



3

AI: Science
over Fiction

Qualitätssicherung beim produktiven Einsatz von KI-Modellen in der Versicherungsbranche

Faktenpapier

Aus der Serie: AI: Science over Fiction

www.bitkom.org

bitkom

Herausgeber

Bitkom
Bundesverband Informationswirtschaft,
Telekommunikation und neue Medien e. V.
Albrechtstraße 10 | 10117 Berlin
T 030 27576-0
bitkom@bitkom.org
www.bitkom.org

Verantwortliches Bitkom-Gremium

AK Artificial Intelligence

Projektleitung

Dr. Nabil Alsabah | Bitkom e. V.

Autorinnen

Janera Kronsbein | Eucon Digital GmbH
Antje Fitzner | Eucon Digital GmbH

Lektorat

Linda van Rennings | Bitkom e. V.

Satz & Layout

Katrin Krause | Bitkom e. V.

Titelbild

© Randall Ruiz | unsplash.com

Copyright

Bitkom 2019

Diese Publikation stellt eine allgemeine unverbindliche Information dar. Die Inhalte spiegeln die Auffassung im Bitkom zum Zeitpunkt der Veröffentlichung wider. Obwohl die Informationen mit größtmöglicher Sorgfalt erstellt wurden, besteht kein Anspruch auf sachliche Richtigkeit, Vollständigkeit und/oder Aktualität, insbesondere kann diese Publikation nicht den besonderen Umständen des Einzelfalles Rechnung tragen. Eine Verwendung liegt daher in der eigenen Verantwortung des Lesers. Jegliche Haftung wird ausgeschlossen. Alle Rechte, auch der auszugsweisen Vervielfältigung, liegen beim Bitkom.

1 Die Digitalisierung stellt Versicherer vor Herausforderungen – KI kann helfen

Kunden erwarten auch in Versicherungsfragen das, was sie aus anderen Lebensbereichen bereits gewohnt sind: sämtliche Leistungen, die bisher manuell erledigt wurden, schnell, einfach und transparent digital abzuwickeln. Diesen Anspruch kann ein Versicherer nur erfüllen, indem entsprechende Tätigkeiten automatisiert werden. Versicherern stehen heute moderne KI-gestützte Systeme zur Verfügung, mit denen sie den Schadenprozess vereinfachen und zielgerichteter steuern können.

Die manuelle Bearbeitung von Schadendokumenten ist zeit- und kostenintensiv. In der Folge vergeht von der Meldung des Schadens durch den Versicherten bis zur ersten Zahlung unnötig viel Zeit. Ein Großteil aller Fälle, die die Versicherer erreichen, sind so genannte »Gut-Fälle«. Das sind Fälle, bei denen keine Prüfung durch einen Sachbearbeiter erforderlich ist. Eucon unterstützt die Versicherer hier bereits seit mehr als 15 Jahren in der Schadenbearbeitung und deren Digitalisierung und Automatisierung. Bisher wurden alle Schadenfälle manuell durch Experten aus Handwerk und Kfz auf Richtigkeit geprüft. Diese Bearbeitung ist zeitintensiv.

Die automatische Bearbeitung von Schadenfällen mit klarer Sachlage hilft Versicherern, Abläufe zu beschleunigen und Fachkräfte noch sinnvoller einzusetzen. So können die Experten ihre Zeit für komplexe und unklare Schadenfälle nutzen. Mithilfe von Algorithmen des maschinellen Lernens können wir die Wahrscheinlichkeit ermitteln, ob es sich bei einem Vorgang um einen »Gut-Fall« handelt. Durch diese Prognose von Prüfbedarfen kann ein Teil der Schäden automatisiert verarbeitet werden. Der manuelle Prüfaufwand wird reduziert und Bearbeitungszeiten werden um bis zu 80% verringert.

Doch trotz der immensen Leistung moderner Systeme bleibt Domänen- und Expertenwissen unverzichtbar. Ein KI-System lernt zunächst einmal anhand von Beispielen und Daten aus abgeschlossenen Fällen. Hierzu wird eine ausreichend große Menge an Daten für die Modellentwicklung und -validierung benötigt. Die Qualität der verfügbaren Daten ist für den maschinellen Lernprozess entscheidend. Nur wenn die Qualität der Daten hoch genug ist, kann das System jene Gemeinsamkeiten ausmachen, die Gut-Fälle aufweisen. Hat das System die Gemeinsamkeiten erlernt, so kann es künftig neue Fälle selbstständig verarbeiten.

»Die KI-gestützte Automatisierung ist der Schlüssel für eine schnelle Abwicklung von Versicherungsfragen.«



2 Eine Qualitätssicherung während des produktiven Einsatzes von KI-Systemen ist unverzichtbar

Die Erstellung von KI-Modellen benötigt eine ausreichend große Menge an Daten. Hierbei handelt es sich um historische Daten, die von Menschenhand bearbeitet wurden. Für die Modellentwicklung gehen wir von der Korrektheit der Daten aus. Eventuelle Fehler durch die manuelle Klassifikation werden vernachlässigt.

In unserem Anwendungsfall handelt es sich dabei um Schadenfälle, die von einem Experten beurteilt wurden.

Wie viele Datensätze benötigt werden, hängt von der Datenqualität und auch dem Ziel des Modells ab. Für das Training wird die zur Verfügung stehende Datenmenge in drei Datensets unterteilt:

- **Trainingsdaten** werden für die eigentliche Modellerstellung, das sogenannte Training eines Modells, verwendet. Sie sind der Großteil der Datenmenge.
- **Validationsdaten** werden im Training genutzt, um verschiedene Parameter des Algorithmus zu optimieren, so z. B. die Anzahl der Knoten in den Schichten des neuronalen Netzes. Hier reicht ein kleiner Teil der Daten.
- **Testdaten** sind dem Modell unbekannt und dienen so dem eigentlichen Test des erstellten Modells.

Ein Modell wird also mit für das trainierte Modell unbekanntem Daten evaluiert. Dabei will man die voraussichtliche Leistung des Systems im produktiven Einsatz vorab testen. Diese Leistungsprognose ist nur dann zuverlässig, wenn die Test- und Trainingsdaten die Realität widerspiegeln.

Da die Modelle mit historischen Daten trainiert werden, beginnt jedes Modell sofort nach der Erstellung zu altern und weicht immer weiter von der Realität ab, je älter es wird. Die aktuellen Daten, die durch ein solches Modell beurteilt werden sollen, können von den Daten in der Modellerstellung abweichen. Somit muss die Modell-Performanz kontinuierlich evaluiert werden, um Veränderungen schnell aufzuspüren – und daraufhin eingreifen zu können.

Die produktive Nutzung eines Modells steht dabei konträr dazu, dass die Datenmenge zur Weiterentwicklung aktuell und repräsentativ gehalten werden soll. Das liegt daran, dass die Vorgänge, die durch ein KI-Modell beurteilt wurden, nicht für die Weiterentwicklung genutzt werden können. Die Verwendung würde dazu führen, dass sich das Modell zunehmend selbst bestätigt und vorhandene Fehler verstärkt werden.

Um eine ausreichend große aktuelle Datenmenge sicherzustellen und die Leistung des eingesetzten Modells zu kontrollieren, ist deshalb eine Qualitätssicherung (QS) notwendig.

3 Die produktive Nutzung und stetige Aktualisierung des ML-Modells stehen konträr zueinander

Wie kann eine solche Qualitätssicherung aussehen? Da es sich bei der Gut-Fall-Prognose um ein KI-Modell zur Prozessautomatisierung handelt, ist es naheliegend, die Qualität durch manuelle Stichprobenprüfungen der Experten abzubilden. Dies entspricht gleichzeitig einer konsistenten Aktualisierung der ursprünglichen Trainingsmenge. Denn das Modell wurde auf historischen Daten trainiert, die über die manuelle Expertenprüfung generiert wurden.

Für die Durchführung der manuellen Stichprobenprüfungen für Vorgänge, die automatisiert hätten verarbeitet werden können, gibt es zwei Ansätze:

- Einerseits kann die manuelle QS als nachträgliche Überprüfung des automatisiert erstellten Ergebnisses ausgeführt werden. In diesem Fall kann der Geschwindigkeitsvorteil im Prozess für alle Vorgänge genutzt werden und die QS besonders effizient erfolgen. Da das Prüfergebnis bekannt ist, ist die QS jedoch nicht vollständig unabhängig und könnte einen Bias in den zukünftigen Trainingsdaten erzeugen.
- Andererseits kann die QS durch die Aussteuerung eigentlich automatisiert zu verarbeitender Vorgänge an die Expertenprüfung direkt im Prozess erfolgen. So kann eine vollkommen unabhängige manuelle Prüfung im regulären Prozessverlauf erreicht werden. Allerdings ist dieser Ansatz aufwendiger, da mehr Vorgänge manuell geprüft werden und der Geschwindigkeitsvorteil nicht mehr vollständig ausgenutzt werden kann.

Bei der Entscheidung, welche der Optionen sinnvoll anzuwenden ist, spielen noch weitere Kriterien eine Rolle:

- Größe der Stichprobe: Für wie viele Vorgänge wird eine manuelle QS benötigt?
- Repräsentativität: Wie werden die Heterogenität und Änderungen in den Daten berücksichtigt?
- Abhängigkeit von der Modellsicherheit: Kann die Effizienz der QS durch diesen Aspekt optimiert werden?

4 Methodik und Intensität der QS sind situativ veränderlich, jedoch nie unverzichtbar

Um die Größe der benötigten Stichprobe zu ermitteln, sind zunächst folgende Fragen zu beantworten:

- Wie groß muss die Trainingsmenge sein, um ein stabiles Modell zu erhalten?
- Wie alt dürfen die Datensätze sein, um ein aktuelles Modell zu gewährleisten?

Sind die grundlegende Größe der Trainingsmenge sowie das maximale Alter der Trainingsdaten bekannt, kann die Anzahl der (z. B. pro Jahr) zu aktualisierenden Datensätze für ein stabiles Modell berechnet werden.

Die Fragestellungen lassen sich nicht allgemein beantworten, sondern hängen von dem individuellen Anwendungsfall ab. Dabei beeinflussen u. a. die Komplexität in der Zielfragestellung, die Anzahl der Input-Dimensionen und die Varietät der Trainingsdaten die Größe der benötigten Trainingsmenge.

Auch der Einfluss des Alters der Trainingsdaten hängt von verschiedenen Faktoren ab. So spielt z. B. die Abhängigkeit der Inputdaten von allgemeinen und saisonalen Trends (z. B. Inflation, Jahreszeiten) eine Rolle. Um diese Fragen für unseren Anwendungsfall zu beantworten, wurden die verfügbaren Daten im Detail analysiert. Zudem wurden verschieden große Teilmengen der verfügbaren Daten für Testmodelle genutzt und so die Datenverfügbarkeit in einem begrenzten Zeitraum simuliert. Die Auswirkungen von Alter und Datenmenge in unserem Anwendungsfall können so untersucht werden.

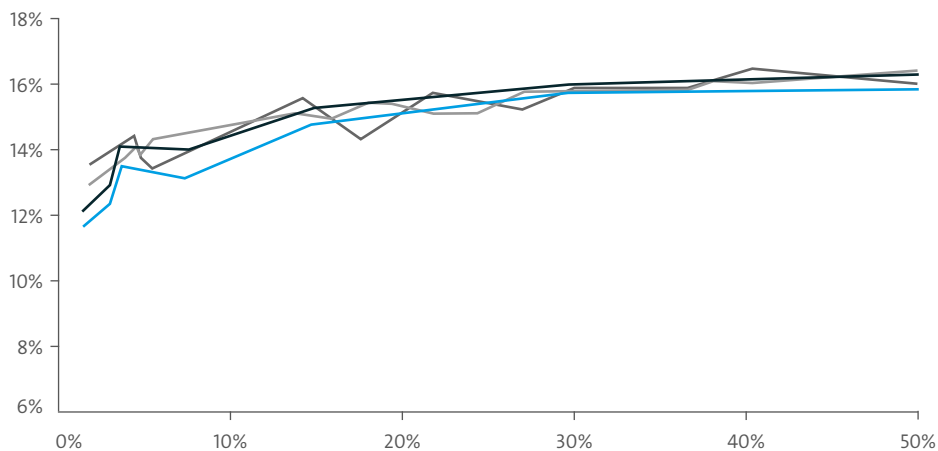


Abbildung 1: Dunkelverarbeitungsrate in % abhängig von der Größe der Trainingsmenge in % relativ zu den zur Verfügung stehenden Daten für vier Modellläufe

Abbildung 1 zeigt ein Beispiel für den Leistungsverlauf eines Modells mit abnehmender Trainingsdatenmenge. Der Anteil der automatisiert verarbeiteten Vorgänge wird als Dunkelverarbeitungsrate bezeichnet.

Es ist erkennbar, dass bis zu einer Reduktion der Trainingsmenge auf ca. 30% der Ausgangsdatenmenge eine relativ konstante Leistung (erkennbar am Prozentsatz der Dunkelverarbeitung) erreicht werden konnte. Bei weiterer Reduzierung der Trainingsdaten hingegen fällt die Dunkelverarbeitungsrate zunehmend, was auf eine dann nicht mehr ausreichende Datenmenge schließen lässt. Da dieser Punkt erst nach erheblicher Reduktion der Datenbasis eintritt, steht für die Gut-Fall-Prognose eine momentan ausreichende Datenmenge zur Erstellung eines aussagekräftigen KI-Modells zur Verfügung [1]. Das Experiment stellt dabei eine Momentaufnahme dar. Eine kontinuierliche Aktualisierung ist, wie oben erläutert, notwendig.

5 Die genutzte Stichprobe muss die Gesamtheit der Daten repräsentieren

Bei der Auswahl der Stichprobe sollte beachtet werden, dass die Gesamtheit der Vorgänge hinsichtlich der für das Modell zum Trainieren verwendeten Faktoren repräsentiert wird.

Als Referenz für die Verteilung ist dabei der aktuelle Zeitraum zu betrachten, für den die Stichprobe entnommen werden soll. Denn im Vergleich zu den historischen Trainingsdaten, kann es zu Veränderungen kommen, z. B. mehr Sturmfälle durch den Klimawandel. Im Kontext der Gut-Fall-Prognose ist das wichtigste Kriterium für die Verteilung der Status, ob es sich um einen Gut-Fall handelt. Denn genau dieser soll prognostiziert werden. Weitere Eigenschaften können z. B. der Schadentyp (Leitungswasser/Sturm/etc.) oder die Beleghöhe (in €) sein. Gerade bei monetären Werten gibt es z. B. durch die Inflation oder neue Kfz-Generationen eine ständige Änderung der Werte, die darüber hinaus in der Datenbereinigung zu berücksichtigen ist.

»Ein KI-Modell gibt eine Wahrscheinlichkeit für den jeweiligen Status als Output zurück: Das Modell ist sich zum Beispiel zu 20%, 50% oder aber 85% sicher, dass es sich um einen Gut-Fall handelt.«

6 Die Auswahl der Stichproben kann auch von der Vorhersagewahrscheinlichkeit abhängen

Ein KI-Modell gibt eine Wahrscheinlichkeit für den jeweiligen Status als Output zurück: Das Modell ist sich zum Beispiel zu 20%, 50% oder aber 85% sicher, dass es sich um einen Gut-Fall handelt. In der produktiven Nutzung wird anschließend über einen Schwellwert entschieden, ob der Output eine ausreichende Sicherheit für die Anwendung des prognostizierten Status darstellt. Vorgänge mit einer ausreichend hohen Wahrscheinlichkeit in der Gut-Fall-Prognose können automatisiert verarbeitet werden. Dabei steigt die Sicherheit des Modells (»Handelt es sich um einen Gut-Fall?«) mit der zurückgegebenen Wahrscheinlichkeit.

Für die QS kann entsprechend der Ansatz gewählt werden, bis zu einem gewissen Wahrscheinlichkeitswert, in dem verhältnismäßig viele Fehler auftreten, eine ausführliche QS durchzuführen. Oberhalb einer hohen Sicherheit genügen nachträgliche Checks, da die Fehlerwahrscheinlichkeit gering ist. Gegebenenfalls kann für diese Vorgänge sogar auf eine QS verzichtet werden. Denn auch im manuellen Prozess passieren Fehler, sodass eine vollständige Korrektheit bei einem derartig komplexen Anwendungsfall nicht erreicht werden kann. Die durch die fehlende QS zusätzlich erzeugten Fehler hätten in diesem Fall nur marginalen Einfluss auf das Ergebnis und können deshalb vernachlässigt werden.

7 Neben der QS werden auch Erklärmodelle zur Fehlerreduzierung eingesetzt

Neben der Nutzung der aktualisierten Trainingsdaten aus den Ergebnissen der QS zur Erhaltung eines stabilen Modells können identifizierte Fehler des Modells anschließend weiter analysiert werden. Dabei können Verfahren wie LIME Licht in den Entscheidungsprozess eines neuronalen Netzes bringen. Damit kann man jene Faktoren ausmachen, die das Entscheidungsverhalten des ML-Modells besonders stark beeinflusst haben. Zum anderen wird die Begründung des Ergebnisses der manuellen Prüfung durch die Experten analysiert. Aus diesen Analysen werden Rückschlüsse gezogen, in welchen Bereichen Fehler durch das Modell passieren. Dies gibt gleichzeitig Aufschluss darüber, wie gravierend die Fehleinschätzungen sind. Beispielsweise ist eine Fehleinschätzung aufgrund ungenauer Schadennachweise, wobei Preise und Reparaturweg marktüblich und plausibel sind, wesentlich weniger kritisch, als z. B. das Nicht-Ausweisen von Sanierungsanteilen.

Diese Analysen können nicht nur in der QS, sondern auch bereits bei der Entwicklung der Modelle genutzt werden, denn sie helfen Fehlermuster des Modells zu erkennen. Diese können dann beispielsweise durch eine gezielte Verstärkung bestimmter Input-Dimensionen durch Regelwerke korrigiert werden.

»Es gibt mittlerweile Verfahren, die das Entscheidungsverhalten von neuronalen Netzen transparent machen können. So kann man jene Faktoren identifizieren, die die Entscheidung eines ML-Modells besonders stark beeinflusst haben.«

8 Die Datengrundlage ist entscheidend für die Entwicklung eines KI-Modells

Für den Anwendungsfall zur Erkennung von Gut-Fällen im Schadenmanagement wurden verschiedene Möglichkeiten zur Qualitätssicherung aufgezeigt. Zudem wurde verdeutlicht, welche Kriterien in die Auswahl einer Option (oder eine Kombination) einfließen sollten.

Wird beispielsweise in der Analyse festgestellt, dass die aktuelle Datenbasis für das weitere Modell-Training eher ausgebaut werden sollte, ist auf eine umfassende und möglichst unabhängige QS zurückzugreifen, um eine qualitativ hochwertige Trainingsmenge aufzubauen.

Stellt sich die vorhandene Trainingsmenge als umfassend und wenig anfällig z. B. für Trendänderungen dar (dies hängt vom Anwendungsfall und den Möglichkeiten der Approximation der Input-Dimensionen ab), kann die einfachere QS über nachträgliche Prüfungen gewählt werden. Die Auswahl der Methode sollte langfristig überprüft und an geänderte Rahmenbedingungen angepasst werden.

Dabei ist das Wissen der Experten der manuellen Prüfung essentiell für die Entwicklung und Qualitätssicherung der Algorithmen zur Automatisierung. Denn ohne manuelle Prüfung in Stichproben ist keine nachhaltige Automatisierung möglich.

9 Literaturverzeichnis

1. Janera Kronsbein und Christina Blanke, Modellerwartungen vs. Realität – Monitoring von Modellperformance, Vortrag auf der M3 Konferenz für Machine Learning und künstliche Intelligenz, Mai 2019
2. [↗Marco Túlio Ribeiro](#), [↗Sameer Singh](#), [↗Carlos Guestrin](#), »Why Should I Trust You?«: Explaining the Predictions of Any Classifier, HLT-NAACL Demos 2016, DOI:10.1145/2939672.2939778
3. LIME package on github: [↗https://github.com/marcotcr/lime](https://github.com/marcotcr/lime)

Autorinnen



Antje Fitzner

Antje Fitzner hat in Nijmegen Physik und Astronomie studiert und anschließend in Kopenhagen in Geophysik promoviert. Jetzt ist sie Teamleiterin des Data-Science-Teams bei der Eucon Digital GmbH, das sich mit den verschiedensten KI-Projekten im Unternehmen beschäftigt.



Janera Kronsbein

Janera Kronsbein ist Produktmanagerin für die KI-Lösungen der Eucon Digital GmbH. Nach ihrem Wirtschaftsmathematikstudium hat sie sich bei Eucon bereits aus verschiedenen Perspektiven z. B. als Data Scientist mit den Themen Prozessautomatisierung und -digitalisierung beschäftigt.



Bitkom vertritt mehr als 2.700 Unternehmen der digitalen Wirtschaft, davon gut 1.900 Direktmitglieder. Sie erzielen allein mit IT- und Telekommunikationsleistungen jährlich Umsätze von 190 Milliarden Euro, darunter Exporte in Höhe von 50 Milliarden Euro. Die Bitkom-Mitglieder beschäftigen in Deutschland mehr als 2 Millionen Mitarbeiterinnen und Mitarbeiter. Zu den Mitgliedern zählen mehr als 1.000 Mittelständler, über 500 Startups und nahezu alle Global Player. Sie bieten Software, IT-Services, Telekommunikations- oder Internetdienste an, stellen Geräte und Bauteile her, sind im Bereich der digitalen Medien tätig oder in anderer Weise Teil der digitalen Wirtschaft. 80 Prozent der Unternehmen haben ihren Hauptsitz in Deutschland, jeweils 8 Prozent kommen aus Europa und den USA, 4 Prozent aus anderen Regionen. Bitkom fördert und treibt die digitale Transformation der deutschen Wirtschaft und setzt sich für eine breite gesellschaftliche Teilhabe an den digitalen Entwicklungen ein. Ziel ist es, Deutschland zu einem weltweit führenden Digitalstandort zu machen.

**Bundesverband Informationswirtschaft,
Telekommunikation und neue Medien e.V.**

Albrechtstraße 10
10117 Berlin
T 030 27576-0
F 030 27576-400
bitkom@bitkom.org
www.bitkom.org

bitkom