Christian Eckert, Daniela Giesinger, Felix Müller und Antonia Schöning

Machine Learning in der Berufsunfähigkeitsversicherung?

Eine Analyse von Risikofaktoren

1. Einleitung und Hintergrund

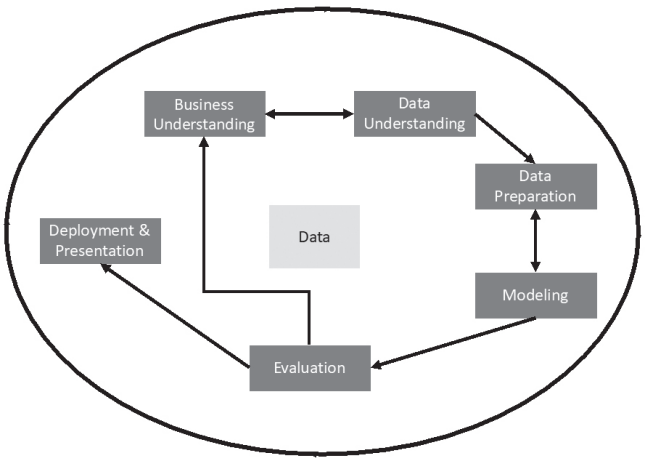
Die fortschreitende digitale Transformation führt dazu, dass auch in Versicherungsunternehmen Data Science eine immer größere Rolle spielt. Da Data Science eine interdisziplinäre Wissenschaft zwischen Mathematik/Statistik, Informatik und jeweiliger Fachexpertise ist, sind Aktuar*innen in Versicherungsunternehmen für entsprechende Aufgaben besonders gut geeignet. Auch in ihren bisherigen Tätigkeiten haben sich Aktuar*innen in einem interdisziplinären Themenfeld bewegt und besitzen bereits die benötigte Fachexpertise. Die Bedeutung von Data Science für Aktuar*innen spiegelt sich auch darin wider, dass die DAV eine Zusatzqualifikation zum „Certified Actuarial Data Scientist (CADS)" anbietet. Zudem organisiert die DAV seit 2020 jährlich eine Data Science Challenge, die Aktuar*innen dabei unterstützt, sich vermehrt mit datenwissenschaftlichen Fragestellungen und maschinellem Lernen zu beschäftigen. Im Jahr 2021 ist die Einsendung von Unterlagen bis einschließlich 31.08.2021 möglich. Detaillierte Informationen finden Sie auf der DAV-Website.

Dieser Artikel basiert auf unseren Analysen im Zusammenhang mit der Data Science Challenge 2020, die im Rahmen unserer Tätigkeit bei der Nürnberger Versicherung entstanden sind. Unser Vorgehen erläutern wir im Folgenden anhand des Cross Industry Standard Process for Data Mining (CRISP-DM) (siehe Abbildung 1). Dieser ist ein typischer Workflow für Data Science/KI-Projekte. CRISP-DM ist ein Regelkreis der aus den Phasen Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation und Deployment & Presentation besteht.

2. Business Understanding

Wird ein*e Versicherungsnehmer*in einer Berufsunfähigkeitsversicherung berufsunfähig, entstehen dem Versicherungsunternehmen hohe Kosten. Die Selektion und Bewertung der Risiken bei Vertragsabschluss, aber auch Präventionsmaßnahmen spielen daher eine große Rolle. In den letzten Jahren sind immer häufiger psychische Krankheiten der Grund für Berufsunfähigkeit. Wir gehen daher der Forschungsfrage nach, welche Rahmenbedingungen psychische Krankheiten begünstigen, die dann zu einer Berufsunfähigkeit führen? D.h., was sind relevante (und auch: was sind keine relevanten) Einflussfaktoren auf die Berufsunfähigkeit und welcher Zusammenhang besteht? Diese Fragestellungen sind für Versicherungsunternehmen aus verschiedener Hinsicht interessant. Zum einen in der Tarifierung. Einflussfaktoren/Risikofaktoren sollten Einfluss auf die Prämie haben. Gleichzeitig genügt es aber, sich auf die wesentlichen und relevanten Einflussfaktoren zu fokussieren, was gegebenenfalls eine Reduktion von Fragen im Antragsprozess ermöglicht. Dies führt zu einer verbesserten Customer-Experience. Des Weiteren können die Ergebnisse auch relevanten Mehrwert hinsichtlich potenzieller Präventionsmaßnahmen bieten. Präventionsmaßnahmen, die Risikofaktoren positiv beeinflussen, sind besonders erfolgversprechend. Aber auch für den Einzelnen sind unsere Ergebnisse bedeutend. Niemand möchte psychisch erkranken und berufsunfähig werden.

Abbildung 1:
Der Regelkreis CRISP-DM
Quelle: Eigene Darstellung in Anlehnung an Shearer (2000)



3. Data Understanding

Um einen ersten Einblick zu erhalten, haben wir uns für öffentliche Daten vom National Center for Health Statistics (USA) entschieden. Wir haben bewusst externe Daten bezogen, um auch Faktoren berücksichtigen zu können, die von Versicherungsunternehmen gegebenenfalls bisher nicht erhoben werden. Das National Health and Nutrition Examination Survey (NHANES) stellt ein für die USA repräsentatives Sample im Zeitraum von 1999 bis 2018

dar. Dort werden alle zwei Jahre äußerst umfangreiche Informationen über den Gesundheitszustand (insbesondere über die Psyche) und die Lebensumstände sowie das Arbeitsverhältnis inklusive Berufsunfähigkeit von rund 5000 Befragten erhoben.

Wir verwenden die Befragungsergebnisse der Jahre 2005 bis 2016.[1] Da die Fragebögen alle paar Jahre angepasst werden, haben wir uns bei den verwendeten Daten auf sechs Kapitel der Befragung beschränkt, die konsistente Fragen über den betrachteten Zeitraum beinhalten. Die ausgewählten Kapitel sind „Blutdruck und Cholesterin", „Demografische Variablen", „Körperliche Verfassung", „Schlafstörungen", „Rauchverhalten" und „Gewichtshistorie" (Kapitel BPQ, DEMO, PFQ, SLQ, SMQ, WHQ). In den sechs Kapiteln beschränken wir uns auf eine Auswahl von Fragen, die sich auf die Vergangenheit beziehen oder nicht von der Tatsache beeinflusst werden, ob aktuell eine Berufsunfähigkeit vorliegt. Zum Beispiel wird die Frage zum aktuellen Einkommen vermutlich maßgeblich davon beeinflusst, ob der Befragte einer regulären Arbeit nachgehen kann oder nicht. Dieses Vorgehen ist notwendig, um eine ähnliche Situation, wie bei einem Abschluss einer BU-Versicherung zu schaffen, bei dem die relevanten Fragen zum Anfang des Versicherungsverhältnisses und zeitlich vor dem möglichen Eintritt der Berufsunfähigkeit gestellt werden.

Unsere Forschungsfrage wollen wir beantworten, indem wir die in den Daten vorhandene abhängige Variable „Berufsunfähigkeit" (im Original „Limitations keeping you from working") mit den Ausprägungen „ja" und „nein" durch in den Daten vorhandene unabhängige Variable erklären. Genauer wird hier im Fragebogen die Frage gestellt, ob eine Langzeiterkrankung in Form von einer körperlichen, geistigen oder emotionalen Beeinträchtigung die Ausübung einer Arbeit verhindert. Insbesondere sollen keine vorrübergehenden Zustände, wie Erkältungen oder eine Schwangerschaft, betrachtet werden. Die Angaben werden von den Befragten gemacht und stellen eine Selbsteinschätzung dar. Es muss keine ärztlich festgestellte Berufsunfähigkeit vorliegen, wie sie für eine deutsche Versicherung notwendig wäre. Daher geben unsere Ergebnisse nur Hinweise auf mögliche Risikofaktoren, sind aber nicht uneingeschränkt auf einen Versicherungsbestand übertragbar.

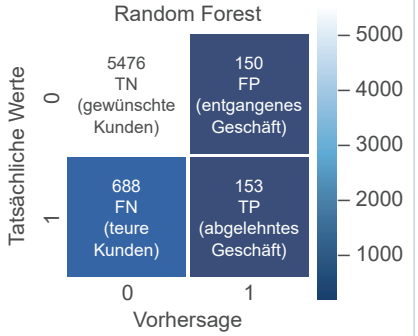
4. Data Preparation

Bevor ein Modell trainiert werden kann, müssen die Daten gründlich aufbereitet werden. Ein sehr wichtiger Schritt ist dabei der Umgang mit fehlenden Daten. Das Ziel sollte sein, möglichst wenige Variablen zu löschen, um die Modellkomplexität nicht zu stark zu reduzieren. Ebenso sollten nicht zu viele Datensätze ausgeschlossen werden, um eine breite Datenbasis zu erhalten. In unserem Fall weisen einige Variablen einen hohen Anteil an fehlenden Werten auf. Die Gründe hierfür sind bei den verschiedenen Variablen unterschiedlich, weswegen wir entsprechend unterschiedliche Lösungsansätze gewählt haben. Insgesamt lassen sich vier Fälle unterscheiden. Vor der Aufbereitung betrachten wir 34 Variablen und 34.163 Datensätze in denen die Zielvariable beantwortet wurde.

Es gibt systematisch fehlende Datensätze, die aufgrund der Art der Datenerhebung per dynamischem Fragebogen fehlen. Bei dynamischen Fragebögen werden Fragen nicht gestellt, wenn es für die befragte Person nur eine mögliche Antwort gibt oder die Frage im Kontext keinen Sinn macht. Übersprungene Fragen werden allerdings als fehlende Werte im Datensatz angegeben. Hierfür gibt es einige Beispiele im Bereich Medikation. Einer Person, bei der noch nie erhöhter Blutdruck diagnostiziert wurde, wird die Frage, ob sie Medikamente gegen erhöhten Blutdruck nimmt, nicht gestellt. In diesem Fall können wir fehlende Datensätze durch logische Schlussfolgerungen ergänzen.

Wenn Daten nicht systematisch fehlen, können numerische Werte beispielsweise durch den Mittelwert oder den Median der gesamten Stichprobe ersetzt werden. Dieses

Abbildung 2:
Confusion Matrix des Random Forest

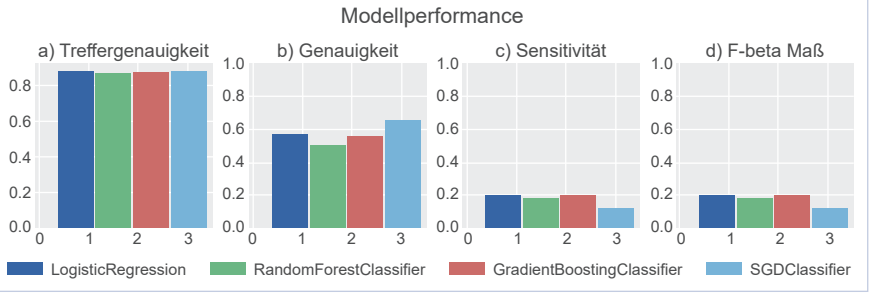


Verfahren haben wir unter anderem im Themenbereich Körpergewicht angewandt. Zum Beispiel wurde die Frage zum aktuellen Körpergewicht sehr häufig beantwortet, die Frage zum Gewicht im Alter von 25 Jahren jedoch seltener. Daher haben wir den mittleren anteiligen Gewichtsunterschied zwischen dem aktuellen Gewicht und dem Gewicht mit 25 Jahren berechnet und diesen Wert wiederum für eine Schätzung fehlender Werte verwendet.

Variablen mit fehlenden Werten, bei denen die ersten beiden Lösungsansätze nicht anwendbar sind, können durch zwei weitere Möglichkeiten bereinigt werden. Entweder kann die Variable entfernt werden oder der Datensatz der Person mit einem fehlenden Wert wird ausgeschlossen. Hierbei haben wir folgende Abwägungen in unsere Entscheidung einfließen lassen: Modelle mit vielen Variablen benötigen beim Modelltraining eine hohe Anzahl von Datensätzen, weswegen Variablen mit vielen fehlenden Werten gelöscht und mit wenigen fehlenden Werten behalten wurden. Die Grenze ist nicht eindeutig definierbar, aber man kann sie etwas verschieben, wenn in der Modelltrainingsphase Probleme auftreten. Letztendlich wollen wir fehlende Werte vermeiden, da nicht

[1] Die Fragen aus den Fragebögen der Jahre 1999–2004 unterscheiden sich von den späteren Fragebögen. Daten der Jahre 2017–2018 lagen zum Zeitpunkt unserer Analysen (August 2020) noch nicht vollständig vor.

Abbildung 3:
Kenngrößen der verschiedenen Modelle auf dem Test-Sample im Vergleich



alle betrachteten Machine-Learning-Modelle mit fehlenden Werten umgehen können und wir für jedes Modell die gleiche Datengrundlage verwenden wollen. Daher haben wir im letzten Schritt alle 12.608 Datensätze von Teilnehmenden, die auch nach der Datenbereinigung noch fehlende Werte hatten, gelöscht. Übrig bleibt ein Datensatz mit 21.555 Datensätzen.

Im ursprüngliche Datensatz waren 14,6 % der Befragten BU, im reduzierten Datensatz noch 13,3 %.

Um auch kategoriale Variablen für unsere Analysen nutzbar zu machen, verwenden wir die sogenannte *One-Hot-Kodierung*. Dabei bekommt jede mögliche Ausprägung einer Variablen eine eigene Variable mit der Kodierung „0: trifft nicht zu" und „1: trifft zu". Angewendet wurde das Vorgehen unter anderem beim Bildungsstatus (fünf Level von *Schulabgänger vor der neunten Klasse* bis *Collegeabsolvent*), beim Beziehungsstatus, aber auch bei der Frage, in welchem Alter mit dem Rauchen begonnen wurde. Der ursprüngliche Datensatz hat bei Nichtrauchern das Alter 0 vermerkt, das so im Zusammenhang mit Einstiegsaltern nicht weiterverarbeitet werden kann. Über Quantile haben wir daher Alter in Intervalle zusammengefasst und dann das Einstiegsalter One-Hot-kodiert, d. h. mit einer Eins das entsprechende Intervall markiert. Nichtraucher haben dabei in allen Intervallen des Einstiegsalters eine Null vermerkt.

Ebenfalls problematisch sind Korrelationen zwischen Variablen, weil das Modell dadurch eine große Komplexität erhält, aber keine bessere Vorhersage liefert und der Lernprozess deutlich erschwert wird. Zum Beispiel sind die Variablen *DoctorHighBloodPressure2* (Zweimalige Diagnose Bluthochdruck) und *DoctorHighBloodPressure* (allgemeine Diagnose Bluthochdruck) stark korreliert, sodass die zweimalige Diagnose als Variable aus dem Datensatz entfernt wird. Einige weitere korrelierende Variablen wurden ebenfalls aus dem Datensatz entfernt.

Die letzten beiden Schritte in der Datenaufbereitung umfassen die Normalisierung der Daten und die Aufteilung des Datensatzes. Da die Wertebereiche unserer Variablen sehr heterogen sind und Algorithmen darauf teilweise sehr empfindlich reagieren, normalisieren wir unsere Daten. Dadurch erzielen wir eine bessere Vergleichbarkeit zwischen Variablen. Wir verwenden dabei die Min-Max-Normalisierung, bei der der Wertebereich einer jeden Variablen auf das Intervall [0,1] umskaliert wird. Um später überprüfen zu können, ob unser Modell auch außerhalb der für die Modellierung verwendeten Daten gut geeignet ist, teilen wir den Datensatz auf. Wir wählen hier eine Aufteilung in Trainings- und Testdatensatz im Verhältnis 70:30.

5. Modeling

Im Folgenden werden vier sehr bekannte Modelle für die Beschreibung von binären Klassifikationsproblemen miteinander verglichen. Die ausgewählten Modelle sind jeweils hinsichtlich Erklärbarkeit, Performanz und Handlichkeit sowie Laufzeit im Vorteil gegenüber vielen anderen möglichen Modellierungsansätzen. Für alle Modelle wurde die Implementierung von scikit-learn (Pedregosa (2011)) verwendet, auf der Homepage des Projekts finden sich auch nähere Modellbeschreibungen.

Die **Logistische Regression** ist ein Standardmodell des maschinellen Lernens, das die Beziehung zwischen abhängigen binären Variablen und unabhängigen (nicht notwendigerweise binären) Variablen mithilfe einer logistischen Zielfunktion modelliert. Die unbekannten Zielfunktionsparameter werden aus den Trainingsdaten geschätzt.[2]

Als zweites Modell betrachten wir den auf Entscheidungsbäumen basierenden **Random Forest**. Hierbei werden mehrere unkorrelierte Entscheidungsbäume kombiniert, um eine Klassifikation vorzunehmen. Als ein Modell des überwachten Lernens werden die einzelnen Entscheidungsbäume parallel auf Teilmengen der Trainingsmenge trainiert. Bei der Bewertung eines neuen Datensatzes werden dann alle Bäume individuell ausgewertet und die am häufigsten angenommene Klasse wird als Ergebnis angenommen. Das Modell ist insbesondere sehr schnell trainiert und auch die Auswertung kann parallelisiert werden.

Ebenfalls auf Entscheidungsbäumen aufbauend ist das dritte Modell, das **Gradient Tree Boosting**. Im Gegensatz zu Random Forest werden allerdings keine individuellen Entscheidungsbäume parallel trainiert, sondern von einem Entscheidungsbaum ausgehend werden in einer Greedy-Prozedur weitere Entscheidungsbäume als zusätzliche Basisfunktionen schrittweise dem ersten Baum hinzugefügt, um die Gesamtvorhersage zu verbessern. In jeder Iteration wird der neu hinzugefügte Baum dann auf den noch fehlerhaften Daten des Hauptbaums trainiert.

Als viertes Modell verwenden wir das sogenannte **Stochastic Gradient**

[2] In der Implementierung von scikit-learn ist die „logistische Regression" ein regularisiertes Regressionsverfahren.

Descent (mit Support Vector Machine), eine stochastische Version des Gradientenverfahrens. Iterativ wird die Zielfunktion verändert, wobei die normalerweise verwendete Richtung des steilsten Abstiegs durch den Gradienten an einem zufällig gewählten Punkt approximiert wird.

Alle auf der Trainingsmenge trainierten Modelle werden zur Evaluierung auf der Testmenge ausgewertet. Für das binäre Klassifikationsproblem (Status BU oder nicht BU/Gesund) bieten sich zur Beurteilung der Modelle quantitative Maße auf Basis von relativen Häufigkeiten – das Einordnen eines Datensatzes in eine richtige oder falsche Klasse – an.

Im Folgenden bezeichnet ein sogenanntes positives Ergebnis den Fall, dass Berufsunfähigkeit vorhergesagt wird. Ein negatives Ergebnis ist der Status Gesund. Es ergeben sich bei jedem Modell für die Klassifikation vier Möglichkeiten:

- Korrekt Positiv (True Positive (TP)): Ein BU-Fall, der vom Modell auch als ein BU-Fall prognostiziert wird.
- Falsch Positiv (False Positive (FP)): Ein Gesunder, der vom Modell als BU-Fall eingestuft wird.
- Korrekt Negativ (True Negative (TN)): Ein Gesunder, der auch vom Modell als gesund erkannt wird.
- Falsch Negativ (False Negative (FN)): Ein BU-Fall, der vom Modell allerdings als gesund eingestuft wird.

Formel 1

$$F_\beta = (1+\beta^2) \cdot \frac{Genauigkeit \cdot Sensitivität}{(\beta^2 \cdot Genauigkeit) + Sensitivität}$$

Auf Basis der Ereignisse können folgende Kenngrößen betrachtet werden:

- Die **Treffergenauigkeit** (Accuracy) beschreibt den Anteil der richtig klassifizierten Fälle im Modell.
- Die **Genauigkeit** (Precision) setzt den Anteil der korrekt vorhergesagten BU-Fälle ins Verhältnis zu allen vorhergesagten BU-Fällen.
- Die **Sensitivität** (Recall) setzt den Anteil der korrekt vorhergesagten BU-Fälle ins Verhältnis zu allen existierenden BU-Fällen.
- Die **Ausfallrate** (Fall-out) ist das Verhältnis zwischen fälschlicherweise als BU klassifizierten und den tatsächlich Gesunden.
- Das **F-beta-Maß** (F-Score) kombiniert als Schätzer die beiden Kenngrößen Genauigkeit und Sensitivität. Durch die Wahl von β lässt sich steuern, welche Größe wie viel Einfluss erhält.

Um die Leistungsfähigkeit der Modelle besser einordnen zu können, führen wir einen naiven Schätzer als Benchmark ein. Als naiven Schätzer wählen wir das sehr einfache Modell, das immer „Gesund“ vorhersagt, unabhängig von der Eingabe. Die Anzahl der richtig als BU eingestuften Fälle beläuft sich bei unserem naiven Schätzer daher auf null, weil alle Fälle als nicht BU eingestuft werden (TP = 0). Außerdem gibt es im Modell keine Nicht-BU-Fälle, die als BU eingestuft wurden (FP = 0). Die Anzahl an Gesunden, die korrekt als nicht BU klassifiziert werden, entspricht dann allen tatsächlich Gesunden im Datensatz. Die Anzahl der fälschlicherweise als gesund eingestuften entspricht der Anzahl an BU-Fällen im Datensatz.

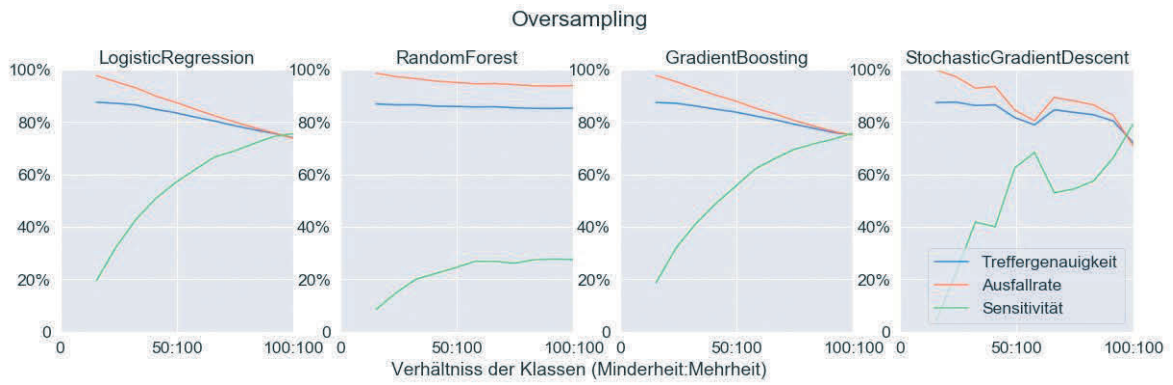
Bereits mit dem naiven Schätzer erreichen wir eine Treffergenauigkeit von etwa 87 %. Wenn die Testmenge einen Versicherungsbestand darstellen und das Modell bei der Antragsannahme eine Rolle spielen würde, würde das insbesondere bedeuten, dass

- von 6467 Anträgen
- keiner abgelehnt und
- alle angenommen und davon
- 841 Kunden BU werden.

Das wäre eine BU-Quote von 13 %.

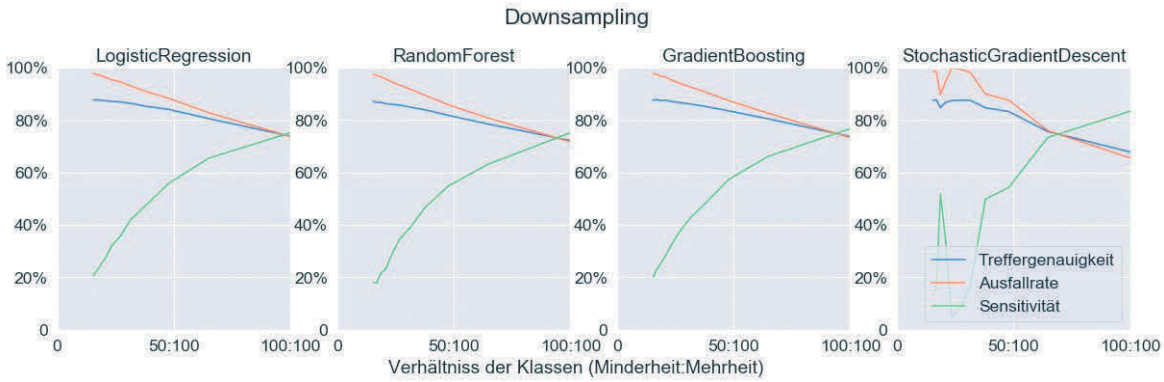
Die Annahme der Anträge könnte mit einem Modell nun so verbessert werden, dass die BU-Quote verringert wird. Dabei würden voraussichtliche BU-Kandidaten gar nicht erst in den Bestand aufgenommen. Die hohe Treffergenauigkeit des naiven Schätzers reflektiert die Tatsache, dass fast 87 % der Personen in der Stichprobe nicht berufsunfähig sind.

Abbildung 4:
Auswirkungen des Oversampling auf die Kenngrößen Treffergenauigkeit, Ausfallrate und Sensitivität.



Auf der x-Achse sind verschiedene Verhältnisse der kleineren Klasse (BU-Fälle) zu der größeren Klasse (Gesunde) dargestellt bis zu einem Verhältnis von 100:100.

Abbildung 5:
Auswirkungen des Downsampling auf die Kenngrößen Treffergenauigkeit, Ausfallrate und Sensitivität.



Auf der x-Achse sind verschiedene Verhältnisse der kleineren Klasse (BU-Fälle) zu der größeren Klasse (Gesunde) dargestellt bis zu einem Verhältnis von 100:100.

Das heißt, dass die zugrunde liegende Datenmenge nicht ausbalanciert ist. Berufsunfähigkeit ist ein seltenes Ereignis, das aber möglichst zuverlässig vorhergesagt werden soll. Jeder Schadensfall in der Berufsunfähigkeitsversicherung ist für ein Versicherungsunternehmen teuer. Unabhängig davon, ob ein Modell zur Vorhersage von Risiken für den Antragsprozess oder für mögliche Präventionsmaßnahmen genutzt werden soll, ist die Sensitivität eine wichtige Kenngröße. Sie gibt an, wie viele der eingetretenen BU-Fälle vom Modell auch richtig erkannt werden. Kunden auf Basis eines Modells abzulehnen, bedeutet andererseits weniger Neugeschäft und langfristig eine kleinere Kundenbasis. Die Genauigkeit der Vorhersage, also das Verhältnis von richtig vorhergesagten BU-Fällen zu allen vorhergesagten BU-Fällen, nimmt daher ebenfalls einen hohen Stellenwert ein.

Das F-beta-Maß als Kombination der beiden Kenngrößen Genauigkeit und Sensitivität kann insbesondere bei unausgeglichener Datenlage – hier mehr Gesunde als BU – ein sinnvoller Ansatz zur Modellbewertung sein. Der Parameter β wurde experimentell bestimmt und ist mit 10 relativ groß, sodass die Sensitivität, die einen großen Einfluss auf die Kosten hat, auch im Fehlerschätzer einen höheren Wert erhält, d. h. hier zehnmal so stark ins Gewicht fällt.

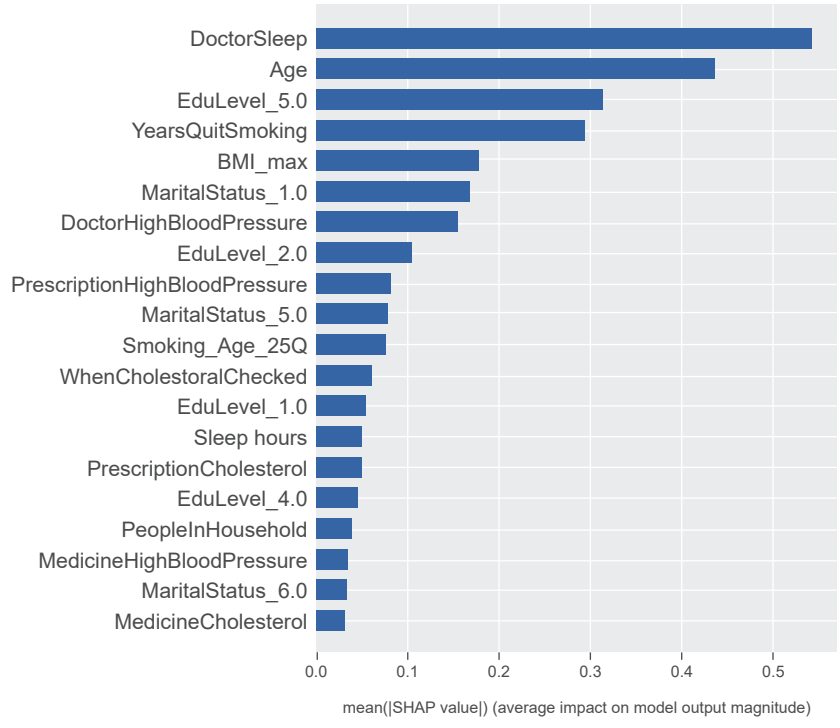
Wenn eine Versicherung auf Basis der Testmenge Anträge bearbeiten würde und dabei die Anträge abgelehnt, die von Random Forest als BU-Fall vorhergesagt werden, würde das bedeuten (siehe Abbildung 2)):

- Von 6467 Anträgen
- würden 303 (4,7 %) abgelehnt (FP + TP),
- 6164 angenommen und
- 688 davon berufsunfähig werden.

Somit würde bei 688/6164 = 11,2 % der Kunden der Versicherungsfall eintreten, was eine leichte Verbesserung zu der naiven Methode wäre.

Wie in Abbildung 3a) zu sehen ist, erzielen alle vier Modelle eine sehr hohe Treffergenauigkeit von mehr als 80 %. An den sehr geringen Sensitivitäten von maximal 20 % (siehe Abbildung 3c)) und den Ergebnissen der Confusion-Matrizen (siehe Abbildung 2 für Random Forest) ist allerdings ersichtlich, dass primär die Nicht-BU-Fälle korrekt klassifiziert werden. Das heißt, dass die Modelle durch die unausgeglichene Datenlage überproportional viele Gesunde vorhersagen, was in der hohen Treffergenauigkeit

Abbildung 6:
Analyse der Einflussfaktoren bei Gradient Tree Boosting anhand der Shapely-Werte



im Vergleich zu den niedrigen Sensitivitäten deutlich wird. Dies verdeutlicht die Bedeutung sich verschiedene Kenngrößen anzuschauen, um ein Modell vollständig bewerten zu können. Das F-beta-Maß ist als Kombination von Genauigkeit und Sensitivität bei allen Modellen niedrig.

Um die unterrepräsentierte Klasse stärker in den Fokus des lernenden Modells zu rücken, können **Resampling-Techniken** bei unausgeglichener Datenlage angewandt werden (vgl. Boyle (2019)). Dabei gibt es zwei Möglichkeiten: Oversampling der Minderheit oder Downsampling der überwiegenden Klasse. Für die Implementierung wurde auf resample von scikit-learn zurückgegriffen.

Beim Oversampling werden zufällige Auszüge der Daten der unterrepräsentierten Klasse mehrfach als Kopie zu den Trainingsdaten hinzugefügt. Abhängig vom Grad des Oversampling kann man in Abbildung 4 die Performance unserer bisher betrachteten Modelle erkennen. Das Random-Forest-Modell wird mit den sich wiederholenden BU-Fällen in der Trainingsmenge kaum besser. Dieses Verhalten ist bekannt und wurde schon vorher in verschiedenen Veröffentlichungen analysiert (vgl. Drumnond (2003)). Die anderen drei Modelle verbessern sich deutlich hinsichtlich Sensitivität, allerdings geht diese Verbesserung zulasten der Treffergenauigkeit und Genauigkeit bei der Vorhersage der BU-Fälle, beide verschlechtern sich mit ausgeglicheneren Daten.

Beim Downsampling wird ein ausgeglichenes Verhältnis von unter- zu überrepräsentierter Klasse erzeugt, indem die verwendete Datenmenge der überrepräsentierten Klasse verkleinert wird. Datenpunkte werden hierbei zufällig ausgesucht und dann entfernt. Auch hier sollte man erst in Trainings- und Testmenge aufteilen, da die Testmenge die Datenbereinigung nicht benötigt. Wir haben nur 13 % BU-Fälle in der Datenmenge, daher müssen wir etwa 83 % der nicht-BU-Fälle aus der Trainingsmenge entfernen, um eine künstliche Gleichverteilung zu erzeugen. Dieses Vorgehen ist nur möglich, wenn überhaupt ausreichend Daten vorhanden sind, da ansonsten die Trainingsmenge insgesamt zu klein werden kann, um ein Modell ausreichend mit Informationen zum Lernen zu versorgen. Beim Downsampling werden alle Modelle deutlich besser in der Sensitivität, siehe Abbildung 5. Das Random-Forest-Modell kommt nun ebenfalls mit der Datenbasis zurecht und die Sensitivität wird auch hier deutlich verbessert.

Am Beispiel der Ergebnisse des Gradient Tree Boosting sind die Ergebnisse für die Versicherung im fiktiven Antragsprozess:

- Von 6467 Anträgen
- würden 2148 (33,2 %) abgelehnt (FP + TP),
- 4319 angenommen und
- 196 davon berufsunfähig werden.

Dadurch ergibt sich eine sehr geringe BU-Quote im Bestand von lediglich 4,5 % im Vergleich zu über 10 % ohne Downsampling-Strategie bzw. beim naiven Vorgehen. Andererseits ist auch der Kundenstamm deutlich kleiner und eine sehr große Zahl von 33,2 % potenzieller Kunden würde direkt im Antragsprozess abgelehnt.

6. Evaluation

In der Evaluation fokussieren wir uns auf die Ergebnisse bezüglich der Einflussfaktoren, gehen auf die Erklärbarkeit der Modelle ein und diskutieren den Mehrwert für die Praxis.

Zur Analyse der Einflussfaktoren der Modelle verwenden wir das Feature „Permutation Importance“ aus dem Python-Paket eli5. Die Idee bei „Permutation Importance“ ist es, dass jeweils ein Attribut entfernt wird, um den Einfluss dieses Attributs auf das Modell zu erkennen. Einzelne Variablen zu entfernen und das ganze Modell neu zu trainieren, kann allerdings sehr großen Rechenaufwand bedeuten. Deswegen werden in diesem Ansatz die Werte der betrachteten Variable durch zufällige Werte ersetzt. Auf diese Weise geht die Information der Variable verloren, ohne das Modell neu trainieren zu müssen.

Ein theoretisches Problem ergibt sich bei der „Permutation Importance“ im Zusammenhang mit „One-hot-kodierten“ Variablen. Hier gilt, dass genau eine Variable den Wert 1 und alle anderen den Wert 0 annehmen. Durch die zufällige Permutation einer Variablen entstehen Parameterkonstellationen, die per Definition nicht möglich sind.

Neben „Permutation Importance“ geben auch die Shapely-Werte aus dem Python-Paket SHAP an, wie viel jedes Feature zu einer bestimmten Vorhersage beiträgt. Abbildung 6 zeigt die Shapely-Werte für Gradient Tree Boosting.

Die relevantesten Einflussfaktoren nach Permutation Importance und SHAP sind insbesondere Schlafprobleme, das Alter und „YearsQuitSmoking“, die Anzahl der rauchfreien Jahre.

Zwar haben wir reale Daten verwendet, allerdings hat es sich in unserem Fall um ein fiktives Projekt im Rahmen der Data Science Challenge gehandelt. Die Fragestellung bezüglich des Einsatzes in einem Versicherungsunternehmen ist daher eher theoretischer Natur. Dennoch wollen wir im Folgenden kurz diskutieren, ob und wie ein solches oder ähnliches Modell konkret eingesetzt werden könnte.

Literaturverzeichnis

Boyle, T. (2019). Dealing with Imbalanced Data – A guide to effectively handling imbalanced datasets in Python.

https://towardsdatascience.com/methods-for-dealing-with-imbalanced-data-5b761be45a18 (abgerufen am 18.04.2021).

Drumnond, C. (2003). Class Imbalance and Cost Sensitivity: Why Undersampling beats Oversampling. In ICML-KDD 2003 Workshop: Learning from Imbalanced Datasets.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research, 12, 2825–2830.

Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Warehousing, 5(4), 13–22.

Bei der Übertragung unserer Ergebnisse auf den Versicherungsbestand eines in Deutschland ansässigen Versicherungsunternehmens sollte man berücksichtigen, dass sich der Begriff und das Verständnis der „Occupational Disability" aus der Studie nicht vollständig mit dem Begriff der Berufsunfähigkeit in Deutschland deckt. Der Fragebogen stellt an dieser Stelle die Frage, ob eine Langzeiterkrankung, in Form von einer körperlichen, geistigen oder emotionalen Beeinträchtigung, eine Arbeit verhindert. Der deutsche BU Begriff ist sehr viel enger gefasst.

In der Regel wird ein Versicherungsunternehmen, das Berufsunfähigkeit versichert, keine so umfangreichen Daten erheben. Daten zu BMI und Rauchverhalten werden meist abgefragt – Daten zum Schlafverhalten eher nicht. Hier könnte unser Modell allerdings Hinweise geben, welche Fragen man im Antragsprozess sinnvollerweise stellen könnte und auch, welche Fragen man nicht unbedingt stellen muss. Die Frage nach der Schlafqualität ist zwar schwer überprüfbar – allerdings ist die Frage so formuliert, dass der Betroffene bereits mit einem Arzt über Schlafprobleme gesprochen hat. Dies könnte man auch im branchenüblichen Prozess gegebenenfalls bereits überprüfen.

Darüber hinaus ergibt sich für Versicherungsunternehmen immer das Problem, dass aus dem eigenen Datenbestand keine Informationen über die Personen gesammelt werden können, deren Anträge abgelehnt wurden. Hier bietet es sich an, zusätzlich zu den unternehmensinternen Daten, wie in unserer Analyse externe Daten zu nutzen. Dadurch könnten Erkenntnisse abgeleitet werden, ob man die falschen Personen ablehnt.

Weiter sieht man in der Branche schon in mehreren Ansätzen die Möglichkeit Daten der Kunden zu erheben, während die Versicherung besteht. So wäre es möglich, sowohl auf Unternehmens- als auch auf Kundenseite einen Mehrwert zu schaffen, indem man sich verstärkt als Gesundheitspartner seiner Kunden versteht und so frühzeitig eingreift, um manche Berufsunfähigkeit zu verhindern. Hier können weitergehende Analysen eine Priorisierung von Präventions- und Rehabilitationsmaßnahmen liefern.

7. Fazit

Unsere Ergebnisse bezüglich der Risikofaktoren zeigen, dass insbesondere das Schlafverhalten, das Alter und das Rauchverhalten einen großen Einfluss auf die Wahrscheinlichkeit haben, berufsunfähig zu werden. Das Rauchverhalten und das Alter sind bekannte Risikofaktoren, die bereits häufig in der Tarifierung verwendet werden. Dies zeigt, dass unsere Ergebnisse plausibel sind und zu dem bisherigen Wissen passen. Zudem erweitern unsere Ergebnisse aber das bekannte Wissen zu Risikofaktoren, weil beispielsweise das Schlafverhalten hier einen weiteren wesentlichen Risikofaktor darstellt. Sicherlich ist es kaum möglich das Schlafverhalten in der Tarifierung zu berücksichtigen, allerdings gewinnen Präventivmaßnahmen, die das Schlafverhalten verbessern (bspw. Apps, die das Schlafverhalten tracken und Feedback geben), an Attraktivität. Basierend auf unseren ersten Ergebnissen könnten zukünftige Analysen noch weitere Daten des NHANES nutzen (u. a. Examination Data, Laboratory Data) und Zusammenhänge zwischen unabhängigen Variablen im Detail untersuchen (bspw. Schlafverhalten, Schichtdienst, Berufe und Ausbildungsniveau).

Im Fokus dieses Projektes stand aber nicht ausschließlich die Forschungsfrage, sondern auch die Auseinandersetzung mit verschiedenen Verfahren und Methoden. Diesbezüglich verdeutlichen unsere Analysen den Mehrwert von Resampling-Verfahren, die bei unserem unausgewogenen Datensatz zu einer deutlichen Verbesserung der Ergebnisse führten. Zudem ist nicht immer der komplexeste Machine-Learning-Algorithmus derjenige, der Zusammenhänge am besten beschreibt. In unserem Fall hat die logistische Regression zu ähnlichen oder sogar besseren Ergebnissen geführt im Vergleich mit beispielsweise Random Forests oder Gradient Tree Boosting. Einer gesamtheitlichen Evaluation der Modelle unter Berücksichtigung der Erklärbarkeit kommt daher eine große Bedeutung zu.

Prof. Dr. Christian Eckert ist Professor für Versicherungs- und IT-Management an der Hochschule Coburg und Aktuar DAV. Seine Forschungsgebiete sind Actuarial Data Science, Risikomanagement und Digitale Transformation im Versicherungsbereich.

Felix Müller ist Data Scientist bei DATEV eG. Seine Haupttätigkeit ist die Entwicklung von Automatisierungsservices mit dem Schwerpunkt auf Klassifikationsverfahren.

Daniela Giesinger ist Leiterin der Abteilung Aktuariat im Bereich Mathematik Leben der NÜRNBERGER Versicherung und seit 2013 Aktuarin (DAV).

Dr. Antonia Schöning ist als Data Scientist bei der Siemens AG im Bereich Smart Infrastructure tätig. Sie entwickelt Anwendungen im Themenfeld Internet of Energy für die IoT Plattform MindSphere.
Sie ist Mathematikerin und Aktuarin DAV und engagiert sich als Dozentin für Actuarial Data Science in der Aus- und Weiterbildung der DAV.