

# Modellierung von Übergangswahrscheinlichkeiten für Storno und Beitragsfreistellung

Zusammenfassung der Masterarbeit an der Universität Ulm

Laura Bader

## Motivation

Im Regelfall kalkuliert ein Versicherungsunternehmen die Beiträge unter der Annahme, dass ein Vertrag über die volle Vertragsdauer aktiv und somit beitragspflichtig ist. Jedoch muss für die Kapitalstrom- und Liquiditätsplanung berücksichtigt werden, dass sich Versicherungsnehmer abweichend von dieser Annahme verhalten können. 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. 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 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 sowie die Modellierung der Kündigung zuvor beitragsfrei gestellter Verträge in einem gemeinsamen Modell.

## Modellstrukturen und Methodik

Grundsätzlich wird in dieser Arbeit ein Mehrzustandsmodell mit den möglichen drei Zuständen der Verträge aktiv, beitragsfrei und storniert betrachtet. Das Ziel besteht darin, auf Grundlage der Kovariablen einer Beobachtung im einem Datensatz, d.h. Informationen bezüglich eines Vertrags, möglichst genaue Vorhersagen treffen zu können, wie wahrscheinlich eine Änderung des Zustands für diesen Vertrag im betrachteten Beobachtungszeitraum ist. Dazu wird für jede Beobachtung im Datensatz basierend auf dem Zustand zu Beginn einer Beobachtungsperiode (entweder aktiv oder beitragsfrei) die Wahrscheinlichkeit für die möglichen Zustände am Ende der Periode (aktiv, beitragsfrei oder storniert) geschätzt. Mithilfe der in dieser Arbeit vorgestellten Modellstrukturen werden die Wahrscheinlichkeiten der Zustandsänderungen modelliert. Eine solche Änderung kann entweder die Beitragsfreistellung eines zuvor aktiven Vertrages oder die Kündigung eines zuvor aktiven bzw. beitragsfreigestellten Vertrages bedeuten. Folglich werden durch die vorgestellten Modellstrukturen drei mögliche Zustandswechsel der Verträge modelliert. Dazu werden in der Arbeit maßgeblich logistische Regressionsmodelle verwendet. Diese Modellklasse bietet den Vorteil, dass die Ergebnisse leicht zu interpretieren sind. Als Referenzmodell dient dabei jenes Modell, das dem herkömmlichen Ansatz entspricht, bei dem die Übergangswahrscheinlichkeiten der Zustandswechsel völlig unabhängig voneinander geschätzt werden. So wird für jede mögliche Zustandsänderung eines Vertrages ein separates binomiales logistisches Regressionsmodell angepasst. Dadurch wird beispielweise die Wahrscheinlichkeit der Beitragsfreistellung eines Vertrages völlig unabhängig von der Wahrscheinlichkeit einer Vertragskündigung modelliert. Dieser Ansatz bringt jedoch den Nachteil, dass mögliche gemeinsame Einflussfaktoren für Storno bzw. Beitragsfreistellung von Verträgen nicht berücksichtigt werden können. Aus diesem Grund wird in der Arbeit zusätzlich ein multinomiales logistisches Regressionsmodell vorgestellt, mit welchem die Übergangswahrscheinlichkeiten aller möglicher Zustandsänderungen mit einem gemeinsamen Modell geschätzt werden können. Mithilfe dieser Modellstruktur können gemeinsame Treiber der Zustandsänderungen von Versicherungsverträgen modelliert werden. Eine weitere in dieser Arbeit eingeführte Modellstruktur ermöglicht dies ebenfalls. Dabei werden mehrere binomiale logistische Regressionsmodelle hierarchisch angeordnet, sodass sich die finale Schätzung der Wahrscheinlichkeit des Zustandswechsels eines Vertrages aus den Schätzungen mehrerer binomialer logistischer Regressionsmodelle zusammensetzt. Die

ähnlichen Einflussfaktoren mehrerer Zustandswechsel werden dabei durch einen zweistufigen Prozess herausgearbeitet, bei dem zunächst grundsätzlich die Wahrscheinlichkeit eines Zustandswechsels geschätzt wird, bevor die Wahrscheinlichkeit für den tatsächlichen Endzustand des Vertrages geschätzt wird. Bei allen beschriebenen Modellen wird zudem beleuchtet, wie durch L1-Regularisierung die Komplexität der Modelle reduziert werden kann und wie sich diese Reduktion der Komplexität auf die Güte der vorhergesagten Übergangswahrscheinlichkeiten auswirkt. 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 werden zusätzlich neuronale Netze zur Modellierung der Übergangswahrscheinlichkeiten herangezogen. Sowohl die Modellstrukturen basierend auf logistischer Regression, als auch das neuronale Netz werden auf den Datensatz eines europäischen Lebensversicherers angepasst. Dies dient 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 werden verschiedene Varianten der Modellanpassung, wie beispielsweise mit und ohne Regularisierung, untersucht. Zur Anpassung der Modelle wird der Datensatz in einen zufälligen Trainings- und Testdatensatz aufgeteilt 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. Zur Quantifizierung der Prognosegüte aller in dieser Arbeit diskutierten Modelle wird die Devianz als Gütemaß verwendet. Zudem wird die Anzahl der Parameter (ungleich null) sowie die Laufzeit zur Modellanpassung zwischen den einzelnen Modellen verglichen.

## **Ergebnisse**

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 unbekanntem 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. 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. 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.