**Universität Ulm**

Fakultät für Mathematik und

Wirtschaftswissenschaften

Titel unserer Seminararbeit

Seminararbeit

In Wirtschaftsmathematik

vorgelegt von

Denise Falk, Simon Hofer, Hannah Knehr

am 18. 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, zukünftig einen großen Einfluss auf wichtige Industrien zu haben. Doch nicht nur in der 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 korrektes Training angewiesen sein, da KI in Gebieten eingesetzt 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

@Denise

## Counterfactual explanations

Wie bereits zuvor verdeutlicht, zielt XAI darauf ab, Ergebnisse und Entscheidungen von KI für EndnutzerInnen erklärbar zu machen. Um zu garantieren, dass die Erklärungen der XAI angenommen werden, muss man jedoch verstehen, wie Menschen Entscheidungen treffen. 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 gelieferten 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 BenutzerInnen 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

Wir machen es aus Kundensicht

Ziel : Kunde sollen mit Hilfe der Counterfactual Erklärung verstehen warum sie den Kredit nicht bekommen haben und das AI Modell verstehen. Im besten Fall bekommen Sie eine erklärung wie sie ihr verhalten ändern können um Ihre Chancen zu erhöhen : (Verstehen + im Besten Fall konstruktive Erklärung )

Um was geht es

Datensatz: 90657 Einträge/Kunden

Anzahl features: 27 (davon x raus)

Welche features: ‘ID’, “Customer ID’, ‘Month’, 'Age', ‘SSN’, 'Annual\_Income', 'Monthly\_Inhand\_Salary', 'Num\_Bank\_Accounts', 'Num\_Credit\_Card', 'Interest\_Rate', 'Num\_of\_Loan', ‘Type of Loan’, 'Delay\_from\_due\_date', 'Num\_of\_Delayed\_Payment', 'Changed\_Credit\_Limit', 'Num\_Credit\_Inquiries', 'Outstanding\_Debt', 'Credit\_Utilization\_Ratio', 'Total\_EMI\_per\_month', 'Amount\_invested\_monthly', 'Monthly\_Balance',

discrete\_features = ['Occupation', 'Credit\_Mix', 'Credit\_History\_Age', 'Payment\_of\_Min\_Amount', 'Payment\_Behaviour'

‘Credit Score’

Wie ist Situation

Was wollen wir erreichen ? Was ist die Aufgabe ?

## Datenaufbereitung

@ Denise

* Statistiken

@Hannah

* unlogische Werte droppen (Customer rausschemsien)
* 3 Spalten droppen
* Nan droppen
* Credit Score – ganzer Datensatz ohne Zielvariable
* Continous / discret
* Standard skaliert (continous)
* Discrete features encoded (discret) - Ordinal / one hot encoding (0/1 System) / Label encoder (BAUT ALLES AUF CARE – AUTOREN AUF)
* Übereinstimmmung Get\_to\_know\_data
* Aufteilen Loan\_Type

## Auswahl AI System mit Parametertuning etc.

Auswahl AI

@Denise

Random Forest

Nach der Entscheidung, den Random Forest als zugrundeliegendes AI System zu nutzen, haben wir ein Parametertuning angewendet. Dies ermöglicht nicht nur eine Verbesserung der Performance, sondern auch einen zusätzlichen Gewinn an Robustheit. Genutzt haben wir den Random Forest Klassifikator des sklearn-Paketes in Python. Für das Parametertuning haben wir für folgende Parameter verschiedene Werte ausgetestet:

-n\_estimators: Gibt die Anzahl der Bäume im Random Forest an. Hierfür haben wir die Werte 100, 200 und 300 getestet.

-max\_depth: Gibt die maximale Tiefe des Baumes an. Hierfür haben wir die Werte "Nicht bestimmt", 5 und 10 getestet.

-min\_samples\_split: Bestimmt die mindestanzahl an Trainingsdaten, die erforderlich sind, um einen Knoten zu teilen. Hierfür haben wir die Werte 2, 5 und 10 getestet.

-min\_samples\_leaf: Bestimmt die Mindestanzahl an Trainingsdaten, die erforderlich ist, um einen Endknoten zu bilden.

Diese Parameterwerte wurden in allen Kombinationen getestet, wir haben dabei von 81 (3\*3\*3\*3) Kombinationen diejenige verwendet, welche die beste Performance aufweisen konnte. Für die Evaluation wurde ein 5-Folds Cross-Validation-Verfahren angewendet, um weiter sicherzustellen, dass wir robuste und genauere Schätzungen erhalten.

Resultierend aus dem Parametertuning/ 5-Folds Cross-Validation-Verfahren haben sich die folgenden Parametereinstellungen nach unserem Parametertuning als bestmöglich ergeben:

{'n\_estimators': 300'max\_depth': "Keine Angabe", 'min\_samples\_leaf': 1, 'min\_samples\_split': 2}

Durch das Parametertuning konnten wir eine Accuracy von 79.02% erreichen. Als zweites Gütemaß des Random Forest haben wir das F1-Measure berechnet. Hier haben wir eine Genauigkeit von 78.98% vorzuweisen, was auf eine ausgewogene Leistung hinsichtlich der Klassifikation unserer Daten deutet.

Für die post-hoc XAI Methode haben wir uns für Methoden entschieden, welche kontrafaktischen Erklärungen erzeugen. Da wir uns zum Ziel gesetzt haben, dass sich unsere Erklärungen an den Endkunden richten sollen, empfinden wir kontrafaktische Erklärungen als ideal. Zum einen, weil Nutzer auf Basis der sich veränderten Attribute verstehen können, welche Werte wichtig für die Entscheidung des AI Modells sind (schafft Transparenz und Verständnis), zum anderen aber auch, weil diese kontrafaktischen Ergebnisse im besten Fall auch umsetzbar für den Kunden sind. Eine umsetzbare kontrafaktische Erklärung kann im Idealfall ein Wegweiser für den Kunden sein, sodass dieser nicht nur Verständlichkeit für das AI Modell gewinnt, sondern auch direkt handeln kann, um das gewünschte Resultat zu erzielen.

Als Methode der kontrafaktischen Erklärungen haben wir uns für CARE nach Rasouli (Quelle) entschieden. CARE ist ein Ansatz, welcher auf einer Paretooptimierung basiert, d.h es fließen verschiedene Kriterien bei der Suche nach einem geeigneten foil (der generierten, synthetischen kontrafaktischen Erklärungsinstanz) in die Optimeirung ein. Dabei können bis zu 7 Zielfunktionen in das Optimierungproblem Einfluss nehmen:

Fest verankert sind die Zielfunktionen "outcome","distance" und "sparsity", welche Teil des Moduls "validity" sind. Dieses Modul ist immer aktiv. Die Zielfunktion "outcome" achtet darauf, dass der Foil in der richtigen Zielklasse landet. "Distance" ist eine Zielfunktionen, die den Abstand zwischen Fact und Foil berechnet (Minimierungsproblem). "Sparsity" zählt die Anzahl an Attributen, die sich von Fact zu Foil unterscheiden (Minimierung).

Neben dem "validity" Modul gibt es weitere Module, die aktiviert werden können (und somit in das Optimierungsproblem einfließen). "Soundness" ist ein Modul, bei dem durch die zwei Zielfunktionen "proximity" und „connectedness“ sichergestellt werden soll, dass der foil nahe an den Instanzen des Datensatzes liegt, d.h, dass es sich tendentiell um eine typische Instanz handelt und nicht einem Ausreißer entspricht. Desweiteren sollen die Änderungen von fact zu foil einem kohärenten, sinnvollen Pfad an Werten entlang gehen.

Im dritten Modul „coherency“, welches aus der gleichnamigen Zielfunktion besteht. Hierbei soll durch Beachtung der Korrelationen zwischen den Variablen sichergestellt werden, dass die Änderungen zwischen Fact und Foil kohärent sind. Das vierte und letzte Modul, „actionability“ soll gewährleisten, dass die Änderungen aus Nutzersicht machbar sind. Hierbei kann man Restriktionen für Wertebereiche einzelner Variablen setzen, um sicherzustellen, dass die Änderungen auch machbar sind.

Neben der vielschichtigen Beachtung verschiedener Konstrukte, welche wichtig sind bei der Generierung von kontrafaktischen Erklärungen, hat uns des weiteren überzeugt, dass CARE einen gradienten-freien Optimierer verwendet. Diese haben den Vorteil, dass sie auf natürlichem Wege mit kategorischen Variablen arbeiten können. In unserem Use-Case haben wir einige davon (Occupation, Loan Type, etc.), weshalb wir dieses Kriterium als wichtig ansehen. Zusätzlich kann man CARE vorgeben, wie viele foils aus einem fact entstehen sollen. Das ist insofern vorteilhaft, dass es sein kann, das wir zwar einen validen Foil haben (d.h in der richtigen Zielklasse), dieser aber nicht umsetzbar ist. Bei mehrern foils kann es hingegen schon eher sein, dass wir einen machbaren finden. Dadurch kann man weitere Einblicke in das AI System erhalten, da man anhand der verschiedenen Attributsänderungen zwischen den foils die AI Entscheidungsgrundlage ablesen kann. Final gilt noch zu erwähnen, dass CARE model-agnostisch ist, d.h wir können es unabhängig vom zugrundeliegenden AI Modell verwenden.

Nachdem wir nun die Datenaufbereitung absolviert und den Random Forest trainiert haben, haben wir noch CARE an unseren Use Case angepasst. Hierfür haben wir die Zieklassenzuweisung von CARE technisch angepasst. Im Originalansatz wird diejenige Klasse als Zielklasse verwendet, welche bei der Prognose des AI Models die 2.höchste Klassenwahrscheinlichkeitszurordnung annimmt. Für unsere Zwecke ist es aber sinnvoller, diese Zurodnung klar zu formulieren. So gelten folgende Zurodnungen für die Zieklasse:

Fact = „Poor“ 🡪 Foil = „Standard“, Fact = “Standard”🡪 Foil = “Good”, Fact = “Good” 🡪Foil = “Standard”. Zusätzlich wurde ein “Probability Thresh” von 0.45 festgelegt. Der Foil muss mindestens diese Wahrscheinlichkeit vom zugrundeliegenden AI Modell in die Zielklasse erhalten. Um die Ergebnisse zu speichern, haben wir einen Loop implementiert, in welchem die Foils in einer pickle-Datei gespeichert werden.

Da es für unsere Zwecke sinnvoll ist, alle Module zu aktivieren, haben wir auch noch Restriktionen implementiert.

XAI System

@Simon

* Accuracy
* Parameterisierung
* @Simon

## Auswahl XAI System mit Modifikationen (Restriktionen etc)

@Denise / @Hannah

* Vorgehen kurz beschreibene
* Zu viele Restrictions – keine genauen Ergebnisse
* Nötige beschränkt
* Einzelne Attribute eingehen

## Auswertung

@ Simon

- 1/ 2 richtig gute / richtig schlechte Counterfactual

- limitieren

- nicht immer gute Counterfactuals

- Matthias / Obi bzw. Lars 🡪 synthetischer Daten

# Diskussion

@ Denise / Hannah / Simon

* Faktenbasiert noch keine Interpretation
* Findet manchmal sehr gut / nicht immer
* Laufzeit
* Accuracy
* Nicht immer optimal (Wegargumentieren)
* Kategorische Variablen, Binärsystem
* Vorteile nochmal betonen
* Wie passt es zum Use Case

# Fazit/Limitationen

* Interpretation
* Use Case eingehen
* Was muss Bankberater beachten / Wissen
* Wie geht der Bankberater damit um

Limitationen

* Grenzen von Care

Weitere Forschungsideeen

# 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