Maximilian Euthum, Prof. Dr. Ralf Korn, Prof. Dr. Alfred Müller und Prof. Dr. Matthias Scherer

Data Science: Nun zu 99,5 % ohne Daten?

Sind Sie gerade über unseren bewusst provokanten Titel gestolpert? Leider ist die skizzierte Situation oft näher an der Realität im Bereich Data Science für Aktuar*innen, als wir uns dies vorstellen und eingestehen möchten. Um die Forschung und Lehre in dieser wichtigen Disziplin zu verbessern, arbeitet die DGVFM an einem ambitionierten Datenbankprojekt, für das die Kooperation mit der Versicherungsindustrie extrem wichtig ist und das für beide Seiten lohnend sein kann.

Zielsetzung: gute (empirische) Forschung im Bereich Data Science und praxisnahe Ausbildung

In der versicherungsmathematischen Praxis sind wir es gewohnt, mit großen Datenmengen zu arbeiten. Die vermeintlich simple Frage nach den richtigen Kovariablen in einem GLM zur Schadensvorhersage beschäftigt Aktuar*innen jeder Versicherung. Arbeitsgrundlage sind üblicherweise die Daten aus dem eigenen Versicherungsbestand, der letztendlich modelliert und in die Zukunft projiziert werden soll. Weiten wir nun den Blick auf die universitäre Forschung und Lehre, so wird schnell ersichtlich, dass viele Forschungsgruppen aktuell keinen Zugang zu realistischen Daten besitzen. Als Konsequenz findet dann oft rein theoretische Forschung statt, oder es werden immer dieselben, weil frei verfügbaren Datensätze als Beispiel herangezogen. Als Klassiker seien hier die „Danish Fire Insurance Claims" genannt. Es versteht sich von selbst, dass diese Ausgangssituation kein Katalysator für erfolgreiche Forschung ist. Einwenden könnte man nun, dass es durchaus Kooperationen zwischen Unternehmen und Universitäten gibt, die es einzelnen Wissenschaftler*innen ermöglicht, auf firmeninterne Datensätze zuzugreifen und damit zu arbeiten. Das ist zwar prinzipiell richtig, doch löst es ein grundsätzliches Problem guter empirischer Forschung nicht: Diese muss im Peer Review-Verfahren verifizierbar sein und idealerweise lassen sich die entwickelten Methoden und Hypothesen von anderen Wissenschaftler*innen reproduzieren und verbessern. Aktuell ist dies zu 99,5 % unmöglich, um den Bogen zum plakativen Titel zu spannen.

Es ist selbstredend auch in der Lehre wenig motivierend, wenn zwar die Theorie verschiedener Machine-Learning-Methoden erläutert wird, aber der praktische Einsatz ausbleibt, an dem sich die zahlreichen in der Praxis auftretenden Detailprobleme bei der Umsetzung von ML-Methoden und ihr dafür nötiges Finetuning demonstrieren ließe. Mit aktuellen, anwendungsrelevanten Datensätzen aus dem Bereich der Versicherungswirtschaft könnte auch frühzeitig im Studium ein Interesse an der Aktuarstätigkeit geweckt werden, das über das rein Fachliche der Vorlesung hinausgehen und die Attraktivität der Versicherungsbranche als Arbeitgeber deutlich erhöhen würde.

Der Gordische Knoten

Ganz offensichtlich ist es schon aus (guten!) datenschutzrechtlichen Gründen unmöglich, den Zugang auf tatsächliche Versicherungsbestände freizugeben. Das ist allerdings auch nicht notwendig. Weder sind personen- oder unternehmensbezogene Attribute eines Versicherungsvertrags für die Forschung relevant, noch, von welchem Versicherer ein Datensatz stammt. Idealerweise ist die Datengrundlage für ein Forschungsprojekt auch nicht der Bestand eines einzelnen Versicherers, da dieser aufgrund des spezifischen Profils (Wo regional stark? Ansprache welcher Kundengruppe?) potenziell eine nicht repräsentative Datengrundlage besitzt, sondern ein aggregiertes Portfolio mehrerer Versicherer. Und um weitere Bedenken zu zerstreuen: Es gibt bewährte statistische Verfahren, die aus realen Datenbeständen fiktive Datenbestände kreieren, die identische statistische Eigenschaften haben. Aus unserer Sicht ist der Gordische Knoten daher weniger im Bereich des Datenschutzes zu sehen, als vielmehr in der (bisher) nicht vorhandenen Datenbank und der Bereitschaft, diese zu befüllen. Die Zielsetzung der DGVFM ist es, quasi auf „neutralem Boden" ein Forum für den Datentausch anzubieten.

Vision: das DGVFM-Datenbank-Projekt

Grundsätzlich ist das Ziel die Erstellung einer Datenbank mit Datensätzen aus verschiedenen Gebieten der Versicherungsmathematik. Auf dieser Datengrundlage können dann einerseits neue Modelle und Methoden entwickelt werden, andererseits kann damit empirisch gearbeitet werden, um konkrete aktuarielle Fragestellungen zu lösen. Ein großer Vorteil solch einer Datenbank ist die Möglichkeit der Überprüfung und Reproduktion der veröffentlichten Erkenntnisse. Außerdem kann die Arbeit mit ausgesuchten Datensätzen die Ausbildung wie oben beschrieben maßgeblich verbessern. Nicht zuletzt können die Datensätze auch der Weiterbildung der unternehmensinternen Aktuar*innen dienen, indem z. B. die DAV-Fachgruppe Actuarial Data Science durch ihre Nutzung realitätsnahe Weiterbildungsveranstaltungen anbieten kann.

Auch für die beteiligten Unternehmen ergeben sich zahlreiche Vorteile. Eigene Ansätze und Modelle können anhand eines anderen Datensatzes überprüft werden. Veröffentlichte Erkenntnisse können übernommen und an das eigene Unternehmen angepasst werden. Das Interesse an empirischer Forschung im aktuariellen Bereich steigt und die Ausbildung der Universitätsabsolvent*innen verbessert sich. Dieser kooperative Ansatz ist in anderen Branchen, z. B. dem Bankwesen, durchaus verbreitet und Sammlungen von ausgefallenen Krediten oder operationale Schäden machen Forschungsfragen möglich, die mit den Daten eines einzelnen Unternehmens nur schwer umzusetzen sind.

Nächste Schritte

Neben einem prinzipiellen Interesse seitens der Versicherungsindustrie müssen technische, rechtliche und inhaltliche Fragen geklärt werden. Dazu haben wir eine DGVFM Arbeitsgruppe eingerichtet, in der sich auch Vertreter*innen interessierter Versicherungsunternehmen engagieren können. Die Arbeitsgruppe erarbeitet aktuell ein Konzept, das die folgenden Punkte beinhaltet:

1. Technische Fragen hinsichtlich der Datenbank (Datenformat, wo administriert, technischer Zugriff?)
2. Inhaltliche Fragen (Daten welcher Versicherungssparten sind von Interesse? Was ist ein Minimalkonsens an Kovariablen? Wie erfolgt eine konsistente Aggregation?)
3. Datenschutzrechtliche Fragestellungen und Anonymisierungstechniken
4. Zugangsberechtigung (Welche Unternehmen/Universitäten haben Zugriff?)

Beschreibung bereits existierender, versicherungsmathematischer Datensätze

Um sicherzustellen, dass mit der Erstellung einer eigenen Datenbank keine redundante Aufgabe erledigt wird, wurde in einem vorgelagerten Schritt bereits systematisch nach existierenden Datensätzen gesucht. Diese Arbeit wurde hauptsächlich von Maximilian Euthum geleistet, der dabei von den Mitgliedern Korn, Müller und Scherer des DGVFM-Vorstands unterstützt wurde. Sie wird nachstehend zusammengefasst. Der ausführliche Bericht kann als PDF von www. aktuar.de heruntergeladen werden.

Das entstandene Dokument „Datenbanken mit finanz- und versicherungsmathematischem Bezug: Beschreibung und Zugriff auf kostenfreie Quellen“ ist eine Aufstellung von in der Versicherungsmathematik anwendbaren und öffentlich zugänglichen Daten. Datensätze, Quellen und Verweise sind dabei unmittelbar im PDF verlinkt. Einige Datensätze sind Teil eines Paketes in der Software R. Ist dies der Fall, so wird das Abrufen der Daten mittels R-Konsole erklärt. Manche Datensätze sind eigenständige Tabellen, andere wiederum Ansammlungen von Daten oder Datenbanken. In der Dokumentation enthalten sind spartenübergreifende Datensätze, von Mortalitätsdaten über klassische Kfz-Schadenfälle bis hin zu Cyberrisiken und Strommarktdaten. Selbstredend, dass diese Sammlung keinen Anspruch auf Vollständigkeit erhebt und die Autoren über Hinweise für Ergänzungen dankbar sind.

Datensätzen auf Schadensbasis sind allgemeine Kennzahlen von Versicherungsunternehmen gegenübergestellt. Gewisse Datensätze eignen sich zur versicherungsmathematischen Modellierung, andere liefern interessante Informationen über den Versicherungsmarkt. Einige Datensätze beziehen sich auf den deutschen Versicherungsmarkt, andere betreffen beispielsweise den US-Markt, weitere Statistiken wiederum sind global. In das Dokument aufgenommen wurden nur öffentlich zugängliche, kostenfreie Datensätze oder Datenbanken, nur selten ist eine Registrierung notwendig. Oft handelt es sich um aggregierte Daten, die nicht mit der Granularität der Bestandsdaten aus Versicherungsunternehmen konkurrieren können.

Die Beschreibung der Datensätze unterscheidet sich von Fall zu Fall. Gemeinsam ist jedoch ein Steckbrief mit den Angaben zu:

- Art: knappe Beschreibung des Datensatzes,
- Quelle: Ursprung des Datensatzes, Link zu einer Webseite oder einem R-Paket,
- Datenformat: Exceldaten, csv, html usw.,
- Sprache: Englisch oder Deutsch,
- Zugriff: Registrierung notwendig – ja/nein,
- Dateiumfang: Größe in Datenpunkten/benötigter Speicherplatz,
- Zeithorizont: zeitliche Abdeckung der Daten.

Auf den Steckbrief folgt jeweils eine detailliertere Erläuterung des Datensatzes mit einem Auszug der Daten. Wenn vorhanden, werden auch einschlägige Publikationen genannt, die den entsprechenden Datensatz beinhalten oder verwenden.

Aus der Aufstellung wird deutlich, dass zwar einige Datensätze zu Forschungszwecken frei verfügbar sind, aber Kernaufgaben aus der aktuariellen Praxis (Bewertung, Risikomessung) damit nicht zufriedenstellend umgesetzt werden können. Auch ist bei vielen internationalen Datensätzen unklar, ob und wie sich erzielte Erkenntnisse auf die Situation in Deutschland übertragen lassen.

Wir sind der festen Überzeugung, dass wir bei Erfolg dieser Initiative sowohl auf dem Forschungs- als auch auf dem Ausbildungssektor einen Mehrwert für die deutsche Versicherungswirtschaft erzielen können, allerdings auch nur dann, wenn wir von der Branche den gewünschten Input an Daten erhalten, um Actuarial Data Science demnächst mit Daten betreiben zu können.