QuantAkt

Menschen fragen, Maschinen antworten. Wie kann man komplexe Modelle erklären?

7. Dezember 2023

q_x-Club Zürich

Guido Grützner (guido.gruetzner@quantakt.com)

Agenda

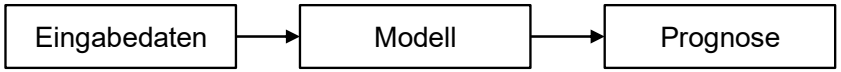
- Statistische Modelle
- Warum braucht es Erklärbarkeit?
- Beispiele für Erklärbarkeitsmethoden
- Fazit

Um welche komplexen Modelle geht es?

- Alle reden über KI, keiner über das Wetter! Warum?
- Es gibt zwei fundamental verschiedene Kategorien von Modellen

Naturwissenschaftlich-kausales Modell

- Beispiel Wetterprognose: Bestimmung der Niederschlags-wahrscheinlichkeit morgen in Zürich



- Temperatur
- Feuchtigkeit
- Luftdruck
- ……

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla)\mathbf{v} = -\frac{1}{\rho}\nabla\rho + \nu\Delta\mathbf{v} + f(\boldsymbol{x}, t)$$

Schnee: 5%
Regen: 45%

Statistisches Modell

- Beispiel Unfallprognose: Bestimmung der Unfallwahrscheinlichkeit eines Fahrers

- Fahrzeug
- Wohnort
- Schadenklasse
- ….

$$\log \mu = \beta_0 + \beta_1 X_1 + \cdots + \beta_d X_d \qquad p = 0.1‰$$

- Das Wettermodell
  - Basiert auf Naturgesetzen,
  - bildet kausale Zusammenhänge ab,
  - Prognosen werden aus Daten durch Differentialgleichungen gewonnen.

- Das Unfallmodell
  - Basiert auf beobachteten Daten
  - bildet statistische Assoziationen ab,
  - Prognosen werden aus dem besten Fit an die Beobachtungen gewonnen.

Statistisch, Klassisch, Machine Learning

- Dieser Vortrag beschäftigt sich mit der Erklärbarkeit von statistischen Modellen
  - Praktisch alle aktuariellen Modelle sind von dieser Art
  - Machine Learning/Statistical Learning Modelle sind auch von dieser Art
- In diesem Vortrag unterscheide ich nicht zwischen «klassischen» und «Machine Learning» Modellen
  - Meiner Meinung nach gibt es keinen fundamentalen Unterschied
  - Den Unterschied machen Modellcharakteristika wie die Komplexität und das Einsatzgebiet des Modells
- Statistische Modelle haben grosse Vorteile gegenüber naturwissenschaftlich-kausalen Modellen
  - Man kann sehr gute Prognosen abgeben OHNE das man die zu Grunde liegenden Phänomene vollständig versteht oder abbildet
  - Sie sind im Allgemeinen viel einfacher zu erstellen und viel schneller zu berechnen als naturwissenschaftlich-kausale Modelle

Konsequenzen für die Erklärbarkeit

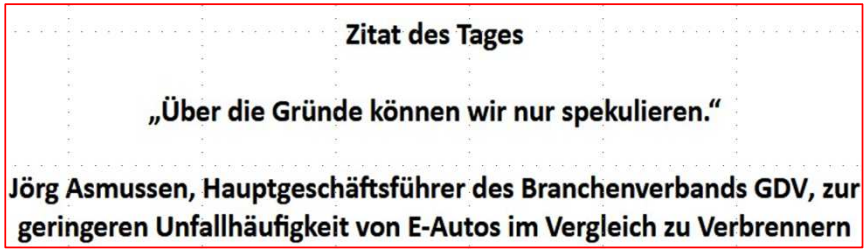
- Erklärungen in einem Wettermodel

> Ein Tief zieht heute von Deutschland zum Balkan. Auf seiner Rückseite steuert es aus Norden feuchte Polarluft zur Alpennordseite. […] Am Donnerstag steuert ein neues Tief mit Kern über der Biskaya eine Warmfront zur Schweiz.

MeteoSwiss, 28.November

- Die kausale Abbildung der Zusammenhänge erlaubt kausale Erklärungen.
- Die internen Elemente des Modells stehen in einem unmittelbaren Zusammenhang mit der abgebildeten Wirklichkeit.

- Statistisches Modell ???



Versicherungsmonitor vom 26. Oktober 2023

- Da ein statistisches Modell nur statistische Assoziationen abbildet sind keine kausalen Erklärungen möglich.
- Die internen Elemente des Modells (im Beispiel die Koeffizienten $\beta_0, \beta_1, \ldots \beta_d$) haben keinen direkten Bezug zu Phänomenen der abgebildeten Wirklichkeit.

Agenda

- Statistische Modelle
- **Warum braucht es Erklärbarkeit?**
- Beispiele für Erklärbarkeitsmethoden
- Fazit

Erklärbarkeit versus Nachvollziehbarkeit (Transparency)

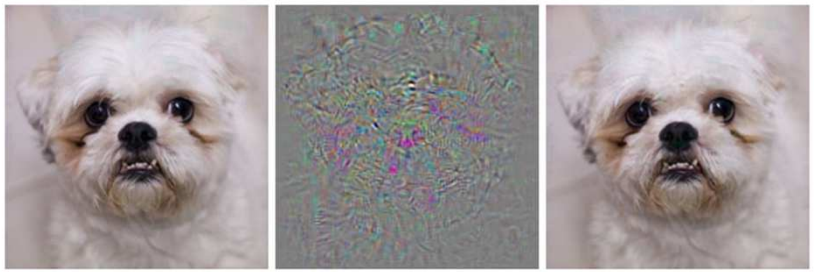
- Es ist sinnvoll zwischen der Erklärbarkeit des Modells im engeren Sinne und der Nachvollziehbarkeit des Modellierungsprozesses insgesamt zu unterscheiden
- Nachvollziehbarkeit («Transparency» in [1]) umfasst unter anderem
  - Angemessenheit des Modells für den Einsatzzweck
  - Auswahl der Daten (aktuell, korrekt, repräsentativ, …)
  - Technische Korrektheit der Implementierung
  - Dokumentation und Revisionssicherheit
- Ein juristisches «Recht auf Erklärung» bezieht sich sicherlich auf beide Aspekte also Erklärbarkeit und Nachvollziehbarkeit
- Die Unterscheidung ist aber dennoch sinnvoll, weil
  - die auftretenden Probleme und dann nötigen Methoden zur Lösung sehr verschieden sind.
  - Methoden zur Nachvollziehbarkeit sind Anwendern aus regulierten Bereichen wohl vertraut (und Routine?).
- Annahme im Folgenden: Nachvollziehbarkeit ist vollständig und lückenlos gegeben.
- Im Weiteren: Fokus auf Erklärbarkeit im engeren Sinne!

[1] EIOPA, , 2021
«ARTIFICIAL INTELLIGENCE GOVERNANCE PRINCIPLES»

Braucht es denn Erklärbarkeit?

- Annahmen:
  1. Wir sind nur an guter Prognose interessiert.
  2. Die Qualität der Prognosen des Modells ist tipp-topp.
  3. Daten, Modell und Prozess drumherum sind perfekt nachvollziehbar.
- These: «Let the data speak for itself!»
  - Jeder Datensatz, der interessiert, kann ja in das Modell eingespeist werden.
  - Man erhält dazu die passende Prognose.
  - Schon weiss man alles über das Modell, was relevant ist.
- Wozu braucht man dann «Erklärungen»?
- Das Problem:
  - Egal wie gross der Datensatz ist: Er ist immer endlich
  - Egal wieviel man testet: Man kann nur endlich viele Kombinationen von Inputs prüfen
- Die Frage bleibt immer: Wie reagiert das Modell auf Daten, auf die es nicht kalibriert/getestet wurde?
  - Im Machine Learning nennt man das «Generalisierung»
  - In der Mathematik heisst das «Extrapolation»
- Die folgenden zwei Beispiele zeigen, dass auch sehr gute Modelle versagen können!

Adversarial examples – Klassifikation von Bildern



Zitat aus der Veröffentlichung:
We find that applying an imperceptible non-random perturbation to a test image, it is possible to arbitrarily change the network's prediction. These perturbations are found by optimizing the input to maximize the prediction error.

We term the so perturbed examples "adversarial examples".

Ian Goodfellow et. al. «Intriguing properties of neural networks», 2014
https://www.youtube.com/watch?v=CIfsB_EYsVI

Adversarial policies oder «Wie man Go-Programme besiegt»

Demis Hassabis

Lee Sedol

Kellin Pelrine, Go-Amateur

We attack the state-of-the-art Go-playing AI system KataGo by training adversarial policies against it, achieving a >97% win rate against KataGo running at superhuman settings.

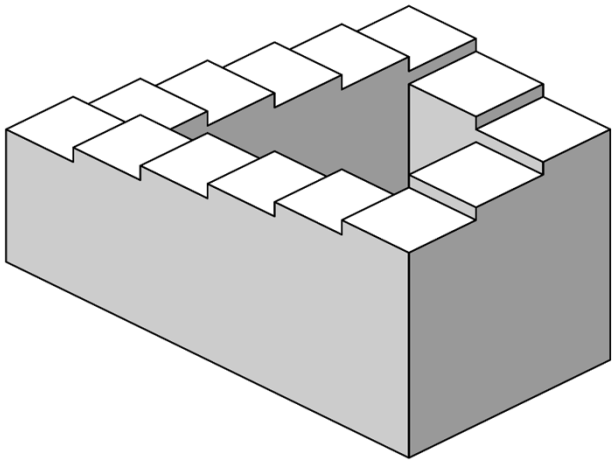
Our adversaries do not win by playing Go well. Instead, they trick KataGo into making serious blunders. Our attack transfers zero-shot to other superhuman Go-playing AIs, and is comprehensible to the extent that human experts can implement it without algorithmic assistance to consistently beat superhuman AIs.

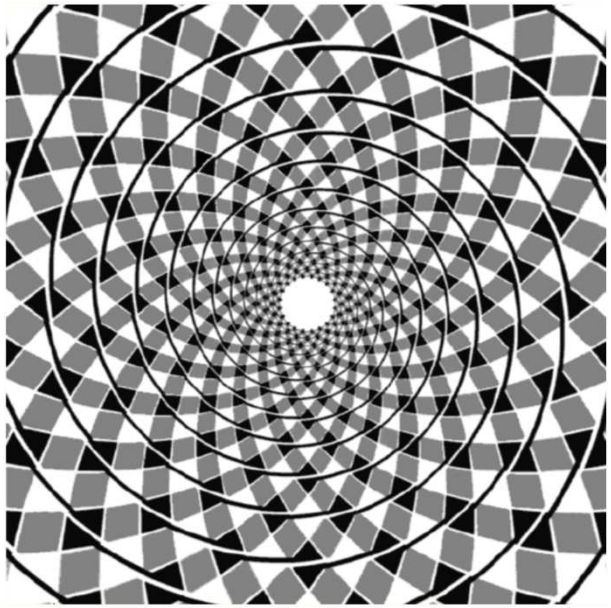
The core vulnerability uncovered by our attack persists even in KataGo agents adversarially trained to defend against our attack. Our results demonstrate that even superhuman AI systems may harbor surprising failure modes.

Tony T Wang et.al., 2023
"Adversarial Policies Beat Superhuman Go AIs"
https://goattack.far.ai/

Nur zur Erinnerung!

Es gibt auch Adversarial Examples für biologische neuronale Netze!





https://michaelbach.de/ot/

Fazit: Es braucht mehr als nur Prognose!

- Diese Beispiele beleuchten ein fundamentales Problem: Statistischen Modellen fehlen «Leitplanken» wie Naturgesetze oder logische Prinzipien
- Gute Qualität – sogar auf grossen Trainings- und Testdatensätzen – kann daher Fehlverhalten fernab dieser Daten nicht ausschliessen. Daher ist ein globales Verständnis des Modells erforderlich.
- Das Problem wird umso grösser je komplexer das Modell wird.
- Komplexität kann dabei gemessen werden durch
  - Die Dimensionalität des Inputraumes also z.B. der Anzahl der Input-Variablen.
  - Der Dimensionalität des Hypothesenraumes also z.B. der Anzahl der freien Parameter des Modells.

- Über diese fundamentale Notwendigkeit hinaus ist Erklärbarkeit aber auch nützlich
  - Zum Debugging/Test des Modells
  - Zur Selektion geeigneter Modelle und Inputdaten
  - Zur Unterstützung von Business-Cases
  - Bei rechtlichen Aspekten wie Verantwortungsübernahme, Haftung, Fairness.

Agenda

- Statistische Modelle
- Warum braucht es Erklärbarkeit?
- **Erklärbarkeitsmethoden**
- Fazit

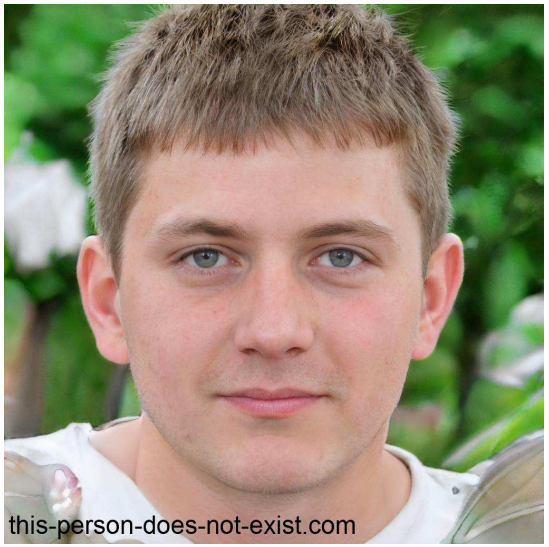
Definitionen von Erklärbarkeit/Interpretierbarkeit

- Einige Definitionsversuche in der Literatur
  - An interpretable system is a system where a user cannot only see but also study and understand how inputs are mathematically mapped to outputs. (Addadi und Berrada 2018)
  - You could describe a model as interpretable if you can comprehend the entire model at once. (Lipton 2016)
  - A model is explainable when its internal behaviour can be directly understood by humans (interpretability) or when explanations (justifications) can be provided for the main factors that led to its output." (European Banking Authority, 2020)
- Diese Definitionen (und viele mehr) haben eins gemeinsam: Sie sind zirkulär!
  - Ein undefinierter Begriff „interpretable/explainable" wird ersetzt durch einen anderen undefinierten Begriff „understandable" „compehensible" usw.
- Es gibt keine allgemein anerkannte und praktikable Definition von erklärbar/interpretierbar!

- Die Konsequenz: Unsicherheit
  - Erklärbarkeit ist nicht messbar/operationalisierbar was faktenbasierte Diskussionen darüber schwierig macht.
  - Wenn ein Gesetz oder eine Aufsicht «Erklärbarkeit» verlangt, ist völlig unklar, wie das zu erreichen ist.
  - Für Hersteller und Verwender von statistischen Modellen ist nicht klar, welche Modelle, wann als ausreichend erklärt gelten.
- Konsequenz für den Vortrag (und evtl. Empfehlung für Anwender/Nutzer?): Beispielhaft und pragmatisch vorgehen!
  - Beispiele geben für Methoden zur Erklärung.
  - Auf natürliche Fragen und passende Antworten konzentrieren.
  - Ein Modell gilt dann als ausreichend erklärt, wenn die Erklärung Kunden, Anwender oder Bundesrichter überzeugt.

Erklärungsmethode: Counterfactual explanations

Luca, 21 Jahre, Plattenleger (EFZ),
Will sich sein erstes Auto kaufen



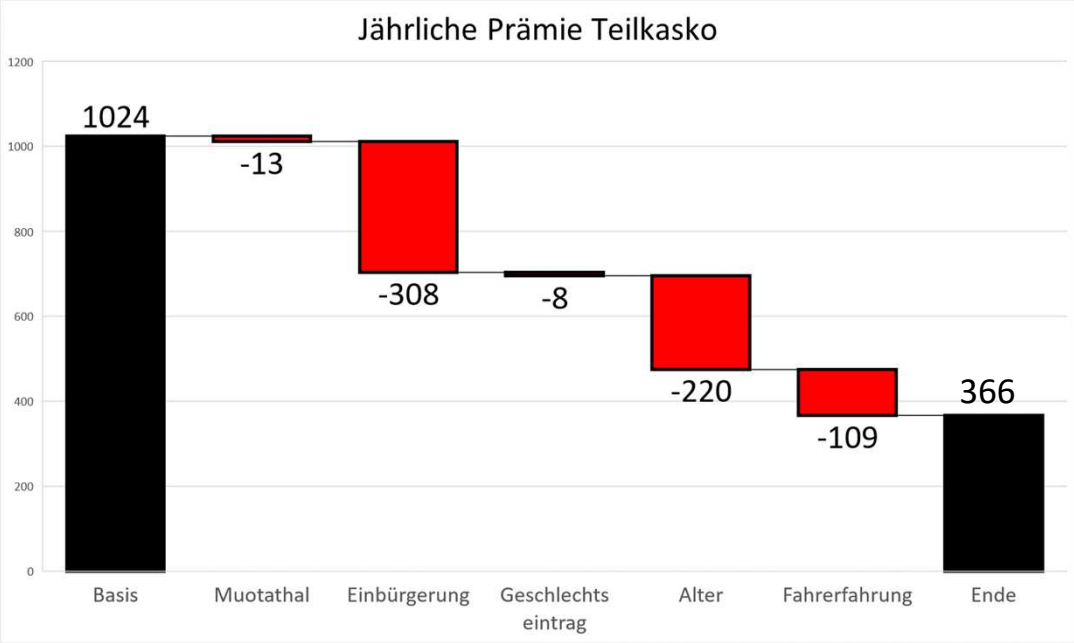
Subaru BRZ 2.4 Sport, 172kW

Aber: Prämie Teilkasko CHF 1024 !

Lucas Erklärungsproblem bzw. seine naheliegende Frage:
Warum ist meine Prämie so hoch??

Erklärungsmethode Counterfactual Explanations

- Sein Ansatz: Counterfactual Explanations oder «Was wäre wenn?»
  - Er rechnet mit einem On-Line Tarifrechner einfach ein paar Varianten seiner Daten.



- Seine (hypothetischen) Massnahmen
  - Von 8051 Schwamendingen zur Oma nach 6436 Muotathal ziehen
  - Endlich die Einbürgerung Albanien -> CH anpacken
  - Geschlechtseintrag M -> F: Vielleicht doch nicht für CHF 8 ?
  - Alter: von 21 auf 41 erhöhen
  - Zeit seit Fahrprüfung: von 1 Jahr auf 21 Jahre erhöhen.

- Counterfactuals sind eine lokale Methode
  - Erklärt wird eine ganz bestimmte Prognose bzw. Modell-Entscheidung.
- Für Counterfactuals braucht es nur die Modell-Outputs
  - Die Interna des Modells können völlig verborgen bleiben (Blackbox-Ansatz)
  - Es braucht sogar nur wenige Werte in der Nachbarschaft des relevanten Outputs.

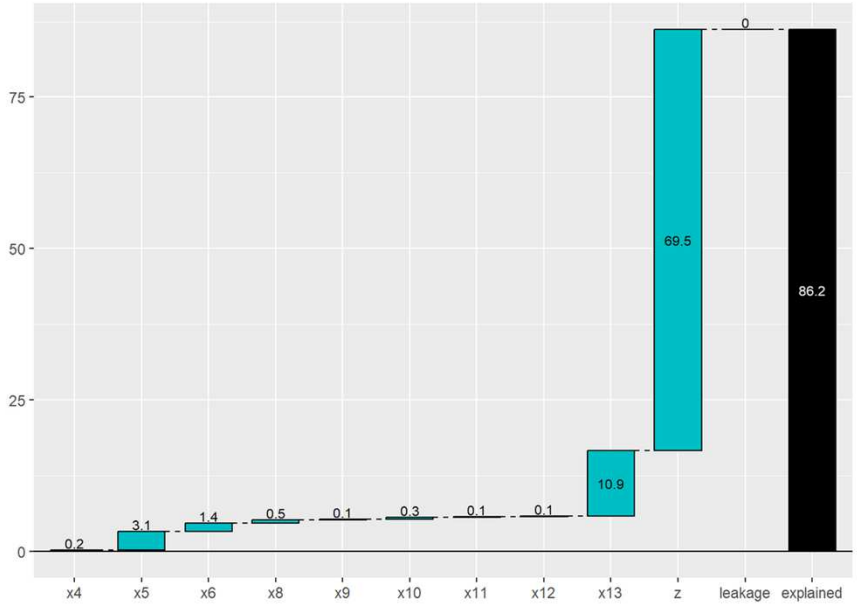
Erklärbarkeitsmethoden - Übersicht

- Es gibt inzwischen ein Standard-Repertoire von (etwa einem Dutzend?) Erklärbarkeitsmethoden.
- Dies sind mathematische Verfahren/Berechnungen, die typischerweise liefern:
  - Visualisierungen, d.h. Teile eines hoch-dimensionalen Modells werden auf 2-D, 3-D Sichten projiziert.
  - Kennzahlen, die z.B. die «Wichtigkeit» der Inputs beschreiben sollen.
  - In diesem Sinne sind die Methoden auch immer Vereinfachungen.
- Diese Standard-Methoden sind gut beschrieben
  - Christoph Molnar: https://christophm.github.io/interpretable-ml-book/
  - C. Lorentzen, M. Mayer «Peeking into the Black Box» https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3595944
  - Und natürlich die Notebooks und Ausarbeitungen unserer DAV-Arbeitsgruppe (erscheinen in Bälde)
- Viele der Methoden sind Modell-Agnostisch. D.h. eine bestimmte Methode funktioniert für alle Arten von Modellen.
  - Es gibt aber auch modellspezifische Ansätze, z.B. für baumbasierte Modelle.
- Erklärbarkeitsmethoden kommen inzwischen integriert in die Standardpakete für die jeweiligen Modelle.
  - Damit ist ein einfacher und niedrigschwelliger Zugang möglich.
  - Grundsätzlich kann die jeder sofort verwenden.

Erklärbarkeitsmethode: Varianzallokation

- Varianzallokation (Permutation Feature Importance, Sobol Index) beschreibt die Wichtigkeit der Inputvariablen für die Vorhersage.
  - Modell-agnostische Methode für Regressionsprobleme
  - Benötigt nur Vorhersagen des Modells keine interne Struktur
  - Praktikabel auch bei grossen Datenmengen und vielen Input-Variablen
- Grundidee: Ordne das R^2 bzw. den Mean Square Error des Modells den einzelnen Inputs zu.
  - Mit einem Schätzer bestimmt man, um wieviel sich das R^2 verschlechtert, wenn man jeweils einen bestimmten Input ignoriert.
  - Unter bestimmten Voraussetzungen sagt einem die Theorie dann, dass sich die einzelnen R^2 Werte zum Gesamtwert aufaddieren.
- Die Methode ist eine globale Erklärungsmethode.
- Nachteil der Methode: Setzt unabhängige Inputs voraus!

Beispiel: Erklärung eines internen Modells zur Berechnung der Solvenz.



Siehe
- «Neuronale Netze treffen auf Least Squares Monte Carlo» Link
- R-Script «main_effects» unserer DAV-Arbeitsgruppe (Veröffentlichung in Bälde)

Kritik der standardisierten Erklärbarkeitsmethoden

- Standard-Erklärbarkeitsmethoden sind leicht anwendbar und hilfreich. In vielen Fällen reichen sie aber nicht aus!
  - Beispiel: Weder lokale noch globale Erklärungen sind brauchbar, wenn man sich für die 1% Tail-Prognosen seines Modells interessiert.
  - Die Methoden sind technisch. Ein echtes Verständnis der Ergebnisse erfordert daher fachliche Kenntnis der Anwender.
  - Die Voraussetzungen zur Anwendung der Methoden sind oft sehr restriktiv. Z.B. funktionieren wichtige Methoden nur mit unabhängigen Inputs.
- Standard-Methoden können von problemspezifischen Ansätzen ablenken.
  - Im schlimmsten Fall: Alibi Effekt statt «Gehirn einschalten».
- Der Einsatz von Standard-Erklärungsmethoden scheint mir oft etwas schlampig zu sein. Es fehlt gelegentlich an «aktuarieller Sorgfalt» und wichtige Aspekte werden nicht berücksichtigt.
  - Was will ich mit der Erklärung erreichen?
  - Ist die Methode dokumentiert? Habe ich die Methode verstanden?
  - Sind die Voraussetzungen für den Einsatz der Methode erfüllt?
  - Wie kommuniziere ich die Ergebnisse und ihre Unsicherheit?

Kritik der standardisierten Erklärbarkeitsmethoden

- Die Interaktion von Menschen mit statistischen Modellen ist kein mathematisches Problem, sondern ein psychologisch/soziales Phänomen.
  - Diese Interaktion («Human centred XAI») ist noch wenig untersucht
  - Es scheint mir noch weitgehend unklar zu sein, wann und wie eine Erklärbarkeitsmethode Endanwendern wirklich hilft.
- Die Art der Erklärung und ihre Präsentation muss sich an den Bedürfnissen des Anwenders und der konkreten Anwendung orientieren.
  - Je nach Zielgruppe hat Erklärung einen anderen Zweck.
  - Entwickler: Test und Modellverbesserung
  - Entscheidungsträger: Zusammenhang der Erklärung mit der Entscheidung
  - Endanwender: Anfechtung einer Entscheidung oder Regressansprüche
  - Aufsicht, Audit: Compliance mit Richtlinien
- Erklärbarkeitsmethoden können sogar «falsches Vertrauen» erzeugen und so die Qualität von Entscheidungen verschlechtern.
  - In einer Studie [1] bewirkten schlecht oder falsch verstandene Erklärungen, dass Anwender dem Modell vertrauten und Fehler des Modells übersahen.

[1] «Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making», Zheng, Liao and Bellamy, Proceedings FACCT2020.

Agenda

- Statistische Modelle
- Warum braucht es Erklärbarkeit?
- Erklärbarkeitsmethoden
- Fazit

Rolle der Aktuare

- Aktuare haben ideale Voraussetzungen, statistische Modelle im Versicherungskontext zu verstehen und zu erklären.
  - Lange Tradition im Umgang mit Daten und statistischen Modellen.
  - Im Gegensatz zu Laien und Juristen haben sie ein mathematisches Verständnis.
  - Im Gegensatz zu reinen Machine Learning-Experten und Data Scientists verstehen sie das Fachgebiet.
- Durch ihre Ausbildung und Erfahrung können sie mehr als bloss Standard-Methoden aufrufen.
- Wichtige Aspekte rund um den Betrieb und Einsatz statistischer Modelle sind Aktuaren schon aus anderen Bereichen bekannt.
  - Z.B. Fragen und Anforderungen rund um die Nachvollziehbarkeit.
- Sie können (und sollten!) eine gewisse «aktuarielle Sorgfalt» in den Umgang mit statistischen Modellen einbringen.
  - Einordnen der Zweckmässigkeit von Modellen und Methoden,
  - Prüfen der Voraussetzungen und Grenzen,
  - Interpretation und Kommunikation der Ergebnisse.

Fazit:

- Die Erklärbarkeit von statistischen Modellen wird immer ein schwieriges Thema sein.
- Diese Schwierigkeit sollte aber niemanden davon abhalten diese Modelle einzusetzen.
- Erklärbarkeit ist nur ein Aspekt von vielen rund um den Einsatz von statistischen Modellen.
- Standardisierte Erklärungsmethoden sind hilfreich. Sie werden aber nicht alle Probleme lösen.
- Die Interaktion zwischen Menschen, Modellen und die Rolle der Erklärung dabei ist noch wenig verstanden.
- Aktuare sind dazu prädestiniert, statistische Modelle im Kontext ihrer Verwendung zu erklären.

- Letztendlich muss und kann man Grenzen der Erklärbarkeit akzeptieren. Denn was ist die Alternative?
  - Das menschliche Gehirn ist auch (momentan noch?) eine Blackbox.
  - Menschen sind notorisch schlecht darin, sich selbst und anderen Gründe für ihr Handeln zu nennen.
  - Wir haben gelernt, auch damit umzugehen.

Mitglieder der DAV-Arbeitsgruppe «Erklärbare KI»

Manuel Grbac
Guido Grützner
Simon Hatzesberger
Martin Hüttemann
Benjamin Müller
Zoran Nikolić
Janusch Rentenatus
Anja Schmiedt (Leitung)
Simon Steinbach
Corinna Walk
Florian Walla

Voraussichtliche Veröffentlichung
des Berichts: Im 1.Halbjahr 2024

Fragen Bemerkungen Kommentare ???

Fragen im Nachgang auch gerne an mich:

guido.gruetzner@quantakt.com