Hochschule Augsburg University of Applied Sciences

Seminararbeit

Fakultät für Informatik

Studienrichtung
Informatik

Simon Hellbrück
Explainable Artificial Intelligence

Betreuer: Professor Thomas Rist
Masterseminar: Advanced Topics in Artificial Intelligence
Abgabe der Arbeit am: 16. Juli 2020

Hochschule für angewandte
Wissenschaften Augsburg
University of Applied Sciences

An der Hochschule 1
D-86161 Augsburg

Telefon +49 821 55 86-0
Fax +49 821 55 86-3222
www.hs-augsburg.de
info@hs-augsburg.de

Fakultät für Informatik
Telefon +49 821 5586-3450
Fax +49 821 5586-3499

Verfasser:
Simon Hellbrück
simon.hellbrueck@hs-augsburg.de

Inhaltsverzeichnis

Abkürzungsverzeichnis

AI Artificial Intelligence

ALE Accumulated Local Effects

BETA Black Box Explanations through Transparent Approximations

HTML Hypertext Markup Language

ICE Individual Conditional Expectation

KI Künstliche Intelligenz

LIME Local Interpretable Model-Agnostic Explanations

PDP Partial Dependence Plot

SHAP SHapley Additive exPlanations

XAI An eXplainability toolbox for machine learning

1 Einleitung

Neben Anwendungen Künstlicher Intelligenz, welche seit einigen Jahrzehnten in unterschiedlichen Gebieten Verwendung finden [1], koexistiert auch die Problematik der Komplexität einzelner Teilbereiche, die zu jener Disziplin gehören [2]. Der Begriff Explainable Artificial Intelligence (AI) wird in Fachkreisen für das Gebiet verwendet, in welchem der Versuch unternommen wird, komplex konstruierte Software, die sich durch Erfahrungswerte selbst optimieren kann, für den Menschen zugänglich und verständlich zu machen [3]. Bestimmte Teilbereiche der Künstlichen Intelligenz wie die Wahrscheinlichkeitsaussage und deren Modellfindung, in denen beispielsweise Wörter auf hochdimensionale Vektoren abgebildet werden, stellen für den Menschen kaum noch zu verstehende Abstraktionen dar [4]. Auch auotonomes Fahren, Gesichtserkennung, Sprachassistenten und Empfehlungssysteme stellen Anwendungen des täglichen Lebens dar, die für Menschen nur teilweise bis überhaupt nicht verstanden werden können [5]. In manchen Fällen mag es nicht von sonderlich großer Relevanz sein, einen Algorithmus vollends verstehen zu können, der aus einer komplexen Datenmenge ein bestimmtes Ergebnis berechnet, das ein Mensch – zumindest nicht in derselben Zeit – hätte berechnen können. Handelt es sich jedoch um moralisch-ethisch bedenkliche Fragestellungen wie dem Trolley-Problem im Bereich des autonomen Fahrens oder um Entscheidungen im medizinischen Bereich durch Empfehlungssysteme, müssen Methode und Lösungswege, die zu einem bestimmten Ergebnis geführt haben, nicht nur transparent und nachvollziehbar, sondern auch dergestalt sein, dass sie mit gesellschaftlichen Werten kompatibel sind [6, 7]. Genau hier knüpft Explainable AI mit dem Ziel an, komplexe Entscheidungswege offenzulegen, so dass sie nachvollzogen und verstanden werden können.

2 Einordnung

Erklärbare Künstliche Intelligenz lässt sich als Metawissenschaft bezeichnen, in der die Nachvollziehbarkeit maschineller Lernverfahren für den Menschen im Fokus steht. Sie koexistiert nicht nur mit Künstlicher Intelligenz, sondern ist ein Produkt ihrer selbst und wird bedingt durch jahrzehntelangen Fortschritt im Bereich maschineller Lernverfahren [4].
Zumeist sind in der (moderneren) Fachliteratur im Bereich der erklärbaren Anwendungen der Künstlichen Intelligenz Deep-Learning-Modelle Gegenstand der Betrachtung, jedoch diskutierten und diskutieren Experten potentiell erklärbare Expertensysteme, Wissenssysteme sowie Entscheider-Unterstützungssysteme unter den folgenden Begriffen:

- Explainable Expert Systems [8],
- Explanations in Knowledge Systems [9],
- Explainable Decision Support Systems [10]

3 Einsatzgebiete

Relevanz besitzen vor allem Applikationen in kritischen Bereichen, in denen schwer nachvollziehbare Black-Box-Modelle potentiell schwerwiegende Entschei-

dungen treffen. Aber auch Anwendungen der Künstlichen Intelligenz im täglichen Leben wie selbstlernende Algorithmen bei der Sprach- oder Gesichtserkennung sollen keine Black-Boxen sein [2].

4 Strategien

Wenngleich es einer standardisierten Vorgehensweise bei der Vereinfachung von Black-Box-Modellen in Fachkreisen mangelt und unterschiedliche Strategien existieren [11], ist der Ansatz, die unterschiedlichen Vorgehensweisen ganz allgemein in Post-hoc- beziehungsweise Ante-hoc-Methoden zu unterteilen, relativ weit verbreitet [4, 12, 13].

4.1 Übersicht

Tabelle 1 gibt einen Überblick über die erwähnten Methoden.

System	Funktionsweise	Beispiele
Ante-hoc	Verwendung von Modellen, welche von Natur aus White-Box-Ansätze sind.	Lineare Regression, Logistische Regression, fuzzy inference, Entscheidungsbäume, Entscheidungsregeln, Generalized Additive Model (GAM), Generalized Linear Models (GLM), RuleFit
Post-hoc	System wird als agnostisches Modell verstanden, in welchem einzelne Entscheidungen erklärbar und reproduzierbar sein sollen.	Black Box Explanations through Transparent Approximations (BETA), Partial Dependence Plot (PDP), Individual Conditional Expectation (ICE), Accumulated Local Effects (ALE), Feature Interaction, Permutation Feature Importance, Global Surrogate, Local Interpretable Model-Agnostic Explanations (LIME), Scoped Rules (Anchors), Shapley Values, SHAP (SHapley Additive exPlanations)

Tabelle 1: Methoden zum Verständnis von Entscheidungswegen [6, 14]

An dieser Stelle soll nicht unerwähnt bleiben, dass neben agnostischen Modell-Ansätzen auch spezifische existieren. Da diese jedoch durch ihre starke Abhängigkeit von dem zu interpretierenden Modell wenig Flexibilität aufweisen, sind agnostische Modelle – unabhängig von der Komplexität des zugrunde liegenden Machine-Learning-Modells – zu bevorzugen [15]. In den folgenden Kapiteln werden die Vorgehensweisen LIME, PDP sowie ICE, die in Tabelle 1 unter *Post-hoc* aufgeführt sind, weiter ausgeführt.

4.2 LIME

Wie die Bezeichnung der Vorgehensweise bereits suggeriert, liegt der Fokus darauf, Erklärungen konkreter Entscheidungen auf bestimmte Bereiche eines

Systems zurückzuführen und nicht etwa, das System in seiner Gesamtheit zu verstehen.

4.2.1 Übersicht

Seitdem das Framework 2016 von Marco Ribeiro [16] vorgestellt worden ist, stößt LIME in Fachkreisen auf reges Interesse [17, 18, 19]. Hinter Black-Box-Modellen verstecken sich meist komplexe, mehrdimensionale Funktionen, deren Verhalten in ihrer Gesamtheit nur schwer zu erfassen sind. Reduziert man jedoch das Modell auf einen konkreten Ein-/Ausgabebereich, verringert sich auch die Komplexität. Das heißt, dass eine Beobachtung des Black-Box-Modells durch ein lokales Modell approximiert wird. Da LIME ein sogenanntes Surrogat-Modell darstellt, werden Vorhersagen eines zugrundeliegenden Black-Box-Modells angenähert, nicht aber dessen Logik. Hierbei beschränkt sich das Framework auf Modelle, die tatsächlich interpretierbar sind, wie etwa lineare/logistische Regression oder Entscheidungsbäume. Eine Vorhersage wird nun verständlich, da ein Bereich des – sonst nicht erklärbaren – Black-Box-Modells durch das Surrogat ersetzt worden ist [14].

4.2.2 Algorithmisches Vorgehen

LIME modifiziert eine Datenstichprobe, indem Merkmalswerte justiert werden, anschließend wird die Auswirkung auf den Output betrachtet. Grundsätzlich wird der Frage nachgegangen, warum gewisse Vorhersagen getroffen wurden und welche Parameter hierfür verantwortlich waren [20].
Zunächst wird eine zu erklärende Beobachtung x gewählt. Abhängig von der Verteilung der Merkmale im Datensatz, werden von LIME neue, künstliche Beobachtungen erzeugt. Zuvor muss die Verteilung aller Merkmale jedoch geschätzt werden. Der dann entstandene Datensatz besteht aus den synthetischen Beobachtungen sowie den Vorhersagewerten des Black-Box-Modells. Daraufhin werden die Differenzen zwischen künstlicher Beobachtung und tatsächlicher gewichtet. Ebenjene Gewichtungen stellen das Fundament dar, auf dem LIME ein erklärbares Modell trainiert, wobei die Zielgröße die Vorhersage des Black-Box-Modells darstellt [21]. Hierbei kann eine gute lokale Annäherung an eine Beobachtung erreicht werden – diese kann jedoch global nicht mehr gültig sein beziehungsweise in die Irre führen. Der von Marco Ribeiro [16] verwendete Begriff *locally faithful* deutet diese Problematik bereits an.

4.2.3 Beispiel

Abbildung 1 illustriert die Problematik bei hochdimensionalen, nicht-linearen Funktionen, die einem Black-Box-Modell zugrunde liegen können.

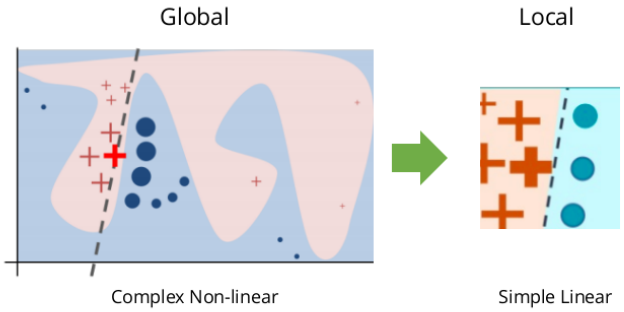


Abbildung 1: (Beispielhafte) lokale Approximation durch LIME [22]

Rein formal können lokale Surrogat-Modelle eingeschränkter Interpretierbarkeit folgendermaßen dargestellt werden [14].

$$explanation(x) = arg \min_{g \in G} L(f, g, \pi_x) + \Omega(g) \tag{1}$$

Das erklärende Modell für eine beliebige Instanz x ist dasjenige, das die Differenz zwischen Vorhersage und eigentlichem Wert durch die Verlust-Funktion L sowie die Model-Komplexität durch $\Omega(g)$ möglichst gering hält. G enthält mögliche Modelle (Lineare Regression, Entscheidungsbäume etc.). Das Proximitätsmaß π_x definiert den Bereich um die betrachtete Instanz x. In der Praxis jedoch optimiert LIME lediglich die Verlust-Funktion L, während die Komplexität durch den User bestimmt wird – etwa durch die Maximalzahl an Features, die ein lineares Regressions-Modell benutzen kann [20].

4.3 PDP

Vielerorts werde moniert, dass maschinelle Lernverfahren nicht zu verstehende Blackboxen seien, deren Arbeitsweise – unabhängig vom jeweiligen Datensatz – nicht eingesehen werden könne. Das daraus resultierende Problem, dass weder ein Verständnis vorherrscht geschweige denn potentielle Probleme des Modells identifiziert werden können, lässt sich laut Rogelio Escalante [23] darauf zurückführen, dass partielle Abhängigkeitsdiagramme zu wenig Verwendung finden beziehungsweise zu wenige Anwender sich mit ihnen auskennen.

4.3.1 Übersicht

Ebenjene Diagramme zeigen, wie einzelne unabhängige Variablen die Vorhersage eines Modells beeinflussen. Solche Plots können verwendet werden, um beispielsweise folgende Fragen zu klären:

- Bestehen geschlechtsbedingte Gehaltsunterschiede ausschließlich aufgrund des Geschlechts oder auch wegen eines unterschiedlichen Bildungshintergrundes oder Berufserfahrung [24]?
- Sind gesundheitliche Unterschiede zwischen zwei Gruppen auf unterschiedliche Ernährungsgewohnheiten oder auf andere Faktoren zurückzuführen [24]?

- Welche Einflussfaktoren korrelieren besonders stark mit den Verkäufen von bestimmten Produkten (wie beispielsweise Fahrrädern)? Die Temperatur, die Windgeschwindigkeit oder die Luftfeuchtigkeit [14]?

Die letztgenannte Fragestellung wird in einem Beispiel am Ende dieses Kapitels anhand eines Beispiels illustriert.

4.3.2 Algorithmisches Vorgehen

Der Ansatz *Partial Dependence Plot* beziehungsweise partielles Abhängigkeitsdiagramm zeigt, inwiefern eine oder zwei unabhängige Variablen im Zusammenhang zur Vorhersage stehen. Alle restlichen unabhängigen Variablen werden herausgerechnet. Daraufhin kann eine Aussage darüber getroffen werden, in welcher Beziehung die Zielgröße zur unabhängigen Variable steht: lineare oder monotone Relation? Oder gar komplexer? Wird die Strategie beispielsweise auf eine lineare Regression angewandt, werden partielle Abhängigkeitsdiagramme logischerweise eine lineare Beziehung darstellen. Die partielle Abhängigkeitsfunktion für Regressionen ist folgendermaßen definiert [14].

$$\hat{f}_{x_S}(x_S) = E_{x_C}[\hat{f}(x_S, x_C)] = \int \hat{f}(x_S, x_C) d\mathbb{P}(x_C) \tag{2}$$

Der partiellen Abhängigkeitsfunktion f werden Parameter x_S übergeben, für welche die Funktion geplottet werden soll. Die Parameter x_C stellen weitere Features dar, die im Machine-Learning-Modell f verwendet werden. Für gewöhnlich enthält S lediglich ein oder zwei Features, von denen in Erfahrung gebracht werden soll, welchen Einfluss sie auf die Vorhersage des Modells besitzen. Die Feature-Vektoren x_S und x_C stellen kombiniert den gesamten Vorrat an verfügbaren Feature-Werten dar. Indem die Vorhersage des Modells über die Verteilung der Features in x_C marginalisiert wird, kann eine partielle Abhängigkeit zwischen den Merkmalen in S und der Vorhersage bestimmt werden. Werden nun auch die restlichen Merkmale marginalisiert, erhält man eine Funktion, die vollständig auf Merkmalen von S beruht. Somit berechnet die partielle Abhängigkeitsfunktion – unter Berücksichtigung aller möglichen Ausprägungen der Merkmale in S – eine gemittelte Modellvorhersage. Um das Integral in 2 anzunähern, wird die folgende Approximation verwendet [14].

$$\hat{f}_{x_S}(x_S) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}(x_S, x_C^{(i)}) \tag{3}$$

Die Funktion $\hat{f}_{x_S}$ erhält die Features in S und berechnet den gemittelten marginalen Effekt auf die Vorhersage des Modells. x_S enthält die restlichen Ausprägungen der Variablen im Datensatz, die nicht betrachtet werden, n steht für die Anzahl der vorhandenen Beobachtungen im Datensatz. Eine Annahme von PDP lautet, dass die Merkmale in C und S nicht miteinander korrelieren. Ist diese Annahme verletzt, werden die geplotteten Datenpunkte nur wenig Sinn ergeben. Für Klassifizierungen von Modellen, deren Output Wahrscheinlichkeiten sind, liefert PDP eine Wahrscheinlichkeit für eine gegebene Klasse mit unterschiedlichen Ausprägungen des Merkmals in S. Um multiple Klassen in den Griff zu bekommen, lässt sich schlichtweg ein Plot pro Klasse illustrieren. Die Methode der partiellen Abhängigkeiten ist eine globale Methode, da alle Instanzen berücksichtigt werden und eine Aussage über die globale Beziehung zwischen einem Merkmal und einer Vorhersage getroffen wird [14].

4.3.3 Beispiel

Abbildung 2 zeigt, welchen Einfluss einzelne Variablen auf den Verkauf von Fahrrädern besitzen.

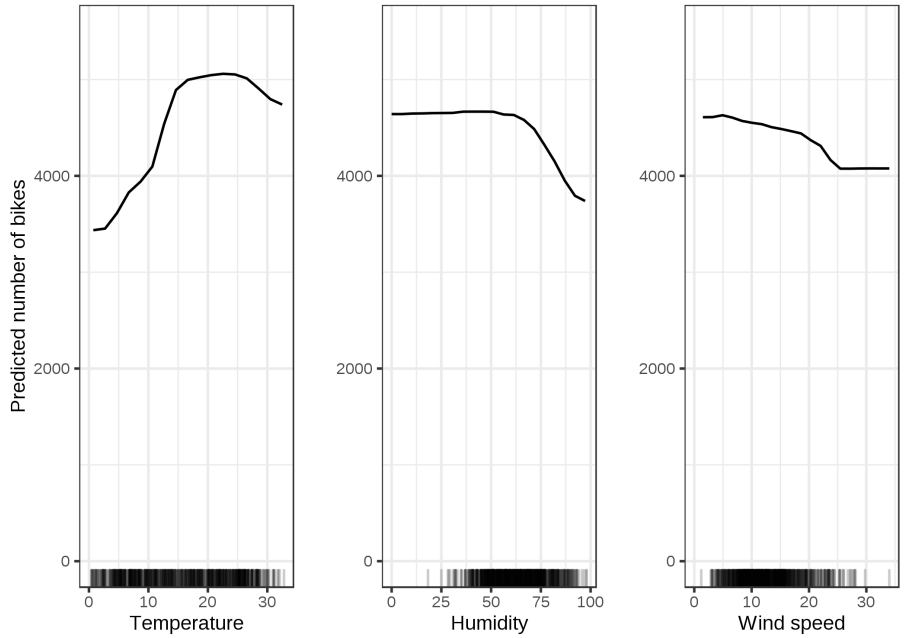


Abbildung 2: PDP-Plots für Verkäufe von Fahrrädern [14]

Die unabhängigen Variablen Temperatur, Luftfeuchtigkeit und Windgeschwindigkeit stehen in obiger Abbildung in Zusammenhang mit der Vorhersage der Verkäufe. Der größte Einflussfaktor scheint die momentane Temperatur zu sein. Aus dem ersten Plot in Abbildung 2 lässt sich schließen, dass die Verkäufe bis zu einer Temperatur von circa 25 Grad Celsius zunehmen und anschließend abnehmen.

4.4 ICE

Während die im vorherigen Kapitel behandelten partiellen Abhängigkeitsdiagramme den gemittelten Effekt eines Merkmals anzeigen und somit als globale Methode zu verstehen sind, können ICE-Plots als deren Äquivalent gesehen werden [25].

4.4.1 Übersicht

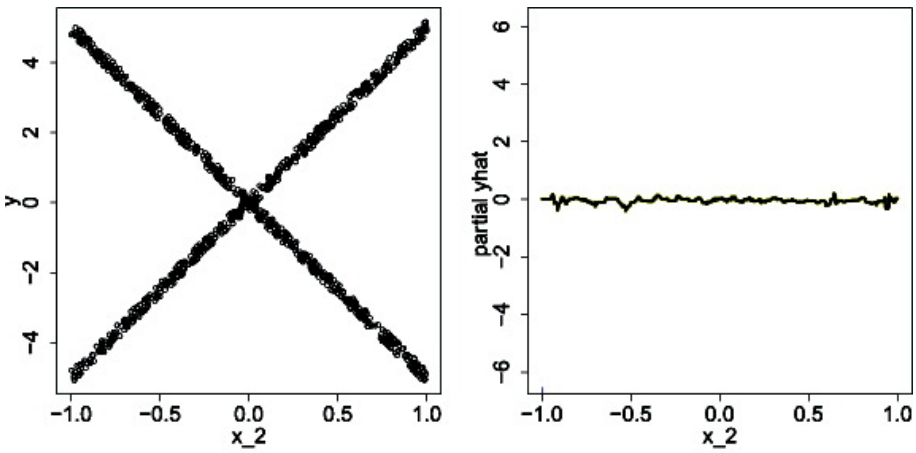
ICE-Plots zeigen eine Gerade pro Instanz, wobei deutlich wird, wie sich die Vorhersage einer Instanz in Abhängigkeit von einem Merkmal ändert. Ein ICE-Plot illustriert eine Vorhersage in Abhängigkeit von einem Merkmal für jede Instanz separat. Daraus resultiert ein Diagramm, das jeweils eine Gerade für eine Instanz zeigt, während partielle Abhängigkeitsdiagramme lediglich eine gemittelte Gerade zeigen. Die Gerade in einem partiellen Abhängigkeitsdiagramm lässt sich somit als Mittelung aller Instanzen eines ICE-Plots verstehen [14].

4.4.2 Algorithmisches Vorgehen

Werte einer Linie und einer Instanz können berechnet werden, indem alle restlichen Merkmale bestehen bleiben, während Varianten ebenjener Instanz erstellt werden, indem der Wert des Merkmals durch Werte aus einem Raster ersetzt wird und Vorhersagen mit dem entsprechenden Black-Box-Modell für jene neu erstellten Instanzen getroffen werden. Das Ergebnis ist eine Reihe von Punkten für eine Instanz mit dem Merkmalswert aus dem Raster und den jeweiligen Vorhersagen. Rein formal lässt sich die Vorgehensweise folgendermaßen definieren: Für jede Instanz $\{(x_S^{(i)}, x_C^{(i)})\}_{i=1}^N$ wird die Kurve $\hat{f}_S^{(i)}$ gegen $x_S^{(i)}$ geplottet, während $x_C^{(i)}$ gleich bleibt [14].

4.4.3 Beispiel

Abbildung 3 verdeutlicht die Schwächen von partiellen Abhängigkeitsdiagrammen gegenüber ICE-Plots.



(a) Scatterplot of Y versus X_2 (b) PDP

Abbildung 3: Interaktionen von Merkmalen wird durch PDP nicht klar [25]

Der Vorteil, der sich aus der Verwendung von ICE-Plots ergibt, besteht darin, individuelle Erwartungen anstelle von Teilabhängigkeiten zu betrachten. Partielle Abhängigkeitsdiagramme können eine durch Interaktionen erzeugte heterogene Beziehung verschleiern. PDPs können verdeutlichen, wie die durchschnittliche Beziehung zwischen einem Merkmal und der betreffenden Vorhersage aussieht. Dies funktioniert jedoch nur dann, wenn die Wechselwirkungen zwischen den Merkmalen (für die partielle Abhängigkeitsdiagramme berechnet werden) und den anderen Merkmalen schwach sind. Im Falle von Interaktionen zwischen Merkmalen ermöglicht ein ICE-Diagramm einen deutlich besseren Einblick in die Daten [14].

4.5 Fazit

Die größte Herausforderung bei der Entwicklung eines Modells besteht darin, es so zu gestalten, dass es einfach genug ist, um von Laien verstanden werden zu können, und dennoch so ausgefeilt, dass es den zugrunde liegenden Daten

genügt. Dementsprechend kann das Mittel der Wahl nicht immer jenes sein, komplexe Modelle vollends zu ersetzen. Eine Abwägung ist hierbei notwendig. Ein Modell, das einen inhärent erklärbaren Ansatz mit einer komplexen Black-Box-Methode kombiniert, ohne dessen Vorhersagegenauigkeit einzuschränken, kann in vielen Fällen sinnvoller sein [26].

5 Marktpotentiale

Da immer fortschrittlichere maschinelle Lernverfahren entwickelt werden und sich deren Entscheidungen auf unser tägliches Leben auswirkt (bezüglich Kreditwürdigkeit, Rekrutierung, Gesundheitsfürsorge oder des Transportwesens), ist der Bedarf an verständlichen algorithmischen Verfahren als ansteigend zu betrachten [27].

6 Relevante Journals/Konferenzen

Unter dem Begriff *FATML (Fairness, Accountability, and Transparency in Machine Learning)* wird ein Bereich des maschinellen Lernens verstanden, der Anwendern und Forschern durch regelmäßig stattfindende Konferenzen eine Möglichkeit zum Austausch von Berechenbarkeit, Transparenz und Fairness maschineller Lernverfahren anregen soll [28]. Im Folgenden werden einige Konferenzen/Workshops aufgeführt, welche regelmäßig die Thematik Explainable AI zum Thema haben beziehungsweise hatten.

6.1 ACM FAccT Conference

Seit 2018 findet jährlich die *ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)* statt, auf welcher regelmäßig zahlreiche Vorträge zum Thema erklärbarer Intelligenz gehalten werden [29].

6.2 Explainable AI xAI

In diesem interdisziplinären Workshop kommen internationale Experten zusammen, die daran interessiert sind, Maschinenentscheidungen transparent, interpretierbar, reproduzierbar, rückverfolgbar, reaktiv, verständlich und ethisch verantwortbar zu gestalten [30].

6.3 CVPR-19

Dieser Workshop soll Forscher, Ingenieure und Praktiker aus der Industrie zusammenbringen, deren Themenfeld Interpretierbarkeit, Sicherheit und Zuverlässigkeit Künstlicher Intelligenz umfasst. Es wird erwartet, dass gemeinsame Anstrengungen in dieser Richtung die Black-Box der Deep Neural Nets (DNNs) transparenter macht.

6.4 FAT/ML

Von 2014 bis 2018 fand jährlich die Konferenz *Fairness, Accountability, and Transparency in Machine Learning* an unterschiedlichen Orten statt [31].

7 Tools und Tutorials

Im Bereich der erklärbaren Künstlichen Intelligenz lassen sich einige Tools und Tutorials auf GitHub finden. Nachfolgend werden einige kurz vorgestellt und deren Zweck erläutert.

7.1 Interactive Studio for Explanatory Model Analysis

Das modelStudio-Paket automatisiert die erklärende Analyse von Vorhersagemodellen für maschinelles Lernen. Es lassen sich erweiterte interaktive und animierte Modellerklärungen in Form einer serverlosen Hypertext Markup Language (HTML)-Seite mit nur einer Codezeile generieren. Da dieses Tool modellagnostisch ist, besteht eine Kompatibilität mit vielen Black-Box-Vorhersagemodellen und Frameworks [32].

7.2 An eXplainability toolbox for machine learning (XAI)

XAI ist eine Bibliothek für maschinelles Lernen, deren Kern die Erklärbarkeit Künstlicher Intelligenz ist. XAI enthält verschiedene Tools, mit denen Daten und Modelle analysiert und ausgewertet werden können. Die XAI-Bibliothek wird vom Institut für ethische Künstliche Intelligenz und maschinelles Lernen verwaltet und wurde auf der Grundlage der 8 Prinzipien für verantwortungsbewusstes maschinelles Lernen entwickelt [33].

7.3 Anchorj

Das Projekt Anchorj stellt eine Java-Implementierung des Anchors-Erklärungsalgorithmus für maschinelle Lernmodelle dar und basiert auf dem Paper *Anchors: High-Precision Model-Agnostic Explanations* von Marco Ribeiro [34].

7.4 Machine Learning Tutorials and Articles

Das Projekt *Machine Learning Tutorials and Articles* beinhaltet zahlreiche Artikel bezüglich maschineller Lernverfahren, inklusive Artikel über die Erklärbarkeit ebenjener Verfahren [35].

7.5 Picasso

Picasso ist ein kostenloses Open-Source-DNNs-Visualisierungstool, mit welchem sich Okklusions- und Ausnahmekarten erstellen lassen. Es handelt sich hierbei um eine Flask-Anwendung, die ein Deep-Learning-Framework mit einer Reihe von Standard- und benutzerdefinierten Visualisierungen zusammenfügt. Es können sowohl die integrierten Visualisierungen verwendet als auch eigene hinzufügt werden. Picasso wurde für die Arbeit mit neuronalen Netzen mit Keras und Tensorflow entwickelt. Darüber hinaus werden Tensorflow- und Keras MNIST-Checkpoints sowie ein Keras VGG16-Checkpoint angeboten [36].

7.6 TensorWatch

TensorWatch ist ein Debugging- und Visualisierungstool, das für Data Science, Deep Learning und Reinforcement Learning von Microsoft Research entwickelt

wurde. Es funktioniert in Jupyter Notebook, um Echtzeitvisualisierungen eines maschinellen Lerntrainings anzuzeigen und verschiedene weitere relevante Analyseaufgaben für Modelle und Daten auszuführen [37].

8 Bisherige Erfolge

Eine Behörde des Verteidigungsministeriums der Vereinigten Staaten unterscheidet zwischen den folgenden Ansätzen mit dem Ziel, Deep-Learning-Modelle nachvollziehbarer zu gestalten. Hierbei konzentriert sich die Behörde stark darauf, maschinelle Lernverfahren auch für Laien verständlich zu gestalten und nicht lediglich für Experten. Die im Projekt *Explainable Artificial Intelligence* arbeitenden Teams gehen nach drei strategischen Konzepten vor [2]:

- **Deep explanation**: Hierbei soll das Deep-Learning-Model so geändert werden, dass es erklärbar wird. Eine bekannte Vorgehensweise besteht darin, weitere Elemente in den verschiedenen Schichten neuronaler Netze hinzuzufügen, um die komplexen Verbindungen, die während der Trainingsphase entstehen, besser verstehen zu können.
- **Building more interpretable models**: Deep-Learning-Modelle werden bei dieser Strategie durch andere Modelle Künstlicher Intelligenz ersetzt, die von Natur aus erklärbar sind. Beispiele hierfür sind probabilistische, relationale Modelle oder fortschrittliche Entscheidungsbäume. Jene Verfahren also, die bereits vor dem Aufkommen von neuronalen Netzen Verwendung gefunden haben.
- **Model induction**: Unter diesem agnostischen Ansatz wird ein neuronales Netz als Black-Box verstanden. Es wird mit dem Input experimentiert und versucht, über den Output Rückschlüsse auf das Verhalten des Netzes zu ziehen.

Die Vorgehensweise `building more interpretable models` der amerikanischen Behörde lässt sich nach Tabelle 1 den Ante-hoc-Strategien zuschreiben, während die Methode der `model induction` unter den Methoden der Post-hoc-Strategien zu verorten ist. Die Methode `deep explanation` fällt etwas aus dem Raster, da hier nicht direkt ein Modell Gegenstand der Optimierung ist.

9 Aktuelle Forschungsfragen

Zwar wird die Thematik erklärbarer maschineller Lernverfahren in Fachkreisen mannigfaltig diskutiert, jedoch mangelt es an konkreten Handlungsempfehlungen und Anwendungsfällen, wie die Transferleistung von Erkenntnissen aus Black-Box-Modellen auf Problemstellungen erfolgen kann, wenn der Zusammenhang und die Interaktion nicht weiter bekannt sind. Dass sich maschineller Lernverfahren bedient werden kann, um statistische Verfahren zu optimieren, ist hinlänglich bekannt [38, 39, 40]. Einen ersten umfangreichen Vorschlag, wie eine potentielle Integration reproduzierbar und unabhängig von der jeweiligen Komplexität aussehen könnte, unterbreitet Christophe Krech in seiner Arbeit *Erklärbarkeit maschineller Lernverfahren* [21]. An diesem Ansatz gilt es weiterzuarbeiten.

10 Einstiegspunkte

Da es – wie bereits erwähnt – einer einheitlichen Definition beziehungsweise Strategie bezüglich der Herangehensweise an Black-Box-Modelle in der Fachliteratur mangelt, gestaltet es sich denkbar schwierig, einen bestimmten Einstiegspunkt zu empfehlen. Nichtsdestotrotz sei an dieser Stelle auf das Buch *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable* von Chrstoph Molnar [14] sowie auf die Abschlussarbeit von Christophe Krech [21] hingewiesen, deren Ausarbeitungen gut geeignet sind, sich einen ersten Überblick zu verschaffen, wenngleich weitere Strategien wie die von Bahador Khaleghi [26] existieren, die in diesem Zusammenhang inhaltlich eine andere Richtung einschlagen.

Literatur

[1] of Dartmouth College T. Artificial Intelligence (AI) Coined at Dartmouth; 2020. Online; abgerufen am 26. April 2020. Verfügbar unter: https://250.dartmouth.edu/highlights/artificial-intelligence-ai-coined-dartmouth.

[2] Holzinger A. Inside DARPA's effort to create explainable artificial intelligence; 2019. Online; abgerufen am 05. April 2020. Verfügbar unter: https://bdtechtalks.com/2019/01/10/darpa-xai-explainable-artificial-intelligence/.

[3] Turek M. Explainable Artificial Intelligence (XAI); o. J. Online; abgerufen am 22. Juni 2020. Verfügbar unter: https://www.darpa.mil/program/explainable-artificial-intelligence.

[4] Holzinger A. Explainable AI (ex-AI); 2018. Online; abgerufen am 04. April 2020. Verfügbar unter: https://gi.de/informatiklexikon/explainable-ai-ex-ai/.

[5] Rai A. Explainable AI: From black box to glass box. Journal of the Academy of Marketing Science. 2020;48(1):137–141.

[6] Holzinger A, Biemann C, S Pattichis C, B Kell D. What do we need to build explainable AI systems for the medical domain?; 2017.

[7] Glomsrud JA, Ødegårdstuen A, Clair ALS, Smogeli Ø. Trustworthy versus Explainable AI in Autonomous Vessels. In: Proceedings of the International Seminar on Safety and Security of Autonomous Vessels (ISSAV) and European STAMP Workshop and Conference (ESWC) 2019. Sciendo; 2020. S. 37–47.

[8] Swartout WR, Cecile L. Paris, and Johanna D. Moore. Design for explainable expert systems. IEEE Expert. 1991;6(3):58–64.

[9] Chandrasekaran B, Swartout W. Explanations in knowledge systems: the role of explicit representation of design knowledge. IEEE expert. 1991;6(3):47–49.

[10] Sachan S, Yang JB, Xu DL, Benavides D, Li Y. An Explainable AI Decision-Support-System to Automate Loan Underwriting. Expert Systems with Applications. 2019 11;144:113100.

[11] Sokol K, Flach P. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency; 2020. S. 56–67.

[12] Gandhi P. Explainable Artificial Intelligence; 2019. Online; abgerufen am 30. April 2020. Verfügbar unter: https://www.kdnuggets.com/2019/01/explainable-ai.html.

[13] Srihari S. Explainable Artificial Intelligence; 2019. Online; abgerufen am 30. April 2020. Verfügbar unter: https://cedar.buffalo.edu/srihari/talks/IISc-XAI.pdf.

[14] Molnar C. Interpretable Machine Learning. A Guide for Making Black Box Models Explainable; 2020. Online; abgerufen am 09. April 2020. Verfügbar unter: https://christophm.github.io/interpretable-ml-book/pdp.html.

[15] Tanner G. Introduction to Machine Learning Model Interpretation; 2019. Online; abgerufen am 09. April 2020. Verfügbar unter: https://gilberttanner.com/blog/introduction-to-machine-learning-model-interpretation.

[16] Ribeiro MT, Singh S, Guestrin C. 'Why should I trust you?' Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining; 2016. S. 1135–1144.

[17] Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org; 2017. S. 3319–3328.

[18] Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. ACM computing surveys (CSUR). 2018;51(5):1–42.

[19] Tolomei G, Silvestri F, Haines A, Lalmas M. Interpretable predictions of tree-based ensembles via actionable feature tweaking. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining; 2017. S. 465–474.

[20] Hulstaert L. Understanding model predictions with LIME; 2018. Online; abgerufen am 10. April 2020. Verfügbar unter: https://towardsdatascience.com/understanding-model-predictions-with-lime-a582fdff3a3b.

[21] Krech C. Erklärbarkeit maschineller Lernverfahren [Masterarbeit]. Hochschule Darmstadt; 2019.

[22] Joseph M. Interpretability part 3: opening the black box with LIME and SHAP; 2019. Online; abgerufen am 22. Juni 2020. Verfügbar unter: https://www.kdnuggets.com/2019/12/interpretability-part-3-lime-shap.html.

[23] Escalante R. Machine Learning: Partial Dependence Plots; 2018. Online; abgerufen am 25. April 2020. Verfügbar unter: https://www.kaggle.com/rogelioem/machine-learning-partial-dependence-plots.

[24] Becker D. Partial Dependence Plots; 2020. Online; abgerufen am 30. April 2020. Verfügbar unter: https://www.kaggle.com/dansbecker/partial-plots.

[25] Goldstein A, Kapelner A, Bleich J, Pitkin E. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. Journal of Computational and Graphical Statistics. 2015;24(1):44–65.

[26] Khaleghi B. The How of Explainable AI: Explainable Modelling; 2019. Online; abgerufen am 23. April 2020. Verfügbar unter: https://towardsdatascience.com/the-how-of-explainable-ai-explainable-modelling-55c8c43d7bed.

[27] Waltl B, Vogl R. Explainable artificial intelligence the new frontier in legal informatics. Jusletter IT. 2018;4:1–10.

[28] Bertani-Økland MA. What is FATML and why should you care; 2019. Online; abgerufen am 25. April 2020. Verfügbar unter: https://medium.com/grensesnittet/https-medium-com-mab-55055-what-is-fatml-and-why-should-you-care-dfb36e51f2f4.

[29] Conference AF. ACM FAT* Conference 2020; 2020. Online; abgerufen am 25. April 2020. Verfügbar unter: https://facctconference.org/2020/index.html.

[30] Group H. Explainable AI xAI 2020; 2020. Online; abgerufen am 23. April 2020. Verfügbar unter: https://human-centered.ai/explainable-ai-2020/.

[31] FAT/ML. Fairness, Accountability, and Transparency in Machine Learning; 2018. Online; abgerufen am 25. April 2020. Verfügbar unter: https://www.fatml.org/.

[32] Oriented M. Interactive Studio for Explanatory Model Analysis; 2020. Online; abgerufen am 25. April 2020. Verfügbar unter: https://github.com/ModelOriented/modelStudio.

[33] for Ethical Machine Learning TI. An eXplainability toolbox for machine learning; 2019. Online; abgerufen am 25. April 2020. Verfügbar unter: https://github.com/EthicalML/xai.

[34] Ribeiro MT, Singh S, Guestrin C. Anchors: High-precision model-agnostic explanations. In: Thirty-Second AAAI Conference on Artificial Intelligence; 2018. .

[35] Fabien M. Machine Learning Tutorials and Articles; 2019. Online; abgerufen am 25. April 2020. Verfügbar unter: https://github.com/maelfabien/Machine_Learning_Tutorials.

[36] Henderson R. Picasso: A free open-source visualizer for Convolutional Neural Networks; 2017. Online; abgerufen am 25. April 2020. Verfügbar unter: https://medium.com/merantix/picasso-a-free-open-source-visualizer-for-cnns-d8ed3a35cfc5.

[37] Microsoft. Welcome to TensorWatch; 2019. Online; abgerufen am 25. April 2020. Verfügbar unter: https://github.com/microsoft/tensorwatch.

[38] Hall P, Gill N. Introduction to Machine Learning Interpretability. O'Reilly Media, Incorporated; 2018.

[39] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in neural information processing systems; 2017. S. 4765–4774.

[40] Welling SH, Refsgaard HH, Brockhoff PB, Clemmensen LH. Forest floor visualizations of random forests. arXiv preprint arXiv:160509196. 2016;.