Universität Ulm

Fakultät für Mathematik und Wirtschaftswissenschaften

Institut für Versicherungswissenschaften

Modellierung von Übergangswahrscheinlichkeiten für Storno und Beitragsfreistellung

Masterarbeit

in Wirtschaftsmathematik

vorgelegt von
Laura Bader
am 13.09.2022

Gutachter

Prof. Dr. Jochen Ruß
Prof. Dr. Hans-Joachim Zwiesler

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1. Einleitung

Im Regelfall kalkuliert ein Versicherungsunternehmen die Beiträge unter der Annahme, dass ein Vertrag über die volle Vertragsdauer aktiv und somit beitragspflichtig ist. Für die Beitragsberechnung ist diese Annahme sinnvoll. Jedoch muss für die Kapitalstrom- und Liquiditätsplanung berücksichtigt werden, dass sich Versicherungsnehmer abweichend von dieser Annahme verhalten können.[1]
Somit ist es im Risikomanagement von Versicherungsunternehmen von enormer Wichtigkeit, dass Risiken, die vom Versicherungsnehmer ausgehen, adäquat modelliert und quantifiziert werden. Zu diesen Risiken zählen insbesondere das sogenannte Storno und die Beitragsfreistellung.

Dabei definiert man diese beiden Ausscheideordnungen wie folgt: Unter dem Stornorisiko versteht man das Risiko, dass ein Versicherungsnehmer einen Vertrag vor Ablauf der vereinbarten Vertragsdauer kündigt. Unter Beitragsfreistellung versteht man eine vom Versicherungsnehmer induzierte Vertragsänderung, die das Ende von regelmäßigen Beitragszahlungen beinhaltet, wobei auch die Versicherungssumme entsprechend angepasst wird.

Die Modellierung der Wahrscheinlichkeit einer Beitragsfreistellung bzw. Kündigung eines Vertrages erfolgt häufig durch verallgemeinerte lineare Modelle oder durch die Whittaker-Henderson Methode. Beide Verfahren sehen vor, diese Risiken unabhängig voneinander zu modellieren. So wird beispielsweise die Wahrscheinlichkeit der Beitragsfreistellung unabhängig von der Wahrscheinlichkeit der Kündigung eines Vertrages geschätzt. Jedoch ist es naheliegend, dass Beitragsfreistellung und Storno ähnlichen Einflüssen unterliegen, was allerdings durch die unabhängigen Modellierungen nicht erfasst wird.

Im Kontrast zu diesen herkömmlichen Ansätzen werden Storno und Beitragsfreistellung in dieser Arbeit gemeinsam modelliert. Konkret werden die Eintrittswahrscheinlichkeiten der beiden Risikoereignisse gemeinsam modelliert. Dies erfolgt mithilfe von Modellen, durch

[1] Vgl. Reck et al. (2022) [15]

welche die Übergangswahrscheinlichkeiten der Zustandsänderungen der Verträge modelliert werden können. Eine solche Zustandsänderung kann beispielsweise die Beitragsfreistellung eines zuvor beitragspflichtigen (aktiven) Vertrags bedeuten. Die in dieser Arbeit vorgestellten Ansätze ermöglichen dabei die Modellierung von Storno und Beitragsfreistellung zuvor aktiver Verträge und die Modellierung der Kündigung zuvor beitragsfrei gestellter Verträge in einem gemeinsamen Modell. Ausgehend davon wird evaluiert, wie vorteilhaft solch eine Modellierung im Vergleich zum herkömmlichen Ansatz ist.

Zur Modellierung der Übergangswahrscheinlichkeiten werden in dieser Arbeit maßgeblich logistische Regressionsmodelle verwendet. Diese Modellklasse bietet den großen Vorteil, dass die Ergebnisse leicht zu interpretieren sind. Mithilfe einer $\mathcal{L}^1$-Regularisierung kann zudem die Komplexität der Modelle (Anzahl der Modellparameter ungleich null) reduziert werden.
Konkret wird sich dafür in dieser Arbeit am Ansatz zur Regularisierung von logistischen Regressionsmodellen nach Reck et al. (2022) [15] orientiert. In diesem Ansatz wird die Form der $\mathcal{L}^1$-Regularisierung in Abhängigkeit von den Skalen der vorhandenen Kovariablen (ordinal oder nominal) gewählt.
Bezüglich der Modellierung der Übergangswahrscheinlichkeiten erzielt die Regularisierung eine implizite Auswahl der wichtigsten Kovariablen. Außerdem können durch den gewählten Ansatz nach Reck et al. (2022) [15] mögliche Trends innerhalb der Merkmalsausprägungen der Kovariablen erkannt werden.

Häufig geht die Interpretierbarkeit eines Modells jedoch auf Kosten der Genauigkeit der Prognosen. Aus diesem Grund untersucht diese Arbeit zudem, inwiefern genauere Prognosen der Übergangswahrscheinlichkeiten durch weniger interpretierbare Modelle erreicht werden können. Dazu werden neuronale Netze eingeführt und deren Ergebnisse diskutiert.

Die Datengrundlage für die Analysen dieser Arbeit stammt dabei von einem großen europäischen Lebensversicherer, der Geschäfte in vier Ländern tätigt. Dieser Datensatz stimmt grundsätzlich mit dem Datensatz in Reck et al. (2022) [15] überein. Anders als in der vorliegenden Arbeit wurde in Reck et al. (2022) [15] allerdings nur die Kündigung zuvor beitragspflichtiger Verträge betrachtet.

Insgesamt ist die Arbeit wie folgt aufgebaut: In Kapitel 2 werden zunächst die Grundlagen der binomialen und multinomialen logistischen Regression eingeführt. Zudem werden die verschiedenen Arten der $\mathcal{L}^1$-Regularisierung vorgestellt, bevor die Modellstrukturen, die auf logistischer Regression basieren, beschrieben werden.

Dazu werden vier Modelle eingeführt, die die zentrale Grundlage für diese Arbeit bilden und Übergangswahrscheinlichkeiten für mögliche Zustandsänderungen der Verträge (bspw. eine Änderung von *aktiv* zu *beitragsfrei*) modellieren. Die Modelle lassen sich wie folgt beschreiben:

1. Das erste Modell bedient sich an den Ideen von Hanewald et al. (2018) [10]. Es schätzt die Übergangswahrscheinlichkeiten unabhängig voneinander mithilfe mehrerer binomialer logistischer Regressionsmodelle. Folglich entspricht dieses Modell dem herkömmlichen Ansatz und dient als Referenz zur Bewertung der Güte der weiteren Modellstrukturen.
2. Ein Modell schätzt die Übergangswahrscheinlichkeiten mittels multinomialer logistischer Regression.
3. Zwei weitere Modelle basieren auf hierarchisch angeordneten binomialen logistischen Regressionsmodellen.

Darüber hinaus werden neuronale Netze, sowie deren Aufbau und die Theorie ihrer Modellanpassung erläutert. Abschließend wird im zweiten Kapitel die Devianz als Gütemaß für die Evaluation der Prognosegüte der Modellstrukturen eingeführt. Auf Grundlage dieses Gütemaßes werden die zu vergleichenden Modelle hinsichtlich ihrer Vorhersagegenauigkeit im weiteren Verlauf der Arbeit bewertet.

Das dritte Kapitel dient der Beschreibung des oben genannten Datensatzes. Neben der Erläuterung der Grundstruktur des Datensatzes finden sich in diesem Kapitel auch Analysen jener Variablen, die die im Mehrzustandsmodell enthaltenen Zustände kodieren. Darüber hinaus wird in Kapitel 3 auf die Bearbeitung des Datensatzes eingegangen. Jedoch war diese aufgrund der ausgiebigen Vorarbeit von Reck et al. (2022) [15] nur in sehr geringem Umfang notwendig.

In Kapitel 4 werden die Ergebnisse, die durch die Modellstrukturen nach Anpassung auf den Datensatz erzielt werden, vorgestellt und bewertet. Die Bewertung der Modelle erfolgt dabei insbesondere auf Grundlage der Devianz, der Dauer zur Modellanpassung und der Komplexität des resultierenden Modells.
Auch in diesem Kapitel liegt der Fokus auf den logistischen Regressionsmodellen. In diesem Zusammenhang werden zunächst die Details der Modellanpassung erläutert. Beispielsweise wird hier die Zuordnung der Arten der $\mathcal{L}^1$-Regularisierung zu den Kovariablen im Datensatz beschrieben. Ausgehend davon werden die vier bereits genannten Modellstrukturen miteinander verglichen und ihre Stärken und Schwächen diskutiert.

Das Kapitel schließt mit der Auswertung der Ergebnisse des neuronalen Netzes. Dabei werden die Prognosen des neuronalen Netzes mit jenen der logistischen Regressionsmodelle verglichen, wobei die Ergebnisse bei diesem Vergleich insbesondere vor dem Hintergrund der Modellinterpretierbarkeit diskutiert werden.

Abschließend werden in Kapitel 5 die wichtigsten Ergebnisse dieser Arbeit zusammengetragen. Im Zuge dessen werden weitere Modellierungsansätze vorgeschlagen, die Gegenstand möglicher zukünftiger Analysen sein könnten.
Im Anhang befinden sich einige zusätzliche Grafiken, die im Rahmen der Analysen erstellt wurden. Diese Grafiken wurden dorthin im Sinne der Lesbarkeit des Hauptteils dieser Arbeit verlegt.
Alle Abbildungen in den nachfolgenden Kapiteln wurden speziell für die Ausführungen und Analysen in dieser Arbeit erstellt.

2. Grundlagen und Modellstrukturen

In diesem Kapitel wird die Modellierung von Übergangswahrscheinlichkeiten thematisiert. Konkret werden zwei Klassen von Modellstrukturen betrachtet: logistische Regressionsmodelle und (künstliche) neuronale Netze, wobei der Fokus in dieser Arbeit und somit auch in diesem Kapitel auf den logistischen Regressionsmodellen liegt.
In den ersten beiden Abschnitten werden zunächst die theoretischen Grundlagen der binomialen und multinomialen logistischen Regression sowie der $\mathcal{L}^1$-Regularisierung eingeführt. Diese Konzepte werden für die Modellierung der Übergangswahrscheinlichkeiten benötigt. Darauf aufbauend werden im dritten Teil des Kapitels die Modellstrukturen, die zur Modellierung der Wahrscheinlichkeiten verwendet werden und die auf logistischer Regression beruhen, vorgestellt. Konkret werden eine Modellstruktur basierend auf unabhängigen binomialen logistischen Regressionsmodellen, ein multinomiales logistisches Regressionsmodell sowie hierarchisch strukturierte binomiale logistische Regressionsmodelle vorgestellt.
Die theoretischen Grundlagen neuronaler Netze werden daran anschließend beleuchtet. Abschließend wird auf die Gütemaße eingegangen, auf deren Grundlage die Modelle in Kapitel 4 evaluiert und verglichen werden.

2.1. Logistische Regression

Unter logistischer Regression versteht man eine Form der Regressionsanalyse, mit der die Wahrscheinlichkeit modelliert werden kann, dass die abhängige Variable einer Beobachtung im Datensatz eine bestimmte Ausprägung besitzt.
Die Zielsetzung ist dabei analog zu jener von anderen Regressionsmodellen: Zum einen soll das an die zugrunde liegende Datenmenge am besten angepasste Modell gefunden werden. Gleichzeitig soll dieses Modell aber nicht zu komplex werden und interpretierbar bleiben.
Im Gegensatz zum linearen Regressionsmodell ist die abhängige Variable bei der logistischen Regression diskret. Das Ziel logistischer Regressionsmodelle ist dabei jedoch nicht

die Modellierung der Realisierungen der abhängigen Variable. Stattdessen wird die Wahrscheinlichkeitsverteilung der abhängigen Variable modelliert.

In der Praxis bedeutet dies das Folgende: Unter der Annahme, dass

- die endliche Menge $\mathcal{K}$ die möglichen Ausprägungen der abhängigen Variable Y beschreibt und
- $x \in \mathbb{R}^p$ Realisierungen der unabhängigen Variablen sind

so schätzt die logistische Regression für jede mögliche Ausprägung $k \in \mathcal{K}$ die Wahrscheinlichkeit, dass die durch x zu erklärende abhängige Variable die Ausprägung k besitzt. Gibt es für die abhängige Variable im Datensatz insgesamt nur zwei mögliche Ausprägungen, spricht man von *binomialer logistischer Regression*, andernfalls von *multinomialer logistischer Regression*.

2.1.1. Binomiale logistische Regression

Sei die Menge der möglichen Ausprägungen der abhängigen Variable Y im Folgenden durch $\mathcal{K} = \{0, 1\}$ gegeben.

Grundsätzlich handelt es sich bei logistischen Regressionsmodellen um Spezialfälle verallgemeinerter linearer Modelle. Bei verallgemeinerten linearen Modellen werden für die Beobachtungen im Datensatz der sogenannte lineare Prädiktor $\eta = \beta_0 + x^T\beta$ und der Erwartungswert der abhängigen Variable $\mathbb{E}(Y|x)$ durch eine Link-Funktion g zueinander in Bezug gesetzt, d.h.

$$g(\mathbb{E}(Y|x)) = \eta = \beta_0 + x^T\beta$$

Mit $\beta \in \mathbb{R}^p$ wird hier der Koeffizientenvektor und mit $\beta_0 \in \mathbb{R}$ der sogenannte Intercept des Modells bezeichnet. Durch x ist weiterhin der Merkmalsvektor einer Beobachtung, der Realisierungen der unabhängigen Variablen enthält, gegeben. Für die Verteilung der abhängigen Variable Y wird dabei häufig angenommen, dass diese Teil der Exponentialfamilie ist. Bei der Exponentialfamilie handelt es sich um Verteilungen einer bestimmten Form, die im Zusammenhang mit verallgemeinerten linearen Modellen vorteilhafte Eigenschaften erfüllen.[2]

[2] Vgl. Fahrmeir et al. (2009) [6], S.190f.

Im Fall der binomialen logistischen Regression wird für die abhängige Variable Y angenommen, dass diese Bernoulli-verteilt ist, wobei die Bernoulli-Verteilung Teil der Exponentialfamilie ist. Der Erwartungswert der Bernoulli-verteilten abhängigen Variable Y wird dann durch die sogenannte Logit-Link-Funktion mit dem linearen Prädiktor verknüpft, wobei die Logit-Link-Funktion durch $g(p) = \log\left(\frac{p}{1-p}\right),\ p \in (0,1)$, gegeben ist.

Unter der Annahme $Y|x \sim Ber(p)$ mit $\mathbb{P}(Y = 1|x) = p$ gilt $\mathbb{E}(Y|x) = p$. Somit folgt mit $\mathbb{P}(Y = 0|x) = 1 - \mathbb{P}(Y = 1|x)$, dass

$$g(p) = g(\mathbb{E}(Y|x)) = g(\mathbb{P}(Y = 1|x)) = \log\left(\frac{\mathbb{P}(Y = 1|x)}{1 - \mathbb{P}(Y = 1|x)}\right) \tag{2.1}$$
$$= \log\left(\frac{\mathbb{P}(Y = 1|x)}{\mathbb{P}(Y = 0|x)}\right) = \beta_0 + x^T\beta = \eta$$

wobei der Quotient $\log\left(\frac{\mathbb{P}(Y=1|x)}{1-\mathbb{P}(Y=1|x)}\right)$ häufig als *Log-Odds* oder *Logit* bezeichnet wird.[3] Aufgrund des Logits und der Verteilungsannahme $Y|x \sim Ber(p)$ bezeichnet man diesen Fall der Regression als *binomiale logistische Regression*.

Grundsätzlich gilt: Je höher der Logit ist, desto höher ist die Wahrscheinlichkeit, dass die Zufallsvariable Y bei gegebenem Merkmalsvektor x die Ausprägung 1 annimmt. Durch Umformen der Gleichung (2.1) nach $\mathbb{P}(Y = 1|x)$ bzw. $\mathbb{P}(Y = 0|x)$ erhält man dann die Wahrscheinlichkeiten der Zugehörigkeit zu Kategorie 1 bzw. 0 für den gegebenen Merkmalsvektor x einer Beobachtung. Bei der binomialen logistischen Regression ergeben sich diese Wahrscheinlichkeiten folglich durch

$$\mathbb{P}(Y = 1|x) = \frac{\exp(\beta_0 + x^T\beta)}{1 + \exp(\beta_0 + x^T\beta)} \text{ bzw.}$$
$$\mathbb{P}(Y = 0|x) = \frac{1}{1 + \exp(\beta_0 + x^T\beta)} = 1 - \mathbb{P}(Y = 1|x)$$

wobei der Koeffizientenvektor β und der Intercept β_0 mittels Maximum-Likelihood-Methode geschätzt werden.[4] Dazu wird die Log-Likelihood-Funktion maximiert bzw. die negative Log-Likelihood-Funktion l minimiert.

[3] Vgl. Agresti (2002) [3], S.166
[4] Vgl. Friedman et al. (2010) [7], Hosmer et al. (2013) [12]

Für einen Datensatz $\mathcal{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ ergibt sich diese Funktion als

$$\begin{aligned} l(\beta_0, \beta) &= -\sum_{i=1}^{N} \Big(y_i \log\Big(\mathbb{P}(Y_i = 1|x_i; \beta_0, \beta)\Big) + (1 - y_i) \log\Big(\mathbb{P}(Y_i = 0|x_i; \beta_0, \beta)\Big)\Big) \\ &= -\sum_{i=1}^{N} \Big(y_i(\beta_0 + x_i^T \beta) - \log\Big(1 + \exp(\beta_0 + x_i^T \beta)\Big)\Big) \end{aligned}$$

wobei $x_1, \ldots, x_N \in \mathbb{R}^p$ Merkmalsvektoren und $y_i \in \mathcal{K},\ i = 1, \ldots, N$, Realisierungen der abhängigen Variable beschreiben.

Die Interpretation der daraus resultierenden Koeffizienten erfolgt ähnlich wie bei der linearen Regression: Erhöht sich der Wert der j-ten erklärenden Variable einer Beobachtung und gilt für den zugehörigen Koeffizienten $\beta_j > 0$, so erhöht sich auch der Logit, womit die Wahrscheinlichkeit des Eintritts des Ereignisses mit Ausprägung 1 steigt. Für $\beta_j < 0$ steigt wiederum die Wahrscheinlichkeit des Gegenereignisses, falls die j-te erklärende Variable erhöht wird.

2.1.2. Multinomiale logistische Regression

Im Gegensatz zur binomialen logistischen Regression sind bei der *multinomialen logistischen Regression* mehr als zwei mögliche Ausprägungen der abhängigen Variable Y der Beobachtungen möglich.[5] Im Folgenden sei deshalb durch $\mathcal{K} = \{1, \ldots, K\}$ die Menge der $K > 2$ möglichen Ausprägungen der abhängigen Variable Y gegeben.

Die traditionelle Art der multinomialen logistischen Regression erweitert das Verfahren in Abschnitt 2.1.1, indem $K - 1$ binomiale logistische Regressionsmodelle angepasst werden. Diese ergeben sich für den Merkmalsvektor $x \in \mathbb{R}^p$ einer Beobachtung im Datensatz als

$$\log\left(\frac{\mathbb{P}(Y = l|x)}{\mathbb{P}(Y = K|x)}\right) = \beta_{0l} + x^T \beta_l, \qquad l = 1, \ldots, K - 1 \tag{2.2}$$

woraus die Wahrscheinlichkeiten für die möglichen Ausprägungen $l \in \mathcal{K}$ der abhängigen Variable mit

$$\mathbb{P}(Y = l|x) = \frac{\exp(\beta_{0l} + x^T \beta_l)}{1 + \sum_{k=1}^{K-1} \exp(\beta_{0k} + x^T \beta_k)},\ l = 1, \ldots, K - 1 \tag{2.3}$$

[5] Vgl. Friedman et al. (2010) [7], Hosmer et al. (2013) [12], Vincent und Hansen (2013) [20]

und

$$\mathbb{P}(Y = K|x) = \frac{1}{1 + \sum_{k=1}^{K-1} \exp(\beta_{0k} + x^T \beta_k)} \tag{2.4}$$

folgen.[6] Durch $\beta_l \in \mathbb{R}^p,\ l = 1, \ldots, K-1$, sind dabei die Koeffizientenvektoren und durch $\beta_{0l},\ l = 1, \ldots, K-1$, die Intercepts des Modells gegeben, welche mittels Maximum-Likelihood-Methode geschätzt werden. Obwohl wie in (2.2) häufig die Wahrscheinlichkeit der Ausprägung K als Nenner der Logits verwendet wird, wären hier auch die Wahrscheinlichkeiten der anderen möglichen Ausprägungen in $\mathcal{K}$ denkbar.[7]

In einem symmetrischeren Ansatz in Anlehnung an Friedman et al. (2010) [7] und Zhu und Hastie (2004) [21] werden die Wahrscheinlichkeiten für die Ausprägungen $l \in \mathcal{K}$ einer Beobachtung mit Merkmalsvektor x durch

$$\mathbb{P}(Y = l|x) = \frac{\exp(\beta_{0l} + x^T \beta_l)}{\sum_{k=1}^{K} \exp(\beta_{0k} + x^T \beta_k)},\ l = 1, \ldots, K \tag{2.5}$$

bestimmt, wobei für jede der K Klassen ein separater Koeffizientenvektor $\beta_l \in \mathbb{R}^p$ bzw. Intercept $\beta_{0l} \in \mathbb{R}$ mittels Maximum-Likelihood-Methode ermittelt wird. Dieser Ansatz findet unter anderem Anwendung in der Implementierung in statistischer Software, wie zum Beispiel in [2]. Diese Software wird auch zur Anpassung der Modelle aus Kapitel 2.3 verwendet.

Zur Schätzung der Parametervektoren und Intercepts in (2.5) wird die negative Log-Likelihood-Funktion l auf Grundlage eines Datensatzes $\mathcal{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ mit Merkmalsvektoren $x_1, \ldots, x_N \in \mathbb{R}^p$ und Realisierungen der abhängigen Variable $y_i \in \mathcal{K},\ i = 1, \ldots, N$ minimiert. Diese Funktion lautet

$$\begin{aligned} l(\{\beta_{0l}, \beta_l\}_{l=1}^K) &= -\sum_{i=1}^{N} \sum_{l=1}^{K} z_{il} \log\Big(\mathbb{P}(Y_i = l|x_i; \beta_{0l}, \beta_l)\Big) \\ &= -\sum_{i=1}^{N} \left(\sum_{l=1}^{K} z_{il}(\beta_{0l} + x_i^T \beta_l) - \log\Big(\sum_{l=1}^{K} \exp(\beta_{0l} + x_i^T \beta_l)\Big)\right) \end{aligned}$$

wobei $z_{il} = 1$ gilt, falls $y_i = l$ und $z_{il} = 0$ sonst.[6]

[6] Vgl. Friedman et al. (2010) [7], Hastie et al. (2009) [11], S.119
[7] Vgl. Hastie et al. (2009) [11], S.119

Um die daraus resultierenden Koeffizienten interpretieren zu können, muss zunächst der Zusammenhang der Ansätze in (2.3) bzw. (2.4) und (2.5) hergestellt werden.
Ersetzt man in (2.5) die Koeffizientenvektoren β_l für $l = 1, \ldots, K$ durch $\tilde{\beta}_l = \beta_l - \beta_K$ und die Intercepts β_{0l} durch $\tilde{\beta}_{0l} = \beta_{0l} - \beta_{0K}$, so erhält man wieder die Form wie im ursprünglichen Modell in den Gleichungen (2.3) und (2.4). Die Interpretation der Koeffizienten im traditionellen Ansatz erfolgt dann wie die Interpretation der Koeffizienten bei der binomialen logistischen Regression in Abschnitt 2.1.1:

- Steigt der Wert der j-ten erklärenden Variable einer Beobachtung, so steigt der Logit der Klasse $l \in \{1, \ldots, K-1\}$, falls für den zugehörigen Koeffizienten $\beta_{jl} > 0$ gilt. Somit hat eine Erhöhung der j-ten erklärenden Variable der Beobachtung den Effekt, dass die Ausprägung l der abhängigen Variable dieser Beobachtung gegenüber der Ausprägung K wahrscheinlicher wird.
- Bei einer Erhöhung des Werts der j-ten erklärenden Variable der Beobachtung steigt zudem die Wahrscheinlichkeit gegenüber der Referenzklasse K für diejenige Ausprägung $l \in \{1, \ldots, K-1\}$ am stärksten, für die der Koeffizient β_{jl} am größten ist.
- Gilt hingegen für alle Koeffizienten $\beta_{jl} < 0,\ l = 1, \ldots, K-1$, so steigt die Wahrscheinlichkeit, dass die Beobachtung die Ausprägung K besitzt, wenn die j-te erklärende Variable erhöht wird.

Ausgehend davon können die Koeffizienten, die im symmetrischen Ansatz geschätzt werden, wie folgt interpretiert werden: Falls $\beta_{jl} - \beta_{jK} > 0$ gilt, d.h. $\beta_{jl} > \beta_{jK}$, so steigt im Falle einer Erhöhung der j-ten erklärenden Variable einer Beobachtung die Wahrscheinlichkeit für die Ausprägung l der abhängigen Variable gegenüber der Ausprägung K. Je größer die Differenz $\beta_{jl} - \beta_{jK}$ ist, desto stärker steigt die Wahrscheinlichkeit der Ausprägung l. Umgekehrt wird die Ausprägung K gegenüber der Ausprägung l im Falle einer Erhöhung der j-ten erklärenden Variable wahrscheinlicher, falls $\beta_{jl} - \beta_{jK} < 0$ gilt.

2.2. $\mathcal{L}^1$-Regularisierung

Wie bereits einführend in Abschnitt 2.1 erwähnt, ist ein großer Vorteil von logistischen Regressionsmodellen, dass die Modelle interpretierbar sind. Bei einer steigenden Anzahl unabhängiger Variablen besteht jedoch die Gefahr, dass diese Interpretierbarkeit verloren geht. *Lasso* (kurz für *least absolute shrinkage and selection operator*) ist eine Methode, die durch Verwendung von $\mathcal{L}^1$-Regularisierung einen Teil der Koeffizienten gleich null setzt

und dadurch die Interpretierbarkeit des Modells erhöht.
Da im Datensatz, der in Kapitel 3 vorgestellt wird, ausschließlich kategoriale Variablen enthalten sind, wird nachfolgend insbesondere Bezug auf Formen der $\mathcal{L}^1$-Regularisierung genommen, die sich für diese Art von Kovariablen eignen.

Analog zum von Tibshirani (1996) [17] eingeführten Lasso für lineare Regressionsmodelle kann $\mathcal{L}^1$-Regularisierung auch bei logistischen Regressionsmodellen verwendet werden. Dazu wird der $\mathcal{L}^1$-Bestrafungsterm

$$G_b(\beta) := \sum_{j=1}^{J} g_j(\beta^{(j)})$$

bei binomialer logistischer Regression bzw.

$$G_m(\beta) := \sum_{k=1}^{K} \sum_{j=1}^{J} g_j(\beta_k^{(j)})$$

bei multinomialer logistischer Regression mit der zu minimierenden negativen Log-Likelihood-Funktion addiert. Hierbei wird mit J die Anzahl der Kovariablen der Beobachtungen im Datensatz und mit K die Anzahl der möglichen Ausprägungen der abhängigen Variable der Beobachtungen bezeichnet. Insbesondere wird der Index j in g_j verwendet, um die verschiedenen Formen der $\mathcal{L}^1$-Bestrafung, die in den Abschnitten 2.2.1 und 2.2.2 eingeführt werden, kenntlich zu machen. Diese müssen für jede der J Kovariablen mit $p_j,\ j = 1, \ldots, J$, Merkmalsausprägungen spezifiziert werden, wodurch für den Koeffizientenvektor der j-ten Kovariable $\beta^{(j)} \in \mathbb{R}^{p_j}$ gilt.

Geht man wie in den Kapiteln 2.1.1 und 2.1.2 von Merkmalsvektoren $x \in \mathbb{R}^p$ aus, so gilt $J \leq p$ und $p_1 + \ldots + p_J = p$. Dem liegt zugrunde, dass kategoriale Kovariablen mithilfe sogenannter Dummy-Variablen kodiert werden. Dies erhöht, ausgehend von der ursprünglich im Datensatz enthaltenen Zahl an Kovariablen, die Anzahl der tatsächlichen unabhängigen Variablen, die für das logistische Regressionsmodell verwendet werden.

Für die negative Log-Likelihood-Funktion bei binomialer logistischer Regression ergibt sich mit Regularisierung

$$\tilde{l}_\lambda(\beta_0, \beta) = l(\beta_0, \beta) + \lambda G_b(\beta)$$

und bei multinomialer logistischer Regression

$$\tilde{l}_\lambda(\{\beta_{0k}, \beta_k\}_{k=1}^K) = l(\{\beta_{0k}, \beta_k\}_{k=1}^K) + \lambda G_m(\beta)$$

wobei die Stärke der Regularisierung durch den Parameter $\lambda \geq 0$ gesteuert werden kann.[8] Für $\lambda = 0$ erhält man so das ursprüngliche logistische Regressionsmodell ohne Regularisierung. Für $\lambda \to \infty$ werden die Koeffizienten des Modells durch die Regularisierungs-Methode zunehmend gleich null gesetzt. Dadurch ist im Extremfall lediglich noch der Intercept β_0 als einziger Koeffizient ungleich null im Modell enthalten. Dies ist der Fall, da der Intercept nicht in den Bestrafungsterm $G_b(\beta)$ bzw. $G_m(\beta)$ eingeht.

2.2.1. Lasso

Nach der ursprünglichen Form der $\mathcal{L}^1$-Regularisierung nach Tibshirani (1996) [17] erfolgt die Bestrafung der Koeffizienten der j-ten Kovariable im Datensatz mit dem zugehörigen Koeffizientenvektor $\beta^{(j)} \in \mathbb{R}^{p_j}$ durch den Lasso-Term

$$g_R(\beta^{(j)}) := \sum_{i=1}^{p_j} |\beta_i^{(j)}|$$

Einerseits resultiert das Aufaddieren dieses Terms auf die negative Log-Likelihood-Funktion darin, dass einige Parameter gleich null gesetzt werden. Somit werden Merkmale mit wenig oder keinem Einfluss auf die abhängige Variable aus dem Modell entfernt, was die Interpretierbarkeit des Modells erhöht. Andererseits werden diejenigen Parameter, die im Modell enthalten bleiben, in ihrer betragsmäßigen Größe reduziert. Dadurch wird der Einfluss einzelner unabhängiger Variablen auf die abhängige Variable verringert. Dies mindert die Gefahr des sogenannten *Overfittings*, d.h. der Überangepasstheit des Modells auf den Datensatz, auf den es angepasst wird (Trainingsdatensatz). Durch die Vermeidung von Overfitting kann im Allgemeinen die Vorhersagekraft und die Verallgemeinerungsfähigkeit des Modells verbessert werden.

2.2.2. Fused Lasso und Trend Filtering

Das im vorigen Abschnitt 2.2.1 eingeführte Lasso bestraft die Absolutbeträge der Koeffizienten, vernachlässigt dabei jedoch eine mögliche ordinale Struktur der Merkmalsausprä-

[8] Vgl. Friedman et al. (2010) [7]; Hastie et al. (2009) [11], S.125

gungen einer kategorialen Kovariable. Beispielsweise kann eine Variable mehrere Ausprägungen besitzen, denen eine gewisse Anordnung unterstellt werden kann. Für die Modellierung kann es nützlich sein, diese ordinale Struktur zu berücksichtigen. Deshalb stellt der folgende Abschnitt zwei modifizierte Arten der $\mathcal{L}^1$-Regularisierung vor. Ziel dieser Regularisierungen ist die angesprochene Berücksichtigung von ordinalen Strukturen und die verbesserte Modellierung möglicher Abhängigkeiten zwischen den Merkmalsausprägungen einer kategorialen Kovariable.

Das von Tibshirani et al. (2005) [18] eingeführte *Fused Lasso* bestraft die Differenz der Werte der Koeffizienten benachbarter Merkmalsausprägungen einer kategorialen Kovariable. Somit erfolgt die Regularisierung der Koeffizienten der j-ten Kovariable im Datensatz mit p_j Ausprägungen durch den Term

$$g_F(\beta^{(j)}) := \sum_{i=2}^{p_j} |\beta_i^{(j)} - \beta_{i-1}^{(j)}|$$

Diese Form der $\mathcal{L}^1$-Regularisierung hat zum Vorteil, dass adjazente Kategorien einer Kovariable mit wenigen Beobachtungen tendenziell ähnliche Koeffizienten besitzen, womit ein Oszillieren der Werte der Koeffizienten und damit der späteren Modellvorhersagen verhindert wird. Zudem kommt es einer Verschmelzung der benachbarten Kategorien gleich, wenn die Kategorien ähnliche Koeffizienten besitzen. Dies bringt zusätzlich den Vorteil, dass die Auswahl der korrekten Anzahl an Kategorien bei der Kodierung stetiger Kovariablen als kategoriale Variablen (vgl. Abschnitt 3.1) vor der Anpassung des Modells an Bedeutung verliert und dem Modell überlassen wird.[9]

Eine weitere Form der $\mathcal{L}^1$-Regularisierung besteht im sogenannten *Trend Filtering*, wobei die Differenzen der Steigungen der Koeffizienten benachbarter Kategorien bestraft werden. Dazu wird der Regularisierungsterm

$$g_T(\beta^{(j)}) := \sum_{i=3}^{p_j} |\beta_i^{(j)} - 2\beta_{i-1}^{(j)} + \beta_{i-2}^{(j)}|$$

verwendet.[10] Da nicht die Änderungen der Koeffizienten benachbarter Kategorien sondern die Änderungen ihrer Steigungen bestraft werden, kann das Modell monotone Strukturen in den Ausprägungen einer Kovariable erkennen. Im Gegensatz zum Fused Lasso lässt die-

[9] Vgl. Reck et al. (2022) [15]
[10] Vgl. Kim et al. (2009) [13], Tibshirani (2014) [19]

se Form der Regularisierung benachbarte Kategorien jedoch nicht miteinander verschmelzen.[11]

2.3. Modellierung von Übergangswahrscheinlichkeiten

Die folgenden Kapitel dienen der Erläuterung der Modellstrukturen, mithilfe derer Übergangswahrscheinlichkeiten bei Zustandswechseln im Mehrzustandsmodell modelliert werden können. Auf die Zustandswechsel, die modelliert werden sollen, wird genauer in Abschnitt 2.3.1 eingegangen. Die konkreten Modellstrukturen werden in den darauffolgenden Kapiteln 2.3.2 bis 2.3.4 vorgestellt und analysiert, wobei jedes der Modelle auf logistischer Regression basiert. Das unabhängige Modell in Abschnitt 2.3.2 und die in Abschnitt 2.3.4 eingeführten hierarchischen Modelle beruhen auf binomialer logistischer Regression, während in Teil 2.3.3 ein multinomiales logistisches Regressionsmodell vorgestellt wird.

Das unabhängige Modell entspricht dem aktuellen Stand der Forschung wie beispielsweise in Hanewald et al. (2018) [10] oder Renshaw und Haberman (1995) [16], bei dem für jeden möglichen Zustandswechsel im Mehrzustandsmodell ein separates Modell angepasst wird. Die Anpassung von jedem der einzelnen Modelle erfolgt dabei völlig unabhängig von der Anpassung der Modelle der anderen Zustandswechsel.

Es ist jedoch nicht auszuschließen, dass mehrere Zustandswechsel unter dem Einfluss gleicher oder ähnlicher Faktoren stehen. Deshalb bietet es sich an, auf Modellstrukturen zurückzugreifen, die Abhängigkeiten der Zustandswechsel bei der Modellierung der Übergangswahrscheinlichkeiten berücksichtigen. Während beim multinomialen Modell nur ein einziges Modell zur Modellierung aller Übergangswahrscheinlichkeiten herangezogen wird, werden bei den hierarchischen Modellen mehrere binomiale Regressionsmodelle verwendet, welche hierarchisch angeordnet werden. Das bedeutet, dass sich die finalen Schätzungen der Übergangswahrscheinlichkeiten aus den Vorhersagen der stufenweise angeordneten binomialen Regressionsmodelle zusammensetzen. Die folgenden Abschnitte sollen neben dem Aufbau der Modelle zudem mögliche Vor-und Nachteile dieser Modellstrukturen beleuchten.

[11] Vgl. Reck et al. (2022) [15]

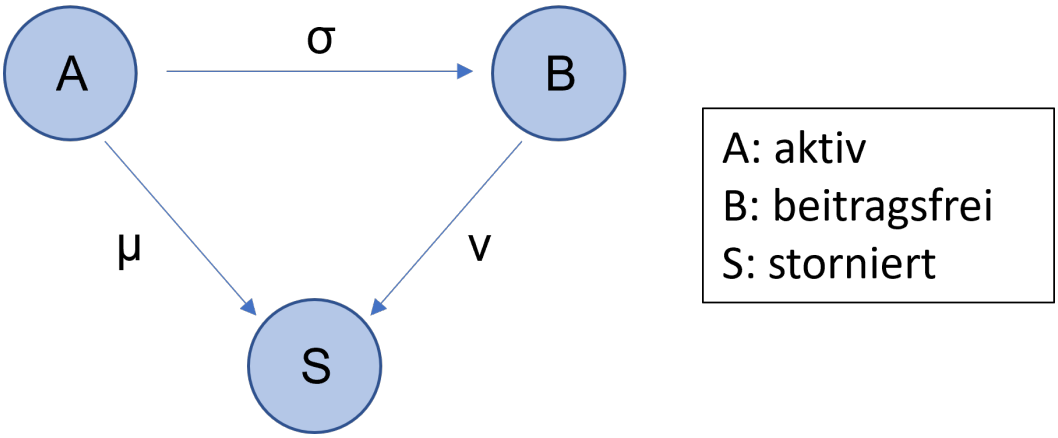


Abbildung 2.1.: Mehrzustandsmodell mit den möglichen Zustandsübergängen

2.3.1. Grundstruktur der Modelle

Um ein besseres Verständnis für die in den Abschnitten 2.3.2 bis 2.3.4 vorgestellten Modellstrukturen zu erlangen, ist es hilfreich, die in Kapitel 1 eingeführte Grundproblematik genauer zu studieren. Grundsätzlich wird in dieser Arbeit ein Mehrzustandsmodell mit den möglichen drei Zuständen der Verträge *aktiv* (A), *beitragsfrei* (B) und *storniert* (S) betrachtet. Das Ziel besteht darin, auf Grundlage der Kovariablen einer Beobachtung im Datensatz, d.h. Informationen bezüglich eines Vertrags, Vorhersagen treffen zu können, wie wahrscheinlich eine Änderung des Zustands für diesen Vertrag im betrachteten Beobachtungszeitraum (ein Jahr) ist.

Dazu wird für jede Beobachtung basierend auf dem Zustand Y_0 zu Beginn einer Beobachtungsperiode (entweder *aktiv* oder *beitragsfrei*) die Wahrscheinlichkeit für die möglichen Zustände Y_1 am Ende der Periode (*aktiv*, *beitragsfrei* oder *storniert*) geschätzt. Folglich werden für jede Beobachtung im Datensatz mithilfe der Modelle die drei Wahrscheinlichkeiten $\mathbb{P}(Y_1 = A|Y_0)$, $\mathbb{P}(Y_1 = B|Y_0)$ und $\mathbb{P}(Y_1 = S|Y_0)$ bestimmt, wobei

$$\mathbb{P}(Y_1 = A|Y_0) + \mathbb{P}(Y_1 = B|Y_0) + \mathbb{P}(Y_1 = S|Y_0) = 1$$

gelten muss.

Die Abbildung 2.1 veranschaulicht das Mehrzustandsmodell mit den möglichen Zustandswechseln, die durch die in den nachfolgenden Kapiteln vorgestellten Modelle modelliert werden. Ist ein Vertrag zu Beginn eines Jahres *aktiv*, so gibt es die möglichen Zustandswechsel zu *beitragsfrei* oder zu *storniert*, welche durch die Übergangswahrscheinlichkeiten

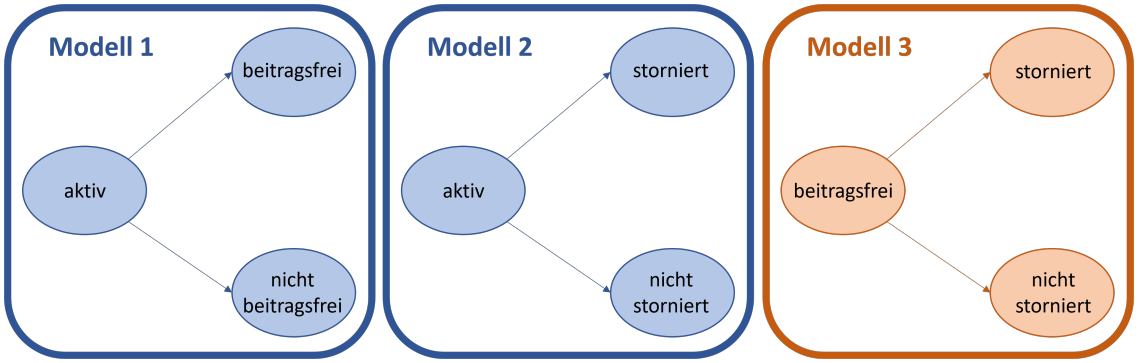


Abbildung 2.2.: Unabhängige Modellstruktur

$\sigma = \mathbb{P}(Y_1 = B|Y_0 = A)$ bzw. $\mu = \mathbb{P}(Y_1 = S|Y_0 = A)$ modelliert werden. Für einen Vertrag, der zu Beginn *beitragsfrei* ist, wird lediglich die Übergangswahrscheinlichkeit $\nu = \mathbb{P}(Y_1 = S|Y_0 = B)$ für einen Wechsel zu *storniert* modelliert. Dem liegt zugrunde, dass im Datensatz keine Beobachtung für einen Wechsel von *beitragsfrei* nach *aktiv* enthalten ist (siehe Kapitel 3) und ein solcher Zustandswechsel in der Praxis auch eher unwahrscheinlich ist. Nachfolgend werden Modelle vorgestellt, welche der Schätzung bzw. Modellierung der Übergangswahrscheinlichkeiten σ, μ und ν dienen.

2.3.2. Unabhängige Modelle

Vor allem im Bereich der Krankenversicherung finden Mehrzustandsmodelle häufig Anwendung, um Änderungen im Gesundheitszustand von Versicherten modellieren zu können.[12] Dazu wird für alle möglichen Zustandsänderungen ein separates Modell angepasst, wobei die Modellanpassung eines einzelnen Modells jeweils unabhängig von der Anpassung der Modelle der anderen Zustandsänderungen erfolgt (siehe Abbildung 2.2). Aber auch für Lebensversicherungen kann ein solches Mehrzustandsmodell verwendet werden.

Im vorliegenden Fall bedeutet das, dass drei binomiale logistische Regressionsmodelle für die Übergangswahrscheinlichkeiten σ, μ und ν angepasst werden. Dazu werden die Beobachtungen (x_i, y_i), $x_i \in \mathbb{R}^p$, $y_i \in \{A, B, S\}$, $i = 1, \ldots, N$, aus dem Datensatz $\mathcal{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ zunächst nach ihren Anfangszuständen (*aktiv* oder *beitragsfrei*) in zwei Datensätze $\mathcal{D}_A$ und $\mathcal{D}_B$ unterteilt. Die Modelle für $\sigma = \mathbb{P}(Y_1 = B|Y_0 = A)$ und $\mu = \mathbb{P}(Y_1 = S|Y_0 = A)$ werden dann nur auf Datenpunkten $(x_i, y_i) \in \mathcal{D}_A$, $i = 1, \ldots, N_A$, trainiert, für die der Anfangszustand *aktiv* ist (d.h. $Y_0 = A$). Zudem muss in

[12] Vgl. Hanewald et al. (2018) [10], Renshaw und Haberman (1995) [16]

Abhängigkeit des Zustandswechsels mit Anfangszustand *aktiv*, der modelliert werden soll, bei der Modellanpassung die Zielgröße im Datensatz angepasst werden (bspw. $Y_1 = B$ und $Y_1 = \bar{B}$ oder $Y_1 = S$ und $Y_1 = \bar{S}$). Das Modell für $\nu = \mathbb{P}(Y_1 = S|Y_0 = B)$ wird ausschließlich auf Beobachtungen $(x_i, y_i) \in \mathcal{D}_B,\ i = 1, \ldots, N_B$, mit Anfangszustand *beitragsfrei* (d.h. $Y_0 = B$) trainiert. Somit gilt $\mathcal{D} = \mathcal{D}_A \dot{\cup} \mathcal{D}_B$ und $N_A + N_B = N$.

Für die Übergangswahrscheinlichkeiten σ und μ einer Beobachtung aus dem Datensatz $\mathcal{D}_A$ mit Merkmalsvektor $x \in \mathbb{R}^p$ folgt

$$\begin{aligned}\mathbb{P}(Y_1 = l|Y_0 = A, x) &= \frac{\exp(\beta_{0j} + x^T\beta_j)}{1 + \exp(\beta_{0j} + x^T\beta_j)} \\ \mathbb{P}(Y_1 = \bar{l}|Y_0 = A, x) &= \frac{1}{1 + \exp(\beta_{0j} + x^T\beta_j)} = 1 - \mathbb{P}(Y_1 = l|Y_0 = A, x)\end{aligned}$$

wobei $l \in \{B, S\},\ \beta_j \in \mathbb{R}^p$ und $j \in \{1, 2\}$ gilt. Tritt nicht der Endzustand l ein, so wird dies mit $\bar{l}$ bezeichnet. Die Koeffizienten mit Index $j = 1$ dienen dabei der Modellierung des Zustandswechsels von *aktiv* nach *beitragsfrei* (d.h. $l = B$), während mithilfe der Koeffizienten mit Index $j = 2$ der Zustandswechsel von *aktiv* nach *storniert* modelliert wird (d.h. $l = S$). Die Übergangswahrscheinlichkeit ν für eine Beobachtung aus dem Datensatz $\mathcal{D}_B$ mit Merkmalsvektor $x \in \mathbb{R}^p$ ergibt sich analog als

$$\begin{aligned}\mathbb{P}(Y_1 = S|Y_0 = B, x) &= \frac{\exp(\beta_{03} + x^T\beta_3)}{1 + \exp(\beta_{03} + x^T\beta_3)} \\ \mathbb{P}(Y_1 = \bar{S}|Y_0 = B, x) &= \frac{1}{1 + \exp(\beta_{03} + x^T\beta_3)} = 1 - \mathbb{P}(Y_1 = S|Y_0 = B, x)\end{aligned}$$

Insgesamt hat das Modell $3p + 3$ Koeffizienten, welche mittels Maximum-Likelihood-Methode ermittelt werden. Wird bei den drei Modellanpassungen jeweils zusätzlich $\mathcal{L}^1$-Regularisierung verwendet, führt dies zu einer erheblichen Senkung der tatsächlichen Anzahl an Parametern, da durch die Regularisierung viele Koeffizienten gleich null gesetzt werden.
Für die Wahrscheinlichkeiten, dass kein Zustandswechsel für eine Beobachtung im Datensatz mit Merkmalsvektor x beobachtet wird, gilt abhängig vom Anfangszustand Y_0 der Beobachtung

$$\begin{aligned}&\mathbb{P}(Y_1 = A|Y_0 = A, x) = 1 - \mathbb{P}(Y_1 = B|Y_0 = A, x) - \mathbb{P}(Y_1 = S|Y_0 = A, x) \text{ bzw.} \\ &\mathbb{P}(Y_1 = B|Y_0 = B, x) = 1 - \mathbb{P}(Y_1 = S|Y_0 = B, x)\end{aligned} \tag{2.6}$$

Letztere Gleichung folgt, da kein Zustandswechsel von *beitragsfrei* nach *aktiv* beobachtet wird, d.h. $\mathbb{P}(Y_1 = A|Y_0 = B,\ x) = 0$ für alle Beobachtungen aus dem Datensatz $\mathcal{D}_B$ gilt.

Aus der ersten Gleichung in (2.6) geht dabei ein großer Nachteil der unabhängigen Modellstruktur hervor: Es besteht grundsätzlich die Möglichkeit, dass das Modell zur Modellierung von $\sigma = \mathbb{P}(Y_1 = B|Y_0 = A)$ und jenes zur Modellierung von $\mu = \mathbb{P}(Y_1 = S|Y_0 = A)$ für einzelne Beobachtungen aus dem Datensatz $\mathcal{D}_A$ Wahrscheinlichkeiten schätzen, die in Summe einen Wert größer als eins ergeben. Für eine solche Beobachtung mit Merkmalsvektor x gilt dann $\mathbb{P}(Y_1 = B|Y_0 = A, x) + \mathbb{P}(Y_1 = S|Y_0 = A, x) > 1$, woraus mit Gleichung (2.6) $\mathbb{P}(Y_1 = A|Y_0 = A, x) < 0$ folgt. Dies verletzt die Definition einer Wahrscheinlichkeit nach Kolmogorow.

Ein weiterer Nachteil der Modellstruktur liegt in der fehlenden Abhängigkeit der einzelnen binomialen logistischen Regressionsmodelle zur Schätzung der Übergangswahrscheinlichkeiten. Die Modelle jedes Zustandswechsels werden vollkommen unabhängig voneinander angepasst, obwohl möglicherweise mehrere Zustandswechsel durch gleiche bzw. ähnliche Faktoren beeinflusst werden. In den nachfolgenden beiden Abschnitten werden deshalb Modelle präsentiert, die mögliche gemeinsame Einflüsse der Zustandswechsel bei der Modellanpassung berücksichtigen.

2.3.3. Multinomiales Modell

Im Gegensatz zum unabhängigen Modell in Kapitel 2.3.2 und den hierarchischen Modellen in Kapitel 2.3.4 wird bei der multinomialen Modellstruktur nur ein multinomiales logistisches Regressionsmodell angepasst, um alle Übergangswahrscheinlichkeiten der Zustandswechsel zu modellieren (siehe Abbildung 2.3). Dazu wird dem Kovariablenvektor einer Beobachtung mit ursprünglich p Kovariablen der jeweilige Anfangszustand $Y_0 \in \{A, B\}$ der Beobachtung als weiteres Merkmal hinzugefügt.

Die Wahrscheinlichkeiten für die möglichen Merkmalsausprägungen *aktiv*, *beitragsfrei* und *storniert* der abhängigen Variable einer Beobachtung aus dem Datensatz mit Merkmalsvektor $x \in \mathbb{R}^{p+1}$ ergeben sich mittels multinomialer logistischer Regression durch

$$P(Y_1 = l|x) = \frac{\exp(\beta_{0l} + x^T\beta_l)}{\sum_{k=1}^{3} \exp(\beta_{0k} + x^T\beta_k)},\ l \in \{1, 2, 3\}$$

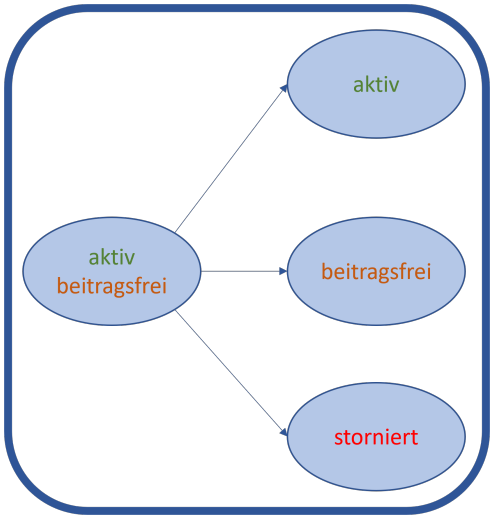


Abbildung 2.3.: Multinomiales Modell

Die möglichen Zustände $\{A, B, S\}$ am Ende eines Beobachtungsjahres werden in diesem Modell numerisch mit $\{1, 2, 3\}$ kodiert.

Durch die Maximum-Likelihood-Methode werden für das Modell $3(p + 1) + 3$ Koeffizienten geschätzt. Wie bei der unabhängigen Modellstruktur kann jedoch mittels $\mathcal{L}^1$-Regularisierung die tatsächliche Komplexität des Modells reduziert werden, da einige Koeffizienten gleich null gesetzt werden.

Der Vorteil dieses Modells besteht darin, dass nur ein Modell angepasst werden muss, mit welchem alle möglichen Übergangswahrscheinlichkeiten sowie die Wahrscheinlichkeiten des Verbleibs in einem Zustand modelliert werden können. Zudem wird durch die gemeinsame Log-Likelihood-Funktion zur Bestimmung der Koeffizienten für die Modellierung der Übergangswahrscheinlichkeiten aller Zustandswechsel eine Abhängigkeit zwischen den Zustandswechseln geschaffen, sodass mögliche gemeinsame Einflussfaktoren der Zustandswechsel berücksichtigt werden können.
Das Modell lässt jedoch grundsätzlich $\mathbb{P}(Y_1 = A|Y_0 = B) > 0$ zu, was im vorliegenden Fall gegebenenfalls zu etwas schlechteren Ergebnissen bei der Modellierung der Übergangswahrscheinlichkeiten führen kann. Tatsächlich ist dieser Effekt aber marginal, da, wie in Kapitel 4.2 genauer beleuchtet wird, die Wahrscheinlichkeiten, die das Modell für $\mathbb{P}(Y_1 = A|Y_0 = B)$ schätzt, verhältnismäßig gering sind.

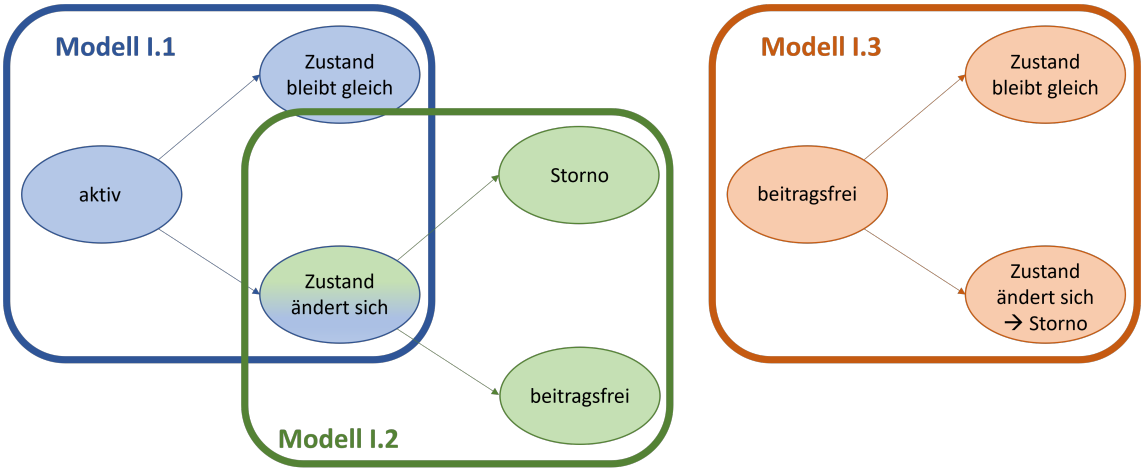


Abbildung 2.4.: Hierarchisches Modell mit Aufteilen der Daten nach dem Anfangszustand (Modell I)

2.3.4. Hierarchisch strukturierte Modelle

Eines der Ziele dieser Arbeit ist es, Modelle zu untersuchen, die ähnliche Einflussfaktoren der Zustandswechsel berücksichtigen. Das in Abschnitt 2.3.3 diskutierte multinomiale Modell stellt Abhängigkeiten bei der Modellierung der Übergangswahrscheinlichkeiten durch die Schätzung der Parameter mittels einer gemeinsamen Log-Likelihood-Funktion für alle möglichen Zustandswechsel her.

Die beiden im Folgenden vorgestellten hierarchischen Modelle bieten hierzu eine Alternative. Bei diesen beiden Modellstrukturen werden ähnliche Einflussfaktoren mehrerer Zustandswechsel herausgearbeitet, indem in einem zweistufigen Prozess zunächst die Wahrscheinlichkeit eines Zustandswechsels (aber noch nicht die genaue Art des Wechsels) geschätzt wird, bevor die Wahrscheinlichkeit für den tatsächlichen Endzustand eines Vertrags am Ende der Beobachtungsperiode ermittelt wird. Die beiden Modellstrukturen unterscheiden sich dabei in den Datensätzen, auf die sie angepasst werden. Im zuerst vorgestellten Modell wird zunächst nach den Anfangszuständen *aktiv* und *beitragsfrei* getrennt (im weiteren Verlauf als „Modell I“ bezeichnet), während bei der anschließend eingeführten Version des hierarchischen Modells nicht nach den Anfangszuständen unterteilt wird (im Folgenden als „Modell II“ bezeichnet).

Für Modell I werden insgesamt drei binomiale logistische Regressionsmodelle angepasst (siehe Abbildung 2.4). Für die erste Stufe der Modellierung werden die Beobachtungen im Datensatz $\mathcal{D}$ dazu zunächst nach den jeweiligen Anfangszuständen der Beobachtungen wie

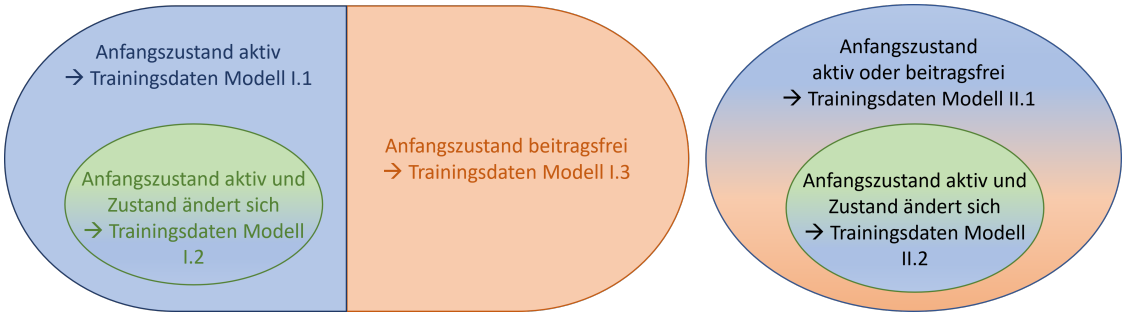


Abbildung 2.5.: Schematische Visualisierung der Trainingsdaten für die hierarchischen Modelle: links für Modell I, rechts für Modell II

in Abschnitt 2.3.2 in zwei disjunkte Datensätze $\mathcal{D}_A$ und $\mathcal{D}_B$ aufgeteilt. Anschließend wird für die Beobachtungen beider Anfangszustände (*aktiv* und *beitragsfrei*) jeweils ein binomiales Modell angepasst (in Abb. 2.4: Modell I.1 und I.3). Dieses schätzt für jede Beobachtung die Wahrscheinlichkeit einer Zustandsänderung, spezifiziert den Endzustand aber noch nicht genauer. Die hierarchische Struktur ergibt sich dann für den Anfangszustand *aktiv*. Hier wird im zweiten Schritt ein weiteres binomiales logistisches Regressionsmodell angepasst, mit welchem die Wahrscheinlichkeiten für die Endzustände *beitragsfrei* bzw. *storniert* im Falle einer Zustandsänderung bei Anfangszustand *aktiv* modelliert werden (in Abb. 2.4: Modell I.2).

Während Modell I.1 auf allen Daten mit Anfangszustand *aktiv* trainiert wird, bestehen die Trainingsdaten für Modell I.2 nur aus jenen Beobachtungen, für die tatsächlich ein Zustandswechsel beobachtet wird (siehe Abbildung 2.5).
Für den Anfangszustand *beitragsfrei* ist die Zweistufigkeit nicht notwendig, da hier im Falle eines Zustandswechsels nur der Wechsel zu *storniert* möglich ist. Dies ist der Tatsache geschuldet, dass es keine Beobachtung für eine Zustandsänderung von *beitragsfrei* nach *aktiv* im Datensatz gibt.

Die Modellstruktur hat für Merkmalsvektoren $x \in \mathbb{R}^p$ ohne $\mathcal{L}^1$-Regularisierung $3(p+1)$ Koeffizienten, was aus jeweils $p+1$ Koeffizienten für jedes einzelne logistische Regressionsmodell resultiert. Die tatsächliche Anzahl an Parametern (ungleich null) kann aber auch bei den hierarchischen Modellen durch Verwendung von $\mathcal{L}^1$-Regularisierung stark reduziert werden.

Im Gegensatz zum gerade eingeführten Modell I werden die Beobachtungen bei Modell II im ersten Schritt nicht nach ihren Anfangszuständen getrennt. Stattdessen wird den Merkmalsvektoren $x \in \mathbb{R}^p$ der Beobachtungen analog zum multinomialen Modell ein zu-

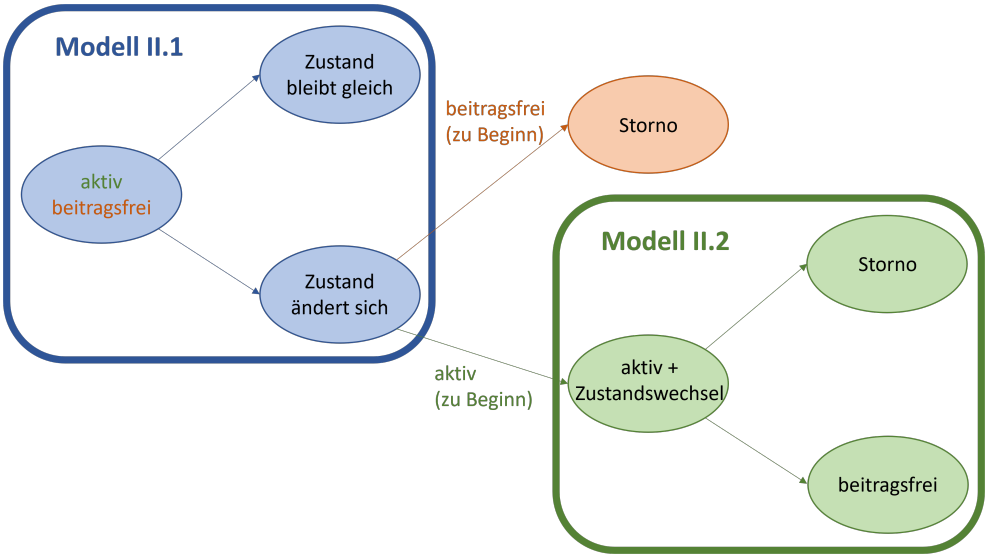


Abbildung 2.6.: Hierarchisches Modell ohne Aufteilen der Daten nach dem Anfangszustand (Modell II)

sätzliches Merkmal hinzugefügt, das den Anfangszustand $Y_0 \in \{A, B\}$ kodiert. Wie auch bei Modell I wird dann für jede Beobachtung zunächst die Wahrscheinlichkeit eines Zustandswechsels geschätzt, bevor anschließend die Wahrscheinlichkeiten für die möglichen Endzustände des Vertrags am Ende einer Beobachtungsperiode ermittelt werden (siehe Abbildung 2.6). Dazu wird Modell II.1 auf allen Beobachtungen im Datensatz trainiert. Danach erfolgt die Anpassung des Modells der zweiten Stufe ausschließlich auf jenen Daten, für die ausgehend vom Anfangszustand *aktiv* tatsächlich ein Zustandswechsel beobachtet wird (siehe Abbildung 2.5).

Insgesamt ergibt sich somit für Merkmalsvektoren $x \in \mathbb{R}^{p+1}$ nach Hinzufügen des Anfangszustands eine Modellstruktur mit $2(p+1)+1$ Parametern. Dabei weist das Modell II.1 $p+2$ Parameter auf, während Modell II.2 nur noch $p+1$ Koeffizienten besitzt. Letzteres ist der Tatsache geschuldet, dass bei Modell II.2 der Anfangszustand *aktiv* der Beobachtungen, auf die das Modell angepasst wird, bereits feststeht und folglich nicht durch ein zusätzliches Merkmal kodiert werden muss.

Neben der klassischen Modellanpassung (vgl. Kapitel 2.1.1) wird für beide hierarchische Modellstrukturen in Abschnitt 4.2.3 zusätzlich diskutiert, inwiefern sich die Berücksichtigung der Ergebnisse der Modelle in der ersten Stufe bei der Modellanpassung in der zweiten Stufe auf die Güte der Modelle auswirkt. Dazu werden die Modelle in der zweiten Stufe mit einer modifizierten Log-Likelihood-Funktion angepasst, welche sich für einen

Datensatz $\{(x_1, y_1), \ldots, (x_{\tilde{N}_A}, y_{\tilde{N}_A})\}$ durch

$$l(\beta_0, \beta) = -\sum_{i=1}^{\tilde{N}_A} p_i \Big(y_i(\beta_0 + x_i^T \beta) - \log\Big(1 + \exp(\beta_0 + x_i^T \beta)\Big)\Big)$$

ergibt. Mit $p_i \in (0, 1)$ wird dabei die Wahrscheinlichkeit eines Zustandswechsels für die Beobachtung i bezeichnet, die mit dem Modell in der ersten Stufe ermittelt wird. Der Wert $\tilde{N}_A$ beschreibt darüber hinaus die Anzahl der Beobachtungen mit Anfangszustand *aktiv*, für die tatsächlich ein Zustandswechsel beobachtet wird. Dies entspricht somit der Anzahl der Beobachtungen, auf denen das Modell in der zweiten Stufe trainiert wird.
Durch die modifizierte Log-Likelihood-Funktion erfolgt bei der Modellanpassung in der zweiten Stufe folglich eine Gewichtung. Dadurch spielen jene Beobachtungen, für die in der ersten Stufe hohe Wahrscheinlichkeiten eines Zustandswechsels prognostiziert werden, bei der Modellanpassung in der zweiten Stufe eine größere Rolle als Beobachtungen mit einer geringeren Wahrscheinlichkeit eines Zustandswechsels. Diese modifizierte Form der Modellanpassung hat das Potential, die Modellierung der Übergangswahrscheinlichkeiten zu verbessern. Jedoch ist dies nur der Fall, wenn die Prognosen des Modells in der ersten Stufe hohe Wahrscheinlichkeiten eines Zustandswechsels für Beobachtungen mit tatsächlich beobachtbaren Zustandswechseln liefern und umgekehrt niedrige Wahrscheinlichkeiten eines Zustandswechsels für Verträge ohne beobachtbaren Zustandswechsel.

Nachfolgend wird beleuchtet, wie sich die Übergangswahrscheinlichkeiten konkret aus den hierarchischen Modellstrukturen ableiten lassen. Mit $\bar{l}$ wird dazu im Folgenden das Ereignis bezeichnet, dass der Zustand l am Ende einer Beobachtungsperiode nicht eintritt.
Die Wahrscheinlichkeiten dafür, dass kein Wechsel des Zustands stattfindet, d.h. die Wahrscheinlichkeiten $\mathbb{P}(Y_1 = A|Y_0 = A)$ und $\mathbb{P}(Y_1 = B|Y_0 = B)$, werden bei Modellstruktur II mit dem Modell in der ersten Stufe II.1 ermittelt. Bei Modell I hingegen wird die Wahrscheinlichkeit für den Anfangszustand *aktiv* mit Modell I.1 ermittelt, während sich jene für den Anfangszustand *beitragsfrei* aus Modell I.3 ergibt. Die Übergangswahrscheinlichkeit $\nu = \mathbb{P}(Y_1 = S|Y_0 = B)$ wird durch die Modelle I.3 bzw. II.1 mit

$$\mathbb{P}(Y_1 = S|Y_0 = B) = \mathbb{P}(Y_1 = \bar{B}|Y_0 = B)$$

geschätzt. Für $Y_0 = A$ setzen sich die Übergangswahrscheinlichkeiten aus den Produkten der Schätzungen der hierarchisch angeordneten binomialen Regressionsmodelle wie folgt

zusammen:

$$\mathbb{P}(Y_1 = S|Y_0 = A) = \mathbb{P}(Y_1 = \bar{A}|Y_0 = A) \cdot \mathbb{P}(Y_1 = S|Y_0 = A, Y_1 = \bar{A})$$
$$\mathbb{P}(Y_1 = B|Y_0 = A) = \mathbb{P}(Y_1 = \bar{A}|Y_0 = A) \cdot \mathbb{P}(Y_1 = B|Y_0 = A, Y_1 = \bar{A})$$

Der jeweils erste Faktor des Produkts wird hierbei durch Modell I.1 bzw. II.1 geschätzt, während der jeweils zweite Faktor aus der Schätzung durch Modell I.2 bzw. II.2 resultiert.

Ein großer Vorteil dieser beiden Modellstrukturen ist, dass gemeinsame Einflussfaktoren der einzelnen Zustandswechseln bei der Schätzung der Übergangswahrscheinlichkeiten berücksichtigt werden können. Dazu wird sowohl bei Modell I als auch bei Modell II zunächst die Wahrscheinlichkeit einer Zustandsänderung ermittelt, bevor der Endzustand des Zustandswechsels konkretisiert wird. Insbesondere Modell II bietet diesen Vorteil, da hier im Gegensatz zu Modell I nicht nur Abhängigkeiten bei der Schätzung von Übergangswahrscheinlichkeiten für Beobachtungen mit Anfangszustand *aktiv*, sondern für alle Beobachtungen berücksichtigt werden können. Zudem müssen bei Modell II nur zwei binomiale logistische Regressionsmodelle angepasst werden, was die Parameteranzahl im Vergleich zu Modell I, aber auch im Vergleich zu den unabhängigen Modellen und dem multinomialen Modell enorm reduziert.
Die Trennung der Beobachtungen nach dem Anfangszustand in Modell I hat wiederum den Vorteil, dass die Wahrscheinlichkeiten der Beobachtungen insgesamt, vor allem aber für Beobachtungen mit Anfangszustand *beitragsfrei*, präziser modelliert werden können.

Ein Nachteil beider Modelle besteht darin, dass in der jeweils zweiten Stufe der hierarchischen Modellierung, d.h. in den Modellen I.2 bzw. II.2, bei der Prognose der Übergangswahrscheinlichkeiten auf Grundlage eines Datensatzes auch eine Schätzung der Wahrscheinlichkeit des Wechsels zu den Endzuständen *beitragsfrei* bzw. *storniert* für jene Beobachtungen erfolgt, für die tatsächlich kein Zustandswechsel beobachtbar ist. Dies kann in der Praxis gegebenenfalls zu Verzerrungen bei der Schätzung der Übergangswahrscheinlichkeiten führen, da die Modelle in der zweiten Stufe nicht auf Beobachtungen angepasst werden, für die kein Zustandswechsel beobachtet wird. Zudem können die Vorhersagen der hierarchischen Modellstrukturen stark negativ beeinträchtigt werden, wenn die Modelle der ersten Stufe, d.h. die Modelle I.1 und II.1, die Wahrscheinlichkeiten eines Zustandswechsel nicht ausreichend präzise abbilden.

2.4. Neuronale Netze

Ein großer Vorteil der logistischen Regressionsmodelle, die in den vorherigen Kapiteln eingeführt wurden, besteht in der Interpretierbarkeit der Modelle. Nichtsdestotrotz wirft dies die Frage auf, inwiefern die Interpretierbarkeit auf Kosten der Prognosegüte der Modelle geht. Aus diesem Grund wird in dieser Arbeit auch untersucht, ob sich *neuronale Netze* zur Modellierung von Übergangswahrscheinlichkeiten eignen und ggf. sogar bessere Ergebnisse erzielen können als logistische Regressionsmodelle. Die Modellklasse der neuronalen Netze hat in den vergangenen Jahren zunehmend an Bedeutung im Bereich des maschinellen Lernens gewonnen. Jedoch sind diese Modelle meist schwer bzw. nicht zu interpretieren, weshalb man häufig von sogenannten *Black-Box-Modellen* spricht.

Grundsätzlich handelt es sich bei (künstlichen) neuronalen Netzen um Modelle, die sich am biologischen Vorbild der neuronalen Netze im Nervensystem von Lebewesen orientieren. Sie können auf eine Vielzahl unterschiedlicher Aufgabentypen (bspw. Regression, Klassifikation) angewendet werden und weisen eine hohe Flexibilität in Bezug auf die gegebene Datengrundlage und konkrete Aufgabenstellung auf. Dadurch bieten sie sich als geeignete Modellklasse für die Modellierung der Übergangswahrscheinlichkeiten an. Die Zielsetzung für diese Modellklasse ist dabei ähnlich wie die der logistischen Regressionsmodelle: Es soll ein Modell gefunden werden, das auf Grundlage eines Input-Datensatzes die Übergangswahrscheinlichkeiten der Zustandswechsel möglichst präzise modellieren kann. In diesem Kapitel werden nun nachfolgend die Grundlagen neuronaler Netze und ihrer Modellanpassung eingeführt. Der Fokus hinsichtlich der Modellarchitektur liegt dabei auf den sogenannten *Multilayer-Perceptrons*.

2.4.1. Multilayer-Perceptron

Bei einem *Multilayer-Perceptron (MLP)* handelt es sich um ein mehrschichtiges Netz, das über eine Eingabe- und eine Ausgabeschicht sowie eine oder mehrere verdeckte Schichten verfügt. Im Folgenden kann stets davon ausgegangen werden, dass ein neuronales Netz mit genau einer verdeckten Schicht betrachtet wird, wobei sich die vorgestellten Grundlagen analog auf eine beliebige Anzahl verdeckter Schichten übertragen lassen. Bei *MLPs* ist jedes Neuron einer Schicht sowohl mit allen Neuronen der vorherigen, als auch mit allen Neuronen der nachfolgenden Schicht verbunden, sofern es eine vorherige bzw. nachfolgende Schicht gibt (vgl. Abb. 2.7). Dabei ist es weder möglich, dass Verbindungen rückwärts

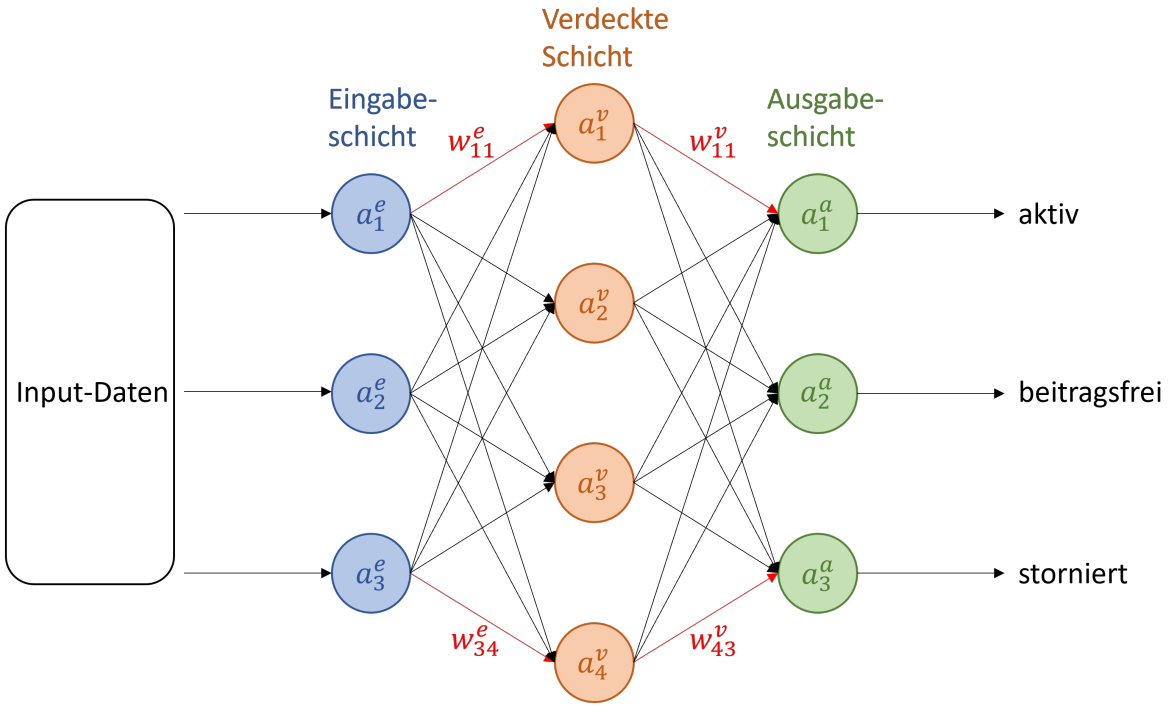


Abbildung 2.7.: Schematischer Aufbau eines *Multilayer-Perceptrons* zur Modellierung des Mehrzustandsmodells mit den Zuständen *aktiv*, *beitragsfrei* und *storniert*

gerichtet oder zyklisch sind, noch dass Verbindungen zwischen Neuronen eine oder mehrere Schichten überspringen.

In der konkreten Anwendung eines Klassifikationsproblems auf einem Datensatz $\mathcal{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ mit Merkmalsvektoren $x_1, \ldots, x_N \in \mathbb{R}^p$ und Realisierungen der abhängigen Variable $y_i \in \mathcal{K}$, wobei $\mathcal{K} = \{1, \ldots, K\}$ die Menge der möglichen Ausprägungen der abhängigen Variable Y beschreibt, folgt für das neuronale Netz, dass die Eingabeschicht p Neuronen und die Ausgabeschicht K Neuronen besitzt. Die Anzahl der Neuronen m in der verdeckten Schicht kann grundsätzlich beliebig gewählt werden. Der Wert a_i^v des i-ten Neurons in der verdeckten Schicht eines *MLPs* ergibt sich als

$$a_i^v = \Phi^v\left(\sum_{k=1}^{p} a_k^e w_{ki}^e + b_i^v\right), \; i = 1, \ldots, m$$

Dabei wird mit Φ^v die Aktivierungsfunktion in der verdeckten Schicht, mit b_i^v der Bias des i-ten Neurons der verdeckten Schicht und mit w_{ki}^e das Gewicht der Verbindung zwischen dem k-ten Neuron in der Eingabeschicht und dem i-ten Neuron in der verdeckten Schicht bezeichnet. Zudem entspricht der Wert des k-ten Neurons a_k^e in der Eingabeschicht dem

Wert des k-ten Eintrags eines Merkmalsvektors in den Input-Daten.
Analog ergibt sich der Wert des i-ten Neurons in der Ausgabeschicht als

$$a_i^a = \Phi^a\Big(\sum_{k=1}^{m} a_k^v w_{ki}^v + b_i^a\Big),\ i = 1, \ldots, K$$

mit der Aktivierungsfunktion der Ausgabeschicht Φ^a und Gewichten w_{ki}^v der Verbindung zwischen dem k-ten Neuron in der verdeckten Schicht und dem i-ten Neuron in der Ausgabeschicht. Mit b_i^a wird der Bias des i-ten Neurons der Ausgabeschicht bezeichnet.[13]

Zusammenfassend wird also ausgehend von der Eingabeschicht zunächst an jedem Neuron das Gewicht jeder eingehenden Kante mit dem Wert des Neurons, von dem die Kante ausgeht, multipliziert. Danach werden diese Produkte aufsummiert und der Bias aufaddiert. Der daraus resultierende Wert wird dann in die Aktivierungsfunktion eingesetzt, woraus sich der Wert des Neurons ergibt. Die Initialisierung der Gewichte und Biases vor der Modellanpassung erfolgt zufällig, wobei häufig eine Normalverteilung oder Gleichverteilung zugrunde gelegt wird. Dies kann in der Praxis bei der Implementierung der Modellstruktur genauer spezifiziert werden. Während des Trainings des neuronalen Netzes, d.h. während der Modellanpassung, werden die Gewichte und Biases dann so lange aktualisiert, bis die abhängige Variable der Beobachtungen im Datensatz ausreichend gut durch das neuronale Netz beschrieben wird.

2.4.2. Aktivierungsfunktionen

Wie bereits in Abschnitt 2.4.1 erläutert, werden die gewichteten Summen aus den Werten der Neuronen der vorherigen Schicht und den Gewichten der Verbindungen zwischen den Neuronen in die Aktivierungsfunktion eingesetzt, um den aktuellen Wert eines Neurons zu erhalten. Das Einsetzen in die Aktivierungsfunktion hat den Vorteil, dass die einzelnen Neuronen eines neuronalen Netzes, und somit auch das Netz insgesamt, nichtlineare Zusammenhänge abbilden können. Dazu werden andere Aktivierungsfunktionen als die *lineare Aktivierungsfunktion* $f(x) = x,\ x \in \mathbb{R}$, verwendet.

Im Folgenden werden mit den Funktionen *ReLU* und *Softmax* jene beiden Aktivierungsfunktionen vorgestellt, die für das neuronale Netz, das zur Modellierung der Übergangswahrscheinlichkeiten verwendet wird, relevant sind. Erstere wird dabei in der Zwischen-

[13] Vgl. Chollet und Allaire (2018) [4]

schicht des neuronalen Netzes verwendet, während die *Softmax*-Aktivierungsfunktion auf die Neuronen der Ausgabeschicht angewendet wird.

Die *ReLU*-Funktion (*ReLU* kurz für *Rectified Linear Unit*) ist durch

$$ReLU(x) = \max(0, x),\ x \in \mathbb{R}$$

definiert.[14] Es handelt sich um eine bei künstlichen neuronalen Netzen häufig verwendete Aktivierungsfunktion. Sie ist deshalb beliebt, weil sie nicht besonders rechenintensiv ist, nach der zufälligen Initialisierung der Gewichte rund die Hälfte aller Neuronen in einer Schicht auf null setzt und vorteilhafte Eigenschaften bei der Modellanpassung aufweist. Dennoch besteht bei der Verwendung der *ReLU*-Aktivierungsfunktion die Gefahr, dass die Modellanpassung enorm verlangsamt bzw. im schlimmsten Fall sogar gestoppt wird. Dem liegt zugrunde, dass für $x < 0$ sowohl die Funktion selbst, als auch ihre Ableitung den Wert $ReLU(x) = ReLU'(x) = 0$ annehmen. Dies kann dazu führen, dass manche bzw. gegebenenfalls sogar alle Gewichte der Verbindungen zwischen Neuronen nicht mehr aktualisiert werden. Zudem ist die *ReLU*-Aktivierungsfunktion an der Stelle $x = 0$ nicht differenzierbar, was für die Anpassung der Gewichte während des Trainings des Modells notwendig ist. Da der Funktionswert der Ableitung an dieser Stelle in der Praxis aber entweder als $ReLU'(0) = 1$ oder $ReLU'(0) = 0$ gewählt werden kann, stellt dies ein untergeordnetes Problem dar.

Zur Modellierung der Übergangswahrscheinlichkeiten in der Ausgabeschicht wird die sogenannte *Softmax*-Funktion verwendet, die durch

$$Softmax(x)_i = \frac{\exp(x_i)}{\sum_{j=1}^{K} \exp(x_j)},\ i = 1, \ldots, K$$

gegeben ist.[15] Sie transformiert einen K-dimensionalen Vektor $x = (x_1, \ldots, x_K)$, sodass $Softmax(x)_i \in (0, 1),\ i = 1, \ldots, K$, und

$$\sum_{i=1}^{K} Softmax(x)_i = \sum_{i=1}^{K} \frac{\exp(x_i)}{\sum_{j=1}^{K} \exp(x_j)} = 1$$

gilt.

[14] Vgl. Glorot et al. (2011) [9]
[15] Vgl. Chollet und Allaire (2018) [4], Dawani (2020) [5]

Die *Softmax*-Aktivierungsfunktion sorgt folglich dafür, dass das neuronale Netz eine gültige Wahrscheinlichkeitsverteilung ausgibt. Aus diesem Grund eignet sie sich als Aktivierungsfunktion der Ausgabeschicht bei Klassifikationsproblemen mit K möglichen Ausprägungen der abhängigen Variable. Es handelt sich dabei um die gleiche Funktion, die auch zur Modellierung der Wahrscheinlichkeitsverteilung bei der multinomialen logistischen Regression in der Gleichung (2.5) in Kapitel 2.1.2 verwendet wird. Somit kann ein neuronales Netz, das nur aus einer Eingabeschicht und einer Ausgabeschicht mit *Softmax*-Aktivierungsfunktion besteht, auch als multinomiales logistisches Regressionsmodell interpretiert werden.

2.4.3. Modellanpassung

Wie bei vielen Algorithmen des maschinellen Lernens ist es das grundlegende Ziel neuronaler Netze, eine Zielgröße möglichst genau auf Basis eines Input-Datensatz zu beschreiben. Bei der Modellanpassung soll, analog zur logistischen Regression, ein optimales Modell gefunden werden, sodass dieses Ziel erreicht wird. Dazu wird das sogenannte *Backpropagation*-Verfahren verwendet. Im Rahmen dieses Verfahrens werden sukzessive die Gewichte zwischen den Neuronen und die Biases angepasst, sodass sich der Unterschied zwischen der Netzwerkausgabe und der tatsächlichen Zielgröße minimiert und das resultierende neuronale Netz die gewünschte Zielvariable präzise beschreibt.[16]

Während der *Backpropagation* wird in jedem Lernschritt des neuronalen Netzes für die Eingabedaten zunächst die Vorhersage des aktuellen Modells berechnet. Anschließend wird der Fehler der Modellvorhersage bzgl. der tatsächlichen Zielgröße quantifiziert. Dabei erfolgt die Bestimmung des Fehlers mithilfe der sogenannten *Verlustfunktion*, die somit die Güte der Ausgabe des aktuellen Modells bemisst.

Im Anschluss an die Quantifizierung des Vorhersagefehlers erfolgt eine Auswertung, wie groß der Beitrag jedes einzelnen Gewichts bzw. Biases zum Gesamtfehler der Netzwerkausgabe ist. Die Gewichte und Biases werden dann schließlich so abgeändert, dass der Fehler, den das neuronale Netz bei der Prognose macht, reduziert wird.

Die Anpassung der Gewichte im *Backpropagation*-Verfahren erfolgt auf Grundlage des Gradientenverfahrens, welches der Bestimmung des Minimums der Verlustfunktion dient. Da dieses in der Praxis bei großen Datensätzen jedoch sehr rechenintensiv ist, wird bei

[16] Vgl. Géron (2017) [8]

der Implementierung ein sogenannter *Optimizer* verwendet. Dieser setzt das Gradientenverfahren in effizienter Form um, sodass das Lernen des neuronalen Netzes im Vergleich zur Modellanpassung auf Grundlage des klassischen Gradientenverfahrens signifikant beschleunigt wird.[17]

Bei Klassifikationsproblemen wird als Verlustfunktion meist die sogenannte *Kreuzentropie* (engl. Crossentropy) verwendet. Für einen Datensatz $\mathcal{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ mit Merkmalsvektoren $x_1, \ldots, x_N \in \mathbb{R}^p$ und K möglichen Ausprägungen der Realisierungen $y_i,\ i = 1, \ldots, N$, der abhängigen Variable erhält man die Kreuzentropie einer einzelnen Beobachtung i als

$$H(z_i, \hat{z}_i) = -\sum_{k=1}^{K} z_{ik} \log(\hat{z}_{ik})$$

Mit $\hat{z}_{ik} = \mathbb{P}(Y = k|x_i)$ wird dabei die vom Modell geschätzte Wahrscheinlichkeit beschrieben, dass eine Beobachtung mit Merkmalsvektor x_i die Ausprägung k besitzt.[18] Zudem gibt die Variable $z_{ik} \in \{0, 1\}$ an, ob die Beobachtung i tatsächlich die Ausprägung k hat, d.h. es gilt $z_{ik} = 1$, falls die Beobachtung i die Ausprägung k besitzt (d.h. $y_i = k$) und $z_{ik} = 0$ andernfalls. Die zu minimierende Verlustfunktion $H_{\mathcal{D}}$ für den kompletten Datensatz $\mathcal{D}$ ergibt sich dann durch die Summe der Kreuzentropien der einzelnen Beobachtungen, d.h. durch

$$H_{\mathcal{D}}(z, \hat{z}) = \sum_{i=1}^{N} H(z_i, \hat{z}_i) = -\sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} \log(\hat{z}_{ik}) \tag{2.7}$$

Bei der Gleichung (2.7) handelt es sich um die gleiche Funktion wie die negative Log-Likelihood-Funktion in Abschnitt 2.1.2, die zur Anpassung des multinomialen logistischen Regressionsmodells minimiert wird. Dies zeigt erneut den Bezug zur Modellklasse der logistischen Regressionsmodelle auf.

Darüber hinaus wird zur Vermeidung der Überanpassung des neuronalen Netzes an den Trainingsdatensatz bei der Modellanpassung häufig eine Methode namens *Dropout* verwendet. Dazu wird in jedem Trainingsschritt ein bestimmter Anteil an Verbindungen zwischen Neuronen auf null gesetzt, womit in jedem Trainingsschritt einige Neuronen nicht zum Lernen des Netzes beitragen. Ähnlich wie bei der $\mathcal{L}^1$-Regularisierung wird dadurch die Überangepasstheit des Modells an den Trainingsdatensatz reduziert, da durch Drop-

[17] Weitere Informationen zu *Backpropagation* und *Optimizern* finden sich in Chollet und Allaire (2018) [4], Géron (2017) [8]

[18] Vgl. Dawani (2020) [5]

out vermieden wird, dass der Einfluss einzelner Neuronen bzw. einzelner Merkmale in den Input-Daten auf die Netzwerkausgabe zu groß wird.[19]

2.5. Gütemaße

Die Güte aller Modellstrukturen wird neben der Anzahl der Parameter ungleich null und der Dauer der Modellanpassung auf den Trainingsdatensatz an der sogenannten *Devianz* bemessen. Diese ist eine Verallgemeinerung der Summe der quadrierten Residuen für diejenigen Fälle, in denen die Modellanpassung mittels Maximum-Likelihood-Methode erfolgt. Berechnet wird die Devianz durch

$$D = -2\log\left(\frac{\mathcal{L}_M}{\mathcal{L}_0}\right) = -2\Big(\log(\mathcal{L}_M) - \log(\mathcal{L}_0)\Big) \tag{2.8}$$

wobei es sich bei $\mathcal{L}_M$ um die maximierte Likelihood-Funktion des angepassten Modells und bei $\mathcal{L}_0$ um jene des saturierten Modells handelt.[20] Bei letzterem handelt es sich um das Modell, das perfekt an die Trainingsdaten angepasst ist, da für jede Beobachtung im Trainingsdatensatz eigene Parameter geschätzt werden.

Auf Grundlage eines Datensatzes mit N Beobachtungen vereinfacht sich die Gleichung (2.8) im binomialen Fall zu

$$D = -2\log(\mathcal{L}_M) = -2\Big(\sum_{i=1}^{N}\Big[y_i\log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\Big]\Big) \tag{2.9}$$

mit $y_i = 1$, falls für die Beobachtung i die Ausprägung 1 beobachtet wird und $y_i = 0$ andernfalls. Zudem bezeichnet $\hat{y}_i = \mathbb{P}(Y = 1|x_i)$ die durch das Modell geschätzte Wahrscheinlichkeit, dass die Beobachtung i mit Merkmalsvektor x_i die Ausprägung 1 besitzt. Für den multinomialen Fall mit K möglichen Ausprägungen der abhängigen Variable folgt

$$D = -2\log(\mathcal{L}_M) = -2\Big(\sum_{i=1}^{N}\sum_{k=1}^{K} y_{i,k}\log(\hat{y}_{i,k})\Big) \tag{2.10}$$

[19] Weitere Informationen zu *Dropout* finden sich in Chollet und Allaire (2018) [4], Géron (2017) [8]
[20] Vgl. [1], Agresti (2002) [3]

wobei $y_{i,k} = 1$ gilt, falls die Beobachtung i die Ausprägung $k \in \{1, \ldots, K\}$ besitzt und $y_{i,k} = 0$ andernfalls.[21] Mit $\hat{y}_{i,k} = \mathbb{P}(Y = k|x_i)$ wird die vom Modell prognostizierte Wahrscheinlichkeit bezeichnet, dass für die Beobachtung i mit Merkmalsvektor x_i die Ausprägung k beobachtet wird.

Den Vereinfachungen in den Gleichungen (2.9) bzw. (2.10) liegt zugrunde, dass die Log-Likelihood-Funktion $\log(\mathcal{L}_0)$ des saturierten Modells gleich null ist, da dieses Modell die Beobachtungen im Trainingsdatensatz perfekt abbilden kann. Im Allgemeinen gilt für die Güte der Modelle: Je geringer die Devianz, desto besser die Anpassung des Modells an den Trainingsdatensatz. Der bestmögliche Wert, den ein Modell hinsichtlich der Devianz erzielen kann, ist $D = 0$.

Bei der Devianz in Gleichung (2.10) handelt es sich um die verdoppelte Kreuzentropie (vgl. Abschnitt 2.4.3), die als Verlustfunktion zur Modellanpassung des neuronalen Netzes dient. Folglich kann die Devianz in Gleichung (2.10), analog wie bei den logistischen Regressionsmodellen, auch als Gütemaß für das neuronale Netz verwendet werden.

[21] Vgl. [2]

3. Datenanalyse

Dieses Kapitel dient der Beschreibung und Analyse jener Daten, auf welche die in Kapitel 2 vorgestellten Modelle angepasst werden. Um die Modellstrukturen in Kapitel 4 zuverlässig evaluieren und vergleichen zu können, ist eine gute Datengrundlage unerlässlich. Diese ist dank einiger Vorarbeit von Reck et al. (2022) [15] bereits vorhanden, wodurch kaum Änderungen daran vorgenommen werden müssen. Im Folgenden wird der Datensatz mit seinen Variablen beschrieben und analysiert. Zudem wird auf vorgenommene Änderungen bzw. neu hinzugefügte Variablen eingegangen.

3.1. Übersicht

Der Datensatz stammt von einem europäischen Lebensversicherer, der in vier Ländern tätig ist. Jede Zeile darin enthält Informationen über den Anfangs- bzw. Endzustand sowie weitere Kovariablen eines Vertrages für einen je einjährigen Beobachtungszeitraum. Somit kann der gleiche Vertrag mehrmals im Datensatz enthalten sein, sofern er über mehrere Jahre beobachtet und erfasst wird. Die Beobachtungen im Datensatz sind folglich nicht komplett unabhängig voneinander, was eine Voraussetzung der logistischen Regression teilweise verletzt. Dennoch wird dies im Folgenden in Anlehnung an Reck et al. (2022) [15] aufgrund der Größe des Datensatzes vernachlässigt. Finden mehrere Zustandsübergänge innerhalb eines Beobachtungsjahres statt, so enthält der Datensatz jeweils nur den Anfangszustand sowie den Endzustand des Vertrages im betrachteten Jahr.

Insgesamt sind im Datensatz $N = 992.215$ Beobachtungen mit $J = 15$ Kovariablen enthalten, wobei keine Beobachtung fehlende Einträge aufweist. Der Beobachtungszeitraum erstreckt sich vom Jahr 2000 bis in das Jahr 2020, womit ein Vertrag höchstens 20-mal im Datensatz vorkommen kann. Bei den erfassten Verträgen handelt es sich um einen sogenannten Run-Off-Bestand, was im vorliegenden Fall bedeutet, dass nach zwölf Beobachtungsjahren, d.h. ab dem Jahr 2012, keine neuen Verträge mehr hinzugefügt wurden.

	Name	Anzahl Kategorien	Beschreibung
1	gender	2	Geschlecht des Versicherungsnehmers
2	country	4	Land des Vertragsabschlusses
3	term_of_insurance	10	Laufzeit des Vertrags
4	premium_duration	9	Dauer der Beitragszahlungen
5	payment_frequency	5	Häufigkeit der Beitragszahlungen pro Jahr (z.B. monatlich/jährlich)
6	yearly_premium	31	Höhe der jährlichen Beitragszahlungen
7	sum_insured	16	Versicherungssumme
8	payment_method	3	Art der Beitragszahlung (z.B. Lastschrift)
9	dynamic	11	dynamische Beitragsanhebung in Prozent
10	age_entry	15	Alter bei Abschluss des Vertrages
11	contract_duration	20	Anzahl der bisherigen Beobachtungsjahre
12	status_b	2	Status zu Beginn eines Beobachtungsjahres
13	status_e	3	Status am Ende eines Beobachtungsjahres
14	insurance_type	2	Art der Versicherung (z.B. traditionell/fondgebunden)
15	native	2	gibt an, ob das Land des Vertragsabschlusses der Nationalität des Versicherungsnehmers entspricht

Tabelle 3.1.: Übersicht über die Variablen mit Beschreibung und Anzahl an Kategorien

Für die restliche Beobachtungszeit werden dann nur noch die bereits bestehenden Verträge beobachtet. Insgesamt sind im Datensatz 167.659 Verträge erfasst, wovon 164.693 zu Beginn *aktiv* und 2.966 *beitragsfrei* sind.

Neben kategorialen Variablen enthält der ursprüngliche Datensatz des Lebensversicherers auch stetige Variablen. Im Datensatz, der für die Modellierungen in dieser Arbeit verwendet wird, sind diese Variablen aber ebenfalls als kategoriale Variablen enthalten. Dadurch können für ursprünglich stetige Variablen komplexere Zusammenhänge abgebildet werden, da mehrere Parameter für dieselbe kategoriale Kovariable geschätzt werden. Bei

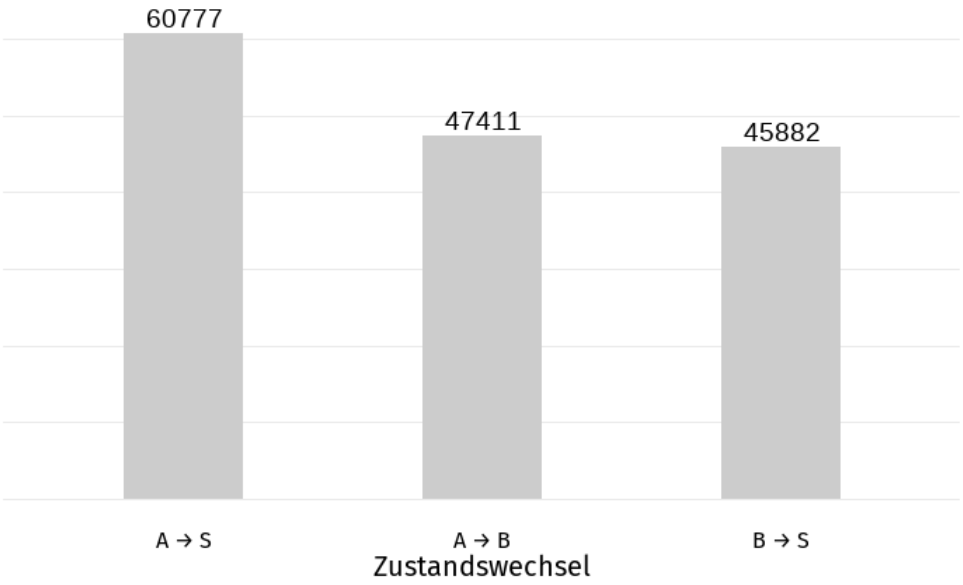


Abbildung 3.1.: Anzahl der beobachteten Zustandswechsel im Datensatz

Verwendung als stetige Variable hingegen würde nur ein Parameter je Kovariable geschätzt werden.[22]

Die Umwandlung der stetigen in kategoriale Variablen erfolgt durch Kategorisierung in Intervalle. Allen Beobachtungen, die in einem Intervall liegen, wird dann die gleiche Merkmalsausprägung zugewiesen. Somit werden die Modelle in dieser Arbeit auf einen Datensatz mit 15 kategorialen Variablen angepasst, über welche die Tabelle 3.1 genauer Aufschluss gibt. Neben der Anzahl der Kategorien je Kovariable sind der Tabelle auch genauere Erläuterungen zu den Variablen zu entnehmen.

Wie aus der Tabelle 3.1 hervorgeht, haben alle im Datensatz enthaltenen Variablen mindestens zwei Merkmalsausprägungen. Dies ist notwendig, da im Falle nur einer Merkmalsausprägung die entsprechende Variable keinen Informationsgehalt besitzen und somit keinen Einfluss auf die Schätzung der Übergangswahrscheinlichkeiten nehmen würde. Von besonderer Wichtigkeit sind die beiden Variablen `status_b` und `status_e`, die den Zustand eines Vertrages zu Beginn bzw. am Ende eines Beobachtungsjahres angeben. Stimmen diese Variablen für eine Beobachtung im Datensatz nicht überein, so wird ein Zustandswechsel beobachtet.

Insgesamt sind im Datensatz 154.070 solcher Zustandswechsel beobachtbar, was ca. 15,53% aller darin enthaltenen Beobachtungen entspricht. Die meisten Zustandsübergänge finden

[22] Vgl. Reck et al. (2022) [15]

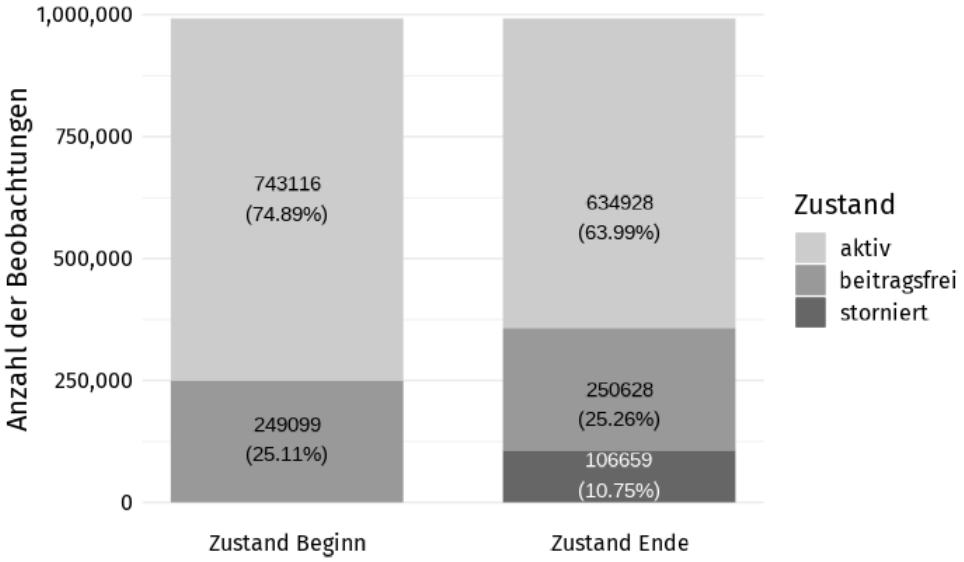


Abbildung 3.2.: Zustände der Verträge im Datensatz zu Beginn und am Ende eines Beobachtungsjahres

dabei von *aktiv* nach *storniert* statt. Die Anzahl der Zustandswechsel von *aktiv* nach *beitragsfrei* bzw. von *beitragsfrei* nach *storniert* bewegen sich ungefähr in der gleichen Größenordnung und liegen jeweils ca. 16.000 Beobachtungen unter der Anzahl an beobachteten Zustandsübergängen von *aktiv* nach *storniert*, wie die Abbildung 3.1 zeigt. Zudem sind sowohl zu Beginn als auch am Ende eines Beobachtungsjahres mit Abstand die meisten Verträge *aktiv*, wie sich der Abbildung 3.2 entnehmen lässt. Darüber hinaus geht aus dem Datensatz hervor, dass es, gemessen am Anteil der Beobachtungen, die zu Beginn *beitragsfrei* sind (25,11%, vgl. Abb. 3.2), mit 29,78% aller Zustandswechsel verhältnismäßig mehr Zustandswechsel bei Anfangszustand *beitragsfrei* gibt als bei Anfangszustand *aktiv*. Somit gibt es ein erhöhtes Stornorisiko, wenn ein Vertrag bereits beitragsfrei gestellt ist. Dies ist durchaus intuitiv, da Beitragsfreistellung als Anzeichen für Schwierigkeiten hinsichtlich des Vertrages eines Versicherten gewertet werden kann, da (vorerst) keine Beitragszahlungen mehr erfolgen.

Ein besonderes Augenmerk in Kapitel 4 liegt auf den beiden Variablen Eintrittsalter (`age_entry`) und der Anzahl der bisherigen Beobachtungsjahre (`contract_duration`), da mithilfe dieser Variablen die Güte der Prognosen der Modelle grafisch evaluiert wird. Die Abbildung 3.3 gibt genaueren Aufschluss über diese beiden Variablen. Konkret lässt sich der Abbildung die Anzahl an Beobachtungen sowie die Anzahl an beobachteten Zustandswechseln je Kategorie entnehmen.

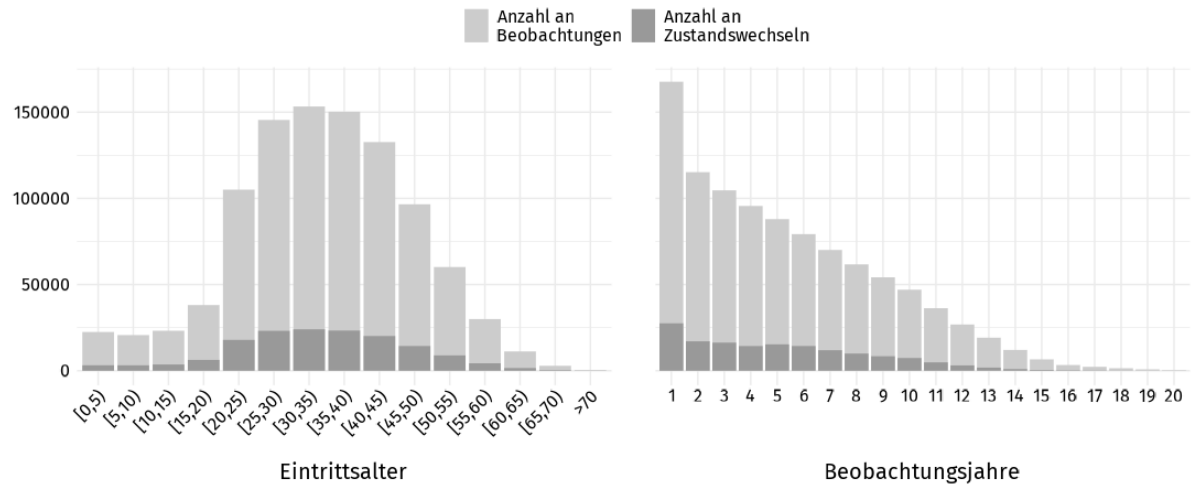


Abbildung 3.3.: Anzahl der Beobachtungen und beobachteten Zustandswechsel je Kategorie für Eintrittsalter (links) und Beobachtungsjahre (rechts)

Auffällig ist hierbei, dass überproportional viele Verträge nur ein Jahr beobachtet werden. Für die Folgejahre nimmt die Anzahl an beobachteten Verträgen recht gleichmäßig ab, wie dem rechten Teil von Abbildung 3.3 zu entnehmen ist. Der Effekt im ersten Beobachtungsjahr ist ebenfalls für die Anzahl der Zustandswechsel erkennbar, wohingegen diese Anzahl im Bereich zwischen zwei und sechs Beobachtungsjahren recht konstant ist und erst anschließend für jedes zusätzliche Beobachtungsjahr weiter abnimmt.

Außerdem fällt auf, dass nur sehr wenige Verträge über einen sehr langen Zeitraum im Datensatz enthalten sind. Insbesondere sind nur rund 15.000 der insgesamt fast 1.000.000 im Datensatz enthaltenen Beobachtungen von Verträgen, die bereits 15 Jahre oder länger beobachtet werden. Dem liegt zugrunde, dass lediglich 6.566 der fast 170.000 erfassten Verträge über einen Zeitraum von mindestens 15 Jahren beobachtet werden. Dementsprechend muss dieser Befund in der grafischen Analyse in Kapitel 4 dahingehend berücksichtigt werden, dass die Ergebnisse für eine hohe Anzahl an bisherigen Beobachtungsjahren möglicherweise ungenau sind. Dies ist auf die wenigen im Datensatz vorhandenen Beobachtungen mit einer hohen Anzahl an Beobachtungsjahren zurückzuführen, auf denen die Modelle trainiert werden können.
Ähnlich verhält es sich mit der Variable Eintrittsalter, da die meisten Beobachtungen ein Eintrittsalter zwischen 15 und 60 Jahren aufweisen. Insofern müssen auch für diese Kovariable die grafischen Analysen für Verträge, die von jüngeren bzw. älteren Menschen abgeschlossen wurden, mit Vorsicht genossen werden.

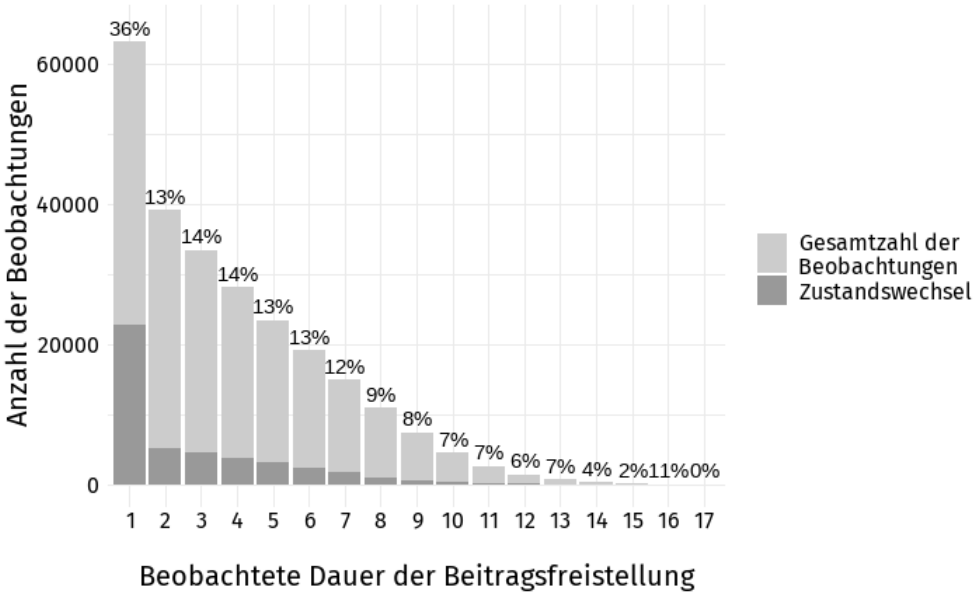


Abbildung 3.4.: Anzahl der Beobachtungen und Zustandswechsel je Beitragsfreistellungsdauer mit dem Anteil der nach dem Beobachtungsjahr stornierten Verträge in Prozent

3.2. Bearbeitung der Daten

Der Datensatz enthält bereits eine Vielzahl an Informationen über die beobachteten Verträge und weist keine fehlenden Einträge auf, sodass kaum Bearbeitung des Datensatzes notwendig ist. So ist auch die Beobachtungsdauer der Verträge in Form der Variable `contract_duration` vorhanden, wobei diese nicht erfasst, wie lange ein Vertrag ggf. bereits als *beitragsfrei* beobachtet wird. Da dies für die Schätzung der Übergangswahrscheinlichkeiten aber durchaus relevant sein kann, wird dem Datensatz eine weitere Variable `beitragsfrei_duration` hinzugefügt, welche die beobachtete Dauer der Beitragsfreistellung angibt. Ist ein Vertrag noch *aktiv*, so nimmt diese Variable den Wert null an.

Die längste beobachtete Dauer einer Beitragsfreistellung im Datensatz beträgt 17 Jahre. Die Abbildung 3.4 gibt Aufschluss über die Anzahl der Beobachtungen je Beitragsfreistellungsdauer sowie die zugehörige Anzahl an Zustandsübergängen, d.h. die Anzahl der Verträge, die während der Beobachtungsperiode storniert werden. Zudem sind der Abbildung die Anteile der Verträge je Beitragsfreistellungsdauer, die nach dem Beobachtungsjahr storniert sind, zu entnehmen. Dies wird in der Abbildung durch die Prozentzahlen angegeben. Auffällig ist, dass mit Abstand die meisten Verträge, nämlich rund 36%, bereits nach einem beobachteten Jahr mit Beitragsfreistellung storniert werden. In den darauffolgen-

den Zeitdauern zwischen zwei und sieben Jahren liegt der Anteil an stornierten Verträgen bei vorangegangener Beitragsfreistellung hingegen recht konstant zwischen 12% und 14%. Folglich ist das Stornorisiko eines Vertrages besonders hoch, wenn die Dauer der bisher beobachteten Beitragsfreistellung bei einem Jahr liegt. Das bedeutet, dass Verträge, die kürzlich beitragsfrei gestellt wurden, einem wesentlich höheren Risiko unterliegen, storniert zu werden als Verträge, die bereits seit längerer Zeit beitragsfrei gestellt sind. Der Einfluss der zusätzlichen Variable `beitragsfrei_duration` auf die Güte der Modellierung der Übergangswahrscheinlichkeiten wird im nachfolgenden Kapitel 4 beleuchtet.

4. Ergebnisse

Dieses Kapitel dient der Vorstellung und Diskussion der Ergebnisse, die durch Anpassung der in Kapitel 2 eingeführten Modelle auf den Datensatz erzielt werden. Im ersten Teil des Kapitels werden Aspekte der konkreten Modellanpassung thematisiert, wie beispielsweise die Daten, die zum Training bzw. Testen der Modelle herangezogen werden. Zudem wird in diesem Teil auf die Art der verwendeten $\mathcal{L}^1$-Regularisierung je Kovariable eingegangen. Anschließend werden die logistischen Regressionsmodelle anhand verschiedener Kennzahlen evaluiert und auf Grundlage dieser Kennzahlen verglichen. Darüber hinaus dienen die unabhängigen logistischen Regressionsmodelle sowie das Modell, das nur aus dem Intercept besteht, als Referenz zur Bewertung der Modelle, die gemeinsame Treiber der Zustandswechsel berücksichtigen. Abschließend werden die Ergebnisse, die mit neuronalen Netzen erzielt werden können, diskutiert und im Kontext der Ergebnisse der logistischen Regressionsmodelle betrachtet.

4.1. Datensatz und Modellanpassung

Dieser Abschnitt bildet die Grundlage für die Auswertung der Ergebnisse in Abschnitt 4.2. Im Folgenden wird dazu zuerst die Auswahl der Trainings- bzw. Testdaten genauer erläutert, bevor auf die konkrete Umsetzung der $\mathcal{L}^1$-Regularisierung bei der Modellanpassung eingegangen wird.

Zur Anpassung der Modelle wird der Datensatz zunächst in einen zufälligen Trainings- und Testdatensatz aufgespaltet unter Beachtung dessen, dass die Endzustände im gleichen Verhältnis in beiden Datensätzen enthalten sein sollen. Die Modelle werden dann auf 75% der im Datensatz enthaltenen Beobachtungen trainiert und auf den restlichen 25% getestet. Der Testdatensatz dient dabei zur Überprüfung, inwiefern *Overfitting*, also eine Überangepasstheit der Modelle auf den Trainingsdatensatz, vorliegt. Indem nicht die

ersten 75% der Datenpunkte sondern zufällig ausgewählte Beobachtungen im Trainingsdatensatz enthalten sind, soll zudem die Abhängigkeit der Datenpunkte reduziert werden, die möglicherweise entsteht, da dieselben Verträge mehrmals im Datensatz enthalten sein können.

Darüber hinaus spielt die Art der $\mathcal{L}^1$-Regularisierung, die für jede Kovariable spezifiziert werden muss, bei der Modellanpassung eine große Rolle. Die Tabelle 4.1 gibt eine Übersicht darüber, wie die Spezifikationen für die Kovariablen gewählt werden. Die Zuordnung der Art der Lasso-Terme zu den einzelnen Kovariablen bleibt dabei bei allen Modellen gleich, um eine maximale Vergleichbarkeit der Modellstrukturen zu schaffen.
Sollen, wie beispielsweise bei der Kovariable Geschlecht, benachbarte Kategorien nicht miteinander verschmolzen werden bzw. ist die Unterstellung eines Trends zwischen den Kategorien einer Kovariable nicht sinnvoll, so werden die Koeffizienten der Kovariable mit dem *regulären Lasso* regularisiert.
Das *reguläre Lasso* wird also auf nominalskalierte Kovariablen angewendet, während *Fused Lasso* bzw. *Trend Filtering* bei ordinalskalierten Merkmalen verwendet werden. Die Zuordnung von *Fused Lasso* bzw. *Trend Filtering* erfolgt nach dem Prinzip von Reck et al. (2022) [15]. Danach werden mit dem *Fused Lasso* die Koeffizienten jener Kovariablen bestraft, bei denen adjazente Kategorien miteinander verschmolzen werden sollen. *Trend Filtering* hingegen wird auf Kovariablen angewendet, bei denen ein Trend zwischen den einzelnen Kategorien vermutet wird.

	Name	**Art**
1	`gender`	regular
2	`country`	regular
3	`term_of_insurance`	trend
4	`premium_duration`	trend
5	`payment_frequency`	fused
6	`yearly_premium`	trend
7	`sum_insured`	trend
8	`payment_method`	regular
9	`dynamic`	trend
10	`age_entry`	fused
11	`contract_duration`	trend
12	`status_b`	regular
13	`insurance_type`	regular
14	`native`	regular
15	`beitragsfrei_duration`	trend

Tabelle 4.1.: Art der verwendeten $\mathcal{L}^1$-Regularisierung je Kovariable im Datensatz

Bei der Verwendung von Regularisierung zur Modellanpassung wird der in Abschnitt 2.2 eingeführte Regularisierungsparameter λ mittels Kreuzvalidierung ermittelt und nach der sogenannten *1-SE-Regel* gewählt. Dadurch erhält man für λ den Wert des Parameters des Modells, das die wenigsten Koeffizienten ungleich null enthält, d.h. die geringste Komplexität aufweist und dessen Fehler gleichzeitig maximal eine Standardabweichung vom Fehler des besten Modells, also dem Modell mit der besten Prognosegüte, entfernt ist.[23] Daraus resultiert ein Modell mit geringer Komplexität, das dennoch Ergebnisse liefert, die ähnlich gut wie die Prognosen des besten Modells mit optimalem Regularisierungsparameter λ sind.

4.2. Auswertung

In diesem Abschnitt werden die Ergebnisse, die durch die verschiedenen Modellanpassungen erzielt werden können, analysiert. Dazu werden die Werte der Gütemaße ausgewertet und zwischen den einzelnen Modellen verglichen. Ist dabei nachfolgend von der Anzahl der Parameter eines Modells die Rede, so ist die Anzahl der Parameter ungleich null gemeint. Im weiteren Verlauf des Kapitels wird insbesondere darauf eingegangen, inwiefern sich die unterschiedlichen Konfigurationen der Modelle, wie beispielsweise mit bzw. ohne Regularisierung, auf die Gütemaße auswirken. Ein besonderes Augenmerk wird auf den Ergebnissen des multinomialen Modells sowie den Ergebnissen der hierarchischen Modelle liegen, wobei die Güte dieser Modellstrukturen auch grafisch veranschaulicht wird. Zudem werden die Ergebnisse der beiden Modellstrukturen mit den unabhängigen Modellen sowie dem Modell verglichen, das nur den Intercept als Parameter enthält. Dies dient der Evaluation, inwiefern die Berücksichtigung möglicher gemeinsamer Treiber der einzelnen Zustandswechsel bei der Modellanpassung eine Verbesserung der Vorhersagen bringen kann.

4.2.1. Überblick

Einen ersten Überblick über die Ergebnisse liefert die Tabelle 4.2, welche die Werte der Gütemaße, die durch die einzelnen Modellanpassungen erzielt werden, enthält. Die zur Anpassung und zum Testen der Modelle verwendeten Daten stimmen dabei für alle Modellstrukturen überein, um eine maximale Vergleichbarkeit der Ergebnisse gewährleisten zu

[23] Vgl. Hastie et al. (2009) [11], S.244

können. Sowohl für die unabhängigen Modelle, als auch für das multinomiale Modell sowie die hierarchischen Modelle sind der Tabelle die Werte der Gütemaße für die Modellanpassung jeweils mit und ohne Regularisierung zu entnehmen. Für das multinomiale Modell bzw. die hierarchischen Modelle sind zudem die Ergebnisse bei Hinzunahme der Variable `beitragsfrei_duration` (Anzahl der Beobachtungsjahre im Zustand *beitragsfrei*) enthalten, wobei hier zur Modellanpassung zusätzlich Regularisierung verwendet wird. Gleiches gilt für die Gewichtung bei der Modellanpassung in der zweiten Stufe der hierarchischen Modelle. Eine detaillierte Analyse der Ergebnisse des multinomialen Modells sowie der hierarchischen Modelle erfolgt in den nachfolgenden Abschnitten 4.2.2 und 4.2.3. Darauf folgt in Teil 4.2.4 der Vergleich der Ergebnisse aller Modelle mit Erörterung möglicher Stärken und Schwächen der Modellstrukturen.

	Devianz[25]			
Modell	**Training**	**Test**	**Parameter**	**Trainingszeit**[26]
Intercept	649.413	216.775	3	
Unabhängige Modelle	302.678	101.420	352	0,40
mit Regularisierung	303.937	101.759	132	50,35
Multinomiales Modell	315.565	105.811	357	0,96
mit Regularisierung	316.254	105.994	122	90,85
mit `beitragsfrei_duration`[24]	314.271	105.410	151	102,29
Hierarchisches Modell I	309.756	103.769	352	0,23
mit Regularisierung	310.986	104.115	124	18,92
mit `beitragsfrei_duration`[24]	308.399	103.323	121	19,12
mit Gewichtung und `beitragsfrei_duration`[24]	308.468	103.333	111	19,00
Hierarchisches Modell II	321.630	107.675	237	0,18
mit Regularisierung	322.494	107.929	99	16,15
mit `beitragsfrei_duration`[24]	316.774	106.096	103	19,51
mit Gewichtung und `beitragsfrei_duration`[24]	316.819	106.100	94	19,54

Tabelle 4.2.: Gütemaße der Modelle bei Anpassung auf den Datensatz

[24] jeweils mit zusätzlicher Regularisierung
[25] je niedriger, desto besser
[26] in Minuten

4.2.2. Multinomiales Modell

Um die Güte des multinomialen Modells zu evaluieren, werden zunächst die Ergebnisse in Tabelle 4.2 beleuchtet. Aus der Tabelle geht hervor, dass sich die Devianz für die Anpassung mit Regularisierung im Vergleich zum Modell ohne Regularisierung sowohl für die Trainingsdaten, als auch für die Testdaten erhöht. Verwendet man für das regularisierte Modell zusätzlich die Variable `beitragsfrei_duration`, so sinkt die Devianz. Insbesondere auf den Testdaten sind die Unterschiede zwischen den Modellanpassungen des multinomialen Modells jedoch marginal.

Darüber hinaus erhöht sich die Trainingszeit für beide Modelle mit Regularisierung stark. Dies ist darauf zurückzuführen, dass zur Bestimmung des optimalen Regularisierungsparameters λ mittels Kreuzvalidierung mehrere Modellanpassungen vorgenommen werden. Im Kontrast dazu muss im nicht regularisierten Fall nur ein Modell angepasst werden.

Die deutlichste Auswirkung der Regularisierung wird in der Anzahl der Parameter ersichtlich. Diese beträgt bei den regularisierten Modellen jeweils weniger als die Hälfte der Parameteranzahl des nicht regularisierten Modells. Damit wird deutlich, dass die Komplexität der Modelle durch Regularisierung tatsächlich enorm reduziert wird, was die Interpretierbarkeit der Modelle erhöht. Ohne Berücksichtigung der Anzahl der Beobachtungsjahre im Zustand *beitragsfrei* (d.h. der Variable `beitragsfrei_duration`) fällt die Reduktion der Parameterzahl sogar noch stärker aus, sodass die Anzahl der Parameter ungleich null nur noch ein Drittel der Parameterzahl des nicht regularisierten Modells beträgt.

Der Effekt der Regularisierung spiegelt sich zudem in den Devianzen wider. Während sich die Devianz des regularisierten Modells im Vergleich zum nicht regularisierten Modell auf den Trainingsdaten erhöht, fällt der Anstieg auf den Testdaten prozentual geringer aus. Dies zeigt, dass sich durch Lasso, wie bereits in Abschnitt 2.2.1 erläutert, die Fähigkeit des Modells zu Verallgemeinern erhöht, da eine Überanpassung an den Trainingsdatensatz vermieden und der Einfluss einzelner erklärender Variablen auf die abhängige Variable reduziert wird.

Abschließend geht aus der Tabelle hervor, dass die Hinzunahme der Variable `beitragsfrei_duration` die Prognosegüte des multinomialen logistischen Regressionsmodells verbessert. Die Devianz liegt hier bei der Modellanpassung mit Berücksichtigung der Beobachtungsjahre im Zustand *beitragsfrei* sowohl auf den Trainingsdaten, als auch auf den Testdaten unter der Devianz des regularisierten Modells ohne die Variable

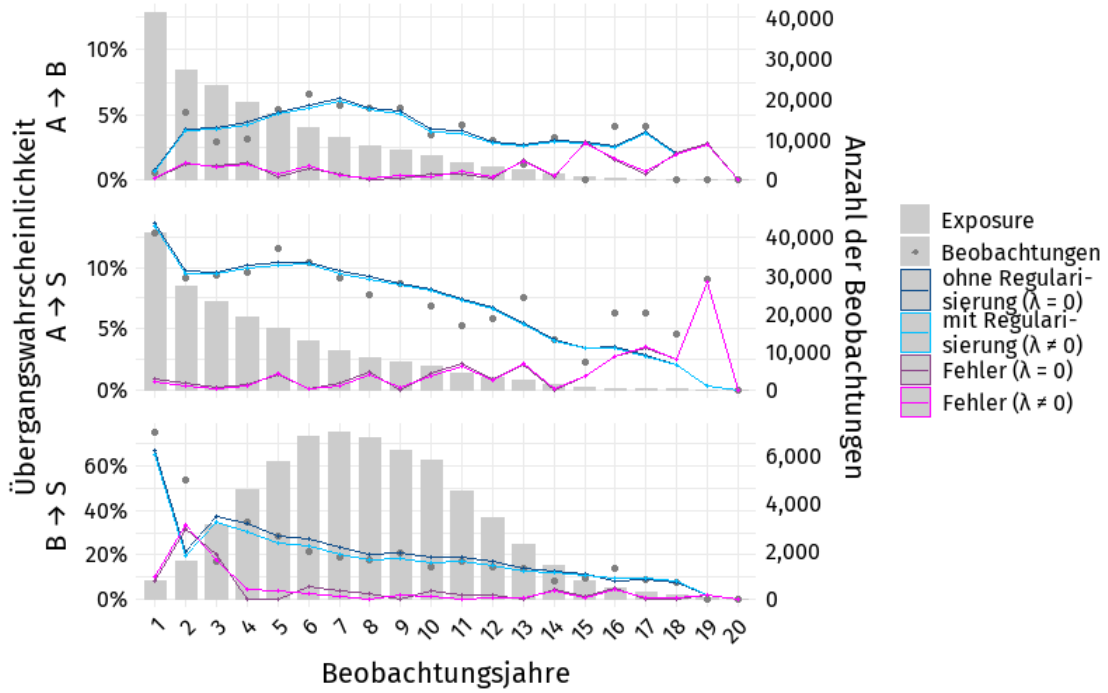


Abbildung 4.1.: Prognosen und Vorhersagefehler des multinomialen Modells für die Anzahl der Beobachtungsjahre mit und ohne Regularisierung auf den Testdaten

`beitragsfrei_duration` und sogar unter jener des nicht regularisierten Modells. Allerdings weist diese Modellanpassung die höchste Trainingsdauer auf.

Die Güte des multinomialen Modells kann grafisch anhand der Abbildung 4.1 beurteilt werden. Diese zeigt die durchschnittlichen geschätzten Übergangswahrscheinlichkeiten sowie die absoluten Prognosefehler auf den Testdaten für die einzelnen Merkmalsausprägungen bezüglich der Anzahl der Beobachtungsjahre mit und ohne Regularisierung. Analoge Abbildungen für die Trainingsdaten sowie in Bezug auf das Eintrittsalter finden sich im Anhang A.1.

Grundsätzlich ist den Abbildungen zu entnehmen, dass die Beobachtungen durch die multinomiale Modellstruktur präzise modelliert werden können, da die Visualisierungen der geschätzten Übergangswahrscheinlichkeiten nahe an den tatsächlich beobachteten Übergangswahrscheinlichkeiten liegen. Auffallend ist jedoch, dass die Schätzungen des regularisierten Modells wesentlich glatter sind als jene des Modells ohne Regularisierung. Insbesondere bei den geschätzten Übergangswahrscheinlichkeiten des Zustandswechsels von *aktiv* nach *storniert* fällt auf, dass das nicht regularisierte Modell zu sehr von den Trainingsdaten abhängt und an diese überangepasst ist (vgl. Abb. A.1). Deshalb bildet dieses

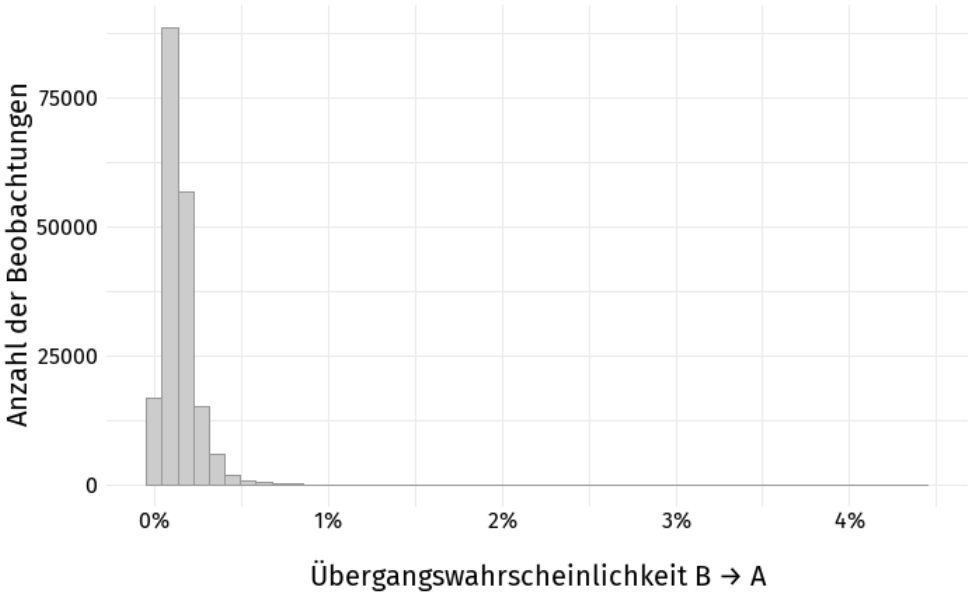


Abbildung 4.2.: Geschätzte Übergangswahrscheinlichkeiten $\mathbb{P}(Y_1 = A|Y_0 = B)$ des multinomialen Modells

Modell ein Rauschen in den Trainingsdaten ab, was sich negativ auf die Prognosefähigkeit bei vorher nicht gesehenen Daten auswirkt. Dies kann grafisch im Vergleich der Abbildung für die Trainingsdaten A.1 und der Abbildung 4.1 für die Testdaten beobachtet werden. Für die Testdaten ist insbesondere beim Zustandswechsel von *beitragsfrei* nach *storniert* beobachtbar, dass der Prognosefehler des Modells ohne Regularisierung höher ist als der des Modells mit Regularisierung. Hinsichtlich der Prognosegüte auf unbekannten Daten kann es folglich von Vorteil sein, die Komplexität des Modells mithilfe von Regularisierung zu reduzieren, um die Vorhersagegenauigkeit zu erhöhen.

Des Weiteren auffallend ist, dass für eine hohe Anzahl an Beobachtungsjahren (mehr als 15 Jahre) sowie für Eintrittsalter jünger als 15 bzw. älter als 60 Jahre die Prognosegüte des Modells stark nachlässt. Dies lässt sich vermutlich darauf zurückführen, dass nur wenige Beobachtungen mit diesen Merkmalsausprägungen im Datensatz vorhanden sind, wie das in den Abbildungen visualisierte Exposure verdeutlicht. Aufgrund der geringen Anzahl an Trainingsdaten ist die Prognosegüte für Beobachtungen mit den entsprechenden Merkmalsausprägungen eher schwach.

Abschließend wird der bereits in Abschnitt 2.3.3 vorgestellte Nachteil des multinomialen Modells, dass für $\mathbb{P}(Y_1 = A|Y_0 = B)$ positive Wahrscheinlichkeiten geschätzt werden, diskutiert. Tatsächlich ist dieser Effekt marginal, wie die Abbildung 4.2 verdeutlicht. Die

maximale Übergangswahrscheinlichkeit, die für den Zustandswechsel von *beitragsfrei* nach *aktiv* unter Verwendung von Regularisierung und der Variable `beitragsfrei_duration` geschätzt wird, beträgt 1,53%. Zudem gilt für rund 85% aller Beobachtungen im Datensatz mit Anfangszustand *beitragsfrei*, dass die geschätzte Übergangswahrscheinlichkeit in den Zustand *aktiv* kleiner ist als 0,25%. Teilt man die Übergangswahrscheinlichkeiten $\mathbb{P}(Y_1 = A|Y_0 = B)$ auf $\mathbb{P}(Y_1 = B|Y_0 = B)$ und $\mathbb{P}(Y_1 = S|Y_0 = B)$ gemäß ihres Verhältnisses auf, sodass $\mathbb{P}(Y_1 = A|Y_0 = B) = 0$ gilt, kann die Devianz marginal um 0,063% gesenkt werden. Dies verdeutlicht nochmals, dass die Auswirkung der geschätzten positiven Wahrscheinlichkeiten für $\mathbb{P}(Y_1 = A|Y_0 = B)$ vernachlässigbar ist und somit zumindest im vorliegenden Fall kaum als Nachteil der multinomialen Modellstruktur betrachtet werden kann.

4.2.3. Hierarchische Modelle

Bei der folgenden Auswertung der hierarchischen Modelle werden die Ergebnisse der hierarchischen Modellstrukturen I und II diskutiert. Diese beiden Modelle unterscheiden sich, wie in Kapitel 2.3.4 beschrieben, darin, ob die Beobachtungen nach ihrem Anfangszustand getrennt werden (Modell I) oder das Modell in der ersten Stufe auf den kompletten Datensatz angepasst wird (Modell II).

Aus der Tabelle 4.2 geht hervor, dass sich bei beiden Modellstrukturen (d.h. Modell I und II) die Devianzen auf den Trainingsdaten und Testdaten bei den regularisierten Modellen im Vergleich zu den nicht regularisierten Modellen erhöhen. Bei zusätzlicher Berücksichtigung der Variable `beitragsfrei_duration` können die Devianzen im Vergleich zu den regularisierten Modellen ohne diese Variable und im Vergleich zu den nicht regularisierten Modellen aber gesenkt werden.

Bei beiden Modellstrukturen werden unter den getesteten Modellanpassungen die besten Devianzen sowohl auf den Trainingsdaten als auch auf den Testdaten erzielt, wenn neben Regularisierung die Anzahl der Beobachtungsjahre im Zustand *beitragsfrei* berücksichtigt wird, jedoch keine zusätzliche Gewichtung bei der Modellanpassung des Modells in der zweiten Stufe erfolgt. Grundsätzlich sind die Unterschiede bezüglich der Devianz innerhalb einer Modellstruktur jedoch marginal.

Im Vergleich der Devianzen zwischen den beiden hierarchischen Modellstrukturen ist auffallend, dass das hierarchische Modell I besser abschneidet als Modell II. Ein möglicher Grund hierfür ist, dass bei Modell I für die Beobachtungen mit Anfangszustand *beitragsfrei*

ein eigenes Modell angepasst wird, das unabhängig von den Modellen der Beobachtungen mit Anfangszustand *aktiv* ist. Dadurch können insbesondere die Zustandswechsel der Beobachtungen mit Anfangszustand *beitragsfrei* präzise modelliert werden, was eine Erklärung der etwas niedrigeren Devianz des Modells I im Vergleich zu Modell II liefern könnte.

Im Hinblick auf die Trainingszeit ist beobachtbar, dass diese für die regularisierten Modelle aufgrund der durch Kreuzvalidierung bestimmten Regularisierungsparameter λ deutlich höher ist als für die nicht regularisierten Modelle. Vergleicht man die Zeiten zur Modellanpassung jedoch mit der Modellstruktur des aktuellen Forschungsstands, d.h. den unabhängigen Modellen, so fällt auf, dass sich die Trainingszeit durch die hierarchischen Modellstrukturen stark reduzieren lässt. Zudem ist der Tabelle 4.2 zu entnehmen, dass sich die Trainingszeiten beider hierarchischer Modellstrukturen für die einzelnen Modellanpassungen in sehr ähnlichen Größenordnungen befinden.

In Bezug auf die Anzahl der Parameter ist festzustellen, dass sich für die regularisierten Modelle, ähnlich wie bereits beim multinomialen Modell, die Anzahl der Parameter im Vergleich zu den nicht regularisierten Modellen enorm reduziert.
Mit Regularisierung reduziert sich bei Modell I die Anzahl der Parameter ungleich null auf jeweils ein Drittel bzw. weniger als ein Drittel der ursprünglichen Parameterzahl des nicht regularisierten Modells.
Bei Modell II kann die Anzahl der Parameter durch Regularisierung, verglichen mit dem Modell ohne Regularisierung, mehr als halbiert werden. Zudem fällt auf, dass Modell II für alle Modellanpassungen deutlich weniger Parameter besitzt als Modell I. Dem liegt zugrunde, dass bei Modell II nur zwei, anstatt wie bei Modell I drei binomiale logistische Regressionsmodelle angepasst werden müssen.
Insgesamt weisen beide Modellstrukturen die geringste Anzahl an Parametern ungleich null auf, wenn die Anzahl der Jahre im Zustand *beitragsfrei* (d.h. die Variable `beitragsfrei_duration`) berücksichtigt und zusätzliche Gewichtung bei der Modellanpassung des Modells in der zweiten Stufe angewendet wird. Die niedrigste Anzahl an Parametern aller logistischer Regressionsmodelle ergibt sich damit für Modell II unter Verwendung von Regularisierung, Gewichtung und Hinzunahme der Variable `beitragsfrei_duration` mit gerade einmal 94 Koeffizienten ungleich null.

Zur grafischen Beurteilung werden im Folgenden insbesondere die Abbildungen 4.3 und 4.4 herangezogen. Diese veranschaulichen die Schätzungen und Prognosefehler der hierarchischen Modelle auf den Testdaten in Abhängigkeit der Anzahl der Beobachtungsjahre

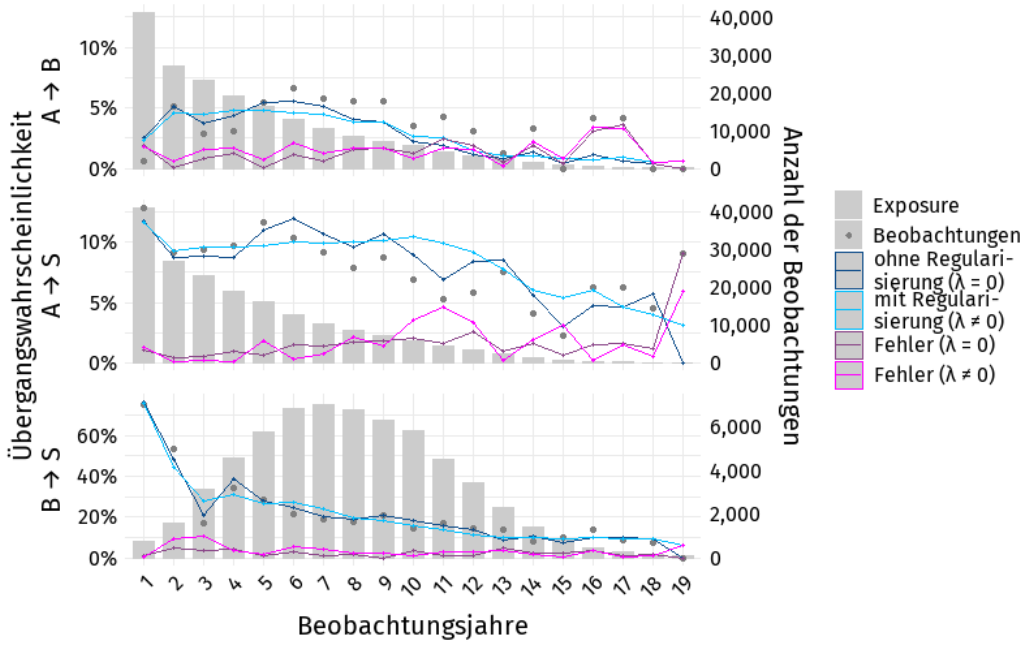


Abbildung 4.3.: Prognosen und Vorhersagefehler des hierarchischen Modells I für die Anzahl der Beobachtungsjahre mit und ohne Regularisierung auf den Testdaten

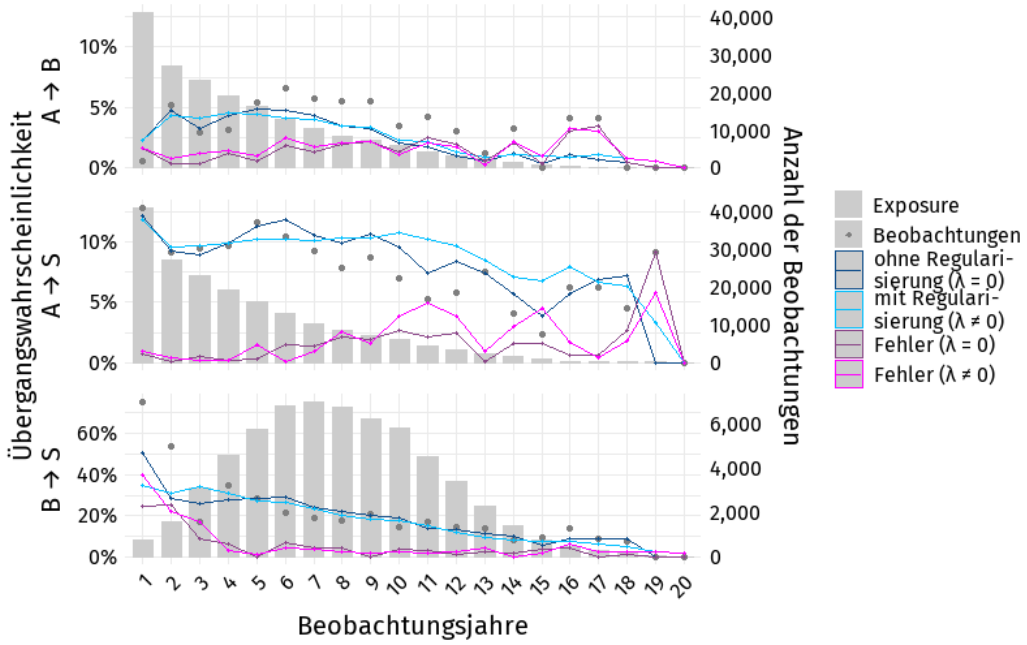


Abbildung 4.4.: Prognosen und Vorhersagefehler des hierarchischen Modells II für die Anzahl der Beobachtungsjahre mit und ohne Regularisierung auf den Testdaten

mit und ohne Regularisierung. Analoge Abbildungen für die Trainingsdaten sowie das Eintrittsalter finden sich im Anhang in Abschnitt A.2.

Anhand der Abbildungen fällt auf, dass die tatsächlich beobachteten Übergangswahrscheinlichkeiten durch beide hierarchischen Modellstrukturen sowohl auf den Trainingsdaten als auch auf den Testdaten präzise modelliert werden können, da die Schätzungen nahe an den tatsächlich beobachteten Übergangswahrscheinlichkeiten liegen. Bei der Visualisierung in Abhängigkeit der Beobachtungsjahre gilt dies insbesondere für den Zustandswechsel von *beitragsfrei* nach *storniert*. Wählt man eine Form der Visualisierung in Abhängigkeit des Eintrittsalters, wie beispielsweise in den Abbildungen A.5 oder A.8, so gilt dieser Befund auch für die anderen beiden möglichen Zustandswechsel.

Bei den Darstellungen bezüglich der Anzahl der Beobachtungsjahre kann beobachtet werden, dass die Wahrscheinlichkeiten des Zustandsübergangs von *aktiv* nach *beitragsfrei* tendenziell unterschätzt werden. Die Übergangswahrscheinlichkeiten des Zustandswechsels von *aktiv* nach *storniert* hingegen werden eher überschätzt. Eine mögliche Erklärung hierfür ist, dass die Modelle in der zweiten Stufe der hierarchischen Modellstrukturen jeweils nur auf Daten trainiert werden, für die bei Anfangszustand *aktiv* tatsächlich ein Zustandswechsel beobachtet wird. Diese Modelle prognostizieren bei der Modellierung der Wahrscheinlichkeiten jedoch auch die Übergangswahrscheinlichkeiten in den Zustand *beitragsfrei* bzw. *storniert* für Beobachtungen, für die tatsächlich kein Zustandswechsel beobachtet wird.

Im Datensatz sind für den Anfangszustand *aktiv* mehr Zustandsübergänge in den Zustand *storniert* als in den Zustand *beitragsfrei* beobachtbar. Deshalb ist es denkbar, dass die Modelle in der zweiten Stufe für Beobachtungen ohne tatsächlich beobachtbaren Zustandswechsel eher den Endzustand *storniert* als den Zustand *beitragsfrei* prognostizieren. Damit folgt, dass diese Modelle die Wahrscheinlichkeiten $\mathbb{P}(Y_1 = S|Y_0 = A, Y_1 = \bar{A})$ tendenziell überschätzen, woraus die Unterschätzung von $\mathbb{P}(Y_1 = B|Y_0 = A, Y_1 = \bar{A})$ resultiert. Dem liegt zugrunde, dass $\mathbb{P}(Y_1 = B|Y_0 = A, Y_1 = \bar{A}) = 1 - \mathbb{P}(Y_1 = S|Y_0 = A, Y_1 = \bar{A})$ für das Modell in der zweiten Stufe gilt.
Der Zustandswechsel von *beitragsfrei* nach *storniert* ist hiervon nicht betroffen, da die zugehörigen Übergangswahrscheinlichkeiten bereits mithilfe der Modelle in der ersten Stufe der hierarchischen Modellstrukturen modelliert werden können.
Bei der Wahl einer anderen Variable auf der x-Achse der Abbildungen fällt dieser Befund des Über- bzw. Unterschätzens von Wahrscheinlichkeiten allerdings geringer aus bzw. ist gar nicht sichtbar (vgl. Abb. A.5 und A.8). Somit könnte dies auch auf die ungünstige

Wahl der Variable `contract_duration` zur Visualisierung der Ergebnisse zurückzuführen sein.

Analog zum multinomialen Modell geht aus den Abbildungen darüber hinaus die Auswirkung der Regularisierung hervor, da die Modellierungen der Übergangswahrscheinlichkeiten durch die regularisierten Modelle wesentlich glatter sind als jene der nicht regularisierten Modelle. Besonders deutlich wird dies anhand der Visualisierungen, die die Prognosen für die Trainingsdaten zeigen, wie beispielsweise in Abbildung A.4 oder A.7. Abschließend ist den Abbildungen zu entnehmen, dass die Güte der Schätzungen für Merkmalsausprägungen mit wenigen Beobachtungen im Datensatz wesentlich schlechter ist als für Merkmalsausprägungen mit vielen Beobachtungen. So werden betragsmäßig höhere Prognosefehler für Beobachtungen mit mehr als 15 Beobachtungsjahren bzw. mit Eintrittsaltern unter 15 oder über 60 Jahren beobachtet. Dies kann erneut auf die geringe Anzahl an Trainingsdaten mit den entsprechenden Merkmalsausprägungen, auf die die Modelle angepasst werden, zurückgeführt werden.

4.2.4. Vergleich der Modelle

In diesem Abschnitt werden die Ergebnisse der logistischen Regressionsmodelle auf Grundlage der Resultate in Tabelle 4.2 miteinander verglichen und die Vorteile bzw. Nachteile der jeweiligen Modellstrukturen diskutiert. Als erste Richtlinie soll dazu das sogenannte *Intercept-Modell* gelten. Dieses Modell besteht nur aus drei Koeffizienten, welche sich durch den jeweiligen Anteil der einzelnen Endzustände an der Gesamtzahl aller Beobachtungen ergeben. Dadurch erhält man für jede Beobachtung im Datensatz die gleichen Wahrscheinlichkeiten für die Endzustände $Y_1 \in \{A, B, S\}$ nach einem Beobachtungsjahr. Aus der Tabelle 4.2 geht hervor, dass grundsätzlich alle in Abschnitt 2.3 eingeführten logistischen Regressionsmodellstrukturen die Devianz gegenüber dem Intercept-Modell halbieren bzw. sogar noch stärker reduzieren können. Dies gilt sowohl für die Trainingsdaten als auch für die Testdaten. Daraus lässt sich ableiten, dass die gewählten logistischen Regressionsmodelle durchaus geeignet sind, um die Übergangswahrscheinlichkeiten der Zustandswechsel zu modellieren.

Von allen untersuchten Modellstrukturen und Modellanpassungen können dabei die unabhängigen logistischen Regressionsmodelle auf den Trainingsdaten sowie auf den Testdaten die niedrigste Devianz erzielen. Nichtsdestotrotz weist diese Modellstruktur auch einige Mängel auf. So muss der ursprüngliche Datensatz zur Modellanpassung in mehrere

Datensätze aufgeteilt werden, um jeden möglichen Zustandswechsel adäquat modellieren zu können. Dadurch ergibt sich je ein Datensatz pro möglichem Zustandswechsel. Zudem muss für jede Art von Zustandsübergang ein separates binomiales logistisches Regressionsmodell angepasst werden. Dies erschwert die Hinzunahme eines neuen Zustandswechsels, da sowohl mehr Trainingszeit aufgewendet werden muss, um noch ein weiteres Modell anzupassen, als auch der ursprüngliche Datensatz erneut bearbeitet und unterteilt werden muss. Zudem erfolgt die Modellanpassung der einzelnen Modelle unabhängig voneinander, sodass gemeinsame Indikatoren eines Zustandsübergangs, die alle Arten von Zustandswechseln gemeinsam haben, nicht erkannt werden können. Die unabhängige Modellstruktur birgt zudem die große Gefahr, negative Wahrscheinlichkeiten für den Verbleib im Zustand *aktiv* zu liefern, da sich diese, wie bereits in Kapitel 2.3.2 beschrieben, durch $\mathbb{P}(Y_1 = A|Y_0 = A) = 1 - \mathbb{P}(Y_1 = B|Y_0 = A) - \mathbb{P}(Y_1 = S|Y_0 = A)$ ergeben. Tatsächlich ist dies zumindest für den vorliegenden Datensatz aber für keine Beobachtung der Fall. So beträgt der maximale Wert, der sich für die Summe der Übergangswahrscheinlichkeiten der Zustandswechsel von *aktiv* nach *beitragsfrei* und *aktiv* nach *storniert* für eine Beobachtung ergibt, $\mathbb{P}(Y_1 = B|Y_0 = A) + \mathbb{P}(Y_1 = S|Y_0 = A) = 0{,}9488$, womit $\mathbb{P}(Y_1 = A|Y_0 = A) > 0$ folgt.

Mithilfe der hierarchischen Modelle bzw. des multinomialen Modells können einige der skizzierten Nachteile der unabhängigen Modellstruktur bei gleichzeitig sehr ähnlicher Devianz und Parameterzahl umgangen werden. Das hierarchische Modell I erzielt nach den unabhängigen Modellen die besten Ergebnisse in Bezug auf die Devianz. Die Modellstruktur ermöglicht dabei, dass der Datensatz für den Anfangszustand *aktiv* nicht getrennt nach den möglichen Endzuständen *beitragsfrei* und *storniert* der Zustandsübergänge betrachtet werden muss. Dadurch können durch das Modell gemeinsame Einflussfaktoren der Zustandswechsel mit Anfangszustand *aktiv* erfasst werden.

Verglichen mit den unabhängigen Modellen besitzt diese Modellstruktur darüber hinaus weniger Parameter bzw. weniger Parameter ungleich null im Falle von Regularisierung. Zudem ist die Zeit, die zur Anpassung des Modells I benötigt wird, ähnlich wie auch beim hierarchischen Modell II, deutlich geringer als die Zeit zur Modellanpassung der unabhängigen Modelle sowie des multinomialen Modells. Dennoch muss beim hierarchischen Modell I der Datensatz zu Beginn nach den Anfangszuständen getrennt werden. Auf der einen Seite führt dies zwar zu tendenziell besseren Ergebnissen bei der Modellierung der Übergangswahrscheinlichkeiten des Zustandswechsels von *beitragsfrei* nach *storniert*. Auf der anderen Seite werden allerdings gemeinsame Einflussfaktoren, die auf alle drei Arten

von Zustandsübergängen zutreffen, beim Zustandswechsel von *beitragsfrei* nach *storniert* nicht berücksichtigt.

Bei der hierarchischen Modellstruktur II hingegen kann das Modell der ersten Stufe auf dem kompletten Datensatz trainiert werden, sodass keine Aufspaltung des Datensatzes nach den Anfangszuständen nötig ist. Dies hat zum einen den Vorteil, dass, anders als bei den unabhängigen Modellen oder Modell I, nur zwei statt drei binomiale logistische Regressionsmodelle benötigt werden, was die Laufzeit der Modellanpassung sowie die Anzahl der Parameter (ungleich null) verringert. Generell hat das hierarchische Modell II dadurch über alle logistischen Regressionsmodellstrukturen hinweg die geringsten Parameterzahlen und zusammen mit Modell I die geringsten Trainingszeiten. Zum anderen können durch diese Modellstruktur die gemeinsamen Treiber der Zustandswechsel aller im Datensatz beobachtbarer Zustandsübergänge berücksichtigt und in einem gemeinsamen Modell erfasst werden. Somit kann eine Abhängigkeit bei der Modellierung der Übergangswahrscheinlichkeiten aller Zustandsübergänge hergestellt werden. Die Devianz auf den Trainings- und Testdatensätzen ist allerdings etwas höher als bei allen anderen diskutierten logistischen Regressionsmodellen, wobei der Unterschied insbesondere zum multinomialen Modell nur marginal ist.

Ein Nachteil, der auf beide hierarchische Modellstrukturen zutrifft, ist, dass ähnlich wie bei den unabhängigen Modellen, bei Hinzunahme eines weiteren Zustandswechsels ein weiteres binomiales logistisches Regressionsmodell in der zweiten Stufe angepasst werden muss. Dazu ist darüber hinaus ein weiterer Datensatz erforderlich, der aus der ursprünglichen Datengrundlage gewonnen werden muss. Trotzdem bleibt bei Hinzunahme eines weiteren Zustandswechsels insbesondere bei Modell II die Eigenschaft erhalten, gemeinsame Einflussfaktoren der Zustandsübergänge zu erfassen und somit gemeinsame Treiber dieser Übergänge bei der Modellierung zu berücksichtigen.

Für die multinomiale Modellstruktur wird nur ein einziges multinomiales logistisches Regressionsmodell zur Schätzung aller Übergangswahrscheinlichkeiten benötigt. Dadurch kann durch Verwendung dieser Modellstruktur der Nachteil, dass bei einem zusätzlichen Zustandswechsel ein weiteres Modell angepasst und ein weiterer Datensatz aus den ursprünglichen Daten abgeleitet werden muss, umgangen werden. Bei einem neuen Zustandswechsel müssen im Datensatz, auf den das Modell angepasst wird, nur jene Variablen, die den Anfangszustand bzw. Endzustand eines Vertrages in einem Beobachtungsjahr kodieren, um die neuen Zustände erweitert werden. Zudem muss der Datensatz beim multinomialen Modell im Gegensatz zu den anderen vorgestellten Modellen, die auf binomialer logistischer

Regression basieren, nicht hinsichtlich der Anfangszustände sowie Endzustände getrennt oder bearbeitet werden. Die Devianz dieser Modellstruktur liegt dabei leicht über der des hierarchischen Modells I sowie über der Devianz der unabhängigen Modelle und leicht unter jener des Modells II und befindet sich somit im Bereich der Devianzen anderer bereits diskutierter Modellstrukturen.

Eine mögliche Erklärung für die etwas höhere Devianz des multinomialen Modells im Vergleich zu den Modellstrukturen, die auf drei binomialen logistischen Regressionsmodellen basieren, könnte zumindest im Falle der Verwendung von Regularisierung darin liegen, dass bei den unabhängigen Modellen bzw. dem hierarchischen Modell I mehr Flexibilität in Bezug auf den Regularisierungsparameter λ besteht. Während beim multinomialen Modell nur ein einzelner Regularisierungsparameter für die komplette Modellstruktur bestimmt wird, werden bei den unabhängigen Modellen bzw. bei Modell I je drei verschiedene Regularisierungsparameter ermittelt. Diese Flexibilität zeigt sich beispielsweise bei den unabhängigen Modellen. Hier weist das Modell zur Modellierung der Übergangswahrscheinlichkeiten $\mathbb{P}(Y_1 = S|Y_0 = B)$ mit $\lambda = 0{,}00417$ einen Regularisierungsparameter auf, der ungefähr 10-mal höher ist als der Parameter $\lambda = 0{,}00047$ des Modells zur Modellierung von $\mathbb{P}(Y_1 = B|Y_0 = A)$. Für das multinomiale Modell ergibt sich als einziger Regularisierungsparameter für alle modellierten Zustandswechsel $\lambda = 0{,}00067$.
In Bezug auf die Dauer zur Anpassung des multinomialen Modells fällt auf, dass diese insbesondere bei Verwendung von Regularisierung deutlich über den Trainingszeiten der anderen Modellstrukturen liegt.
Des Weiteren werden durch das multinomiale Modell auch Übergangswahrscheinlichkeiten für Zustandswechsel geschätzt, die im Datensatz tatsächlich gar nicht beobachtbar sind. Die Analysen in Abschnitt 4.2.2 haben jedoch gezeigt, dass diese Wahrscheinlichkeiten zumindest im vorliegenden Fall marginal sind. Dies kann als positiv hinsichtlich der Eignung des Modells zur Modellierung der Übergangswahrscheinlichkeiten gewertet werden, da durch die Modellstruktur erkannt wird, dass im Datensatz keine Beobachtungen mit einem Zustandswechsel von *beitragsfrei* nach *aktiv* enthalten sind.

Abschließend ist festzuhalten, dass die Komplexität durch Regularisierung bei allen Modellstrukturen enorm reduziert wird, was die Prognosefähigkeit auf zuvor nicht gesehenen Daten erhöht. Zudem zeigt sich für alle Modellstrukturen, dass die zusätzliche Betrachtung der Beobachtungsjahre im Zustand *beitragsfrei* hilfreich ist, da dadurch die Devianz deutlich gesenkt werden kann. Die Prognosefehler der regularisierten logistischen Regressionsmodelle auf den Testdaten können grafisch in Abhängigkeit von der Anzahl der Be-

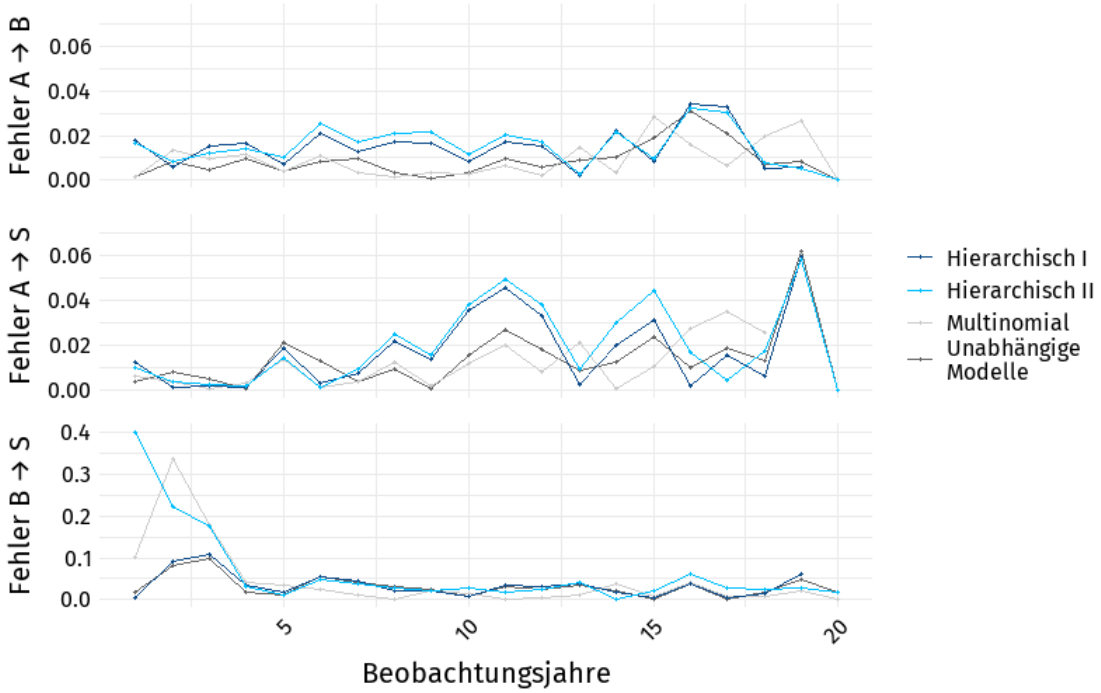


Abbildung 4.5.: Vergleich der Vorhersagefehler auf den Testdaten bei Modellanpassung mit Regularisierung für die Anzahl der Beobachtungsjahre

obachtungsjahre anhand der Abbildung 4.5 verglichen werden. Daraus geht hervor, dass sich die Vorhersagen der beiden hierarchischen Modelle sehr stark ähneln, während zudem eine große Ähnlichkeit zwischen den Modellierungen der Übergangswahrscheinlichkeiten durch das multinomiale Modell bzw. die unabhängigen Modelle besteht.

Darüber hinaus ist nochmals der Befund aus Abschnitt 4.2.3 erkennbar, dass für diese Art der Visualisierung in Abhängigkeit der Anzahl der Beobachtungsjahre die hierarchischen Modelle für die Zustandswechsel von *aktiv* nach *beitragsfrei* bzw. von *aktiv* nach *storniert* tendenziell höhere Prognosefehler aufweisen als die anderen Modellstrukturen. Im Anhang A.3 findet sich eine analoge Abbildung für das Eintrittsalter (Abb. A.10), welche zeigt, dass dies jedoch nur auf die Art der Darstellung zurückzuführen sein könnte. Dies wird ebenfalls durch die Devianz als Gütemaß für die Prognosegüte der Modelle gestützt, welche für die hierarchischen Modellstrukturen in einem ähnlichen Bereich wie die der unabhängigen Modelle bzw. die des multinomialen Modells liegt.

4.3. Neuronale Netze

Dieses Kapitel dient der Evaluation, inwiefern neuronale Netze als *Black-Box-Modelle* bei der Modellierung der Übergangswahrscheinlichkeiten besser abschneiden als die interpretierbaren aber weniger flexiblen logistischen Regressionsmodelle. Um genauer analysieren zu können, wie neuronale Netze im Vergleich zu den logistischen Regressionsmodellen zu bewerten sind, wird im Folgenden zunächst auf die konkrete Modellanpassung des neuronalen Netzes, das zur Modellierung der Übergangswahrscheinlichkeiten verwendet wird, eingegangen. Im Anschluss daran werden die Ergebnisse des Netzes evaluiert und mit den Ergebnissen der logistischen Regressionsmodelle verglichen.

4.3.1. Umsetzung der Modellanpassung

Bei dem verwendeten neuronalen Netz handelt es sich um ein *Multilayer-Perceptron (MLP)* bestehend aus je einer Eingabeschicht, einer verdeckten Schicht und einer Ausgabeschicht. Auch tiefere neuronale Netze, d.h. mit mehreren verdeckten Schichten, wären hier denkbar. Jedoch soll im Folgenden der Fokus auf der Frage liegen, inwiefern bereits sehr simple neuronale Netze bessere Prognosen liefern können als logistische Regressionsmodelle. Zudem hat sich gezeigt, dass die Verwendung komplexerer Modellarchitekturen, die beispielsweise zeitliche Abhängigkeiten besser modellieren können, zu keiner signifikanten Verbesserung hinsichtlich der Prognosegüte führt.

Hyperparameter	**getestete Werte**	**bester Wert**
Anzahl Neuronen (verdeckte Schicht)	$5, 10, \ldots, 150$	105
Aktivierungsfunktion (verdeckte Schicht)	ReLU, tanh	ReLU
Dropout-Rate (verdeckte Schicht)	$0.1, \ldots, 0.5$	0.2
Optimizer	Adam, RMSprop[24]	Adam

Tabelle 4.3.: Getestete Hyperparameter während des Hyperparameter-Tunings

Die Bestimmung der optimalen *Hyperparameter* des neuronalen Netzes, d.h. der Anzahl der Neuronen, der Aktivierungsfunktion und der Dropout-Rate der verdeckten Schicht sowie des Optimizers erfolgt durch ein sogenanntes *Hyperparameter-Tuning*. Dazu werden verschiedene Modellanpassungen untersucht und auf Grundlage ihrer Vorhersagegenauigkeit

[24] Weitere Informationen zum *Adam*-Optimizer finden sich Kingma und Ba (2015) [14] und zum *RMSprop*-Optimizer in Géron (2017) [8]

verglichen, wodurch die geeignetste Modellanpassung gefunden werden soll. Die getesteten und jeweils besten Hyperparameter finden sich in der Tabelle 4.3.

Das beste Modell, das sich auf Grundlage des *Hyperparameter-Tunings* findet, besitzt 153 Neuronen in der Eingabeschicht, 105 Neuronen in der verdeckten Schicht und 3 Neuronen in der Ausgabeschicht. Als Aktivierungsfunktion in der verdeckten Schicht dient die *ReLU*-Funktion und in der Ausgabeschicht die *Softmax*-Funktion. Die Anpassung der Gewichte und Biases erfolgt basierend auf dem *Adam*-Optimizer, wobei in jedem Trainingsschritt 20% aller Gewichte zwischen den Neuronen der Zwischenschicht und der Ausgabeschicht auf null gesetzt werden. Zudem wird das Training des Modells zur zusätzlichen Reduktion von *Overfitting* beendet, sobald der Wert der Verlustfunktion drei Trainingsschritte in Folge nicht weiter reduziert werden kann.

4.3.2. Ergebnisse

	Devianz[26]		
Modell	**Training**	**Test**	**Parameter**
Intercept	94.924	31.590	3
Unabhängige Modelle	45.120	15.450	354
Multinomiales Modell[25]	47.792	16.132	106
Hierarchisches Modell I[25]	46.658	15.724	87
Hierarchisches Modell II[25]	47.620	15.990	76
Neuronales Netz	40.079	14.572	16.173

Tabelle 4.4.: Ergebnisse des neuronalen Netzes mit 60.000 Trainingsdaten und 20.000 Testdaten im Vergleich

Im Folgenden werden die Ergebnisse des neuronalen Netzes ausgewertet und mit jenen der logistischen Regressionsmodelle verglichen. Dies erfolgt zunächst auf Grundlage der Tabelle 4.4, welche die Ergebnisse für ausgewählte Modellstrukturen auf insgesamt 80.000 im Datensatz enthaltenen Beobachtungen aufführt. Grund für die Reduzierung der Datenmenge zur Anpassung des neuronalen Netzes ist die enorme Trainingsdauer, die bei der Modellanpassung auf den vollen Datensatz benötigt worden wäre. Dennoch kann auch bei dieser Teilmenge von Beobachtungen aus dem gesamten Datensatz davon ausgegangen werden, dass die Ergebnisse aufgrund der Datenmenge repräsentativ sind.

[25] mit Regularisierung und Berücksichtigung der Jahre im Zustand *beitragsfrei*
[26] je niedriger, desto besser

Beim multinomialen Modell sowie den hierarchischen Modellen in der Tabelle handelt es sich jeweils um die regularisierte Form unter Berücksichtigung der Anzahl der Jahre im Zustand *beitragsfrei*, da für diese Modellanpassung bei allen Modellstrukturen die besten Ergebnisse hinsichtlich der Devianz erzielt werden. Bei den unabhängigen Modellen hingegen wird keine Regularisierung verwendet, was dem herkömmlichen Ansatz wie in Hanewald et al. (2018) [10] entspricht.

Bezüglich der Ergebnisse ist festzuhalten, dass das neuronale Netz die geringste Devianz unter allen betrachteten Modellstrukturen aufweist. Dies gilt sowohl für die Trainingsdaten als auch für die Testdaten, wobei der Unterschied auf den Testdaten prozentual etwas geringer ausfällt als auf den Trainingsdaten. Sehr auffallend ist die deutlich höhere Parameterzahl des neuronalen Netzes, was auf dessen mehrschichtige Modellstruktur zurückzuführen ist. So ist für jede Verbindung zwischen den Neuronen der vorherigen und nachfolgenden Schicht ein Parameter notwendig, der angepasst werden muss. Zudem wird in der verdeckten Schicht sowie der Ausgabeschicht der Bias angepasst. Aufgrund dieser hohen Anzahl an Parametern wird deutlich, warum neuronale Netze kaum bzw. nicht zu interpretieren sind. Anders als bei den logistischen Regressionsmodellen, bei denen ein einzelner Koeffizient Aufschluss über den Einfluss eines Merkmals auf die Prognose des Modells gibt, ist es bei neuronalen Netzen aufgrund ihrer Komplexität kaum möglich, die Entscheidungsfindung des Modells nachzuvollziehen.

Die Ergebnisse in der Tabelle 4.4 fassen die Stärken und Schwächen neuronaler Netze gegenüber den logistischen Regressionsmodellen somit gut zusammen: Einerseits sind neuronale Netze sehr flexibel einsetzbar und auf eine Vielzahl unterschiedlicher Problemstellungen anwendbar. Aus diesem Grund eignen sie sich gut für reine Prognose- bzw. Modellierungsaufgaben und können hinsichtlich der Devianz bessere Ergebnisse erzielen als die logistischen Regressionsmodelle. Andererseits handelt es sich aufgrund der Komplexität der Modellstruktur aber um *Black-Box-Modelle*, deren Entscheidungsfindung kaum nachvollzogen werden kann.
Die Modellierung weiterer Zustandswechsel ist bei neuronalen Netzen, ähnlich wie beim multinomialen logistischen Regressionsmodell, problemlos möglich. Zur Erweiterung des Modells um einen weiteren Zustandswechsel müssen lediglich den Kovariablen, die die Anfangszustände bzw. Endzustände der Beobachtungen kodieren, die Merkmalsausprägungen der neuen Zustände hinzugefügt werden.

Grafisch kann die Güte des neuronalen Netzes in Abhängigkeit der Anzahl der Beobachtungsjahre auf den Testdaten anhand der Abbildung 4.6 bewertet werden. Die Abbildung

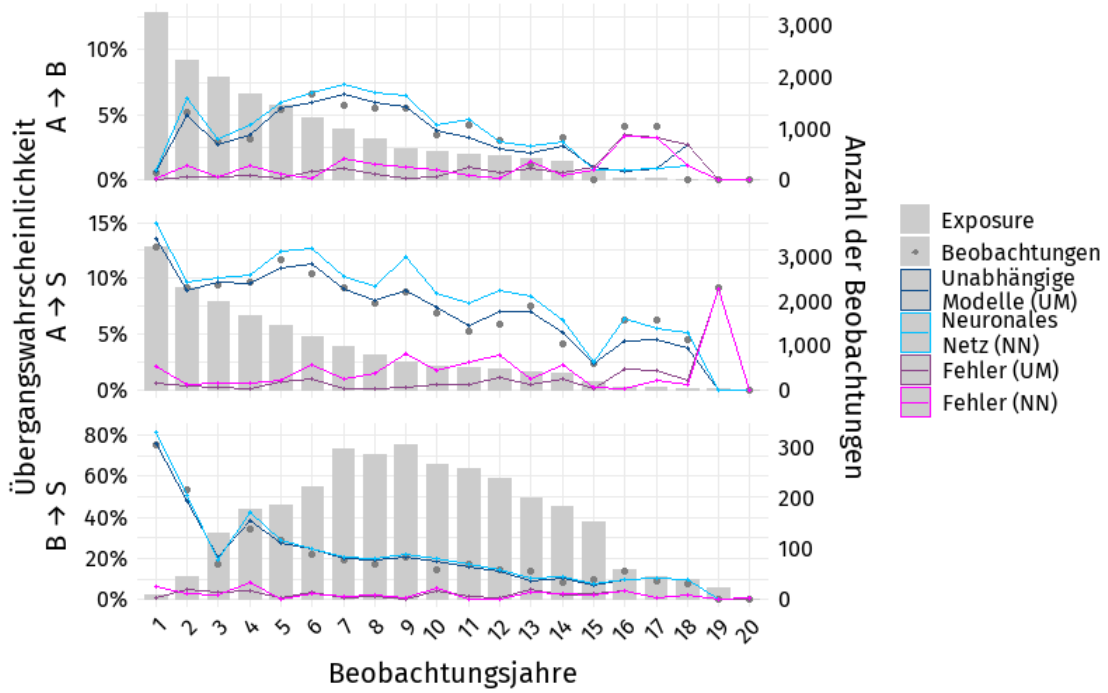


Abbildung 4.6.: Prognosen und Vorhersagefehler des neuronalen Netzes für die Anzahl der Beobachtungsjahre im Vergleich mit den unabhängigen Modellen auf den Testdaten

enthält zudem die Vorhersagen der unabhängigen Modelle ohne $\mathcal{L}^1$-Regularisierung, d.h. des Modells des bisherigen Forschungsstands, sowie die Prognosefehler beider Modellstrukturen. Analoge Abbildungen für das Eintrittsalter sowie die Trainingsdaten finden sich im Anhang A.4.

Aus den Abbildungen geht hervor, dass das neuronale Netz die beobachteten Übergangswahrscheinlichkeiten grundsätzlich genau abbilden kann, da die Schätzungen nahe an den tatsächlich beobachtbaren Übergangswahrscheinlichkeiten liegen. Dennoch fällt auf, dass das neuronale Netz bei Verwendung der vorliegenden Art der Visualisierung in Abhängigkeit der Beobachtungsjahre die Übergangswahrscheinlichkeiten der Zustandswechsel von *aktiv* nach *beitragsfrei* bzw. *aktiv* nach *storniert* tendenziell überschätzt, während die des Zustandsübergangs von *beitragsfrei* nach *storniert* sehr präzise abgebildet werden. Dies könnte, ähnlich wie bereits in den vorigen Abschnitten diskutiert, auf die Art der Darstellung zurückzuführen sein, da der Effekt bei Betrachtung der Übergangswahrscheinlichkeiten bzgl. des Eintrittsalters (vgl. Abb. A.12 und A.13) kaum bzw. nicht erkennbar ist.

5. Zusammenfassung

Das Ziel dieser Arbeit ist es, Modelle vorzustellen und zu evaluieren, mit welchen in einem Mehrzustandsmodell die Übergangswahrscheinlichkeiten für Beitragsfreistellung und Storno von Versicherungsverträgen modelliert werden können. Es ist erwartbar, dass Beitragsfreistellung und Storno durch ähnliche Faktoren beeinflusst werden. Folglich sollten die Modelle für die einzelnen Zustandswechsel der Verträge ähnliche Muster aufweisen. Deshalb lag der Fokus dieser Arbeit insbesondere auf Modellstrukturen, die bei der Modellanpassung gemeinsame Indikatoren der verschiedenen Zustandswechsel bei der Schätzung der Übergangswahrscheinlichkeiten berücksichtigen können.

Im aktuariellen Umfeld wird dazu häufig auf die logistische Regression als Modellklasse zurückgegriffen, deren Grundlagen zu Beginn dieser Arbeit eingeführt wurden. Dabei wurde sowohl die binomiale als auch die multinomiale logistische Regression vorgestellt. Zudem wurde die Theorie der $\mathcal{L}^1$-Regularisierung erläutert. Diese Regularisierung war von entscheidender Bedeutung, um die Komplexität der Modelle zu reduzieren.

Darauf aufbauend wurden die Modellstrukturen zur Modellierung der Übergangswahrscheinlichkeiten, die auf logistischer Regression basieren, beschrieben. Neben dem Modell, bei dem die Übergangswahrscheinlichkeiten der Zustandswechsel völlig unabhängig voneinander mittels binomialer logistischer Regression geschätzt werden, wurde ein multinomiales logistisches Regressionsmodell eingeführt. Zudem wurden zwei Modelle vorgestellt, bei denen mehrere binomiale logistische Regressionsmodelle hierarchisch angeordnet werden.

Ein weiterer Gegenstand der Untersuchungen in dieser Arbeit ist, inwiefern nicht interpretierbare Modelle die Übergangswahrscheinlichkeiten der Zustandswechsel genauer modellieren können als die interpretierbaren logistischen Regressionsmodelle. Deshalb wurden zusätzlich neuronale Netze zur Modellierung der Übergangswahrscheinlichkeiten herangezogen, deren Aufbau, Modellanpassung und Funktionsweise erläutert wurden. Zur Quantifizierung der Prognosegüte aller in dieser Arbeit diskutierten Modelle wurde dabei stets

die Devianz als Gütemaß verwendet, welche ebenfalls zu Beginn dieser Arbeit definiert wurde.

Sowohl die Modellstrukturen basierend auf logistischer Regression, als auch das neuronale Netz wurden auf den Datensatz eines europäischen Lebensversicherers angepasst. Dies diente als Grundlage zur Bewertung der Modelle und zur Klärung der Frage, inwiefern die Berücksichtigung gemeinsamer Treiber der Zustandswechsel bei der Modellanpassung eine Verbesserung der Prognosegüte der Modelle bringen kann. Dabei wurden verschiedene Varianten der Modellanpassung, wie beispielsweise mit und ohne Regularisierung, untersucht. Es hat sich gezeigt, dass die Modellstrukturen, die Abhängigkeiten zwischen den Zustandswechseln bei der Modellanpassung berücksichtigen, hinsichtlich der Prognosegüte etwas schlechter abschneiden als das Modell, bei dem die Übergangswahrscheinlichkeiten der Zustandswechsel unabhängig voneinander geschätzt werden. Dies gilt sowohl für den Trainingsdatensatz, als auch für den Testdatensatz. Dennoch sind die Unterschiede hinsichtlich der Devianz zwischen den Modellstrukturen, die auf logistischer Regression basieren, insbesondere auf den vorher unbekannten Testdaten eher gering. Folglich handelt es sich auch bei den Modellen, die mögliche gemeinsame Einflussfaktoren der Zustandswechsel berücksichtigen, um durchaus geeignete Modellstrukturen zur Modellierung der Übergangswahrscheinlichkeiten.

So bringen die hierarchisch angeordneten binomialen logistischen Regressionsmodelle den Vorteil einer geringeren Anzahl an Parametern ungleich null und einer folglich höheren Interpretierbarkeit. Zudem sind die Zeiten, die zur Modellanpassung benötigt werden, bei dieser Art von Modellstruktur deutlich geringer als bei allen anderen betrachteten Modellen. Das multinomiale Modell wiederum lässt sich im Vergleich zu den Modellen, die auf binomialer logistischer Regression basieren, wesentlich leichter um einen weiteren Zustand erweitern. Dies liegt daran, dass der zusätzliche Zustand lediglich dem Datensatz hinzugefügt werden, das Modell in seiner Struktur aber nicht verändert werden muss.

Des Weiteren haben die Analysen in dieser Arbeit ergeben, dass unter Verwendung von Regularisierung die Prognosegüte aller Modellstrukturen, die auf logistischer Regression beruhen, insbesondere aber auf den Trainingsdaten im Vergleich zu den nicht regularisierten Modellen etwas verschlechtert wird. Dies ist darauf zurückzuführen, dass mithilfe von Regularisierung die Überanpassung des Modells auf den Trainingsdatensatz reduziert wird.
Bei Bearbeitung des Datensatzes durch Hinzunahme der Anzahl der Jahre eines Vertrags im Zustand *beitragsfrei* als weitere Variable konnten die Prognosen der logistischen Regres-

sionsmodelle jedoch wieder verbessert werden. So wurden bei den vorgestellten Modellen, die gemeinsame Indikatoren der Zustandswechsel berücksichtigen können, die besten Ergebnisse hinsichtlich des Gütemaßes Devianz durch die regularisierten Modelle erzielt, bei denen zusätzlich die Anzahl der Jahre im Zustand *beitragsfrei* berücksichtigt wurde.

Abschließend hat sich gezeigt, dass bereits ein sehr einfaches neuronales Netz bessere Ergebnisse in Bezug auf die Prognosegüte erzielen kann als die logistischen Regressionsmodelle. Dennoch sind diese Ergebnisse mit Vorsicht zu genießen, da neuronale Netze im Gegensatz zu den logistischen Regressionsmodellen schwer bzw. nicht interpretierbar sind und bereits sehr einfache neuronale Netze enorm viele Parameter benötigen. Insofern eignen sich neuronale Netze insbesondere für Anwendungen, bei denen der Fokus auf der Güte der Vorhersagen liegt, während auf logistische Regressionsmodelle zurückgegriffen werden sollte, wenn die Entscheidungsfindung und somit die Erklärbarkeit des Modells von Wichtigkeit sind.

Zusammenfassend wurden in dieser Arbeit Strukturen zur Modellierung von Übergangswahrscheinlichkeiten in einem Mehrzustandsmodell untersucht. Dabei sollten insbesondere mögliche gemeinsame Treiber der Zustandswechsel bei der Modellanpassung berücksichtigt werden, wobei sowohl interpretierbare logistische Regressionsmodelle als auch neuronale Netze verwendet wurden. Außerdem wurden die betrachteten Modellstrukturen hinsichtlich ihrer Prognosegüte anhand eines Echtdatensatzes getestet, wofür der Datensatz eines international tätigen Lebensversicherers genutzt wurde.

Ausgehend von den Ergebnissen dieser Arbeit könnte die Verwendung einer ordinalen Struktur bei einem multinomialen logistischen Regressionsmodell Gegenstand weiterführender Analysen sein. Dazu könnte den Zuständen im Mehrzustandsmodell eine aufsteigende Reihenfolge von *aktiv* nach *storniert* unterstellt werden. Zudem könnte untersucht werden, inwiefern die Kombination logistischer Regressionsmodelle und neuronaler Netze die Prognosegüte verbessert. So könnten beispielsweise die Prognosen neuronaler Netze mit in den Datensatz aufgenommen und dem logistischen Regressionsmodell als Kovariable übergeben werden. Alternativ könnten die voneinander unabhängigen Prognosen beider Modellklassen zur finalen Modellierung der Übergangswahrscheinlichkeiten kombiniert werden. In beiden Fällen bliebe dabei eine gewisse Interpretierbarkeit der Modellstruktur durch das logistische Regressionsmodell erhalten.

A. Anhang

A.1. Multinomiales Modell

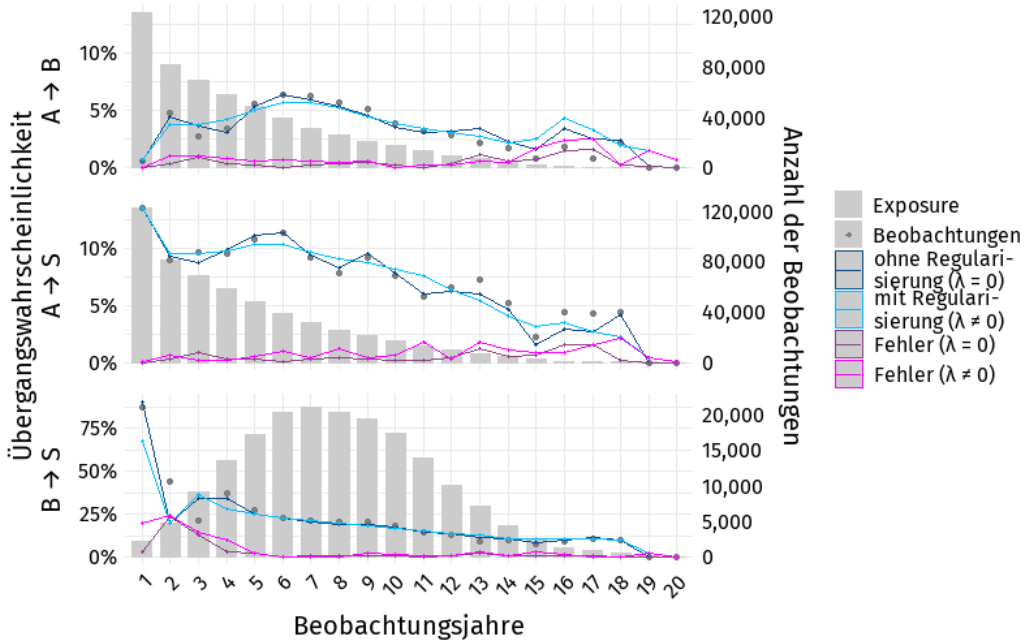


Abbildung A.1.: Prognosen und Vorhersagefehler des multinomialen Modells für die Anzahl der Beobachtungsjahre mit und ohne Regularisierung auf den Trainingsdaten

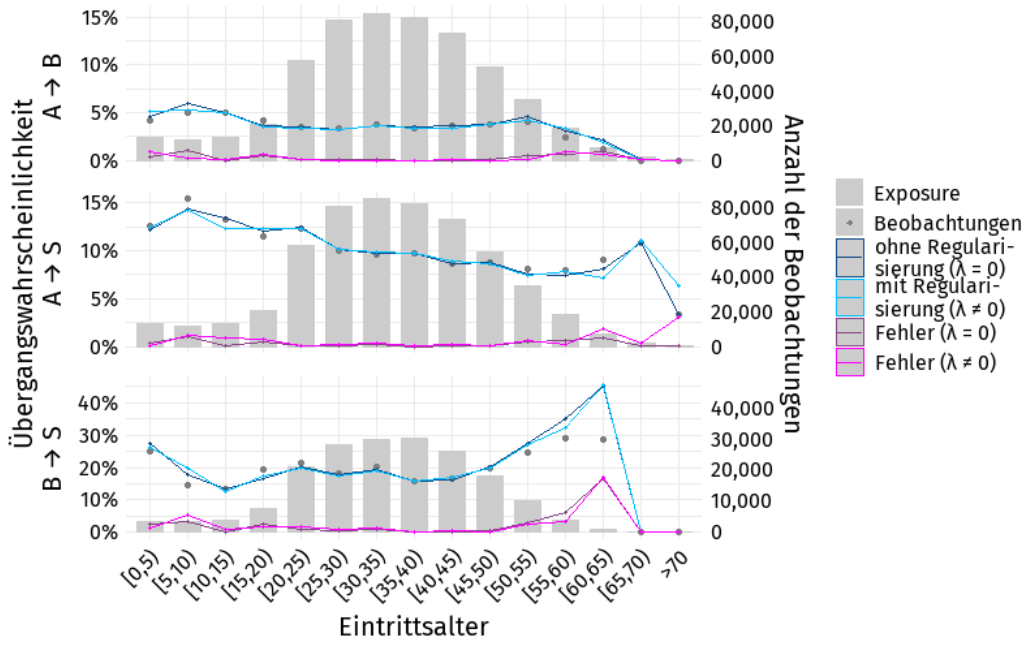


Abbildung A.2.: Prognosen und Vorhersagefehler des multinomialen Modells für das Eintrittsalter mit und ohne Regularisierung auf den Trainingsdaten

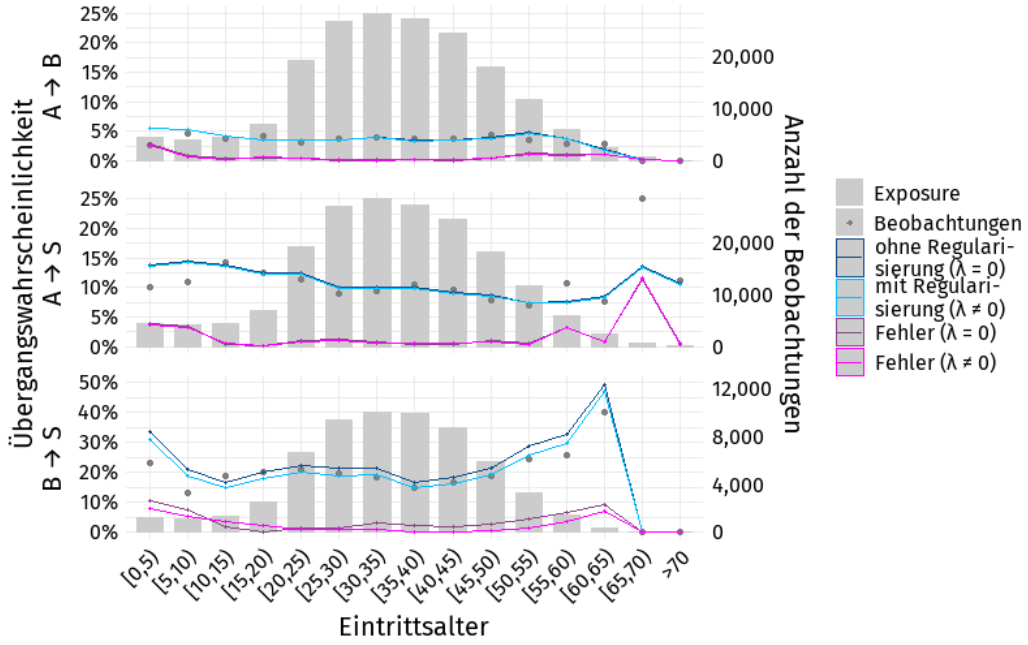


Abbildung A.3.: Prognosen und Vorhersagefehler des multinomialen Modells für das Eintrittsalter mit und ohne Regularisierung auf den Testdaten

A.2. Hierarchische Modelle

A.2.1. Modell I

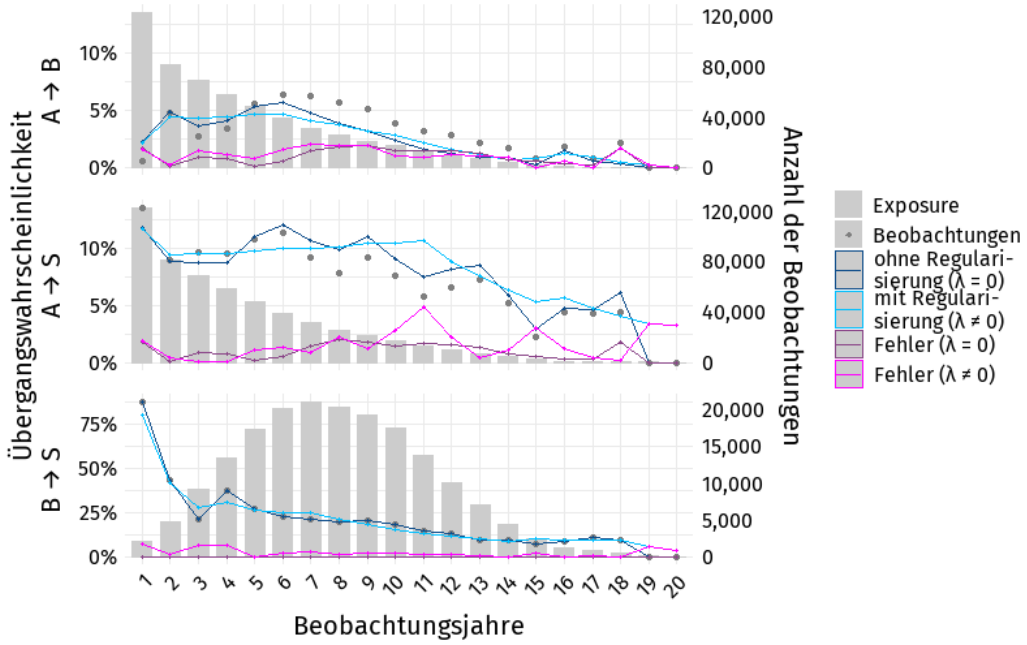


Abbildung A.4.: Prognosen und Vorhersagefehler des hierarchischen Modells I für die Anzahl der Beobachtungsjahre mit und ohne Regularisierung auf den Trainingsdaten

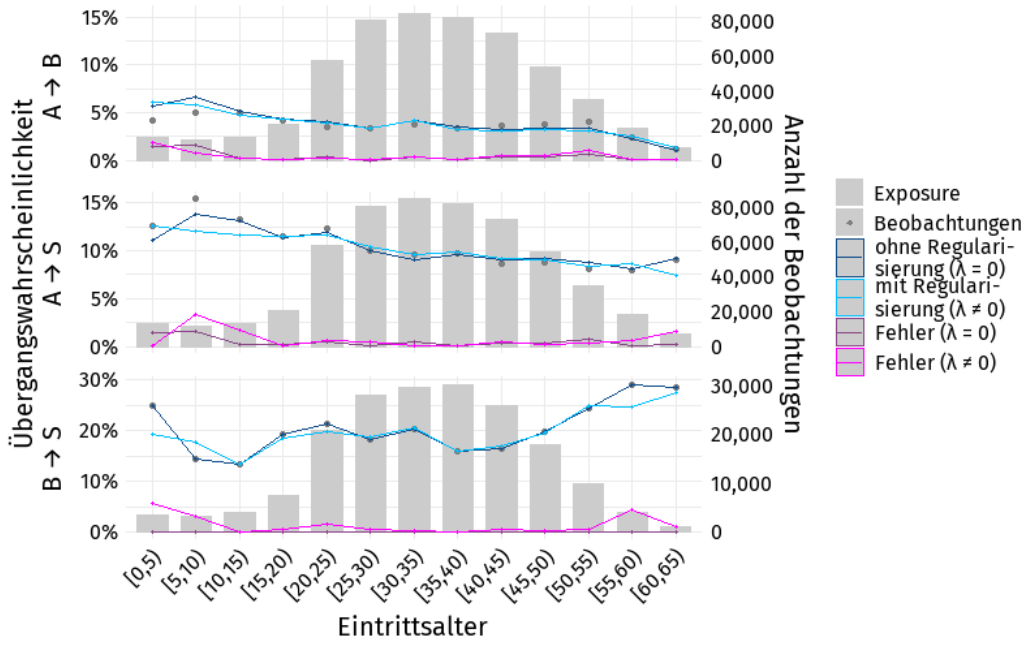


Abbildung A.5.: Prognosen und Vorhersagefehler des hierarchischen Modells I für das Eintrittsalter mit und ohne Regularisierung auf den Trainingsdaten

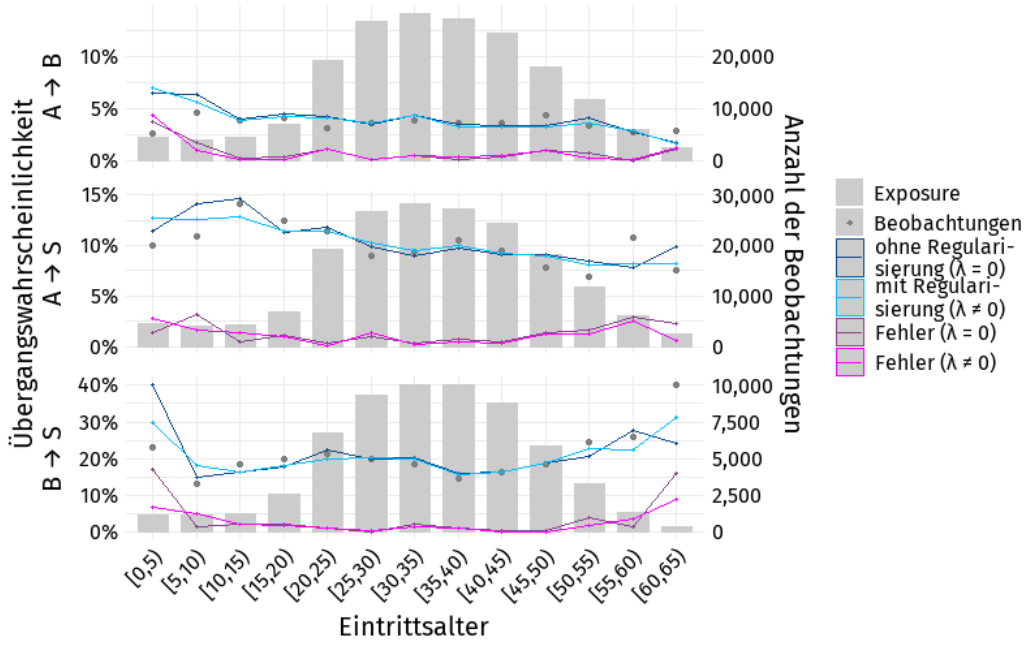


Abbildung A.6.: Prognosen und Vorhersagefehler des hierarchischen Modells I für das Eintrittsalter mit und ohne Regularisierung auf den Testdaten

A.2.2. Modell II

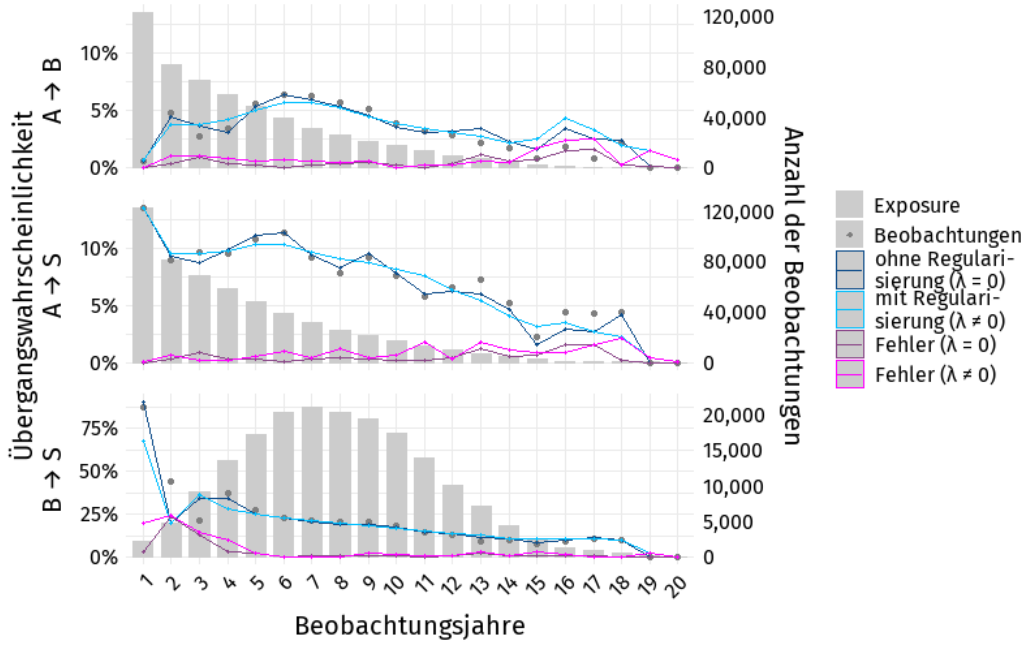


Abbildung A.7.: Prognosen und Vorhersagefehler des hierarchischen Modells II für die Anzahl der Beobachtungsjahre mit und ohne Regularisierung auf den Trainingsdaten

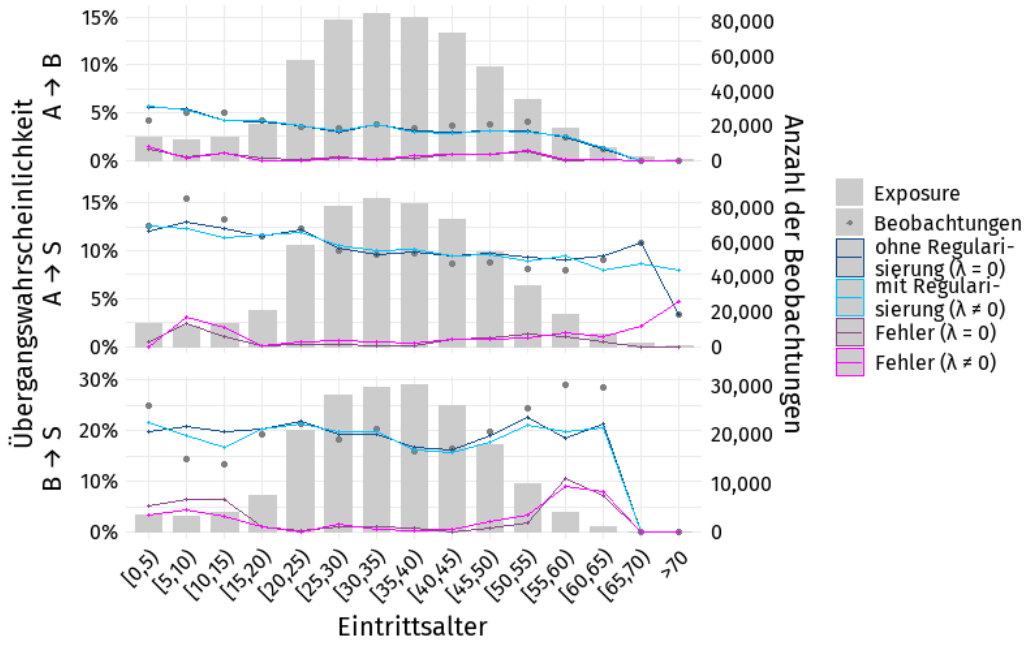


Abbildung A.8.: Prognosen und Vorhersagefehler des hierarchischen Modells II für das Eintrittsalter mit und ohne Regularisierung auf den Trainingsdaten

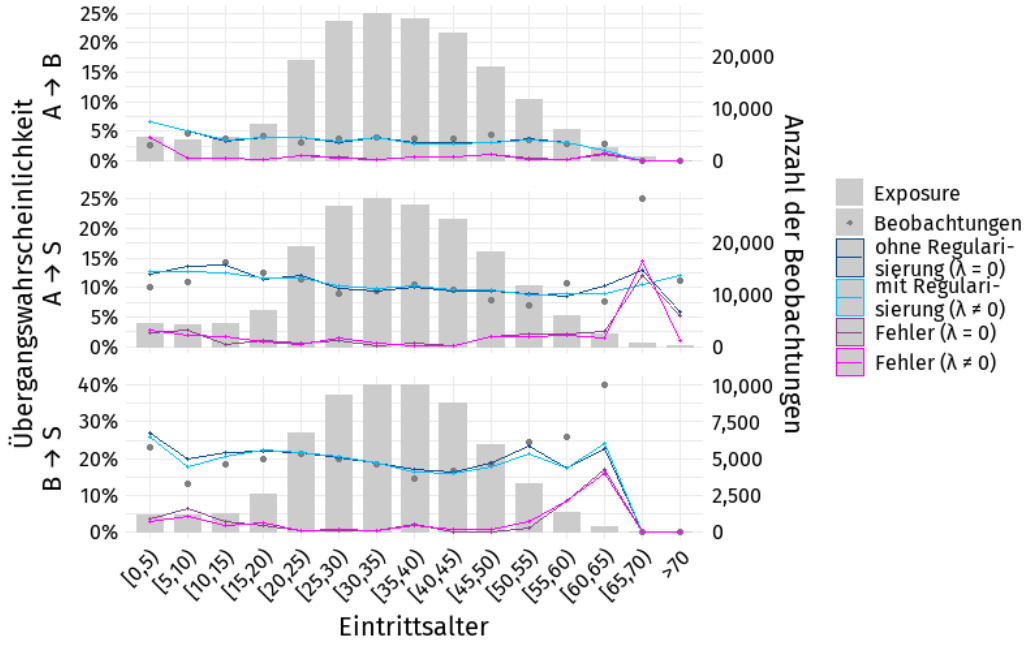


Abbildung A.9.: Prognosen und Vorhersagefehler des hierarchischen Modells II für das Eintrittsalter mit und ohne Regularisierung auf den Testdaten

A.3. Vergleich

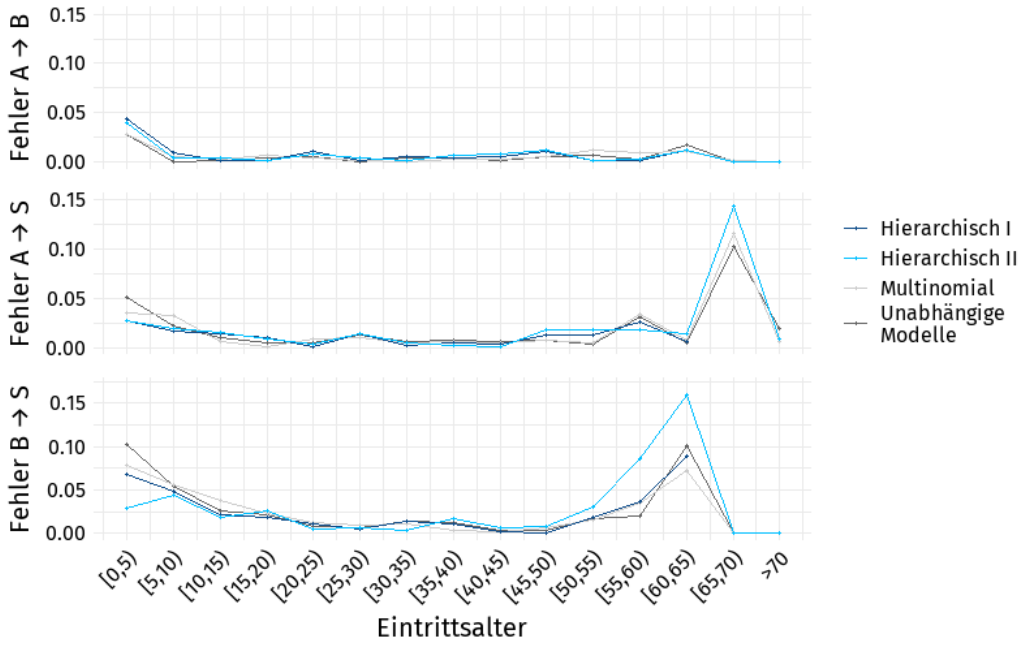


Abbildung A.10.: Vergleich der Vorhersagefehler der Regressionsmodelle für das Eintrittsalter

A.4. Neuronales Netz

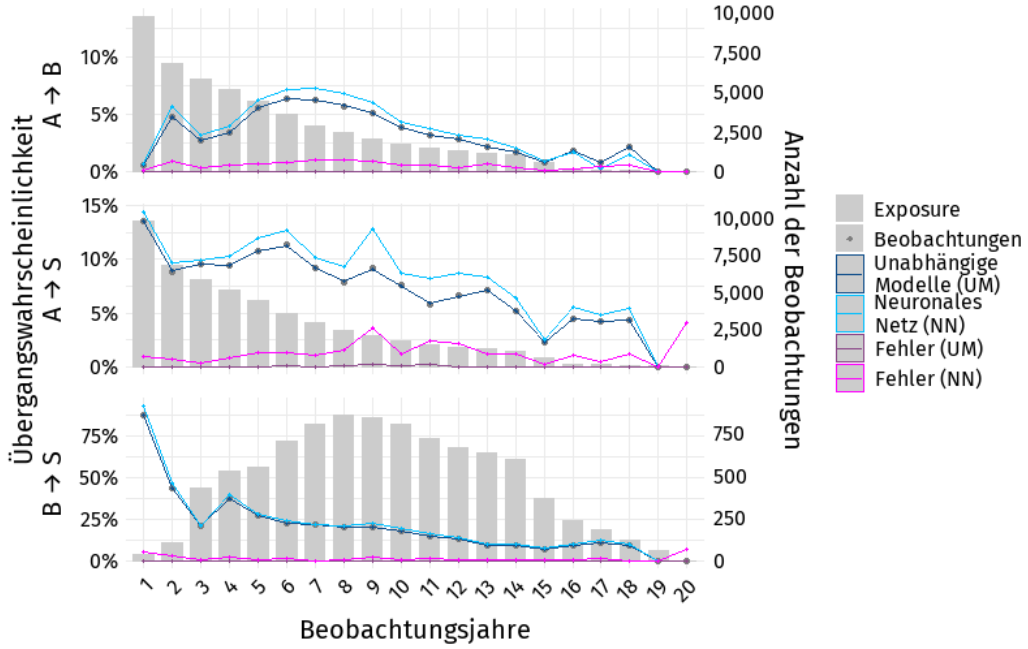


Abbildung A.11.: Prognosen und Vorhersagefehler des neuronalen Netzes für die Anzahl der Beobachtungsjahre im Vergleich mit den unabhängigen Modellen auf den Trainingsdaten

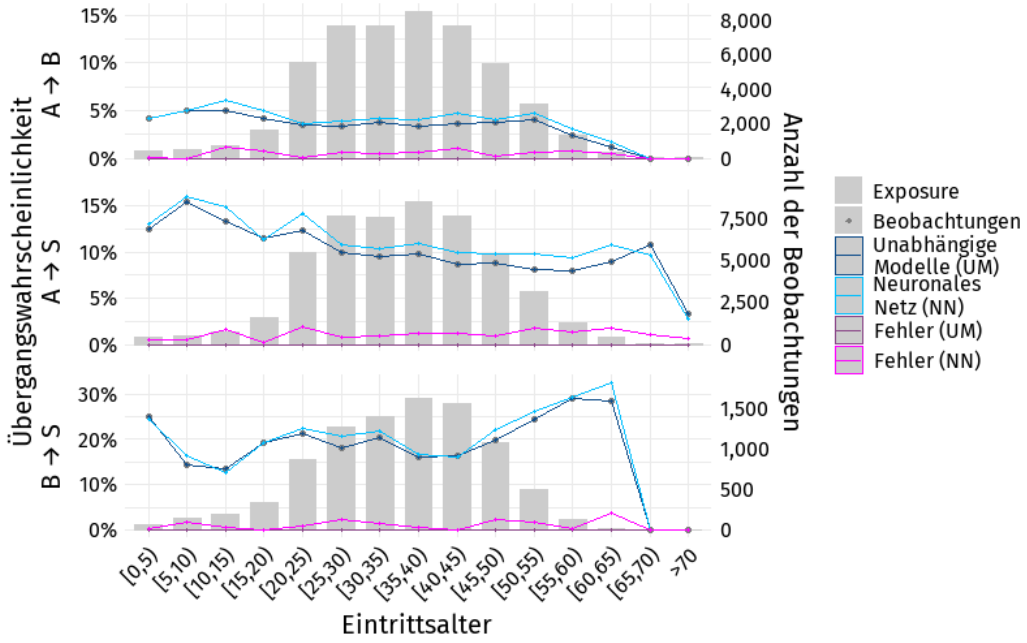


Abbildung A.12.: Prognosen und Vorhersagefehler des neuronalen Netzes für das Eintrittsalter im Vergleich mit den unabhängigen Modellen auf den Trainingsdaten

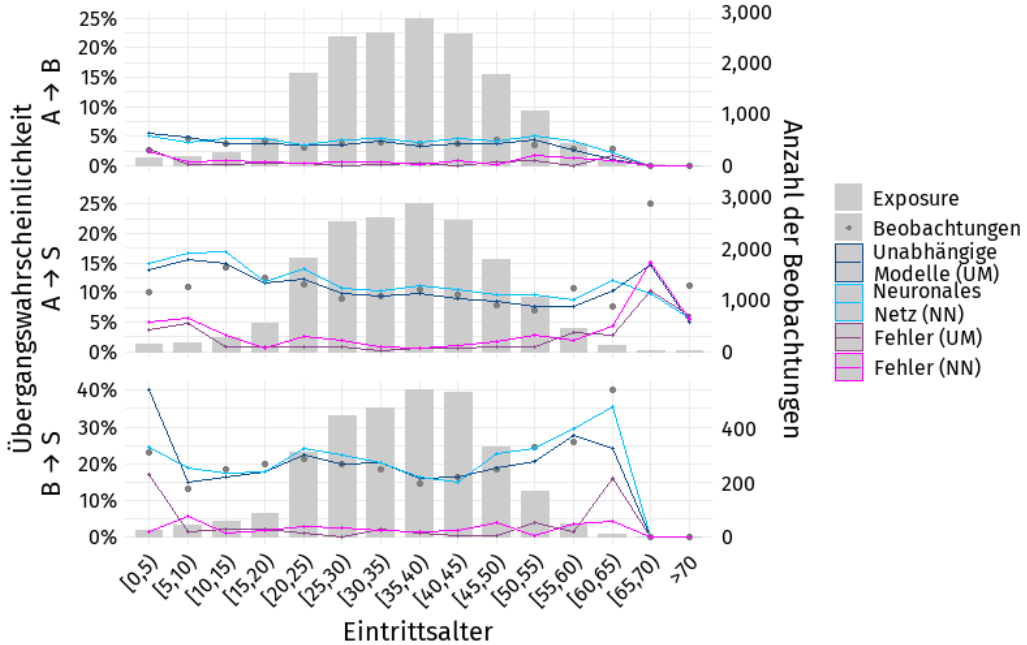


Abbildung A.13.: Prognosen und Vorhersagefehler des neuronalen Netzes für das Eintrittsalter im Vergleich mit den unabhängigen Modellen auf den Testdaten

Literaturverzeichnis

[1] *What is deviance?* https://statisticaloddsandends.wordpress.com/2019/03/27/what-is-deviance/, Abgerufen am 12.04.2022

[2] *Generalized Linear Model (GLM).* https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/glm.html, Abgerufen am 22.03.2022

[3] Agresti, Alan: *Categorical Data Analysis.* Second Edition. John Wiley & Sons, Inc., 2002

[4] Chollet, François ; Allaire, Joseph J.: *Deep Learning with R.* First Edition. Manning Publications, 2018

[5] Dawani, Jay: *Hand-On Mathematics for Deep Learning.* First Edition. Packt Publishing, 2020

[6] Fahrmeir, Ludwig ; Kneib, Thomas ; Lang, Stefan: *Regression - Modelle, Methoden und Anwendungen.* Second Edition. Springer, 2009

[7] Friedman, Jerome ; Hastie, Trevor ; Tibshirani, Robert: *Regularization Paths for Generalized Linear Models via Coordinate Descent.* Vol.33, No.1. Journal of Statistical Software, 2010

[8] Géron, Aurélien: *Hand-On Machine Learning with Scikit-Learn & Tensorflow.* First Edition. O'Reilly Media, Inc., 2017

[9] Glorot, Xavier ; Bordes, Antoine ; Bengio, Yoshua: *Deep Sparse Rectifier Neural Networks.* AISTATS, 2011

[10] Hanewald, Katja ; Li, Han ; Shao, Adam W.: *Modeling multi-state health transitions in China: A generalized linear model with time trends.* Vol.13. Annals of Actuarial Science, 2018

[11] Hastie, Trevor ; Tibshirani, Robert ; Friedman, Jerome: *The Elements of Statistical Learning*. Second Edition. Springer, 2009

[12] Hosmer, David W. ; Lemeshow, Stanley ; Sturdivant, Rodney X.: *Applied Logistic Regression*. Third Edition. Journal of Statistical Software, 2013

[13] Kim, Seung-Jean ; Koh, Kwangmoo ; Boyd, Stephen ; Gorinevsky, Dimitry: *l1 Trend Filtering*. Vol.51, No.2. SIAM Review, 2009

[14] Kingma, Diederik P. ; Ba, Jimmy L.: *Adam: A method for stochastic optimization*. International Conference on Learning Representations (ICLR), 2015

[15] Reck, Lucas ; Schupp, Johannes ; Reuss, Andreas: *Identifying the Determinants of Lapse Rates in Life Insurance: an automated Lasso Approach*. 2022

[16] Renshaw, Albert E. ; Haberman, Steven: *On the graduations associated with a multiple state model for permanent health insurance*. Vol.17. Insurance: Mathematics and Economics, 1995

[17] Tibshirani, Robert: *Regression Shrinkage and Selection via the Lasso*. Vol.58. Journal of the Royal Statistical Society: Series B, 1996

[18] Tibshirani, Robert ; Saunders, Michael ; Rosset, Saharon ; Zhu, Ji ; Knight, Keith: *Sparsity and smoothness via the fused lasso*. Vol.67. Journal of the Royal Statistical Society: Series B, 2005

[19] Tibshirani, Ryan: *Adaptive piecewise polynomial estimation via trend filtering*. Vol.13, No.1. Annals of Actuarial Science, 2014

[20] Vincent, Martin ; Hansen, Niels R.: *Sparse group lasso and high dimensional multinomial classification*. Vol.71. Computational Statistics and Data Analysis, 2013

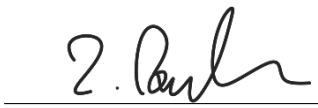
[21] Zhu, Ji ; Hastie, Trevor: *Classification of gene microarrays by penalized logistic regression*. Vol.5, No.3. Biostatistics, 2004

Ehrenwörtliche Erklärung

Ich erkläre hiermit ehrenwörtlich, dass ich die vorliegende Arbeit selbstständig angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht. Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

Ich bin mir bewusst, dass eine unwahre Erklärung rechtliche Folgen haben wird.

Waterloo, den 13.09.2022



(Laura Bader)