



Ein Blick in die Black-Box: Explainable AI (XAI) erklärt

Verena Barth, Wüntsche & Barth, Tobias Goerke, viadee Unternehmensberatung

Machine-Learning-Modelle (ML-Modelle) finden in einem enormen Tempo Einzug in unseren Alltag und werden für Prognosen, datengetriebene Entscheidungen oder Generierung von Inhalten genutzt. Um die gewünschte Genauigkeit der Vorhersagen zu erreichen, werden statt menschenverstndlicher Entscheidungssysteme oftmals komplexe Verfahren wie tiefe neuronale Netze (DNNs) verwendet, deren Entscheidungen aufgrund ihrer inhrenten Komplexitt unverstndlich und nicht nachvollziehbar sind. Der Bereich der Explainable AI (XAI) versucht, das Problem fehlender Transparenz von ML-Modellen zu adressieren und die Ergebnisse fr den Menschen verstndlich zu machen. In diesem Artikel wird das Potenzial von XAI erschlossen und einige exemplarische XAI-Methoden kurz vorgestellt. Im Anschluss werden Kriterien aufgefhrt, die bei der Auswahl einer passenden XAI-Methode zu beachten sind. Abschlieend werden allgemeine Handlungsempfehlungen fr die Sicherstellung interpretierbarer ML-Modelle innerhalb eines Entwicklungs- und Deployment-Workflows gegeben.

Warum XAI? Die Black-Box-Problematik

Für komplexe Aufgaben oder zur Verarbeitung vieldimensionaler Eingabeparameter wie Bild- oder Textdaten, kommen häufig Black-Box-ML-Modelle zum Einsatz. Im Gegensatz zu White-Box-Modellen, wie zum Beispiel linearen/logistischen Regressionen oder kurzen Entscheidungsbäumen, besitzen sie meist einen effizienteren Lernalgorithmus oder erreichen eine höhere Genauigkeit der Vorhersagen durch das Finden komplexerer Entscheidungsgrenzen. Solche Black-Boxes, wie unter anderem Random Forrests, Support Vector Machines (SVMs) und DNNs, sind im Gegensatz zu intrinsisch interpretierbaren White-Box-Modellen aufgrund ihrer Komplexität nicht verständlich und ihre Entscheidungen daher nicht nachvollziehbar.

Trotz ihrer beeindruckenden Leistung zeigen viele Beispiele leider immer wieder die Unvollkommenheit bereits eingesetzter KI-Systeme auf. Diese reichen von geschlechterspezifischen Stereotypen bei der Verarbeitung natürlicher Sprache [6], über die Benachteiligung von Frauen beim automatisierten Einstellungsprozess bei Amazon [7] bis hin zu rassistischen Tendenzen bei einem in den USA eingesetzten Algorithmus, der das Strafmaß durch die Vorhersage der Wahrscheinlichkeit einer erneuten Straftat bestimmt und dunkelhäutige Menschen benachteiligt [2]. Es existieren zahlreiche Sammlungen dokumentierter Fehler von in der realen Welt eingesetzten KI-Systemen; die AI Incident Database (AIID) hat bereits über 1200 Einträge riskanter Zwischenfälle erfasst [17].

Diese Voreingenommenheit oder Diskriminierung des Modells kann vielseitige Ursachen haben wie unangemessene Evaluationsmetriken des Modells, falsche Annahmen über die Daten oder die Verzerrungen (Bias) in ihnen, die oft historische und sozio-technische Probleme der Gesellschaft widerspiegeln. [14]

Regierungen erkennen diese Problematik und fordern Transparenz in KI-Systemen. So stellt die Datenschutzgrundverordnung (DSGVO) Anforderungen an Systeme mit automatisierten Entscheidungsfindungen und der voraussichtlich 2026 inkrafttretende Artificial

Intelligence Act (AI Act) ordnet KI-Anwendungen in Risikoklassen ein.

Angesichts bestehender Probleme und der zunehmenden produktiven Verwendung von ML-Modellen ist es dringend erforderlich, das Problem der mangelnden Transparenz und Nachvollziehbarkeit anzugehen.

Was ist Explainable AI und warum sollte man es anwenden?

Hier kommt Explainable AI ins Spiel: Dieses Forschungsfeld versucht die Interpretierbarkeit von und das Vertrauen in ML-Modelle zu fördern, ohne ihre (Lern-)Leistung einzuschränken. Eine Anwendung von XAI ist notwendig, da eine Erklärung, beziehungsweise Rechtfertigung, der Verhaltensweisen und Entscheidungen für die Nachvollziehbarkeit, Fairness und Sicherheit der Modelle essenziell ist.

Der Erhalt von Erklärungen der Vorhersagen ist zudem hilfreich bei der Fehlerdiagnose im Modell und bei der Identifikation von Verzerrungen in den Daten. Dadurch wird das Modell verbessert und robuster gegenüber Schwachstellen und Angriffen gemacht; außerdem unterstützt es bei der Erkenntnisgewinnung in der Problemdomäne. Damit (personenbezogene) Entscheidungen gerechtfertigt werden können, ist eine Nachvollziehbarkeit je nach Land und Domäne gesetzlich vorgeschrieben.

Erklärbarkeit und Rechtfertigung der Entscheidungen werden umso wichtiger, je komplexer das ML-Modell ist oder je kritischer der Anwendungskontext. Da transparente Modelle intrinsisch interpretierbar sind, können sie als „Explainable AI“ verstanden werden. Nachfolgend fokussieren wir uns auf XAI-Methoden für (Black-Box)-ML-Modelle beliebiger Komplexität, die nachträglich, post-hoc angewendet werden.

Taxonomie der XAI-Methoden

XAI-Methoden unterscheiden sich anhand der Modellart, auf die sie angewendet werden können, und anhand des resultierenden Erklärungsumfangs und -formats:

Es gibt modellspezifische und modellagnostische XAI-Methoden. Modellspezi-

fische sind auf eine bestimmte Modellart beschränkt, während modellagnostische post-hoc, das heißt nach dem Training, auf jeden Modelltypen angewendet werden können. Da modellagnostische Methoden keinen Zugriff auf die Modellinternen haben, liefern sie Erklärungen nur anhand einer Analyse der Ein- und Ausgaben. [16]

XAI-Methoden können zudem nach dem Umfang der erhaltenen Erklärung klassifiziert werden, wobei zwischen globalen und lokalen Methoden unterschieden wird. Lokale Methoden erklären die Gründe einer Modellentscheidung spezifisch für eine einzelne Dateninstanz, die nicht auf eine globale Skala generalisiert werden können. Globale Erklärungen zielen dagegen auf ein Verständnis des Verhaltens des Gesamtsystems ab.

XAI-Methoden lassen sich bezüglich ihrer Erklärungsansätze grob in vier Kategorien einteilen [7]:

1. Visuelle Erklärungen vereinfachen das Verständnis eines Modells durch Visualisierung und eignen sich dadurch auch für Menschen ohne Expertenwissen.
2. Eine Erklärung durch Feature-Relevanz quantifiziert den Einfluss jedes Eingabe-Features auf die Modellvorhersage, indem die Auswirkung der Änderung seines jeweiligen Werts auf das Vorhersageergebnis oder die Modell-Leistung (z. B. die Vorhersagegenauigkeit) beobachtet wird.
3. Bei einer Erklärung durch Wissensextraktion/Vereinfachung werden entweder Regeln für den Entscheidungsprozess unter der Verwendung der Ein- und Ausgaben konstruiert oder das Black-Box-Modell mit einem transparenten, interpretierbaren Modell approximiert.
4. Für eine beispielbasierte Erklärung und den Erhalt eines besseren Modell-Verständnisses werden repräsentative Dateninstanzen des Trainingsdatensatzes ausgewählt. Diese zeigen die inneren Beziehungen und Korrelationen des Modells auf.

Wie funktionieren XAI-Methoden?

XAI-Methoden operieren grundsätzlich mit einem von zwei Ansätzen:

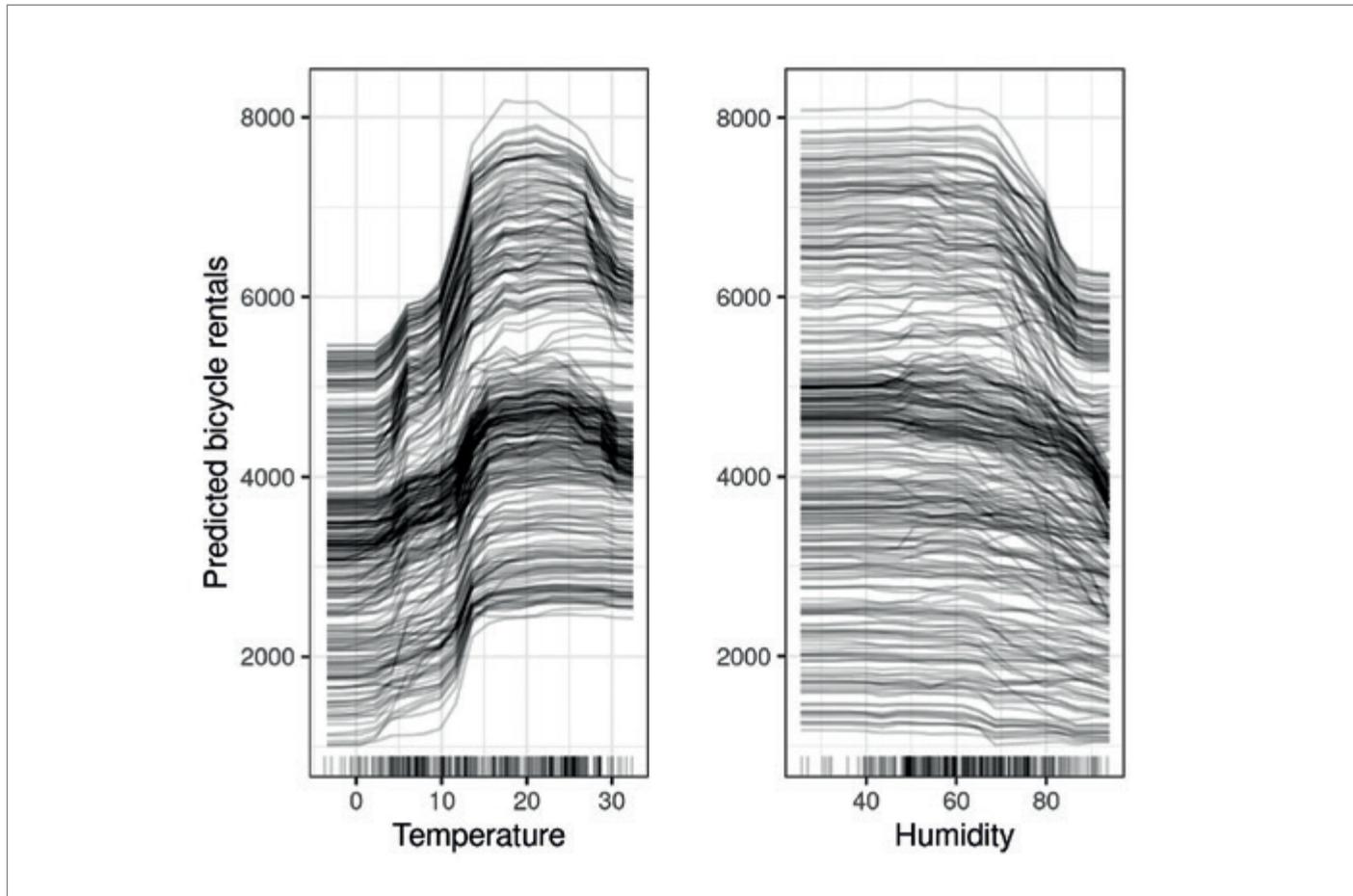


Abbildung 1: ICE der Features „Temperature“ und „Humidity“ am Beispiel des Bike Rental Datasets [16]

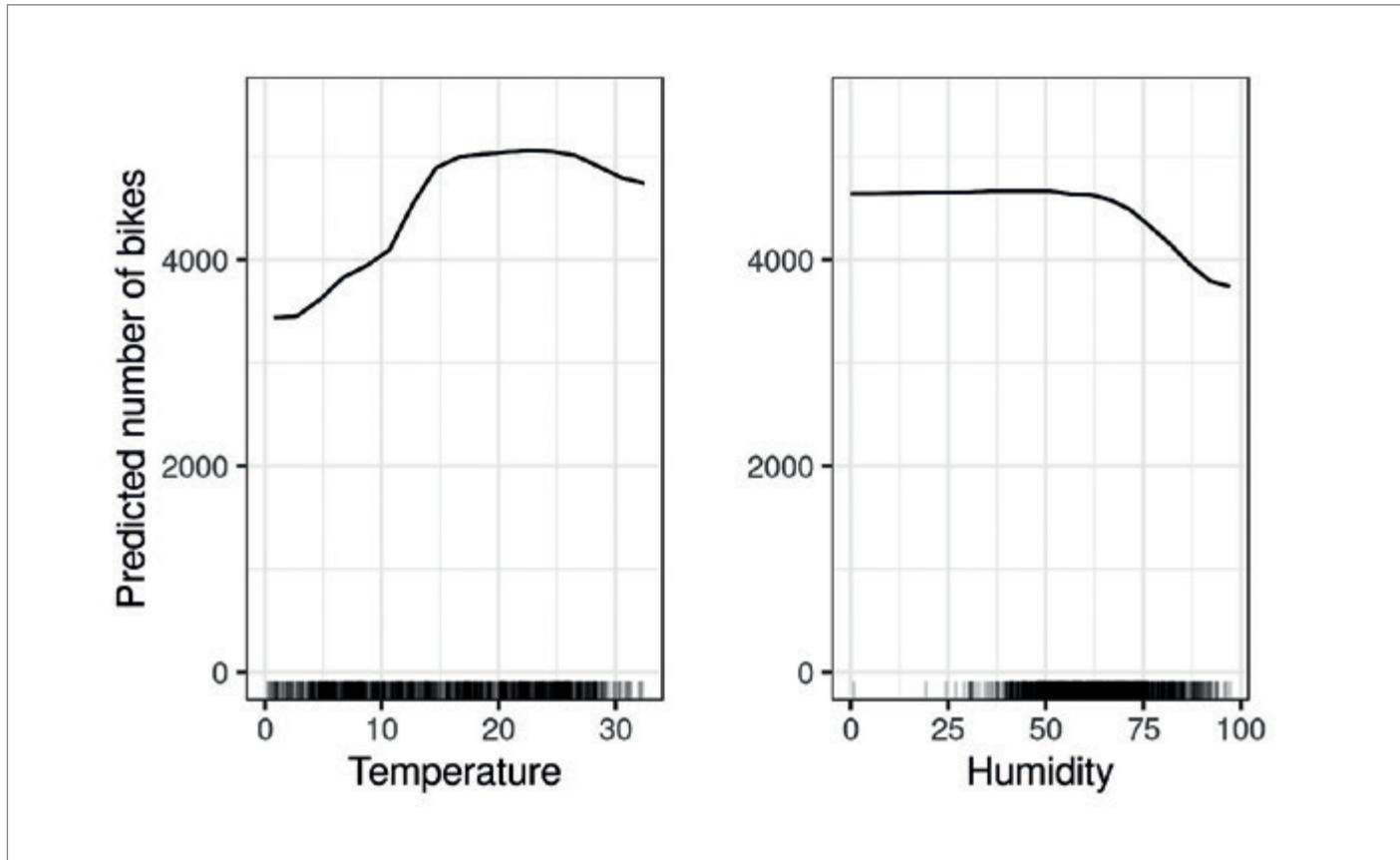


Abbildung 2: PDP der Features „Temperature“ und „Humidity“ am Beispiel des Bike Rental Datasets [16]

Gradientenbasierte Methoden verwenden eine Abwandlung des Backpropagation-Algorithmus und den Gradienten der Ausgabe in Bezug auf die Eingabe, um die wichtigsten Merkmale der Eingabe zu ermitteln. Dabei unterscheiden sie sich hauptsächlich durch die Art der Berechnung des Gradienten und werden vor allem für XAI bei Bilddaten in Form von Saliency Maps verwendet.

Perturbationsbasierte Methoden erklären die Vorhersage eines Modells durch verschiedene durch Perturbationen erzeugte Variationen des Eingabe-Feature-Raums (Störungsmodell). Einfach gesagt: Sie verändern die Modelleingabe und schauen, wie sich das auf die Vorhersage auswirkt. Da sie für die Erklärungsgenerierung keinen Zugriff auf interne Informationen des Modells benötigen, sind sie modellagnostisch und eignen sich für Black-Box-Modelle.

Um modellagnostisch zu bleiben, konzentrieren wir uns im Folgenden auf perturbationsbasierte Methoden.

XAI erklärt

Anhand des *Partial Dependence Plot (PDP)* und der *Individual Conditional Expectation (ICE)* wird die Funktionsweise perturbationsbasierter Methoden erklärt. Anschließend werden die beliebten, exemplarisch ausgewählten Methoden *Shapley Additive Explanations (SHAP)* und *Anchors* kurz vorgestellt und gezeigt, sie dazu beitragen, die „Black-Box-Natur“ von KI-Modellen aufzubrechen und Einblicke in die Entscheidungsfindung zu gewähren. Sie verdeutlichen außerdem Probleme, unter welchen die Qualität der XAI-Erklärungen öfters leidet.

Die für tabellarische Daten anwendbaren XAI-Methoden Partial Dependence Plot und Individual Conditional Expectation erklären beide visuell die Entscheidung des Modells anhand der marginalen Einflüsse einzelner Eingabe-Features und beantworten somit die Frage: „Was ist der Zusammenhang zwischen dem betrachteten, unabhängigen Feature und der Vorhersage?“

PDP visualisiert den globalen Einfluss eines bestimmten Features (am Beispiel des Bike Rental Datasets [8]): „Welchen durchschnittlichen Einfluss hat die Temperatur auf die Anzahl der vermieteten Fahr-

räder?“), während sich ICE auf eine spezifische Dateninstanz fokussiert („Welchen Einfluss hatte die Temperatur vorgestern auf die Anzahl der vermieteten Fahrräder des Tages?“). Beide Methoden generieren Vorhersagen, indem sie den Wert des Features „Temperatur“ verändern während alle anderen Feature-Werte unverändert bleiben. So wird der marginale Effekt isoliert und man erhält Aufschluss darüber, wie sich das Feature auf die Vorhersage auswirkt.

Der ICE-Plot (siehe Abbildung 1) zeigt ähnliche Auswirkungen der Temperatur auf die Fahrradvermietungen über verschiedene Tage, während der globale PDP-Plot (siehe Abbildung 2) den durchschnittlichen Effekt für alle Instanzen darstellt und den Trend von ICE bestätigt: Wenn die Temperatur über 15 Grad ist, werden mehr Fahrräder vermietet, wenn die Luftfeuchtigkeit hoch ist, nimmt die Anzahl an Vermietungen ab.

Da der Feature-Effekt isoliert betrachtet wird, sollten die Eingabe-Features nicht zu stark korrelieren, da andernfalls seltene, unrealistische (oder sogar unmögliche) Dateninstanzen bei der Erklärungserzeugung berücksichtigt werden könnten. Das mindert die Qualität der resultierenden Erklärung oder verfälscht diese. Ein Beispiel hierfür: Sollte die Dateninstanz eine Person darstellen, würde die Veränderung des Features „Gewicht“ auf „41 kg“ eine unmögliche Perturbation ergeben, wenn das Feature „Körpergröße“ den Wert „200 cm“ behält.

Eine weitere populäre XAI-Methode ist Shapley Additive Explanations, die auf einem häufig zitierten Paper [13] basiert und viele verschiedene Erklärungsformate bietet. SHAP beantwortet die Frage, welchen Beitrag der Wert eines oder mehrerer Features zur Vorhersage verglichen mit der durchschnittlichen Vorhersage liefert, zum Beispiel: „Inwieweit wurde meine Vorhersage für die Kredithöhe durch die Tatsache beeinflusst, dass ich vier Bankkonten habe, statt nur der durchschnittlichen Anzahl von zwei Konten?“

Die Methode basiert auf der spieltheoretischen Idee, dass das Vorhersageergebnis fair unter allen Features aufgeteilt wird und, dass für die Bestimmung der Wichtigkeit eines einzelnen Features (Shapley Werte), alle Feature-Kombinationen berücksichtigt werden sollten. SHAP generiert eine Vielzahl an Plots: So lassen

sich sowohl fallspezifische als auch globale Feature Importance Plots (mit der durchschnittlichen Wichtigkeit einzelner Features) erstellen, indem die Shapley Werte einzelner Dateninstanzen kombiniert werden. Die Abbildungen 3 und 4 zeigen Plots am Beispiel des Adult Census Income Datasets [4], das Personen basierend auf ihren Charakteristika klassifiziert, ob sie mehr oder weniger als 50.000 US-Dollar pro Jahr verdienen. In Abbildung 3 sind die für die Vorhersage wichtigsten Features aufgeführt (Feature Importance Plot) und man sieht, dass sie für die Klassifikation beider Klassen etwa gleich ausschlaggebend sind.

(Globale) Dependency Plots wie in Abbildung 4 zeigen, welchen Einfluss der Feature Wert auf die Ausgabe des Modells hat. Sie sind ähnlich wie PDPs, wobei sie nur den möglichen Eingaberaum berücksichtigen (aber Korrelationen in den Eingabedaten trotzdem ignorieren). Durch die vertikale Varianz des Streudiagramms vermitteln sie außerdem die Größe der vorherrschenden Interaktionseffekte der Features. Eine Interpretation dieser visuellen Erklärung ist für Personen ohne datenwissenschaftliche Kenntnisse oftmals nicht trivial, da sie viele Informationen beinhaltet (wie zum Beispiel die Feature Werte, die Varianz der Datenpunkte eines Feature und die Wichtigkeit des Features auf die Vorhersage „Einkommen > 50.000“).

Anchors [18] ist eine weitere beliebte perturbationsbasierte XAI-Methode, welche die kritischsten Bedingungen („Anchors“) identifiziert, die erfüllt sein müssen, damit die Vorhersage eines Modells gültig ist. Sie zerlegt komplexe Modellentscheidungen in leicht verständliche Regeln, die auf möglichst viele Instanzen des Datensatzes zutreffen („Coverage“) und dabei eine möglichst hohe Wahrscheinlichkeit aufweisen, dass ihre Erfüllung zu der prognostizierten Modellentscheidung führt („Precision“).

„Education = Bachelors AND Relationship = Husband AND Occupation = Sales. Precision: 0.95, Coverage 0.02“: Dieses Anchors-Beispiel einer Dateninstanz des Income-Datensatzes mit positiver Vorhersage (> 50 000 Einkommen) bedeutet, dass Personen, auf die diese Regel zutrifft, zu 95% auch eine positive Vorhersage erhalten werden. Leider ist diese Regel sehr spezifisch und trifft mit ei-

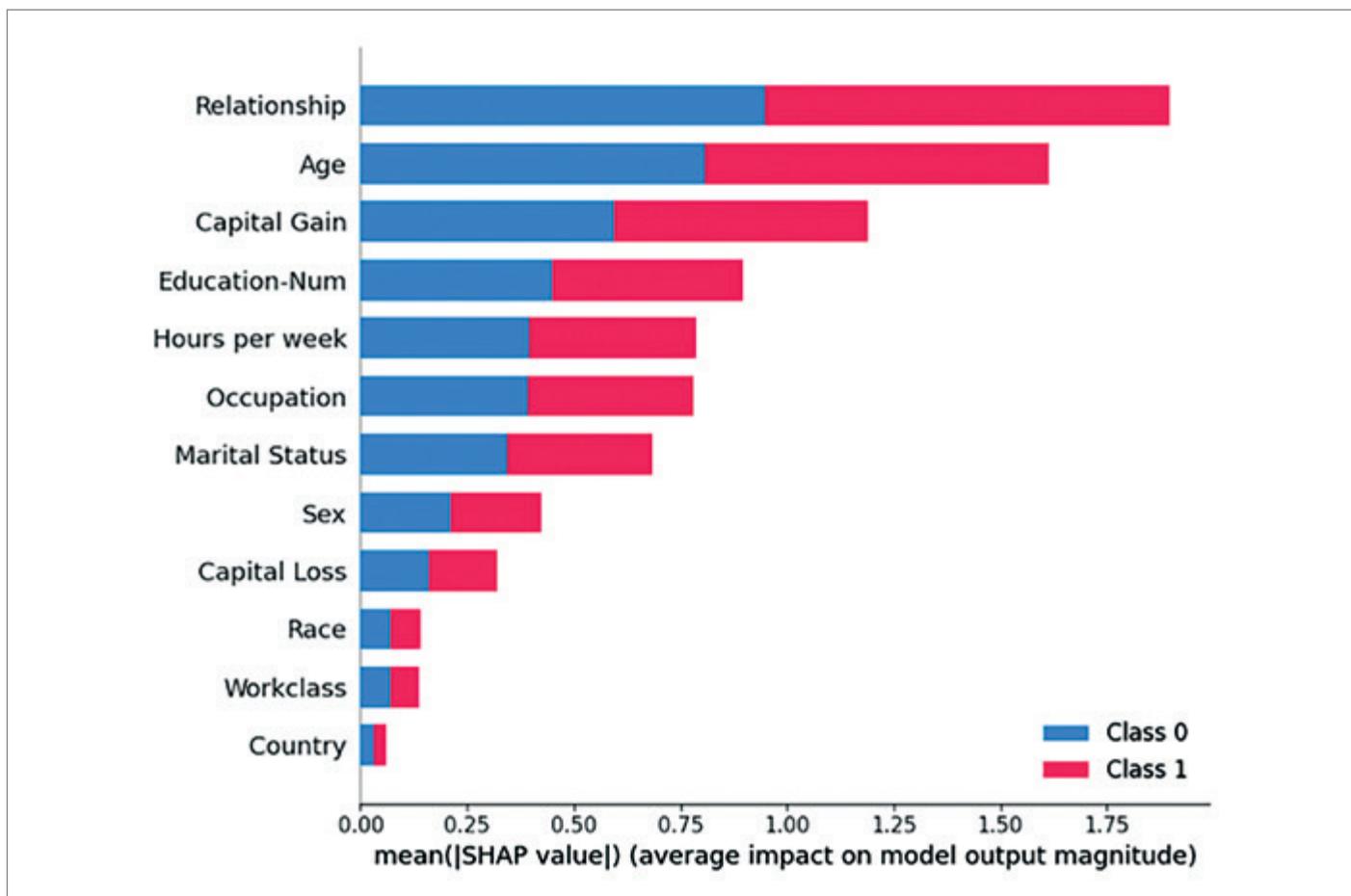


Abbildung 3: Feature Importance Plot am Beispiel des Adult Census Income Datasets (Quelle: Verena Barth)

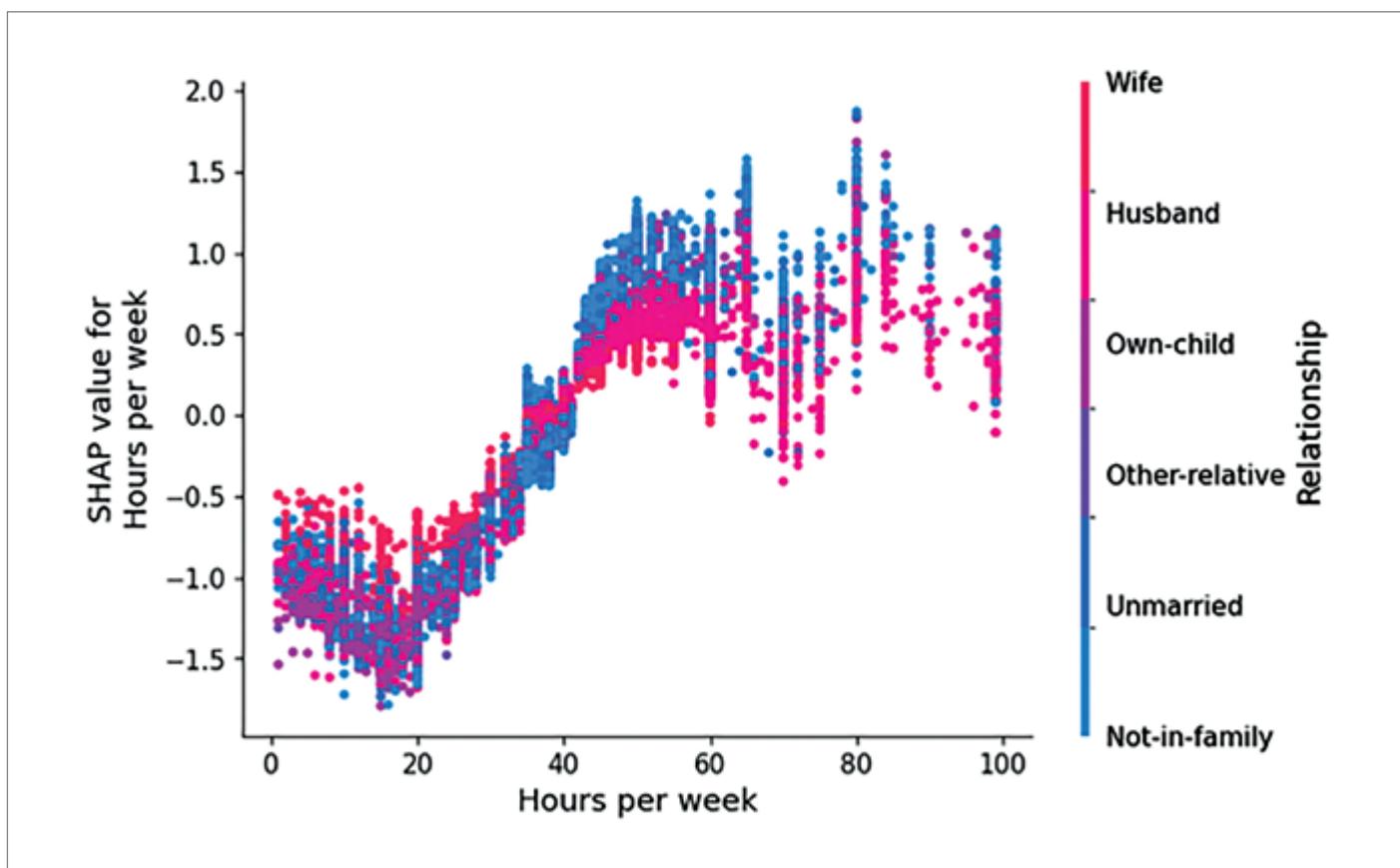


Abbildung 4: Dependence Plot am Beispiel des Adult Census Income Datasets (Quelle: Verena Barth)

ner Coverage von 0.02 nur auf 2% des Datensatzes zu. Eine allgemeingültigere Regel mit einer Coverage von 0.11 und einer trotzdem recht hohen Precision von 0.86 wäre: „*Education = Bachelors AND Relationship = Husband*“.

Worauf muss bei der Methodenauswahl geachtet werden?

Bei der Auswahl von XAI-Methoden sind einige wichtige Überlegungen zu berücksichtigen. Nach wie vor gibt es keine eindeutige Empfehlung, welche Methode(n) in welchem Kontext am besten geeignet ist/sind.

XAI-Methoden können je nach den Erklärungszielen verschiedener Benutzergruppen wie Stakeholdern, Experten und Nutzern ohne Fachwissen, zum Beispiel in Bezug auf Systemdebugging und -validierung, rechtlichen oder ethischen Anforderungen sowie der Vertrauensbildung zugeordnet werden [10]. Die Qualität der resultierenden Erklärungen variiert jedoch je nach Anwendungskontext, da sie kontextabhängig und durch individuelle Überzeugungen und kognitive Verzerrungen stark subjektiv ist, somit ist eine objektive Beurteilung schwierig [15].

Die Güte einer Methode kann daher anhand von Eigenschaften des Modell-, Daten- und Nutzungskontexts definiert werden, die

- die Anwendung der XAI-Methode erschweren oder unmöglich machen,
- aufgrund der algorithmischen Beschriftenheit der XAI-Methode einen negativen, verfälschenden Einfluss auf ein solides, kohärentes und vernünftiges Erklärungsergebnis haben,
- die Interpretierbarkeit der Erklärung mindern oder verkomplizieren. [3]

Dadurch lassen sich Ausschlusskriterien ableiten, die eine Anwendung der Methode für ein gewisses Modell unmöglich machen (wie zum Beispiel die Art der ML-Aufgabe), und eignungsbeeinflussende Kriterien, die einen negativen Einfluss haben (siehe Abbildung 5).

Ausschlusskriterien sind zum Beispiel die Art der ML-Aufgabe (manche Methoden sind nur für Klassifikationen geeignet) oder die Zugriffsmöglichkeit auf die Daten, die Klassenwahrscheinlichkeiten,

die Labels oder Vorverarbeitungsschritte der Daten des zu erklärenden Modells (um beispielsweise Kategorie-Kodierungen nachvollziehen zu können).

Wie anhand von PDP verdeutlicht, muss man bei allen perturbationsbasierten Verfahren zudem darauf achten, dass die Korrelationen unter den Eingabefeatures nicht zu hoch sind, um verfälschte Erklärungen zu vermeiden.

Manche Methoden, zum Beispiel Anchors, verwenden für die Interpretierbarkeit der resultierenden „Wenn-Dann“-Regeln bei kontinuierlichen Features interne Diskretisierungsverfahren (Binning). Die Auswahl eines guten Verfahrens ist schwierig und stark von den Daten abhängig [9]; bei Anchors, das generalisiert anwendbar ist, ist es nicht auf die Daten abgestimmt. Bei sehr schiefen, beziehungsweise ungleichmäßigen Verteilungen können dadurch Entscheidungsgrenzen verschwinden (Informationsverlust), was die Güte der Methode mindert [16]. Die Grenzen können dann entweder sehr nah beieinander liegen (zum Beispiel $29 \leq \text{„Age“} < 32$) oder viel zu weit sein („Age“ > 48), sodass entstehende Erklärungen entweder zu spezifisch oder unpräzise sind.

Außerdem gibt es weitere weiche Kriterien wie Performance-Präferenzen der XAI-Methode, die abhängig von der Häufigkeit und der Performance der Modell-Aufrufe sowie der Anzahl der Features (je mehr Features, desto höher Berechnungsaufwand) ist. Die Anwendung vieler Methoden erzeugt außerdem einen gewissen Einarbeitungs- und Vorbereitungsaufwand.

Zusätzlich zu diesen sich auf die Anwendbarkeit der Methoden beziehenden Faktoren sollte immer auf den Kontext sowie das Daten- oder Domänenwissen des Rezipienten der Erklärung geachtet werden. Zudem sollten Erklärungsbedürfnisse und -ziele erfüllt und sichergestellt werden, sodass die Abbildungen nicht missinterpretiert werden.

Diese Kriterien können bei der Auswahl einer XAI-Methode helfen, aber ...

Wie setzt man XAI erfolgreich ein?

Veröffentlichte Richtlinien zur Förderung der Transparenz von KI-Systemen weisen oft nur auf die Relevanz der Interpretierbarkeit hin, unterstützen allerdings nicht

bei der konkreten Anwendung von XAI. Nachfolgend sind einige Richtlinien für eine Integration transparenzfördernder Maßnahmen in den Entwicklungszyklus zusammengefasst.

Sofern White-Box-Modelle die ML-Aufgabe mit einer guten Leistung meistern können, sind sie in jedem Fall Black-Box-Modellen vorzuziehen [12]. Falls nur ein komplexes Black-Box-Modell infrage kommt, empfehlen der XAI-Forscher Belle (vgl. S. 15-17 [5]) und die Verantwortliche für Responsible AI Kangur [11] in ihren Richtlinien:

- Visualisierungstechniken sollten bereits bei der Datenexploration zur Vermeidung von verzerrten Daten angewendet werden. Diese sollten mindestens für alle wichtigen Features mit Diskriminierungspotenzial angewendet werden.
- Mit einem „local first“-Erklärungsansatz sollten anschließend lokale Methoden am besten auf richtig und falsch vorhergesagte Dateninstanzen angewendet werden, um zu sehen, wie sich kleine Veränderungen auf das Ergebnis auswirken. Gegebenenfalls kann so das Feature Engineering verbessert werden.
- Statt einer ausschließlichen Performancebetrachtung sollten auch globale Feature-Importance-Methoden angewendet werden, um die Vertrauenswürdigkeit des Modells sicherzustellen.
- Auch wenn das Modell in Betrieb ist, sollten die Eingabedaten kontinuierlich auf Änderungen geprüft und überwacht werden, um neuere Entwicklungen widerzuspiegeln, das Modell zu aktualisieren und somit Vertrauen in die Robustheit des Modells zu gewährleisten.

Anhand der Werke von Belle (vgl. S. 16 [5]), Kangur [11] und Leslie (vgl. S. 45-46 [12]), der momentan führenden Experten hinsichtlich XIA-Methoden, kann geschlossen werden, dass ein allgemeiner Konsens darüber besteht, dass eine Kombination mehrerer Erklärungstechniken förderlich für den Erhalt eines besseren Gesamtbilds des Modells und einer vollständigeren Erklärung sei.

Neben der Anwendung von XAI gibt es auch organisatorische Maßnahmen,

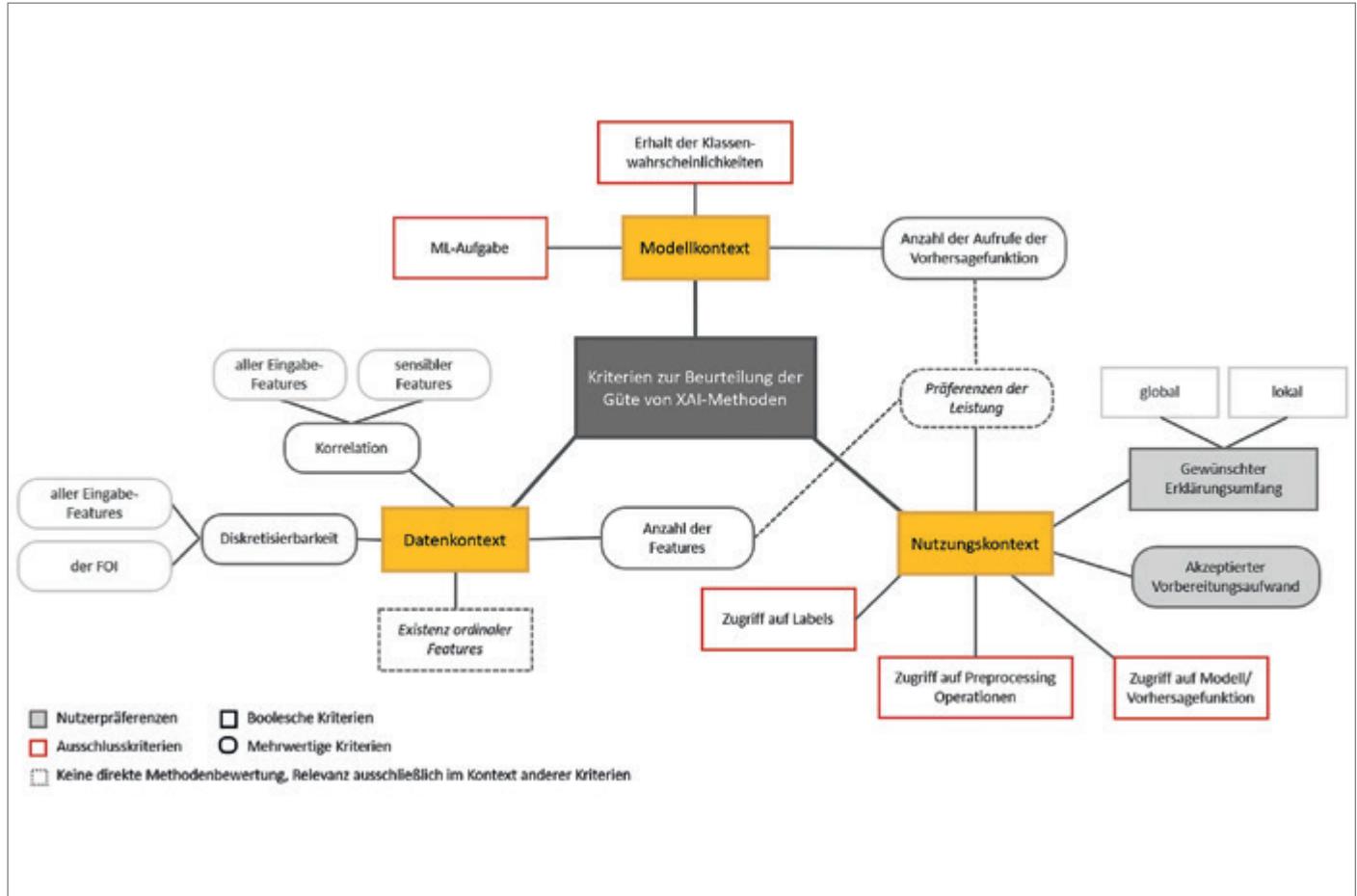


Abbildung 5: Übersicht der für die Beurteilung der Eignung einer XAI-Methode identifizierten Kriterien des Modell-, Daten- und Nutzungskontexts [3]

die zur ante-hoc-Vermeidung von Diskriminierung beitragen: Generell sollte schon bei der Teamzusammenstellung auf Diversität Wert gelegt werden, mit verschiedenen Personengruppen und Experten verschiedener Bereiche. Sie sollten ein gemeinsames Verständnis von Ziel, Kontext, den domänenspezifischen Anforderungen und Auswirkungen der ML-Aufgabe haben.

Hinsichtlich einer Gewährleistung der Nachvollziehbarkeit empfiehlt sich außerdem die Beachtung von MLOps-Praktiken, zum Beispiel der Dokumentation und Archivierung von (Meta-)Informationen von Datensätzen oder ML-Modellen in Model/Dataset Sheets. Es gibt diverse Tools/Plattformen, die Entwicklerinnen und Entwickler dabei unterstützen.

Zusammengefasst fördert eine Anwendung von XAI die essenzielle Interpretierbarkeit von ML-Modellen. Denn nur wenn Menschen maschinellen Entscheidern vertrauen, werden sie diese nutzen können. Mit Beachtung der

oben genannten Richtlinien zur Anwendung von XAI und der aufgeführten Eigenschaften des Modell-, Daten- und Nutzungskontexts zur Auswahl konkreter Methoden kann das Problem mangelnder Transparenz und Nachvollziehbarkeit aktiv angegangen und somit potenzielle Ungleichbehandlung vermieden werden.

Quellen

- [1] Adadi, A. & Berrada, M. (2018), Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI) in IEEE Access vol. 6, pp. 52138–52160.
- [2] Angwin, J., Larson, J., Kirchner, L. & Mattu, S. (2016): Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks., [Online], Verfügbar unter <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> (Zugriff am: 12.12.2023).
- [3] Barth, V. (2021): Konzeption und Implementierung eines Empfehlungssystems für die Auswahl und Anwendung von XAI-Methoden, [Online], verfügbar unter https://github.com/vereba/xai_recom
- [4] Becker, B. & Kohavi, R. (1996): Adult. UCI Machine Learning Repository. <https://doi.org/10.24432/C5XW20>.
- [5] Belle, V. & Papantonis, I. (2021): Principles and Practice of Explainable Machine Learning in Frontiers in big Data vol. 4, 688969.
- [6] Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V. & Kalai, A. T. (2016): Man is to computer programmer as woman is to homemaker? Neural Information Processing Systems (2016).
- [7] Dastin, J. (2018): Amazon scraps secret AI recruiting tool that showed bias against women, [Online], Verfügbar unter <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scaps-secret-ai-recruiting-tool-that-showed-bias-against-women-IDUSKCN1MK08G> (Zugriff am: 12.12.2023).
- [8] Fanaee-T, H. (2013): Bike Sharing Dataset. UCI Machine Learning Repository. <https://doi.org/10.24432/C5W894>.
- [9] Garcia, S., Luengo, J., Sáez, J. A., López, V. & Herrera, F. (2013): A Survey of Discretization Techniques: Taxonomy and Empirical Analysis in Supervised Learning in IEEE Transactions on Knowledge and Data Engineering vol. 25(4), pp. 734–750.
- [10] Hashemi, M. (2023): Who wants what and how: a Mapping Function for Ex-

- plainable Artificial Intelligence. ArXiv abs/2302.03180
- [11] Kangur, A. (2020): Explainable AI in practice: How committing to transparency made us deliver better AI products, [Online], Verfügbar unter <https://towardsdatascience.com/explainable-ai-in-practice-6d82b77bf1a7> (Zugriff am: 12.12.2023).
- [12] Leslie, D. (2019): Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector, SSRN 3403301.
- [13] Lundberg, S. M. & Lee, S.-I. (2017): A Unified Approach to Interpreting Model Predictions, arXiv preprint arXiv:1705.07874.
- [14] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. (2019): A survey on bias and fairness in machine learning, arXiv preprint arXiv:1908.09635
- [15] Miller, T. (2019): Explanation in Artificial Intelligence: Insights from the Social Sciences in Artificial intelligence vol. 267, pp. 1–38
- [16] Molnar, C. (2019): Interpretable Machine Learning, Verfügbar unter <https://christophm.github.io/interpretable-ml-book/>.
- [17] Partnership on AI (2021): Artificial Intelligence Incident Database, [Online], Verfügbar unter <https://incidentdatabase.ai> (Zugriff am: 12.12.2023).
- [18] Riberio, M. T., Singh, S. & Guestrin, C. (2018): Anchors: High precision model-agnostic explanations, Thirty-Second AAAI Conference on Artificial Intelligence.

Über die Autoren

Verena Barth war als IT-Beraterin bei der viadee mit dem Fokus auf Machine Learning und MLOps tätig. Sie setzt sich leidenschaftlich für die Verständlichkeit von komplexen ML-Systemen ein, um eine nachvollziehbare und gewissenhafte Anwendung zu ermöglichen. 2024 gründet sie ein Unternehmen, das mithilfe von LLMs professionelles und personalisiertes Business Coaching skalierbar anbietet.

Tobias Goerke ist Berater bei der viadee IT-Unternehmensberatung. Als Data Scientist liegen seine Schwerpunkte in der Einführung künstlich intelligenter Systeme und der Erforschung verschiedener Verfahren des erklärbaren Maschinenlernens.



Verena Barth
verena.bARTH95@web.de



Tobias Goerke
Tobias.Goerke@viadee.de

DOAG

DOAG
Datenbank
mit Exaday

2023

ON DEMAND

DATENBANK 2023 VERPASST?

Jetzt On-demand-Ticket buchen und
Vortragsaufzeichnungen anschauen!



ALLE ANGEBOTE
IM TICKETSHOP