DAV

DEUTSCHE
AKTUARVEREINIGUNG e.V.

Ergebnisbericht des Ausschusses Lebensversicherung

Big Data in der Lebensversicherung

Köln, den 19. September 2019

Präambel

Die Arbeitsgruppe „Big Data in der Lebensversicherung" des Ausschusses Lebensversicherung der Deutschen Aktuarvereinigung e. V. hat den vorliegenden Ergebnisbericht erstellt.[1]

Zusammenfassung

Der Ergebnisbericht behandelt Fragestellungen zu Methoden aus dem Bereich Big Data, Künstlicher Intelligenz und Machine Learning in der Lebensversicherung und betrifft Aktuare in der Rolle u.a. als verantwortlicher Aktuar, Sachverständiger, oder in versicherungsmathematischer Funktion bei der Ausführung aktuarieller Aufgaben im Rahmen der Produktentwicklung, des aktuariellen Controllings und des Jahresabschlusses eines Lebensversicherers.

Methoden aus dem Bereich Big Data, Künstlicher Intelligenz und Machine Learning stellen den nächsten Schritt in der Entwicklung der für die Versicherungswirtschaft essentiellen Datenanalyse-Verfahren dar. Moderne Verfahren bieten die Möglichkeit, ein weit größeres Volumen an Daten zu verarbeiten, mehr Informationen in die Analysen einfließen zu lassen und auch komplexere Beziehungen zwischen einzelnen Merkmalen abzubilden.

Dieser Bericht erklärt die gängigen Begriffe aus dem Bereich Big Data, Künstlicher Intelligenz und Machine Learning, stellt die aktuellen Entwicklungen dar und beleuchtet Anwendungsfelder in der Lebensversicherungswirtschaft. Zudem werden die Themen Datenquellen, Datengrundlagen und Datenschutz diskutiert sowie auf die zugrunde liegenden statistischen Methoden und gängigen Softwareprogramme eingegangen.

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 Lebensversicherung am 19. September 2019 verabschiedet worden.

---

[1] Der Ausschuss dankt der Arbeitsgruppe *Big Data in der Lebensversicherung* ausdrücklich für die geleistete Arbeit, namentlich Stefan Heyers, Dominique Achard, Dr. Fabian Bohnert, Karin Brendel, Dr. Henning Christ, Andreas Döring, Thomas Gehling, Lukas Hahn, Dr. Michael Hoffmann, Dr. Andreas Kronwald, Bartolomiej Maciaga, Sven Rehmann, Mareike Welter und Sebastian Schirdewahn.

[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.

Inhaltsverzeichnis

1. Überblick über Big Data und Künstliche Intelligenz in der Lebensversicherung

1.1. Hintergrund und Ziele dieses Berichtes

Daten sind der Treibstoff der Versicherungswirtschaft. Seit jeher benötigen die Versicherer Daten über die zu versichernden Risiken, um diese zu messen und zu bewerten. Erst dadurch wird es möglich, einen auskömmlichen Preis für den Risikoschutz zu ermitteln.

Die Verfahren zur Analyse dieser Daten wurden im Laufe der Zeit immer weiter verfeinert. Dies ermöglicht es, heute ein größeres Volumen an Daten zu verarbeiten, mehr Informationen in die Analysen einfließen zu lassen und auch komplexere Beziehungen zwischen einzelnen Merkmalen abzubilden. In diesem Sinne stellen auch die Entwicklungen in den Bereichen Big Data und Künstliche Intelligenz für die Versicherungswirtschaft keine Revolution, sondern eine Evolution dar.

Dieser Bericht soll die Entwicklungen aus den Bereichen Big Data, Künstliche Intelligenz und Machine Learning beleuchten und Anwendungsfelder in Versicherungsunternehmen aufzeigen. Es werden vorwiegend Einsatzmöglichkeiten für Prozesse dargestellt, die charakteristisch für das Lebensversicherungsgeschäft sind. Auf allgemeine Prozesse wie die Kundenansprache oder die Analyse von Cross-Selling-Möglichkeiten wird nicht näher eingegangen.

In diesem ersten Kapitel werden nachfolgend die gängigen Begriffe aus den Bereichen Big Data und Künstliche Intelligenz definiert. Zudem enthält dieses Kapitel einen Überblick zu den in diesem Bericht untersuchten Anwendungsfällen sowie zu weiteren Themen, die im Zusammenhang mit Big Data und Künstlicher Intelligenz diskutiert werden.

In Kapitel 2 werden die Anwendungsfälle im Einzelnen vorgestellt. Der erste Anwendungsfall ist ausführlich gehalten, um dem Leser anhand eines Beispiels konkrete Umsetzungshilfen für einen vollständigen Modellierungsprozess an die Hand zu geben. Die weiteren Anwendungsfälle konzentrieren sich auf die Beschreibung der Problemstellung und Skizzierung von Lösungsansätzen.

Das dritte Kapitel stellt verschiedene für die Versicherungswirtschaft relevante Datenquellen zur Schaffung und Aufbereitung einer geeigneten Datengrundlage vor. Es wird die Eignung dieser Datenquellen diskutiert und auf das Thema Datenschutz eingegangen.

Die Kapitel 4 und 5 enthalten als Anhang weiterführende Informationen zu statistischen Methoden und gängigen Softwareprogrammen.

1.2. Allgemeine Definitionen

Die folgende Darstellung zeigt den Zusammenhang gängiger Begriffe im Umfeld von Big Data:

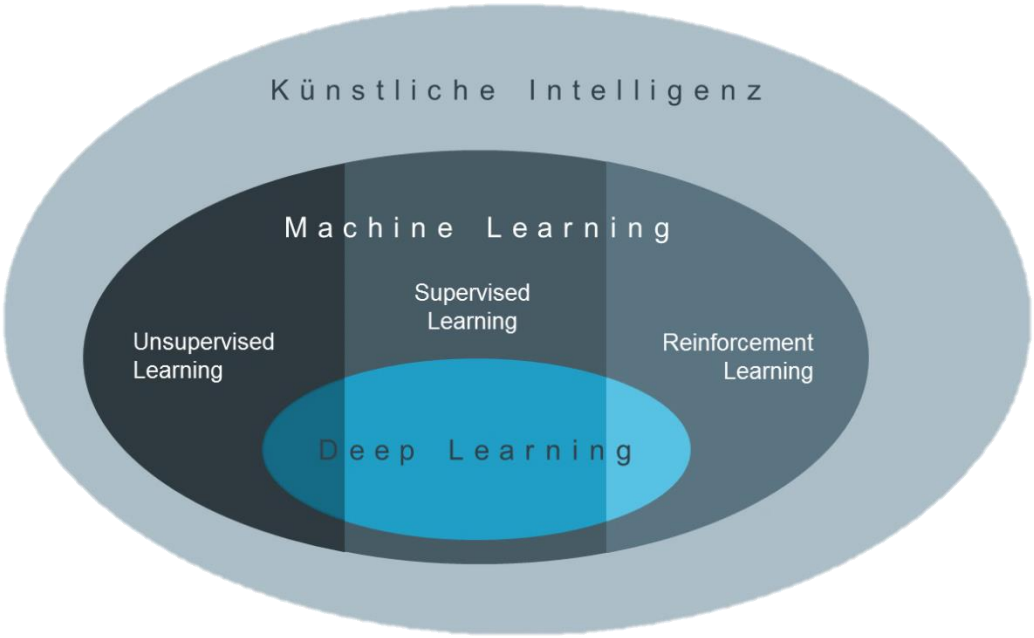

Abbildung 1: Gängige Begriffe im Umfeld von Big Data

- Big Data: Eine einheitliche Definition des hier verwendeten Überbegriffs „Big Data" gibt es nicht. Nach gängiger Auffassung befasst sich Big Data jedoch mit der Frage, wie riesige Datenmengen gespeichert und verarbeitet werden können. Darunter fallen beispielsweise Methoden, mit denen Berechnungen auf mehrere Rechner verteilt und die Teilergebnisse anschließend wieder korrekt zusammengeführt werden können, oder Methoden zur verteilten Datenhaltung auf mehreren Servern. Teilweise wird unter Big Data auch die Arbeit mit unstrukturierten Daten verstanden.
- Data Analytics: Data Analytics bezeichnet die Untersuchung von Datensätzen mit der Zielsetzung, daraus neue Erkenntnisse abzuleiten. Die Bezeichnung meint den Analyseprozess an sich, für den wiederum Techniken und Technologien aus den Bereichen Künstliche Intelligenz, Machine Learning oder Deep Learning eingesetzt werden können.
- Künstliche Intelligenz: Als Künstliche Intelligenz werden Systeme bezeichnet, die Probleme lösen sollen, für deren Lösung – wenn von Menschen gehandhabt – Intelligenz notwendig ist. Der Begriff ist weiter gefasst, als häufig angenommen wird. So fallen darunter u. a. auch regelbasierte Systeme oder systematische Suchen in Lösungsräumen. In der aktuellen Diskussion wird Künstliche Intelligenz oft auf den Bereich Machine Learning beschränkt.
- Machine Learning: Machine Learning ist ein Teilbereich der Künstlichen Intelligenz, bei dem die Lösung eines Problems nicht regelbasiert programmiert, sondern von einem Algorithmus erlernt wird – durch einen autonomen Lernprozess auf Basis von Beispieldaten. Das vom Machine Learning erlernte Modell entspricht aus statistischer Sicht einem Schätzer. Analog zur Statistik wird anhand der Beispieldaten die Abweichung zwischen diesem Schätzer und

dem wahren Wert minimiert – mit Machine Learning werden also mathematische Modelle mittels Optimierungsverfahren kalibriert. Dabei kommen auch bekannte statistische Modelle zur Anwendung, wie z. B. die lineare Regression oder verallgemeinerte lineare Modelle (Generalized Linear Models, GLMs).

- Deep Learning: Deep Learning ist ein Teilbereich des Machine Learnings, bei dem aus den Rohdaten eine Vielzahl von Zwischenergebnissen abgeleitet wird, die unterschiedliche Abstraktionsgrade der enthaltenen Informationen widerspiegeln. Dies wird in der Regel durch den Einsatz von komplexen neuronalen Netzen mit vielen nichtlinearen Transformationen der Inputdaten erreicht.

Eine Einführung in die Themen, die in diesem Bericht behandelt werden, wird im Anhang sowie den dort genannten Literaturverweisen gegeben. Die Lektüre dieses Berichts setzt ein Grundverständnis der dort dargestellten Begriffe voraus.

Eine weitere Besonderheit hinsichtlich Big Data / KI in der Lebensversicherung sind schiefe asymmetrische Verteilungen aufgrund von sehr geringen Häufungen von Ereignissen (Spätstorno, Sterblichkeit, Arbeitsunfähigkeit, …). Für viele Data Science Methoden stellen schiefe Verteilungen ein großes Problem dar und sorgen für eine sehr schwache Vorhersagegüte der Modelle. Die Problematik liegt hierbei besonders darin, dass diese relevanten Ereignisse in gegebenen Daten nicht repräsentativ genug vorhanden sind, sodass Modelle diese Beziehungen nur sehr ungenau lernen können. Vor allem bei Klassifikationsproblemen ist es hier besonders schwierig Klassen mit geringer Datenmenge gut zu klassifizieren. Auf Basis einiger aktueller Studien gibt es für solche Probleme bereits Techniken und Datentransformationen wie Over- oder Undersampling[3], die es Modellen ermöglichen balanciertere Daten für das Training zu verwenden und der Verteilungsschiefe entgegen zu wirken.

[3] https://www.sciencedirect.com/science/article/pii/S1474667016429952

1.3. Überblick Anwendungen

Im folgenden Schaubild werden ausgewählte Anwendungen aus dem Bereich Big Data und Künstliche Intelligenz im Bereich der Lebensversicherungswirtschaft entlang der Wertschöpfungskette dargestellt, inkl. Verweisen auf das jeweils relevante Kapitel in diesem Bericht.

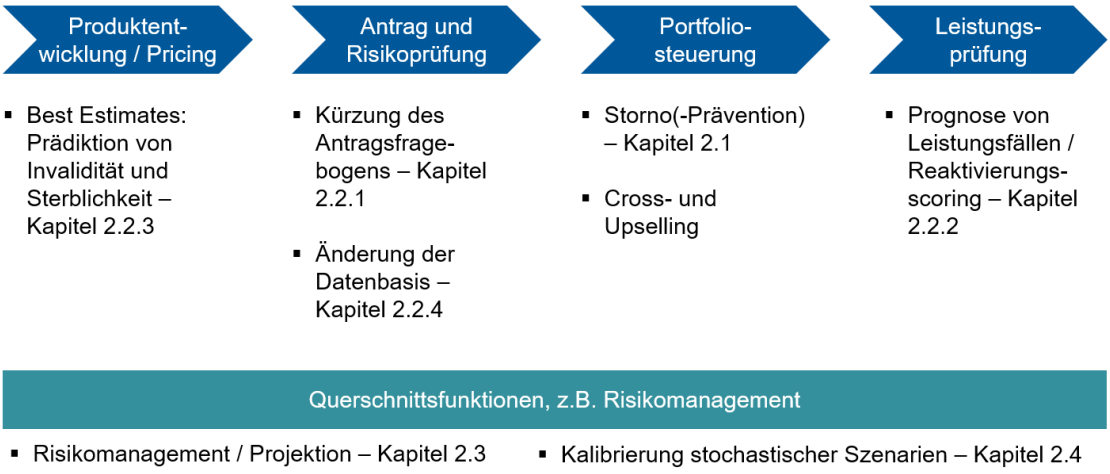

Abbildung 2: Einsatzmöglichkeiten von Big Data und Künstlicher Intelligenz entlang der Wertschöpfungskette

1.4. Aktuelle relevante Diskussionsthemen

Lebensversicherer bieten gesellschaftlich bedeutende Produkte und Dienstleistungen an. Der Zugriff auf Absicherungen gegen biometrische Risiken und Altersvorsorgemöglichkeiten erfüllen elementare Bedürfnisse. Aufgrund dieser Bedeutung und der langfristigen Ausrichtung eines Lebensversicherungsvertrages ist eine gesetzliche Regulierung der Lebensversicherer wie auch der Produkte Standard.

Die Nutzung von Big-Data-Verfahren in der Lebensversicherung *kann* demnach weitreichende Konsequenzen für die Kunden haben. Themen sind hier beispielsweise automatisierte Entscheidungen, Kundenscoring, Diskriminierung, Erklärbarkeit von Entscheidungen und Verbraucherschutz.

Mit der Weiterentwicklung und Verbreitung von Big-Data-Verfahren hat die Anzahl von Stellungnahmen und Publikationen zu ethischen Aspekten der Anwendung von Künstlicher Intelligenz stark zugenommen. Stellvertretend seien genannt:

- Big Data trifft auf künstliche Intelligenz, Bundesanstalt für Finanzdienstleistungsaufsicht (BaFin), 2018,
- Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren, Gesellschaft für Informatik, Berlin, 2018,
- Everyday Ethics for Artificial Intelligence, IBM, 2018,

- High Level Expert Group on Artificial Intelligence, The European Commission, Ethics Guidelines for Trustworthy AI, 2019.

Jedes Versicherungsunternehmen ist selbst für einen angemessenen Einsatz von Big-Data-Verfahren verantwortlich. Abschnitt 3 diskutiert ausgewählte Themen aus den Bereichen Datenschutz und Diskriminierung ausführlicher.

2. Ausgewählte Anwendungsfälle

2.1. Storno

2.1.1. Einführung

Dieses Kapitel beschäftigt sich mit Anwendungsbeispielen von Big Data im Umfeld des Stornoverhaltens von Versicherungsnehmern. Es wird aufgezeigt, wie mit „modernen" prädiktiven Modellen Stornovorhersagen getroffen werden können, die im Vergleich zu klassisch abgeleiteten Stornotafeln sowohl im Umfang als auch in ihrer Art zusätzliche Informationen berücksichtigen und somit eine bessere Steuerung ermöglichen können (z. B. für Werbeaktionen, Bestandsprojektionen oder die Entwicklung von Provisionsmodellen).

Es zeigt sich, dass sich bereits im eingeschränkten Umfeld von Stornoanalysen sehr vielfältige Fragestellungen ergeben, z. B.

- Herleitung von Stornotafeln,
- Modellierung von Frühstorno oder
- Stornoprävention.

Vor dem Erarbeiten eines speziellen Stornomodells muss somit die zugrundeliegende Fragestellung klar definiert sein, da sie Auswirkungen auf die Modellierung, die Anpassung der Parameter und die Auswahl der Daten hat. Die Stornoanalyse besteht dann mindestens aus den folgenden Schritten:

- Definition der Zielsetzung und Maßnahmen,
- Aufbau des Modells zur Stornovorhersage,
- Evaluierung des Modells,
- Evaluierung der Zielerreichung.

Diesen Prozess wollen wir konkret anhand des Anwendungsfalls der Stornoprävention als Beispiel für ein binäres Klassifikationsproblem detailliert vorstellen. Viele Erkenntnisse lassen sich anschließend mit entsprechenden Überlegungen auf andere Probleme übertragen. Weitere Fragestellungen im Rahmen der Stornomodellierung werden beispielhaft in Abschnitt 2.1.7 erläutert.

Will ein Versicherungsunternehmen eine zielführende Stornopräventionsaktion durchführen, stellt sich die Frage danach, welche Zielgruppe angesprochen werden soll. Ein Aspekt ist die Identifizierung von Kunden mit einer hohen Stornowahrscheinlichkeit zur Beantwortung der Fragestellung: „Welche Kunden haben ein erhöhtes Stornopotenzial?" Somit steht nicht der Grund für das Stornieren im Mittelpunkt. Vielmehr ist das Ziel, Kunden abhängig vom Stornopotenzial zu sortieren, sodass anschließend eine gewisse Kundenanzahl (z. B. abhängig vom Budget) zur Stornoprävention ausgewählt werden kann. Dies führt dazu, ein Modell mit einer hohen Güte (vgl. Anhang 4.12) zu entwickeln und erst zweitrangig auf die Interpretierbarkeit des Stornomodells einzugehen.

Eine eingehende Betrachtung dieses Modellierungsproblems wird in den nun folgenden Abschnitten ausgeführt.

2.1.2. Datengrundlage

Die Stornomodellierung zum Zweck der Stornoprävention stellt ein typisches Beispiel für ein überwachtes Lernproblem (vgl. Abschnitt 4.1.6) dar. Hierfür sind sowohl eine Zielvariable, die das zu prognostizierende Ereignis Storno misst, sowie erklärende Merkmale, die einen Zusammenhang mit dessen Eintrittswahrscheinlichkeit aufweisen (könnten), als Basis der Modellierung bereitzustellen. Allgemeine Hinweise zur Schaffung einer geeigneten Datengrundlage finden sich in Kapitel 3.

Für den hier vorgestellten Use Case gilt es vor allem, bereits bei der Definition und Erhebung des Stornokennzeichens das Hauptaugenmerk darauf zu richten, die möglicherweise noch recht abstrakte unternehmerische Zielsetzung in eine wohldefinierte und geeignete mathematisch-statistische Problemstellung – ggf. unter Berücksichtigung bereits geplanter Maßnahmen – zu überführen. Bezüglich der erklärenden Merkmale werden im Falle der Stornomodellierung neben Vertrags- und Kundendaten insbesondere Produkt- und Vermittlerinformationen interessieren (z. B. Garantie- und Rückkaufbedingungen, Provisionshaftung etc.) sowie bei kapitalbildenden Produkten deren Verlauf, ggf. im Vergleich zum Markt.[4] Weiterhin ist zu beachten, dass wesentliche Einflussgrößen auf das Stornoverhalten, beispielsweise aus dem persönlichen Lebensbereich des Versicherungsnehmers, für die Analyse nicht zur Verfügung stehen mögen und der Modellgüte auch dadurch Grenzen gesetzt sind. Aus allen Merkmalen, die als relevant für das Stornoverhalten eingeschätzt werden, sind diejenigen auszuwählen, die letztendlich in ausreichender Qualität vorliegen.

Datenabzug

Für den Use Case des Bestandsstornos spielen historische Daten eine besondere Rolle. Hierunter verstehen wir den rückwirkenden Abzug der Daten zu einem bestimmten Zeitpunkt X in der Vergangenheit, d. h. mit dem Informationsstand, der am Tag X tatsächlich vorgelegen hat.

Soll beispielsweise jeweils zum Stichtag 31.12. auf den dann vorliegenden Bestandsdaten die Stornoanfälligkeit im nächsten Geschäftsjahr für eine zum Anfang des Jahres eingeleitete Kundenansprache untersucht werden, so ist für die Modellbildung der Datenbestand zu einem zurückliegenden 31.12. in Verbindung mit dem beobachteten Stornoverhalten des darauffolgenden Jahres zu nutzen. Werden fälschlicherweise aktuelle Datenbestände für die erklärenden Merkmale verwendet, so liegt in den erklärenden Merkmalen bereits ein Informationsgehalt vor, wie er in der realen Anwendung nicht beobachtbar sein wird. Damit entsteht eine Inkonsistenz zwischen den erlernten Mustern und für die Anwendung benötigten

[4] Towers Watson hat sich in dem Artikel „Predictive Modeling for Life Insurers" beispielhaft mit der Modellierung von Stornoverhalten bei Variable Annuities auseinandergesetzt.

Mustern, die die tatsächliche Modellgüte negativ beeinträchtigt. Auf die zeitliche Kohärenz ist entsprechend zu achten, wenn mehrere zurückliegende Jahre zur Datengrundlage verwendet werden sollen oder wenn das Modell regelmäßig (z. B. zweiwöchentlich) bzw. mit einem zeitlich fixierten Vorlauf zur Einleitung der Präventionsmaßnahmen eingesetzt werden soll.

Datenaufbereitung und -exploration

Bevor mit der technischen Modellierung begonnen werden kann, ist die Datengrundlage zunächst geeignet aufzubereiten und zu analysieren. Dies dient sowohl der Beurteilung der Datenqualität zur vorläufigen Merkmalsauswahl als auch dem für die geeignete Modellwahl und -optimierung zwingend notwendigen Datenverständnis.

Einen ersten Schritt stellt i. d. R. eine deskriptive univariate Analyse mittels statistischer Kennzahlen sowie grafischer Visualisierungen der Zielvariablen und erklärenden Merkmale dar, um u. a.

- den Merkmalstyp (z. B. stetig, diskret oder kategorial),
- die Anzahl an Kategorien bei kategorialen Merkmalen und deren Verteilung (z. B. extremes Ungleichverhältnis zwischen Kategorien),
- die Verteilung stetiger Merkmale (insbesondere deren Schiefe),
- den Befüllungsgrad und ggf. die Natur der fehlenden Daten,
- das Vorliegen von Ausreißern

abzuschätzen. Bivariate Analysen zwischen den erklärenden Merkmalen und der Zielvariable vermitteln zudem einen ersten Eindruck bzgl. wichtiger Einflussfaktoren auf das Stornogeschehen. Paarweise deskriptive Analysen innerhalb der erklärenden Merkmale dienen dazu, Korrelationen zu identifizieren, die zu Problemen während der Modellierung (z. B. Multikollinearität) oder bei der Modellinterpretation (z. B. Aufteilung des gemeinsamen Informationsgehalts auf die einzelnen Attribute) führen können.

Anhand einer solchen Datenexploration stößt man typischerweise auf etwaige Fehler in den Daten, die somit bereits vor der Modellerstellung bereinigt werden können. Sehr schlecht befüllte Variablen sollten an dieser Stelle entfernt bzw. substituiert werden, da sie während der Modellierung prinzipiell schädlich sein können. Je nach Qualität und Granularität der untersuchten Variablen können zur Ableitung einer geeigneten Datengrundlage zudem verschiedene statistisch bzw. fachlich getriebene Anpassungen erfolgen: beispielsweise die Kappung von Ausreißern, die Zusammenfassung einzelner Kategorien mit geringen Fallzahlen oder das Binning/Banding stetiger Merkmale zu kategorialen Attributen. Einen wesentlichen Schritt stellt dabei auch das Feature Engineering dar, bei dem aus den vorliegenden Daten fachlich sinnvolle Attribute abgeleitet werden, die dem Modell als Grundlage einer verbesserten Mustererkennung dienen können.

Bei der Datenaufbereitung ist es allerdings wichtig, zwischen der Herleitung einer qualitativ angemessenen Datengrundlage und bereits modellgetriebenen Anpassungen zu trennen. So hängen beispielsweise Standardisierungen von Merkmalen von den empirisch beobachteten Mittelwerten und Standardabweichungen in den vorliegenden Trainingsdaten ab. Da ein darauf aufbauendes Modell aber auf den derart adjustierten Merkmalen optimiert wird, ist die auf den Trainingsdaten ermittelte Transformation als Teil des finalen Modells zu verstehen (und nicht etwa bei Anwendung auf den dann vorliegenden Daten neu zu bestimmen). Zur Vermeidung von Inkonsistenzen empfiehlt es sich, solche modellgetriebenen Vorverarbeitungsschritte als Teil der technischen Modellierung durchzuführen und diese konsistent erst nach der Aufteilung in Trainings- und Testdaten anzuwenden.

Die durch die Datenexploration gewonnenen Erkenntnisse lassen sich im weiteren Verlauf der Stornomodellierung gewinnbringend nutzen. So kann die Einschätzung zu Ausreißern und fehlenden Daten die Wahl geeigneter Modelle (z. B. Modelle, die mit fehlenden Daten oder Ausreißern robust umgehen können) oder alternativer Vorverarbeitungsschritte (z. B. Kappung von Ausreißern oder Imputation fehlender Daten, siehe Abschnitt 3.6) maßgeblich beeinflussen.

Eine gute Datengrundlage legt den Grundstein für alles, was in den nachstehenden Abschnitten folgt. Ein noch so gut kalibriertes Modell hat keinen Mehrwert, wenn die Datengrundlage unzureichend oder nicht auf die ursprüngliche Fragestellung abgestimmt ist. Im Gegenteil kann eine schlechte Datenbasis zu falschen Erkenntnissen und Entscheidungen führen, die im ungünstigen Fall einen wirtschaftlichen Schaden bedeuten können.

2.1.3. Modellauswahl und Durchführung

Bei der hier betrachteten Stornomodellierung auf Basis von Vergangenheitsdaten handelt es sich um ein überwachtes Klassifikationsproblem, da neben den beschreibenden Merkmalen das tatsächliche Stornogeschehen als kategoriale Zielgröße vorliegt (vgl. Abschnitt 4.1.6). Zur Modellierung kommt somit eine Vielzahl von Klassifikationsverfahren des Predictive Modellings in Frage. Für den einfachsten, aber typischen Fall, dass Storno als binäres Ereignis (storniert vs. nicht storniert) definiert wird, kann sich insbesondere auf den Spezialfall eines binären Klassifikationsmodells konzentriert werden. Die Vorgehensweise lässt sich allerdings auch auf die Modellierung verschiedener Stornoereignisse verallgemeinern, worauf in Abschnitt 2.1.5 kurz eingegangen wird. Für theoretische Grundlagen zu den in diesem Abschnitt besprochenen Modellen sei auf Kapitel 4 (Anhang) verwiesen.

Anforderungen an Modellergebnisse und Ausgabeformat

Bevor nun Modellklassen ausgewählt und mit der technischen Modellierung begonnen wird, ist zunächst zu bedenken, welche Ergebnisse geliefert werden und welches Ausgabeformat benötigt wird, um diese zu bewerten und darauf basierend Entscheidungen zu treffen und entsprechende Maßnahmen durchzuführen.

Typischerweise liefern Klassifikationsmodelle prognostizierte Stornowahrscheinlichkeiten auf Basis des beobachteten Stornogeschehens der Vergangenheit. Bei

vielen kommerziellen oder frei verfügbaren Modellimplementierungen ist häufig als Standardeinstellung eine zusätzliche diskrete Entscheidungsregel anhand des Schwellenwerts 50% zur Einstufung als Stornierer (Wahrscheinlichkeit > 50%) bzw. Nicht-Stornierer (Wahrscheinlichkeit ≤ 50%) vorgesehen[5]. Für viele Fragestellungen ist dennoch die Modellierung der (impliziten) Stornowahrscheinlichkeiten zielführender:

- Stornowahrscheinlichkeiten sind vor allem dann geeignet, wenn diese als Input für ein aktuarielles Modell verwendet oder Kunden gemäß Stornorisiko angeordnet werden sollen. Insbesondere im ersten Fall ist bei den Modellanforderungen darauf zu achten, dass die prognostizierten Stornowahrscheinlichkeiten in ihrer Gesamtheit zu einer adäquaten Wahrscheinlichkeitsverteilung führen.
- Auch im Fall konkreter Maßnahmen auf Basis einer diskreten Einteilung in potenzielle Stornierer und Nicht-Stornierer, z. B. im Rahmen einer vertrieblichen Stornoprophylaxe, eignet sich zunächst die Modellierung von Stornowahrscheinlichkeiten. So kann die Festlegung des Schwellenwerts als diskrete Entscheidungsregel vom eigentlichen Vorhersagemodell fachlich und auch zeitlich entkoppelt werden. Neben der Möglichkeit einer besseren statistischen Güte der binären Entscheidungsregel sind hier vor allem die ökonomische Bewertung und deren entsprechende Übersetzung in einen (ggf. adjustierbaren) Schwellenwert für die Durchführung von Maßnahmen entscheidend.
- Zur Wissensgenerierung ist die Analyse von Einflüssen der beschreibenden Merkmale auf das Stornogeschehen mittels statistischer Kennzahlen oder grafischer Visualisierungen i. d. R. intuitiver und aufschlussreicher, wenn die Veränderung im Stornorisiko in Form der Stornowahrscheinlichkeiten anstelle von diskreten Entscheidungen abgebildet werden kann.

Die Bandbreite der nun zur Modellierung von Stornowahrscheinlichkeiten zur Verfügung stehenden Klassifikationsverfahren ist groß. Die konkrete Modellauswahl sollte auf Basis verschiedener weiterer Kriterien erfolgen, von denen einige im Folgenden exemplarisch diskutiert werden.

[5] Die impliziten Wahrscheinlichkeiten liegen i. d. R. dennoch berechnet vor und können durch ein entsprechendes Funktionsargument, z. B. bei Vorhersagefunktionen wie „predict" in R, für das Ausgabeformat erzeugt werden. Möglich ist aber auch, dass bereits beim Training eines Modells eine entsprechende Einstellung zu wählen ist.

Güte

Wesentliches Entscheidungskriterium für die Modellauswahl ist die von dem Verfahren zu erwartende erreichbare Güte bei der Stornoprädiktion auf ungesehenen Daten. Insbesondere im Fall konkret geplanter Maßnahmen auf den Daten zum Zeitpunkt der Modellanwendung sollte die Fähigkeit, Stornierer möglichst gut von Nicht-Stornierern abzugrenzen und die Ergebnisse in anwendbare Regeln zu übersetzen, das Hauptauswahlkriterium sein. Es ist *a priori* allerdings prinzipiell nicht klar, welche Modellklasse für die konkrete Problemstellung und Datengrundlage am besten geeignet ist.

In der Praxis zeigt sich, dass Stornomodelle häufig auf vergleichsweise strukturierten Daten mit einer überschaubaren Anzahl an Merkmalen entwickelt werden. Die Vorhersagegüte unterliegt i. d. R. einem hohen Maß an Fehlklassifizierungen, da der Entscheidungsprozess eines Kunden hin zu Storno oft nicht oder nur unzureichend in den vorliegenden Daten des Versicherers abgebildet ist. So liegen nur selten detaillierte kundenspezifische Daten zur aktuellen ökonomischen Situation vor, von denen man einen hohen Erklärungsgehalt erwarten würde. Zum anderen wird es anders als bei prominenten Anwendungsfällen von Machine-Learning-Verfahren wie der Bilderkennung eine natürliche Grenze geben, inwiefern Zusammenhänge zwischen Storno und überhaupt erfassbaren Merkmalen vorliegen.

Erfahrungen der letzten Jahre bei Klassifikationsproblemen mit entsprechender Datengrundlage deuten darauf hin, dass in solch einer Ausgangssituation gute Ergebnisse mit hinreichend kalibrierten Gradient Boosting Machines, Random Forests, vergleichsweise einfachen neuronalen Feedforward-Netzen bzw. der Kombination solcher Modelle (Stacking) erzielt werden können. Ist die finale Modellgüte das Hauptkriterium für das Stornomodell, ist es wichtig, nicht nur mit einer Modellklasse, sondern mehreren Modellklassen unterschiedlicher Komplexität zu arbeiten und Resultate zu vergleichen. Einerseits sollen existierende Strukturen hinreichend fein erfasst werden können, gleichzeitig sollen die Ergebnisse mit vertretbarem Aufwand praktisch anwendbar und interpretierbar sein.

Komplexität und Interpretierbarkeit

Die Komplexität des verwendeten Verfahrens bzw. des final entstehenden Modells kann einen Einfluss auf die Modellauswahl haben. Vor allem wenn die Wissensgenerierung zu Stornotreibern, die Erklärbarkeit der Zusammensetzung einer individuellen Stornovorhersage für den Modellanwender oder die notwendige Modellakzeptanz durch weitere Stakeholder (z. B. durch den Vertrieb) Teil der Zielsetzung sind, bedarf es der Möglichkeit, entsprechende Erkenntnisse aus dem Modell ableiten und kommunizieren zu können.

Ein typischer Ansatz ist hier die Verwendung einfacher und unmittelbar interpretierbarer Verfahren wie ein Klassifikationsbaum oder eine logistische Regression. Vor allem für Anwender mit bisher wenig Erfahrung bieten sich diese Modelle für erste Erfolge sowie einfachere Interpretation und Kommunikation an. Aber auch

bei Verwendung komplexerer Verfahren ermöglichen solche Modelle als erster Einstieg in die Modellierung sowohl eine Benchmark für den Mehrwert komplexerer Methoden als auch erste Erkenntnisse über die relevanten Merkmale und mögliche Interaktionen. Neben bewährten deskriptiven Voranalysen können entsprechende Erkenntnisse bei der Merkmalsauswahl und beim Feature Engineering, aber auch beim Abgleich mit bereits vorhandenem Wissen oder Hypothesen zur Datenvalidierung helfen.

Die Nutzung einfacher Modelle geht allerdings i. A. auf Kosten der Modellgüte. Klassifikationsbäume führen mit ihren diskreten Entscheidungsregeln zu unstetigen und stark vereinfachten Vorhersagefunktionen, die vorhandene Muster oft nur unzureichend erfassen können. Regressionsmodelle bergen vor allem in den Rändern das Risiko einer schlechteren Vorhersagegüte. Mit Blick auf die Zielsetzung des Modells ist deshalb abzuwägen, was der Vorteil eines zwar verständlicheren oder nachvollziehbareren Modells ist, wenn es die Wirklichkeit nicht gut abbildet.

Demgegenüber steht mittlerweile eine Vielzahl von sich in den letzten Jahren stark weiterentwickelten Möglichkeiten zur zumindest approximativen Darstellung von Merkmalseffekten und dem Zustandekommen einzelner Vorhersagen zur Anwendung bei komplexeren Modellen zur Verfügung, um gerade aus diesen Verfahren neues Wissen zu generieren, komplexere Modellvorhersagen erläutern zu können und zusehends Ängste abzubauen und Akzeptanz für solche Modelle zu schaffen. Vor allem die Verwendung von baumbasierten Verfahren wie Random Forests und Gradient Boosting Machines ist hierfür geeignet, da beispielsweise mittels modellspezifischer „Explainers" die Zusammensetzung der Stornovorhersage anhand der Features eindeutig abgeleitet und mittels Wasserfall-Charts anschaulich visualisiert werden kann[6]. Allgemein kann dies für komplexe Modelle aber auch durch modellagnostische Verfahren zumindest approximativ erfolgen, z. B. indem auf den Modellvorhersagen ein vollständiger Entscheidungsbaum trainiert und dessen Regelwerk zur Ableitung solcher Vorhersagezusammensetzungen genutzt wird. Alternativ oder ergänzend kann die Erklärung für die Vorhersage einer Beobachtung auch lokal, d. h. im Vergleich zu ähnlichen Kunden oder Verträgen, approximiert werden, z. B. mittels des Algorithmus für Local Interpretable Model-Agnostic Explanations (LIME)[7]. Erkenntnisse über globale Zusammenhänge können neben der Sortierung nach Relevanz mittels Variable Importance Factors (VIF, vgl. Abschnitt 4.13) oder konkret durch Visualisierungen beispielsweise mittels Partial Dependence Plots (PDP, vgl. Abschnitt 4.13.2 bzw. 2.3.2) erfolgen.

Eine alternative (oder auf Basis obiger Erkenntnisse ergänzende) Möglichkeit ist auch die Verwendung der logistischen Regression in Kombination mit Regularisierungsverfahren (Elastic Nets, insbesondere Lasso und Ridge), um die bekannte und gut interpretierbare Regressionsgleichung mit Machine-Learning-Ansätzen zur datengetriebenen Variablenselektion und -dämpfung zu verallgemeinern. Deren

[6] Vgl. z. B. Pakete wie „xgboostExplainer" in R und Python für mit dem Algorithmus „xgboost" erstellte Gradient Boosting Machines.

[7] Vgl. Ribeiro et al. (2016).

Modellgüte kann durch hinreichend gute Effekt- und Interaktionsmodellierung oft an die Güte komplexerer Verfahren heranreichen.

Stabilität und Zeitkonsistenz

Aufgrund von sich über die Zeit änderndem Stornoverhalten – z. B. wegen rechtlicher und ökonomischer Rahmenbedingungen oder auch auf Basis von sich ändernden Beständen und der verfügbaren Datengrundlage – muss ein Stornomodell regelmäßig aktualisiert werden. Dabei sollte eine Stabilität in den Stornovorhersagen über die Zeit sichergestellt werden. Das gilt neben der eigentlichen Stornoprognose (bei unveränderten Merkmalen und Mustern) vor allem auch für die Zusammensetzung der Stornovorhersage, sofern diese zu Interpretationszwecken oder in der anschließenden Maßnahme verwendet wird. Klassifikationsbäume haben aufgrund ihrer hohen Varianz die Tendenz zum Overfitting (vgl. Abschnitt 4.5).

Weitere Kriterien für die Modellauswahl können der Umgang der verschiedenen Verfahren mit einer schlechten Qualität der Daten (z. B. fehlende Werte und Ausreißer) oder deren Auswertungsperformance in der Anwendung sein.

2.1.4. Modelloptimierung

Nach Auswahl einer oder mehrerer Modellklassen gilt es, auf Basis der vorliegenden Vergangenheitsdaten ein bestmögliches Vorhersagemodell abzuleiten. Die Datengrundlage ist wie in Abschnitt 2.1.2 beschrieben spätestens zu diesem Zeitpunkt so aufzubereiten, dass sie dem Informationsstand entspricht, wie er auch bei Anwendung der Maßnahme in der Zukunft vorliegen wird.

Umgang mit Ungleichgewicht

Besonderes Augenmerk beim Training von Stornomodellen gilt dem starken Ungleichgewicht zwischen Stornierern und Nicht-Stornierern. Klassifikationsverfahren sind häufig implizit (z. B. durch die zur Anwendung kommenden Unreinheitsfunktionen bei der Erstellung eines Entscheidungsbaums) sowie explizit (Default-Einstellung für den Split beim Schwellenwert von 50%) auf eine Datengrundlage mit gleichgewichteten Ereignissen ausgerichtet. Ohne eine gesonderte Auseinandersetzung mit dieser Problematik verhindert dies die Herleitung eines geeigneten Modells, da der Algorithmus stets versucht, die Missklassifikationsrate zu minimieren, was in einem trivialen Modell, das nur das Ereignis Nicht-Storno vorhersagt, münden kann.

Umgangen werden kann das Problem u. a. mit einer der folgenden Vorgehensweisen:

- Viele Algorithmen erlauben die Spezifikation von Verlustmatrizen (Angaben von Kosten Stornierer als Nicht-Stornierer bzw. umgekehrt Nicht-Stornierer als Stornierer vorherzusagen), Beobachtungsgewichten und/oder Klassengewichten, die in der beim Modelltraining verwendeten Verlustfunktion berücksichtigt werden. Durch geeignete Wahl hoher Gewichte für Stornierer

bzw. kleiner Gewichte für Nicht-Stornierer kann ein effektives Gleichgewicht zwischen beiden Ereignissen erreicht werden.

- Weitere typische Ansätze sind das Down- oder Upsampling, indem aus der größeren Klasse so viele Beobachtungen wie in der kleineren Klasse zufällig gezogen werden bzw. aus der kleineren Klasse durch Ziehen mit Zurücklegen die gleiche Menge wie in der größeren Klasse erreicht wird. Nachteilig erweist sich hier, dass die Datengrundlage entweder durch Entfernen von Beobachtungen sehr klein und instabil oder aber sehr groß (mit für das Overfitting anfällige Replikationen von Beobachtungen) werden kann. Deshalb sollte das Sampling idealerweise in einen Resampling-Prozess eingebunden werden oder auch abgeleitete Verfahren zur Vermeidung von Replikationen (z. B. SMOTE-Algorithmus[8]) verwendet werden, die aber wiederum mit hohen Rechenzeiten verbunden sein können. Aus theoretischen Gesichtspunkten erscheint die Nutzung von Beobachtungsgewichten sinnvoller, um alle Daten in die Modellbildung einfließen zu lassen, aber keine Sampling-Algorithmen verwenden zu müssen. Zum Teil lässt sich das Down- oder Upsampling aber intelligent in den eigentlichen Trainingsalgorithmus integrieren, z. B. beim Bootstrap der Trainingsdaten für die individuellen Klassifikationsbäume innerhalb eines Random Forests.
- Das Erlernen eines nicht-trivialen Modells kann für einen Algorithmus wie einen Entscheidungsbaum auch erzwungen werden, indem sämtliche Default-Abbruchskriterien abgestellt werden – wie z. B. der minimale Erklärungsgehalt durch einen hinzukommenden Knoten. Nachteilig erweist sich bei diesem Vorgehen, dass die Werte für optimale Hyperparameter stark von typischen Werten abweichen und intensiver getunt werden müssen.
- Die Problematik kann auch durch die Modellwahl berücksichtigt werden. So eignen sich beispielsweise logistische Regressionsmodelle, da diese durch die Minimierung der Likelihoodfunktion auch bei Extremereignissen eine Struktur erlernen.

Zu beachten ist, dass bei Gewichtungs- oder Samplingverfahren die Modellausgabe entsprechend auf eine hypothetische Stornorate von 50% skaliert ist. Während dies für die reine Modellanwendung ausreichend ist, so ist es für die Modellinterpretation vorteilhaft, die entstehenden Stornowahrscheinlichkeiten wieder auf das Ausgangsniveau zu transformieren. Dies kann z. B. anhand einer zum Chancenverhältnis proportionalen Skalierung erfolgen, oder indem die vorläufige Modellausgabe nachkalibriert wird (z. B. Anwendung einer logistischen Regression für die beobachtete Stornozielgröße mit der vorläufigen Modellausgabe als erklärendes Merkmal).

Modelltuning

[8] https://arxiv.org/pdf/1106.1813.pdf

Die Optimierung eines Modells erfolgt anhand der zugrundeliegenden Hyperparameter, die von einer geringen Anzahl bei einfachen Modellen (z. B. Komplexitätsparameter, Baumtiefe und/oder minimale Blattgröße bei Klassifikationsbäumen) bis hin zu einer großen Fülle (z. B. sämtliche Einstellungsmöglichkeiten zu Topologie, Aktivierungen, Optimierungsalgorithmen usw. bei neuronalen Netzen) reicht. Mit Zunahme solcher Hyperparameter geht entsprechend eine Erhöhung des Aufwandes und der benötigten Erfahrung zum Modelltraining einher. Zusätzlich zu den modelleigenen Hyperparametern können i. w. S. auch die zusätzlichen Modellierungsentscheidungen wie der Umgang mit dem Ungleichgewicht oder eine vorgeschaltete Merkmalsreduktion als Hyperparameter angesehen und mit dem folgenden Vorgehen getunt werden.

Die Modelloptimierung sollte nach den bewährten Methoden der statistischen Lerntheorie erfolgen, um eine unverfälschte Schätzung der Modellgüte auf ungesehenen Daten zu erhalten und somit das Overfitting zu kontrollieren. Dazu ist mindestens eine Aufteilung von Trainingsdaten zur Modellerstellung und Testdaten zur Modellbewertung nötig (vgl. Abschnitt 4.1.3). Insbesondere, wenn für mehrere Modellklassen jeweils die zugehörigen Hyperparameter getunt und anschließend die als optimal befundenen Kalibrierungen zur Modellauswahl verglichen werden, ist eine Dreiteilung der Datengrundlage angezeigt. Dabei erfolgt das Modelltraining auf Trainingsdaten und Bewertung, Vergleich und Auswahl der Hyperparameter bzw. allgemein des Modells auf ungesehenen Validierungsdaten. Da die Validierungsdaten implizit Bestandteil der Modelloptimierung werden, ist das Gütekriterium für das optimale Modell i. d. R. zu optimistisch. Ist eine unverfälschte Abschätzung der Modellgüte für die Anwendung in der Praxis notwendig, wird das optimale Modell deshalb abschließend nochmals auf den Trainings- und Validierungsdaten trainiert und auf den bisher noch nicht verwendeten Testdaten bewertet.

Resampling

Wegen des hohen Anteils an Zufälligkeiten in der Modelloptimierung (Datenteilung, stochastische Komponenten in den Machine-Learning-Verfahren, ggf. Samplingverfahren zur Gleichgewichtung der Klassen) unterliegen die Validierungskenngrößen einer hohen Varianz, d. h. eine einfache Trennung von Trainings- und Validierungsdaten führt i. d. R. zu instabilen Ergebnissen und tendiert zu Overfitting. Der Trainings- und Validierungsprozess sollte deshalb mittels Resampling-Verfahren, typischerweise über Kreuzvalidierung, durchgeführt werden (vgl. Abschnitt 4.1.3). Die über alle Validierungsfaltungen berechneten und gemittelten Gütemaße sind zum einen stabiler und können zum anderen anhand der Standardabweichungen bzgl. Instabilität bewertet werden. Ein gängiges Verfahren zum Schutz gegen Overfitting ist es, das gewählte Modell soweit zu vereinfachen (z. B. bei Klassifikationsbäumen die Äste zu stutzen oder bei Random Forests oder Boosting die Baumanzahl zu verringern), bis das Gütekriterium den um eine Standardabweichung reduzierten optimalen Wert erstmals unterschreitet. Weiterhin sollte aufgrund des hohen Ungleichgewichts zwischen Stornierern und Nicht-Stornierern das

Sampling stratifiziert für beide Kategorien erfolgen, um unplausible Ausreißer in den Gütekriterien zu vermeiden.

Modellbewertung

Vor Beginn der Modelloptimierung sollte dazu zunächst die Definition der Modellgüte erfolgen. Hier ist mindestens ein Gütemaß zu bestimmen, das für verschiedene Modelle bzw. Modellkalibrierungen auf ungesehenen Daten ausgewertet und zur Bewertung bzw. finalen Auswahl herangezogen wird. Die Wahl des Gütemaßes sollte neben geeigneten statistischen Eigenschaften für Klassifikationsprobleme vor allem auch die angestrebte Maßnahme berücksichtigen.

Statistisches Gütemaß

Es bietet sich an, während der Modelloptimierung die Modellgüte anhand der prognostizierten Stornowahrscheinlichkeiten und nicht unmittelbar anhand der (u. U. nicht geeigneten) Default-Schwellenwertregel zu bewerten. Ein bewährtes Gütekriterium ist die Fläche unter der Receiver Operating Characteristic (ROC) Curve (Area under the Curve, AUC). Hierüber wird die mittlere Modellgüte über alle möglichen Schwellenwerte hinweg gemessen.

Details zur ROC-Analyse finden sich in Anhang 4.12.2 und 4.12.3. Für unseren konkreten Use Case ergibt sich die entsprechende Vierfeldertafel (Konfusionsmatrix) mit fixem Schwellenwert p beispielhaft zu:

Konfusionsmatrix für den Schwellenwert *p*	**Vorhersage: Storno** (**Stornowahrscheinlichkeit > *p***)	**Vorhersage: kein Storno** (**Stornowahrscheinlichkeit ≤ *p***)
Beobachtung: Storno	**richtig positiv** Stornierer identifiziert, Maßnahmen werden durchgeführt	**falsch negativ** Stornierer nicht identifiziert, fälschlicherweise keine Maßnahme
Beobachtung: kein Storno	**falsch positiv** Nicht-Stornierer als Stornierer identifiziert und fälschlicherweise Maßnahme durchgeführt	**richtig negativ** Nicht Stornierer identifiziert, keine Maßnahmen

Die der ROC-Kurve in Abbildung 3 zugrundeliegende Sensitivität entspricht dann dem Anteil der prognostizierten Stornierer an allen tatsächlichen Stornierern und die Spezifität dem Anteil der nicht-prognostizierten Nicht-Stornierer an allen tatsächlichen Nicht-Stornierern.

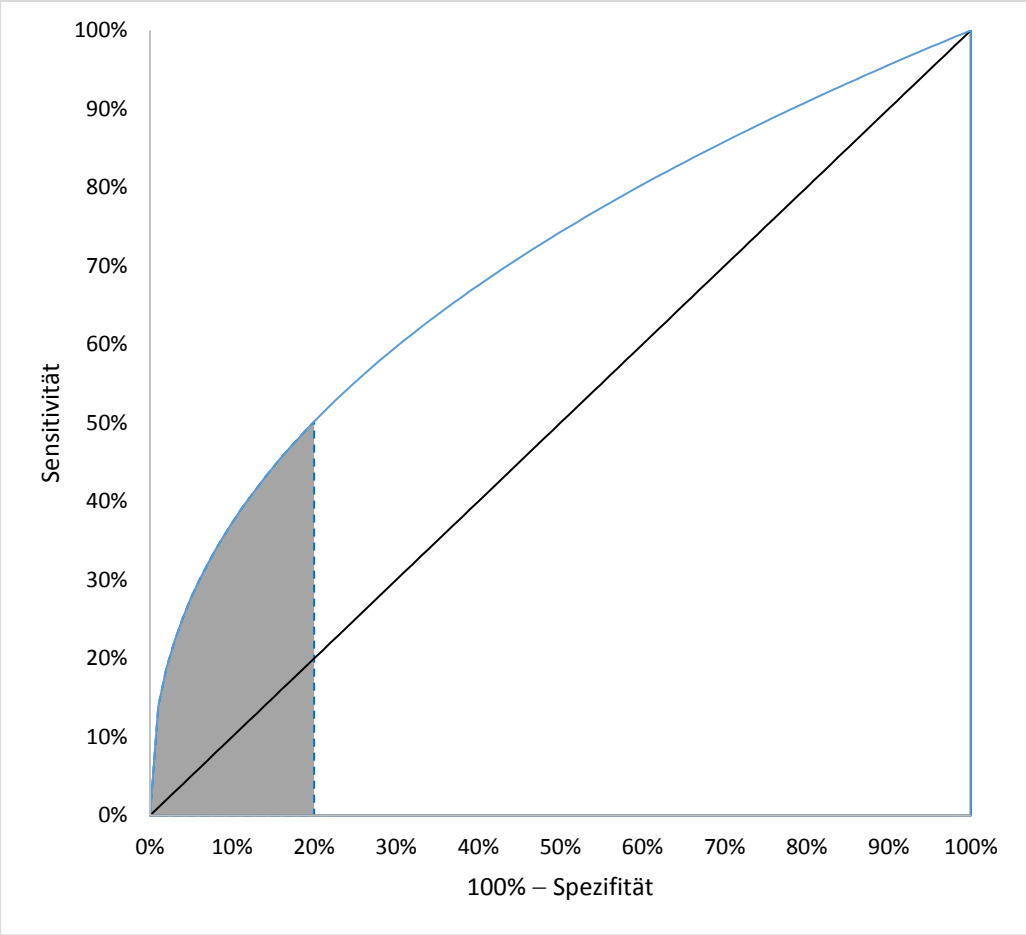

Abbildung 3: ROC-Kurve

Anreicherung um ökonomische Güte

Schon während der Modelloptimierung sollte über die statistische Bewertung hinaus die Erfolgsmessung der angestrebten Maßnahme berücksichtigt werden. Nur ein diesbezüglich trainiertes Modell garantiert, dass die bei der Anwendung zu erzielenden Ergebnisse optimiert werden, während die bestmögliche Kalibrierung anhand eines Gütemaßes aus rein statistischen Gesichtspunkten für den Erfolg der Anwendung gar nicht zwingend geeignet sein muss.

Umfasst die Maßnahme beispielsweise einen Kundenkontakt für alle als Stornierer vorhergesagten Kunden, so sind dessen Kosten für alle positiv vorhergesagten Fälle zu berücksichtigen. Andererseits werden Stornofälle einen Effekt auf die Profitabilität haben, der ebenfalls zu berücksichtigen ist[9]. Dementsprechend sind statistische Kennzahlen allein zur ökonomischen Bewertung des Modells ungeeignet, sondern mit den ökonomischen Auswirkungen zu gewichten. Das Modelltuning sollte dann konsistent anhand der AUC auf der ökonomisch gewichteten ROC-Kurve erfolgen.

Sowohl durch ein vorgegebenes Budget für die angestrebte Maßnahme als auch den aufgrund des Ungleichgewichts in den Daten unweigerlich hohen absoluten

[9] Dabei ist auch zu berücksichtigen, dass Storno eines unprofitablen Kunden vorteilhaft sein kann und nicht verhindert werden sollte.

Fehler zweiter Art bei Ausweitung der Maßnahme (überproportional viele falsch positive Vorhersagen pro korrekt positiver Vorhersage) liegt der zu wählende Schwellenwert für die binäre Klassifikation i. d. R. recht hoch. Damit ist für die Herleitung eines binären Entscheidungsmodells, das für die angestrebte Maßnahme wie dem Kundenkontakt überhaupt in Frage kommt, vor allem der Bereich der ROC-Kurve mit einer geringen Sensitivität, aber hohen Spezifität relevant (typischerweise unten links, siehe Abbildung 3). Für die Modelloptimierung ist der verbleibende Teil der ROC-Kurve prinzipiell nicht ausschlaggebend. Im Gegenteil – das Modell mit der höchsten AUC auf Validierungsdaten vermag möglicherweise auf dem irrelevanten Teil der ROC-Kurve überaus gut, im für die Anwendung relevanten Teil aber vergleichsweise schlecht abschneiden. Die Modell-optimierung sollte somit von vornherein auf den relevanten Teil der ROC-Kurve konzentriert werden.

Dies lässt sich bspw. für die Maximierung der Sensitivität oder der Spezifität erreichen, wenn die jeweils andere Größe für einen sinnvoll vorgegebenen Wert fixiert wird. So kann es anwendungsgetriebene Vorgaben geben, welche Mindestzahl an Stornierern identifiziert oder wie viele Nicht-Stornierer fälschlicherweise kontaktiert werden dürfen. Alternativ kann die Optimierung auch nur für die anteilige AUC für den durch die Fixierung relevanten Bereich der ROC-Kurve erfolgen.

Liegt ein für die Maßnahme verfügbares Budget vor, kann stattdessen zunächst bestimmt werden, wie oft die Maßnahme maximal durchgeführt werden kann. Mit dieser Information kann während der Modelloptimierung für jedes Modell die kleinste Schwellenwertregel bestimmt werden, für die diese Anzahl nicht überschritten wird, und anschließend das Modell mit der größten zugehörigen Sensitivität ausgewählt werden. So lässt sich sicherstellen, dass das Modell bei voller Ausschöpfung des Budgets die beste Stornoprädiktion aufweist.

Eine weitere Möglichkeit besteht in der Maximierung der AUC der Precision-Recall-Curve, also der Abtragung der Präzision gegen die Sensitivität. Diese Darstellung konzentriert sich ausschließlich auf die als Storno vorhergesagten Fälle und misst den Tradeoff zwischen tatsächlich identifizierten Stornierern (Sensitivät) und der erreichten Trefferquote in den positiven Vorhersagen (Präzision). Die Kurve ignoriert anders als die ROC-Analyse die Güte bei der korrekten Identifikation der Nicht-Stornierer. Diese ist für eine Maßnahme wie dem Kundenkontakt zur Stornoprävention von nachgelagerter Bedeutung, sodass die Optimierung der AUC der Precision-Recall-Curve zielführender als die Optimierung der AUC für die ROC-Kurve sein kann.

Anpassung des Schwellenwerts

Durch die Modelloptimierung anhand von übergreifenden Gütekriterien für alle möglichen Schwellenwerte kann die Fixierung des Schwellenwerts nicht nur separat während der Modellentwicklung bestimmt, sondern auch nachträglich adjustiert und bewertet werden. So kann bei geänderten Maßnahmen, geänderten ökonomischen Kosten oder auch bei Kapazitätsengpässen der Schwellenwert hoch- bzw.

herabgesetzt werden und die zugehörigen Auswirkungen aus der ROC-Kurve abgelesen werden.

2.1.5. Modellierung verschiedener Stornoereignisse

Die bisherigen Überlegungen für den binären Fall, dass Stornierer versus Nicht-Stornierer modelliert werden sollen, können auf den Fall mehrerer disjunkter Stornoereignisse (oder auf die Betrachtung alternativer Ausscheideursachen) übertragen werden. Die erwähnten Verfahren sind i. A. in der Lage, Zielgrößen mit mehr als zwei Kategorien zu modellieren, indem pro Kategorie eine zugehörige Wahrscheinlichkeit prognostiziert wird. Im Standardfall wird dann die Kategorie mit der höchsten Vorhersagewahrscheinlichkeit gewählt, allerdings lässt sich die Ableitung des eindeutigen Klassifizierers durch die Bestimmung von Schwellenwerten pro Kategorie analog verallgemeinern.

Zur Modelloptimierung und -bewertung existieren entsprechende Erweiterungen der AUC, indem paarweise AUC (Fläche der ROC-Kurve bei Betrachtung von nur zwei Kategorien) oder 1-vs.-all-AUC (Fläche der ROC-Kurve bei Betrachtung einer Kategorie gegenüber allen anderen Kategorien) gemittelt werden. Analog lässt sich hier die ökonomische Bewertung integrieren, z. B. wenn pro Stornokategorie verschiedene Maßnahmen mit entsprechenden Kosten und erwarteter Profitabilität angestrebt werden.

2.1.6. Praktische Umsetzung und Controlling

Kommunikation der Ergebnisse

Um die Modellergebnisse in Maßnahmen umzusetzen, müssen diese zunächst entsprechenden Entscheidungsgremien als Entscheidungsgrundlage vorgestellt werden. Einfache Metriken zur statistischen Güte sind hilfreich, aber nicht ausreichend. Die in Abschnitt 2.1.3 genannten Interpretationsverfahren helfen weiterhin, Vertrauen auch in komplexe Modelle aufzubauen, indem partielle Zusammenhänge visualisiert und mit bekannten oder erwarteten Mustern verglichen werden können. Der Hauptfokus bei der Kommunikation der Ergebnisse wird auf den vorhergesagten ökonomischen Implikationen der Maßnahmen liegen. Neben einer um die ökonomische Bedeutung angereicherten Konfusionsmatrix des finalen Klassifizierers (vgl. Abschnitt 2.1.4) kann auch die Angabe der entsprechend angereicherten Gini-Koeffizienten oder von Lift-Values helfen, die Auswirkungen der Klassifikation in erwartete Stornierer und Nicht-Stornierer transparent zu machen (Anhang 4.12.1 und 5.3).

Nutzung und Controlling der Ergebnisse in bestehenden Systemen und Prozessen

Eines der Systeme, aus denen die Datengrundlage extrahiert wurde, sollte das Klassifikationsergebnis und die Information zur Maßnahme erfassen können und für die iterative Nutzung der Modellierung Änderungen im Zeitverlauf flexibel zulassen.

Eine Stornopräventionsmaßnahme könnte mittels eines postalischen Kundenkontakts realisiert werden. Hierbei sollte hinterfragt werden, ob es sich als zielführend herausstellt, Kunden mit einem hohen Stornopotenzial anzuschreiben, und – wenn ja – in welcher Form. Es könnte damit genau das Gegenteil bewirkt werden, sodass Kunden durch den Kontakt an ihren Versicherungsvertrag „erinnert" werden und diesen kündigen. In diesem Fall könnte man die Erkenntnisse des Modells z. B. dazu nutzen, bestimmte Kundengruppen aus jeglichem zusätzlichen Kundenkontakt auszuschließen.

Unter anderem aus diesem Grund sollte bei Anwendung des Modells bzgl. einer Maßnahme zur Stornoprävention wie dem Kundenkontakt von vornherein das Controlling der Maßnahme bedacht werden. Ein mögliches Vorgehen besteht darin, die gemäß Modell als Stornierer zu kontaktierenden Kunden in eine Experimentgruppe, die tatsächlich zum Stichtag der Modellanwendung kontaktiert wird, und in eine Kontrollgruppe, für die die Maßnahme bewusst unterlassen wird, einzuteilen. Die Aufteilung in beide Gruppen sollte mindestens zufällig, zur gleichmäßigen Verteilung des Stornorisikos besser aber stratifiziert anhand der Stornowahrscheinlichkeiten erfolgen (z. B. alternierend anhand der geordneten Stornowahrscheinlichkeiten). Nach Ablauf des Zeitfensters, das auch retrospektiv für die Definition und Ableitung der Stornozielgröße genutzt wurde, können die eingetretenen Stornoraten in beiden Gruppen verglichen werden und es lässt sich abschätzen, ob die Maßnahme Storno verhindern konnte (oder ggf. durch die falsch positive Ansprache kontraproduktiv gewirkt hat). Die Beibehaltung einer Kontrollgruppe, für die trotz entsprechender Vorhersage keine Maßnahme durchgeführt wird, ist auch vor dem Hintergrund der Weiterentwicklung des Modells wichtig. Um sich ändernde Muster im Stornoverhalten zu identifizieren, ist eine regelmäßige Neukalibrierung anhand aktueller Daten notwendig. Eine Kontrollgruppe erlaubt dafür die Beobachtung von Storno ohne die potentielle Beeinflussung durch eine bereits durchgeführte Maßnahme und bildet somit die Grundlage für die neuerlichen Trainingsdaten.

2.1.7. Weitere Anwendungsfelder

Abschließend werden in diesem Abschnitt zwei weitere Anwendungsfälle von prädiktiven Algorithmen im Bereich der Stornoanalyse motiviert.

Herleitung von Stornotafeln

Die Herleitung einer möglichst geglätteten Stornotafel für Projektionsrechnungen oder die Berücksichtigung von Storno bei der Zinszusatzreserve stellt ein bekanntes aktuarielles Problem dar. Jedoch hängen klassisch hergeleitete Tafeln i. d. R. ausschließlich von der Vertragslaufzeit ab und werden getrennt nach bestimmten Bestandsgruppen (meistens Gruppen von Tarifen) hergeleitet.

Die Ausführungen der vorigen Abschnitte zur Stornomodellierung motivieren, die ausschließliche Abhängigkeit von der Vertragslaufzeit in der Herleitung von Stornotafeln zu hinterfragen und zu prüfen, von welchen Faktoren die Stornowahr-

scheinlichkeit zusätzlich abhängt und wie darauf aufbauend aussagekräftige, geglättete Tafeln abgeleitet werden können. Dies impliziert zwei aufeinander aufbauende Fragestellungen. Zum einen wird ein Modell gesucht, welches die (z. B. drei) wichtigsten Einflussfaktoren liefert. Darauf aufbauend soll zum anderen eine Tafel abhängig von diesen Faktoren hergeleitet werden.

Für den ersten Schritt bieten sich baumbasierte Verfahren (Entscheidungsbaum, Random Forest, Boosting) an, da es hier eine Vielzahl an Darstellungsarten zur Interpretation des jeweiligen Modells gibt (siehe Abschnitt 2.1.3 „Komplexität und Interpretierbarkeit"), die es erlauben, qualitative und leicht verständliche Aussagen abzuleiten. Hierfür ist die detaillierte Optimierung des Modells (vgl. Abschnitt 2.1.4) im Vergleich zur vorigen Fragestellung sicherlich von geringerer Bedeutung. Stattdessen liegt der Fokus klar bei der Interpretation der Ergebnisse und nicht in der Güte des Modells (auch wenn eine gewisse Mindestgüte zur Validität der Erkenntnisse nicht unterschritten werden sollte). Sind die Haupteinflussfaktoren ermittelt, sollte die Frage der Umsetzbarkeit gestellt werden: Kann das Zielsystem (Projektionsmodell, Verwaltungssystem o. ä.) Tafeln abhängig von den ermittelten Faktoren abbilden oder gibt es die Möglichkeit das Zielsystem dahingehend anzupassen? Ist dies nicht möglich, sollten die nicht abbildbaren Faktoren an dieser Stelle schon ausgeschlossen werden, da sie zur Zielerreichung (der Abbildung von Storno in dem Zielsystem) aus technischen Gründen keinen Beitrag mehr leisten können. Eine Alternative wäre es bereits die Datengrundlage um diese Faktoren zu bereinigen. Jedoch gehen dadurch möglicherweise qualitative Aussagen verloren, die sich eventuell anderweitig nutzen lassen. Es bietet jedoch die Chance, dass andere Faktoren in den Vordergrund rücken, die das Stornoverhalten ebenfalls gut abbilden oder ausgeschlossene Einflussfaktoren mit nur wenig Güteverlust substituieren.

Stehen die Einflussfaktoren fest, die für die Herleitung der Tafel verwendet werden sollen, bietet sich die Anpassung eines generalisierten linearen Modells (GLM) an (im Speziellen bspw. ein logistisches oder Poisson-Regressionsmodell, siehe auch Anhang 4.2). Da dieses Modell alle Inputfaktoren bei der Vorhersage der Stornowahrscheinlichkeit berücksichtigt, ist die vorherige Auswahl der Einflussfaktoren sehr wichtig. Das Modell hat den Vorteil, auf einer gemeinsamen Datengrundlage mehrdimensionale Tafeln abzuleiten und gemeinsame Effekte zu berücksichtigen.

Zur Evaluierung des Modells bietet sich ein Backtest an. Ein Vergleich der tatsächlichen Stornoanzahl eines bestimmten Zeitraums (z. B. eines Jahres) mit der erwarteten Stornoanzahl – zum einen mittels einer klassischen, vertragslaufzeitabhängigen Stornotafel und zum anderen der Stornotafel aus dem GLM – zeigt, welche Vorgehensweise das tatsächliche Stornoverhalten besser abbildet. Es sollte hierfür ein Jahr verwendet werden, welches nicht im Herleitungszeitraum der Tafeln enthalten ist.

Eine Evaluierung der Zielerreichung könnte sich dadurch auszeichnen, dass die neue Tafel in das Zielsystem integriert und plausibilisiert wird, ob sich dadurch bessere Prognosen des Stornos ergeben. Das könnte sich beispielsweise daran zeigen, dass dynamische Änderungen der Bestandsverteilung über die Zeit besser

abgebildet werden als mit einer eindimensionalen Tafel, die ein verändertes Stornoverhalten, z. B. auf Grund einer sich ändernden Altersverteilung im Bestand, nicht berücksichtigt. Durch diese besseren Prognosen lassen sich eventuell Sicherheitszuschläge für Modellunschärfen reduzieren bzw. positivere zukünftige Entwicklungen abbilden, wie es vorher nicht möglich war. Letztendlich kann dies zu einer besseren Solvenzquote oder einer niedrigeren Zinszusatzreserve führen und somit direkte wirtschaftliche Auswirkungen haben.

Modellierung von Frühstorno

Wird ein Vertrag in den ersten Jahren seiner Laufzeit storniert, birgt dies für den Versicherer das Risiko, dass die gezahlte Provision nicht vollständig zurückgefordert werden kann. Hierfür ist es entscheidend, Provisionsmodelle zu schaffen, die auch den Vermittler in die Verantwortung nehmen. Dazu ist es wichtig, das Phänomen Frühstorno besser und tiefergehender zu verstehen. Als Fragestellung ergibt sich, von welchen Faktoren Frühstorno abhängt und wie sich dies für die Modellierung von Provisionsmodellen nutzen lässt. Hierbei ist es wichtig den Fokus schon während des Modellierungsprozesses auf die Optimierung der Provisionsmodelle zu richten, sodass das Modell letztendlich auf Vertriebsmerkmale spezialisiert ist.

Dies beginnt mit der Zusammenstellung der Datengrundlage, da bereits hier die Merkmale festgelegt werden, die in das Modell eingehen und somit potenziell berücksichtigt werden können. Dazu muss das Vertriebsdatenmodell im jeweiligen Versicherungsunternehmen analysiert werden, sodass vertriebsspezifische Daten zusammen mit Vertragsdaten genutzt werden können. Es sollten sich auch Gedanken über verschiedene abgeleitete Merkmale (wie beispielsweise die Stornoquote des Vermittlers in der Vergangenheit, seine Abschlussvolumina oder die Entwicklung der Sollsaldos) gemacht werden. Es könnte ebenfalls interessant sein, vorhandene Klassifizierungen von Vermittlern zu nutzen und dem Modell entsprechende Informationen mitzugeben.

Für eine solche eher qualitative Betrachtung (Ziel eines transparenten, regelbasierten Provisionsmodells) bieten sich – wie im vorangegangenen Beispiel – baumbasierte Verfahren an. Um die Einflussfaktoren aus dem Modell zu extrahieren und für die Provisionsmodelle zu nutzen, ist die Visualisierung der Ergebnisse von entscheidender Bedeutung, wohingegen das Modell nicht bis ins letzte Detail optimiert werden muss, solange die qualitative Aussage valide ist.

Zur Validierung der entstandenen Provisionsmodelle bietet sich erneut ein Backtest an. Durch die Simulation der neuen Provisionszahlungen auf den Bewegungen (Neuzugänge, Storno etc.) eines zu definierenden vergangenen Zeitraums lassen sich Aussagen darüber treffen, wie sich die Provisionsbelastung im Vergleich zu dem damals geltenden Modell dargestellt hätte und ob das neue Provisionsmodell eine Verbesserung darstellen würde. Eine Testphase, in der das neue Provisionsmodell mit allen oder einem Teil der Vermittler getestet wird, stellt sich vermutlich schwierig dar, da die mit dem Vermittler ausgehandelten Provisionsvereinbarun-

gen rechtlich bindend und ständige Neuregelungen für eine vertrauensvolle Zusammenarbeit vermutlich nicht förderlich sind. Eine Evaluierung der Zielerreichung sollte jedoch trotzdem durchgeführt werden. Hierbei könnte eine Beobachterrolle eingenommen werden, in der unerwartete Anpassungen des Marktes (sprich: der Vermittler) durch die neuen Provisionsmodelle beurteilt werden. Diese Erkenntnisse sollten dann auch in eine Neukalibrierung der Stornomodelle eingehen und letztendlich bei Neuverhandlungen von Provisionsregelungen Berücksichtigung finden.

2.2. Modellierung Gesundheitszustand

Die Modellierung der Gesundheit des Versicherungsnehmers im Rahmen einer Lebensversicherung betrifft den Kern des Absicherungsversprechens und der Kalkulation. Nach Vertragsabschluss führt eine definierte Statusänderung der Gesundheit zum vereinbarten Leistungsversprechen, sei es z. B. in der Todesfallabsicherung, Invaliditäts- oder auch Pflegeversicherung.

Angelehnt an die verschiedenen Phasen eines Lebensversicherungsvertrages gibt es somit mehrere Phasen und Zeitpunkte, in denen die Modellierung der Gesundheit wichtig ist. Dies sind insbesondere:

- Werbung / Vertriebsansprache: Kunden, denen aufgrund ihres schlechten Gesundheitszustandes von vornherein keine Lebensversicherung angeboten werden kann, sollten erst gar nicht angesprochen werden.
- Vertragsabschluss / -anbahnung: Die klassische Risikoprüfung zur Entscheidung über Vertragsabschluss und Vertragskonditionen.
- Vertragslaufzeit, kein Leistungsfall: Jeder Kunde wird einen gewissen Verlauf seines Gesundheitsstatus erleben. In Abhängigkeit von vorhandenen Informationen über den Kunden können Vertragsanpassungen / Maßnahmen ergriffen werden, z. B. individuelle Nachversicherungsmöglichkeiten.
- Leistungsfall: Die Entscheidung über einen Leistungsfall in Abhängigkeit von den zur Verfügung gestellten Informationen (z. B. Art der gesundheitlichen Einschränkung) und des Leistungsauslösers der Police.
- Weiterer Verlauf des Leistungsfalls: Üblicherweise erlischt der Leistungsanspruch, z. B. auf Leistungen bei Berufsunfähigkeit (BU), mit Wegfall der Leistungsvoraussetzungen.

Verfahren aus Big Data und Künstlicher Intelligenz können zu jedem Zeitpunkt unterstützen. Dies kann eine Prozessvereinfachung, eine bessere Entscheidungsqualität oder mehr Service für den Kunden bedeuten.

Im Folgenden werden Big-Data-Möglichkeiten anhand der Beispiele „Underwriting – Kürzung der Gesundheitsfragen", „Prognose von Leistungsfällen / Reaktivierungsscoring", „Prädiktion von Invalidität und Sterblichkeit" sowie „Underwriting – Änderung der Datenbasis" skizziert.

Die Beschreibung der operativen Schritte der Modellierung ist hierbei deutlich kürzer gefasst als im vorhergehenden Abschnitt, da sich die grundsätzlichen Vorgehensweisen stark ähneln.

2.2.1. Underwriting – Kürzung der Gesundheitsfragen

Der Abschluss einer Lebensversicherung ist im Vergleich zu vielen anderen Kaufentscheidungen ein überdurchschnittlich komplexer Prozess. Die Gesundheitsfragebögen, in denen Kunden bei biometrischen Produkten Angaben zu ihrem Gesundheitszustand machen müssen, tragen wesentlich zu dieser Komplexität bei. Hier gilt es für Versicherer zwei Zielsetzungen gegeneinander abzuwägen:

Ein umfangreicher Fragebogen verbessert die Einschätzung des zu versichernden Risikos und ermöglicht dem Versicherer die Berechnung einer auskömmlichen Prämie. Der Kunde möchte jedoch möglichst wenige, einfache und angenehme Fragen beantworten.

Möchte ein Unternehmen seinen Gesundheitsfragebogen kürzen, muss es also prüfen, auf welche Fragen verzichtet werden kann, ohne dass dadurch die Güte der Risikoeinschätzung und die resultierende Portfolio-Entwicklung wesentlich negativ beeinträchtigt werden. Machine-Learning-Verfahren können dabei helfen, solche Fragen zu identifizieren.

Datengrundlage

Die wichtigste Datenquelle für diese Analyse sind im Unternehmen vorliegende, von Kunden ausgefüllte Gesundheitsfragebögen mitsamt der zugehörigen Annahmeentscheidung des jeweiligen Risikoprüfers. Neben Gesundheitsangaben sollten auch andere erhobene Merkmale wie z. B. das Alter, der Beruf, das Einkommen oder die gewünschte Versicherungssumme berücksichtigt werden.

Oft werden Gesundheitsangaben in einem abgestuften Verfahren erhoben, indem z. B. bei Vorliegen bestimmter Vorerkrankungen Detailfragebögen nachgesendet werden. Je nach konkreter Fragestellung sollten der allgemeine Fragebogen und die Detailfragebögen getrennt analysiert werden.

Während die einzelnen Gesundheitsfragen und weiteren Merkmale die erklärenden Variablen darstellen, ist die Annahmeentscheidung des jeweiligen Risikoprüfers in diesem Kontext als abhängige Variable anzusehen.

Es sei hier angemerkt, dass das Erreichen der o. g. Datengrundlage nicht unterschätzt werden sollte. Für den Bestand liegen die Dokumente oft nicht maschinennutzbar vor. Und auch für das Neugeschäft ist eine systematische und strukturierte Erfassung durch konsequent digitale Prozesse noch bei weitem nicht überall der Fall.

Die Validierung der Daten und deren explorative Analyse sind wesentlicher Teil der Modellierung. Da sich die Vorgehensweise nicht grundsätzlich von der Vorgehensweise bei Storno unterscheidet, siehe hierzu Abschnitt 2.1.2.

Modelle

Für die Analyse der Optimierungsmöglichkeiten des Gesundheitsfragebogens wird die Annahmeentscheidung modelliert. Dabei gibt es verschiedene Modellierungsansätze. In diesem Bericht wird auf die naheliegende Modellierung als Klassifizierungsaufgabe eingegangen. Im einfachsten Fall beschränkt man sich auf die zwei Klassen „Annahme" und „Ablehnung". Es ist jedoch auch denkbar, statt der Klasse „Annahme" eine feinere Unterscheidung nach Ausschlüssen oder Aufschlägen zu verwenden. Manche Unternehmen fassen dagegen Ablehnungen, Ausschlüsse und Aufschläge in einer Klasse zusammen.

Zur Modellierung der Annahmeentscheidung sollten Modelle verwendet werden, für die die Relevanz der einzelnen Merkmale berechnet werden kann (auch „feature importance" oder „variable importance" genannt, siehe Kapitel 4 sowie für Hinweise zur Berechnung Abschnitt 2.3.2). Die feature importance eines Merkmals gibt an, wie sehr dieses das Modellergebnis – in diesem Fall die Annahmeentscheidung – beeinflusst. In diesem Kontext sind die Merkmale gerade die Fragen des Gesundheitsfragebogens. Merkmale bzw. Fragen mit geringem Einfluss auf die Annahmeentscheidung sind geeignete Kandidaten zur Kürzung des Gesundheitsfragebogens.

Angesichts der Anforderungen an das Modell bieten sich z. B. die folgenden Modellklassen an:

- Baumverfahren (Random Forests oder Boosted Trees)
- Generalized Linear Models (GLMs, insb. logistische Regression)

Diese Modellklassen eignen sich, da sich für sie einerseits die Relevanz der einzelnen Merkmale berechnen lässt und sich mit ihnen andererseits Klassifikationsprobleme modellieren lassen. Weitere Hinweise zu diesen Modellklassen und zur Modellauswahl enthält Kapitel 4.

Hinweise zu Ergebnissen und möglicher Nutzung

Wie bereits erwähnt, ist für die Analyse vor allem interessant, welche Fragen nur geringen Einfluss auf die Annahmeentscheidung haben. Eine geringe feature importance ist dafür ein guter Indikator. Um diese zu bestimmen, sollte das ausgewählte Modell auf Basis der bestehenden Gesundheitsfragebögen und entsprechenden Annahmeentscheidungen trainiert werden. Anschließend sollte eine Feinjustierung der Hyperparameter des Modells erfolgen. Für dieses adjustierte Modell kann nun die feature importance der einzelnen Fragen ausgegeben werden.

Die Methode liefert jedoch keine direkte Entscheidung, welche Fragen tatsächlich aus den Gesundheitsfragebögen entfernt werden können. Diese Entscheidung müssen die Experten des Versicherungsunternehmens selbst treffen – in Abwägung der Vereinfachung des Fragebogens einerseits und der damit einhergehenden Ungenauigkeit in der Risikoprüfung andererseits. Die folgenden vertieften Analysen können die Abwägung unterstützen:

- Wie ist die Vorhersagegüte des Modells?

- Wie hängt die Annahmeentscheidung von einzelnen Fragen des Fragebogens ab? (univariate Analyse)
- Welchen Effekt hätte eine Kürzung des Fragebogens auf die Qualität der Risikoeinschätzung?

Diese Analysen werden nachfolgend kurz beschrieben:

Es ist hilfreich, die Vorhersagegüte des trainierten Modells zu prüfen, auch wenn das Modell nicht tatsächlich für die Vorhersage der Annahmeentscheidung zum Einsatz kommen soll. Die Vorhersagegüte kann z. B. anhand der Konfusionsmatrix oder der Area under the Curve (AUC) geprüft werden. Zeigt sich anhand dieser Maße, dass das trainierte Modell für die Vorhersage der Annahmeentscheidung nicht geeignet ist, so ist auch die berechnete feature importance wenig aussagekräftig. In der Praxis zeigte sich, dass gute Modelle durchaus eine AUC im Bereich von 90 bis 95 Prozent erreichen können.

Wurden nicht relevante Fragen identifiziert, kann weiter plausibilisiert werden, ob im Gesundheitsfragebogen auf diese Fragen verzichtet werden kann. So kann durch eine univariate Analyse für jede Frage geprüft werden, wie sich die Annahmeentscheidung in Abhängigkeit von der Antwort des Kunden ändert. Ist beispielsweise die Anzahl von Annahmen und Ablehnungen über alle Antwortmöglichkeiten hinweg ähnlich verteilt, so ist dies ein weiterer Hinweis darauf, dass die Frage nur einen geringen Einfluss auf die Annahmeentscheidung hat.[10]

Ebenso kann getestet werden, wie ein Verzicht auf bestimmte Fragen die Güte der Risikoprüfung beeinflussen würde. Dazu entfernt man aus den Trainingsdaten alle Informationen, die sich aus den vermeintlich irrelevanten Fragen ergeben. Mit diesen reduzierten Trainingsdaten trainiert man dann ein zweites Modell. Anschließend kann man das zweite Modell mit dem ersten Modell vergleichen, welches auf Basis der vollständigen Informationen trainiert wurde. Gibt es nur kleine Unterschiede in der Vorhersagegüte beider Modelle, spricht dies dafür, dass auf diese Fragen verzichtet werden könnte.

Es sei darauf hingewiesen, dass durch diese analytischen Betrachtungen nicht alle Aspekte der Gesundheitsfragebögen erfasst werden können. Aus diesem Grund sollte das Fachwissen von Experten aus verschiedenen Bereichen (z. B. Risikoprüfung oder Vertrieb) hinzugezogen werden. Das Unternehmen sollte sich beispielsweise zusätzlich folgende Fragen stellen:

- Wie stark ist das zu versichernde Risiko in den Fällen erhöht, welche ohne eine bestimmte Frage angenommen, mit dieser Frage jedoch abgelehnt würden?

[10] Der Umkehrschluss gilt im Übrigen nicht: Auch wenn die univariate Analyse ergibt, dass die Anzahl von Annahmen und Ablehnungen stark von der Antwort auf eine bestimmte Frage des Gesundheitsfragebogens abhängt, folgt daraus nicht unbedingt, dass die Frage auch relevant ist. Es kann sein, dass durch andere Fragen des Gesundheitsfragebogens im Kern die gleiche Information erfasst wird und eine der beiden Fragen damit verzichtbar ist. Solche Abhängigkeiten zwischen einzelnen Merkmalen werden jedoch bei einer univariaten Analyse nicht erfasst.

- Kann der Vertrieb (Makler) zu starker Antiselektion bzgl. eines Merkmals führen?

Praktische Umsetzung und Controlling

Angesichts der Bedeutung der Gesundheitsfragen für das Risikoprofil des Bestandes ist es notwendig, die Auswirkungen einer Anpassung des Gesundheitsfragebogens fortlaufend zu überprüfen. So kann z. B. analysiert werden, ob sich die angesprochene Kundengruppe ändert: Gibt es Verschiebungen bei den Berufsklassen, bei der Höhe der Absicherung oder bei Gesundheitsfaktoren wie dem BMI oder den Antworten auf den Gesundheitsfragebogen?

Ein besonderes Augenmerk sollte auch auf der Entwicklung der (Früh-)Schäden liegen. Unternehmen sollten regelmäßig prüfen, ob es eine Verbindung zwischen den auftretenden Frühschäden und den fehlenden Fragen im Gesundheitsfragebogen gibt.

2.2.2. Prognose von Leistungsfällen / „Reaktivierungsscoring"

In der Leistungsprüfung der Berufsunfähigkeitsversicherung werden zwei wesentliche Fragen bearbeitet: Erfüllt ein gemeldeter Leistungsfall die Definition des Leistungsauslösers? Ist ein bestehender Leistungsfall weiterhin ein Leistungsfall oder liegt eine Reaktivierung vor (bzw. wird diese erwartet)?

Data Analytics kann für beide Fragestellungen wertschaffend eingesetzt werden. Bei der Ersteinschätzung eines Leistungsfalls können Prozesse automatisiert werden und so effiziente, standardisierte und qualitativ hochwertige Entscheidungen herbeigeführt werden.

Bei bestehenden Leistungsfällen kann durch Data-Analytics-Methoden eine Einschätzung erfolgen, welche Leistungsfälle im Vergleich zu anderen Leistungsfällen mit einer größeren Wahrscheinlichkeit bis zum Vertragsende oder innerhalb eines definierten Zeitraumes reaktiviert werden können – auch „Reaktivierungsscoring" genannt. Der Fokus liegt dabei auf jedem Einzelfall. Damit lässt sich beispielsweise eine datenbasierte Priorisierung der Nachprüfungen einführen. Es resultiert ein effektiverer Einsatz der Ressourcen im Schadenmanagement und im besten Falle eine erhöhte Reaktivierungsquote.

Im Folgenden konzentrieren wir uns auf die Erstellung eines Reaktivierungsscorings.

Datengrundlage

Bevor auf konkrete relevante Datenquellen eingegangen wird, gilt zunächst der Hinweis, dass in vielen Unternehmen diese notwendigen Daten nicht oder nur sehr unvollständig digital nutzbar vorliegen. Daraus ergibt sich eine Notwendigkeit für eine Erweiterung der Datenbasis. Die Digitalisierung eröffnet hierbei neue Möglichkeiten. Beispielsweise gibt es erste neue Softwareangebote, mit denen sich komplett digitale regelbasierte Entscheidungsvorgänge in der Leistungsprüfung darstellen lassen. Diese Softwareangebote sind z. B. so gestaltet, dass im Rahmen der operativen Prozesse in der Leistungsprüfung die relevanten Daten direkt strukturiert digital erfasst werden – und damit später für die Modellierung nutzbar werden. Ohne eine erweiterte Datenbasis lässt sich das Potenzial von Big Data und Machine Learning in der Leistungsprüfung nicht heben.

Eine Besonderheit bei der Modellierung von Leistungsfällen ist weiterhin, dass auch bei vollständiger Verfügbarkeit der Daten innerhalb eines Lebensversicherungsunternehmens die Anzahl der beobachteten Fälle gemessen an typischen Beispielen von Big-Data-Anwendungen nicht sehr groß ist, da Reaktivierungen nur innerhalb des relativ kleinen Bestandes der Invaliden auftreten.

Grundlage für Erkenntnisse, die über die „traditionellen Rechnungsgrundlagen" und die damit verbundene Portfoliosichtweise hinausgehen, sind umfassendere Daten. Hierbei sind verschiedene Datenquellen denkbar – grundsätzlich ist eine möglichst breite Datenbasis wünschenswert. Verschiedene Einschränkungen in der

Praxis (Kosten, IT, Digitalisierungsgrad etc.) können zur Priorisierung der genutzten Daten zwingen. Die im Folgenden genannten Datenfelder sind nicht vollständig, sondern stellen eine Auswahl dar und dienen der Illustration:

- Bestandsdaten:
    - Nutzung aller verfügbaren Variablen, z. B. Alter im Leistungsfall, Alter bei Vertragsabschluss, Geschlecht, Beruf bei Vertragsabschluss, Selbstständige / Angestellte / Beamte, Rentenhöhe, Einzel-/ Gruppengeschäft (arbeitnehmer-/arbeitgeberfinanziert), …
- Daten aus der Schadenprüfung:
    - zum Zeitpunkt des Leistungsantrags: Beruf, Jahreseinkommen, BMI, Raucherstatus, …
    - Daten bei Bearbeitung des Leistungsantrags: befristetes Anerkenntnis, medizinische Ursache einer Berufsunfähigkeit, Zeiten der Arbeitsunfähigkeit im Vorfeld der Berufsunfähigkeit, Dauer der Leistungsprüfung, …
    - expertengetriebene Einschätzung zur Prognose des Leistungsfalls
    - BU-Grad (falls vorhanden) oder qualitative Einschätzung der Eindeutigkeit eines BU-Leistungsfalls
- Daten aus dem Underwriting (UW):
    - UW-Entscheidung: Ausschlussklausel, Zuschlag (medizinisch / Freizeit), …
    - Weitere UW-Daten: Vorerkrankungen, ggf. genauer Wortlaut der Antworten auf einzelne Fragen im UW
- Externe Daten:
    - Die VU-internen Daten können beispielsweise mittels regionaler Informationen durch externe Daten angereichert werden. Entscheidend ist hierbei, dass die Matching-Kriterien möglichst spezifisch sind und nicht zu große Personengruppen als Basis für das Matching dienen. So ist z. B. bei einer Bestimmung des sozialen Status über die geografische Information „Bundesland" wenig Aussagekraft zu erwarten. Während bei detaillierteren Adressinformationen durchaus relevante Erklärungskraft gefunden werden kann.
- Weitere interessante Daten sind z. B. Vertriebsdaten.

Modellüberlegungen

Neben Reaktivierungen beeinflusst auch der Invalidentod insgesamt den Verlauf des Invalidenbestandes. Dieser soll aber im Folgenden keine Rolle spielen.

Die aktuellen DAV-Tafeln zur Reservierung einer Berufsunfähigkeitsversicherung enthalten Angaben zu Reaktivierungswahrscheinlichkeiten in Abhängigkeit vom

Eintrittsalter der Berufsunfähigkeit, vom Geschlecht und von der bisherigen Leistungsdauer. Damit lässt sich für ein Kollektiv von Invaliden die Anzahl der erwarteten Reaktivierungen berechnen, ohne dass ein Hinweis darauf gegeben wird, welcher Leistungsfall individuell mit eher hoher oder eher niedriger Wahrscheinlichkeit reaktivieren wird. Es wird also das gesamte Portfolio der Leistungsfälle modelliert.

Im Gegensatz zu den Tafeln zur Reaktivierung soll in dem hier beschriebenen Use Case für jeden Leistungsfall individuell ein Scorewert bzw. eine Reaktivierungswahrscheinlichkeit geschätzt werden. Dafür sollen auch die genannten zusätzlichen Informationen und Daten wie z. B. die Rentenhöhe, die Berufsgruppe resp. der Beruf oder die medizinische Ursache der Berufsunfähigkeit berücksichtigt werden. Die Herleitung von klassischen Reaktivierungstafeln für diese Vielzahl an Inputvariablen ist dabei i. A. nicht notwendig und auch nicht möglich. Ziel ist vielmehr, auf der Grundlage der Beobachtungen das individuelle Reaktivierungsverhalten möglichst gut abzubilden und in einem Scorewert zusammenzufassen.

Hinweise zur Modellauswahl

Zwar ist die Anzahl von Reaktivierungen innerhalb eines durchschnittlichen Unternehmens eher überschaubar aufgrund der geringen Anzahl von Leistungsfällen. Dafür liegen die Reaktivierungswahrscheinlichkeiten für typische Alter im niedrigen einstelligen Prozentbereich[11]. Über die zeitliche Entwicklung ergibt sich damit insgesamt dennoch ein signifikanter Abbau des Invalidenbestandes durch Reaktivierung in einer Größenordnung von 25% bis 50%. Damit sind alle klassischen und neueren Klassifikationsverfahren grundsätzlich für die hier beschriebene Aufgabe geeignet. Im Gegensatz zu seltenen Ereignissen wie Tod oder Invalidität muss bei dem Reaktivierungsscoring nicht auf sehr geringe Wahrscheinlichkeiten für die interessierenden Klassen geachtet werden. Stattdessen ist die Datengrundlage der zu betrachtenden Leistungsfälle in durchschnittlichen Versicherungsunternehmen eher klein, aber für die Herleitung eines Reaktivierungsscorings meistens ausreichend, solange ein hinreichend langer Zeitraum ausgewertet werden kann.

Eine klassische Auswahl geeigneter Verfahren wäre demnach:

- Baumverfahren (Random Forests oder Boosted Trees),
- Diskriminanzanalyse / Generalized Linear Models (GLM, insb. logistische Regression),
- K-Nearest-Neighbor,
- Generalized Additive Models (GAM).

Bestehende Reaktivierungstafeln der DAV können in geeigneter Weise einfließen, z. B. bei Bayes-Verfahren und in den Methoden der Credibility Theory.

[11] Hinweis „Neue Rechnungsgrundlagen für die Berufsunfähigkeitsversicherung DAV 1997", Deutsche Aktuarvereinigung, 2018, Erstveröffentlichung in der DAV-Mitteilung Nr. 11, 1997; Ergebnisbericht: „Überprüfung der DAV 1997 I für Berufsunfähigkeitsversicherungen", DAV, 2013, und Überprüfung Ergebnisbericht, DAV, 2018

Zusätzlich kann der Verlauf des Portfolios der Invaliden auch als klassische Lebensdaueranalyse modelliert werden. Insofern können auch die Methoden der Survival Analysis, evtl. erweitert um neuere Ansätze wie Baumverfahren, zur Anwendung kommen.

Durchführung und Hinweise zu Ergebnissen und möglicher Nutzung

Dieser Anwendungsfall kann in zwei Komponenten separiert werden:

- Bestimmung des Reaktivierungsscores,
- Steuerung der Mitarbeiter im Claims Management durch Priorisierung der Nachprüfungen.

Zur **Bestimmung des Reaktivierungsscores** ist es empfehlenswert, die Analyse in zwei Schritte zu unterteilen: eine explorative Datenanalyse und eine anschließende Analyse mittels moderner statistischer Verfahren wie oben beschrieben.

Schritt 1: Eine explorative Datenanalyse ist unerlässlich, um einerseits ein Datenverständnis zu gewinnen und andererseits Hypothesen für eine zielgerichtete Modellierung in Schritt 2 zu erstellen. Aus herkömmlichen Betrachtungen lassen sich beispielsweise schon sehr interessante Erkenntnisse über die typischen Invaliditätsdauern ableiten, wie exemplarisch am folgenden Beispiel für den Zusammenhang von Invaliditätsdauern und Leistungsursachen gezeigt:

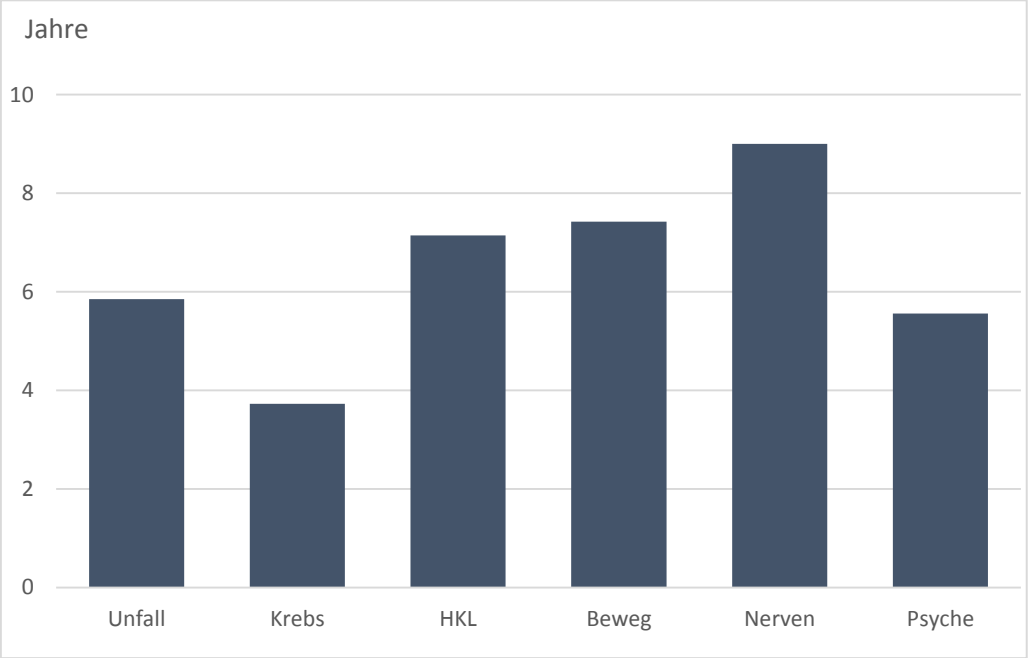

Abbildung 4: Schadendauern: Einfluss der Schadenursache auf Schadendauer. Anzahlgewichtet, geschlossene Fälle, Männer.
Quelle: Rückversicherungsdatenpool

Schritt 2: Die Modellierung mittels weitergehender Methoden erlaubt eine genauere und insbesondere multivariate Bestimmung der Risikofaktoren und ihres Einflusses auf die Reaktivierungswahrscheinlichkeit.

Bei der **Steuerung der Ressourcen der Leistungsmitarbeiter** auf Basis eines wie oben beschriebenen Modells gibt es grundsätzlich zwei Vorgehensweisen:

- Direkte Nutzung des Scores und Abstimmung mit Claims-Experten, wo ein Einsatz der Nachprüfungen am gewinnbringendsten ist.
- Vergleich des Reaktivierungsscores einer Police mit einem Benchmarking-Score für diese Police. Dieser Benchmarking-Score kann beispielsweise auf Basis von Markterfahrung oder auf Basis von Erfahrungen in dem eigenen Portfolio erstellt werden: Wird eine große Reaktivierungswahrscheinlichkeit für eine Police auf Basis des Modells vorhergesagt, aber keine Reaktivierung beobachtet, so könnte eine Nachprüfung ausgelöst werden bzw. eine Empfehlung an den Schadenbearbeiter erfolgen. Die notwendige Markterfahrung für diese Vorgehensweise ist ein weiterer Grund (neben den o. g. notwendigen Fallzahlen) für die Bildung von Pools zum Vorteil aller Marktteilnehmer.

2.2.3. Prädiktive Modelle zu Sterblichkeit und Invalidität

Zunehmende Granularität in der Risikomodellierung im Allgemeinen bzw. beim Pricing im Speziellen und zunehmende Datenverfügbarkeit durch Digitalisierung sind wesentliche Treiber für den Einsatz von Big-Data- und Analytics-Techniken bei der Bestimmung von Mortalität und Morbidität: Durch Data-Analytics-Techniken lässt sich eine simultane Betrachtung aller Risikofaktoren inkl. einer „optimalen" Auswahl der relevantesten Faktoren realisieren. Diese Erkenntnisse lassen sich in granularere Best Estimates umsetzen. Diese wiederum ermöglichen granularere Erkenntnisse beispielsweise im Geschäftsmonitoring, in der Risikomodellierung (Solvency II) und beim Pricing. Letzteres wiederum kann die Basis für die Erschließung attraktiver Kundensegmente sein.

Datengrundlage

Grundsätzlich lässt sich festhalten, dass eine möglichst breite Datenbasis wünschenswert ist (s. Anhang 2: Informatik und Tools). Die Identifikation der signifikanten Risikofaktoren erfolgt dann innerhalb der Modellierung. Je breiter die Datenbasis wird, desto weniger Schadenfälle pro Merkmalskombination sind jedoch zu erwarten. Deshalb sind hier – noch mehr als in den „klassischen" Tafelableitungen – Marktpools wichtig und zum Vorteil aller Teilnehmer.

Modelle

Bei der Vorhersage von Sterblichkeiten und (BU-)Inzidenzen sind primär Verfahren des überwachten Lernens relevant. Klassische parametrische Verfahren wie GLMs haben in diesem Kontext den Vorteil in der Tendenz transparent und gut verständlich zu sein, während nicht-parametrische Verfahren wie Random Forests in der Tendenz robust und flexibel sind, d. h. sich den Strukturen besser anpassen, die den Daten zugrunde liegen. Dies erhöht häufig die Qualität der Vorhersagen.

Prädiktive Modelle können auch auf externen bzw. „neuen" Daten genutzt werden, um für ein Versicherungsunternehmen relevante Ergebnisse zu liefern. Zum Beispiel lässt sich zeigen, dass die Anzahl der Schritte pro Tag unter Berücksichtigung von Alter, Geschlecht, und Raucherstatus deutliche Auswirkung auf die Sterblichkeit hat („high" u. ä. Bezeichnungen beziehen sich in der Folgegrafik auf die Variable „Schritte pro Tag").

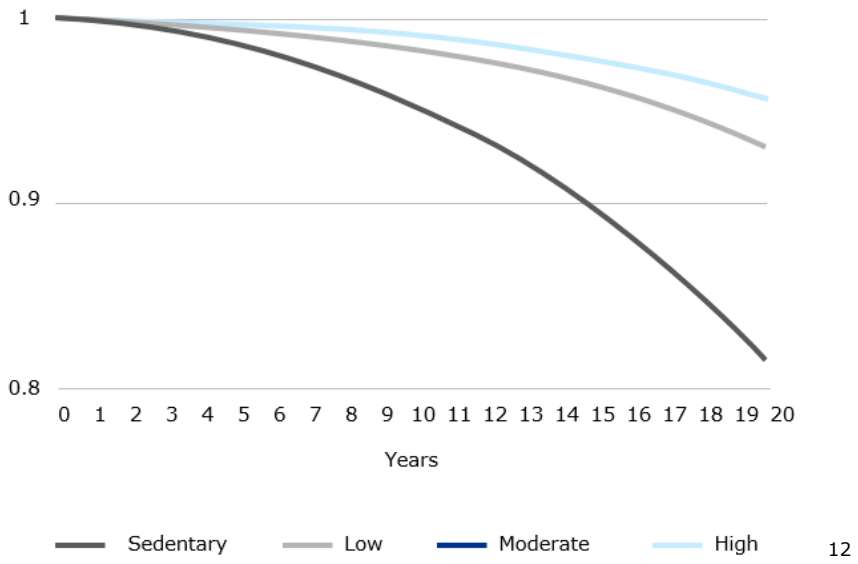

Abbildung 5: Predicted Survival Curve 45-year-old Female Non-smoker https://www.munichre.com/site/marclife-mobile/get/documents_E-889788279/marclife/assset.marclife/Documents/Publications/Stratifying_Risk_Using_Wearable_Data.pdf

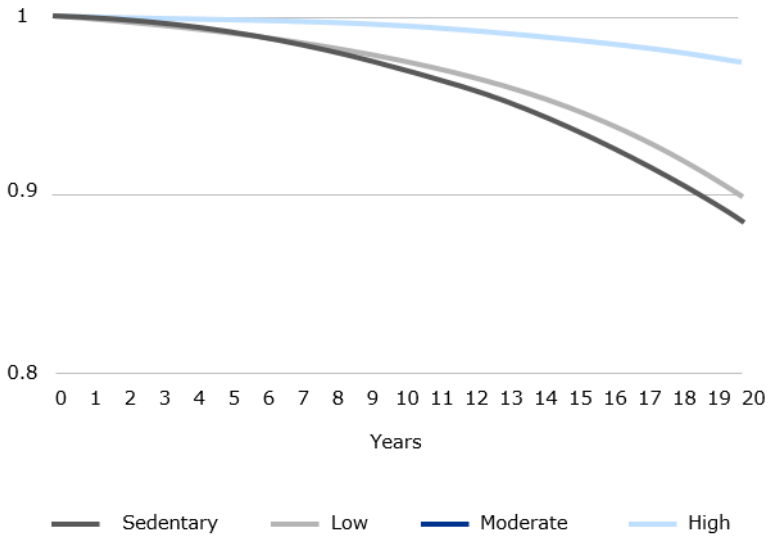

Abbildung 6: Predicted Survival Curve 35-year-old Male Non-smoker

[12] Die hier nicht sichtbare „Moderate"-Kurve liegt unter / hat denselben Verlauf wie die „High"-Kurve, ebenso in Abbildung 6.

Hinweise zu Ergebnissen und möglicher Nutzung

Bei der Anwendung prädiktiver Modelle ist eine sorgsame Interpretation und Überprüfung der Modelle durch Fachexperten erforderlich. Die Verwechslung von Korrelation mit Kausalität kann teuer werden, z. B. wenn sich der Geschäftsmix verschiebt – Kunden können „arbitrieren", beispielsweise durch Wechsel der E-Mail-Adresse eine Prämienreduktion hervorrufen. In der Praxis ist der Übergang von einem „rein" statistischen Modell zu einem Anwendungsmodell nicht zu unterschätzen und verlangt nach (aktuariellem) Expertenwissen. Unter anderem müssen berücksichtigt werden:

- Geschäftsrelevanz,
- Qualitätssicherung,
- Daten / technische Einschränkungen,
- (aufsichts-)rechtliche Aspekte und Reputation.

2.2.4. Underwriting – Änderung der Datenbasis

Bisher ist in der Lebensversicherung der Kunde selbst die primäre Datenquelle für das Underwriting. Risikorelevante Daten liegen jedoch auch bei einer Vielzahl anderer Institutionen vor. Eine Reihe von Unternehmen bietet beispielsweise Cloudumgebungen für medizinische Daten an (z. B. elektronische Patientenakten). Daten von Krankenkassen, Fitnesstrackern, Banken oder aus Social-Media-Plattformen können ebenfalls Rückschlüsse auf die Lebens- und Gesundheitssituation des Kunden zulassen. Sogar Behörden (z. B. die Finanzämter) verfügen für das Underwriting über relevante Informationen wie Alter, Einkommen oder Beruf. Bei einer entsprechenden Integration dieser externen Datenquellen könnte einerseits die Erfassung beim Kunden entfallen oder reduziert werden und damit der Antragsprozess verschlankt werden. Andererseits könnten sich andere Daten als bessere Risikoprädiktoren erweisen als die bisher erfassten Merkmale.

Eine große Hürde bei der Anbindung dieser externen Datenquellen sind die rechtlichen Rahmenbedingungen. Sollen neue Informationen bzw. Merkmale für die Risikobewertung verwendet werden, stehen Unternehmen im Falle von Modellen des überwachten Lernens zudem vor der Herausforderung, Trainingsdaten herzuleiten. Denn diese Trainingsdaten müssen bereits eine Abbildung der erweiterten Datensätze auf die Zielvariable enthalten. Erste Unternehmen versuchen bereits bestimmte Kundenangaben aus anderen Quellen abzuleiten, z. B. indem aus Kundenbildern auf das Alter und den BMI des Kunden geschlossen wird. Langfristig könnte es sich jedoch als zielführender erweisen, solche neuen Datenquellen direkt zur Risikobewertung heranzuziehen, anstatt mit ihnen etablierte Risikomerkmale zu schätzen. So könnten beispielsweise Bilder durch das Erscheinungsbild der Haut oder der Augen bessere Rückschlüsse auf die Gesundheit zulassen als das Alter oder der BMI.

2.3. Risikomanagement / Projektion

In diesem Abschnitt werden ausgewählte Anwendungsmöglichkeiten von Big-Data- bzw. Machine-Learning-Verfahren im Kontext des Risikomanagements bzw. aktuarieller Projektionen diskutiert.

Dazu wird der aktuarielle Modellierungsprozess in die drei Teilschritte

1) Modellinput und dessen Erzeugung
2) Modellprojektionen
3) Modelloutput und Validierung

zerlegt, für die jeweils Anwendungsfälle diskutiert werden. Hierbei ist anzumerken, dass der Prozessschritt 2) nicht völlig losgelöst von den anderen beiden Prozessschritten betrachtet werden kann.

2.3.1. Einordnung und allgemeine Bemerkungen

Im Folgenden werden wir uns zunächst auf das überwachte Lernen fokussieren und als Beispiel Projektionsergebnisse eines aktuariellen Projektionsmodells vorhersagen. Anwendungen des unüberwachten Lernens werden anschließend ebenfalls besprochen, z. B. die Gruppierung von Modellpunkten in der Modellinput-Verarbeitung.

Bemerkungen zum überwachten Lernen und Abgrenzung zu bisherigen Anwendungen des Machine Learning außerhalb des Aktuariats

In den meisten Anwendungsfällen des Machine Learning außerhalb des Aktuariats werden die zu untersuchenden Daten aus „realen" Prozessen erzeugt. Beispiele in der Versicherungsbranche sind z. B. die gesammelten Daten, wenn ein Versicherungsnehmender (VN) den Versicherungsvertrag abschließt (oder auch nicht) oder kündigt (siehe z. B. auch der Abschnitt zur Stornoanalyse). Diese Prozesse haben gemeinsam, dass die unterliegenden Entscheidungsregeln, ob ein VN einen Vertrag abschließt oder kündigt, nicht bekannt sind. Um die zugrundeliegenden Zusammenhänge zu analysieren, Regeln zu extrahieren und diese anschließend für Vorhersagen anzuwenden, können maschinelle Lernverfahren auf die vorliegenden Datensätze angewandt werden. An dieser Stelle könnte man die gelernten Modelle ausschließlich als Vorhersagemodelle verstehen, die mit Inputdaten beliefert werden und anschließend zu einer Vorhersage kommen.

Im folgenden Abschnitt soll nun auch diskutiert werden, wie man Machine Learning auf Daten anwenden kann, die aus Prozessen stammen, deren Regeln vordefiniert sind. Versteht man dann das gelernte Modell als reines Prognosemodell, so ist der Nutzen (abgesehen von potentiellen Geschwindigkeitsvorteilen) des Modells fragwürdig. Im iterativen Prozess der Vorhersagemodelloptimierung werden wir aber diskutieren, dass man derartige Modelle durch geeignetes Befragen dazu bringen kann, die gelernten Regeln bzw. Zusammenhänge preiszugeben. So kann man das gelernte Modell als Tool sehen, welches die unterliegenden Zusammenhänge zwi-

schen dem Modell-Input und dem Modell-Output erlernt. Diese Regeln können anschließend durch geeignete Verfahren abgefragt und mit den eingestellten Regeln im aktuariellen Projektionsmodell verglichen werden. Durch diese Herangehensweise können verschiedene Fragen diskutiert werden:

- Welches sind die Haupttreiber, die die Outputgröße maßgeblich beeinflussen?
- Wie können Inputgrößen codiert bzw. vereinfacht werden, ohne dass die Vorhersagequalität darunter leidet? (Reicht es beispielsweise aus, die 10-jährige risikofreie Spotrate zu kennen oder muss die ganze Zinsstrukturkurve bekannt sein?)
- Sind die gelernten Regeln so, wie man sie erwartet, und führen diese zu den erwarteten Wirkungen im Modell? (Eine der Kernfragen einer jeden Modellvalidierung)

Genau diese Fragen möchten wir im Folgenden im Kontext der Validierung und Beschleunigung eines aktuariellen Modells diskutieren.

2.3.2. Einsatz von Machine Learning zur Erzeugung von Projektionsergebnissen und Validierung

In diesem Abschnitt diskutieren wir, wie Machine Learning eingesetzt werden kann, um aktuarielle Vorhersagen zu treffen, dadurch aktuarielle Projektionen zu beschleunigen und aktuarielle Modelle zu validieren.

Dazu soll der vorliegende Datensatz bestehen aus

- dem Input eines aktuariellen Modells, d. h. der Datensatz beinhaltet Informationen über den Aktiv- und Passivbestand sowie ggf. Managementregeln[13], und
- dem Output des Modells wie z. B. dem ökonomischen Eigenkapital nach Solvency II (SII) oder einer anderen Bilanzgröße.

Im Gegensatz zu „klassischen" Datensätzen, die z. B. für Storno- oder Kaufwahrscheinlichkeitsvorhersagen genutzt werden, handelt es sich somit hier um einen Datensatz, der synthetisch, d. h. durch den Einsatz eines Modells, erzeugt worden ist. Der Unterschied zu anderen, „klassischen" Datensätzen ist also, dass die Regeln, die den Input mit der Vorhersagegröße verbinden, bereits bekannt sind.

Dies wirft die Frage auf, welche Vorteile es bietet, ein Vorhersagemodell zu trainieren, um den Modelloutput selbst vorhersagen zu können? Warum sollte nicht das Modell selbst direkt verwendet werden, um Modelloutputs zu generieren?

Zusätzlich stellt sich die Frage, wie die verwendeten Modellinputs codiert werden sollten, um einen geeigneten Input für einen Lernalgorithmus darzustellen. Beispiele sind hier die Codierung der Passivseite („Reicht es aus, die initialen Rückstellungen pro Rechnungszinsgeneration bereitzustellen?"), der Aktivseite („In

[13] Wie die Codierung aussehen kann, wird im Folgenden noch allgemein diskutiert werden.

welcher Granularität muss die Aktivseite gegeben sein, um gute Vorhersagen treffen zu können?") oder auch der Zinskurven, Managementregeln, Annahmen zum dynamischen Versicherungsnehmerverhalten etc.

Die Beantwortung dieser Frage ist nicht trivial und erfordert einen iterativen Prozess des Feature Engineerings und der Modelloptimierung. Das Spannende ist, dass die Beobachtungen und Antworten, die in dieser Analyse gewonnen werden können, zu neuen Erkenntnissen über die Zusammenhänge verschiedener Versicherungsprodukte und die Wichtigkeit einzelner versicherungstechnischer Annahmen führen und auch quantitative Aussagen über die Güte von Input-Repräsentationen (z. B. 10-jährige Spotrate vs. gesamte Zinskurve) getroffen werden können.

Eine Auswahl an Erkenntnismöglichkeiten, aber auch an Tools, die auf dem Weg zum Erkenntnisgewinn verwendet werden können, wird im folgenden Abschnitt näher beleuchtet.

Beschleunigung von Vorhersagen

Nehmen wir an, es wurde bereits ein selbstlernender Algorithmus trainiert, der mit hinreichender Güte die Eigenmittel nach SII[14] vorhersagen kann. Dann ist die Erwartung an dieses Modell, dass es dazu verwendet werden kann schnellere Vorhersagen zu treffen, als es das „vollständige, aktuarielle Projektionsmodell" könnte. Das heißt, anstatt in Zukunft das vollständige aktuarielle Vorhersagemodell zu verwenden, um beispielsweise zu analysieren, wie sich eine geänderte Assetallokation auf die Eigenmittel nach SII auswirkt (als einen Teil der SII-Quote), kann das in der Regel schneller ablaufende, gelernte Vorhersagemodell befragt werden.[15] Dieser Geschwindigkeitsvorteil kann direkt genutzt werden, um Vorhersagen im Zusammenhang mit „nested stochastics" zu treffen (für das derzeit u. a. „Least-Squares-Monte-Carlo"-Methoden zum Einsatz kommen).

In diesem Zusammenhang muss aber auch Folgendes beachtet werden: Machine-Learning-Algorithmen können und sollten nicht als Lösung für nicht-performante aktuarielle Projektionsmodelle verstanden werden. Um eine geeignete Vorhersagegüte von Vorhersagealgorithmen sicherstellen zu können, muss eine ausreichende Anzahl an „Trainings- und Validierungsdatensätzen" bereitgestellt werden. Da diese Datensätze vom aktuariellen Modell selbst erzeugt werden müssen, ist es unerlässlich, in einem ersten Schritt die Performance des nicht-performanten aktuariellen Projektionsmodells zu verbessern, um ausreichend Datenmaterial zu erzeugen.

Zusammenfassend lässt sich also sagen, dass der geplante Einsatz von Machine-Learning-Verfahren der Beschleunigung von aktuariellen Modellen dienlich ist:

[14] oder einer anderen Größe

[15] Wobei natürlich beachtet werden muss, dass einige Vorhersagemodelle (wie z. B. Random Forests) nicht gut extrapolieren, d. h. nicht auf Bereiche der Inputdaten angewandt werden sollten, die das Modell vorher nie gesehen hat.

- Einerseits kann die Planung des Einsatzes von Machine Learning als Anreiz gesehen werden, die aktuariellen Modelle selbst zu beschleunigen, um eine möglichst hohe Anzahl an Trainings- und Validierungsmaterial bereitzustellen.
- Andererseits kann das gelernte Modell als Proxy-Modell eingesetzt werden, um Ad-hoc-Analysen oder „What-if"-Analysen durchführen zu können.

Neben den allgemeinen Herausforderungen des Machine Learning müssen im Kontext des Proxy-Modells natürlich auch die üblichen Fragen nach Vorhersagegrößen (SII-Eigenmittel oder BEL, Barwerte oder Cashflows) und nach Gütekriterien (wann ist das Proxy-Modell „gut genug"?) diskutiert werden.

Unserer Einschätzung nach stellen sich Aufwand und Nutzen dieser Methode dabei wie folgt dar:

- Initialaufwand für die Beschleunigung bisheriger Modelle: sehr hoch,
- Initialaufwand für die Implementierung von Proxy-Modellen: mittel bis hoch
- Nutzen: sehr hoch.

Nutzen von zufriedenstellenden trainierten Vorhersagemodellen für Modellvalidierung und Risikomanagement

In diesem Abschnitt diskutieren wir, wie

- der Prozess aussieht, ein erfolgreich trainiertes Vorhersagemodell zu erzeugen,
- das trainierte Modell anschließend verwendet werden kann, um aktuarielle Projektionsmodelle zu validieren, und
- das Modell für das Risikomanagement genutzt werden kann.

Dazu werden zunächst die folgenden Fragen diskutiert:

- Was bedeutet es, ein zufriedenstellendes Vorhersagemodell trainiert zu haben?
- Was kann man aus dem Vorhersagemodell-Optimierungsprozess lernen?
- Welche Tools können verwendet werden, um zu einem zufriedenstellenden Vorhersagemodell zu kommen und dieses zu interpretieren?

Gehen wir z. B. davon aus, dass das Vorhersagemodell die Eigenmittel nach Solvency II vorhersagen soll. Es liegt nahe, die Frage nach zufriedenstellendem Training allgemein mit „wenn das Modell gute Vorhersagen trifft" zu beantworten. Doch wie kann man diese Frage quantifizieren?

Die Vorhersagegüte lässt sich für stetige Vorhersagegrößen mit Hilfe diverser **Gütekriterien** messen, wie z. B. dem „Mean Squared Error" (MSE, deutsch: mittlere quadratische Abweichung), dem „Mean Absolute Error" oder auch dem „R^2-Bestimmheitsmaß", welches mit dem MSE verwandt ist. Sind die Vorhersagen auf dem Validierungs- und Testdatensatz nun gemäß dem ausgewählten Gütekriterium

in einem akzeptablen Rahmen (welcher selbst auch im Vorfeld für den Anwendungszweck definiert werden muss), so kann man davon ausgehen, ein zufriedenstellendes Vorhersagemodell trainiert zu haben.

Um zu diesem Ziel zu gelangen, genügt es in der Regel nicht, einen linearen Prozess zu verfolgen, der bei der Bereitstellung eines Datensatzes startet, anschließend das Modell trainiert, validiert und final das Modell für Vorhersagen freigibt. Vielmehr handelt es sich beim Machine Learning um einen iterativen Prozess. Dieser iterative Prozess führt dann, wenn er geeignet eingesetzt wird, zu einem Erkenntnisgewinn, wie im Folgenden exemplarisch dargestellt wird.

Dazu stellen wir uns zunächst vor, dass alle getesteten Vorhersagemodelle (wie z. B. lineare Regression, Random Forest, neuronale Netze) nicht in der Lage sind, aus dem vorliegenden Datensatz die entscheidenden Zusammenhänge zu erlernen. Alle verwendeten Modelle sind also nicht in der Lage, Vorhersagen akzeptabler Güte zu treffen. Dies führt zu folgender Frage: Ist der verwendete Datensatz etwa ungeeignet, um die abhängige Variable (im Beispiel die Eigenmittel) vorherzusagen? Wenn der Datensatz z. B. ausschließlich aus der Summe der HGB-Buchwerte der Aktiva als auch der Summe der versicherungstechnischen Verpflichtungen nach HGB besteht, so liegt diese Vermutung für das Beispiel nahe. Dies zeigt, dass diese beiden Größen nicht ausreichend sein müssen, um gute Vorhersagen der Eigenmittel zu treffen. Natürlich hätte ein Aktuar diese Aussage auch ohne „Machine Learning" treffen können, indem er seinen Sachverstand (häufig auch „Expert Judgement" genannt) verwendet. Allerdings ist der Vorteil an diesem Vorgehen eindeutig, dass man nahezu gänzlich auf sogenanntes „Expert Judgement" verzichten kann, da man mithilfe eines geeigneten Verfahrens (hier Machine Learning) diese Aussage nachgewiesen hat. Handelt es sich bei der Auswahl der Features um kompliziertere Faktoren, so ist die Antwort auf die Frage, welche Features verwendet werden sollten, nicht so einfach zu beantworten. In diesem Fall kann der Aktuar das Machine Learning als Möglichkeit verwenden, die Vermutungen zu untermauern.

In einem nächsten Schritt liegt es nun nahe, den Datensatz um weitere Attribute zu erweitern („Bottom-Up"-Auswahl der Inputfeatures), beispielsweise durch eine Erhöhung der Granularität der Daten (z. B. versicherungstechnische Rückstellungen nach Rechnungszins oder Assetklasse) oder das Hinzufügen von zusätzlichen Informationen (z. B. Zinskurven). Im Beispiel der Zinskurve stellt sich die Frage, ob die gesamte Zinskurve als Input verwendet werden sollte oder ob nicht auch ausgewählte Stützpunkte (z. B. zu Laufzeiten von 5, 10, 15 und 20 Jahren) ausreichend sind, um die Zinskurve zu charakterisieren. Diese Frage lässt sich nun, wie oben beschrieben, in einem iterativen Prozess analysieren. Einerseits kann man als Dateninput die gesamte Zinskurve auswählen, andererseits nur ausgewählte Stützpunkte. Am Ende entscheidet die Vorhersagegüte beider Inputs darüber, welche Variante sich als „am besten" herausstellt.

Als Gegensatz zu diesem „Bottom-Up"-Vorgehen in der Auswahl der Datenattribute („Feature Selection") kann man natürlich auch einen „Top-Down"-Ansatz wählen. In diesem Ansatz zeigt man dem Vorhersagemodell in der Trainingsphase alle zur

Verfügung stehenden Attribute. Anschließend werden diejenigen Features selektiert, die die größte Vorhersagekraft haben. Doch wie kann nun die Vorhersagekraft eines Attributs quantifiziert werden?

Im Falle der linearen Regression wird die Wichtigkeit von Attributen meist über zwei Tests definiert:

1. Mithilfe eines statistischen Tests (Nullhypothese: Der Regressionskoeffizient ist null) wird geprüft, ob der Koeffizient der linearen Regression tatsächlich einen signifikanten Einfluss auf die Vorhersagegröße hat.[16]
2. Falls der Regressionskoeffizient einen signifikanten Einfluss aufweist, kann mit Hilfe des Wertes des Koeffizienten auf die „Wichtigkeit" im Vergleich zu den anderen Attributen geschlossen werden.[17]

Verwendet man nun andere Vorhersagemodelle, weil diese z. B. zu besseren Vorhersagen führen, so gibt es ebenfalls Methoden, um die Vorhersagekraft eines Attributs zu quantifizieren.[18] Im Englischen wird die im Folgenden diskutierte Methode meist im Zusammenhang mit „Feature Importance" diskutiert (vgl. Abschnitt 4.13.1).

Stellen wir uns als Beispiel vor, dass die Eigenmittel bekannt sind für eine Reihe an verschiedenen Inputwerten.[19] Nehmen wir z. B. an, dass die Inputgrößen durch

- aggregierte Versicherungssummen des Bestandes („sum_insured"),
- mittlere Alter des Bestandes („duration_if"),
- mittlere Eintrittsalter der VN („age_at_entry"),
- Geschlechtermix des Bestandes („gender") und
- 10- und 20-jährige Spotrates („yield_10y" und „yield_20y")

gegeben sind.

Nachdem das Modell trainiert worden ist, kann anschließend die Feature Importance für die verschiedenen Datenattribute berechnet werden. Ein typischer Output der Analyse sieht wie in Abbildung 7 dargestellt aus.

[16] Wie immer gilt für den statistischen Test natürlich, dass die Annahmen der linearen Regression gelten müssen (Gauß-Markov-Annahmen), damit der Test gültig ist. Sollten diese für den Datensatz verletzt sein, stellt sich die Frage nach der Korrektheit des oben angesprochenen Hypothesentests.

[17] Hierbei muss dringend auf die Skalierung der Inputgrößen geachtet werden, da der Wert des Koeffizienten nicht unabhängig von der Skalierung der Inputgröße ist.

[18] Die im Folgenden vorgestellte Methode lässt sich auch auf lineare Regression anwenden.

[19] In den Abbildungen in diesem Abschnitt wurde von einem vollständig synthetisch erzeugten Datensatz ausgegangen, in dem die Eigenmittel nur von den Variablen „sum_insured" und „duration_if" abhängt. Die Variable „yield_10y" ist mit „duration_if" korreliert.

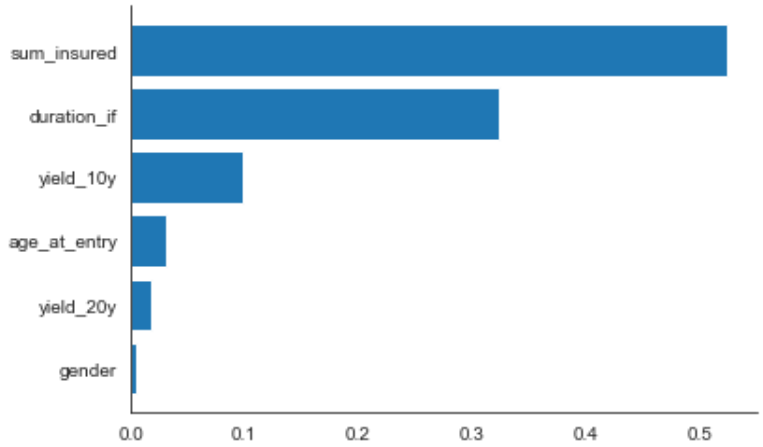

Abbildung 7: Feature Importance Plot eines synthetisch erstellten Datensatzes

An diesem Beispiel sieht man, dass das Feature „sum_insured" am wichtigsten ist. Anschließend folgen „duration_if" und weit abgeschlagen „yield_10y". Die Inputs „age_at_entry", „yield_20y" und „gender" haben in diesem Beispiel kaum eine Aussagekraft. Es stellt sich nun also die Frage, wie man die Feature Importance definieren bzw. bestimmen kann.

Einerseits kann man natürlich eine bereits implementierte Routine verwenden.[20] Alternativ kann man die Feature Importance auch wie folgt selbst berechnen. Zunächst berechnet man die aktuelle Vorhersagegüte des Modells basierend auf dem unveränderten, vorliegenden Datensatz. Angenommen, ein Datenattribut hat eine sehr hohe Vorhersagekraft. In diesem Fall müsste sich die Vorhersagequalität des Modells drastisch verschlechtern, sobald das Attribut aus dem Vorhersagemodell entfernt wird. Man könnte nun auf die Idee kommen, Vorhersagemodelle zu trainieren, bei dem jeweils eines der Attribute entfernt wird, um die Wichtigkeit desjenigen Attributs zu prüfen. Da das Trainieren eines Modells allerdings rechnerisch sehr aufwändig ist, sieht man in der Regel davon ab und bedient sich stattdessen des folgenden Tricks: Man verwendet das bisher trainierte Modell, das alle Datenattribute verwendet. Um nun den Effekt des „fehlenden" Datenattributs zu simulieren, korrumpiert man das Datenattribut, indem man das Attribut im vorhandenen Datensatz zufällig mischt, sodass sämtliche Zusammenhänge zwischen dem durchmischten Attribut und den anderen Attributen „verschleiert" worden sind. Handelt es sich nun um ein äußerst wichtiges Attribut, das eine starke Vorhersagekraft hat, so müsste nach dem Durchmischen des Attributs die Vorhersagekraft zurückgegangen sein. Hat das durchmischte Attribut keinerlei Vorhersagekraft, so dürfte sich die Vorhersagekraft des Modells mit dem durchmischten Attribut nicht geändert haben. Genau so geht man nun also vor: Man sucht sich ein Attribut aus, durchmischt das betroffene Attribut im Datensatz, wobei alle anderen Datenattribute aber unverändert (!) bleiben. Anschließend vergleicht man die Vorhersagequalität des „durchmischten" Datensatzes mit der Vorhersagequalität des originalen Datensatzes. Dieses Vorgehen wiederholt man, bis man jedes Attribut einmal

[20] Im „scikit-learn"-Model der Sprache „Python" gibt es z. B. für Random Forests das Attribut „feature_importances_"

durchmischt hat, wobei aber alle anderen Attribute dann unverändert bleiben müssen. Anschließend kann man die Reduktion der Vorhersagekraft vergleichen, um ein „Importance Ranking" der Inputvariablen zu erstellen, vgl. Abbildung 7.

Diese Feature Importance kann man nun verwenden, um beispielsweise die Vorhersagekraft des Modells zu erhöhen, indem man „unwichtige" Attribute aus dem Modell entfernt. Gleichzeitig kann man aber auch Erkenntnisse gewinnen, die zur Modellvalidierung verwendet werden können. So würde man anhand der Abbildung 7 lernen, dass die gesamte Versicherungssumme im Bestand einer der Haupttreiber für die Eigenmittel nach Solvency II sind. Anschließend folgt das mittlere Bestandsalter und anschließend der 10-jährige Zins. Weiterhin beobachtet man, dass die anderen Größen einen weit geringeren Einfluss auf die Eigenmittel haben.

- Aus Vorhersageperspektive würde man nun die Größen Eintrittsalter, 20-jährigen Zins und Geschlecht aus dem Modell entfernen, um die Vorhersagegüte zu verbessern.[21]
- Aus Validierungsperspektive würde man sich die Frage stellen: Macht es Sinn, dass die gesamte Versicherungssumme des Bestandes der größte Treiber für die Eigenmittel sind? Warum ist das so? Wenn dies der größte Treiber ist, würde sich die Vorhersagegüte verbessern, wenn man die Versicherungssumme pro Rechnungszinsgeneration zur Verfügung stellt?
    - Diese strukturierte Vorgehensweise der Modellvalidierung kann zu einer systematischeren Analyse und Validierung des Projektionsmodells führen, da sich neue Sichtweisen und Erkenntnisse auf das aktuarielle Projektionsmodell ergeben könnten, die ansonsten ggf. im Verborgenen geblieben wären.
    - Diese Analyse kann ebenfalls im Nachhinein verwendet werden, um festzulegen, welcher Modellinput sehr genau/detailliert hergeleitet werden muss und welcher Modellinput auch „gröber" hergeleitet werden kann, da man ja gezeigt hat, dass zum Verständnis des Modelloutputs ggf. gröbere Inputs ausreichen.
    - Gleichzeitig kann die Wichtigkeit einzelner Inputgrößen mit (bekannten) Modellsensitivitäten verglichen werden, um so eine weitere Validierung des Modells zu ermöglichen.
- Aus Risikomanagementsicht kann dieser Zusammenhang interpretiert werden als: Welche Größe ist der größte Treiber für die Eigenmittel? D. h. was sollte als Erstes verändert werden, um die Eigenmittel zu optimieren?

Ähnlich zur Verwendung der Feature Importance zur Identifikation von Abhängigkeiten können auch sogenannte „Dendrograms" verwendet werden. Diese dienen dazu, Gemeinsamkeiten zwischen verschiedenen Inputvariablen zu erkennen. Auf diese Analysemethode wird in diesem Abschnitt aber nicht weiter eingegangen.

[21] Bzw. um zu testen, ob sich die Modellgüte dadurch verbessert.

Nachdem die Feature Importance analysiert worden ist, zu besseren Modellvorhersagen geführt hat und ggf. auch bereits neue Blickwinkel auf das Modell erzeugt hat, kann das Vorhersagemodell im Anschluss im Sinne von „Partial Dependence Plots" befragt werden. Hierbei werden die Zusammenhänge der abhängigen Variablen von den Prädiktoren analysiert. Von dieser Analysemethode kann einerseits die Modellvalidierung profitieren, da Fragen wie „Verändert sich mein Modelloutput so, wie es erwartet wird?" analysiert werden können. Andererseits können aber auch Fragen des Risikomanagements wie z. B. nach Abhängigkeiten der Risiken vom Bestand oder Kapitalmarkt beantwortet werden.

Eine typische Vorgehensweise für die Analyse der Abhängigkeit des Modelloutputs von einer Inputvariable besteht darin, die abhängige Variable (hier die Eigenmittel) bzgl. der Veränderung in einem erklärenden Attribut darzustellen, wobei die Darstellung auf dem vorliegenden Datensatz basiert. In dem in diesem Beispiel verwendeten synthetischen Datensatz sieht ein solcher Plot wie folgt aus.

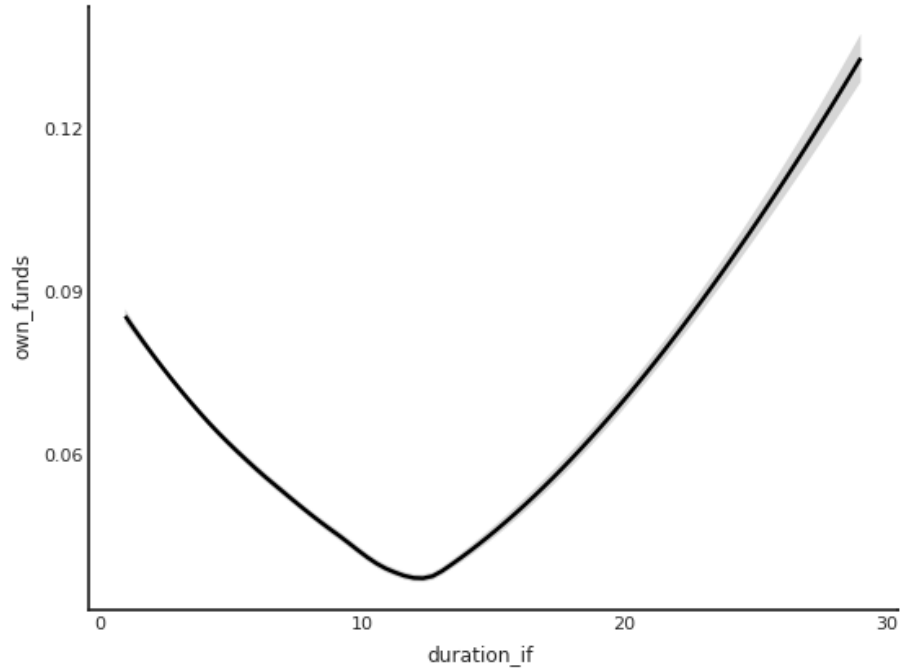

Abbildung 8: Abhängigkeit der Eigenmittel vom mittleren Alter des Bestands

In diesem Beispiel werden für alle verfügbaren Datensätze die Eigenmittel („own_funds") in Abhängigkeit vom mittleren Bestandsalter dargestellt („duration_if") und die vorhandenen Datenpunkte mittels einer lokal gewichteten Regression geglättet. Diese Abbildung basiert also ausschließlich auf den Datenpunkten des vorliegenden Datensatzes. Bemerkenswert an dieser Abbildung ist das Minimum der Eigenmittel in Abhängigkeit vom Bestandsalter, das nach einer näheren Analyse verlangt.

Abbildung 8 lässt nun folgende zwei Schlüsse zu, wenn man die Regel „je größer das mittlere Bestandsalter, desto geringer die Eigenmittel" erwarten würde:

- Entweder ist das aktuarielle Projektionsmodell nicht korrekt, sodass dieses Minimum ein Artefakt darstellt,

- oder aber das aktuarielle Projektionsmodell ist korrekt und das Minimum kann durch eine Korrelation mit einem anderen Feature verstanden werden.

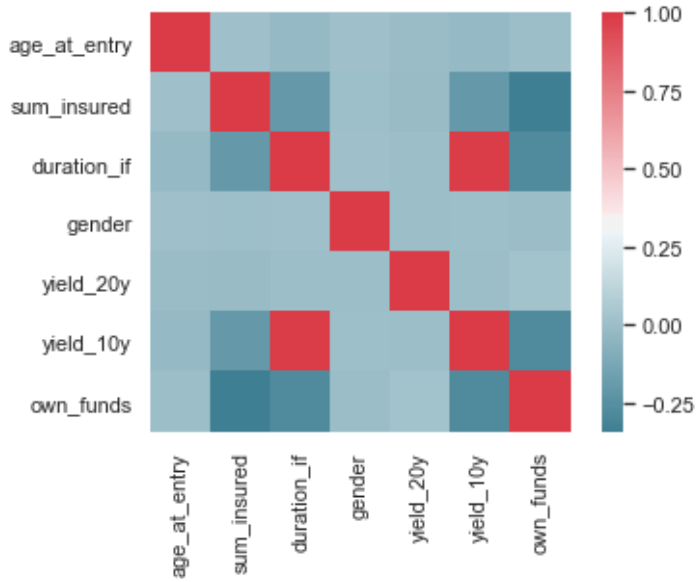

Abbildung 9: Korrelationsmatrix des zugrundeliegenden Datensatzes

Betrachtet man die Korrelationsmatrix des unterliegenden Datensatzes aus Abbildung 9 , so stellt man fest, dass das mittlere Bestandsalter mit der Versicherungssumme korreliert ist. Wenn nun höhere Bestandsalter mit kleineren Werten der Versicherungssumme korreliert sind, und eine geringere Versicherungssumme mit höheren Eigenmitteln korreliert ist (z. B. weil es sich um defizitäre Policen handelt), so könnte das Minimum in Abbildung 8 genau durch Korrelationen in den Inputdaten erklärt werden.

Dass dies auch der Fall ist, zeigt die folgende Abbildung 10.

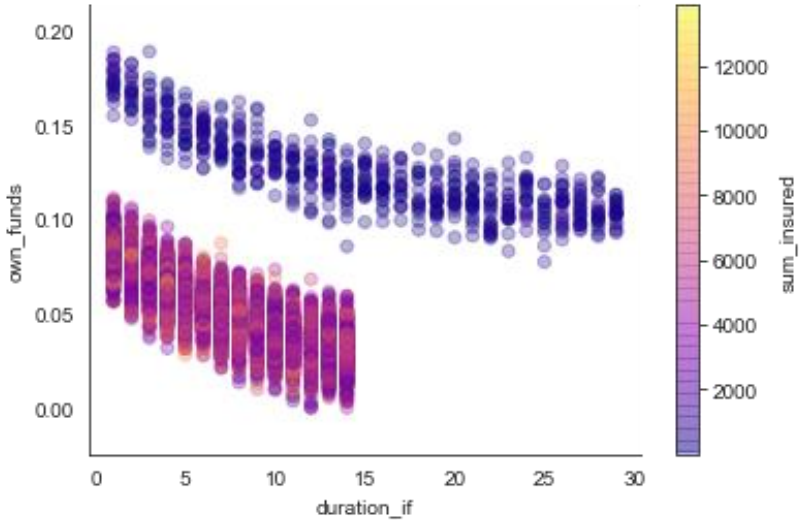

Abbildung 10: Zusammenhang zwischen Eigenmitteln, Bestandsalter und Versicherungssumme

Diese beobachteten Zusammenhänge haben bereits nach einer relativ komplexen Analyse verlangt, um den Datensatz als „plausibel" bezeichnen zu können. Aus

diesem Grund stellt sich die Frage, ob es nicht auch „direktere" Analysemöglichkeiten gibt, die solche Korrelationen in den Inputdaten beseitigen. Dies würde zu einer einfacheren Analyse und damit auch potentiell zu einem besseren Verständnis des Modells führen.

Genau für diesen Analysezweck können die „Partial Dependence Plots" verwendet werden, vgl. Abbildung 11 und Abbildung 12.

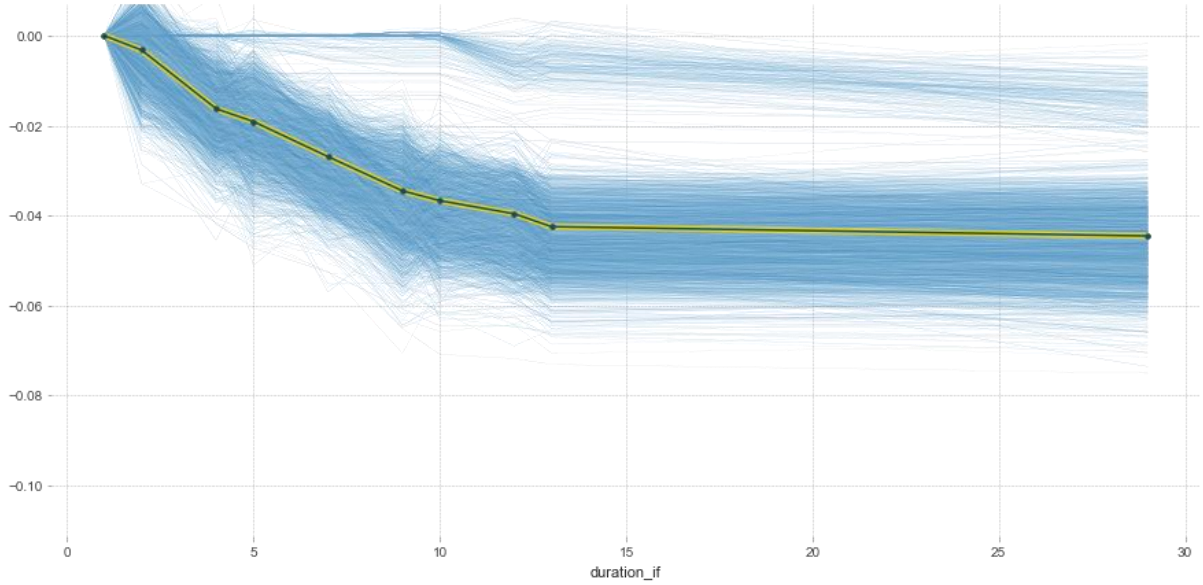

Abbildung 11: Partial Dependence Plot für "duration_if"

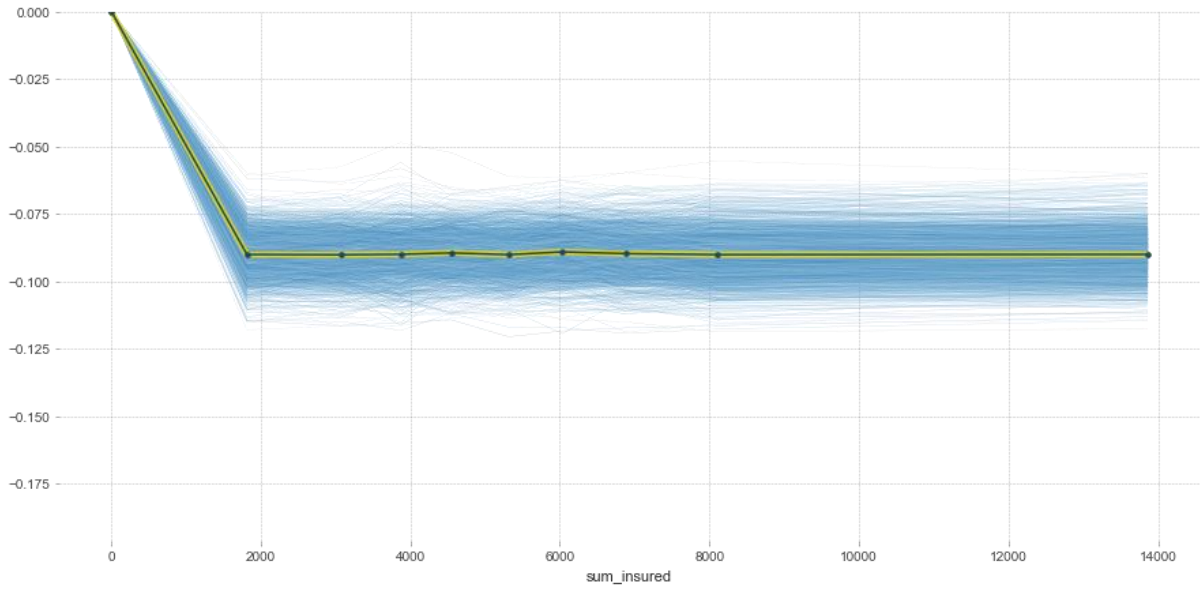

Abbildung 12: Partial Dependence Plot für "sum_insured"

Diese Plots stellen die Abhängigkeit des Modelloutputs (hier gleich den Eigenmitteln) von der Versicherungssumme und des Bestandsalters dar (vgl. Abschnitt 4.13.2). Dabei stellen die blauen Linien gerade die Vorhersagen für jeden einzelnen Datenpunkt dar, bei dem der Wert der unabhängigen Variablen eingestellt worden ist. Hierbei ist zu beachten, dass der Startwert der Eigenmittel zu „0" definiert worden ist. Die gelbe Linie stellt dann die „gemittelte" Vorhersage („Randverteilung") dar. Durch diese Technik lassen sich intrinsische Korrelationen in den Inputdaten herausmitteln und die reine Abhängigkeit der abhängigen Variablen von den unabhängigen Variablen ermitteln.

Anhand der obigen Darstellungen lassen sich nun also die folgenden Aussagen für die Abhängigkeit der „own_funds" von der „duration_if" und „sum_insured" treffen: Je größer das mittlere Bestandsalter, desto niedriger sind die „own_funds" bei ansonsten unveränderten Inputgrößen. Durch den Einsatz des Vorhersagemodells und der oben beschrieben Technik sieht man direkt, dass der „wahre" Zusammenhang zwischen den „own_funds" und der „duration_if" monoton ist – ganz im Gegensatz zu der ersten Vermutung, die man ggf. nach Studium der Abbildung 8 gewonnen hätte. Für die Versicherungssumme lässt sich folgende Beobachtung machen: Erhöht man die Versicherungssumme des Bestands, hält aber alle anderen Größen konstant, so erwartet man zunächst eine Abnahme der Eigenmittel. Anschließend erwartet man, dass die Eigenmittel ab einer gesamten Versicherungssumme von 2000 Einheiten saturiert. Dies ist eine Erkenntnis, die anschließend im Rahmen einer Modellvalidierung diskutiert werden kann – so könnte man sich z. B. fragen, warum die Eigenmittel nicht weiter absinken.

Kommt man zu dem Schluss, dass diese Beobachtungen valide sind, so kann man diese auch im Rahmen des Risikomanagements einsetzen: So würde man zu dem (vielleicht erstaunlichen) Schluss kommen, dass die Versicherungssumme kaum einen Einfluss mehr auf die Eigenmittel hat, sobald sie größer als 2000 Einheiten ist. Gleichzeitig scheint ein älterer Bestand sich negativ auf die Eigenmittel auszuwirken. Da man nun diese beiden Einflussfaktoren isoliert voneinander betrachten kann, kann man natürlich anschließend ein fundierteres Risikomanagement betreiben, als wenn man sich von Abbildung 8 hätte leiten lassen und zum Schluss gekommen wäre: Je älter der Bestand, desto besser für meine Eigenmittel.

Zum Schluss sei noch erwähnt, dass sich anhand der „Partial Dependence Plots" auch weitere Erkenntnisse wie eine Segmentierung der Inputdaten gewinnen lassen, vgl. z. B. die zwei verschiedenen Klassen an blauen Graphen in Abbildung 12. Die Diskussion dieser Beobachtung bleibt dabei dem interessierten Leser überlassen.

In Summe schätzen wir den Initialaufwand zum Aufbau der entsprechenden Infrastruktur und Analyseprozesse als sehr hoch ein. Wie anhand der Beispiele deutlich werden dürfte, schätzen wir aber auch den daraus resultierenden Nutzen als sehr hoch ein.

2.4. Kalibrierung stochastischer Szenarien zur Bewertung von Optionen und Garantien

2.4.1. Kurze Beschreibung

Für die stochastische Bewertung von Optionen und Garantien werden marktkonsistente und risikoneutrale Szenarien verwendet. Üblicherweise verwenden mittelständische Versicherungsunternehmen für deren Erzeugung ein 1-Faktor-Hull-White-Modell. Dieses Modell kann in der Praxis nur an einer stark begrenzten Anzahl ausgewählter Stützstellen (Swaptionvolatilitäten) kalibriert werden. Dies wirft unter Umständen die Frage auf, welche Stützstellen zu wählen sind, und verursacht weiteren analytischen Aufwand. Andererseits ist die Implementierung und Kalibrierung dieses Modells im Vergleich zu anderen Zinsmodellen überschaubar.

Eine Alternative zum 1-Faktor- stellt das 2-Faktor-Hull-White-Modell dar. Hier können bis zu fünf Parameter für die Kalibrierung an Marktdaten verwendet werden, sodass die damit erzeugten stochastischen Szenarien den aktuellen Kapitalmarkt besser widerspiegeln. Die Kalibrierung des 2-Faktor-Hull-White-Modells stellt allerdings eine deutlich größere Herausforderung dar. Bei dieser Aufgabe können Methoden der Künstlichen Intelligenz helfen.

2.4.2. Datengrundlage

Als Datengrundlage sind eine für die Kalibrierung verwendete Zinskurve sowie Swaptionvolatilitäten nötig.

2.4.3. Modelle

Für die Kalibrierung der Modelle sind Optimierungsalgorithmen nötig, die die freien Parameter so bestimmen, dass sie die Szenarien die Marktdaten widerspiegeln. In Abhängigkeit der Komplexität des Optimierungsproblems lassen sich einfache Algorithmen wie das eindimensionale Newtonverfahren und Non-negative-Least-Squares-Algorithmen bis hin zu komplexeren mehrdimensionalen Verfahren einsetzen.

Eine Alternative zur klassischen Kalibrierung über Optimierungsalgorithmen stellt die Kalibrierung mit neuronalen Netzen dar. Dieses Verfahren kann unabhängig vom zugrundeliegenden Zinsmodell verwendet werden. Hierbei kann das neuronale Netz entweder mit historischen oder synthetischen Trainingsdaten erstellt werden. Für den Einsatz von historischen Daten werden neben Marktdaten auch die optimierten Parameter des entsprechenden Zinsmodells benötigt. Diese lassen sich mit den oben beschriebenen Optimierungsalgorithmen bestimmen, wobei einiges an Implementierungsaufwand und finanzmathematisches Grundwissen notwendig ist. Eine weitere Möglichkeit, die diesen Implementierungsaufwand umgeht, besteht darin, Trainingsdaten durch Variation der zu optimierenden Parameter zu erzeugen. Dazu generiert man Szenarien mit verschiedenen zu optimierenden Parameter und bestimmt daraus die relevanten Marktwerte. Diese Marktwerte werden dann zusammen mit den zu kalibrierenden Optimierungsparametern als Trainingsdaten verwendet. Der Vorteil dieses Vorgehens ist, dass ein trainiertes Netz den Kalibrierungsvorgang unabhängig vom gewählten Zinsmodell beschleunigen kann. Dieses trainierte Netz kann dann auch verwendet werden, um Szenarien für Stresstests zu berechnen.

2.4.4. Hinweise zu Ergebnissen und möglicher Nutzung

Eine allgemeine Aussage, inwieweit sich die bessere Abbildung des Kapitalmarktes auf die Bewertung der Optionen und Garantien auswirkt, ist kaum möglich. Bei Kalibrierung des 1-Faktor-Hull-White-Modells an eine mittlere Swaptionlaufzeit beobachtet man häufig eine Überschätzung der Marktpreise bei kurzen Laufzeiten und eine Unterschätzung bei langen Laufzeiten. Wie sich diese Abweichung in der Bewertung fortpflanzt, hängt jedoch vom jeweiligen Passivbestand ab. Demnach hängt auch die Auswirkung einer Verbesserung der Marktkonsistenz durch Verwendung eines 2-Faktor-Hull-White-Modells vom jeweiligen Passivbestand ab.

Da das oben beschriebene Verfahren unabhängig vom Zinsmodell verwendet werden kann, wäre es möglich die Komplexität des Zinsmodells weiter zu erhöhen und damit die Marktkonsistenz noch weiter zu verbessern. So wäre es beispielsweise denkbar neben ATM- auch an OTM-Swaptions zu kalibrieren bzw. untere Zinsschranken bereits bei der Generierung der Szenarien zu berücksichtigen.

2.5. Clustering von Bestandsdaten – Bestandsverdichtung und allgemeine Bestandsauswertungen

2.5.1. Bestandsverdichtung – Beschreibung

Mit wachsenden Anforderungen an stochastische Bewertungen der großen Lebensversicherungsportfolien stoßen Unternehmen auf ein Laufzeitproblem. Es sollen grundsätzlich Millionen Verträge unter Tausenden Zinsszenarien über mehrere Jahrzehnte und unter zahlreichen Annahmen projiziert werden. Ziel der Bestandsverdichtung ist die Erzeugung eines möglichst kleinen Teilbestandes, der die gleichen Eigenschaften besitzt wie der Originalbestand.

2.5.2. Datengrundlage

Als Basis der Bestandsverdichtung dienen unternehmenseigene Datensysteme, in denen die Informationen über alle Policen geführt werden. Solche Daten sollten im Idealfall für das stochastische Modell aufbereitet werden. Es kann passieren, dass dabei Informationen durch Vereinfachungen verloren gehen, da nicht alle Besonderheiten im stochastischen Modell abgebildet werden. In diesem Fall soll eine als Teil der Bestandsverdichtung entstandene Unschärfe zwischen deterministischen und stochastischen Ergebnissen möglichst ausgeglichen werden.

2.5.3. Modelle

Als Grundlage der Bestandsverdichtung dient die Idee der Gruppierung ähnlicher Daten in Clustern. Es soll möglichst hohe Homogenität innerhalb eines Clusters und hohe Heterogenität zwischen den Clustern gewährleistet werden. In jedem Cluster wird eine möglichst niedrige Anzahl an repräsentativen Policen gesucht, die die Eigenschaften des gesamten Clusters möglichst nah abbildet. Beim Clustering wird anhand mehrerer Kenngrößen im Verlauf der Projektion eine Verlustfunktion definiert und für die Optimierung verwendet. Als Referenz für die Optimierung dienen die Ergebnisse der deterministischen Projektion des Originalbestands.

Die initiale Auswahl der repräsentativen Policen erfolgt meistens zufällig. Das Clustering wird durch den dafür entworfenen Algorithmus iterativ geändert und verfeinert. Für die Lösung dieses Optimierungsproblems können sowohl klassische Lösungsalgorithmen (ggf. ihre modernere Versionen) wie z. B. Simplex als auch für diese Aufgabe speziell entworfene Methoden aus dem Gebiet der Metaheuristik verwendet werden.

2.5.4. Hinweise zu Ergebnissen und möglicher Nutzung

Die größte Hürde bei der Optimierung ist die hohe Dimension und ggf. Unschärfe zwischen deterministischen und stochastischen Modellen. Um die Lösung einer solchen Optimierung in Echtzeit sicherzustellen, sollte der Algorithmus verbessert und richtig parametrisiert werden. Dabei spielt die Definition der Kenngrößen für die

Optimierung sowie für Toleranzen bzgl. der akzeptablen Abweichung von Referenzwerten eine wichtige Rolle.

2.5.5. Allgemeine Bestandsauswertungen – Beschreibung

Noch allgemeiner liegen bereits aufbereitete und strukturierte Daten zu Kunden und deren Verträgen in vielen Datenbanken und Tabellen bei Versicherungen vor. Diese Daten können auf nicht im Vorfeld konkret definierte Fragestellungen analysiert werden. Mögliche Ziele sind:

- Monitoring der Kundenstruktur im Zeitverlauf,
- Gruppierung ähnlicher Kunden oder Verträge in Segmente,
- Identifikation von allgemeinen Auffälligkeiten in den Bestandsdaten.

Im Gegensatz zu den meisten sonstigen Beispielen in diesem Bericht befindet man sich im Bereich des unüberwachten Lernens und es werden nur existierende Datenquellen herangezogen.

2.5.6. Datengrundlage und Modelle

Die bestehenden Datenbanken sind üblicherweise relationale Datenbanken und für Big-Data-Methoden damit bestens geeignet. Die Datenqualität und Datenkonsistenz existierender Datenbanken ist häufig höchst unterschiedlich. Vor Analysen ist dies geeignet zu überprüfen und bei Bedarf zu verbessern.

Die Identifikation von Auffälligkeiten und die Segmentierung bestehender Daten gehören zur Gruppe der Clustering-Verfahren. Dimensionsreduzierende Verfahren können ähnlich zu einer Bestandsverdichtung die wichtigsten Faktoren in einem Datensatz herleiten. Aufgrund der unspezifischen Zielsetzung muss eine Bewertung der Ergebnisse mit externer Fachkenntnis erfolgen. Eine automatisierte Nutzung ist nicht empfehlenswert.

K-Means-Clustering und Hierarchical Clustering sind beispielhafte Verfahren, aber auch Ansätze mit neuronalen Netzen (z. B. Autoencoder) können zur Anwendung kommen.

3. Daten und Datenschutz

Die Auswahl der Daten hängt von der Definition der Fragestellung, der Auswahl des Modells und den geplanten Maßnahmen ab. Bei einem überwachten Lernproblem (vgl. Abschnitt 4.1.6) sollte grundsätzlich im ersten Schritt die Abbildung der Zielvariable geklärt werden. Hierbei muss der Transfer von der meist abstrakten Fragestellung hin zu einem konkret in den Bestandsführungssystemen abgebildeten Datenelement geleistet werden. Gibt es nicht *die* eine Variable, die die Zielvariable abbildet, kann es auch sinnvoll sein, mehrere Datenfelder zu verbinden, um die gewünschte Information der Zielvariable zu erzeugen. Die Datenqualität sollte bei der Auswahl der Zielvariable von zentraler Bedeutung sein, da eine schlechte Qualität zu keinem guten Modell führen kann. Je nach Fragestellung und deren Umsetzung in die Zielvariable werden für die meisten überwachten Lernprobleme (binäre) Klassifikationsmodelle oder Regressionsmodelle zur Anwendung kommen.

Neben der Definition der Zielvariablen sind erklärende Merkmale, die die Zielgröße beeinflussen oder bei einem unüberwachten Lernproblem (vgl. Abschnitt 4.1.6) als Grundlage dienen, zu identifizieren und die entsprechenden Datenquellen zu definieren. Die Systemlandschaft der Versicherungsunternehmen hält wesentliche Merkmale bereit (siehe unten) und auch externe Datenquellen stehen zur Verfügung.

Klassischerweise liegen in Versicherungsunternehmen strukturierte Daten aus den Bestandsführungssystemen vor. Zum Vertragsabschluss sind die Daten meist in guter Qualität vorhanden, wobei der Umfang der erfassten Merkmale von Tarifgeneration zu Tarifgeneration unterschiedlich sein mag und insbesondere Verträge, die schon länger im Bestand sind, häufig mit einer dünneren Datenbasis einhergehen. Wichtig ist außerdem zu unterscheiden, ob Merkmale den Zustand bei Vertragsabschluss oder -änderung abbilden bzw. fortlaufend aktualisiert werden. Gerade die Personendaten enthalten viele Daten, die über den Zeitverlauf variabel sind (wie z. B. Anzahl Kinder, Familienstand oder sogar Beruf), in den Systemen jedoch nicht stets gemäß dem aktuellen Status abgebildet sind. Dies muss bei der Modellierung beachtet werden, um Fehlinterpretationen und -schlüsse zu vermeiden.

Neben strukturierten Daten aus den o. g. klassischen Systemen werden zunehmend auch semi-/unstrukturierte Daten erhoben, aus denen für die Modellierung relevante strukturierte Merkmale extrahiert werden können. Hierbei lassen sich z. B. Daten aus Angebotssystemen, Risikoprüfungs- und Leistungsprüfungstools nutzen. Ebenso denkbar sind Daten, die typischerweise vom Vertrieb erhoben werden, wobei zukünftig auch Daten aus Onlinegeschäft oder Apps eine größere Rolle spielen können.

Je nach Fragestellung und Aussagekraft der unternehmenseigenen Daten kann es sinnvoll sein, weitere externe Datenquellen heranzuziehen. Dies können jegliche Statistiken (etwa vom statistischen Bundesamt) sein, die über bestimmte Schlüssel (z. B. die Postleitzahl oder Adresse bei Regionalkennzahlen) an den Bestand angelesen und verwendet werden können. Auch Pooldaten von Rückversicherern bieten die Möglichkeit bessere Modelle zu kalibrieren und Brancheninformationen

zu berücksichtigen. Weiterhin mögen Marktdaten wie Entwicklungen an den Finanzmärkten oder Arbeitslosenzahlen als erklärende Merkmale dienen.

3.1. Datenschutz

Die im Mai 2018 endgültig in Kraft getretene Datenschutzgrundverordnung (DSGVO)[22] hat ausführliche Regelungen zur Verarbeitung personenbezogener Daten aufgestellt, die innerhalb der EU einheitlich anzuwenden sind. Relevant sind hierbei aus Sicht der Nutzung der Versicherungsnehmerdaten für Big-Data-Zwecke zunächst vier wesentliche Aspekte:

1. Einschränkung der Nutzung: Die Verarbeitung personenbezogener Daten darf nur nach zweckgebundener expliziter Einwilligung der Person erfolgen, sofern keine anderen gesetzlichen Erlaubnistatbestände erfüllt sind. Daten sollen sparsam verwaltet werden und die Erhebung soll bei der Person erfolgen. Insbesondere Gesundheitsdaten unterliegen noch einmal besonderen Schutzvorschriften.
2. Informationspflichten: Die betroffene Person ist u. a. über die Nutzung etwaiger externer Datenquellen für personenbezogene Daten sowie über automatisierte Entscheidungsfindung aus einem Modell ohne menschliche Einflussnahme zu informieren.
3. Maßnahmen des Datenschutzes: Personenbezogene Daten sollen durch Sicherungsmaßnahmen wie Pseudonymisierung, Beschränkung des Zugangs etc. umfassend geschützt werden.
4. Entsprechende Aufwände sind zu berücksichtigen und müssen dem Zweck angemessen sein. Dies gilt in fast allen Bereichen der DSGVO: für die Informationspflichten, die Schutzmaßnahmen, bei der (hypothetischen) Identifizierung von Einzelpersonen aus den Daten etc.

In Abwesenheit etablierter Rechtsprechung rund um diese Konzepte ergibt sich für den Anwendungsfall der Big-Data-Modellierung die Herausforderung, in Absprache mit Rechtsabteilung, Datenschutzbeauftragtem und anderen Bereichen ein Verfahren zu erarbeiten, das den entsprechenden Umgang mit den Daten ermöglicht.

Zwei Ansätze werden zur Modellierung hierbei in der Praxis verfolgt:

1. Zu statistischen Zwecken dürfen personenbezogene Daten in zumindest pseudonymisierter Form verwendet werden [Art. 5 Abs. 1 lit. e) DSGVO], auch über den ursprünglichen Zweck und über die Dauer der notwendigen Speicherung hinaus. Dies könnte eventuell auch für externe personenbezogene Daten ohne Informationspflicht der Einzelperson gelten, da hier die Informationspflicht unverhältnismäßig wäre.

   Allerdings gibt es hier auch eine Widerspruchsmöglichkeit für den Versicherungsnehmer, die ggf. operationale Schwierigkeiten aufwerfen könnte.

[22] https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:32016R0679&from=DE

2. Für eine Person oder einen Bereich ohne Zugriff auf identifizierende Merkmale zu Versicherungsnehmern wie Namen, Anschrift, ggf. genaues Geburtsdatum etc. könnte auch innerhalb eines Unternehmens die Situation von anonymen Daten hergestellt werden, in der diese Person oder der Bereich die den Daten zugrunde liegenden Individuen nur mit unverhältnismäßig hohem Aufwand oder unter Verstoß gegen Datenschutzvorschriften des Unternehmens identifizieren könnte.

   Innerhalb dieses Bereichs könnten die Daten dann als anonym und nicht den Vorschriften der DSGVO unterliegend betrachtet werden.

Für die Umsetzung der Analyseprojekte empfiehlt es sich entsprechend, für die Modellierung nur pseudonymisierte, aus Sicht der durchführenden Personen möglichst sogar anonymisierte Daten zu verwenden.

Da die verschiedenen Anwendungsbeispiele und Use Cases in diesem Dokument ohnehin häufig abteilungsübergreifende Abstimmungs- und Datentransferprozesse notwendig machen, dürfte die Pseudonymisierung in der Ursprungsabteilungen und Transfer von lediglich anonymisierten Daten für die Modellierung aber in der Regel keine wesentliche Hürde darstellen.

Im Zeitpunkt einer eventuellen Operationalisierung der Modelle muss selbstverständlich auf personenbezogene Daten zum (identifizierten) Versicherungsnehmer zurückgegriffen werden, sodass hier Zweckgebundenheit, Informationspflichten bzgl. externer Daten etc. zu überprüfen sind. Hierbei sind Detailgrad und Ausmaß der Informationen im Rahmen der Verhältnismäßigkeit sorgfältig abzuwägen. Es bietet sich an, sich entsprechend schon zum Zeitpunkt der Modellierung Gedanken darüber zu machen, ob alle technisch erhältlichen Daten des Modells für den letztlichen Einsatz auch verfügbar sein werden.

Sollte es im operationalen Modell tatsächlich zu einer vollständig automatisierten Entscheidungsfindung und Maßnahme kommen, sind entsprechende Informationspflichten und auch Überprüfungsoptionen des Versicherungsnehmers zu beachten. In der Praxis werden Modelle häufig jedoch eher zur Unterstützung einer menschlichen Einzelentscheidung dienen.

3.2. Interne Datenquellen

3.2.1. Bestandsdaten

Diese lassen sich weitestgehend den vier Gruppen

- Produktdaten
  (z. B. Produkttyp, Garantienkonzept, Art der Risikoabdeckung, Stornobedingungen),
- Vertragsdaten
  (z. B. Vertragsdauer, Vertragsvolumen bzw. Versicherungssumme, aktuelle

Fondsperformance, Beitragsart, Zugangsjahr, Rechnungszins, Zuschläge bzw. Ausschlüsse),

- Personen-/Kundendaten
  (z. B. Alter, Geschlecht, Beruf, Wohnort, Familienstand, BMI, Einkommen),
- Vermittlerdaten
  (z. B. Provisionsmodell, Vertriebsweg)

zuordnen.

Sowohl die Quantität wie auch die Qualität der Daten ist recht stark von Beginn und Art der Police abhängig – bei modernen BU-Versicherungen wird man sehr gute Daten zum Beruf haben, bei Kapitalversicherungen des Altbestandes hingegen nicht.

Bei tarifierungsrelevanten Daten wird die korrekte Erfassung und Speicherung besser qualitätsgesichert als bei Daten, die aus rein statistischen Zwecken erfasst werden.

Bestandsdaten werden häufig schon in geeigneter Form in den Poolmeldungen für den Verband oder für Rückversicherer aufbereitet, so dass diese Meldungen eine sehr geeignete Ausgangsbasis für eigene Analysen darstellen.

Klassischerweise erfolgen allerdings Auswertungen zu einem Teilbestand nur in dessen Bestandsdaten selbst. Interessante Auswertungen können sich jedoch ergeben, wenn diese um entsprechende Angaben zu weiteren Lebensversicherungspolicen derselben Person ergänzt werden können, wozu ein Personenkennzeichen notwendig ist.

3.2.2. Daten aus Antrags- und Leistungsprüfung

Daten aus der Antrags- und Leistungsprüfung sind häufig detaillierter als in der Bestandsverwaltung, wenn sie denn strukturiert vorliegen. Hier ergeben sich interessante Auswertungen, wenn die Daten in pseudonymisierter Weise weitergegeben werden können, da sie selbst bei strukturierter Erfassung nicht über die Zeit und die Produkte hinweg vergleichbar sein dürften, aber einigen Erkenntnisgewinn versprechen.

Im Falle der einfachen Archivierung von Textdokumenten könnte ein sehr komplexes Big-Data-Projekt in der automatisierten Informationsextraktion mithilfe von Machine-Learning-Verfahren bestehen.

3.2.3. Verlaufsdaten zu Zahlungsverhalten und Kundenkommunikation

Grundsätzlich stehen den Versicherern weit mehr Daten zur Verfügung, als klassischerweise genutzt werden. Ein Beispiel sind Angaben zum Zahlungsverhalten unter den laufenden Verträgen (z. B. Verzug in den Prämienzahlungen).

Je nach Datenorganisation stehen dem Versicherer auch Angaben zur Kundenkommunikation zu Verfügung: etwa Datum, Art und Ergebnis der Kommunikation sowie initiierende Partei – sowohl aus der Antragsstrecke als auch im Policenverlauf.

Mit solchen Daten ließe sich klären, ob und wo beispielsweise eine proaktive Kundenansprache auch negative Folgen haben kann.

3.2.4. Vermittlerdaten

Eine Sonderstellung kommt den Daten des Vermittlers zu, da diese je nach Vertriebsorganisation nicht als intern betrachtet werden können. In jedem Fall werden hier häufig vertrieblich interessante Angaben zum VN gesammelt, die auch in der Bestandsanalyse von Interesse sein können (aktuelle Berufstätigkeit, Familienstand usw.).

3.3. Externe, zustimmungspflichtige personengebundene Daten

3.3.1. Bankdaten

Seit September 2019 müssen Banken nach PSD2 ihre Daten auf Verlangen des Kunden mit dafür lizenzierten dritten Firmen teilen.[23]

Die finanzielle Risikoprüfung ließe sich durch diese Daten verifizieren und automatisieren. Besonders bei Berufsunfähigkeitsversicherungen liegt ein Einfluss zwischen versicherter Rente, Gehalt und den beobachteten Inzidenzen nahe. Auch Erhöhungsklauseln bei Veränderungen des Lebensstils lassen sich prüfen, erkennen und definieren.

Bei kontinuierlichem Monitoring von Einkommen, Vermögen und Konsum lässt sich die individuelle Risikolage sehr gut bestimmen. Auf den Bedarf des Kunden maßgeschneiderte Lösungen könnten angeboten werden.

3.3.2. Kreditwürdigkeit

Gegebenenfalls könnte eine bei Vertragserstellung erhobene Einstufung der Kreditwürdigkeit (Schufa, Creditreform) in Bewertungen einfließen, wenn die verwendete Klausel dies zulässt.

3.3.3. Social Media

Die weite Verbreitung der regelmäßigen Nutzung sozialer Medien in Deutschland lässt deren Verwendung zunächst attraktiv erscheinen (siehe Abbildung 13).

[23] PSD2 (Second Payment Services Directive, Zweite Zahlungsdiensterichtlinie vom 25.11.2015).

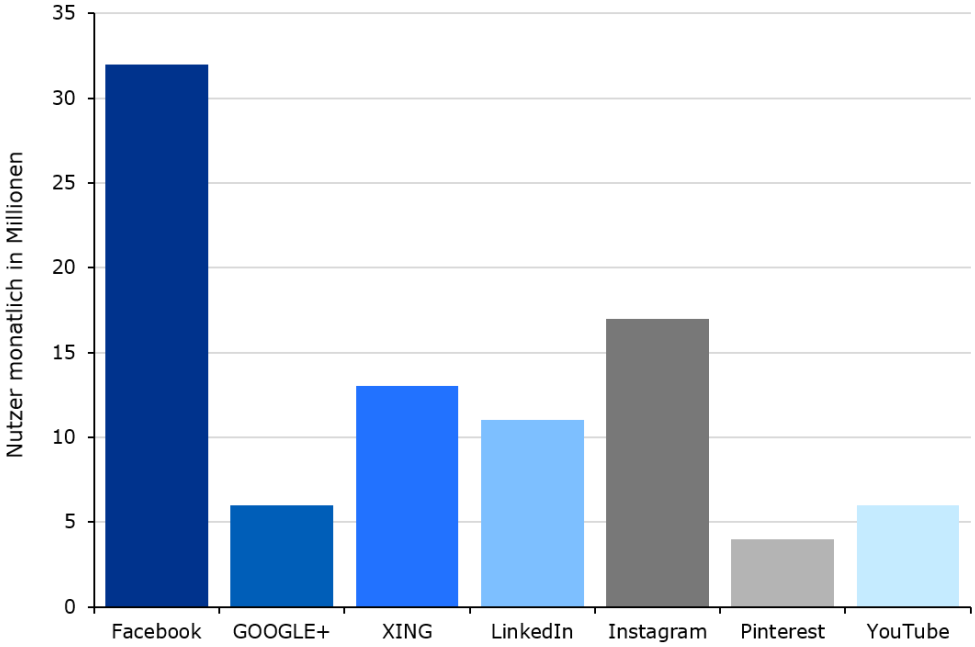

Abbildung 13: Monatl. Nutzung Sozialer Medien (Quelle: https://www.kontor4.de/beitrag/aktuelle-social-media-nutzerzahlen.html)

Allerdings können nur nach entsprechender Mitwirkung des Versicherungsnehmers detaillierte personenbezogene Daten erhoben werden.

Nutzt der Versicherer beispielsweise auf eigenen Websites „Social Login" – Funktionen der entsprechenden Netzwerke, kann er grundsätzlich auf Profilinformation des Versicherungsnehmers zugreifen und hierdurch in gewissem Umfang strukturierte Zusatzinformationen erhalten.

Stimmt der Versicherungsnehmer darüber hinaus einer virtuellen „Freundschaft" mit dem Versicherer zu, erlauben grundsätzlich neuronale Netze die Verarbeitung der unstrukturierten Daten der geteilten Informationen.

Neben dem entsprechenden Aufwand für die Beschaffung der Daten wären hier aber die Nutzungsbedingungen der entsprechenden Betreiber zu beachten. Interessanterweise untersagt Facebook etwa die Nutzung der Daten für Zugangsbeschränkungen zu Produkten, wie nach dem Fall von firstcarquote.com bekannt wurde.

3.3.4. Elektronische Patientenakten

Eher perspektivisch interessant dürften elektronische Patientenakten der GKV und PKV werden. Diese sind mit einem allgemeinen Standard derzeit noch Zukunftsmusik. Zurzeit bestehen bereits einige kassen- und versichererübergreifende Initiativen wie Vivy, deren Akzeptanz bei Versicherten und somit deren Abdeckung der Bevölkerung und der Historie sich noch erweisen müssen. Bis dahin sind Zugriffe auf die digitalen Patientendaten letztlich nur unstrukturiert möglich und der Aufwand entsprechender Nutzung dürfte sehr hoch sein.

3.4. Öffentlich zugängliche personenbezogene Daten

Auch wenn personenbezogene Daten öffentlich zugänglich sind, bedeutet dies nicht, dass sie ohne weiteres durch den Versicherer verwendet werden dürfen. Neben den Erwägungen zum Datenschutz (Details in Abschnitt 3.6) sind normalerweise Nutzungsbedingungen des Anbieters zu beachten.

3.4.1. Social Media

Ein gewisser Teil der Informationen sozialer Netzwerke sind prinzipiell ohne die explizite Zustimmung des Versicherungsnehmers zugänglich. Dies betrifft insbesondere berufsbezogene soziale Medien, bei denen die Teilnehmer gerade von der Veröffentlichung profitieren.

Nur der geringste Teil davon liegt allerdings in einem öffentlich zugänglichen Bereich. Im Normalfall wird zumindest ein Nutzerkonto des Versicherers für den Zugang zu Informationen brauchbarer Detailtiefe notwendig sein. Deren Nutzung unterliegt aber dann wie im zustimmungspflichtigen Fall (s. Abschnitt 3.3.3) den Regularien des Betreibers.

3.4.2. Internet

Zuletzt könnten natürlich allgemeine Internet-Fundstellen zur Person genutzt werden. Zu beachten sind allerdings selbst hier die Nutzungsbedingungen der genutzten Suchmaschine. Aufgrund der Schwierigkeiten der zweifelsfreien Zuordnung von Fundstelle zur versicherten Person und der Heterogenität der Ergebnisse erscheint deren Nutzen letztlich zweifelhaft.

3.5. Öffentliche anonymisierte Datensätze

Auch in Deutschland werden zunehmend Statistiken bzw. anonyme Daten der öffentlichen Hand veröffentlicht, oder stehen im Wege einer Auftragsauswertung zur Verfügung. Datenquellen sind beispielsweise:

- Statistisches Bundesamt
- Bundesagentur für Arbeit
- Deutsche Rentenversicherung
- Deutsches Institut für Wirtschaftsforschung mit dem Socioeconomic Panel (SOEP)

Daneben werden mit zunehmender Popularität von Machine Learning auch versicherungsspezifische Daten zu Trainings- oder Wettbewerbszwecken veröffentlicht, etwa zu den Kaggle-Wettbewerben von Prudential zu Underwriting und Storno.

3.6. Datenaufbereitung und Vervollständigung fehlender Daten (Imputation)

Bevor die Daten analysiert werden können, müssen sie in der Regel aufbereitet werden, was einen Großteil des Arbeitsaufwands darstellt:

Zunächst müssen natürlich eventuell unstrukturierte Daten strukturiert werden, manuell oder je nach Fallzahl durch Methoden der künstlichen Intelligenz unterstützt. Teilweise wird eine Kosten-Nutzen-Abwägung ergeben, dass vorhandene Daten doch nicht verwendet werden können.

Schließlich müssen die verschiedenen strukturierten Datensätze zusammengefasst werden, wozu eindeutige Schlüssel benötigt werden. Bei Lebensversicherungsprodukten kann es hier zu Problemen kommen, denn ein Vertrag kann eine VN und mehrere VP (Partnerrisikoversicherungen, Hinterbliebenenrenten) haben. Eine VP kann durch mehrere Verträge abgesichert sein. Ein Vertrag kann mehrere Risiken (Zusatzversicherungen) haben sowie mehrere Vertragsteile, die sich z. B. im Rechnungszins unterscheiden. Je nach gewähltem Schlüssel sind Angaben auf diese Ebene hin zu aggregieren und Entscheidungen zu konkurrierenden Angaben zu treffen, wenn z. B. auf die VP aggregiert wird, aber der BMI der VP sich zu den Abschlusszeitpunkten der einzelnen Verträge verändert hat.

Schließlich sind fehlende Daten ein alltägliches Problem in der Datenanalyse, welches allerdings häufig nicht erschöpfend behandelt wird[24]. Gängige Ad-hoc-Bereinigungsverfahren (siehe unten) führen meist zu Modellfehlern, die in Abhängigkeit des Umfangs fehlender Daten in Kauf genommen werden.

Bei fehlenden Daten größeren Ausmaßes sollte sich grundsätzlich über die Abhängigkeit des Fehlens von der Größe selbst und der anderen Prädiktoren Gedanken gemacht werden:

- Ist das Fehlen unabhängig von der Variable selbst und den anderen Variablen („missing completely at random", MCAR), kann der Wert mit einfachen Mitteln ergänzt werden – falls verfahrenstechnisch überhaupt notwendig –, etwa als zusätzliche Ausprägung „Unbekannt" bei kategoriellen Variablen oder mittels Mean Imputation, also der Ergänzung des Mittelwerts. Hierbei wird die Sicherheit der Vorhersage jedoch überschätzt.

  Alternativ könnten auch lediglich vollständige Datensätze ausgewertet werden („Complete Case Analysis", „Listwise Deletion"), wobei dabei natürlich entsprechend Informationsgehalt aus etwaigen anderen durchaus vorhandenen Variablen verloren geht.

- Ist das Fehlen unabhängig von der Variable selbst, aber abhängig von anderen Variablen („missing at random", MAR), dann würden die einfachen Verfahren das Ergebnis verzerren. Hier müssen fortgeschrittene Verfahren angewendet werden. So sollte zumindest eine Imputation ausgehend von einem Vorhersagemodell für die fehlende Variable ins Auge gefasst werden. Besser – wenn auch im Gesamtprozess aufwändiger – ist eine Multiple Imputation, bei der in mehreren Durchgängen imputiert, modelliert und schließlich ein Ensemble gebildet wird. Dies wird in einigen statistischen Paketen umgesetzt.

[24] Z. B. Stef van Buuren, Flexible Imputation of Missing Data, https://stefvanbuuren.name/fimd/

- Ist das Fehlen abhängig von der Variable selbst, so ergeben sich letztlich natürlich keine echten Heilungsverfahren und eventuelle Modelle müssen bezüglich ihrer Aussagekraft unter Einbeziehung des Modellfehlers betrachtet werden. Hier kann eine Imputation sogar zu grundsätzlichen Fehlern führen, z. B. im Falle der Angabe zum Rauchverhalten: Solange dies noch nicht für die Tarifierung von Risikolebensversicherungen relevant war (und das Rauchverhalten somit unbekannt), war der Raucheranteil höher. Würde man nun mit den o. g. Verfahren das Raucher/Nicht-Raucher-Kennzeichen imputieren, ergäbe sich eine überschätzte Sterblichkeit.

Zuletzt ergibt sich in einem Iterationsprozess mit der Modellierung häufig noch die Notwendigkeit, zahlenmäßig kleine Ausprägungen gewisser Variablen zusammenzufassen, um das Modell zu verbessern. Ansonsten kann es vorkommen, dass die einzelnen Klassen nicht entsprechend berücksichtigt werden.

3.7. Antidiskriminierung

In der öffentlichen und politischen Diskussion zum Einsatz von Algorithmen bzw. Künstlicher Intelligenz ist Antidiskriminierung ein zentrales Thema. Viele Verbraucherschutzorganisationen fordern, dass einzelne Personengruppen nicht durch Algorithmen diskriminiert werden. Dabei kann die Diskriminierung entweder bereits in den Vorgaben der Programmierer oder bei Machine-Learning-Algorithmen in den Trainingsdaten enthalten sein. Wenn Algorithmen ein diskriminierendes Verhalten aufweisen, ist das also meistens auf ein vorgelagertes Problem zurückzuführen. In der Debatte entsteht jedoch häufig der falsche Eindruck, dass es für den Einsatz von Algorithmen neuer Antidiskriminierungsvorgaben bedarf.

Was unter Diskriminierung zu verstehen ist und dass Diskriminierung verboten ist, ist allerdings bereits gesetzlich geregelt. Diese Regelungen gelten unabhängig vom Medium auch für den Einsatz von Algorithmen und Künstlicher Intelligenz. Verbraucher sind also bereits umfassend vor Diskriminierung geschützt - z. B. durch

- das Allgemeine Gleichbehandlungsgesetz,
- das Gesetz gegen den unlauteren Wettbewerb,
- das Gendiagnostikgesetz,
- die EU-Datenschutzgrundverordnung,
- das Bundesdatenschutzgesetz oder
- den Gleichbehandlungsgrundsatz im Versicherungsaufsichtsgesetz.

Befeuert wird die Diskussion durch bekannt gewordene Entscheidungssysteme, die systematisch bestimmte Personengruppen benachteiligt haben. Prominentes Beispiel ist ein System zur Verbrechensprognose, das eine dunkle Hautfarbe als das entscheidende Kriterium für eine hohe Rückfallwahrscheinlichkeit für Straftaten verwendete. Ein anderes System war überzeugt, dass die Begriffe „Arzt" und „Mann" im gleichen Verhältnis stehen würden wie „Krankenschwester" und „Frau", statt dem korrekten Wortpaar „Ärztin" und „Frau".

Wie solch ein unerwünschtes Verhalten von Entscheidungssystemen identifiziert und korrigiert werden kann, ist Gegenstand der Forschung zur Fairness von Algorithmen. Es existieren verschiedene Konzepte von Fairness. In der Versicherungswirtschaft ist das Prinzip der sachgerechten Differenzierung verankert. Das heißt, Personen werden unterschiedlich behandelt, wenn sie Unterschiede bei relevanten Merkmalen aufweisen.

In der wissenschaftlichen Literatur gibt es eine Vielzahl von Definitionen für Fairness-Kennzahlen, die zur Analyse von Entscheidungssystemen eingesetzt werden können.[25] Ein übliches Kriterium ist ob unterschiedliche Personengruppen von Fehlentscheidungen des Algorithmus in gleichem Ausmaß betroffen sind. Dieses wurde unter anderem in einer Studie zu Diskriminierungstendenzen eines Risiko-Scores zu Wiederholungstaten bei verurteilten Straftätern genutzt[26]:

Black Defendants		
	Low	High
Survived	990	805
Recidivated	532	1369
FP rate: 44.85		
FN rate: 27.99		

White Defendants		
	Low	High
Survived	1139	349
Recidivated	461	505
FP rate: 23.45		
FN rate: 47.72		

Abbildung 14 - Fehler 1. und 2. Art nach Hautfarbe der Verurteilten bei Klassifikation in Wiederholungswahrscheinlichkeiten hoch und niedrig.

Grundsätzlich ist es nicht ausreichend, das diskriminierende Kriterium selbst (hier: Hautfarbe) nicht in die Modellierung eingehen zulassen. Falls es nämlich stark korreliert ist mit anderen Kriterien, sucht sich das Modell diese Kriterien unabhängig von ihrem eigenen erklärenden Gehalt sozusagen als Ventil, und es kommt zu mittelbarer Diskriminierung.

Stattdessen sollten explizite Verfahren angewendet werden. Im einfachsten Fall bestehen diese in einer entsprechenden Umgewichtung der Trainingsfälle derart, dass in den geschützten Gruppen jeweils gleiche Verhältnisse der abhängigen Variablen erzeugt werden (z.B. gleiche Wahrscheinlichkeit eines Leistungsfalls für Männer und Frau). Praktisch erreicht man dies entweder durch Berücksichtigung von expliziten Gewichten im Modell oder durch geeignetes Up-/Down-Sampling der Trainingsfälle.

[25] siehe z. B. Studie „Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren" der Gesellschaft für Informatik e. V., S. 37 ff., bzw. http://www.francescobonchi.com/algorithmic_bias_tutorial.html

[26] https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

4. Anhang 1: Statistische Methoden

4.1. Grundlagen der Lerntheorie

Einführend gilt es einige grundlegende Begrifflichkeiten einzuführen, welche die Basis der hier behandelten statistischen Lernmethoden bilden.

4.1.1. Abhängige und unabhängige Variablen

Als *abhängige Variable* wird auf eine Zielgröße verwiesen, welche von einer anderen, nämlich einer *unabhängigen Variablen*, beeinflusst wird. Die unabhängige Variable wird verwendet, um eine Prognose für die abhängige Variable zu bestimmen. In der Literatur finden sich hierzu Synonyme für die abhängige Variable wie *erklärte Variable* oder *Zielgröße* und für die unabhängige Variable Synonyme wie *erklärende Variable* oder *Prädiktorvariable*.

4.1.2. Trainingsdaten und Trainingsfehler, Testdaten und Testfehler

Als *Trainingsdaten* oder auch *Trainingsbeobachtungen* $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ wird der Teil der vorliegenden Daten bezeichnet, der zum *Trainieren* der zur Anwendung stehenden statistischen Methode dient. Die Trainingsdaten beschreiben also sowohl unabhängige Variablen x_i als auch abhängige Variablen y_i. Dabei wird das statistische Modell $\hat{f}$ mit diesen Trainingsdaten gespeist, um die Modellparameter zu bestimmen. Dieser Prozess wird auch als *Modellfitting* bezeichnet, da das Modell bzw. dessen Parameter an die Daten *gefittet* und somit berechnet werden.

Als *Trainingsfehler* wird zunächst die Abweichung zwischen den vom Modell mittels der unabhängigen Trainingsdaten vorhergesagten Werte $\hat{f}(x_i) = \hat{y}_i$ und den abhängigen Trainingsdaten y_i bezeichnet. Dieser Fehler gibt demnach an, wie zuverlässig das trainierte Modell die tatsächlichen Trainingsdaten abbildet.

In der Praxis werden hier verschiedene Verlustfunktionen (loss functions) verwendet, etwa 0-1-Verlust für Klassifikationsprobleme $L(x, y) = \delta_{x,y}$, oder üblicherweise die quadratische Abweichung $L(x, y) = (x - y)^2$ für Regressionen. Der Trainingsfehler ergibt sich dann zu

$$Err(f) = \frac{1}{n} \sum L(x_i, \hat{y}_i).$$

Für den 0-1-Verlust bei Klassifikationen steht hier die Accuracy (Anteil richtiger Vorhersagen), für die quadratische Abweichung bei Regressionen der Mean Squared Error (MSE).

Als *Testdaten* werden die Daten bezeichnet, welche nicht zum Trainieren des Modells verwendet wurden und zum Testen der Modellzuverlässigkeit dienen.

Durch das Zurückhalten von Testdaten soll sichergestellt werden, dass das Modell auch mit anderen, „neuen" Daten gut funktioniert. Dazu wird analog zum Trainingsfehler der Fehler auf den Testdaten berechnet und bewertet.

Falls mehrere Modelle parallel trainiert werden sollen, um das beste Modell zu verwenden, sollte als weitere Ebene noch Validierungsdaten genutzt werden, um hiermit den jeweiligen Testfehler für die Modellauswahl zu bestimmen.

Es gibt verschiedene Verfahren zur effizienten Ausnutzung der Daten als Trainings- und Testdaten.

4.1.3. Validierungsstrategien

Der *Validation Set Approach* stellt ein simples Verfahren dar, bei dem die Beobachtungsmenge in einen Trainingsdatensatz und einen Validierungsdatensatz zufällig aufgeteilt wird. Das zu validierende Modell wird auf den Trainingsdatensatz *gefittet* und zur Prognose der unabhängigen Variablen des Validierungsdatensatzes verwendet. Anschließend wird wie oben beschrieben der Testfehler berechnet.

Ein Vorteil dieser Methode ist ihre Einfachheit. Nachteilig hingegen ist zum einen die hohe Empfindlichkeit des Schätzwertes für den Testfehler gegenüber der zufälligen Aufteilung in den Trainings- und den Validierungsdatensatz. Zum anderen kann die Tatsache, dass nur eine Stichprobe für die Schätzung des gesamten Testfehlers verwendet wird, zu einer Überschätzung des tatsächlichen Fehlers führen.

Bei einer k-fold Cross-Validation werden die Beobachtungen in k Gruppen ungefähr gleicher Größe aufgeteilt. Die erste Gruppe bildet den Validierungsdatensatz und das Modell wird auf die restlichen $k-1$ Gruppen *gefittet*. Der Testfehler wird mittels des Validierungsdatensatzes, also der ersten Gruppe, bestimmt. Diese Vorgehensweise wird für alle k Gruppen wiederholt und es ergeben sich k Ergebnisse für den Testfehler. Der Schätzwert für die sogenannte k*-fold Cross-Validation* ergibt sich durch die Mittelung aller Testfehler.

Typischerweise werden für k Werte wie $k = 5$ oder $k = 10$ verwendet. Der Spezialfall, dass $k = n$ gilt, also der Validierungssatz aus nur einer Beobachtung besteht, wird *Leave One Out Cross Validation* genannt. Hierbei (mit k=5 bzw. 10) ergibt sich ein signifikanter Rechenaufwandsvorteil gegenüber der *Leave-One-Out Cross-Validation*-Methode aufgrund einer geringeren Anzahl von *Modellfits* (5 bzw. 10 an Stelle von n).

Im Fall von Klassifikationsproblemen mit starkem Ungleichgewicht in den Ausprägungen der Zielgröße kann das Sampling der k Gruppen zusätzlich stratifiziert, d. h. separat für jede Teilmenge der Daten mit identischer Kategorie von y, ausgeführt werden. Dadurch wird eine konstante Verteilung der Zielgröße über alle Gruppen erreicht, die vor einem zu starken Einfluss des Zufallsfehlers schützt und somit die Validierungsergebnisse stabilisieren kann.

Bei Ensembleverfahren die Bagging (Bootstrap Aggregating, s. 4.6.1) verwenden, entsteht durch das Sampling für die Erstellung der einzelnen Teilmodelle (beispielsweise Bäume) ein kleiner Testdatensatz, der eben nicht zum Training herangezogen wurde. Dieser kann natürlich ebenfalls für Validierungszwecke verwendet werden. Der entsprechende Testfehler wird als *Out-of-bag error* (OOB) bezeichnet.

4.1.4. Overfitting

Von *Overfitting* wird gesprochen, wenn ein statistisches Modell einen sehr großen Testfehler relativ zu einem kleinen Trainingsfehler aufweist. In einem solchen Fall ist das Modell zu stark auf die spezifischen Muster der Trainingsdaten geprägt und abstrahiert zu wenig – man sagt auch „Das Modell lernt die Daten (auswendig)". Dies kann an zu vielen Parametern, zu wenig Trainingsdaten und zu vielen Trainingsdurchläufen liegen. Je nach Modell finden sich in der Literatur unter dem Stichwort „Regularization" viele Methoden zur Bekämpfung von Overfitting (siehe auch Abschnitt 4.6 und 4.9). Darüber hinaus sind manche Verfahren wie Random Forest (s. 4.6.2) von vorneherein relativ robust gegenüber Overfitting.

4.1.5. Regression und Klassifikation

Für die vorliegenden Methoden wird zwischen zwei Arten von Variablen unterschieden: Quantitative oder qualitative (kategoriale) Variablen. Quantitative Variablen nehmen metrische Zahlenwerte an, wohingegen qualitative Variablen Werte kategorialer Art sind, beispielsweise 0 oder 1 bei einer Zuordnung in zwei Klassen. Probleme quantitativer Art sind Regressionsprobleme; Probleme qualitativer oder kategorialer Art sind Klassifikationsprobleme.

4.1.6. Überwachtes und Unüberwachtes Lernen

Beim sogenannten überwachten Lernen (engl. *Supervised Learning*) existieren Informationen über unabhängige als auch abhängige Variablen, sodass eine Ermittlung eines Modells, mit dem eine Prognose der abhängigen Variablen mittels Information der unabhängigen Variablen direkt möglich ist.

Beim unüberwachten Lernen (engl. *Unsupervised Learning*) ist die Problemstellung anspruchsvollerer Art, da keine Informationen über die abhängigen Variablen vorliegen.

In diesem Kapitel werden hauptsächlich Methoden des überwachten Lernens beschrieben, wobei beispielsweise *künstliche neuronale Netze* auch auf Problemstellungen des unüberwachten Lernens anwendbar sind.

4.2. Grundlegende Regressionsverfahren

4.2.1. Lineare Regression

Die Methode der *linearen Regression* eignet sich für die quantitative Prognose einer Variablen und ist weit verbreitet im Bereich der statistischen Lernmethoden. Viele Anwendungen bauen auf dieser Methodik auf, wie weiter unten gezeigt wird. In der simplen Form gilt für eine lineare Beziehung:

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

wobei ein Fehlerterm ε existiert, der normalverteilt ist. Die Schätzwerte der Parameter β_0, β_1 werden typischerweise durch Minimierung der kleinsten Quadrate ermittelt, so dass

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \beta_1 \bar{x}$$

mit $\bar{y} = \frac{1}{n}\sum_{i=1}^{n}(y_i)$ und $\bar{x} = \frac{1}{n}\sum_{i=1}^{n}(x_i)$ gilt.

4.2.2. Multiple Lineare Regression

In der Praxis existieren häufig mehr als eine unabhängige Variable. Eine Alternative zu einer mehrfachen separaten Verwendung der linearen Regressionsmethode ist die *multiple lineare Regression* für p unabhängige Variablen:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

Auch hier können die Parameter β_i mittels der Methode der kleinsten Quadrate bestimmt werden. Sie werden so gewählt, dass

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i2} - \hat{\beta}_1 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2$$

minimal ist.

4.2.3. Logistische Regression

Die *logistische Regression* ist ein lineares Verfahren, mittels dem die Wahrscheinlichkeit ermittelt werden kann, mit der Y einer bestimmten Klasse angehört. Das logistische Modell simuliert die Beziehung zwischen $p(X) = \Pr(Y = 1|X)$ und X, so dass $p(X)$ die Werte 0 oder 1 annimmt. Typischerweise gilt dabei:

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}.$$

Das *Fitten* des Modells kann mittels der *Maximum-Likelihood*-Methode durchgeführt werden. Im Vergleich zur linearen Regression (siehe oben) stellt das logistische Modell eine geeignetere Methode dar, die Wahrscheinlichkeit für eine bestimmte Klassifizierung abzuschätzen. Es lässt sich außerdem zeigen, dass folgende Beziehung gilt:

$$log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X.$$

Es existiert also eine lineare Beziehung zwischen der linken Seite der obigen Gleichung und X.

4.2.4. Lineare Diskriminanzanalyse

Auch bei diesem Verfahren ist es das Ziel, die Wahrscheinlichkeit einer bestimmten Klassenzugehörigkeit zu ermitteln. Hier wird die Verteilung der einzelnen unabhängigen Variablen X für jede mögliche Klasse Y modelliert. Mittels des Satzes von Bayes kann dann umgekehrt der Schätzwert für $\Pr(Y = k|X = x)$ bestimmt werden. Vorteilhaft gegenüber dem Verfahren für die logistische Regression ist, dass im

Falle einer klaren Trennung der möglichen Klassen die zu bestimmenden Modellparameter eine höhere Stabilität aufweisen. Besonders bei einer Anzahl von mehr als zwei Klassen kommt dieses Verfahren bevorzugt zum Einsatz.

4.2.5. Generalisierte Additive Modelle

Das *Generalisierte Additive Modell* (GAM) ist eine Erweiterung der linearen (multiplen) Regression, indem es nicht-lineare Zusammenhänge zulässt.

Betrachte ein multiples lineares Regressionsmodell:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i.$$

Als Erweiterung wird nun jede lineare Komponente $\beta_j x_{ij}$ durch eine nicht-lineare Funktion $f_j x_{ij}$ ersetzt. Das Modell erhält die Form:

$$y_i = \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \varepsilon_i.$$

Für jedes X_j wird separat ein f_j bestimmt und anschließend addiert. Es existiert eine Vielzahl an Möglichkeiten, um f_j zu bestimmen, z. B. *polynomiale Regression*, *Step-Funktionen*, *Spline-Methoden* und *lokale Regression*. Für eine detaillierte Beschreibung zur genauen Anwendung sei auf *James et al.* (2013) verwiesen.

Für den Fall, dass die abhängige Zufallsgröße Y aus der Klasse der Exponentialfamilie stammt, spricht man auch von einem *Generalisierten Linearen Modell* (GLM).

Ein klarer Vorteil des GAMs ist die Möglichkeit, nicht-lineare Zusammenhänge zu erfassen. Außerdem ermöglicht die Additivität des Modells stets eine individuelle Analyse der einzelnen X_j bzgl. Y. In der garantierten Eigenschaft der Additivität liegt allerdings auch eine Limitierung des Modells, da dadurch mögliche wichtige Interaktionen unerkannt bleiben können.

4.3. Bayes-Klassifizierungsverfahren

Ziel dieser Methode zur Lösung eines Klassifizierungsproblems ist die Minimierung des Testfehlers. Dabei wird jede Testbeobachtung der Klasse mit der höchsten Zugehörigkeitswahrscheinlichkeit zugeordnet. Als Inputinformation werden dabei die unabhängigen Variablen x_0 verwendet. Jede Testbeobachtung mit unabhängigem Variablenvektor x_0 wird derjenigen Klasse j zugeordnet, dass die bedingte Wahrscheinlichkeit

$$\Pr(Y = j | X = x_0)$$

maximal ist. Im Allgemeinen ist der Fehler nach dem Bayes-Klassifizierungsverfahren gegeben durch $1 - \mathrm{EW}\left(\max_j \Pr(Y = j|X)\right)$, wobei EW für den Erwartungswert steht.

4.4. K-Nearest Neighbor

Die K-*Nearest Neighbor*-Methode ist sowohl auf Regressions- als auch auf Klassifizierungsprobleme anwendbar. Als erstes wird auf die Klassifikation eingegangen.

Typischerweise ist die bedingte Wahrscheinlichkeit von Y bzgl. gegebenen X, wie sie im Bayes-Klassifizierungsverfahren verwendet wird, nicht bekannt. Bei der Methode des K-Nearest Neighbor-Klassifizierer wird ein *Schätzwert* für die bedingte Wahrscheinlichkeitsverteilung von Y bzgl. gegebenen X ermittelt. Eine Beobachtung wird basierend auf der höchsten *geschätzten* Klassenzugehörigkeitswahrscheinlichkeit klassifiziert.

Sei K eine positive ganze Zahl und x_0 eine Testbeobachtung. Es gilt, eine Anzahl K von Trainingsbeobachtungen in der nächsten Umgebung N_0 von x_0 zu bestimmen. Dann wird die bedingte Wahrscheinlichkeit für eine Zugehörigkeit zur Klasse j geschätzt als der Anteil von Punkten in N_0, die der Klasse j angehören:

$$\Pr(Y = j|X = x_0) = \frac{1}{K}\sum_{i\epsilon N_0} I(y_i = j),$$

wobei $I(y_i = j) = 1$ falls $y_i = j$ und $I(y_i = j) = 0$ falls $y_i \neq j$. Mittels dieser Schätzung wird die Testbeobachtung x_0 gemäß dem Bayes-Verfahren klassifiziert.

Bei Anwendung auf Regressionsprobleme besitzt diese Methode gegenüber anderen Regressionsmethoden den Vorteil, dass sie keine parametrische Form des Modells $f(X)$ voraussetzt und damit flexibler ist.

Das Regressionsverfahren ist dem für die Klassifizierung sehr ähnlich. Für einen gegebenen Wert von K und einen Prognosepunkt x_0, findet die Methode des K-Nearest Neighbour die K Trainingsbeobachtungen innerhalb N_0, die am nächsten an x_0 sind. N_0 ist dabei eine Umgebung von x_0 definiert durch K am nächsten liegenden Punkten x_0 in den Trainingsdaten. Als Maß für die Nähe kann hier der euklidische Abstand verwendet werden. Dann wird $f(x_0)$ anhand des Mittelwertes aller Trainingsprognosen in N_0 ermittelt:

$$\hat{f}(x_0) = \frac{1}{K}\sum_{x_i\epsilon N_0} y_i.$$

4.5. Baumverfahren

Bei sogenannten *Baumverfahren* wird der Raum der unabhängigen Variablen in mehrere einfache Segmente aufgeteilt, die in sich bezüglich der abhängigen Variablen möglichst homogen sind. Diese Aufteilung wird anhand von sogenannten Splitregeln vorgenommen.

Bei der Verwendung von Entscheidungsbäumen muss verstärkt auf die Gefahr von Overfitting geachtet werden, sodass verschiedene Einstellungen der Hyperparameter (maximale Baumtiefe, minimale Anzahl Datensätze pro Blatt) getestet werden sollten. Generell empfiehlt sich für das Training von Entscheidungsbäumen ein ausgewogenes Set an Trainingsdaten, d. h. die Klassen sollten in den Trainingsda-

ten in ungefähr gleichen Anteilen vertreten sein. Der Einsatz von Ensemblemethoden wie Random Forests hat gegenüber einzelnen Entscheidungsbäumen einige Vorteile. So reduziert man damit Probleme mit Instabilität der gelernten Modelle bei kleineren Änderungen der Trainingsdaten und Probleme mit dem Lernen suboptimaler Modelle.

4.5.1. Entscheidungsbäume

Entscheidungsbäume werden zur Klassifikationsmodellierung herangezogen. Grundlage ist die Abfrage eines Kriteriums an jedem Knoten des Baumes, welches final zur Klassifizierung jedes Datenteils führt. Im Allgemeinen können Entscheidungsbäume mit fehlenden und nichtnumerischen Merkmalswerten, mit geringen Trainingsdaten und mit nichtlinearen Zusammenhängen umgehen.

Ziel des Entscheidungsbaums ist eine qualitative Prognose. Dabei wird angenommen, dass jede Beobachtung zur am häufigsten vorkommenden Klasse von Trainingsbeobachtungen im jeweils zugehörigen Gebiet gehört. Grundlage für die Bildung eines Entscheidungsbaums ist eine rekursive binäre Splittingmethode. Als Kriterium für einen binären Split wird der *Klassifikationsfehler* verwendet, welcher den Anteil der Trainingsbeobachtungen darstellt, der nicht zu der häufigsten vorkommenden Klasse gehört:

$$E = 1 - \max_k \hat{p}_{mk},$$

wobei $\hat{p}_{mk}$ den Teil der Trainingsbeobachtungen aus dem m-ten Gebiet und aus der k-ten Klasse darstellt. Alternativen zum *Klassifikationsfehler* sind der *Gini-Index* und die *Kreuzentropie*, welche sich als sensibler gegenüber *Knotenreinheit* erweisen. Beim sogenannten *Zurückschneiden* des Entscheidungsbaums empfiehlt sich der Klassifikationsfehler als Methode, falls die Prognosegenauigkeit des Endbaums das Ziel ist.

4.5.2. Regressionsbäume

Regressionsbäume werden bei stetigen abhängigen Variablen verwendet. Für die Konzipierung eines Regressionsbaums wird folgendermaßen vorgegangen. Gegeben seien p Inputvariablen und eine abhängige Variable für jede der N Beobachtungen mit (x_i, y_i) für $i = 1,2, \dots, N$ und $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$. Gegeben seien weiter M Gebiete $R_1, R_2, \dots, R_M$ und die abhängige Variable wird anhand einer Konstanten c_m in jedem Gebiet simuliert:

$$f(x) = \sum_{m=1}^{M} c_m I \ (x \epsilon R_m).$$

Durch Anwendung des Kriteriums der Minimierung der Summe der kleinsten Quadrate zwischen y_i und $f(x_i)$ ergibt sich für die beste Schätzung der Konstanten c_m:

$$\hat{c}_m = ave(y_i | x_i \epsilon R_m),$$

also der Mittelwert über alle y_i im Gebiet R_m. Für einen binären Split ist die Methode der kleinsten Quadrate jedoch nicht geeignet. Als geeigneter Algorithmus bietet sich folgende Vorgehensweise an:

Für eine Splitting-Variable j und einen Splitting-Punkt s lässt sich ein Paar von Halbebenen definieren:

$$R_1(j,s) = \{X | X_j \leq s\} \text{ und } R_2(j,s) = \{X | X_j > s\}.$$

Man sucht nach Splitting-Variable j und Splitting-Punkt s, welche lösen:

$$\min_{j,s} \left[\min_{c_1} \sum_{x_i \epsilon R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \epsilon R_2(j,s)} (y_i - c_2)^2 \right].$$

Für alle j, s gilt für die innere Minimierung:

$$\hat{c}_1 = ave(y_i | x_i \epsilon R_1(j,s) \text{ und } \hat{c}_2 = ave(y_i | x_i \epsilon R_2(j,s).$$

Dieser Prozess wird für den gesamten Datensatz wiederholt. Es existieren unterschiedliche Strategien, um die Baumgröße zu bestimmen. Zu große Bäume tendieren zum *Overfitting*, zu kleine könnten hingegen wichtige Merkmale unberücksichtigt lassen. Eine Möglichkeit ist einen großen Baum zu generieren und diese mittels *Cost-Complexitiy Pruning* zu verkleinern.

4.6. Ensemble Methoden

Die Idee von Ensemble Methoden ist es, mittels einer Kombination mehrerer einzelner (simpler) Modelle eine verbesserte Prognose zu erreichen. Potentielle Nachteile dieser Vorgehensweis sind eine erschwerte Interpretierbarkeit der Modelle sowie erhöhte Rechenkosten. Hierzu existieren unterschiedliche Verfahrensweisen.

4.6.1. Bagging

Das Prinzip des *Baggings* ist es, mittels *Bootstrapping* eine Vielzahl B von Trainingsdatensätzen b zu erzeugen, so dass auf alle $b \, \epsilon \, B$ die gewünschte statistische Lernmethode $\hat{f}$ angewendet wird. Anschließend wird beispielsweise über alle Ergebnisse der Mittelwert gebildet, so dass sich eine aggregierte Prognose ergibt:

$$\hat{f}_{bag}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}^{*b}(x).$$

Dieses Verfahren ist also eine Komposition von Bootstrapping und *Aggregation*, woraus sich der Name der Methode zusammensetzt. Der verbesserte Effekt ist eine Verringerung der Varianz der Prognose im Vergleich zu dem Fall, wenn die Methode nur auf einen Datensatz angewendet wird.

4.6.2. Random Forests

Die Methode des *Random Forests* ist eine weitere Verbesserung des Baggings von Baumverfahren. Es führt zu einer Dekorrelation der Bäume. Bei Entscheidungsbäumen wird beispielsweise bei jedem Baumsplit eine zufällige Teilmenge von m

Prädiktoren aus der Gesamtmenge aller Prädiktoren p betrachtet, typischerweise $m = \sqrt{p}$ oder $\mathrm{p}/3$. Durch dieses Verfahren wird eine geringere Korrelation der Prädiktoren erreicht, was eine Verringerung der Varianz über einen Entscheidungsbaum bedeutet.

In der Praxis bedeutet dies, dass die Random Forests kaum ins Overfitting laufen, wenn sie nicht zu tief (d. h. mit zu vielen Knoten pro Baum) konzipiert werden.

4.6.3. Boosting

Das *Boosting*-Verfahren ist ähnlich zu dem des Baggings, nur wird eine statistische Methode iterativ verbessert, indem sie stets auf ihrer vorherigen Version aufbaut. Es gilt deshalb als Methode des langsamen Lernens und somit als erfolgsversprechend. Im Falle von Regressionsmethoden werden beispielsweise die Residuen der Modellprognose als Information für eine iterative Modellverbesserung verwendet. Dabei wird das Modell anstatt anhand der abhängigen Variablen $\boldsymbol{Y}$ auf die Residuen der vorherigen Prognose *gefittet*. Anfangs wird dafür $\hat{f}(x) = 0$ gesetzt und für die Residuen gilt $r_i = y_i$ für alle Trainingsdaten. Das Ergebnis $\hat{f}^b$ wird mit einem sogenannten Shrinking-Faktor λ zu $\hat{f}$ hinzuaddiert, so dass sich der neue Iterationswert ergibt:

$$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(\mathrm{x}).$$

Dabei ist $b \epsilon B$ und B die Menge aller Iterationsschritte.

Es ergeben sich neue Residuen:

$$r_i \leftarrow r_i - \lambda \hat{f}^b(\mathrm{x})$$

Dieses Verfahren wird für $b = 1,2,\ldots,B$ wiederholt, so dass sich für das *geboostete* Modell ergibt:

$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x).$$

Die Wahl des Shrinking-Faktors λ erlaubt eine Kontrolle der Lerngeschwindigkeit. Bei größeren Werten für λ ist der Lernprozess langsamer und effektiver. Allerdings resultiert dies in einer größeren Menge an Iterationsschritten und erhöht dadurch die Berechnungskosten.

Bei Boosting-Verfahren muss sehr sorgfältig auf die Gefahr des Overfittings geachtet werden.

4.7. Support Vector Machine

Die *Support Vector Machine* dient zur Klassifizierung zweier Klassen und ist eine Erweiterung des *Support Vector Classifiers*, der wiederum eine Weiterführung des *Maximal Margin Classifiers* ist.

Gegeben sei ein $n \times p$-Datensatz X bestehend aus n Trainingsbeobachtungen der Dimension p, welche in die binäre Klassifikationsmenge {-1,1} fallen. Außerdem

existieren Testbeobachtungen x^* der Dimension p. Ziel ist es, ein Klassifizierungswerkzeug für die Testbeobachtungen basierend auf den Trainingsbeobachtungen in Form einer Hyperebene als Trenninstrument zu erzeugen.

Für eine Hypereben dieser Form gilt:

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) > 0, \qquad i = 1, \dots, n,$$

wobei y_i die kategorialen abhängigen Variablen darstellen. Grafisch interpretiert bedeutet dies, dass jede Testbeobachtung einer Klasse zugeordnet wird abhängig davon, auf welcher Seite der Hyperebene sie liegt.

4.7.1. Maximal Margin Classifier

Beim Prinzip des *Maximal Margin Classifiers* wird eine optimale Hyperebene dergestalt bestimmt, dass diese den größtmöglichen Abstand zu den Trainingsbeobachtungen besitzt. Die gesuchte Hyperebene löst folgendes Optimierungsproblem:

$$\max_{\beta_0, \beta_1, \dots, \beta_p} M,$$

$$\sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) > M, \quad \forall\, i = 1, \dots, n.$$

In den meisten Fällen ist jedoch eine Klassifizierung mittels einer Hyperebene, welche die Beobachtungen vollständig in zwei Hälften teilt, ungünstig. So kann eine Trennung in dieser Art und Weise zu einer sehr hohen und unerwünschten Empfindlichkeit gegenüber einzelnen Beobachtungen führen. Daher kann die Wahl einer Hyperebene auch so erfolgen, dass diese die zwei zur Wahl stehenden Klassen nicht perfekt trennt. Diese Vorgehensweise erhöht die Robustheit gegenüber einzelnen bestimmten Beobachtungen und führt zu einer verbesserten Klassifizierung der meisten Trainingsbeobachtungen.

4.7.2. Support Vector Classifier

Der *Support Vector Classifier* verfolgt diesen Ansatz. Die Hyperebene wird in diesem Fall so gewählt, dass sie das folgende modifizierte Optimierungsproblem löst:

$$\max_{\beta_0, \beta_1, \dots, \beta_p, \varepsilon_1, \dots, \varepsilon_n} M,$$

$$\sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) > M(1 - \varepsilon_i), \quad \forall\, i = 1, \dots, n,$$

$$\varepsilon_i \geq 0, \sum_{i=1}^{n} \varepsilon_i \leq C, \;\; C \geq 0$$

Der Support Vector Classifier ist beschränkt auf Fälle mit linearen Trennungen zweier Klassen. Im Falle nicht-linearer Grenzen findet die *Support Vector Machine* Anwendung.

4.7.3. Support Vector Machine

Bei der *Support-Vector-Machine*-Methode kommt eine *Kernelfunktion* $K(x_i, x_{x'})$ zum Einsatz, welche die Ähnlichkeit zweier Beobachtungen quantifiziert; $(x_i, x_{x'})$ sind dabei zwei Beobachtungen. Die *Kernelfunktion* kann beispielsweise eine polynomiale Form besitzen:

$$K(x_i, x_{i'}) = (1 + \sum_{j=1}^{P} x_{ij} x_{x'j})^d,$$

wobei d eine positive ganze Zahl darstellt. Mit $d > 1$ wird die Trennlinie zwischen den Klassen nicht-linear und flexibler wählbar. Der Kernel wird als verallgemeinerte Form des Skalarprodukts zweier Beobachtungen $(x_i, x_{x'})$ verwendet, welcher stets Teil der Lösung des Optimierungsproblems beim Support Vector Classifier ist.

4.8. Künstliche neuronale Netze

Künstliche neuronale Netze sind mehrschichtige nicht-lineare statistische Modelle, welche typischerweise durch ein Netzwerkdiagramm (Abb. 13) dargestellt werden. Für eine K-Klassen-Klassifizierung besitzt das Netzwerk die binäre Zielmessungen $Y_k,\ k = 1, .., K$ am oberen Rand. Die Merkmale Z_m werden mittels Linearkombinationen der Eingabewerte abgeleitet und die Zielmessungen Y_k werden als Linearkombinationen der Z_m modelliert:

$$Z_m = \sigma(\alpha_{0m} + \alpha_m^T X), \qquad m = 1, \dots, M,$$

$$T_k = \beta_{0k} + \beta_k^T Z, \qquad k = 1, \dots, K,$$

$$f_k(X) = g_k(T), \qquad k = 1, \dots, K,$$

wobei $Z = (Z_1, Z_2, \dots, Z_M)$ und $T = (T_1, T_2, \dots, T_K)$. Dabei ist außerdem $\sigma(v)$ eine Aktivierungsfunktion. Je nach Anwendung sind $sigmoid = 1/(1 + e^{-v})$ und $ReLU = \max(0, v)$ beliebte Kandidaten für $\sigma(v)$. Die Z_M werden auch als versteckte Werte bezeichnet, da sie nicht direkt beobachtbar sind.

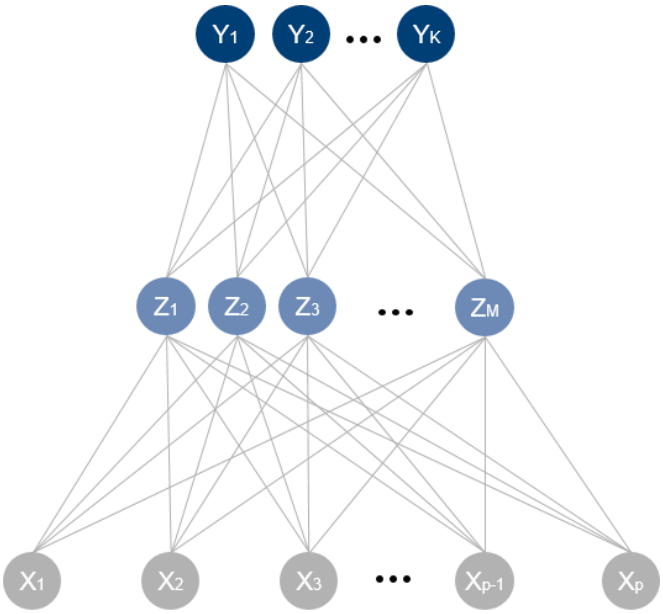

Abbildung 15: Schematische Darstellung eines neuronalen Netzwerkdiagramms mit einer versteckten Schicht (Hastie et al. 2009).

Neurale Netze verloren nach gewisser Popularität in der Anfangszeit der künstlichen Intelligenz zugunsten anderer hier genannter Verfahren an Bedeutung. Ein echtes Revival hat dieses Verfahren erst mit dem sogenannten *Deep Learning* erfahren. Hier werden sehr tiefe neuronale Netze mit je nach Anwendung sehr speziellen Topografien verwendet (etwa für Bilderkennung verschiedene Convolutional Neural Networks, oder Recurrent Neural Networks in Video- oder Spracherkennung).

4.9. Shrinkage-Methoden

Mittels *Shrinkage-Methoden* lässt sich die Varianz der geschätzten Koeffizienten in Regressionsmodellen deutlich senken, was beispielsweise die Gefahr von *Overfitting* senkt.

4.9.1. Ridge-Regression

Bei dieser Anwendungstechnik werden die sogenannten *Ridge*-Regressionskoeffizienten β^R bestimmt, dessen Werte folgenden Term minimieren:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2,$$

wobei der erste Term die Fehlerquadratsumme des linearen Regressionsmodells darstellt. Der zweite Term stellt einen sogenannten Strafterm dar, der im Falle sehr kleiner Werte der $\beta_1, \ldots, \beta_p$ gegen Null strebt. Der Parameter $\lambda \geq 0$ dient zur Kontrolle der relativen Beträge der beiden Terme auf den Schätzwert des Regressionskoeffizienten. Im Falle $\lambda = 0$ hat der Strafterm keinen Effekt und die *Ridge*-Regression ist gleich dem Schätzwert mittels der Methode der kleinsten Quadrate.

4.9.2. The Lasso

Im Unterschied zur *Ridge*-Regression basiert das *Lasso*-Verfahren prinzipiell nicht auf der Verwendung aller unabhängigen Variablen. Dies wird durch eine Auswahl von Variablen erreicht, was ein Vorteil gegenüber dem obigen Verfahren darstellt. Dadurch sind die Modelle i. A. deutlich einfach zu interpretieren. In ähnlicher Weise wird folgender Term mittels geeignet-gewählter *Lasso*-Koeffizienten $\hat{\beta}_\lambda^L$ minimiert:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|,$$

wobei $\sum|\beta_j| = \|\beta\|_1$ eine l_1-Norm vom Vektor β darstellt.

4.10. Dimensionsreduktionsverfahren

Bei diesen Verfahren ist das Ziel, die Anzahl der Dimensionen eines Datensatzes ohne signifikanten Verlust von Informationsgehalt zu reduzieren. Hier werden die unabhängigen Variablen transformiert und für ein Modell der kleinsten Quadrate verwendet.

Seien $Z_1, Z_2, \dots, Z_M,\ M < p$ Linearkombinationen der p originalen unabhängigen Variablen und es gilt:

$$Z_m = \sum_{j=1}^{p}\Phi_{jm}X_j,$$

wobei $\Phi_{1m}, \Phi_{2m}, \dots, \Phi_{pm},\ m = 1, \dots, M$. Das lineare Regressionsmodell hat folgende Form:

$$y_i = \theta_0 + \sum_{m=1}^{M}\theta_m z_{im} + \varepsilon_i, i = 1, \dots, n,$$

wobei $\theta_0, \theta_1, \dots, \theta_M$ die Regressionskoeffizienten darstellen und das Modell mittels der Methode der kleinsten Quadrate gelöst wird.

Die Dimensionsreduktion kommt dadurch zustande, dass nur $M + 1$ Koeffzienten $\theta_0, \theta_1, \dots, \theta_M$ geschätzt werden anstatt $p + 1$ Koeffzienten $\beta_0, \beta_1, \dots, \beta_p$ mit $M < p$.

4.10.1. Hauptkomponentenanalyse

Eine konkrete Methode, um einen Datensatz mit niedriger Dimension abzuleiten ist die Hauptkomponentenanalyse. Dabei wird der Datensatz in sogenannte Hauptkomponenten zerlegt, welche unterschiedlich viel der erklärten Varianz der Daten darstellen – die erste Hauptkomponente erklärt den größten Anteil.

Die Hauptkomponentenanalyse für die Regressionsanwendung sieht eine Konstruktion von M Hauptkomponenten $Z_1, \dots, Z_M$ vor, welche als unabhängige Variablen in einem linearen Regressionsmodell verwendet werden. Ziel dabei ist es, mit nur einer geringen Anzahl von Hauptkomponenten die Varianz im Datensatz und die Beziehung zu den abhängigen Variablen zu erfassen. Die Verwendung von nur

wenigen Hauptkomponenten liefert dabei ein ausreichend gutes approximatives Ergebnis.

Die erste Hauptkomponente eines Satzes von Zufallsvariablen $X_1, \dots, X_p$ lässt sich als normalisierte Linearkombination der Variablen darstellen:

$$Z_1 = \Phi_{11}X_1 + \Phi_{21}X_2 + \cdots + \Phi_{p1}X_p.$$

Aufgrund der Normalisierung gilt:

$$\sum_{j=1}^{p} \Phi_{j1}^2 = 1.$$

Für die Bestimmung der $\Phi_{11}, \Phi_{21}, \dots, \Phi_{p1}$ gilt es folgendes Optimierungsproblem beispielsweise mittels einer Eigenvektorzerlegung zu lösen:

$$\max_{\Phi_{11},\Phi_{21},\dots,\Phi_{p1}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \Phi_{j1} x_{ij} \right)^2 \right\}.$$

Darauf aufbauend können alle restlichen Hauptkomponenten bestimmt werden.

4.11. Ereigniszeitanalysen / Survival analysis

Im Gegensatz zu klassischen aktuariellen Methoden, bei denen im Wesentlichen die Wahrscheinlichkeit eines binären Ereignisses untersucht wird, interessiert sich die Ereignisdatenanalyse für die Dauer bis zum Eintreten eines Ereignisses.

In den aktuariellen Anwendungen wird diese häufig primär alters- und nicht wie in der Medizinstatistik laufzeitabhängig sein. Die entsprechenden Wanderungsbewegungen im Portfolio berücksichtigt man in der aktuariellen Ereigniszeitanalyse durch die Modellierung des Portfolios als (links) abgeschnittene, rechtszensierte Daten.

Ein Vorteil der entsprechenden Verfahren kann, je nach Datengrundlage und Implementation, in geringeren Datenvolumina und Rechenzeiten liegen, da die Ausgangsdaten nicht zunächst nach Alter und ggf. anderen Zeitdimension geschnitten werden müssen. Des Weiteren werden in entsprechenden Schätzungen nicht Fehler bzgl. der einzelnen Jahre bzw. Alter für sich, sondern bzgl. der gesamten Überlebenszeit minimiert.

Bezüglich der Ereigniszeitanalyse gibt es eine Vielzahl an Modellen, die geschlossene Formeln für die Überlebensdauer bzw. Sterblichkeit annehmen, so genannte parametrische Modelle, wie etwa Gompertz-Makeham. Hier besteht die Aufgabe im Wesentlichen in der Schätzung der Parameter auf Grundlage der Daten.

Im Folgenden sind nur die nicht-parametrischen Ansätze beschrieben.

4.11.1. Kaplan-Meier

Der bekannte Kaplan-Meier-Schätzer ist ein Maximum-Likelihood-Schätzer der Überlebensfunktion $S(t) = P(\tau > t)$ mit dem Todeszeitpunkt τ zu einer gegebenen

Abfolge von Ereignissen (t_i, c_i) zu Zeitpunkten t_i, entweder zensiert ($c_i = 1$, etwa Storno, Ablauf) oder nicht-zensiert ($c_i = 0$, z.B. Tod), und ergibt sich zu

$$\hat{S}(t) = \prod_{i:\, t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

mit der Anzahl der Tode d_i im und Lebenden n_i bis zum Zeitpunkt t_i.

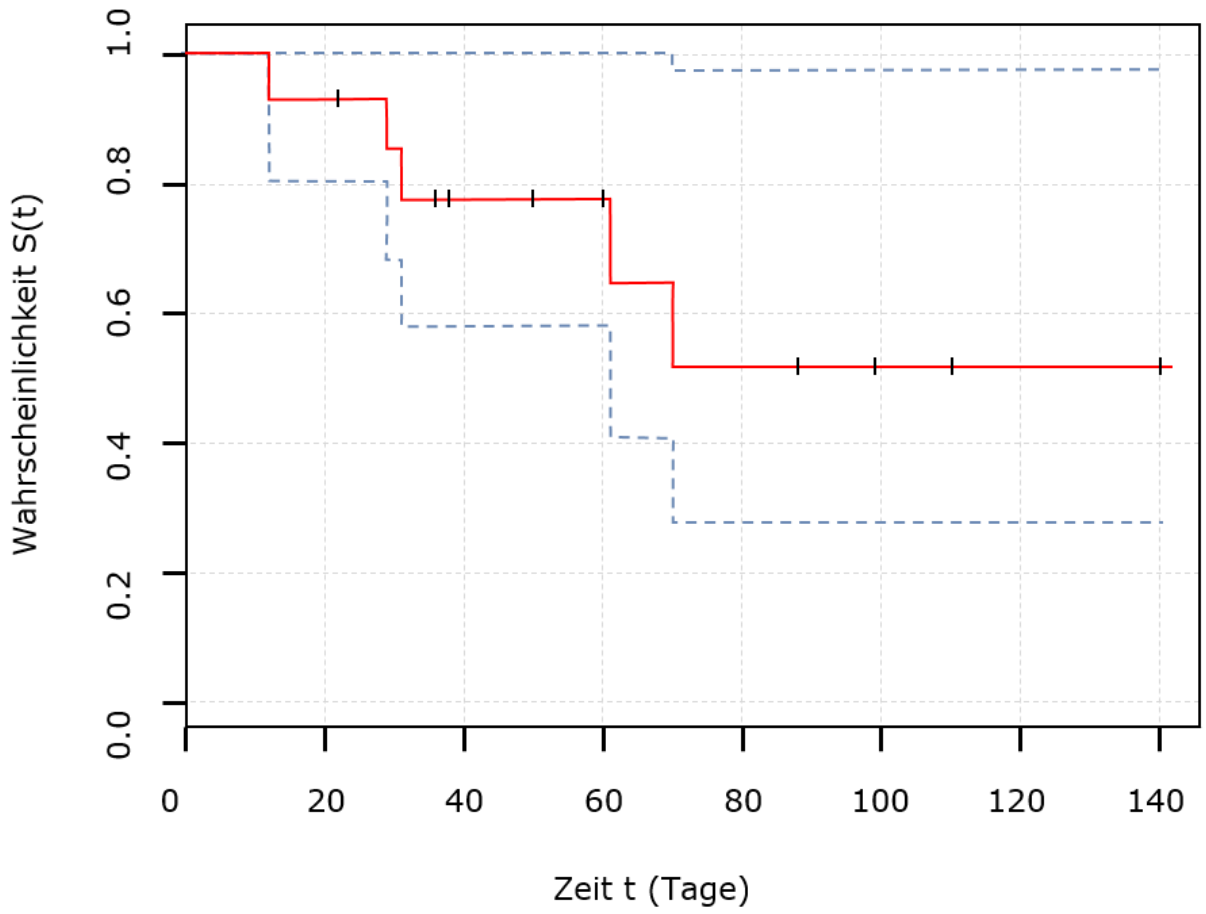

Abbildung 16: Kaplan-Meier-Schätzer

4.11.2. Cox-Regression

Bezüglich mehrerer unabhängiger Variablen geht die Cox-Regression davon aus, dass die ursprüngliche Intensität $\mathrm{h(t)}$ (in diesem Zusammenhang Hazard-Funktion genannt) über den Parameter $\beta = (\beta_1, \ldots, \beta_m)$ wie folgt von m Kovariablen $z_1, \ldots, z_m$ beeinflusst wird:

$$h(t, z) = h(t)e^{\beta z}.$$

Der Parameter β kann hierbei nicht direkt als Maximum-Likelihood-Schätzer für die Überlebensdauer bestimmt werden, da in den verschiedenen Zeitpunkten ohne Tode der Einfluss der Kovariablen unbestimmt ist. Stattdessen wird eine partielle Maximum-Likelihood geschätzt für die bedingten Wahrscheinlichkeiten, dass zu einem bestimmten Todeszeitpunkt von allen Überlebenden genau dieses Risiko stirbt (und vorausgesetzt, dass zu einem Zeitpunkt auch nur höchstens ein Risiko stirbt).

Das ursprüngliche Modell wurde dann dahingehend erweitert, dass auch mehrere Tode zum gleichen Zeitpunkt erlaubt werden, dass die Kovariablen zeitabhängig sein können und dass abgeschnittene Daten erlaubt werden.

Wie bei GLMs sind die Ergebnisse der Regression intuitiv verständlich. So lässt sich bei normalisierten Werten für die Kovariablen $z = (z_{1,}\dots, z_m)$ deren relativer Einfluss auf die Sterblichkeit direkt an den Werten für e^{β_i} ablesen, vergleichbar mit der Feature Importance von Baumverfahren. Bei dichotomen Variablen $z_i \in \{0,1\}$ ist e^{β_i} der multiplikative Faktor auf die Sterblichkeit für das Vorliegen des entsprechenden Einflusses $z_i = 1$.

Cox-Regressionen sind implementiert in SAS, R (z. B. „survival"-Paket) und Python („lifelines"-Bibliothek). Der Kaplan-Meier-Schätzer lässt sich hierbei als Baseline Hazard mit ausgeben.

4.11.3. Random Survival Forests

Eher jüngeren Datums sind die Verknüpfung von Ereigniszeitanalysen mit baumbasierten Verfahren. Eine gewisse Popularität haben hier Random Survival Forests (RSF) erfahren, letztlich eine Regression per Random Forest auf die Überlebensdauer unter Berücksichtigung der zensierten Daten.

Von Anwendung und Nutzen stehen diese im Verhältnis zur Cox-Regression wie lineare Modelle zu Random Forests, d. h. nichtlineare Effekte und Wechselwirkungen etc. können wesentlich schneller modelliert werden.

In der Interpretation steht genau wie für normale Random Forests die Feature Importance zur Verfügung.

Abweichungen zu den üblichen Random Forests als Regressionsverfahren bestehen bezüglich der Splitting Rules (z. B. logrank split für RSF) und der Fehlermaße (Harrell's C(oncordance)-Index als Standard).

4.12. Gütemaße

4.12.1. Gini-Index

Der Gini-Index beruht auf der Idee, die Modellgüte auf der Basis des Ranges, z. B. von 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

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

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

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 (in diesem Beispiel für Risiken) wird auf Basis der Lorenzkurve berechnet.

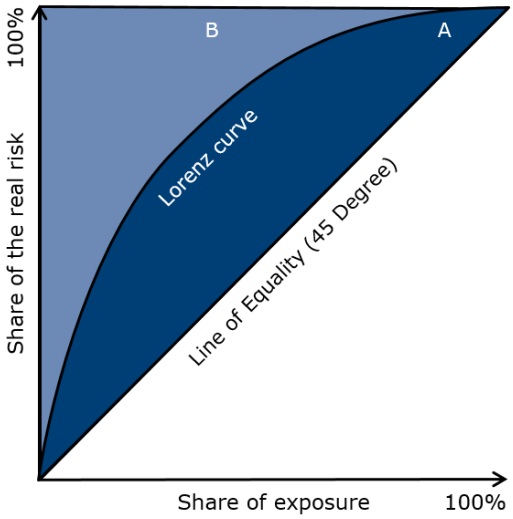

Abbildung 17: Lorenzkurve

Wenn die Fläche in Abbildung 17 zwischen der Diagonalen und der Lorenzkurve A ist und die Fläche über der Lorenzkurve B, dann ist der Gini-Index definiert als $G = \frac{A}{(A+B)}$. Da $A + B = 0.5$, 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

$$G = 2\int_0^1 L\,(x)dx - 1.$$

Zur Bestimmung der Lorenzkurve in der Praxis ordnet man die Zeilen des Datensatzes absteigend nach der Zielvariablen (z.B. der Modellvorhersage), bestimmt für die x-Achse die Position der Datenzeile im Datensatz als Perzentil und für die y-Achse das Verhältnis von laufender Summe der Zielvariable zur Gesamtsumme über den Datensatz.

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 Zielvariable kann durch folgende Normierung verallgemeinert werden:

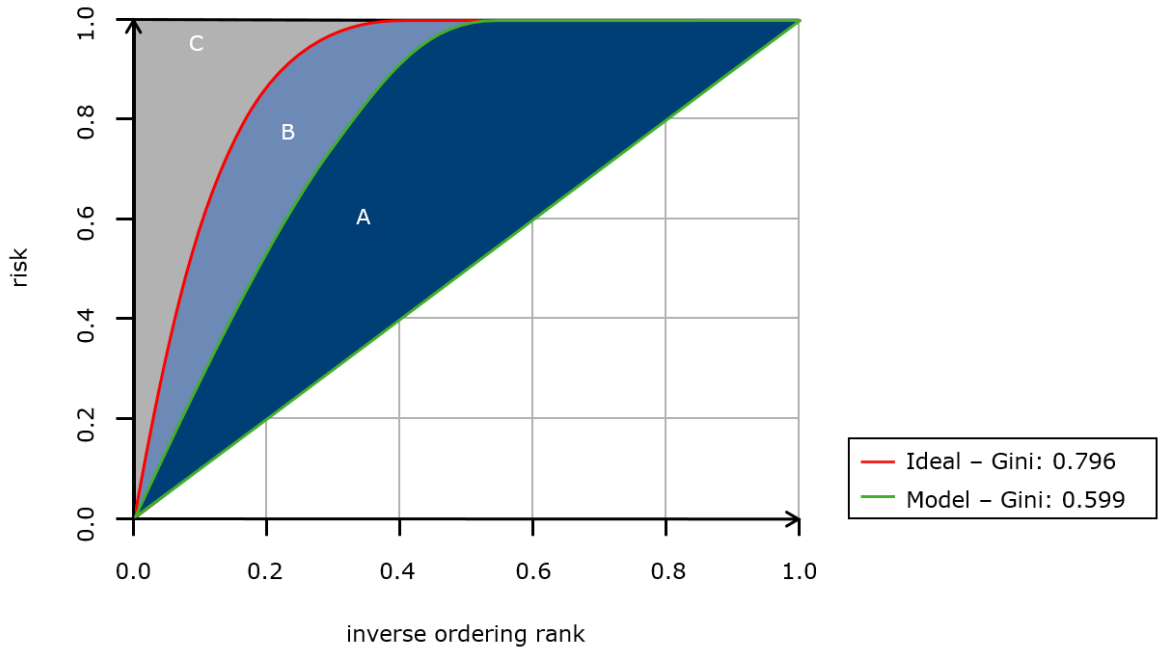

Abbildung 18: Gini-Index

Nach Sortierung der Zielvariablen absteigend nach der Höhe definiert man drei Flächen im obigen Graphen (vgl. Abbildung 18):

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 den Gini-Index zu normalisieren, betrachtet man $G = \widetilde{G_M}/G_I$, das Verhältnis von idealem Gini-Index zum modellierten Gini-Index. Daraus folgt direkt: $\tilde{G} = \frac{\mathrm{A}}{(\mathrm{A+B})}$. $\tilde{G}$ wird als normalisierter Gini-Index bezeichnet. Dieser Index liegt zwischen 0 und 1 und ist umso höher, je besser das Modell die Daten beschreibt.

Um bei einem Modell auf Overfitting zu prüfen, wird der (normalisierte) Gini-Index sowohl auf den Trainings- als auch auf den Testdaten berechnet. Die Differenz der beiden Werte ist ein Maß für das Overfitting des Modells auf den Trainingsdaten.

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

4.12.2. Konfusionsmatrix

Eine Konfusions- oder auch Wahrheitsmatrix bzgl. eines Klassifikationsmodells besteht aus den (absoluten) Häufigkeiten der tatsächlich beobachteten gegenüber den vorhergesagten Klassen und hat die Form:

Vorhersage Beobachtung	Klasse 1	Klasse 2	Klasse 3	Klasse 4
Klasse 1				
Klasse 2				
Klasse 3				
Klasse 4				

Je höher die Anzahl an Fällen auf der Diagonalen, desto besser ist die Modellanpassung an die Daten. Die Art der Fehlklassifikationen lässt sich aus den Feldern außerhalb der Diagonale ablesen. Anhand der Konfusionsmatrix lassen sich verschiedene Gütemaße, z. B. die übergreifenden, aber auch klassenspezifische und ggf. um unterschiedliche Kosten angereicherte Fehlklassifikationsraten berechnen.

Häufig betrachtet man die Konfusionsmatrix im Zusammenhang mit binären Klassifikationsproblemen der Art „ja/nein" (z. B. auch bzgl. der Zugehörigkeit zu einer bestimmten Klasse wie im obigen Fall oder im Rahmen der Regression eines Wahrscheinlichkeitswerts, wenn ein Schwellenwert gewählt wird, ab dem eine Vorhersage gemacht werden soll). In diesem Fall ist folgende Terminologie typisch:

	Vorhersage: ja	Vorhersage: nein
Beobachtung: ja	**richtig positiv**	**falsch negativ**
Beobachtung: nein	**falsch positiv**	**richtig negativ**

Aus dieser Konfusionsmatrix lassen sich neben der Missklassifikationsrate Kennzahlen zur Modellgüte rechnen, etwa Sensitivität (Anteil der korrekt positiven Vorhersagen an allen positiven Beobachtungen) und Spezifität (Anteil der korrekt negativen Vorhersagen an allen negativen Beobachtungen). Eine weitere geeignete Maßzahl, insbesondere bei Klassifikationsproblemen mit starkem Ungleichgewicht, ist die Präzision als Anteil der korrekt positiven Vorhersagen an allen positiven Vorhersagen.

4.12.3. AUC

Betrachtet man für eine Klassifikationsaufgabe die Sicherheit der Vorhersage der Klassenzugehörigkeit oder bestimmt alternativ die Wahrscheinlichkeit der Zugehörigkeit als Regressionsproblem, entsteht die Receiver-Operating-Characteristic-

Kurve (ROC-Kurve), wenn über alle Schwellenwerte (0 bis 1) für die Sicherheit/Wahrscheinlichkeit die Richtig-Positiv- gegen die Falsch-Positiv-Rate abgetragen wird (Sensitivität gegen 1-Spezifität).

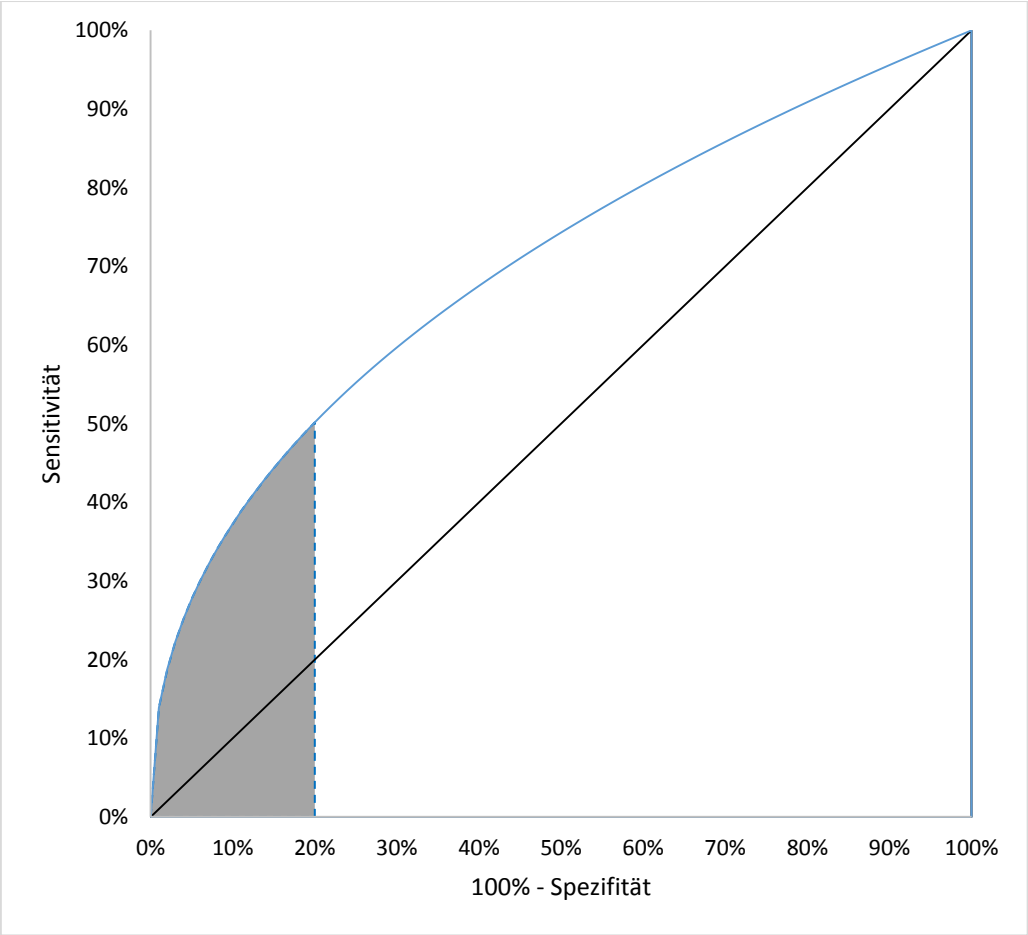

Abbildung 19: ROC-Kurve

Je weiter die Kurve oberhalb der Diagonalen liegt, desto besser ist das Modell. Dies wird über alle Schwellenwerte mit der Fläche unterhalb der ROC-Kurve (Area under the Curve, AUC) als Maßzahl zusammengefasst. Hier entspräche ein Wert von 1 einer perfekten Vorhersage, ein Wert von 0,5 einem Modell, das nicht besser als der Zufall abschneidet.

Wie beim Gini-Index ist wichtig im Auge zu behalten, dass die Maßzahl AUC nicht die absolute Höhe der Vorhersage über den Datensatz beurteilt, sondern die Differenzierung anhand der verschiedenen Kriterien. Ein Modell, bei dem die Vorhersagewahrscheinlichkeit monoton geändert wird, hat dieselbe AUC.

4.13. Interpretationsverfahren

Aus verschiedenen Gründen sind reine Black-Box-Verfahren zur Vorhersage von Zielvariablen häufig nicht ausreichend, auch wenn deren (quantitativ gemessene) Qualität sehr gut ist. Es kann einerseits darum gehen, dass Hauptzusammenhänge plausibilisiert sollen, um grobe systematische Fehler auszuschließen. Andererseits werden solche Modelle auch genutzt, um mit ihnen nicht-lineare Zusammenhänge und Erklärungen für bestimmte Effekte zu finden und gar nicht in erster Linie ein

möglichst präzises Modell zu bestimmen. Hierzu dienen verschiedene Interpretationsverfahren.

Zu guter Letzt erwarten zunehmend Regulatoren und andere Parteien von Anwendern von Maschinenlernverfahren, dass sie deren Wirkungsweise validieren und erklären können, sowie gegebenenfalls über diese Schritte Diskriminierung ausschließen können.

4.13.1. Feature Importance

Die „Feature Importance", also die Bedeutung der unabhängigen Variablen für die Vorhersage bei baumbasierten Verfahren kann durch im Wesentlichen zwei verschiedene Methoden berechnet werden.

Einerseits sehr performant ist die „Mean Decrease in Impurity (MDI)" oder Gini-Importance, bei der die Verbesserung des Gini-Koeffizienten in den einzelnen Knotenpunkten durch die entsprechende unabhängige Variable (Feature) als Splitting-Kriterium in den Entscheidungsbäumen über den Datensatz und das Ensemble aggregiert wird.

Andererseits üblicher, aber wesentlich langsamer, ist die „Perturbation Importance". Hier wird gemessen, wie sich die Vorhersage verschlechtert, wenn die entsprechende unabhängige Variable wertmäßig permutiert wird.[27] Hiervon gibt es Varianten, wie etwa die Permutation genau erfolgt.

Die Merkmale mit der höchsten Feature Importance finden sich bei beiden Verfahren nicht notwendigerweise im Wurzelbereich des Entscheidungsbaumes. Die Auswahl eines Merkmals als Entscheidungskriterium für einen Knoten erfolgt immer lokal, während die Relevanz der Merkmale im Nachhinein global berechnet wird. Zudem weicht das Maß zur Bestimmung des Entscheidungskriteriums für einen Knoten meist von dem Maß zur Bestimmung der Relevanz der Merkmale ab.

Die Relevanz misst bei Klassifikationsaufgaben, wie gut die Trennung der Datensätze gemäß einem bestimmten Merkmal auch die Klassen voneinander trennt. Beispiel:

Betrachten wir die Entscheidungskriterien Geschlecht und Alter für einen Datensatz, der 80 Anerkennungen und 20 Ablehnungen enthält. Die Aufgliederung nach dem Geschlecht (siehe Abb. 18) trägt nicht dazu bei, die Klassen „Anerkennung" und „Ablehnung" besser voneinander zu trennen. Bei Männern und Frauen sind Anerkennungen und Ablehnungen zu gleichen Anteilen vertreten wie in der Grundgesamtheit. Eine Unterscheidung nach Alter (siehe Abb. 18) hingegen macht deutlich, dass bei Jüngeren die Klasse „Anerkennung" und bei Älteren die Klasse „Ablehnung" deutlich überrepräsentiert sind. In diesem Beispiel würde für das Geschlecht eine niedrige Relevanz und für das Alter eine hohe Relevanz berechnet werden.

[27] Die besondere Bedeutung der Permutation Importance besteht darin, dass sie modellunabhängig, also auch für ganz andere Modelltypen berechnet werden kann.

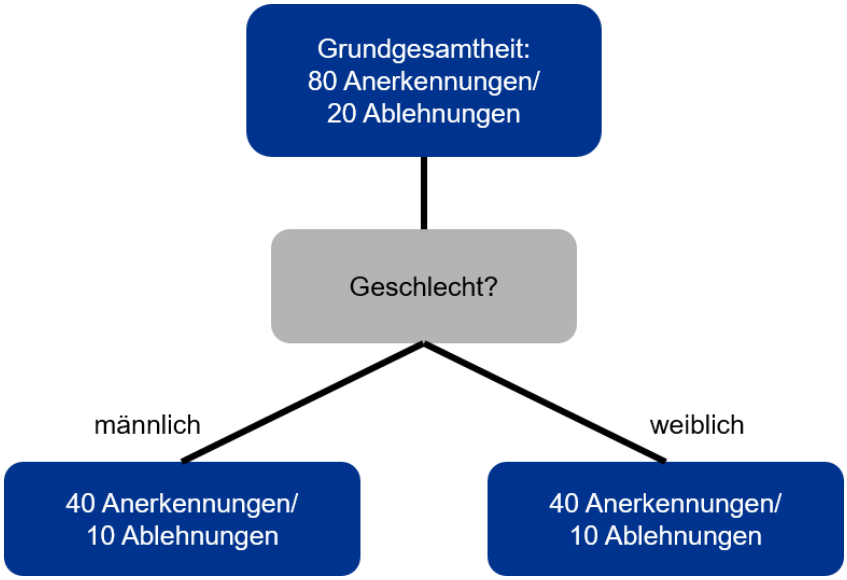

Abbildung 20: Feature Importance Bsp. 1

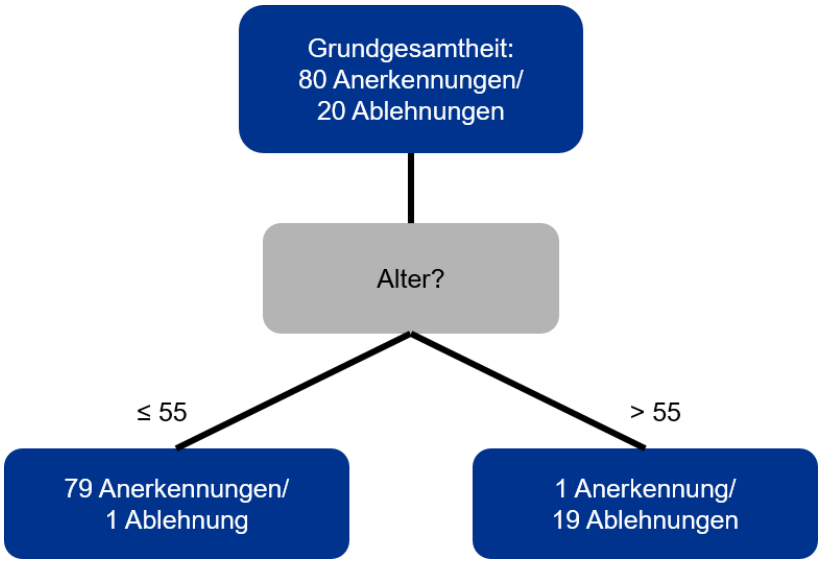

Abbildung 21: Feature Importance Bsp. 2

4.13.2. Partial Dependence Plots

Diese Plots stellen die Abhängigkeit des Modelloutputs von einer (oder mehreren) erklärenden Variablen als bedingte Randverteilung der abhängigen Variablen von der (oder den) gewählten unabhängigen Variablen dar.
Sie werden wie folgt erstellt: Zunächst wird die zu analysierende unabhängige Variable innerhalb ihres Wertebereichs oder innerhalb bestimmter Grenzen schrittweise variiert. Für jeden Wert wird der Mittelwert der Vorhersage über alle Datensätze genommen, wobei der Wert der unabhängigen Variablen entsprechend geändert wird.

Erweiterungen des Plots bestehen darin, auch die einzelnen Vorhersagewerte pro Datensatz als Linie zu plotten (s. Abbildung 11 und Abbildung 12), oder bei kategoriellen unabhängigen Variablen die Vorhersagen als Box-Plots darzustellen, um die Streuung der Vorhersagen darzustellen.

4.13.3. Andere Verfahren

Explainable AI ist ein Thema, das noch stark in Entwicklung ist. Es entstehen oder etablieren sich deshalb noch laufend weitere Verfahren.

Grundsätzlich kann man hier Verfahren unterscheiden, die globale Erklärungen liefern wollen und solche die lokal, also im Datenpunkt selbst dessen Vorhersagewert erklären.

Zu den globalen Verfahren gehören die o.g. Partial Dependence Plots. Diese haben den Nachteil, dass hier auch Kombinationen verwendet werden, die in der Realität nicht vorkommen (können) wenn z.B. zwei unabhängige Variablen stark korreliert sind. Hier können alternativ sog. ALE (Accumulated Local Effect) Plots verwendet werden, bei denen der Wertebereich der zu untersuchenden Variablen in Abschnitte eingeteilt wird und anschließend nur für die Datenpunkt im entsprechenden Abschnitt die Variable selbst für die Vorhersage mit den Randwerten des Abschnitts gesetzt wird um den Einfluss der Variable zu bestimmen. Unmögliche oder unwahrscheinliche Kombinationen werden hierdurch vermieden. Allerdings besteht dieses Verfahren nur für kontinuierliche Variablen.

Andere Verfahren liefern zu einzelnen Datenpunkten Einflussgrößen für den konkreten Vorhersagewerts im Datenpunkt. Hierzu gehören insbes. LIME und SHAP.

Bei LIME (Local Interpretable Model-Agnostic Explanations) wird mit Datenvariationen rund um den Datenpunkt ein besser erklärbares Modell wie ein einfacher Entscheidungsbaum, ein GLM o.ä. mit den Vorhersagen des eigentlichen Modells trainiert. Anschließend wird das erklärbare Modell genutzt um die lokalen Effekte des eigentlich Modells zu beschreiben.

SHAP wiederum ist eine Weiterentwicklung von Shapley Values, spieltheoretisch hergeleitete Einflüssen der Variablen auf die Abweichung von der Durchschnittsvorsage des Modells. Für Details sei auf die umsetzenden Pakete und das Literaturverzeichnis verwiesen.

4.14. Modellgovernance

Selbstverständlich müssen spätestens operativ genutzte Modelle, aber auch sinnvollerweise präfinale Modelle im Entscheidungsprozess einer ordentlichen Governance unterliegen.

U.a. sollten Modellergebnisse exakt reproduzierbar sein. Um dies zu erreichen genügt es als besondere Herausforderung im Data Science Umfeld nicht, nur Ausgangsdaten und Source-Code der Analyse selbst zu archivieren bzw. weiter zu geben. Zur Herstellung einer einheitlichen Ausgangslage gehören auch folgende Aspekte:

- Benutzte Bibliotheken in den entsprechenden Versionen samt aller Abhängigkeiten
- Benutzte Seeds für zufallsabhängige Verfahren (Training/Test-Splits auf Basis von Sampling, Cross-Validation, Analysen die Bagging nutzen, etc.) – insbesondere müssen diese explizit im Code gesetzt werden
- Individuelle Konfiguration der Entwicklungsumgebung, insofern hierdurch bestimmte Pakete automatisch geladen oder andere Ausführungsbestimmungen gesetzt werden wie dies etwa in R Studio durch die .RProfile-Datei möglich ist
- Start nur in definiertem Ausgangsstatus, d.h. in neuer / eigener Sitzung der Analyse-Umgebung (etwa nach Clear Workspace in R Studio)

Grundsätzlich sollte natürlich das Ergebnis des Modells und seiner Analyse nicht wesentlich von den Seeds und kleinen Änderungen der Tuning-Parameter abhängen, ansonsten wäre von schlechter Modellqualität auszugehen. Trotzdem ist damit zu rechnen, dass Ergebnisse im Detail bei etwa lediglich selben Ausgangsdaten und selbem Modelltyp (z.B. Random Forest) abweichen.

4.15. Literaturverzeichnis

James G et al., An Introduction to Statistical Learning with Applications in R, Springer 2013

Hastie T, Tibshirani R, Friedman J, The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition, Springer 2009

Ribeiro M T, Singh S, Guestrin C, "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 2016, https://arxiv.org/abs/1602.04938.

Richards S J, Survival Models for Actuarial Work, Note for the CMI, 2011, http://www.richardsconsulting.co.uk/Survival%20Models%20for%20Actuarial%20Work.pdf

Towers Watson, Predictive Modeling for Life Insurers: Application of Predictive Modeling Techniques in Measuring Policyholder Behavior in Variable Annuity Contracts, 2010.

Ishawaran H et al, Random Survival Forests, The Annals of Applied Statistics 2008, Vol. 2, No. 3, 841–860, https://arxiv.org/pdf/0811.1645.pdf

Christoph Molnar, Interpretable Machine Learning, https://christophm.github.io/interpretable-ml-book/

5. Anhang 2: Informatik und Tools

5.1. Software, Tools und Bibliotheken

5.1.1. Anwendungen für Datenmanagement

Im Wesentlichen betreffen Datenanalysen in der Lebensversicherung keine echten Big Data, die spezielle Datenmanagementtechniken voraussetzen wie bspw. Hadoop. Unstrukturierte Daten etwa in Form von Texten oder gescannten Dokumenten z. B. aus Risiko- oder Leistungsprüfung können vom Volumen her im Regelfall problemlos in den üblichen Dateisystemen, strukturierte Daten in den vorhandenen Datenbanken, Data Warehouses oder ebenfalls in eigenen Dateien gespeichert werden.

Bezüglich des Schutzes personenbezogener Daten sind natürlich entsprechende technische Voraussetzungen zur Zugangsbeschränkung einzuhalten.

5.1.2. Anwendungen für Datenbereinigung und Analysen

Grundsätzlich stehen hier selbstverständlich verschiedenste Open-Source- und kommerzielle Systeme zur Auswahl. Bei Versichern sind dies verbreitet SAS, R oder insbesondere in letzter Zeit verstärkt Python.

Eine Sonderrolle kommt SAS dadurch zu, dass in diesem System eine eigene physische Datenverwaltung eingeschlossen ist, die einerseits in gewissem Umfang andere Anwendungen für das Datenmanagement ersetzen kann, andererseits (um einen Preis in der Laufzeit) erlaubt, Analysen auf Datenmengen weit oberhalb des Umfangs des Hauptspeichers des Rechners durchzuführen. R und Python können lediglich in-memory arbeiten, sodass entweder entsprechend großer Hauptspeicher etwa auf einem Server zur Verfügung stehen muss oder die Ausgangsdaten vorab in entsprechende Pakete zerlegt werden müssen.

Alle diese Systeme haben einen Basisfunktionsumfang, der sich durch Pakete oder Programmierumgebungen erweitern lässt. Während bei SAS ein großer Umfang an Funktionen in der Grundausstattung standardisiert und direkt nutzbar ist (außer etwa Machine-Learning-Verfahren oder auch eine Vielzahl von Datenbankanbindungen, für die jeweils zusätzliche Lizenzgebühren anfallen), lebt die Anwendung von R und Python gerade von einer Vielzahl im Grunde nicht-standardisierter und zum Teil redundanter Pakete, die von anderen Nutzern zur Verfügung gestellt werden. Insbesondere neuere akademische statistische Methoden werden dabei häufig zunächst in R implementiert, Machine-Learning- oder Textanalyseverfahren dafür in Python. Der Nachteil ist, dass je nach Anwendungsfall verschiedene Pakete in Frage kommen können, sodass im Zweifel ein höherer Rechercheaufwand besteht, und sich erst nach und nach Standards herausbilden.

Python:

Für *Python* werden generell die Bibliotheken *numpy* (Numerisches *Python*) sowie *pandas* (Panel Data) verwendet. Für statistische Methoden mit Verbindung zum Machine Learning bietet *Python* das Toolkit *scikit-learn* an.

Hierin sind nicht nur die Lernverfahren selbst, sondern auch Hilfsroutinen wie etwa die Aufteilung der vorliegenden Daten in einen Trainingsanteil und einen Testanteil als Funktion *train_test_split()* enthalten, welche eine zufällige Aufteilung vornimmt.

R:

Die Programmiersprache und -umgebung R stammt aus dem Anwendungsbereich Statistik. Daher sind Funktionalitäten von numpy und pandas bereits enthalten. Caret und MLR sind mit scikit-learn vergleichbare Pakete, welche verschiedene Hilfsroutinen zur Modellierung mitbringen, und verschiedene zugrundeliegende Machine-Learning-Pakete in einheitlicher Weise ansprechbar machen.

In R setzt sich zunehmend eine Verwendung des Tidyverse durch, einer Sammlung von Paketen wie tidyr, dplyr, ggplot2 etc., die eine gewisse Philosophie mit einander teilen. Am auffallendsten wird hierdurch die ursprüngliche R-Sprache durch Pipes zur Datenmanipulation ergänzt. Diese ermöglichen in Zusammenhang mit den Prozeduren zur Datenmanipulation die Darstellung sukzessiver Schritte als leicht lesbare Kochrezeptstruktur, etwa

```
data <- data %>%
  filter(!is.na(Cabin)) %>%
  select(Cabin)
```

statt

```
data <- select((filter(data,!is.na(Cabin)),Cabin)
```

oder in reinem R

```
data <- data$Cabin[!is.na(data!Cabin))]
```

um nur die Variable „Cabin" mit deren nicht-leeren Einträgen zu behalten.

Neben den oben gebrauchten Funktionen select() und filter sind mutate() (definieren einer neuen Variablen), `%>%`group_by()`%>%`summarise() (Berechnung von zusammengefasste Statistiken wie Median und Summe einzelner Kategorien), dplyr-joins (Zusammensetzen zweier Datensätze mit Hilfe von Schlüsseln) sowie ggplot (grafische Darstellung der Daten) nützlich, um das Statistische Lernen vorzubereiten.

5.2. Statistische Auswertung

Im Folgenden werden mögliche Anwendungsklassen in Python mittels des *scikit-learn*-Tools für die oben-beschriebenen statistischen Methoden aufgelistet.

Methode	***Python***
Lineare Regression	`sklearn.linear_model.LinearRegression()`

Multiple lineare Regression	`sklearn.linear_model.LinearRegression()` (?)
Logistische Regression	`sklearn.linear_model.LogisticRegression`
Lineare Diskriminanz-analyse	`sklearn.discriminant_analysis.Linear-DiscriminantAnalysis`
Generalisierte Additive Modelle	[nicht implementiert, z.B. pyGAM Bibliothek]
Bayes-Klassi-fizierungs-verfahren	`sklearn.naive_bayes`
K-Nearest Neighbor	`sklearn.neighbors.KNeighborsClassifier` `sklearn.neighbors.KNeighborsRegressor`
Baumverfahren	`sklearn.tree.DecisionTreeClassifier` `sklearn.tree.DecisionTreeRegressor`
Bagging (Ensemble-Methoden)	`sklearn.ensemble.BaggingClassifier` `sklearn.ensemble.BaggingRegressor`
Random Forests (Ensemble-Methoden)	`sklearn.ensemble.RandomForestClassifier` `sklearn.ensemble.BaggingRegressor`
Boosting (Ensemble-Methoden)	`sklearn.ensemble.AdaBoostClassifier` `sklearn.ensemble.AdaBoostRegressor`
Support Vector Classifier Support Vector Machine	`sklearn.svm.`**`SVC`** `sklearn.svm.`**`SVR`**
Künstliche neuronale Netze	`sklearn.neural_network.`**`MLPClassifier`** `sklearn.neural_network.`**`MLPRegressor`**
Ridge-Regression The Lasso (Shrinkage Method)	`sklearn.linear_model.`**`Ridge`** `sklearn.linear_model.`**`Lasso`**
Hauptkomponenten-analyse (Dimensionsreduktion)	`sklearn.decomposition.`**`PCA`**
Validierung Validation Set Approach	

Leave-One-Out Cross-Validation	`sklearn.model_selection.cross_validate`
k-fold Cross-Validation	`sklearn.model_selection.LeaveOneOut` `sklearn.model_selection.KFold`
Ereigniszeitanalysen	`[nicht implementiert,` Bibliotheken scikit-survival oder life lines; bisher keine Implementation von Random Survival Forests in Python]

Zur Umsetzung in R sei für Caret auf https://topepo.github.io/caret/available-models.html verwiesen. Hier gibt es oft mehrere Varianten und Implementierungen der Methoden (z. B. 16 Varianten von Random Forests). Für MLR findet sich die Dokumentation unter https://mlr.mlr-org.com/articles/tutorial/integrated_learners.html. Hier sind auch die Ereigniszeitanalysen implementiert.

5.3. Visualisierung

In SAS sind umfangreiche Möglichkeiten zur flexiblen oder auch schnellen Erstellung von Grafiken enthalten. In R und insbesondere Python werden dafür verschiedenste Pakete genutzt je nach Präferenz und Zielmedium. Für Grafiken, die in sogenannten Notebooks verwendet werden, wird häufig bokeh bzw. rbokeh genutzt, um eine gewisse Interaktivität zu erreichen. Ansonsten bietet sich in R ggplot2 an, für das es in Python neben dem Standard Matplotlib ebenfalls eine Entsprechung gibt: ggplot.

Visualisierung ist auch eine wichtige Komponente der Datenanalyse – einerseits in der explorativen Datenanalyse, bei der die Ausgangsdaten untersucht werden, andererseits in der Validierung von Modellen. Hier werden neben den oben genannten Kennzahlen auch folgende Darstellungen für eine Einschätzung der Ergebnisse genutzt:

- Decile plot / Lift chart
- Partial Dependence Plot
- Feature Importance

5.4. Literaturverzeichnis

https://scikit-learn.org/stable/

https://mlr.mlr-org.com/

https://topepo.github.io/caret/index.html

https://www.soa.org/globalassets/assets/files/research/projects/research-pred-mod-life-huet.pdf

5.5. Abbildungsverzeichnis