Entwicklung eines datenbasierten Vorhersagemodells für die erwarteten Schäden in der privaten Krankenversicherung

Zusammenfassung der Bachelorarbeit an der Universität Ulm

David Neuhäusler

Motivation und Daten

Ein wesentlicher Teil der Kosten eines privaten Krankenversicherers sind Versicherungsleistungen für Versicherungsverträge. In Deutschland erstatteten Krankenversicherer im Jahr 2019 im Mittel Versicherungsleistungen von 3.264 Euro pro Versicherungsvertrag. Deshalb ist es für einen privaten Krankenversicherer von Interesse, ein möglichst gutes Modell zu entwickeln, das die erwarteten Versicherungsleistungen in der Zukunft modellieren kann. Die Bachelorarbeit zielt darauf ab, ein datenbasiertes Einjahresmodell für die erwarteten Schäden in der privaten Krankenvollversicherung zu entwickeln. Das heißt, es sollen die erwarteten Schäden im kommenden Geschäftsjahr basierend auf historischen Daten modelliert werden.

Gegeben ist ein Datensatz mit Vertragsdaten zu einem Tarif der privaten Krankenvollversicherung. Der Datensatz basiert auf dem Bestand eines deutschen privaten Krankenversicherers und beinhaltet 25.000 Datensätze mit 27 Merkmalen je Vertrag. Neben Merkmalen zum Alter, Geschlecht und Beruf der Versicherten enthält der Datensatz quantitative Informationen zu Versicherungsprämien, Erstattungen und Selbstbehalten für vier aufeinanderfolgende Geschäftsjahre. Davon sind acht Merkmale als Quotient anderer Merkmale berechnet.

Die Modellidee ist es, ein Generalisiertes Lineares Modell (GLM) für die erwartete Schadenhöhe im kommenden Geschäftsjahr zu trainieren und anschließend mittels Merkmaltransformationen (Feature Engineering) und Lasso-Regularisierung zu verbessern. Datenanalysen und Modellierung werden in Python umgesetzt.

Statistische Methoden

Im ersten Teil der Arbeit wird die theoretische Herleitung zu den eingesetzten statistischen Methoden beschrieben. Im Kontext von Exponentialfamilien und Maximum-Likelihood Maximierung werden Generalisierte Lineare Modelle sowie Lasso-Regularisierung formal definiert. Ausgehend von diesen Definitionen und Beispielen werden der Fisher-Score Algorithmus als Lösungsverfahren für ein GLM vorgestellt und Kennzahlen zum Vergleich von GLMs erklärt.

Analyse der Zielgröße und Modellidee

Zunächst wird die Zielgröße „Schaden im kommenden Geschäftsjahr“ analysiert. Die Analyse fokussiert sich auf positive Merkmalausprägungen der Zielgröße, weil 47,2% der Datensätze keine Schäden im kommenden Geschäftsjahr aufweisen. Da eine Beschreibung des Merkmals anhand einer Gammaverteilung plausibel erscheint, werden mittels einer Maximum-Likelihood Maximierung mögliche Skalen- und Formparameter der Verteilung bestimmt. Ein QQ-Plot sowie Mittelwert-Varianz Plots gruppiert nach Alter der versicherten Person und Laufzeit des Vertrags unterstützen den Ansatz, die Schadenhöhen im kommenden Geschäftsjahr mittels einer Gammaverteilung zu charakterisieren.

So wird zur Modellbildung ein zweischrittiges Vorgehen gewählt. Ein erstes Teilmodell soll dazu dienen, die Wahrscheinlichkeit eines positiven Schadens im kommenden Geschäftsjahr zu modellieren. Als Modellansatz wird eine Logistische Regression verwendet, die ein Spezialfall eines GLM für binäre Zielvariablen ist. Ein zweites Teilmodell soll dazu dienen, die erwarteten Schadenhöhen im kommenden Geschäftsjahr zu modellieren unter der Bedingung, dass sie positiv sind. Dazu wird ein Gamma GLM genutzt. Anschließend werden die beiden Teilmodelle im Sinne eines multiplikativen Gesamtmodells verknüpft. Für jeden Eingabedatensatz liefert das Gesamtmodell eine Aussage über Schadenwahrscheinlichkeit und erwartete Schadenhöhe.

Modellbildung

Zur Modellierung der Schadenwahrscheinlichkeit im kommenden Geschäftsjahr wird eine Logistische Regression verwendet. Zunächst

werden die Regressionskoeffizienten der Logistischen Regression zu den 25 gegebenen Prädikatorvariablen im Trainingsdatensatz bestimmt. Um dieses *naive Modell* zu verbessern, werden Transformationen für Merkmale des gegebenen Datensatzes durchgeführt, das sogenannte Feature Engineering. Das Hinzufügen von transformierten Merkmalen zum Datensatz ermöglicht eine bessere Anpassung des Modells an die Trainingsdaten. Dies wird durch Randverteilungsplots veranschaulicht, die den modellierten und tatsächlichen Einfluss einzelner Merkmale auf die Zielgröße miteinander vergleichen. Um Overfitting zu vermeiden, werden zwei unterschiedliche Ansätze zur Begrenzung der Anzahl der Prädikatorvariablen angewendet und deren Vorgehen sowie Ergebnisse danach miteinander vergleichen.

Im ersten Ansatz wird ausgehend von den Merkmalen im *naiven Modell* je eine weitere Modellvariante mit je einer einzelnen zusätzlichen Merkmaltransformation gefittet. Dabei werden Polynomtransformation zweiten (poly2) und dritten Grades (poly3), Logarithmus-Transformation (log) sowie Aufteilen eines Merkmals in zwei Merkmale anhand eines Schwellenwerts (split) angewendet. Somit besitzen diese neuen Modellvarianten stets 26 beziehungsweise 27 Prädikatorvariablen. Pro Merkmal wird nur diejenige Transformation gespeichert, dessen Modellvariante im Vergleich das größte Pseudo R^2 besitzt. Anschließend wird der zugrundeliegende Datensatz des *naiven Modells* für jedes Merkmal um diese gespeicherten Merkmaltransformationen erweitert. Der Datensatz besitzt nun 42 Merkmale. Um die Anzahl der Prädikatorvariablen wieder zu reduzieren, werden alle Merkmale entfernt, deren standardisierte Regressionskoeffizienten infolge eines Wald-Tests zu einem Signifikanzniveau von 5% als einflusslos erachtet werden.

Im zweiten Ansatz wird für jedes quantitative Merkmal im zugrundeliegenden Datensatz des *naiven Modells* eine Polynomtransformation fünften Grades und eine Logarithmus-Transformation durchgeführt. Auf diesen Datensatz mit 135 Merkmalen wird eine Logistische Regression mit Lasso-Regularisierung angewendet, wobei der Hyperparameter für den Einfluss des Penalisierungsterms mittels Kreuzvalidierung auf den Testdaten bestimmt wird. Anschließend werden alle Merkmale entfernt, deren

standardisierte Regressionskoeffizienten einen Absolutwert von kleiner als 0.03 besitzen.

Die Modellierung der erwarteten Schadenhöhe im kommenden Geschäftsjahr mit einem Gamma GLM erflogt analog zur Methodik, wie für die Schadenwahrscheinlichkeit dargestellt.

Modellergebnisse und Fazit

Die zwei Ansätze zur Auswahl von Merkmaltransformationen werden für beide Zielgrößen miteinander verglichen und auf den Testdaten evaluiert. Die modellierten und tatsächlichen Schadenwahrscheinlichkeiten werden mittels Randverteilungsplots gegenübergestellt. Es ist keine systematische Abweichung zwischen beobachteten und modellierten Schadenwahrscheinlichkeiten zu beobachten. Die Modellresultate beider Ansätze zur Schadenhöhe weisen eine relative Abweichung von weniger als 0.23% auf in Bezug auf die beobachtete tatsächliche Schadenhöhe im Testdatensatz auf.

Die Arbeit schließt mit einem Vorschlag für ein Gesamtmodell, das für einen Eingabedatensatz eine Schadenwahrscheinlichkeit sowie eine Schadenhöhe für das kommende Geschäftsjahr modelliert. Das Produkt der Schätzungen aus beiden Teilmodellen ergibt die final geschätzte Schadenhöhe. Die Modelle beider Ansätze werden gegenübergestellt und analysiert. Die Modellvariante mit Lasso-Regularisierung weist unter allen Modellansätzen den niedrigsten mittleren quadrierten Fehler auf den Testdaten auf.

Um die Güte der entwickelten Modelle einzuordnen, sollten diese in einem nächsten Schritt mit bereits in der Praxis verwendeten Modellen verglichen werden, was die Bachelorarbeit allerdings nicht mehr beinhaltet. Außerdem könnten mit den Daten andere Machine Learning Modellansätze wie Regressionsbäume, neuronale Netze oder Ensemblemethoden ausgearbeitet werden.