**Universität Ulm**

Fakultät für Mathematik und

Wirtschaftswissenschaften

Titel unserer Seminararbeit

Seminararbeit

In Wirtschaftsmathematik

vorgelegt von

Denise Falk, Simon Hofer, Hannah Knehr

am 11. Juni 2023

**Gutachter**

Lars Moestue

Dr. Andreas Obermeier

**Inhaltsverzeichnis**

[Tabellenverzeichnis IV](#_Toc137379823)

[Abkürzungsverzeichnis V](#_Toc137379824)

[2 Einleitung 6](#_Toc137379825)

[3 Theoretischer Teil 7](#_Toc137379826)

[3.1 Artificial Intelligence (AI) und Explainable Artificial Intelligence (XAI) 7](#_Toc137379827)

[4 Analyse Part 8](#_Toc137379828)

[4.1 Use Case 8](#_Toc137379829)

[4.2 Datenaufbereitung 8](#_Toc137379830)

[4.3 Auswahl AI System mit Parametertuning etc. 8](#_Toc137379831)

[4.4 Auswahl XAI System mit Modifikationen (Restriktionen etc) 8](#_Toc137379832)

[4.5 Auswertung 8](#_Toc137379833)

[5 Diskussion 9](#_Toc137379834)

[6 Fazit/Limitationen 10](#_Toc137379835)

[Anlage 11](#_Toc137379836)

[Literaturverzeichnis 12](#_Toc137379837)

[7 Ehrenwörtliche Erklärung 13](#_Toc137379838)

Abbildungsverzeichnis

Abbildung 4.1: Feature Importance 4

[Abbildung 4.2: Lime Werte 5](#_heading=h.lnxbz9)

[Abbildung 4.3: Shap Werte 6](#_heading=h.lnxbz9)

[Abbildung 4.4: Explanation mit jeweils einer Variable 8](#_heading=h.lnxbz9)

[Abbildung 4.5: Explanation mit zwei Variablen 8](#_heading=h.lnxbz9)

[Abbildung 4.1: Feature Importance 1](#_heading=h.lnxbz9)

# Tabellenverzeichnis

# 

# Abkürzungsverzeichnis

|  |  |  |
| --- | --- | --- |
| Künstliche Intelligenz. | *KI* |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

# Einleitung

Künstliche Intelligenz (KI) hat mittlerweile zahlreiche Branchen erreicht und revolutioniert. Der Finanzsektor bildet hierbei keine Ausnahme, denn die Fähigkeit von KI-Systemen, große Datenmengen zu analysieren und fundierte Entscheidungen zu treffen, ist für Banken und Finanzinstitute zunehmend wertvoll geworden. Vor allem bei Kreditvergabeprozessen wird gern auf KI zurückgegriffen. Die inhärente Komplexität von KI-Modellen gibt jedoch häufig Anlass zu Bedenken hinsichtlich ihrer Transparenz und Interpretierbarkeit. Diese Faktoren sind im Bankensektor von großer Bedeutung, da Entscheidungen über die Bewilligung von Krediten erhebliche Auswirkungen auf das Leben und das finanzielle Wohlergehen Einzelner haben können. Ein mangelndes Verständnis der Faktoren, die von KI-Modellen berücksichtigt werden, kann dazu führen, dass KundInnen sich ungerecht behandelt fühlen. Dies kann das Vertrauen in den Entscheidungsprozess der Bank beeinträchtigen und sich negativ auf die KundInnenzufriedenheit und -loyalität auswirken.

In den letzten Jahren hat man zunehmend erkannt, dass erklärbare KI-Techniken (Explainable Artificial Intelligence; XAI) notwendig sind, um diese Probleme zu lösen. XAI zielt darauf ab, Einblicke in den Entscheidungsfindungsprozess von KI-Modellen zu geben und ihre Ergebnisse sowohl für Fachleute als auch für Endnutzer verständlicher und interpretierbar zu machen. Durch den Einsatz von XAI kann die Kluft zwischen der "Blackbox"-Natur herkömmlicher KI-Algorithmen und dem Wunsch nach Transparenz und Verantwortlichkeit überbrückt werden.

In dieser Seminararbeit werden wir den Einsatz von KI und XAI bei der Analyse eines Datensatzes untersuchen, um festzustellen, ob BankkundInnen ein Kredit gewährt werden sollte oder nicht. Durch die Anwendung eines XAI-Modells auf unseren Kreditgenehmigungsklassifikator wollen wir klare und verständliche Erklärungen für die getroffenen Entscheidungen liefern. Dadurch sollen die KundInnen nicht nur nachvollziehen können, warum ein Kreditantrag bewilligt oder abgelehnt wurde, sondern auch Bereiche identifizieren, in denen sie ihr Finanzprofil verbessern können. Darüber hinaus können XAI-Techniken dabei helfen, mögliche Verzerrungen im Entscheidungsprozess des KI-Modells zu erkennen und abzumildern, um faire und gerechte Kreditentscheidungen zu gewährleisten.

# Theoretischer Teil

## Artificial Intelligence (AI) und Explainable Artificial Intelligence (XAI)

@Denise

KI funktioniert mithilfe von Algorithmen, die mithilfe von Daten Strukturen erlernen, Aufgaben erledigen oder Entscheidungen treffen können. Es wird ihr nachgesagt, in Zukunft einen großen Einfluss auf wichtige Industrien zu haben. Doch nicht nur in Zukunft wird KI eine große Rolle spielen- bereits heute findet sie beispielsweise großflächig im Bankensektor, im Bildungswesen, in der Industrie oder im Gesundheitswesen Anwendung. Auch in Privathaushalten ist KI nicht mehr wegzudenken- *Beispiel Alexa/Siri, TikTok, … ?*

Dies liegt vor allem an den zahlreichen Vorteilen, die die Nutzung einer KI mit sich bringt: Erhöhte Perfomance bei Problemlösungen, das Erkennen von Mustern oder der Umgang mit großen Datenmengen zählen unter anderem dazu. Dass das Training von KIs mit entsprechenden Datensätzen jedoch extrem wichtig ist, zeigt sich spätestens seit *Amazon's hiring software (Daston, 2018)/ Rückfälligkeits KI in USA bei Straftätern*. In Zukunft werden wir noch mehr auf dieses korrekte Training angewiesen sein, da KI in Gebieten eingesetzt werden wird, in denen Fehler lebensgefährlich sein können und Rechenschaftspflicht, Vertrauen und ethische Überlegungen von entscheidender Bedeutung sind, wie zum Beispiel in der Medizin oder autonomen Systemen.

Daher ist es wichtig, dass Menschen verstehen können, wie eine KI funktioniert und ihre Entscheidungen trifft. Im Moment ist dies aufgrund des sogenannten „Blackbox“-Charakters von KI noch kompliziert.

Um diesem Problem entgegenzuwirken, haben sich ForscherInnen aus verschiedenen Disziplinen, wie der Informatik und ihrem Unterbereich Mensch-Computer-Interaktion, Psychologie und Kognitionswissenschaft damit beschäftigt, Methoden und Ansätze zu entwickeln, die die Herausforderung der Erklärbarkeit von KI-Systemen angehen. Das Ziel von XAI ist es, Erklärungen für das "Wie und Warum" hinter den Entscheidungen einer KI zu liefern. Kritische Anforderungen an XAI-Systeme sind daher Handlungsfähigkeit, algorithmische Transparenz, Kausalität, Kohärenz, Verständlichkeit, Vertrauen, Fairness, Treue, Informativität, Bewusstsein für den Datenschutz, Übertragbarkeit, Vertrauenswürdigkeit und Verständlichkeit. Um Erklärungen zu generieren, die für Menschen greifbar sind, sind diese Anforderungen wesentlich. Die Forschung zu XAI und ihre Entwicklung sind jedoch noch nicht beendet. Letztlich geht es darum, ein Gleichgewicht zwischen der Vorhersagekraft komplexer KI-Modelle und der Fähigkeit zu finden, sinnvolle Erklärungen zu liefern, die es den Nutzern ermöglichen, KI-Systemen zu vertrauen, sie zu verstehen und effektiv mit ihnen zu interagieren.

## AI und XAI in der Finanzbranche

## Counterfactual explanations

Wie bereits zuvor verdeutlicht, zielt XAI darauf ab, Ergebnisse und Entscheidungen von KI für EndnutzerInnen erklärbar zu machen. Dabei muss man verstehen, wie Menschen Entscheidungen treffen, um später zu garantieren, dass sie die Erklärungen der XAI annehmen. Miller (2019) beschreibt in seiner Forschung zu counterfactual explanations drei Beobachtungen:

1. Bei Erklärungen geht es nicht nur darum, Zusammenhänge und Ursachen aufzuzeigen: Erklärungen müssen den spezifischen Kontext und die Umstände, unter denen sie gegeben werden, berücksichtigen.
2. Die Relevanz von Ursachen hängt vom Kontext ab: Wenn etwas geschieht, kann es viele verschiedene Gründe oder Ursachen dafür geben. Nicht alle diese Ursachen sind jedoch wichtig oder relevant für die vorliegende Situation. Der oder die Erklärende wählt die für den spezifischen Kontext am relevantesten Ursachen aus.
3. Interaktion und Diskussion sind wichtig: Der oder die Erklärende und die Person, die die Erklärung erhält, können unterschiedliche Perspektiven und Meinungen haben. Sie können sich auf Diskussionen einlassen, um die Erklärung besser zu verstehen und zu verfeinern. Diese Interaktion ermöglicht einen sinnvolleren und effektiveren Informationsaustausch.

Ein erklärungsfähiges XAI-System muss diese drei Punkte unbedingt berücksichtigen. Auf diese Weise kann sichergestellt werden, dass die bereitgestellten Erklärungen nicht nur korrekt, sondern auch relevant für die jeweilige Situation und offen für Diskussionen und Klärungen sind.

Counterfactual explanations zielen darauf ab, die Frage zu beantworten: "Was hätte passieren können, wenn etwas anders gewesen wäre?" Diese Erklärungen bieten Einblicke in die kausalen Beziehungen zwischen Inputs und Outputs eines maschinellen Lernmodells.

Im Kontext von XAI beinhaltet eine kontrafaktische Erklärung die Identifizierung der minimalen Änderungen an den Eingabevariablen oder Merkmalen eines Modells, die zu einer anderen Vorhersage oder einem anderen Ergebnis führen würden. Durch die Untersuchung dieser kontrafaktischen Szenarien können die Benutzer ein besseres Verständnis für die Faktoren gewinnen, die den Entscheidungsprozess des Modells beeinflussen. Kontrafaktische Erklärungen sind besonders nützlich, wenn es um komplexe Modelle geht, wie z. B. tiefe neuronale Netze, die aufgrund ihrer undurchsichtigen Entscheidungsprozesse oft als Blackboxen betrachtet werden. Durch die Bereitstellung kontrafaktischer Erklärungen bieten XAI-Systeme eine besser interpretierbare und transparentere Möglichkeit zu verstehen, wie das Modell zu einer bestimmten Vorhersage oder Entscheidung gelangt ist.

*12.06.23: Ich würde hier in den kommenden Tagen noch etwas mehr dazu scheiben.*

# Analyse Part

## Use Case

@Denise

## Datenaufbereitung

@Hannah

## Auswahl AI System mit Parametertuning etc.

Auswahl AI

@Hananh

XAI System

@Denise

* Accuracy
* Parameterisierung
* @Simon

## Auswahl XAI System mit Modifikationen (Restriktionen etc)

@Simon / @Hannah

## Auswertung

@ Hannah

- 1/ 2 richtig gute Counterfactual

- Matthias / Obi bzw. Lars 🡪 synthetischer Daten

# Diskussion

@ Denise / Hannah / Simon

* Faktenbasiert noch keine Interpretation
* Laufzeit
* Accuracy
* Nicht immer optimal (Wegargumentieren)
* Kategorische Variablen, Binärsystem
* Vorteile nochmal betonen

# Fazit/Limitationen

* Interpretation
* Grenzen von Care @Denise

# Anlage

# Literaturverzeichnis

|  |
| --- |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |

# Ehrenwörtliche Erklärung

Wir erklären hiermit ehrenwörtlich, dass wir die vorliegende Arbeit mit dem Titel

Titel der Arbeit

selbständig angefertigt haben; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht. Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

Wir sind uns bewusst, dass eine unwahre Erklärung rechtliche Folgen haben wird.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Ort, Datum Vorname Nachname

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Vorname Nachname

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Vorname Nachname