – oder das gewählte Verfahren für das Problem nicht geeignet ist.

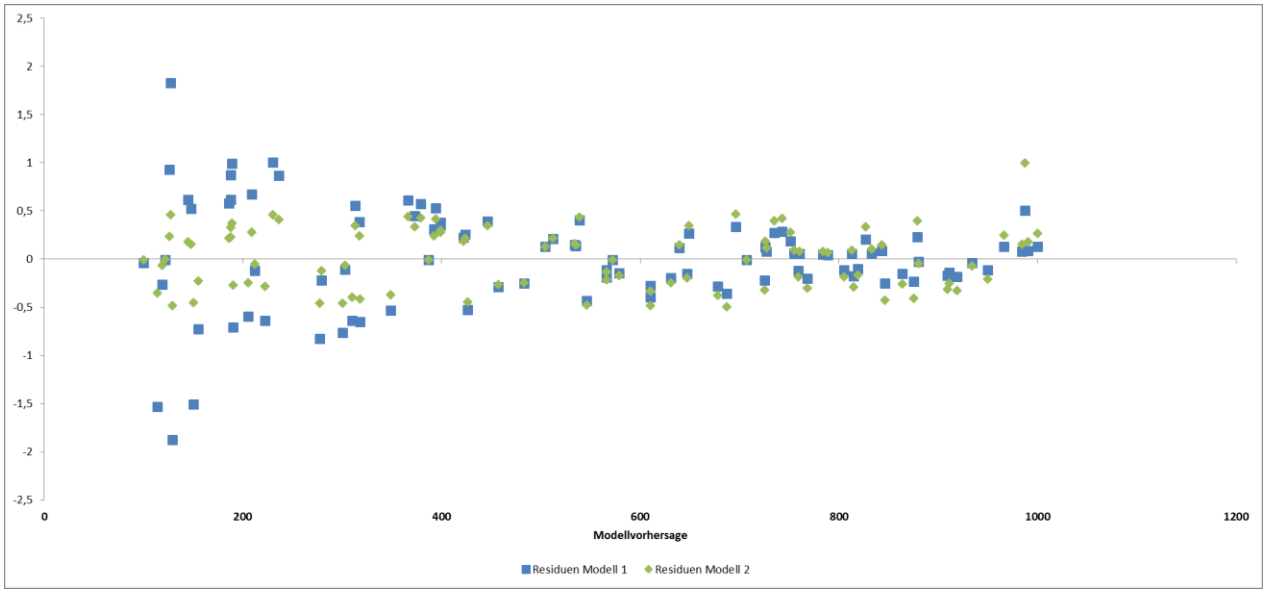


Abbildung 7: Beispielhafte Darstellung Residualanalyse

5.2.7. Surrogate Models

Ein *Surrogate Model* (deutsch: *Stellvertretermodell*) ist ein einfaches Modell, das dabei helfen soll, ein komplexes Modell zu interpretieren. Es wird auf den Prädiktionen des komplexen Modells trainiert (*Remodellierung*[5]). Die Erkenntnisse aus dem einfach zu interpretierenden Modell werden auf das komplexe Modell übertragen.

So kann z. B. der Output eines komplexen *GBM* mit einem einfachen Baumverfahren remodelliert werden. Die Verschachtelungen des einfachen Baums deuten auf die einflussreichsten nicht linearen Effekte hin. Ebenso kann das *GBM* mit einem GLM remodelliert werden, um die wichtigsten beschreibenden Merkmale und Zweifach-Interaktionen zu verstehen.

Kritisch zu hinterfragen ist, wie gut das Surrogate Model das komplexe Modell repräsentiert. Durch die Vereinfachung gehen wesentliche Teile des komplexen Modells verloren – sonst würde das komplexe Modell auch keinen Mehrwert liefern können.

5.2.8. Mehrdimensionale Grafiken

Residuen und beliebige weitere Größen können auch mehrdimensional dargestellt werden. Bekannte Techniken dafür sind Heatplots, 3D-Gebirge und (aggregierte)

[5] Das Modellieren eines Modelloutputs bezeichnet man als Remodellierung.

Scatterplots. Damit lassen sich beispielsweise die Interaktionseffekte eines Modells zwischen zwei Merkmalen aufzeigen.

5.3. Gütemaße

5.3.1. Gütemaße ohne explizite Verteilungsannahme

Summe der absoluten Abweichungen und der Abweichungsquadrate

Bei metrischen Variablen bietet es sich grundsätzlich an, Fehler anhand der Abweichungen zwischen Prognose $\hat{y}_i$ und Beobachtung y_i der Zielvariablen mit dem *Sum of squared errors (SSE)*.

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

zu messen (oder alternativ mit $\sum_{i=1}^{n} |y_i - \hat{y}_i|$). Bei der Verwendung des *SSE* (bzw. der Summe der Absolutbeträge) als Gütekriterium für Modelle ist zu beachten, dass bei einigen Verfahren, beispielsweise beim linearen Modell bzw. beim GLM mit Normalverteilung und identischer Linkfunktion, gerade der *SSE* zur Parameterschätzung minimiert wird. Somit kann dieses Kriterium höchstens als ein Kriterium unter mehreren zum Vergleich der Güte unterschiedlicherer Modellansätze herangezogen werden.

Sind Volumina v_i vorhanden, ist es notwendig, diese in die Gütemaße einzubeziehen – z. B. mit

$$\sum_{i=1}^{n} v_i (y_i - \hat{y}_i)^2.$$

Der *SSE* ist insbesondere ein Beispiel einer Verlustfunktion, die wegen des oben beschriebenen Zusammenhangs mit der linearen Regression, der Symmetrie und der mathematischen Handhabbarkeit eine besondere Stellung einnimmt. Allerdings können sich bei der Bewertung von Modellen auch andere Funktionen, die einen möglichen ökonomischen Verlust nach schlechter Modellwahl bewerten, als geeignet erweisen.

Kreuzvalidierung (Cross-Validation)

Um *Overfitting* zu untersuchen, werden die Maße zur Analyse der Modellgüte auf den Trainings- und Testdaten verglichen. Die Differenz der (volumenadjustierten) Gütemaße ist ein Messwert für die Stabilität des jeweiligen Modells. Abgesehen davon, dass diese Differenz u. U. von der Wahl der Zerlegung in Trainings- und Testdaten abhängt, führt die Reduktion der Daten in den Trainingsdaten – und bei unsymmetrischer Aufteilung erst recht in den Testdaten – zu einer höheren Modellvarianz. Diese Verzerrung kann durch Simulationsexperimente quantifiziert werden.

Um dieses Problem zu umgehen, kann die Validierung auf Testdatensätzen als k-fache Kreuzvalidierung (*k-fold cross validation*) durchgeführt werden. Dazu wird die Datenmenge in *k* Teile zerlegt, damit jedes Segment einmal als *holdout sample* dienen kann. In der Praxis wählt man etwa *k = 5* bzw. *k = 10*. Der verbleibende Rest sind dann die Modellierungsdaten, so dass *k* Modelle zu rechnen und zu bewerten sind. Je höher *k*, desto stärker sind die Trainingsdaten und daher die Modelle korreliert. Als Kennzahl, z. B. den „mittleren quadratischen Fehler" (*MSE:= SSE/n*), betrachtet man entweder den Mittelwert der *k MSE* für die Parameterauswahl einer Modellfamilie oder das Minimum der *MSE* zum Vergleich unterschiedlicher Modelle.

Gini-Index

Der *Gini-Index* beruht auf der Idee, die Modellgüte auf der Basis des Ranges der Risiken zu bestimmen. Hierzu werden die Beobachtungen der Zielvariablen $y_i, i = 1, \dots, n$ und zugehörige Prognosewerte $\hat{y}_i$ absteigend nach den Prognosewerten geordnet. Dann nennt man die stückweise lineare Kurve, die die Punkte (a_i, b_i) mit

$$a_0 := 0, a_i = \frac{i}{n}$$

$$b_0 := 0, b_i = \frac{\sum_{k=1}^{i} y_k}{\sum_{k=1}^{n} y_k}$$

verbindet, die *Lorenzkurve* (auch *Gainkurve*) des Modells zur Stichprobe der y_i. Die ideale *Lorenzkurve* ergibt sich, wenn man die Stichprobenwerte absteigend anordnet (ohne die Prognosewerte).

Der *Gini-Index* wird auf Basis der *Lorenzkurve* berechnet.

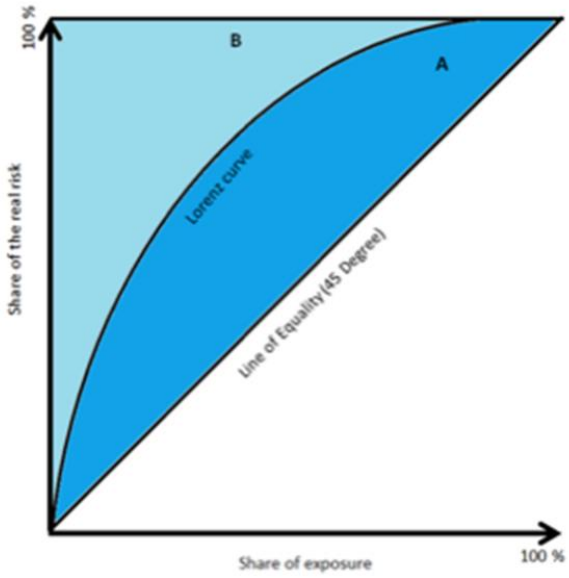


Abbildung 8: Beispielhafte Darstellung Lorenzkurve und Berechnung des Gini-Index

Wenn die Fläche zwischen der Diagonalen und der *Lorenzkurve* gleich *A* ist und die Fläche über der *Lorenzkurve B*, dann ist der *Gini-Index* definiert als *G:=A/(A+B)*. Da *A + B = 0.5* ist, berechnet sich der *Gini-Index* als *G=2A*, oder *G=1–2B*.

Wenn die *Lorenzkurve* durch die Funktion *Y=L(X)* dargestellt wird, so ergibt sich der Wert von *B* mittels Integration als:

$B = 1 - \int_0^1 L\,(X)dX$, bzw. der *Gini-Index* als $G = 2\int_0^1 L\,(X)dX - 1$.

In der Praxis ordnet man die Zeilen des Datensatzes absteigend nach der Modellvorhersage und bestimmt die *Lorenzkurve* als Treppenfunktion z. B. der Centile dieser Ordnung als Anteile des zugehörigen Volumens der Zielvariablen.

Der Datensatz wird wie üblich in eine Trainings- und eine Teststichprobe geteilt. Der *Gini-Index* wird sowohl auf der Trainings- als auch auf der Testmenge berechnet. Die Differenz der beiden Werte ist ein Maß für das sogenannte *Overlearning* des Modells auf der Trainingsmenge.

Die *Gainkurve* und den *Gini-Index* kann man für jedes Modell des überwachten Lernens berechnen. Allerdings kann der absolute Wert des Index stark schwanken, je nachdem welche Risikostruktur modelliert wird. Der Vergleich verschiedener Modellansätze zu einem Datensatz mit einer Targetvariable kann durch folgende Normierung verallgemeinert werden:

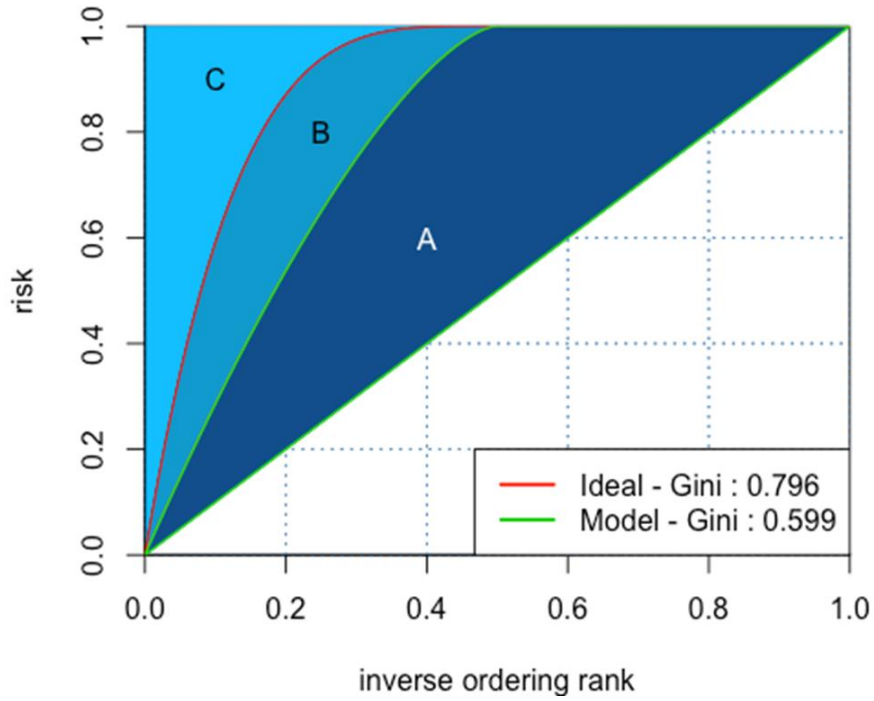


Abbildung 9: Beispielhafte Darstellung ideale Lorenzkurve inkl. Gini-Indizes

Nach Sortierung der Zielvariablen absteigend nach der Höhe definiert man drei Flächen im obigen Graph:

A: die Fläche zwischen der Modellkurve und der Diagonale,

B: die Fläche zwischen dem idealen Modell und dem aktuellen Modell,

C: die Fläche zwischen dem idealen Modell und der oberen Begrenzung.

Um so den *Gini-Index* zu normalisieren, betrachtet man $\tilde{G}$ = *G_M/G_I*, das Verhältnis von idealem zum modellierten *Gini-Index*. Daraus folgt direkt: $\tilde{G}$ = A/(A+B). $\tilde{G}$ wird als *normalisierter Gini-Index* bezeichnet. Dieser Index liegt zwischen *0* und *1* und ist umso höher, je besser das Modell die Daten beschreibt. Wie oben beschrieben wird auch diese Kennzahl auf Modell- und Trainingsdaten berechnet, um zu prüfen, inwiefern das Modell overfitted wurde.

Bemerkung:

Der *Gini-Index* bewertet nur die Ordnung der Modellierung, nicht die Werte an sich, ein aus einem erwartungstreuen Model durch Skalierung mit einem Faktor *F > 1* gewonnenes Modell $M_F := M * F$, welches die Zielvariable systematisch überschätzt, hätte denselben *Gini-Index* wie *M*.

Baumverfahren: Node impurity

Bei Baumverfahren werden Splits vorgenommen, die zu einer größeren Homogenität in den entstehenden Knoten führen. Als Maß der Homogenität bzw. Inhomogenität wird die *node impurity* verwendet. Das ist im Falle stetiger Zielvariablen die mittlere quadratische Abweichung vom Mittelwert des Knotens, bei kategorialen Daten werden der *Gini-Index*, die *Entropie* oder der *Missklassifikationsfehler* verwendet.

Literatur:

[10] Hastie, Tibshirani, Friedman (2009): The Elements of Statistical Learning, Abschnitte 9.2.2 und 9.2.3

Random Forests: Variablenwichtigkeit

Neben der Prognose kann auch ein Verfahren zur Bestimmung der Variablenwichtigkeit durchgeführt werden. Nach der Erstellung eines Baums T_b werden die Prognosen auf dem nicht verwendeten Teil der Stichprobe (*out of the bag – OOB*) bestimmt und die Prognosegüte abgespeichert (z. B. Summe der Fehlerquadrate). Dann werden die Ausprägungen des *j*-ten Merkmals in der *OOB*-Stichprobe zufällig permutiert und anschließend erneut die Prognosegüte von T_b berechnet. Die Verschlechterung der Prognosegüte wird dann über alle Bäume des Random Forests gemittelt und ergibt die Wichtigkeit des Merkmals *j*. Je größer die Verschlechterung ausfällt, umso größer die Wichtigkeit.

Literatur:

[10] Hastie, Tibshirani, Friedman (2009): The Elements of Statistical Learning, Abschnitt 15.3.2

5.3.2. Gütemaße mit expliziter Verteilungsannahme

Bei parametrischen statistischen Modellen gibt es verschiedene Kennzahlen zur Beurteilung der Güte eines Modells, welche durch die parametrische Modellbeschreibung ermittelt werden können. Der Vorteil dieser Gütemaße liegt darin, dass spezielle Eigenschaften der Modelle verwendet werden, wenn die Modelle die Daten gut beschreiben.

Wir beschränken uns hier auf die GLMs. Eine GLM-Familie ist durch die zugehörige *Varianzfunktion V* und eine *Linkfunktion g* bestimmt. Ein in dieser Familie enthaltenes statistisches *Modell M* ist durch folgende Angaben festgelegt: Ein bekanntes *Gewicht w*, ein bekannter *Vektor x* von bestimmten erklärenden Variablen, ein globaler *Dispersionsparameter* ϕ (in manchen Familien fix) und ein *Parameter-Vektor* β bezüglich *x*. Eine nach diesem *Modell M* verteilte *Zufallsvariable Y* erfüllt die Verteilungsaussagen

$$\mu = E(Y) = g^{-1}(x^T\beta) \quad und \quad Var(Y) = \frac{\phi}{w} \cdot V(\mu)$$

Durch die Angaben von Erwartungswert und Varianz ist die komplette Verteilung der Zufallsvariablen *Y* innerhalb der betrachteten GLM-Familie eindeutig bestimmt (durch Spezifikation der Dichte bzgl. eines gemeinsamen Maßes), oft arbeitet man aber nur mit den obige Formeln für Erwartungswert und Varianz, ohne die genaue Verteilung zu spezifizieren (*Quasilikelihood-Ansatz*).

Der bekannteste und älteste Spezialfall der GLM*s* sind die *Linearen Modelle (LM)*. Diese Familie ist durch die konstante Varianzfunktion $V = 1$ und die identische Linkfunktion $g = id$ charakterisiert. Die Verteilungen dieser Familie sind Normalverteilungen. Der Erwartungswert *E(Y)* hängt dann linear von den Parametern β ab.

In der Schaden-/Unfallversicherung spielen die Linearen Modelle keine Rolle. Bei der Erstellung von Risikomodellen im Rahmen der Tarifierung wird meistens als Linkfunktion der natürliche Logarithmus $g = ln$ verwendet, um eine multiplikative Abhängigkeit von den erklärenden Merkmalen zu erhalten. Anstelle der Normalverteilungen werden z. B. die Poissonverteilungen (zur Modellierung der Schadenhäufigkeit), die Gamma-Verteilungen (für den Schadendurchschnitt) oder die Tweedie-Verteilungen (für den Schadenbedarf) verwendet. Diese Verteilungen sind alle GLM*s* und durch die Potenz-Varianzfunktionen $V(\mu) = \mu^p$ mit $p = 1$ für Poisson, $p = 2$ für Gamma und *p* zwischen 1 und 2 für Tweedie charakterisiert.

Die standardmäßige Grundannahme in der GLM-Modellierung ist, dass Trainingsdaten aus *N* Realisierungen y_i von unabhängigen Zufallsvariablen Y_i vorliegen, welche von einer bekannten GLM-Familie mit *Varianzfunktion V* und *Linkfunktion g* erzeugt wurden. Die Zufallsvariablen Y_i hängen noch von bekannten *Gewichten* w_i, bekannten erklärenden Variablen x_i und unbekannten gemeinsamen Parametern ϕ und β ab, welche anhand der vorliegenden Realisierungen y_i geschätzt werden. Das so beschriebene Gesamtmodell *M* hängt von den verwendeten erklärenden Variablen ab. In der Regel liegen viele verschiedene erklärende Variablen vor, von denen in einem Modell *M* einige zur Beschreibung der Daten ausgewählt werden.

In theoretischen Überlegungen wird teilweise abweichend (aber realitätsnah) angenommen, dass die Daten y_i von *wahren Verteilungen* Z_i erzeugt wurden, welche nicht notwendigerweise zur vorgegebenen GLM-Familie gehören, aber weiterhin von den bekannten Größen w_i und von der Gesamtheit x_i aller erklärenden Variablen abhängen. In diesem Fall spricht man von „Missspezifikation". Die Schätzung der Parameter eines Modells *M* kann dann als „beste Approximation" der wahren Verteilungen innerhalb der betrachteten GLM-Familie interpretiert werden.

Die Parameter ϕ und β werden separat geschätzt, β mit der Maximum-Likelihood(ML)-Methode unter Fixierung von ϕ. Für die zur Schätzung von β verwendete Loglikelihood-Funktion der Variablen Y_i verwendet man oft die Bezeichnung $l_i(\mu_i; y_i)$, da der Erwartungswert μ_i über die Beziehung

$$\mu_i = E(Y_i) = g^{-1}(x_i^T \beta)$$

mit β zusammenhängt. Die Loglikelihood-Funktion des Gesamtmodells *M* ergibt sich aufgrund der Unabhängigkeitsvoraussetzung dann als Summe der Einzelbeiträge:

$$l(\mu; y) = \sum_i l(\mu_i; y_i)$$

Jede einzelne Loglikelihood-Funktion $l_i(\mu_i; y_i)$ nimmt ihr Maximum im (eventuell unerreichbaren) Fall $\mu_i = y_i$ an. Das zugehörige (eventuell fiktive) Gesamtmodell M_s wird dann „saturiertes" Modell genannt, es erklärt die Daten vollkommen.

Nach Einsetzen des ML-Schätzers $\hat{\beta}$ bzw. $\hat{\mu}$ in die Loglikelihood-Funktion des Modells *M* erhält man deren Maximalwert

$$l_{max}(M) = l(\hat{\mu}; y).$$

Zur Beurteilung der Modellgüte eines Modells *M* betrachtet man die hieraus abgeleitete *Devianz D(M)* und *skalierte Devianz D*(M)* des *Modells M*:

$$D^*(M) = 2 \cdot \big(l(y; y) - l(\hat{\mu}; y)\big) = 2 \cdot \big(l_{max}(M_s) - l_{max}(M)\big) \; und \;\; D(M) = \phi \cdot D^*(M)$$

Da der Parameter ϕ bei der Schätzung von β festgehalten wird, ist der ML-Schätzer $\hat{\beta}$ genau der Schätzer mit minimaler Devianz. Die *Gesamt-Devianz D(M)* des Modells lässt sich als Summe von Einzeldevianzen $d_i(\hat{\mu}_i; y_i)$ darstellen. Im Linearen Modell ergibt sich der bekannte quadratische Abstand zwischen Daten und Schätzung:

$$d_i(\hat{\mu}_i; y_i) = (\hat{\mu}_i - y_i)^2$$

Die Modellauswahl in GLM*s* erfolgt klassisch mit Hilfe der Devianz. Dabei können zwei Modelle M_1 und M_2 der Familie aber nur dann verglichen werden, wenn sie ineinander „geschachtelt" sind, d. h. ein Modell „gröber" als das andere ist. Man nennt das Modell M_2 eine Vergröberung des Modells M_1, wenn die in M_2 verwendeten erklärenden Variablen eine Teilmenge der in M_1 verwendeten erklärenden Variablen sind. Die Anzahl p_1 der Parameter von M_1 ist dann größer als die Anzahl p_2 der Parameter von M_2, und es gilt $D^*(M_2) \geq D^*(M_1)$. Unter Voraussetzung der Gültigkeit des

Modells folgt dann die Differenz $D^*(M_2) - D^*(M_1)$ der Devianzen einer χ^2-Verteilung mit $p_1 - p_2$ Freiheitsgraden.

Bei der Modellauswahl kann es zum sogenannten *Overfitting* kommen, indem das Modell sehr nah an den Trainingsdaten ist und damit auf unabhängigen neuen Daten eine schlechte Vorhersagequalität aufweist. Im oben erwähnten Extremfall des saturierten Modells M_s werden die Daten vom Modell exakt repliziert. Dieser Extremfall kann insbesondere eintreten, wenn es genauso viele Parameter wie Daten gibt. In diesem Extremfall nimmt die Devianz das absolute Minimum 0 an.

Um eine gute Vorhersagequalität auf unabhängigen Daten zu bekommen, muss daher die Flexibilität des Modells beschränkt werden. Dies wird z. B. durch das *Lasso-Verfahren* erreicht.

Ein alternativer Ansatz mit gleicher Zielrichtung ist die Verwendung von sogenannten „Informationskriterien", die in der Modellbeschreibung verwendete Information minimieren wollen. Die bekanntesten Informationskriterien sind das *Akaike-Informationskriterium (AIC)* und das *Bayessche Informationskriterium (BIC)*. Diese sind für ein Modell *M* anhand des Maximums $l_{max}(M)$ der Loglikelihood-Funktion und der *Parameteranzahl p(M)* sowie (beim *BIC*) der *Anzahl N der Beobachtungen* definiert:

$$AIC(M) := -2 \cdot l_{max}(M) + 2 \cdot p(M)$$
$$BIC(M) := -2 \cdot l_{max}(M) + p(M) \cdot \ln(N)$$

Ein weiterer Vorteil gegenüber der Devianz ist, dass mit diesen Kriterien beliebige Modelle und nicht nur geschachtelte Modelle verglichen werden können.

Im Spezialfall der Linearen Modelle erhält man für das AIC mit der hier üblichen Bezeichnung σ^2 für den Dispersionsparameter ϕ den Ausdruck

$$AIC(M) = \frac{1}{\sigma^2} \cdot \sum_i (\hat{\mu}_i - y_i)^2 + 2 \cdot p(M) + const.$$

Im Fall eines bekannten Dispersionsparameters stimmt damit das *AIC* bis auf eine additive Konstante mit der Kennzahl *Mallows*' C_p überein:

$$C_p(M) = \frac{1}{\sigma^2} \cdot \sum_i (\hat{\mu}_i - y_i)^2 + 2 \cdot p(M) + N$$

Mallows' C_p wird aber oft auch in nichtparametrischen Situationen verwendet.

Es gibt verschiedene Herleitungen dieser Informationskriterien. Die übliche Ableitung des *AIC* verwendet die sogenannte *Kullback-Leibler-Distanz* zwischen der wahren und der modellierten Verteilung. Hierbei wird vorausgesetzt, dass die wahre Verteilung zur Modellfamilie gehört. Eine analoge Ableitung ohne diese Annahme führt zum *Takeuchi-Informationskriterium (TIC)*, dessen Formel aber nicht so einfach darstellbar ist und im GLM-Kontext von der Verteilungsfamilie abhängt. Eine alternative Ableitung des *AIC* stammt von Stone über die Kreuzvalidierung. Das *BIC* wird meist über einen *Bayes-Ansatz* abgeleitet.

Weiterhin wird in der Ableitung des *AIC* und des *TIC* nur eine asymptotische Aussage getroffen. Bei kleinen Datenmenge wird daher oft eine korrigierte Version des *AIC* verwendet:

$$cAIC(M) = -2 \cdot l_{max}(M) + 2 \cdot p(M) + 2 \cdot \frac{p(M) \cdot (p(M)+1)}{N - p(M) - 1}$$

Es gibt diverse Vergleichsstudien zur Verwendung dieser Informationskriterien in der Modellauswahl, aber keinen eindeutigen Sieger.

Literatur:

[24] Claeskens, Hjort (2010): Model Selection and Model Averaging

[25] Efron, Hastie (2016): Computer Age Statistical Inference

[26] Bühlmann, van der Geer (2011), Statistics for High-Dimensional Data

[27] Lv, Liu (2016): Model selection principles in misspecified models

5.4. Statistische Tests

5.4.1. Log-Likelihood-Test

Beschränkt man sich auf GLM, steht zur Modell- und Variablenbeurteilung die klassische statistische Testtheorie zur Verfügung. Hierbei wird die bereits im vorangehenden Abschnitt dargestellte Asymptotik (*Wilks Theorem*) ausgenutzt, die besagt, dass unter der Nullhypothese (hier: die beiden geschachtelten Modelle unterscheiden sich nicht) die Log-Likelihood-Ratio-Teststatistik beider Modelle asymptotisch einer χ^2-Verteilung mit $p_1 - p_2$ Freiheitsgraden folgt. Etwas anders als im vorangehenden Abschnitt wird hier jedoch das *Intercept Model* (oder in Versicherungssprache das *Einheitspreismodell*) gegen ein komplexeres Modell getestet. Ist der assoziierte p-Wert unter einer gewünschten Schranke, wird die Nullhypothese abgelehnt und das komplexe Modell bevorzugt. In der Praxis ist die Beurteilung von Modellen mit Hilfe der χ^2-Verteilungsasymptotik und der Einsatz derselben für schrittweise Regression und statistischem Lernen nicht unumstritten. Es gibt keine Bestrafung für den Einsatz von komplexen Modellen. Pinheiro und Bates zeigten in einer Arbeit aus 2000, dass die Log-Likelihood-Ratio-Teststatistik in simulierten Modellen keiner χ^2-Verteilung genügt, was fallbezogen auf eine schwache Asymptotik hinweisen könnte. In der Modellierungspraxis entstehen oft „merkwürdige" Modelle: Variablen mir sehr viel Merkmalsausprägungen werden vom Test bevorzugt – die zusätzliche Komplexität wird nur wenig bestraft. Auch gilt die Log-Likelihood-Teststatistik nur für geschachtelte Modelle, was in der Praxis zu sehr einschränkend ist.

5.4.2. Vuong-Test für nicht geschachtelte Modelle

Der *Vuong-Test* (Vuong 1989) für nicht geschachtelte Modelle (im Gegensatz zu geschachtelten bzw. überlappenden Modellen) wird das Modell bevorzugen, welches die *Kullback-Leibler-Distanz* gegen ein „perfektes", aber unbekanntes Modell minimiert. Hierbei werden Resultate aus der statistischen Informationstheorie benutzt: das „perfekte" Modell beschreibt ohne Informationsverlust – gemessen als Entropie – die Daten (beispielsweise könnte das die wahre aber unbekannte Verteilung sein).

Vuong formulierte die Kullback-Leibler Distanz wie folgt:

$$KLIC := E_0[ln\, h_0(Y_i \mid X_i)] - E_0[ln\, f\,(Y_i\,;\, ß_i\,)]\,.$$

Hierbei ist h_0 die „perfekte" bedingte Dichte von Y_i mit gegebenem X_i.

Da das Modell zu bevorzugen ist, das die Distanz zum Erwartungswert der Entropie des unbekannten Modells $E_0[ln\, h_0(Y_i \mid X_i)]$ minimiert, muss der Term $E_0[ln\, f\,(Y_i\,;\, ß_i\,)]$ maximiert werden. Das Quellencodierungstheorem stellt sicher, dass $KLIC$ nicht negativ ist und es formuliert, dass das unbekannte „perfekte" Modell nicht weiter vereinfacht werden kann, ohne Informationen zu verlieren. Daher ist auch jedes Modell, das im Sinne der $KLIC$ Distanz näher am „perfekten" Modell liegt, zu bevorzugen: es ist ein Modell gefunden, das fast so gut wie das „perfekte Modell" die Daten beschreibt.

Wichtig ist, dass wir das „perfekte" Modell nicht kennen und auch nicht ausrechnen müssen. Die Formulierung des „perfekten" Modells dient allein zur Begründung des folgenden Optimalitätskriteriums: bei rivalisierenden Modellen ist dasjenige zu bevorzugen, bei dem die individuellen Log-Likelihoods signifikant größer sind.

Bei dieser Herangehensweise wird die Verwendung des Quotienten der beiden Log-Likelihoods als Bewertungsmaßstab für Modelle nicht statistisch, sondern mit Hilfe der Informationstheorie begründet.

Vereinfacht ausgedrückt formuliert der *Vuong-Test*, dass bei nicht geschachtelten Modellen unter der Nullhypothese die standardisierte Log-Likelihood-Ratio-Teststatistik approximativ standardnormalverteilt ist. (Nullhypothese: Beide Modelle erklären „gleich gut", das heißt, der Erwartungswert des Quotienten der Log-Likelihoods beider Modelle ist gleich null.)

Dass der Test nur für nicht geschachtelte Modelle gilt, beschränkt die Anwendung gerade in der Tarifentwicklung, in der man in der Regel ausgehend von einem Modellkern alternative zusätzliche Variablen bewerten möchte. In der gleichen Arbeit zeigt Vuong aber auch, dass unter gewissen Voraussetzungen der Test auch für Modelle mit geschachtelten bzw. sich überlappenden Variablen anwendbar ist. Die Log-Likelihood-Ratio-Teststatistik folgt dann asymptotisch einer gewichteten χ^2-Verteilung.

Beim *Vuong-Test* motiviert die *Kullback-Leibler-Distanz* einen klassischen statistischen Test, während diese Distanz neben der Entropie als Ergebnisse der Informationstheorie insbesondere auch zur Herleitung von Gütemaßen im Zusammenhang mit Statistical-Learning-Verfahren benutzt werden (siehe *5.3 Gütemaße*).

Literatur:

[28] Vuong (1989): Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses
(Oder auch: https://en.wikipedia.org/wiki/Vuong%27s_closeness_test)

[29] Pinheiro, Bates (2000), Mixed-Effects Models in S and S-PLUS

5.4.3. Distribution free test

Clarke stellt in einer Arbeit von 2003 eine Modifikation des *Vuong-Tests* dar. Es handelt sich um einen modifizierten nicht-parametrischen Vorzeichentest, angewandt auf den Quotienten der beiden Log-Likelihoods. Die Formulierung der Nullhypothese („beide Modelle sind gleich gut") bezieht sich auf den Median und nicht auf den Erwartungswert des Quotienten der Likelihoods: die Wahrscheinlichkeit, dass der Quotient der beiden Log-Likelihoods positiv ist, beträgt 0,5. Die Nullhypothese wird verworfen, wenn die Anzahl der positiven einzelnen Quotienten der Log-Likelihoods in einem kritischen Bereich liegt.

Literatur:

[30] Clarke (2007): A Simple Distribution-Free Test for Nonnested Model Selection

5.5. Bayes-Faktoren

Für zwei Modelle M_1 und M_2 soll anhand des Datensatzes D entschieden werden, welches die in den Daten realisierte Wirklichkeit am besten beschreibt. Dafür werden im Bayesianischen Ansatz die A-posteriori-Modellwahrscheinlichkeiten $P(M_2|D)$ und $P(M_1|D)$ verglichen. Mit dem Satz von Bayes und den A-priori-Wahrscheinlichkeiten der beiden Modelle $P(M_1)$ und $P(M_2)$ gilt dann:

$$\frac{P(M_2|D)}{P(M_1|D)} = \frac{P(D|M_2)}{P(D|M_1)} \frac{P(M_2)}{P(M_1)}.$$

Den Quotienten der Wahrscheinlichkeiten $P(D|M_2)$ und $P(D|M_1)$ bezeichnet man als *Bayes-Faktor*:

$$B^{2,1} := \frac{P(D|M_2)}{P(D|M_1)}.$$

Sind keine A-priori-Wahrscheinlichkeiten bekannt, werden die zu vergleichenden Modelle oftmals als gleich wahrscheinlich angenommen. Damit genügt der *Bayes-Faktor* zur relativen Beurteilung der beiden Modelle.

Als Faustregel für den Modellvergleich gilt: Ist der *Bayes-Faktor* größer als *exp(5)* ≈ *148*, so ist die Evidenz für Modell M_2 gegenüber Modell M_1 sehr stark. Folglich ist M_2 gegeben die Daten entschieden zu präferieren.

Für die Berechnung eines *Bayes-Faktors* betrachten wir zunächst die marginale Likelihood-Funktion nur eines Modells M:

$$P(D|M) = \int_{-\infty}^{\infty} P\,(D|\beta, M)P(\beta|M)d\beta.$$

Dabei bezeichnet $P(D|\beta, M)$ die Likelihood-Funktion von Modell Mmit dem Parametervektor β, der der A-priori-Dichtefunktion $P(\beta|M)$ unterliegt. Dieser Ausdruck kann mittels der Laplace-Methode approximiert werden. Die Anwendung der Laplace-Integralapproximation erfordert, die Likelihood-Funktion des Modells am bayesianischen *MAP-Schätzer*, dem Modus der A-posteriori-Verteilung, zu evaluieren. In der Bayesianischen Statistik liegt dem *MAP-Schätzer* die (grobe) Indikatorfunktion als Verlustfunktion zugrunde. Wird zusätzlich eine nicht informative a-priori-Verteilung angenommen, so kann der *MAP-Schätzer* durch den ML-Schätzer ersetzt werden.

Unter weiteren Annahmen führt die Rechnung näherungsweise zu:

$$2\ln P(D|M) \approx 2l_{max}(\beta) - S,$$

wobei $l_{max}(\beta)$ die Log-Likelihood evaluiert am ML-Schätzer und S ein Strafterm ist. Der Term S ist dabei nicht einfach strukturiert und die Ermittlung kann in der Praxis sehr aufwändig sein.

Für den *Bayes-Faktor* der beiden Modelle M_1 und M_2 ergibt sich damit:

$$2\ln B^{2,1} \approx 2\left(l_{max,2}(\beta_2) - l_{max,1}(\beta_1)\right) - S_2 + S_1.$$

Obwohl diese Darstellung die Interpretation als „korrigiertes Likelihood-Verhältnis" nahelegt, ist die Natur dieses Ansatzes der Vergleich von a posteriori Wahrscheinlichkeiten. Im Gegensatz zum Likelihood-Ratio-Test ist der Vergleich der Modelle auch nicht auf genestete Modelle beschränkt. Die Ermittlung ist aber oft mit hohem Aufwand verbunden, weswegen in der Regel Näherungen zum Einsatz kommen.

Mit zusätzlichen Annahmen lassen sich die obigen Ausdrücke asymptotisch weiter vereinfachen und man erhält zum Beispiel das bekannte *Baysianische Informationskriterium* BIC:

$$-2\ln P(D|M) \approx -2l_{max}\left(\hat{\beta}\right) + p\ln n =: BIC.$$

Dabei ist $\hat{\beta}$ der geschätzte Parametervektor, p die Anzahl der genutzten Parameter und n die Größe des verwendeten Datensatzes. Die letzte Approximation zeigt, wie das *BIC* über einen Bayes-Ansatz motiviert werden kann. Allerdings spiegeln sich im *BIC* weder die A-priori-Verteilung noch die zugrundeliegende Verlustfunktion des Bayesianischen Modells wider.

Die Differenz zweier BICs ist ein approximierter, logarithmierter *Bayes-Faktor*:

$$BIC_1 - BIC_2 = 2\left(l_{max,2}(\beta_2) - l_{max,1}(\beta_1)\right) - (p_2 - p_1)\ln n \approx 2\ln B^{2,1}$$

Eine weitere Anwendung findet sich im *R-package Bayes Model Averaging (BMA)*:

Die A-posteriori-Verteilung einer interessierenden Größe *Q* ist das gewichtete Mittel der A-posteriori-Verteilungen von *Q* gegeben die Modelle, gewichtet mit den A-posteriori-Modellwahrscheinlichkeiten:

$$P(Q|D) = \sum_k P\,(Q|D, M_k)P(M_k|D)$$

Hierbei können die a posteriori Modellwahrscheinlichkeiten unter Nutzung des *BIC* approximiert werden:

$$P(M_k|D) = \left.exp\left(-\frac{1}{2}\,BIC_k\right)P(M_k)\right/\sum_l exp\left(-\frac{1}{2}\,BIC_l\right)P(M_l)$$

R-package:

`Bayesian Model Averaging (BMA)`
https://cran.r-project.org/web/packages/BMA/index.html

Literatur:

[31] Raftery (1993): Approximate Bayes Factors and Accounting for Model Uncertainty in Generalized Linear Models

[32] Anderson, Burnham (2004): Multimodel Inference - Understanding AIC and BIC in Model Selection

5.6. Modellgüte bei Klassifikationen

Im Gegensatz zur dem Aktuar bekannten Regression stellen Klassifikationsmethoden eine echte Erweiterung und damit Bereicherung des methodischen Werkzeugkastens dar – dies bedingt indes allerdings auch die Feststellung, dass geläufige

quantitative Gütemaße für Fittings sich für Klassifikationen, insbesondere Binärunterscheidungen, in der Regel nicht eignen. Zuvorderst deswegen, weil es keine Verteilungsannahme gibt, die einen kanonischen Schätzervergleich erlauben würde.

Die in diesem Abschnitt vorgestellten Gütemaße sind daher im Wesentlichen aus der Bayes'schen Statistik abzuleiten, wobei die Nullhypothese verteilungsfrei ausfallen muss, also konzeptionell mit einer zufälligen Klassifikation verglichen wird. Während sich die jeweils aus der Kontingenztafel abgeleiteten Gütemaße *Confusion Matrix* und *Kappa-Koeffizient* auch für Mehrklassenprobleme eignen, ist die Nutzbarkeit von beispielsweise *ROC* auf Binärklassifikation beschränkt. Zur Bewertung des Modells bzw. seiner Prädiktionsgüte verhält es sich in der Klassifikation wie bei anderen Methoden auch stets so, dass nur eine Betrachtung mehrerer Gütemaße eine sinnvolle Einschätzung erlaubt.

5.6.1. Confusion Matrix, Fehlerrate, Sensitivität, Spezifität

Die *Confusion Matrix* ist die einfachste Form der Kontingenztafel für die binäre Klassifikation. Sie ist ein tabellarischer Vergleich der Referenzdaten mit den korrespondierenden Ergebnissen der Klassifizierung zur quantitativen Performancebewertung von Klassifikatoren. Bei der binären Klassifikation mit den Klassen positiv (+) und negativ (−) werden die Anzahlen der Vorhersagen in den Untermengen richtig positiv (*TP = true positive*), richtig negativ (*TN = true negative*), falsch positiv (*FP*) und falsch negativ (*FN*) bestimmt. Dabei fällt jede Vorhersage nach Anwendung der Klassifikation in genau eine Kategorie. Zusätzlich wird die Anzahl *P* und *N* positiver Vorhersagen und negativer Vorhersagen ermittelt, welche die Berechnung abgeleiteter Benchmarks erlaubt.

		beobachtete Klasse	
		+	−
vorhergesagte Klasse	+	TP	FP
	−	FN	TN
Summe		P	N

Abbildung 10: Schema der Confusion Matrix

Neben der Betrachtung der reinen Fehlerrate *ER = (FP + FN) / (N + P)*, welche quantifiziert, wie groß die relative Korrektheit der Vorhersage ist, werden zur Messung der absoluten Korrektheit die Maße *Sensitivität S = TP / P = TP / (TP + FN) = 1 − Typ 2* und *Spezifität SPC = TN / N = TN / (TN + FP) = 1 − Typ 1* verwendet.

Darüber hinaus wird oftmals der *F1-Score F1 = 2TP / (2TP + FP + FN)* herangezogen, um die relevanten (richtig positiv klassifizierten) von den irrelevanten (richtig negativ klassifizierten) Vorhersagen abzugrenzen.

Literatur:

[34] Makhoul, Kubala, Schwartz, Weischedel (1999): Performance measures for information extraction

[35] Baeza-Yates, Ribeiro-Neto (1999): Modern Information Retrieval

5.6.2. ROC-Kurve und AUC

Die *ROC-Kurve* (englisch: *receiver operating characteristic*) ist ein Standardtool zur Visualisierung der Performance bei der binären Klassifizierung: der Anteil der *TP* an den *P* (*true positive rate TPR*) wird gegen den Anteil der *FP* an den *N* (*false positiv rate FPR*) geplottet.

Die Kurve zeigt also gleichzeitig die zwei verschiedenen Fehler für alle möglichen Schwellenwerte, die zur Entscheidung genutzt werden: *(x,y) = (FPR , TPR) = (1 – Spezifität , Sensitivität) = (Typ 1, 1 – Typ 2)*. Eine ideale *ROC-Kurve* verläuft nahe der oberen linken Ecke *(1, 0)*.

Die *Abbildung 11 z*eigt die *ROC-Kurve* von einem logit-Modell für eine Trainings- und Teststichprobe.

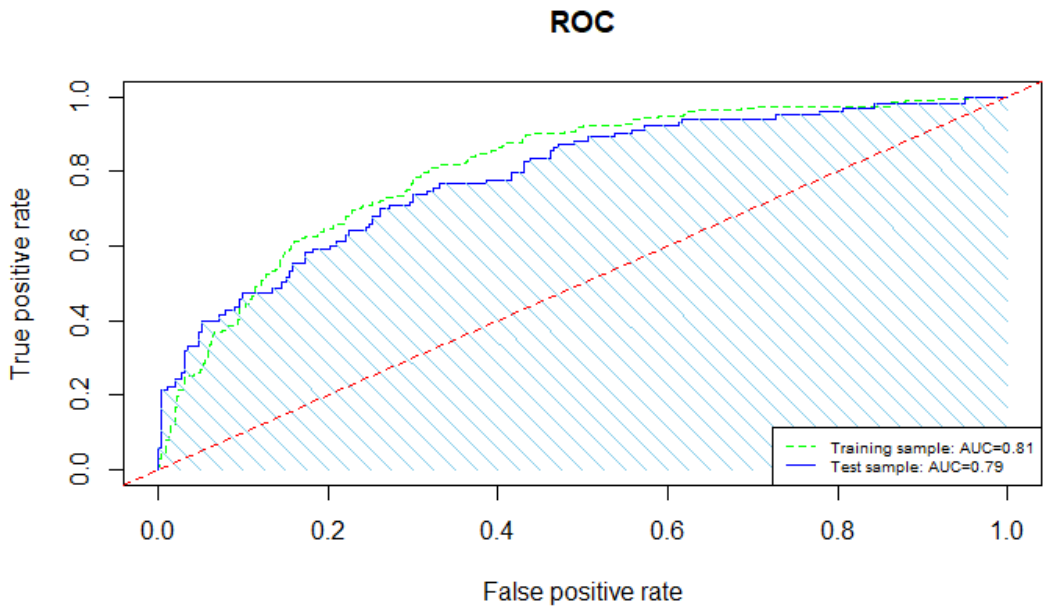


Abbildung 11: Beispielhafte Darstellung ROC-Kurve und AUCs für Trainings- und Testdaten

Als Maß für die *Overall Performance* der Klassifikation wird die Fläche unter der *ROC-Kurve* herangezogen – *area under the curve (AUC*). Je größer der *AUC*, umso besser das Modell.

Für den Modellvergleich verschiedener Klassifikatoren eignen sich *ROC* und *AUC* gut. Anhand der *ROC-Kurve* kann auch ein optimaler Trennwert (*cut point*) ermittelt werden. Allerdings ist zu beachten, dass der *AUC* nur eine Aussage über die korrekte Reihenfolge liefert – er ist nicht geeignet, die prädiktiven Werte zu beurteilen.

R-packages:

```
library(ROCR)
library(pROC)
```

Literatur:

[12] James et al. (2014): An Introduction to Statistical Learning, Kapitel 4.4.3.

5.6.3. Konfidenzintervall der Fehlerrate (Clopper-Pearson)

Als Gütemaß kann neben der *Fehlerrate ER = (FP + FN) / (N + P)* ebenso die *Accuracy ACC = 1 – ER* verwendet werden. Hierbei geht man davon aus, dass die Wahrscheinlichkeit einer korrekten Klassifikation bei allen zu betrachtenden Fällen konstant ist. Insofern unterstellen wir eine Bernoulli-Verteilung *B(1, p)* mit Erfolgswahrscheinlichkeit *p* für jede Beobachtung. Zur Quantifizierung der Unschärfe des Punktschätzers $\widehat{p} = \frac{TP+TN}{N+P}$ wird das Konfidenzintervall nach *Clopper-Pearson* mit geeigneten Quantilen der *F-Verteilung* zum Konfidenzniveau α konstruiert:

$$P\left(\widehat{p_u} \leq p \leq \widehat{p_o}\right) \geq 1 - \alpha$$

mit

$$\widehat{p_u} = \frac{x}{x + (n - x + 1)\, F_u} \; mit \; F_u = F_{\left(2(n-x+1),2x,1-\frac{\alpha}{2}\right)}, x \coloneqq TP + TN \; und \; n \coloneqq N + P,$$

$$\widehat{p_o} = \frac{(x+1)\, F_o}{n - x + (x+1)\, F_o} \; mit \; F_o = F_{\left(2x+2,2n-2x,1-\frac{\alpha}{2}\right)}.$$

Als zusätzliches Gütemaß wird die *Balanced Accuracy* $BACC = \frac{1}{2}\frac{TP}{FN+TP} + \frac{1}{2}\frac{TN}{TN+FP}$ definiert. Die *BACC* wird verwendet, wenn die Daten sehr ungleich verteilt sind, also eine Ausprägung deutlich überwiegt. Dadurch ergibt sich ein Bias in der *Accuracy*, welcher die Güte des Klassifikationsmodells als zu optimistisch einstuft.

Literatur:

[36] Clopper, Pearson (1934): The use of confidence or fiducial limits illustrated in the case of the binomial

[37] von Collani, Dräger (2001): Binomial Distribution Handbook for Scientists and Engineers

[38] Brodersen, Ong, Stephan, Buhmann (2010): The balanced accuracy and its posterior distribution

5.6.4. Kappa-Koeffizienten

Der *Kappa-Koeffizient* ist ein statistisches Maß zur Ermittlung der zufallskorrigierten Übereinstimmung mehrerer (unabhängiger) Beurteiler, die eine bestimmte Anzahl von Objekten mit Hilfe eines vorgegebenen Kategoriensystems klassifizieren. *Kappa* gibt das Verhältnis zwischen dem Anteil von Übereinstimmungen und dem maximalen Anteil möglicher Übereinstimmungen an.

Das *Cohens Kappa* ist eine gängige Statistik zum Messen der Übereinstimmung der Einstufungen zwischen zwei Beurteilern:

$\kappa = \frac{p_0 - p_c}{1 - p_c}$,

mit der gemessenen relativen Übereinstimmung p_0 der beiden Beurteiler und der zufällig erwarteten relativen Übereinstimmung p_c. Das *Fleiss-Kappa* ist eine Verallgemeinerung des *Cohens Kappa* für mehrere Beurteiler.

Kappa kann Werte im Bereich von *-1* bis *+1* annehmen. Je höher der *Kappa*-Wert ist, desto höher ist die Übereinstimmung:

- Bei *Kappa < 0* ist die Übereinstimmung geringer als die erwartete zufällige Übereinstimmung.
- Bei *Kappa = 0* entspricht die Übereinstimmung der erwarteten zufälligen Übereinstimmung.
- Bei *Kappa = 1* liegt eine vollkommene Übereinstimmung vor.

Das *Kappa* berücksichtigt lediglich die Beobachtungen, bei denen eine vollständige Übereinstimmung vorliegt, also die Hauptdiagonale der Kontingenztafel. Dahingehend gehen bei der Erweiterung, dem *gewichteten Kappa*, alle Zellen der Kontingenztafel in die Berechnung ein, wobei die Hauptdiagonale das höchste Gewicht erhält.

R-package:

```
psych, function: cohen.kappa
```

Literatur:

[39] Cohen (1960): A coefficient of agreement for nominal scales

[40] Cohen (1968): Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit

[41] Fleiss (1971): Measuring nominal scale agreement among many raters

Literaturverzeichnis

[1] Internationale ASTIN-Gruppe (2015): ASTIN Big Data: **Data Analytics Working Party – Phase 1,** April 2015

Original-Speicherort im Internet:
http://www.iaa-astin.org/ → Documents → ASTIN Big Data/Data Analytics Working Party – Phase 1 Paper- April 2015

Im SharePoint-Raum der Arbeitsgruppe:
https://aktuar.de/mein-arbeitsraum/schadenversicherung/ag-tarifierungsmethodik/Documents/ASTIN_Data_Analytics_Final_20150518.pdf

[2] Chen, Zhang. (2014): **Data-intensive applications, challenges, techniques and technologies: A survey on Big Data**. Information Sciences 75, Seiten 314–347.

[3] Dodson (2014): **Big Data, Big Hype?** Wired.
https://www.wired.com/insights/2014/04/big-data-big-hype/

[4] Morawetz (2016): **Der telematische Irrweg der Kfz-Versicherung** [Newsletter der GenRe]

Original-Speicherort im Internet:
http://media.genre.com/documents/kfz1603-de.pdf

Im SharePoint-Raum der Arbeitsgruppe:
https://aktuar.de/mein-arbeitsraum/schadenversicherung/ag-tarifierungsmethodik/Documents/Telematischer%20Irrweg.pdf

[5] Frey, Schönfelder, Wellisch (2016): **Anwendung von maschinellem Lernen in der Tarifierung**, Der Aktuar 02/2016, Seiten 51–54.

[6] Wikipedia (2016): **Big Data**.

Original-Speicherort im Internet:
https://de.wikipedia.org/wiki/Big_Data

[7] Forrester Research Inc. (2016): **Pragmatische Definition**.

Original-Speicherort im Internet:
http://blogs.forrester.com/mike_gualtieri/12-12-05-the_pragmatic_definition_of_big_data

[8] Fachgrundsatz der deutschen Aktuarvereinigung e. V. (2016): **Berufspflichten des Aktuars bei der Tarifgestaltung in der Schadenversicherung**

Original-Speicherort im Internet:
https://aktuar.de/unsere-themen/fachgrundsaetze-oeffentlich/2016-01-19-Hinweis-Tarifierung.pdf

[9] Hastie, Tibshirani (1986): **Generalized Additive Models**, Statistical Science (1), Seiten 297-310.

[10] Hastie, Tibshirani, Friedman (2009**): The Elements of Statistical Learning**, Springer-Verlag.

[11] R-package gbm: **Generalized Boosted Regression Models**.
https://cran.r-project.org/web/packages/gbm/gbm.pdf

[12] James, Witten, Hastie, Tibshirani (2014): **An Introduction to Statistical Learning**, Springer-Verlag.

[13] Breiman, Friedman, Olshen, Stone (1984): **Classification and Regression Trees**, Taylor & Francis Ltd.

[14] Quinlan (1993): **C4.5: Programs for Machine Learning**, Morgan Kaufmann Publishers.

[15] Wu, Kumar, Quinlan, Ghosh, Yang, Motoda, McLachlan, Ng, Liu, Yu, Zhou, Steinbach, Hand, Steinberg (2008): **Top 10 algorithms in data mining**, Knowl. Inf. Syst., 2008 (14), 1-37, DOI 10.1007/s10115-007-0114-2.

[16] Nagarajan, Scutari, Lèbre (2013): **Bayesian Networks in R with Applications in Systems Biology**, Use R!, Vol. 48, Springer (US).
http://www.bnlearn.com/book-useR/

[17] Scutari, Denis (2014): **Bayesian Networks with Examples in R**. Texts in Statistical Science, Chapman & Hall/CRC (US).
http://www.bnlearn.com/book-crc/

[18] Salakhutdinov, Mnih, Hinton (2007): **Restricted Boltzmann machines for collaborative filtering**, Proceedings of the 24th international conference on Machine learning - ICML '07, 2007, Seiten 791–798.

[19] Hinton (2010): **A Practical Guide to Training Restricted Boltzmann Machines**, UTML TR 2010–003, University of Toronto.
http://www.cs.toronto.edu/~hinton/absps/guideTR.pdf

[20] Wikipedia: **Deep Learning**.
https://en.wikipedia.org/wiki/Deep_learning

[21] Kramer (2011): **Dimensionality Reduction by Unsupervised K-Nearest Neighbor Regression**, Soft Computing, 19(6), DOI: 10.1109/ICMLA.2011.55.

[22] Ohlsson, Johansson (2010): **Non-Life Insurance Pricing with Generalized Linear Models**, Springer, Heidelberg Dordrecht London New York.

[23] DAV-Arbeitsgruppe Tarifierungsmethodik (2. Auflage, 2015): **Aktuarielle Methoden der Tarifgestaltung in der Schaden-/Unfallversicherung**, Verlag Versicherungswirtschaft, Karlsruhe.

[24] Claeskens, Hjort (2010): **Model Selection and Model Averaging, Cambridge University Press**, 2010.

[25] Efron, Hastie (2016): **Computer Age Statistical Inference, Cambridge University Press**, 2016.

[26] Bühlmann, van der Geer (2011), **Statistics for High-Dimensional Data**, Springer, 2011.

[27] Lv, Liu (2016): **Model selection principles in misspecified models**, arXiv, 2016.

[28] Vuong (1989): **Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses**, Econometrica, Ausgabe 2/1989, Seiten 307–333.

Oder auch:
https://en.wikipedia.org/wiki/Vuong%27s_closeness_test

[29] Pinheiro, Bates (2000**), Mixed-Effects Models in S and S-PLUS**, Springer-Verlag, ISBN 0-387-98957-9, Seiten 82–93.

[30] Clarke (2007): A **Simple Distribution-Free Test for Nonnested Model Selection**.

http://kevinaclarke.org/SDFT.pdf

[31] Raftery (1993): **Approximate Bayes Factors and Accounting for Model Uncertainty in Generalized Linear Models**, 1993.

http://www.stat.washington.edu/research/reports/1993/tr255.pdf

[32] Anderson, Burnham (2004): **Multimodel Inference - Understanding AIC and BIC in Model Selection**, 2004.

http://sortie-nd.org/lme/Statistical%20Papers/Burnham_and_Anderson_2004_Multimodel_Inference.pdf

[33] R-package BMA: **Bayesian Model Averaging**.

https://cran.r-project.org/web/packages/BMA/index.html

[34] Makhoul, Kubala, Schwartz, Weischedel (1999): **Performance measures for information extraction**, Proceedings of DARPA Broadcast News Workshop, Herndon, VA, Februar 1999.

[35] Baeza-Yates, Ribeiro-Neto (1999): **Modern Information Retrieval, New York: ACM Press**, Addison-Wesley. ISBN 0-201-39829-X, Seiten 75 ff.

[36] Clopper, Pearson (1934): **The use of confidence or fiducial limits illustrated in the case of the binomial**, Biometrika, 26, DOI: 10.2307/2331986, Seiten 404–413.

[37] von Collani, Dräger (2001): **Binomial Distribution Handbook for Scientists and Engineers**, Birkhaeuser Boston.

[38] Brodersen, Ong, Stephan, Buhmann (2010): **The balanced accuracy and its posterior distribution**, International Conference on Pattern Recognition.

http://ong-home.my/papers/brodersen10post-balacc.pdf

[39] Cohen (1960): **A coefficient of agreement for nominal scales, Educational and Psychological Measurement**. 20, 1960, Seiten 37–46.

[40] Cohen (1968): Weighted kappa: **Nominal scale agreement with provision for scaled disagreement or partial credit**, Psychological Bulletin, 1968, Seiten 213–220.

[41] Fleiss (1971): **Measuring nominal scale agreement among many raters**, Psychological Bulletin, Vol. 76, No. 5, 1971, Seiten 378–382.