Journalbeitrag

zum Thema

Einsatzes von Explainable Artificial Intelligence zur Unterstützung von Machine-Learning Prognosen in der Immobilienbranche

von

Emir BegovacProf.Dr. Adem Alparslan

Zusammenfassung: Der praxisorientierte Einsatz von künstlicher Intelligenz (KI) in Unternehmen nimmt einen zunehmend hohen Stellenwert ein. Gerade in der Disziplin des Machine Learning (ML) lassen sich bedeutende Erfolge innerhalb der letzten Jahre für viele Bereiche beobachten. Das Machine Learning ML insbesondere in der Vorhersagemodellierung eine wichtige Rolle, um Unternehmen einen wertschöpfenden Geschäftsprozess und den damit verbundenen Wettbewerbsvorteil zu ermöglichen. Viele Nutzer empfinden ML-Algorithmen als nicht verständliche Blackbox, bei denen Entscheidungen und Ergebnisse unklar und undurchsichtig sind. Die erklärende künstliche Intelligenz (XAI) kann hier Abhilfe schaffen, indem die Ergebnisfindung von ML-Modellen interpretierbarer, transparenter und verständlicher dargestellt wird. Die überlegte Anwendung von XAI-Methoden soll Unternehmen ebenso die Möglichkeit verschaffen, neue Perspektiven und neues Wissen im Rahmen der Vorhersagemodellierung erschließen. Nachfolgend wird ein Zwischenstand und eine Übersicht des derzeitigen XAI-Forschungsstandes wiedergegeben. Ergänzt wird der Beitrag durch einen praxisorientierten Einsatz von diversen XAI-Methoden. Auf Basis der Ergebnisse wird eine kritische Begutachtung hinsichtlich des Einsatzes von XAI-Methoden für Unternehmen durchgeführt.

Schlüsselwörter: künstliche Intelligenz · maschinelles lernen · erklärende künstliche Intelligenz · Rashomon-Effekt · Vorhersagemodellierung ·

Abstract The practical application of artificial intelligence (AI) in businesses is gaining increasing importance. Particularly in the field of machine learning (ML), significant successes have been observed in various areas over the past years. ML, especially in predictive modeling, plays a crucial role in enabling companies to create value-added business processes and gain a competitive advantage. Many users perceive ML algorithms as inscrutable black boxes, where decisions and outcomes are unclear and opaque. Explainable artificial intelligence (XAI) can provide a solution by making the decision-making process of ML models more interpretable, transparent, and understandable. The deliberate use of XAI methods should also offer companies the opportunity to explore new perspectives and knowledge within the context of predictive modeling. The following provides an overview of the current state of XAI research. Additionally, the article includes a practical application of various XAI methods. Based on the results, a critical assessment of the use of XAI methods in businesses is conducted.

Keywords: artificial intelligence· Machine Learning · explainable artificial intelligence · rashomon effect · predictive modelling ·

1 Einleitung

In der Ära der Digitalisierung wird der Einsatz von künstlicher Intelligenz (KI) in Unternehmen immer bedeutender. Laut einer Deloitte-Studie halten 94% der weltweit befragten Unternehmen den Einsatz von KI für ihren Erfolg für wesentlich, wobei 76% dieser Unternehmen im Vergleich zu 2021 um 17% mehr in KI investiert haben (Deloitte, 2022). Eine gemeinsame Studie von Ernst&Young und dem Zentralen Immobilien Ausschuss bestätigt diesen Trend auch in der deutschen Immobilienbranche, wo 35% der befragten Unternehmen mehr als 5% ihres Jahresumsatzes in Digitalisierungsmaßnahmen investieren (ZIA & EY Real Estate, 2022). Besonders im Bereich der Immobilienbranche wird maschinelles Lernen (ML) für die Vorhersagemodellierung von Verkaufspreisen als bedeutender Mehrwert betrachtet (Zhou, 2020).
Die Implementierung von ML stellt jedoch viele Unternehmen vor eine Herausforderung. Komplizierte ML-Algorithmen werden als Blackbox-Modelle bezeichnet, da die Ergebnisfindung für den außenstehenden Dritten mathematisch undurchsichtig ist. Das blinde Vertrauen in die intransparente Ergebnisfindung von ML-Modellen kann die Anwendbarkeit und Legitimität des Einsatzes stark schmälern. Infolgedessen nimmt die Nachfrage nach verständlichen ML-Modellen signifikant zu (Adadi & Berrada, 2018; Cajias, 2021). Gartner (2022) prognostiziert, dass 85% der KI-Projekte zu ungenauen Ergebnissen führen werden, hauptsächlich aufgrund mangelnden Wissens über die Ergebnisfindung von KI. Der Einsatz von erklärender künstlicher Intelligenz (XAI) soll Unternehmen helfen, die Ergebnisfindung von ML-Modellen interpretierbarer und verständlicher zu gestalten. Es soll mithilfe von XAI das Vertrauen in die KI gestärkt und einen transparenten und kontrollierten Einsatz in wertschöpfenden Geschäftsprozessen ermöglicht werden. Der damit eingehende Wettbewerbsvorteil ermöglicht Unternehmen eine fortschrittliche Arbeitsweise und fördert gleichzeitig die Akzeptanz und das Vertrauen in den Einsatz von ML in der Unternehmenssphäre (Gunning et al., 2019; Zolanvari et al., 2021).

1.1 Künstliche Intelligenz und Maschinelles Lernen

Der gegenwärtige Einsatz von KI findet in verschiedenen Branchen diverse Anwendungsmöglichkeiten. Dabei ermöglicht der Einsatz von KI beispielsweise intelligente Spracherkennung, datengetriebene Problemlösungen oder mathematische Vorhersagemodellierungen (Mittal & Sharma, 2021). Die Entwicklung der KI durchlief drei Phasen: Die Erste kombinierte Computer und Sensoren zur Datenerfassung und -verarbeitung. In der zweiten Phase wurden Big Data und KI zusammengeführt, um die Interaktion zwischen Daten und KI zu ermöglichen. Die dritte Phase, inspiriert von den vorherigen, beinhaltet das Sammeln von Daten, um die KI lernen und Handlungen wie automatisierte Produktion und Robotik ausführen zu lassen (Liu, 2022). KI ermöglicht den effizienten Abgleich großer Datenmengen durch Datenverarbeitung und Mustererkennung, die in ML Anwendung münden (Mockenhaupt, 2021). Bereits im Jahr 1959 beschrieb Arthur Samuel die Disziplin des ML als ein Fachgebiet, das Computern die Fertigkeiten des Lernens verleiht, ohne eine explizite Programmierung zu erfahren. Mitchell beschrieb im Jahr 1997 ML aus einer technischen Sicht als „ein Computerprogramm [..] das aus Erfahrungen E in Bezug auf eine Aufgabe T und ein Maß für die Leistung P lernt, wenn seine durch P gemessene Leistung bei T mit der Erfahrung E anwächst“ (Geron, 2018). Im Rahmen des ML werden Algorithmen genutzt, deren Aufgabe es ist, Muster und Regelmäßigkeiten in Datensätzen zu erkennen. Die daraus entwickelten

Lösungen befähigt den Algorithmus aus den erschlossenen Erkenntnissen Wissen zu generieren. Dies ermöglicht im Umkehrschluss nicht nur die Bewältigung von Wenn-Dann-Ereignissen, sondern die Lösung von Umstrukturierungsproblemen (Mockenhaupt, 2021). Der Einsatz von ML kann in verschiedene Bereiche unterteilt werden, darunter einfache Aufgaben mit hohem manuellem Aufwand und festen Regeln. ML-Algorithmen reduzieren die Kodierung in komplexen Aufgaben und sind anpassungsfähig an sich ändernde Umgebungen. Letztendlich bieten die Algorithmen die Möglichkeit eines Erkenntnisgewinns, der anhand komplexer Aufgabenstellungen und großen Datenmengen zu erzielen ist. Dabei lassen sich die ML-Systeme durch die Anzahl und Art der Überwachung beim Training einordnen (Geron, 2018). ML weist ein breites Methodenspektrum auf. Daher konzentriert sich dieser Beitrag im Rahmen der Regression auf die Besonderheiten und Anforderungen des überwachten Lernens, welcher Algorithmen wie Explainable Bosting Machine (EBM), K-nearest Neighbor (KNN), Extreme Gradient Boosting (XGB) und künstliche neuronale Netze (ANN) umfasst.

1.2 Erklärende künstliche Intelligenz

Erklärende künstliche Intelligenz ermöglicht es, die Ergebnisfindung von KI-Modellen transparent und interpretierbar darzustellen, indem sie die Gründe des Modells für getroffene Entscheidungen offenlegt (Barredo Arrieta et al., 2020 ; Gunning et al., 2019). Die Definition von Interpretierbarkeit und Erklärbarkeit ist mathematisch anspruchsvoll, wobei nicht-mathematische Ansätze die Fähigkeit des menschlichen Verständnisses betonen, die Ursachen von Entscheidungen zu verstehen. Andererseits wird die Interpretierbarkeit und Erklärbarkeit mit der Fähigkeit des menschlichen Verständnisses gleichgesetzt, Ergebnisse von Modellen auf Basis der bestehenden Daten Vorhersagen zu können (Kim et al., 2016; Miller, 2019). Die bestehende XAI-Methoden weisen einen unterschiedlichen Umfang an Erklärung auf, die dem Nutzer geboten werden (Das & Rad, 2020). XAI-Methoden variieren im Umfang der bereitgestellten Erklärungen, wobei zwischen interpretierbaren, spezifischen und agnostischen Modellen unterschieden wird. Letzteres stellt einen Ansatz dar, der die Erklärung des ML-Modells separat und getrennt vom Algorithmus ausführt. Der Einsatz findet überwiegend nach dem eigentlichem Modell-Training statt. Generell haben agnostische Modelle den Vorteil, dass eine flexible Auswahl von ML-Algorithmen stattfinden kann und Erklärungen diverser ML-Modelle miteinander verglichen werden können. Spezifische Methoden können nur auf einen bestimmten Typen angewendet werden, wie beispielsweise auf Baumartige- oder Deeplearning Algorithmen. Die Anwendung von spezifischen Modellen findet ebenso nach dem Training oder innerhalb der Evaluierung statt (Ribeiro et al., 2016). Interpretierbare Modelle werden auch als Whitebox-Modelle beschrieben und bilden das komplementär zu den Blackbox-Modellen. Diese agieren für den Menschen nach verständlichen Mustern, Regeln und Entscheidungen (Loyola-Gonzalez, 2019). Agnostische XAI-Methoden können lokal oder global sein, wobei globale Methoden das Verständnis fördern, in welcher gesamtheitlichen Beziehung die unabhängigen Variablen mit dem Modellergebnis und demnach der zu vorhersagenden Zielvariable stehen. Die globale Perspektive birgt jedoch das Risiko von Annäherungswerten, die nicht auf einzelne Beobachtungen zutreffen könnten. Hierfür werden lokale Interpretationen eingesetzt, die das Vorhersageergebnis für eine einzelne oder ähnlich gruppierte Beobachtungen interpretierbar machen. Die

Untersuchung einer einzelnen Instanz unterstützt dabei die Förderung des Vertrauens hinsichtlich der Möglichkeit für Nutzer, Modellergebnisse auf Faktoren wie ethische Korrektheit, Fairness, Stabilität, Abhängigkeiten und Plausibilität zu untersuchen (Barredo Arrieta et al., 2020). Im Kontext von ML-Algorithmen ist es weit verbreitet, dass Blackbox-Modelle eine hohe Vorhersagegenauigkeit aufweisen, die Interpretierbarkeit der Modelle jedoch schmälert. Für den nachfolgende Beitrag ist es elementar von Bedeutung zu verstehen, dass viele Artikel fälschlicherweise von einer höheren Genauigkeit der Modelle sprechen, obwohl eigentlich die Performance gemeint ist. Demnach erzielen Blackbox-Modelle nach den Autoren eine bessere Leistung als durchsichtige Whitebox-Modelle (Hall & Gill, 2018; Saeed & Omlin, 2021; Sahakyan et al., 2021; Z. Zhang et al., 2022). Die Genauigkeit kann in Rahmen der Klassifikation benannt werden, jedoch nicht in den übrigen Disziplinen des ML. Die Einordnung der verwendeten Modelle innerhalb der Performance- und Interpretierbarkeitsspektrum ist der Abbildung 1 zu entnehmen und erfolgt dabei nach den Erkenntnissen von Barredo et al. (2020) und Nori et al. (2019).

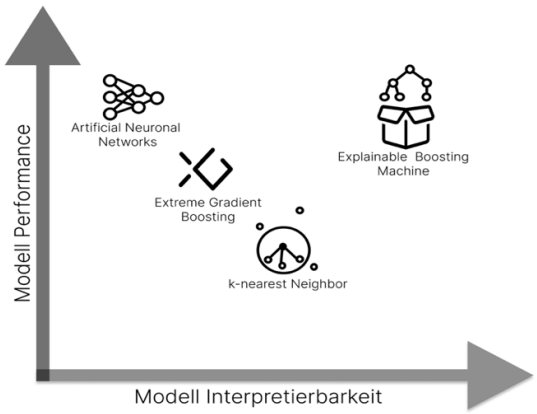

Abbildung 1: Performance gegen Interpretierbarkeit (Barredo Arrieta et al., 2020, Nori et al., 2019)

2 Untersuchungsrahmen

- Der Datensatz

Der verwendete Datensatz *Housing prices Ames, Iowa* wird seitens OpenIntro zur Verfügung gestellt. Die Daten sind von Dean De Cock (2011) im Jahre 2011 erhoben worden und sollen den Einsatz von ML-Algorithmen anhand von Echtdaten ermöglichen, die nachfolgend empirische Analysen über den Immobilienmarkt der Stadt Ames zulassen. Die Daten sind unter folgendem Weblink erreichbar: https://openintro.org/data/index.php?data=ames

- Methodik der Literaturrecherche und Ergebnisse

Für die Evaluierung der Modelle wurde eine Literaturrecherche nach vom Brocke (2009) durchgeführt, um einflussreichen Variablen für die Immobilienpreisfindung anhand der Literatur zu identifizieren. Hierfür sind 61 Journalartikel, 41 Magazinaufsätze und19 Konferenzpaper untersucht worden. Relevante Literatur wird anhand eines Hirsch-Index über 30 oder eines Q-Rankings von Q1 oder Q2 für 2023 ausgewählt. Reports werden als signifikant eingestuft, sofern diese aus Institutionen mit staatlichem Bezug entspringen. Diese Limitationen sollen den Untersuchungsrahmen mit qualitativ hochwertiger Literatur stärken und zur Ergebnisbewertung dienen. Eine wesentliche Erkenntnis ist, dass derzeit vorwiegend noch hedonische

Ansätze zur Preisbestimmung von Immobilien gewählt werden. Hochaktuell werden nicht nur hedonische- und ML-Methoden separat untersucht, sondern ebenso die Synergieeffekte der Kombination beider Ansätze. Dabei liegt ein Forschungsschwerpunkt darauf, wie Big Data sich mit dem hedonischen Konzept mittels ML vereinen lässt (Potrawa & Tetereva, 2022; Wei et al., 2022; Zaki et al., 2022). In den letzten Jahren wird vermehrt auf Algorithmen gesetzt, die eine sehr gute Performance bei der Prognose erzielen. Dabei fokussierten sich die diversen Untersuchungen hauptsächlich darauf, die Genauigkeit des ML-Modells sukzessiv zu verbessern (Alfaro-Navarro et al., 2020; Alzain et al., 2022; Bin et al., 2017; Pai & Wang, 2020; Pinter et al., 2020; Quang et al., 2020; Rampini & Re Cecconi, 2021; Zhou, 2020). Einige Studien untersuchen Whitebox-Modelle, Teilfunktionen von XAI-Methoden oder die Vorteile von XAI für die Vorhersage von Immobilienpreisen. Seitens diverser Autoren wird ein hoher Forschungsbedarf geäußert, dem der nachfolgende Beitrag nachkommen wird (Cajias, 2021; Dimopoulos & Bakas, 2019; Lorenz et al., 2021; Potrawa & Tetereva, 2022).

- Entwicklung und Validierung der Vorhersagemodelle

Die Entwicklung von 4 Vorhersagemodellen erfolgt anhand des Cross Industry Standard Process for Data Mining Prozesses. Für diesen Beitrag sind die wesentlichen konzentrierten Phasen des Modells das Data Understanding, die Data Preparation, das Modeling und die Evaluation (Shafique & Qaiser, 2014). Darüber hinaus werden klare Leistungsmetriken definiert, um die Ergebnisse verschiedener ML-Algorithmen vergleichbar zu machen. Dies ermöglicht eine aussagekräftige Beurteilung der XAI-Methoden anhand generalisierbarer Modelle. Die Verwendeten Metriken entsprechen dabei dem Mean absolute Error, Root mean squared Error, mean average per cent Error und dem adjusted R^2 (Qi et al., 2020; Willmott & Matsuura, 2005; de Myttenaere et al., 2016; Chicco et al., 2021). Zur Validierung der Modelle werden die Ergebnisse des Trainingsdatensatzes mit den Werten der Testdaten verglichen (Carvalho et al., 2019).

- XAI-Methoden

Die 3 verwendeten XAI-Verfahren beschränken sich dabei auf Shapley Additive Explanations (SHAP), Local interpretable modelagnostic explanations (LIME) und die InterpretML Bibliothek. Nach dem aktuellen Stand existiert keine einheitlich definierten Metriken, die eine Messung der Qualität der Erklärbarkeit von XAI ermöglichen. Dies erschwert folglich die Evaluierung hinsichtlich der Erklärungsqualität. Hierfür werden nachfolgend die XAI-Methoden hinsichtlich der Eigenschaften der Stabilität, Trennbarkeit, Konsistenz und Erklärungsähnlichkeit untersucht. Dies soll eine Gegenüberstellung der eingesetzten Methoden ermöglichen und als Grundlage zur Gesamtheitlichen Evaluation dienen. Die definierten Eigenschaften in der Tabelle 1 beruhen auf den Erkenntnissen der Autoren Molnar (2019) Molnar et al., (2022), Sokol & Flach, (2020) und Semenova et al. (2022).

Tabelle 1: Erklärungseigenschaften

Erklärungseigenschaften	**Beschreibung**
Stabilität	Für den gleichen Datenpunkt wird stets die gleiche Erklärung generiert
Trennbarkeit	Unterschiedliche Datenpunkte führen zu unterschiedlichen Erklärungen

Konsistenz	Ähnliche Erklärungen werden für minimal unterschiedliche Dateninstanzen generiert
Erklärungsähnlichkeit	Ähnliche Erklärungen mithilfe unterschiedlicher XAI-Verfahren
Berechnungszeit	Dauer der Berechnungszeit der XAI-Methoden

3 Ergebnisse Modelentwicklung

ber alle Leistungsmetriken hinwegalle Modelle für die Trainings- und Testdatensatz gute Ergebnisse auf und somit auch ein Maß an Generalisierbarkeit.

Tabelle 1:

Trainingsdaten				
	MAE	**MAPE**	**RMSE**	**adjusted R2**
ANN	12.490,70	0,075	26.969,37	0,87
KNN	15.606,72	0,095	26.360,04	0,88
XGB	12.256,29	0,077	17.033,15	0,95
EBM	14.243,16	0,089	19.968,41	0,93
Testdaten				
ANN	14.173,09	0,081	26.049,45	0,90
KNN	18.344,66	0,103	29.823,08	0,87
XGB	15.247,09	0,089	25.321,29	0,90
EBM	15.892,23	0,094	23.843,25	0,91

Dadurch kann im nächsten Schritt die Erklärbarkeit durch XAI bei validierten Modellen untersucht werden, wobei die Trainingsdaten als Basis dienen. Dies soll die Anzahl an Dateninstanzen verringern, da die Berechnungszeit der meisten XAI-Methoden linear mit der Anzahl an mehr Datenpunkten steigen. Für die globale Erklärung von XAI lässt ich festhalten, dass alle Modelle im Bereich der einflussreichsten Variablen überwiegend Gemeinsamkeiten aufweisen. Unterschiede bestehen hauptsächlich in der Höhe der berechneten Beiträge, die nicht nur innerhalb der Methoden, sondern auch zwischen den Methoden erkennbar sind. Es gibt auch Differenzen in der Berechnungszeit der verschiedenen Methoden, wobei selbst innerhalb der SHAP-Methode je nach Algorithmus Unterschiede zu erkennen sind. Eine übermäßige Bevorzugung eines Skalenniveaus kann nicht beobachtet werden. Ein Auszug der Ergebnisse der globalen und lokalen Betrachtung ist der Tabelle 2 zu entnehmen.Die lokale Erklärbarkeit wird mittels der ersten beiden Datenpunkte des Trainingsdatensatzes untersucht. Im Gegensatz zur globalen Beobachtung zeigen sich hier diverse Unterschiede. Es ist festzustellen, dass die einflussreichsten Variablen für jedes Modell unterschiedlich nach SHAP bewertet werden. Sowohl die Reihenfolge als auch der quantitativ ausgewiesene Einfluss der Variablen variieren je nach Modell. Die lokale LIME-Betrachtung weist im Vergleich zu SHAP mehr Ähnlichkeiten zur komplementären globalen Beobachtung auf. Die aufgeführten

einflussreichen Variablen der Modelle entsprechen weitgehend der globalen Beobachtung. Die Berechnungsdauer beläuft auf zwischen fünf bis zehn Minuten und ist demnach wesentlich schneller als SHAP. Es ist jedoch zu beachten, dass ein programmiertechnischer Mapping-Aufwand im Vorfeld betrieben werden muss, um die gewünschten Ergebnisse zu erzielen. Diese erhöht die reale Berechnungszeit pro Algorithmus um 5 Minuten. Die Verwendung des lokalen InterpretML-Moduls weist im Rahmen der einflussreichsten Variablen Ähnlichkeiten zu den Ergebnissen der lokalen SHAP-Betrachtung auf. Dies ist darauf zurückzuführen, dass die lokale Methode von InterpretML dem KernelExplainer von SHAP entspricht, und somit die Ergebnisse oft identisch sind, insbesondere im Hinblick auf das XGB- und EBM-Modell. Dies trifft auch auf die Beitragshöhen der Variablen zu. Lediglich das Verhältnis an quantitativen und kategorialen Variablen des EBM-Modells unterscheidet sich im Vergleich zu den Ergebnissen anderer XAI-Methoden. Insgesamt lässt sich jedoch im Rahmen der Berechnungszeit beobachten, dass das EBM-Model neben dem XGB- und KNN-Model am schnellsten performt.

Abbildung 2: Ergebnisse Globalen und lokalen Beobachtung

SHAP Global					
Algorithmus	**Top 3 Variablen**	**Beitragshöhe Top 3**	**Beitrag übrigen Variablen**	**Top 10 Verhältnis qualitativ / kategorial**	**Berechnungsdauer**
ANN	Gr Liv Area Overall Qual TotalSF	11.604,97 $ 8.136,93 $ 7.985,08 $	126.208,61	5 / 5	1 Std 12 Min
KNN	Gr Liv Area Overall Qual TotalSF	5.011,06 $ 4.310,02 $ 4.131,30 $	4.9682,88	6 / 4	0 Std 25 Min
XGB	TotalSF Overall Qual Year Built	19.329,30 $ 15.418,98 $ 4.794,59 $	22.160,17	7/3	0 Std 34 Min
EBM	Gr Liv Area TotalSF Overall Qual	5.283,90 $ 4.615,64 $ 4.568,94 $	7.9754,74	6 / 4	3 Std 58Min
LIME Global					
ANN	Pool Qc Gr Liv Area TotalSF	29.054,13 $ 28.667,78 $ 20.065,33 $	-	4/6	0 Std 20 Min
KNN	TotalSF Gr Liv Area Garage Area	77.080,43 $ 22.132,35 $ 18.985,63 $	-	4/6	0 Std 15 Min
XGB	TotalSF Overall Qual Bath	50.053, 89 $ 25.409,99 $ 12.897,75 $	-	5/5	0 Std 12 Min
EBM	Garage Qual TotalSF Gr Liv Area	17.315,90 $ 15.384,68 $ 15.110,86 $	-	4/6	0 Std 14 Min
InterpretML					
KNN	Pool Qc TotalSF Total BsmtSF	110.880,30 $ 92.504,82 $ 77.531,36 $	-	5/5	0 Std 0,5 Min
XGB	TotalSF Overall Qual 1st Flr SF	14.0745,40 $ 120.058,40 $ 64.592,96 $	-	6/4	0 Std 0,5 Min
EBM	TotalSF Gr Liv Area Bath	4.187,17 $ 4.119,42 $ 3.770,83 $		6/4	0 Std 0,5 Min

SHAP lokal 1					
Algorithmus	**Top 3 Variablen**	**Beitragshöhe Top 3**	**Beitrag übrigen Variablen**	**Top 10 Verhältnis qualitativ / kategorial**	**Berechnungsdauer**
ANN	Neighborhood_Crawford Gr Liv Area Bldg Type_Duplex	19.757,26 $ 12.047,59 $ -9.332,59 $	-25.540,46 $	2 / 8	1 Std 12 Min
KNN	Year Remod/Add Garage Year Build Fireplaces	-7.705,57 $ -6.171,17 $ -5.635,38 $	-11.715,39 $	5 / 5	0 Std 25 Min
XGB	Overall Qual Neighborhood_Crawford Year Remod/Add	-15.975,10 $ 11.245,33 $ -6.953,71 $	15.860,14 $	4 /6	0 Std 34 Min
EBM	Garage Cars Garage Area Gr Liv Area	12.477,28 $ 12.317,56 $ 8.342,71 $	-31.462,87 $	4/6	3 Std 58Min
LIME lokal 1					
ANN	Pool Qc Gr Liv Area TotalSF	-34.753,34 $ 28.200,26 $ 20.063,75 $	-	3/7	0 Std 10 Min
KNN	Pool Qc Gr Liv Area TotalSF	-47.780,00 $ 16.191,91 $ 12.439,27 $	-	4/6	0 Std 6 Min
XGB	TotalSF Overall Qual Bath	50.429,00 $ -25.501,43 $ -13.380,83 $	-	5/5	0 Std 5 Min
EBM	Garage Qual Gr Liv Area TotalSF	-17.240.01 $ 16.276,26 $ 15.796,24 $	-	5/5	0 Std 6Min
InterpretML lokal 1					
KNN	Fireplaces Kitchen Qual Year Remod/Add	- 4.243,86 $ - 4.036,13 $ - 3.756,07 $	-	4/6	0 Std 10 Min
XGB	Overall Qual Neighborhood_Crawford Kitchen Qual	-16.519,64 $ 11.634,05 $ - 6.609,72 $	-	4/6	0 Std 8 Min
EBM	Garage Area Garage Cars Neighborhood_Crawford	12838,34 $ 12573,79 $ 8244,05 $	-	3/7	0 Std 1 Min

SHAP lokal 2					
Algorithmus	**Top 3 Variablen**	**Beitragshöhe Top 3**	**Beitrag übrigen Variablen**	**Top 10 Verhältnis qualitativ / kategorial**	**Berechnungsdauer**
ANN	Land Slope_Sev Bsmt Exposure_Gd Kitchen Qual	16.327,76 $ 13.159,48 $ 13.080,93 $	-18.786,42 $	1 / 9	1 Std 12 Min
KNN	Fireplaces Kitchen Qual Bath	16.486,13 $ 5.367,53 $ 5.042,63 $	16.191,75 $	5 / 5	0 Std 25 Min
XGB	Overall Qual Fireplaces Bsmt Exposure_Gd	14.174,41 $ 7.325,87 $ 6.473,00 $	-1.633,86 $	4 /6	0 Std 34 Min
EBM	Kitchen Qual Fireplaces Land Slope_Sev	17.450,09 $ 10.891,55 $ 8.225,49 $	4.580,38 $	4 /6	3 Std 58Min
LIME lokal 2					
ANN	Pool Qc Overall Qual Bldg Type_Fam	32.183,29 $ 17.318,58 $ 16.464,79 $	-	2/8	0 Std 8 Min
KNN	Pool Qc Overall Qual Kitchen Qual	46.860,41 $ 10.579,92 $ 10.533,53 $	-	4/6	0 Std 6 Min
XGB	Overall Qual Garage Cond Bath	25.746,40 $ 19.724,73 $ 13.347,87 $	-	2/8	0 Std 6 Min
EBM	Garage Qual Kitchen Qual Land Slope_Sev	19.368,48 $ 17.366,09 $ 16.210,82 $		4/6	0 Std 6 Min
InterpretML lokal 2					
KNN	Fireplaces Kitchen Qual Fireplace Qual	9.824,07 $ 9.329,74 $ 4.468,13 $		5/5	0 Std >1 Min
XGB	Overall Qual Fireplaces Bsmt Exposure_Gd	-14.673,06 $ 7.158,36 $ 6.532,65 $	-	4/6	0 Std >1 Min
EBM	Kitchen Qual Fireplaces Bsmt Exposure_Gd	17.152,43 $ 10.596,74 $ 8.279,41 $		3/7	0 Std >1 Min

3.1 Evaluation der Erklärungseigenschaften

Die Resultate der Analyse zur Stabilität, Trennbarkeit und Konsistenz sind in Tabelle 4 aufgeführt. Bezüglich der Stabilität fällt auf, dass nahezu alle angewendeten XAI-Verfahren für die lokale Analyse diese Eigenschaft verletzen. Eine Ausnahme bilden das XGBoost-Modell unter Anwendung des TreeExplainers und das EBM-Modell unter Anwendung des Glasbox Explainers. Hierbei zeigen die Gewichtungen der Variablen Ähnlichkeiten zum initialen Durchlauf, stellen jedoch keine identische Erklärung dar. Diese Beobachtung erstreckt sich auf alle Modelle und stellt die Herausforderung eines instabilen Modells dar. Im Rahmen der Trennbarkeit ist anhand der lokalen Analysen der Beobachtung 1 und 2 bereits zu erkennen, dass alle XAI-Verfahren unter Anwendung aller Algorithmen das Kriterium erfüllen. Im Hinblick auf das Konsistenzkriterium werden die Resultate weiter geprüft. Dabei zeigt sich, dass minimale Änderungen der Datenpunkte zu ähnlichen Resultaten in den Beobachtungen führen. Es ist entscheidend, welche Variable angepasst wird. So bewirken geringfügige Anpassungen einflussreicher Variablen größere Veränderungen in der Erklärung als die Modifikation anderer Variablen, deren Beiträge für die Vorhersage weniger einflussreich sind. Diese Beobachtung erstreckt sich auf sämtliche Modelle und XAI-Anwendungen.

Tabelle 2: Ergebnisse Stabilität, Trennbarkeit und Konsistenz

⊗ erfüllt o teilweise erfüllt x nicht erfüllt

XAI-Verfahren	**Stabilität**	**Trennbarkeit**	**Konsistenz**
SHAP Explainer	x	⊗	o
SHAP TreeExplainer	⊗	⊗	o
LIME	x	⊗	o
InterpretML - KernelExplainer	x	⊗	o
InterpretML – Glassbox Explainer	⊗	⊗	o

Abschließend erfolgt die Auswertung der Erklärungsähnlichkeit. Hierbei zeigt sich global betrachtet, dass alle Modelle durch die angewendeten XAI-Verfahren ähnliche Erklärungen liefern, die sich jedoch punktuell unterscheiden. Bei genauer lokaler Betrachtung für die Beobachtungen 1 und 2 wird deutlich, dass die Gewichtungen der Variablen je nach Modell und XAI-Verfahren weniger Ähnlichkeiten aufweisen und größtenteils keine übereinstimmende Erklärung bieten. Demnach unterscheidet sich die Ergebnisfindung der Modelle im übergreifenden Vergleich erheblich. In Abbildung 2 wird dies anhand des ersten Datenpunktes des Trainingsdatensatzes für das XGB-Modell verdeutlicht. Hierbei wird deutlich, dass für den gleichen Datenpunkt unter Anwendung desselben Modells unterschiedliche Gewichtungen und Erklärungen seitens SHAP und LIME ausgegeben werden. Während LIME die Garagenqualität mit einem negativen Beitrag bewertet, gibt SHAP an, dass dieselbe Variable positiv zur Ergebnisfindung beiträgt.

Wortwörtlich bedeutet dies, dass das Ergebnis des TreeExplainers für das XGB-Modell die Garagenqualität, die mit einer 3 angegeben ist und somit einer durchschnittlichen Garagenqualität entspricht, als positiv für die Ergebnisfindung einstuft und somit zur Erhöhung des Immobilienpreises um 652 $ beiträgt. Für den gleichen Datenpunkt bewertet LIME die durchschnittliche Garagenqualität als negativen Faktor und impliziert dadurch eine Senkung des Preises um 3540 $. Hier lässt sich eindeutig ein Widerspruch in den Erklärungen erkennen. Angesichts der ähnlich guten Vorhersageergebnisse der multiplen Modelle wird das soeben beschriebene Phänomen als Rashomon-Effekt definiert. Dieser Effekt impliziert, dass verschiedenen statistische- oder ML-Modelle zu ähnlich guten Vorhersageergebnisse für denselben Datensatz führen können. In der vorliegenden Beobachtung manifestiert sich der Rashomon-Effekt nicht nur hinsichtlich der Vielfalt von guten Modellen, sondern auch innerhalb der Vielfalt von guten Erklärungen (Breiman, 2001; Leventi-Peetz & Weber, 2023).

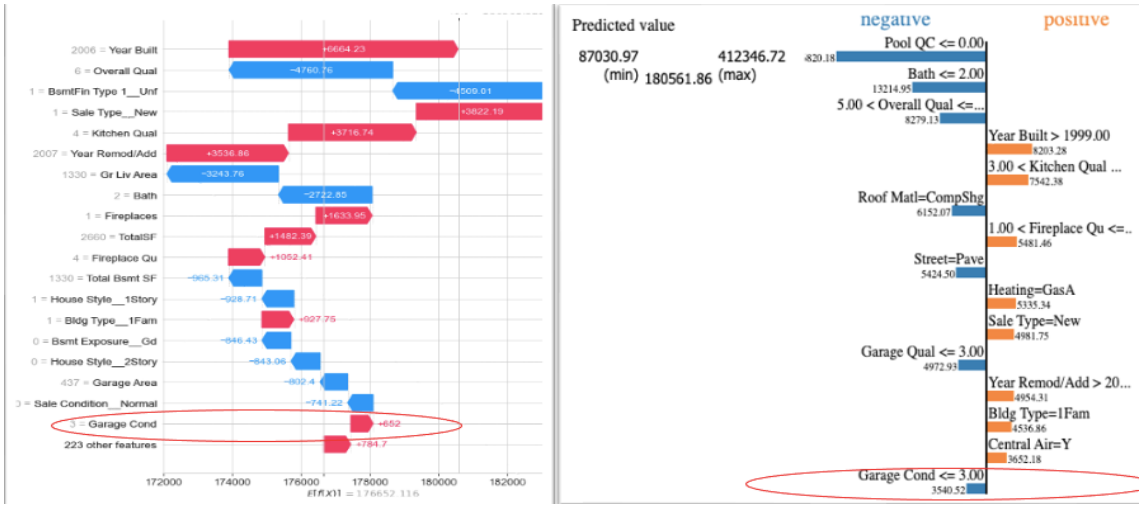

Abbildung 2: Rashomon Effekt für das XGB-Modell

4 Beurteilung der Forschungsergebnisse

Zunächst werden einige positiven Aspekte der Ergebnisse in Kurzform wiedergegeben:

- XAI bietet durch die Gewichtung von Variablen einen wesentlichen Vorteil zur Deutung Ihres Beitrages für die Ergebnisfindung. Globale Betrachtungen schaffen ein generelles Verständnis der Variablenbedeutung, während lokale Analysen die Treiber für der Modellentscheidung für einzelne Datenpunkte identifizieren. Die Quantifizierung der Variablenbeiträge erlaubt nicht nur Vorhersagen, sondern auch die Messung der Inputs-Auswirkungen (Hall et al., 2017; Owen & Prieur, 2017)
- XAI ermöglicht die Nutzung von Optimierungspotenzialen, die Identifizierung von Modellfehlern und die Untersuchung struktureller Schwächen (Fryer et al., 2021; Gerlings et al., 2021)
- Die Gewichtung der Variablenbeiträge erlaubt die Prüfung von Fairness, ethischer Korrektheit und Risikokontrolle. Unter Beachtung des Rashomon-Effektes ist das ein notwendiger Schritt, um die Entscheidung für ein Modell nicht nur auf Basis der Performance zu treffen. Die ethische

Korrektheit von KI wird in der Forschung und Praxis betont, mit XAI als Lösungsansatz (Gerlach et al., 2022; John-Mathews, 2022,; Vale et al., 2022; G. Zhang, 2022)

- Variablenbeiträge sind besonders wichtig in rechtlichen, cybersicherheits- und medizinischen Kontexten.
- Das EBM-Modell erreicht ähnlich gute Leistung wie Blackbox-Modelle, was Fragen zur Notwendigkeit von XAI für Blackbox-Modelle aufwirft. Befürworter argumentieren, dass die Transparenz bereits im Ansatz des Modells verankert sein sollte, um von Anfang an einen Faktor für die Vertrauensbildung zu schaffen und nicht erst nachträglich beizusteuern (Rudin, 2019)

4.1 Stabilität, Trennbarkeit, Konsistenz und Berechnungszeit

Eine wesentliche Erkenntnis der Ausarbeitung ist, dass bei erneutem Durchlauf der genutzten XAI-Verfahren die Ergebnisse der lokalen Erklärung variieren. Alvarez-Melis und Jaakkola (2018) unterstützen die Beobachtung, indem ihre Untersuchung zeigt, dass dieselben Erklärungen für zwei nahezu identische Datenpunkte unterschiedliche Ergebnisse produzieren können. Dies ist auf die Unsicherheit der modellagnostischen Methoden zurückzuführen, die aus der Zufälligkeit der Stichproben, der Variation in der Nähe zu den Stichproben und der Variation in der Glaubwürdigkeit des erklärten Modells für verschiedene Datenpunkte resultiert (Y. Zhang et al., 2019). Zwar unterscheiden sich die Erklärungen der XAI-Methoden begrenzt und weisen den Variablen ähnliche Beitragshöhen zu, jedoch können bereits geringe Änderungen zu unterschieden in den Ergebnissen führen. Demnach sind die Ergebnisse jedes Berechnungsdurchlaufes mit Vorsicht zu beobachten und kritisch zu hinterfragen. Hinzu kommt, dass im Rahmen der Stabilität von Wichtigkeit ist, anhand erneuter Durchläufe die stabilen Variablen zu identifizieren. Es ist ebenso zu beachten, dass zuverlässige und stabile XAI-Ergebnisse ebenfalls durch mögliches vorhandenes und nicht ausreichend bekämpftes Datenrauschen beeinträchtigt werden können (Ribeiro et al., 2016; Rosenfeld, 2021). Im Gegensatz dazu sind die Ergebnisse des TreeExplainers stabiler als die der anderen XAI-Verfahren, da die baumartige Grundstruktur robuster gegenüber Veränderungen im Datensatz ist und nicht auf Annäherungen basiert (Lundberg et al., 2020). Dies trifft auch auf die lokale Erklärbarkeit des EBM-Modells zu, wenn der Glassbox-Explainer genutzt wird (Rudin, 2019). Für die Trennbarkeit zeigt sich, dass alle XAI-Verfahren unter allen Modellen diese in den Ergebnissen vereinbaren. Für die Konsistenz erfüllen die Modelle teilweise die Bedingungen. Wie bereits beschrieben, führen minimale Änderungen zu keinen schwerwiegenden Änderungen in der Erklärung. Dies ist jedoch abhängig von der betrachteten Variable sowie dem verwendeten Algorithmus. So zeigt sich, dass gerade die baum- oder distanzartigen Algorithmen stärker auf Änderungen in den Daten reagieren (Arsov et al., 2019). Eine weitere Herausforderung stellt dabei die Berechnungszeit der einzelnen Methoden dar. So unterscheiden sich diese je nach Algorithmus und verwendeter XAI-Methode. Dabei weist die Berechnung des TreeExplainers unter Verwendung des XGB-Modells die schnellste Berechnungszeit vor. Der Grund hierfür ist, dass die Berechnung nicht exponentiell verläuft, sondern polynominal. Ein weiterer Faktor für eine höhere Berechnungszeit ist die Komplexität des Algorithmus. Da die Erklärbarkeit der Ergebnisfindung beispielsweise des KNN-Modells mit einem einfacheren Aufwand für SHAP verbunden ist als für das ANN-Modells, fällt die Berechnungszeit dementsprechend geringer aus (Yang, 2021).

4.2 Erklärungsähnlichkeit

Gerade im Hinblick auf die Erklärungsähnlichkeit rückt besonders der Rashomon-Effekt in den Vordergrund. Dieser stellt eine gravierende und offensichtliche Herausforderung im Hinblick auf die Multiplizität von Vorhersagergebnissen und deren Erklärbarkeit dar. Gerade für systemkritische Entscheidungen und Branchen ist eine transparente und nachvollziehbare Ergebnisfindung von ML-Modellen unabdingbar (Gerlings et al., 202; Sahakyan et al., 2021; Zolanvari et al., 2021). In der ersten Instanz verhilft der Einsatz von XAI zu einer Verringerung des Rashomon-Effekts bezüglich der Multiplizität von guten Modellen. Jedoch zeigt gerade die lokale Analyse, dass die Erklärung der XAI-Methoden Wiedersprüche und Unterschiede aufweisen. Demnach lässt sich beobachten, dass eine Multiplizität von guten Erklärungen besteht und das widersprüchliche Erklärungen je nach Modell und XAI-Verfahren auftreten können.

(Leventi-Peetz & Weber, 2023). Dies stellt zusätzlich eine erhebliche Herausforderung für die praktische Umsetzung dar, insbesondere in Bezug auf die Akzeptanz, das Vertrauen und das Verständnis für Erklärung der ML-Prognosen. Zusätzlich können Aggregationseffekte oder bestehende Multikollinearität von wesentlichem Einfluss sein, die zusätzlich nochmals die Erklärungsähnlichkeit gravierend beeinflussen. Ebenso erschwert die Aggregation von mehrerer Entscheidungsfaktoren, bestimmte Entscheidungen verstehen oder erklären zu können. Ein Problem, das sich hierbei ergeben kann, ist, dass die Aggregation von Faktoren einen Verlust an Feinheit und Detailgenauigkeit bei der Erklärung der Ergebnisfindung bedeuten kann. Es kann ebenfalls schwierig sein, eine Übersicht über die Gesamtheit der Entscheidungsfaktoren zu gewinnen. Daher ist es essenziell eine geeignete XAI-Methode zur Aggregation der Entscheidungsfaktoren zu wählen, die ein ausgewogenes Verhältnis zwischen Verständlichkeit und Detailgenauigkeit ermöglicht (Aas et al., 2021; Kenny et al., 2021).

4.3 Bekämpfung der Herausforderungen

Trotz vieler guter Modelle bleibt der verbreitete Ansatz, verschiedene XAI-Verfahren zu vergleichen. Wenn keine ähnlichen Erklärungen vorliegen, sind kritische Überlegungen entscheidend. Es gibt verschiedene Ansätze, um spezifischen Herausforderungen zu begegnen. Li et al. (2023)testen die neue XAI-Methode "mean centroid prediction difference" (PredDiff), die die Konsistenz und Berechnungszeit im Vergleich zu SHAP wesentlich verbessert. Sovrano & Vitali (2023) haben die "Degree of Explainability" (DoX)-Metrik entwickelt, um den Grad der Erklärbarkeit des XAI-Outputs anzuzeigen. DoX bewertet die Vollständigkeit, nicht den Wahrheitsgehalt. Der Datenpool bildet das Fundament des DoX-Wertes und eignet sich besonders für juristische Themen. Auch der Einsatz von Whitebox-Modellen wird seitens diverser Autoren erneut betont, um den Rashomon-Effekt punktuell entgegentreten zu können (Rudin, 2019; Semenova et al., 2022). Gerade die Herausforderung der widersprüchlichen Erklärungen zeigt auf, wie dringend Lösungsansätze in Form von Metriken oder Gütekriterien benötigt werden. Ein nicht existierender wissenschaftlicher Konsens bezüglich einer einheitlichen Methode mit standardisierten Metriken wird perspektivisch die Erfolge der Forschung von XAI negativ beeinflussen. Ein Vergleich auf Basis messbarer Kriterien ist entscheidend, um die Eignung und Leistung der verschiedenen XAI-Methoden zu beurteilen. Obwohl XAI den Rashomon-Effekt hinsichtlich der Vielfalt von guten Modellen

reduzieren kann, entstehen neue Herausforderungen, wie die Existenz von mehreren guten Erklärungen und die möglichen Inkonsistenzen zwischen ihnen. Die Schwere des Rashomon-Effekts kann damit nur minimiert, nicht eliminiert werden (Barredo Arrieta et al., 2020; Leventi-Peetz & Weber, 2023; Weerts et al., 2019). Die Definition von Metriken kann dem Beispiel von Konfidenzintervallen folgen, um einen Vertrauensbereich für die Erklärbarkeit zu schaffen. Denkbar ist auch die Definition von einheitlichen Leistungsmetriken, die angeben, wie vertrauenswürdig die Erklärung unter Annahme des Rashomon-Effektes oder der Multikollinearität ist. Der kontrafaktische Erklärungsansatz von Dandl et al. (2020) und Mothilal et al. (2020) könnte eine mögliche Richtung vorgeben. Es gibt vielversprechende Ansätze, die qualitative und quantitative Metriken definieren. Die Forschung ist sich einig, dass fehlende Einheitlichkeit bei der Einordnung der Erklärbarkeit das Vertrauen und Verständnis der Nutzer sukzessive verringern wird (Hoffman et al., 2018; Hsiao et al., 2021; Mohseni et al., 2021; Rosenfeld, 2021; Sisk et al., 2022).

5 Implikationen für die Praxis

Die Erkenntnisse dieser Arbeit sind in der Immobilienbranche und darüber hinaus anwendbar. Der Einsatz von XAI ermöglicht die Identifizierung wirtschaftskonformer Modelle, die juristische, ethische und risikobedingte Anforderungen erfüllen und den wirtschaftlichen Erfolg steigern können. Durch Interpretierbarkeit und Transparenz können wirtschaftliche Schäden reduziert werden, indem Missstände frühzeitig erkannt werden. Die Instabilität der Erklärungen stellt eine Herausforderung dar, die durch die Verwendung von baumbasierten und Whitebox-Modellen eingeschränkt werden kann. Globale und lokale Erkenntnisse können fachübergreifend genutzt werden, um für bisher undurchsichtige ML-Implementationen zu sensibilisieren. Die Visualisierung von Variablenbeiträgen ermöglicht eine transparente Prognose und schafft eine datenfundierte Argumentationsgrundlage für die Bestimmung der weiteren Strategieausrichtung. Im Kontext des Rashomon-Effekts stellt die Vielzahl an guten Erklärungen eine Herausforderung dar, da widersprüchliche Ergebnisse die Akzeptanz und das Vertrauen in die Ergebnisfindung von ML-Modellen signifikant senken können. Die fehlende Einheitlichkeit von Metriken zur Evaluierung der Erklärbarkeit führt zu unkorrekten Argumentationsgrundlagen, die in Fehlentscheidungen resultieren wirtschaftliche Verluste herbeiführen können. Es ist entscheidend, einheitliche Metriken für den XAI-Einsatz zu entwickeln, um die Erklärbarkeit in Forschung und Wirtschaft zu bewerten und eine fundierte Beurteilung sicherzustellen. Eine Stagnation in diesem Bereich könnte das Verständnis zwischen Mensch und Maschine beeinträchtigen und schwerwiegende Folgen für die praktische Implementierung haben.

6 Zusammenfassung und Fazit

Der Beitrag hebt diverse Möglichkeiten hervor, wie der praktische Einsatz von XAI-Methoden das Vertrauen und die Akzeptanz in ML-Prognosen fördern kann. Die Studie vergleicht erstmalig interdisziplinär vier ML-Modelle unter Anwendung von drei XAI-Methoden im Kontext der Performance-Interpretierbarkeitsdebatte. Die ausgewählten Algorithmen decken die gesamte Bandbreite von transparenten Whitebox-Modellen bis hin zu undurchsichtigen Blackbox-Modellen ab. Die Ergebnisse werden anhand der Erklärungseigenschaften diskutiert. Die Ergebnisse zeigen, dass der Einsatz von XAI

die Interpretierbarkeit von ML-Modellen verbessern kann, aber auch auf gravierende Herausforderungen hinweist. Der Rashomon-Effekt kann hinsichtlich der Multiplizität von guten Modellen mittels XAI reduziert werden. Simultan werden multiple gute Erklärungen erschaffen, die punktuelle Widersprüche in der Ergebnisfindung aufweisen und somit die Entscheidungsfindung für systemkritische Abwägungen maßgeblich behindern. Der Zugewinn an Akzeptanz, Verständlichkeit und Transparenz, den Unternehmen durch den Einsatz von XAI-Methoden erfahren können, wird erneut nichtig gemacht. Dennoch bietet XAI Unternehmen einen klaren Vorteil, indem es transparente ML-Ergebnisse in kritische Entscheidungsprozesse integriert. Durch die Anwendung von XAI und ML-Techniken können Unternehmen ihre Geschäftsentscheidungen beeinflussen und erfolgreich im Markt agieren. Konsolidiert ermöglicht eine transparente Ergebnisfindung faktenbasierte Erkenntnisse, die fundierte Entscheidungen im Wettbewerbsumfeld fördern und den wirtschaftlichen Erfolg skalieren.

7 Ausblick und Forschungsbedarf

- Uniforme Metriken

Wie bereits ausgiebig und mehrfach hervorgehoben, ist der Bedarf an einheitlichen und fundierten Metriken von großer Relevanz. Die Entwicklung von Metriken und Methoden zur Überprüfung und Validierung von XAI-Verfahren ist nicht nur für die Verständlichkeit von Relevanz, sondern wird ebenfalls aus ethischer und rechtlicher Sicht als signifikant gedeutet.

- Menschenzentrierte Ansätze

Es ist ebenso notwendig, menschlich-zentrierte Ansätze für die Interaktion mit XAI-Systemen zu erforschen, indem weitere Befragungen erhoben werden, wie es durch Sisk et al. (2022) bereits getätigt wird. Außerdem liegt der derzeitige Fokus auf Large Language Models wie ChatGPT, die einer breiten Nutzerschaft im privaten und beruflichen Umfeld die Möglichkeit bietet, mit KI erstmalig zu interagieren. Gerade hierbei ist der Einsatz von XAI notwendig, um insbesondere in Bezug auf "halluzinierende" Informationen, die ethisch fragwürdige oder fehlerhafte Informationen darstellen, die Transparenz der Ergebnisfindung sicherzustellen. Es bedarf weitere Intensive Forschung wie es bereits durch Datta & Dickerson (2023) und Liao & Vaughan Liao & Vaughan (2023) angestoßen wird.

Literaturverzeichnis