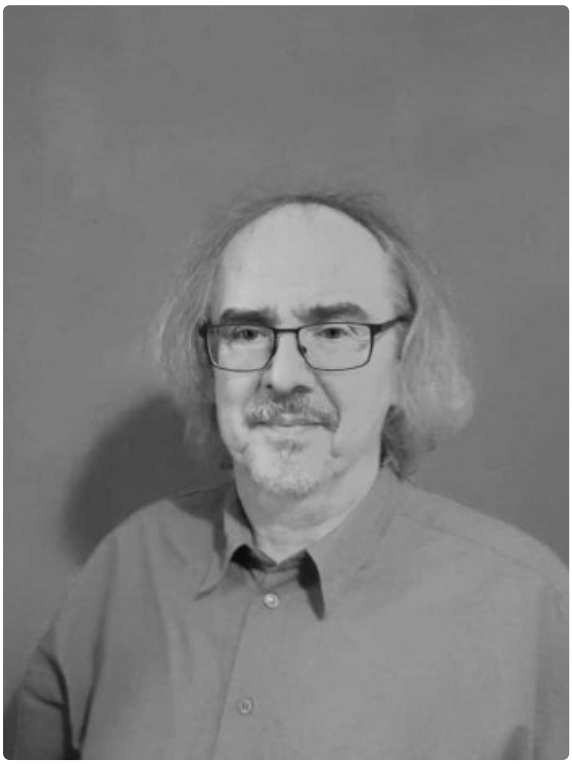
xAI / Explainable AI: Vorhersagen eines KI-Modells einfach akzeptieren? – die KI-Serie (Teil 2)

Vom Einsatz moderner Technologien wie der künstlichen Intelligenz (KI) versprechen sich Finanzdienstleister schlankere Prozesse, bessere Prozessergebnisse und dadurch langfristig Steigerungen von Umsatz und Gewinnmarge. Dabei sollten aber die Risiken nicht außer Acht bleiben. Für die richtige Anwendung von KI-Modellen und um fehlerhafte Ergebnisse erkennen und korrigieren zu können, ist ein gewisses Verständnis für komplexe Algorithmen notwendig. Aus regulatorischer Sicht ist dies sogar unbedingte Voraussetzung für die Zulässigkeit von KI-Entscheidungen.

von Dr. Stefan Rieß ist Leading Software Engineer und Dr. Paul Günther ist Managing Consultant PPI

Dr. Paul Günther ist Managing Consultant bei PPI Quelle: PPI



Dr. Stefan Rieß ist Leading Software Engineer bei PPI

Quelle: PPI

Hat die britische Comedy-Serie „Little Britain“ die Zukunft vorweggenommen? Dort können Zuschauer immer wieder erleben, wie die Sachbearbeiterin einer Bank beziehungsweise eines Reisebüros die Anfrage eines Kunden in ihren Computer tippt und nach minimaler Wartezeit mit Verweis auf die IT ablehnt. Immer häufiger durchlaufen Anliegen von Bankkunden im Verlauf der Bearbeitung eine KI-Anwendung. Deren Votum ist dann ein mitentscheidendes Kriterium bei Kreditanfragen, Kartenanträgen, Kontoeröffnungen oder Ähnlichem. Das ist mit ein Grund, warum der Gesetzgeber im Begriff ist, Regularien zum Einsatz von KI im Bereich Finanzdienstleistungen zu etablieren.

Risikobasierter Ansatz des Gesetzgebers

Die Europäische Kommission hat das Thema in ihrem Entwurf für eine „EU-Verordnung zur Regulierung künstlicher Intelligenz (KI)“ berücksichtigt. Dieser soll einen übergreifenden, risikobasierten Ansatz für den KI-Einsatz sowohl in der Wirtschaft als auch in der öffentlichen

Verwaltung darstellen. Die Prämisse: Je höher das Risiko negativer Auswirkungen einer KI-Entscheidung ist, desto stärker greift der Gesetzgeber ein, bis hin zu einer Untersagung von KI-Anwendungen. Grob umrissen ergeben sich folgende Zuordnungen:

- Verbot von KI-Anwendungen mit extremen Risiken
- Regulierung bei KI-Anwendungen mit hohen Risiken
- Verpflichtung zu Transparenz bei Systemen mit geringen Risiken
- Selbstregulierung der jeweiligen Körperschaften bei risikolosen Anwendungen

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

Prüfungen auf Kreditwürdigkeit werden nach dem Verordnungsentwurf als Anwendungen mit hohem Risiko klassifiziert. Um sie einsetzen zu dürfen, ist eine höchstmögliche Transparenz der getroffenen Entscheidungen erforderlich. Ist eine solche nicht gegeben oder verstößt die Anwendung anderweitig gegen die EU-Vorschrift, sieht diese Bußgelder von bis zu sechs Prozent des Jahresumsatzes vor, maximal jedoch 30 Millionen Euro. Für die entsprechende Überwachung und Durchsetzung sind die nationalen Aufsichtsbehörden zuständig, in Deutschland die Bundesanstalt für Finanzdienstleistungsaufsicht (BaFin).

Nationale Vorgaben existieren bereits

Nicht zuletzt durch die Aktivitäten der EU getrieben haben BaFin und Bundesbank bereits Vorgaben für den KI-Einsatz veröffentlicht. Diese finden sich im Papier „Big Data und künstliche Intelligenz: Prinzipien für den Einsatz von Algorithmen in Entscheidungsprozessen“. Die Vorgaben sind deutlich weniger streng als die Überlegungen der EU-Kommission. Aber auch hier ist klar herauszulesen, dass die Banken gehalten sind, die Qualität ihrer Datengrundlage fortlaufend zu überwachen, um weitgehend auszuschließen, dass Verzerrungen oder ein Bias die Ergebnisse der jeweiligen Anwendung beeinflussen. Dies gilt nicht nur für die Trainingsdaten, sondern für alle auch zukünftig gesammelten Informationen.

Bias verzerrt Ergebnisse

Das Papier der BaFin gibt bereits klare Hinweise auf die häufigsten Probleme bei der Verwendung von KI-Algorithmen. Ein Bias, englisch für Vorurteil oder Verzerrung, entsteht durch Fehler in der Datenerhebung. Zur Illustration: Wird etwa ein Modell, das allgemeingültige Aussagen über den Bevölkerungsdurchschnitt liefern soll, mit den Ergebnissen einer ausschließlich an Vormittagen per Festnetz durchgeführten Telefonumfrage trainiert, sind die Ergebnisse nicht verwertbar. Denn alle Personen, die keinen Festnetzanschluss haben oder zum fraglichen Zeitpunkt erwartbar nicht zu Hause sind, fallen durch das Raster der Untersuchung. Die dadurch bedingte Verzerrung von Alters- und Sozialstruktur verhindert repräsentative Ergebnisse.

Testdaten und Anwendung müssen zusammenpassen

Ein ähnlicher Effekt mit anderer Ursache tritt ein, wenn die Testdaten für das Anlernen des Modells und die späteren tatsächlichen Anwendungsfälle fachlich nicht zusammenpassen. Ein praktisches Beispiel wäre ein Entscheidungsmodell für die Kreditvergabe, das mit Datensätzen folgender Kundengruppen trainiert wird:

Frauen im Alter von 20 bis 40 Jahren und Kredithöhen von 50.000 bis 100.000 Euro

40- bis 60-jährige Männer mit Kredithöhen unter 50.000 Euro

Wird diese KI-Anwendung für die Risikoeinschätzung bei einer 50-jährigen Frau mit einem Kreditwunsch von 30.000 Euro verwendet, so wird das Ergebnis vermutlich falsch sein. Denn das Modell arbeitet in diesem Fall außerhalb seines Gültigkeitsbereiches, da die verlangte Vorhersage vom Training nicht erfasst wird.

Dieses Beispiel macht eindrücklich klar, wie wichtig es ist, die Funktionsweise und Grenzen von KI zu verstehen und diese auch jedem Anwender zu vermitteln. Denn während der Softwareingenieur und auch die IT-Abteilung nachvollziehen können, in welchen Grenzen der Algorithmus zuverlässig arbeitet, ist dies dem Sachbearbeiter in der Kreditabteilung nicht ersichtlich. Da Menschen meist kein intuitives Verständnis von Statistiken und Wahrscheinlichkeiten haben, auf denen KI aber grundsätzlich aufsetzt, können die Ergebnisse eines KI-Modells der Intuition des Bearbeiters auch einmal zuwiderlaufen. Ohne Erklärung der Resultate besteht dann die Gefahr, dass die Akzeptanz für den Einsatz der Technologie als Ganzes bei allen Beteiligten schwindet. Zudem haben Betroffene ein Recht auf Auskunft, warum eine Entscheidung so gefällt wurde.

Know your algorithm

Daraus ergibt sich logischerweise die Forderung nach einer Erklärbarkeit der KI-Ergebnisse. Um die Qualität der getroffenen Vorhersagen einzuschätzen, müssen Anwender die Eigenheiten des Algorithmus kennen und verstehen. Gerade bei hochkomplexen Anwendungen aus dem Bereich Deep Learning stößt das normale menschliche Abstraktionsvermögen aber an Grenzen. Hier liefert das Forschungsfeld der explainable AI (xAI) Methoden und Techniken, sich mit den eigenen Daten vertraut zu machen und die Zusammenhänge der Einzeldaten untereinander zu verstehen. Zur Analyse des Einzelfalls lassen sich dann die entsprechenden xAI-Verfahren heranziehen, so dass am Ende eine nachvollziehbare Begründung ähnlich der durch einen Menschen gegebenen vorliegt.

KI engmaschig überwachen

KI-Algorithmen sind keine Allheilmittel. Es gibt durchaus Fragestellungen, bei denen trotz aller Sorgfalt bei der Datenerhebung wichtige Entscheidungskriterien fehlen. Umso wichtiger ist das Verständnis für die tatsächliche Funktionsweise der Modelle. Die dahingehenden regulatorischen Forderungen sind nicht nur verständlich, sie stellen auch eine Chance für Finanzdienstleister dar, die eigenen KI-Anwendungen kritisch zu hinterfragen. Denn ein Algorithmus auf diskriminierender Testdatenbasis ist nicht nur ein ethisches und rechtliches Problem, sondern wird auch das Geschäft negativ beeinflussen. Schließlich verhindern verfälschte Vorhersagen positive Entscheide oder schließen gar bestimmte Kundenkreise aus. Die notwendige Überwachung der Prozesse und Resultate gehört jetzt und in Zukunft zu den Aufgaben einer verantwortungsvoll handelnden Bank. Hilfestellung leisten dabei die Methoden der xAI, auf die der letzte Teil der Serie detailliert eingehen wird.

Dr. Stefan Rieß und Dr. Paul Günther, PPI ∎

Im dritten Teil unserer Serie ‘Explainable AI (xAI)’ beleuchten wir verschiedene Methoden zur Erläuterung von KI-Entscheidungen, sowohl vor als auch nach ihrer Anwendung, und zeigen die Notwendigkeit, dass diese Erklärungen verständlich und nachvollziehbar sein müssen, um die Akzeptanz von KI-Technologie zu erhöhen.