xAI / Explainable AI: Methoden und Techniken der xAI im Detail – die KI-Serie (Teil 3)

Das Fachgebiet der explainable AI (xAI) will schwer nachvollziehbare Entscheidungen von KI-Anwendungen transparent machen. Relativ einfache, logische Algorithmen lassen sich verständlich erklären, auch vor der ersten Verwendung eines Programms. Aber dieser sogenannte Ante-hoc-Ansatz scheitert spätestens bei Deep-Learning-Modellen an deren Komplexität. Hier bleibt nur die nachträgliche Erläuterung der gefundenen Lösung, also post-hoc. Ein Überblick über die wichtigsten Methoden der beiden xAI-Kategorien.

von Marc-Nicolas Glöckner, Senior Manager und Severin Bachmann, Consultant PPI

Marc-Nicolas Glöckner, Senior Manager PPI Quelle: PPI

Severin Bachmann ist Consultant bei PPI Quelle: PPI

Wie lauten wahrscheinlich die Lottozahlen der kommenden Woche? Wo bleibt die Kugel beim Roulette liegen? Der Mensch hat einen Hang dazu, selbst absolute Zufälle nachvollziehen oder gar vorhersagen zu wollen. Oder aber zumindest die Bestätigung zu erhalten, dass es sich eben um ausschließliche Fügungen des Schicksals handelt. Die Ergebnisse, zu denen Anwendungen im Bereich künstliche Intelligenz (KI) kommen, sind vieles, nur kein Zufall. Aber vorhersehbar und nachvollziehbar? Gerade bei hochkomplexen KI-Modellen wirkt der Algorithmus wie eine Black Box, die am Ende irgendein Ergebnis auswirft.

Gründe für das Ergebnis darlegen

Unter intransparenten Ergebnissen leidet die Akzeptanz der Technologie als solche, denn unverständliche Entscheidungen rufen schnell Reaktanz hervor. Und zwar bei allen Beteiligten: Betroffenen, Anwendern und am Ende auch dem Gesetzgeber. Schließlich muss Letzterer die Verbraucher vor Benachteiligungen schützen. Also gilt es, nicht nur Erklärungen zu finden, sondern diese auch verständlich darzulegen. Genau hier beginnt das Feld der explainable AI (xAI) als Teilgebiet der KI. Ihr Ziel: detailliert beschreiben, wie die verschiedenen Variablen die Prognosen von KI-Algorithmen in welcher Form und Intensität beeinflussen.

SERIE: explainable AI (xAI)

Teil 1: Mitentscheidend für den Erfolg von KI-Anwendungen: explainable AI

Teil 2: Vorhersagen eines KI-Modells einfach akzeptieren?

Teil 3: **Methoden und Techniken der xAI im Detail**

Ante-hoc für transparente Modelle

Für transparente KI-Modelle genügen sogenannte Ante-hoc-Ansätze. Diese Anwendungen sind so aufgebaut, dass die Betrachtung der jeweiligen Struktur und Daten bereits ausreicht, um die Funktion und am Ende auch die getroffene Entscheidung zu erklären. Beispiele sind Regressionsmodelle und Decision Trees. In vielen Fällen wird für die genauere Erklärung die Analyse der Trainingsdaten hinzugefügt. Diese gilt es, so gut wie möglich zu beschreiben.

Zum Einsatz kommen dabei unter anderem:

- die explorative Datenanalyse
- eine Standardisierung der Datensatzbeschreibung
- die Datensatz-Zusammenfassung
- erklärbares Feature Engineering

Darüber hinaus gibt es Forschungsprojekte, die zwar komplexere Black-Box-Algorithmen verwenden, während der Trainingsphase aber versuchen, das Lernen so zu monitoren, dass das resultierende Modell keine Black Box bleibt. Solche Versuche basieren vorwiegend auf Decision Trees und kleineren neuronalen Netzen. Allerdings stoßen alle eingesetzten Instrumente an ihre Grenzen, sobald wirklich komplexe Black-Box-Modelle, die in ihren Prognosen oftmals treffsicherer sind, erklärt werden sollen. Hier sind Post-hoc-Ansätze deutlich vielversprechender. Sie versuchen anhand des Ergebnisses zu erklären, wie das Modell zu diesem gelangt ist.

Post-hoc für komplexe Fälle

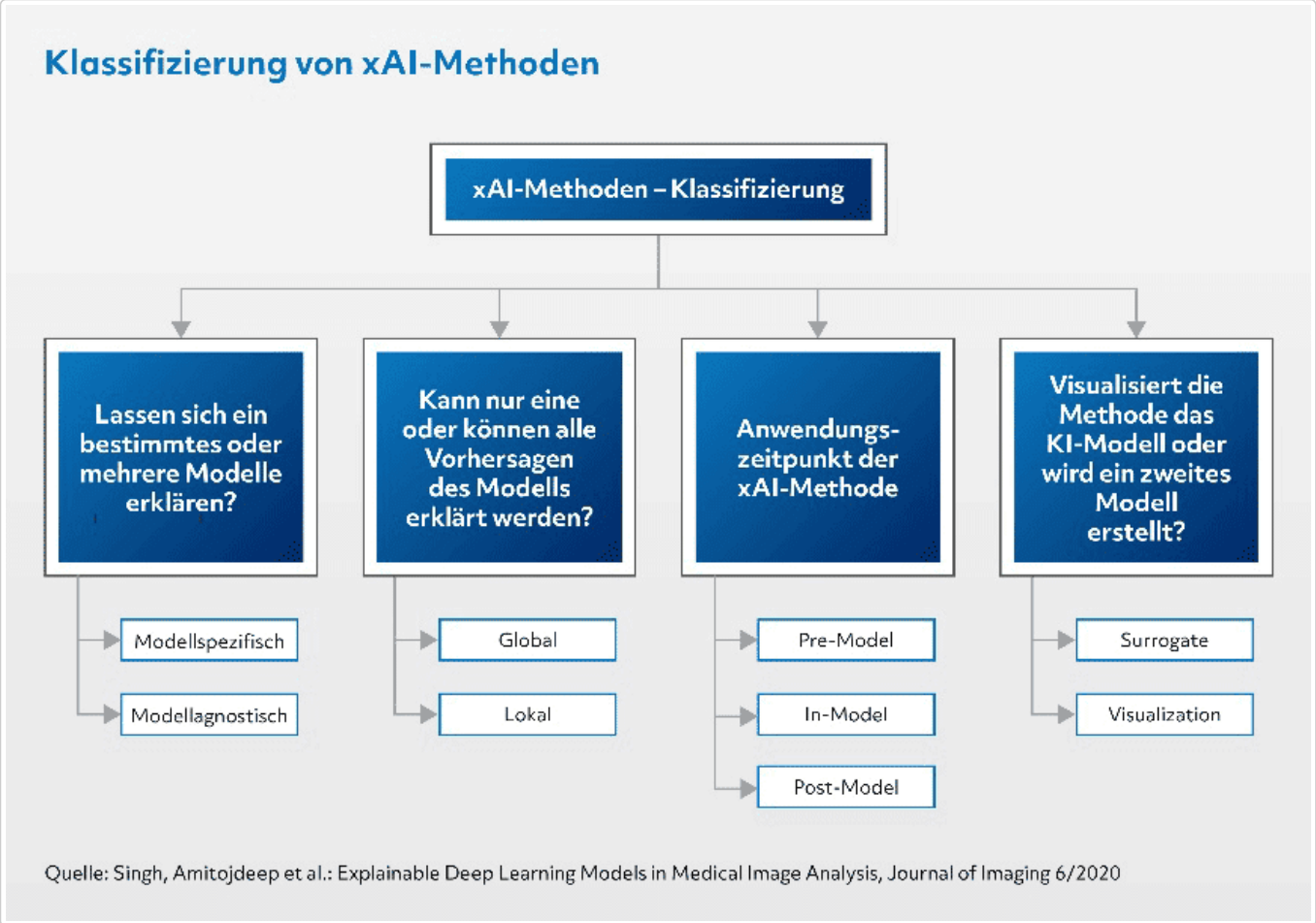
Generelles Merkmal aller xAI-Methoden im Bereich Post-hoc ist, dass sie erst nach der Erstellung des Modells auf dieses aufsetzen und die Entscheidungen transparent machen. Grundsätzlich lassen sich die unterschiedlichen Herangehensweisen in modellagnostische oder modellspezifische und lokale oder globale Techniken einteilen.

Global: Hier wird herausgearbeitet, welche Merkmale für das verwendete Modell insgesamt besonders wichtig sind und welche den größten Einfluss auf die Genauigkeit haben.

Lokal: Diese Erklärung beschreibt, wie ein Modell zu einer ganz bestimmten einzelnen Vorhersage gelangt ist.

Modellspezifisch: Diese Erklärungsansätze öffnen die Black Box und untersuchen die Modellinterna, sind aber aufgrund ihrer Spezifika immer nur auf ein bestimmtes Modell anzuwenden.

Modellagnostisch: Hier wird die Beziehung zwischen Input-Output-Paaren trainierter Algorithmen untersucht, ohne dabei auf die internen Strukturen einzugehen. Daher sind modellagnostische Techniken grundsätzlich auf alle KI-Modelle anwendbar, haben aber den Nachteil, kein Licht in die Black Box zu bringen.



Quelle: PPI

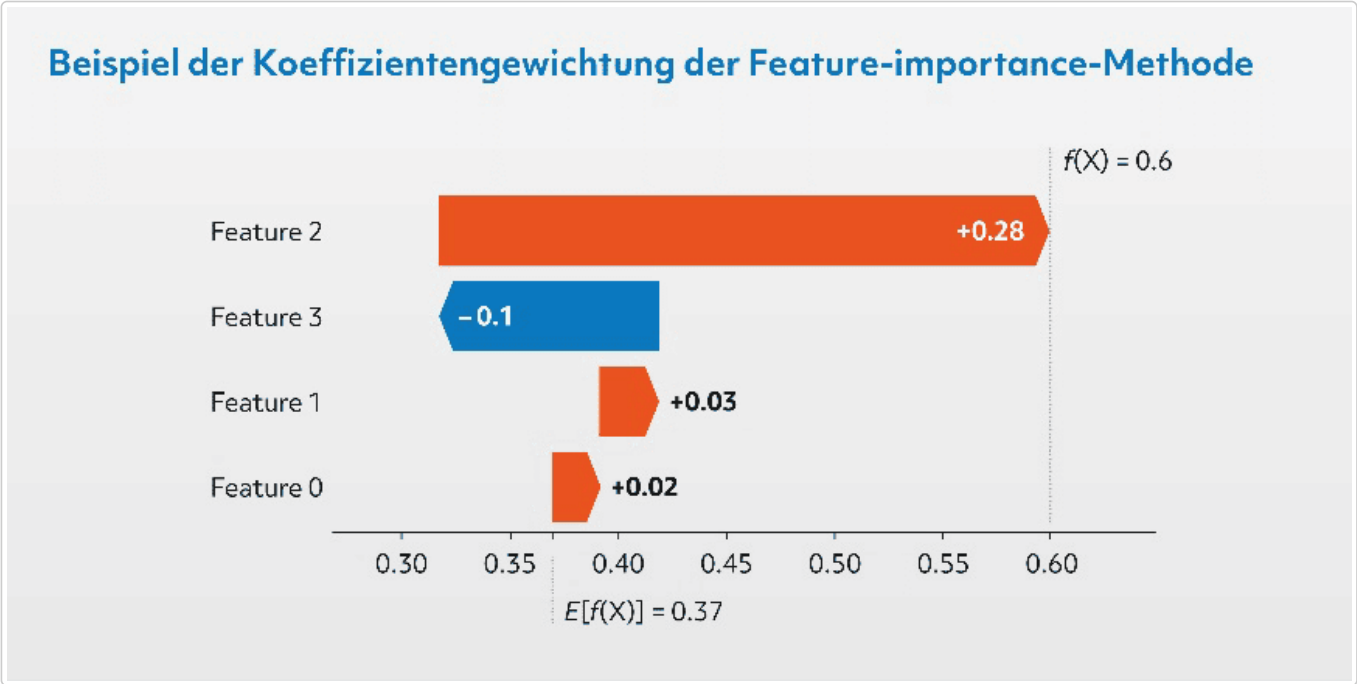
Im Folgenden werden einige Beispiele für Post-hoc-Methoden und ihre Einordnung in diese Taxonomie der xAI-Techniken betrachtet.

Surrogate Models

Hier wird ein zweites, vereinfachtes KI-Modell auf eine lokale Region der Eingabedaten des eigentlichen Modells trainiert, so dass es bei gleichem Input zu möglichst ähnlichen Ergebnissen kommt. Dieses Referenzmodell ist so transparent gestaltet, dass der Nutzer Akzeptanz und Vertrauen gegenüber den tatsächlichen Ergebnissen des komplexen Algorithmus zeigt. Surrogate Models sind immanent modellagnostisch – ein Beispiel ist der Trepan-Algorithmus, der einen Decision Tree als Approximationsmodell nutzt. Lokal kommen beispielsweise gewichtete lineare Regressionen um einen Point of Interest zum Einsatz. Surrogate-Modelle erreichen aufgrund ihrer Einfachheit nicht die Treffsicherheit des ursprünglichen Modells. Und selbst wenn, so bleiben sie eine Schätzung des echten Modells an einem ausgewählten Punkt.

Feature Importance

Die Messung der Wichtigkeit einzelner Variablen ist insofern schwierig, weil diese zumeist miteinander korreliert sind. Dieses Dilemma lässt sich vereinfachen, indem die einzelne Variable ausgeklammert und die sich ergebenden Auswirkungen gemessen werden. Daraus ergibt sich im Umkehrschluss die Bedeutung für das Gesamtergebnis. Modellagnostisch, aber vor allem lokal kommt die aus der Spieltheorie zur gerechten Bestimmung eines Gewinners stammende SHapley Additive exPlanation (SHAP) zum Einsatz. Hier werden einzelne Variablen für jede mögliche Kombination der anderen Faktoren ausklammert und der Durchschnitt der Ergebnisveränderungen berechnet. SHAP liefert sehr stabile Ergebnisse, ist aber extrem rechenintensiv.



Quelle: PPI

Ergebnisauszug einer Beispielberechnung von SHapley-Values. Die Darstellung zeigt, ob der vorhergesagte Wert durch Feature 0–3 positiv (rot) oder negativ (blau) beeinflusst wird.

Decision Rules
Hier werden mittels Entscheidungsregeln die Beziehungen zwischen Inputvariablen und Ergebnissen von KI-Modellen hergestellt. Die Regel ist dabei eine Funktion über die Inputmerkmale. Diese führt über eine geordnete Entscheidungsliste oder einen ungeordneten Entscheidungssatz zum Ergebnis. Etwa in einer Logik wie „Wenn Variable A gleich Zustand X und Variable B kleiner als Y, dann ist das Ergebnis Z“. Für echte Black-Box-Modelle ist diese Herangehensweise aber meist zu rechenaufwendig und die Beschreibung in Regeln für die echten Beziehungen nicht ausreichend. Die Anchor-Methode versucht, der wahren Komplexität durch das Extrahieren von lokalen Regeln entgegenzuwirken. Allerdings steigt mit der Zahl an lokalen Regeln natürlich wieder die Unübersichtlichkeit, und die Erklärbarkeit wird erschwert.

Counterfactual
Bei dieser Methode wird erklärt, welche Veränderungen einzelner Variablen nötig wären, um ein bestimmtes Ergebnis zu erzielen. Praktisches Beispiel ist eine KI-Kreditanwendung, welche einen Kundenantrag auf Basis der Faktoren Höhe des Gehalts, Dauer der Beschäftigung und Ausreizung des Dispos ablehnt. Bei der Counterfactual-Erklärung wird hierzu eine Aussage getroffen, was an diesen Größen geändert werden müsste, um eine positive Kreditentscheidung zu erhalten – also beispielsweise die Dispo-Ausnutzung zu reduzieren.

Breites, hochdynamisches Forschungsfeld

Es existiert eine ganze Reihe vielversprechender Ansätze, unter denen sich aber zumindest derzeit kein zukünftiger Gold-Standard abzeichnet. Große Player wie Google haben sich einzelnen Methoden angenommen und sind dabei, diese zu einer möglichst unkomplizierten Anwendung weiterzuentwickeln. Dabei geht es auch darum, am Ende zu Out-of-the-box-Lösungen zu kommen oder alternativ zumindest den Programmieraufwand in Grenzen zu halten. Für jeden Einzelfall klare, sofort eingängige Erklärungen wird aber kein Tool liefern können, die Ergebnisse bedürfen immer der Interpretation innerhalb der jeweiligen methodischen Grenzen. Auch bleibt jedes Mal die grundsätzliche Frage zu beantworten: Werden die Daten

erklärt oder aber das Modell? Daraus ergibt sich immanent die Grenze von xAI und damit auch von KI. Für die Masse der Modelle müssen am Ende aber standardisierte Methoden verfügbar sein, mit denen die Anwender die Ergebnisse mit vertretbarem Aufwand schnell interpretieren und nachvollziehbar überprüfen können.

Marc-Nicolas Glöckner und Severin Bachmann, PPI ∎