**Universität Ulm**

Fakultät für Mathematik und

Wirtschaftswissenschaften

Titel unserer Seminararbeit

Seminararbeit

In Wirtschaftsmathematik

vorgelegt von

Denise Falk, Simon Hofer, Hannah Knehr

am 19. Juni 2023

**Gutachter**

Lars Moestue

Dr. Andreas Obermeier

**Inhaltsverzeichnis**

[Tabellenverzeichnis IV](#_Toc138058127)

[Abkürzungsverzeichnis V](#_Toc138058128)

[2 Einleitung 6](#_Toc138058129)

[3 Theoretischer Teil 7](#_Toc138058130)

[3.1 Artificial Intelligence (AI) und Explainable Artificial Intelligence (XAI) 7](#_Toc138058131)

[3.2 AI und XAI in der Finanzbranche 8](#_Toc138058132)

[3.3 Kontrafaktische Erklärungen 8](#_Toc138058133)

[4 Analyse Part 10](#_Toc138058134)

[4.1 Use Case 10](#_Toc138058135)

[4.2 Datenaufbereitung 11](#_Toc138058136)

[4.3 Auswahl AI System mit Parametertuning etc. 12](#_Toc138058137)

[4.4 Auswahl XAI System mit Modifikationen (Restriktionen etc) 15](#_Toc138058138)

[4.5 Auswertung 17](#_Toc138058139)

[5 Diskussion 18](#_Toc138058140)

[6 Fazit/Limitationen 19](#_Toc138058141)

[Anlage 20](#_Toc138058142)

[Literaturverzeichnis 21](#_Toc138058143)

[7 Ehrenwörtliche Erklärung 22](#_Toc138058144)

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

[Tabelle 1: 1](#_heading=h.35nkun2)

# 

# 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 Kunden und 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 (Englisch: 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 und Endnutzerinnen verständlicher und interpretierbar zu machen. Durch den Einsatz von XAI kann die Kluft zwischen der "Blackbox[[1]](#footnote-2)"-Natur herkömmlicher KI-Algorithmen und dem Wunsch nach Transparenz und Verantwortlichkeit überbrückt werden.

In dieser Seminararbeit wird der Einsatz von KI und XAI bei der Analyse eines Datensatzes untersuch, um festzustellen, ob BankkundInnen ein Kredit gewährt werden sollte oder nicht. Durch die Anwendung eines XAI-Modells auf den Kreditgenehmigungsklassifikator sollen klare und verständliche Erklärungen für die getroffenen Entscheidungen geliefert werden. 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

## Künstliche Intelligenz (AI) und Erklärbare künstliche Intelligenz (XAI)

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 korrektes Training noch mehr an Bedeutung gewinnen, 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 Blackbox-Charakters von KI noch kompliziert.

Um diesem Problem entgegenzuwirken, haben sich Forscher und 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.

## KI und XAI in der Finanzbranche

Der Einsatz von KI und XAI in der Finanzbranche hat aufgrund seines Potenzials, verschiedene Prozesse und Dienstleistungen zu verbessern, erheblich an Beliebtheit gewonnen. Die Fähigkeit von KI-Algorithmen, große Datenmengen zu analysieren (darunter zum Beispiel Transaktionsaufzeichnungen, KundInnenverhaltensmuster und externe Faktoren), ermöglichen es zum Beispiel, verdächtige Aktivitäten und potenziellen Betrug zu erkennen. KI-Modelle können die Kreditwürdigkeit beurteilen, Ausfallwahrscheinlichkeiten vorhersagen und Kreditrisiken auf der Grundlage historischer Daten bewerten, was eine genauere Risikobewertung ermöglicht. KI-gestützte Chatbots und virtuelle Assistenten bieten personalisierten Kundensupport, können bearbeiten Routineanfragen und Produktempfehlungen geben. Bei der Automatisierung von Compliance-Prozessen kann KI dabei helfen, große Mengen an regulatorischen Daten zu analysieren, Muster zu identifizieren und potenzielle Verstöße aufzudecken. Weiterhin können Markttrends, historische Daten und Kundenpräferenzen analysiert werden, um personalisierte Anlageberatung und Portfolioverwaltung anzubieten.

XAI-Techniken liefern dabei immer Erklärungen für die von der KI getroffenen Entscheidungen und ermöglichen es Banken, Finanzinstituten, deren Angestellten und ErmittlerInnen, die Identifizierungsfaktoren zu verstehen.

Insgesamt ermöglicht die Integration von KI und XAI im Bankensektor es, die betriebliche Effizienz zu verbessern, das KundInnenerlebnis zu steigern, die Einhaltung von Vorschriften zu gewährleisten und die Transparenz und das Vertrauen zwischen Banken und ihren KundInnen zu fördern. Wichtig ist jedoch, ein Gleichgewicht zwischen der Nutzung der KI-Funktionen und der Gewährleistung einer ethischen und verantwortungsvollen Nutzung von KundInnendaten und Entscheidungsprozessen zu finden.

## Kontrafaktische Erklärungen

Um das Blackbox-Problem, das bei der Anwendung von KI häufig auftritt, zu lösen, gibt es mittlerweile mehrere Ansätze. Zur Bearbeitung der Fragestellung in diesem Seminar wurde der Ansatz der kontrafaktischen Erklärungen angewendet. Kontrafaktische Erklärung 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.

Um zu garantieren, dass die Erklärungen einer XAI angenommen werden, muss man verstehen, wie Menschen Entscheidungen treffen. Miller (2019) beschreibt in seiner Forschung zu kontrafaktischen Erklärungen 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.

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 NutzerInnen 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.

Ein XAI-Modell, das sich kontrafaktischer Erklärungen bedient, enthält Fakten und Folien. Dabei stellt ein Fakt einen Datenpunkt aus dem gegebenen, tatsächlichen Datensatz dar, den wir erklären wollen. Im Bankensektor könnte ein Fakt zum Beispiel der Kreditantrag eines bestimmten Kunden oder einer bestimmten Kundin sein.

Eine Folie hingegen stellt ein hypothetisches oder alternatives Szenario dar, das sich von den Fakten unterscheidet. Es handelt sich dabei um eine modifizierte Version des Sachverhalts, die dabei hilft zu verstehen, wie die Änderung bestimmter Attribute zu einem anderen Ergebnis oder einer anderen Entscheidung hätte führen können. Folien werden erstellt, um Was-wäre-wenn-Szenarien zu erforschen und Einblicke in die Art und Weise zu gewähren, wie sich Änderungen der Variablen auf das gewünschte Ergebnis auswirken könnten. Im Beispiel des Kreditantrages könnte eine Folie zum Beispiel durch die Änderung von bestimmten Attributen, wie z. B. der Erhöhung des Einkommens generiert werden.

Das Ziel der Erstellung von Folien ist es, Erklärungen dafür zu liefern, warum ein bestimmtes Ergebnis eingetreten ist oder wie es unter anderen Umständen hätte anders ausfallen können. Durch den Vergleich von Fakten und Folien können Einblicke in die Faktoren gewonnen werden, die zu dem Ergebnis beigetragen haben.

Um zufriedenstellende Erklärungen zu garantieren, heben Forscher und Forscherinnen verschiedene Merkmale hervor. Obwohl die Meinungen über die Wichtigkeit der entsprechenden Merkamale auseinander gehen, werden folgende Kriterien wiederkehrend genannt: Kohärenz, Einfachheit, Allgemeinheit und Relevanz. Kohärenz bezieht sich auf die Übereinstimmung von Erklärungen mit früheren Überzeugungen, und Realismus ist für das Erreichen von Kohärenz entscheidend. Einfachheit und Allgemeinheit bedeuten, dass einfachere Erklärungen und Erklärungen, die mehr Ereignisse erklären, bevorzugt werden. Gültigkeit, Handlungsfähigkeit, Sparsamkeit, Nähe der Datenvielfalt und Kausalität werden als Kriterien für Folien in kontrafaktischen Erklärungen diskutiert.

Leider werden in den bestehenden Ansätzen zur Erzeugung kontrafaktischer Erklärungen oft nicht alle diese Kriterien erfüllt. Weiterhin erleben diese Ansätze Herausforderungen beim Umgang mit gemischten Daten (kategorisch und numerisch). Daher ist es nötig, in Zukunft besonders am Umgang mit kategorischen Variablen zu arbeiten.



# Analyse Part

## Use Case

Der Anwendungsfall dreht sich um die Implementierung eines KI-basierten Kreditgenehmigungssystems in einem Bankinstitut. Durch den Einsatz eines KI-Klassifikators in Verbindung mit einem XAI-Modell soll das System fundierte Entscheidungen über die Kreditvergabe an KundInnen treffen können. Dabei soll das System eine Datenbank analysieren, die die Daten von KundInnen enthält. Jeder Eintrag in der Datenbank enthält 27 Attribute, und liefert so eine umfassende Bewertung der Kreditwürdigkeit der KundInnen. Nach der Klassifizierung soll den KundInnen mit Hilfe des ausgewählten XAI-Modells Empfehlungen gegeben werden, wie sie ihre Bonität verbessern können. Dieser Ansatz bietet mehrere Vorteile, darunter eine höhere Genauigkeit bei der Entscheidungsfindung über die Kreditvergabe, Transparenz und die Möglichkeit, sinnvolle Erklärungen für die Genehmigung oder Ablehnung eines Kredits zu liefern.

Die gewählte Methode hilft vor allem den KundInnen des Finanzinstitutes, denn sie bekommen mittels kontrafaktischer Erklärungen genaue Hinweise, warum ein Kreditantrag eventuell abgelehnt wird. Das System liefert daraufhin explizite und realistische Vorschläge, die Bonität zu verbessern. Diese Methode ermöglicht es den KundInnen zu verstehen, wie die XAI ihre Entscheidungen trifft und nimmt ihr den Blackbox-Charakter. SieInneinleiten, dieen

Der genutzte Datensatz enthält 90.657 Einträge von entsprechend vielen KundInnen. Alle Einträge setzen sich aus 27 Attributen zusammen, die da sind: ID, KundInnen-ID, Monat, Alter, Sozialversicherungsnummer (Englisch: Social Security Number, SSN), Beruf, Jahreseinkommen, Monatsgehalt, Anzahl Bankkonten, Anzahl Kreditkarten, Zinssatz, Anzahl Darlehen, Art des Darlehens, Verspätung vom Fälligkeitsdatum, Anzahl verspäteter Zahlungen, Geändertes Kreditlimit, Anzahl Kreditanfragen, Kreditmix, Ausstehende Schulden, Kreditauslastungsquote, Kredithistorie Alter, Zahlung von Mindestbetrag, Gesamte Rückzahlung pro Monat (Englisch: equated monthly installment, EMI), Monatlich angelegter Betrag, Zahlungsverhalten, Monatlicher Saldo, Bonität.

## Datenaufbereitung

Bevor die Daten den Klassifikationsmodellen übergeben werden können, müssen sie zunächst aufbereitet und bereinigt werden. Es wurden folgende vorbereitende Schritte auf den Datensatz angewandt, um eine sinnvolle Klassifizierung der Daten zu ermöglichen:

Im ersten Schritt erfolgt eine intensive Betrachtung der Daten, um mögliche Unstimmigkeiten zu identifizieren. Dabei werden Einträge (KundInnen) mit unlogischen, unvollständigen oder fehlerhaften Attributen aus dem Datensatz entfernt. Ebenso werden Einträge entfernt, bei denen Werte fehlen oder unvollständige Informationen vorliegen. Im verbleibenden Datensatz ist somit jede Zeile vollständig, über jede\*n verbleibende\*n KundIn sind damit alle möglichen Informationen verfügbar. Dieser Plausibilitätscheck stellt sicher, dass nur valide und vollständige Daten für die weiteren Schritte der Analyse verwendet werden.

Im zweiten Schritt werden die Spalten ID, Customer\_ID, SSN und Month aus dem Datensatz entfernt, da sie keinen Einfluss auf die Kreditvergabe haben. Diese Angaben sind nur von administrativer Bedeutung.

Im dritten Schritt werden bestimmte Attribute weiter angepasst, um sinnvolle Ergebnisse zu erzielen. Das Attribut "Type\_of\_Loan" enthält eine Auflistung verschiedener Kreditarten. Um den Einfluss dieses Attributs auf die Kreditvergabe besser messen zu können, wird für jede Kreditart ein neues Attribut implementiert, das angibt, ob die entsprechende Kreditart vorliegt oder nicht. Anschließend wird das Attribut "Type\_of\_Loan" gelöscht. Das Attribut "Credit\_History\_Age" ist in Form von Jahren und Monaten dargestellt. TODO.

Im vierten Schritt wird die Zielvariable in Form des Attributs "Credit\_Score" aus dem Datensatz extrahiert.

Im fünften Schritt werden die Attribute in stetige und diskrete Attribute unterteilt. Die stetigen Attribute sind: TODO. Die diskreten Attribute sind: TODO.

Im sechsten Schritt erfolgt die Skalierung der stetigen Attribute. Dabei kommt eine Standartskalierung zum Einsatz. Die Standardisierung schließt Effekte aus, die die unterschiedlichen Skalen auf das Ergebnis des Random Forests haben könnten. Es wird verhindert, dass unterschiedlich große Skalenbereiche unverhältnismäßig große Auswirkungen auf das Ergebnis haben.

Die diskreten Attribute werden hingegen kodiert. Viele Algorithmen für maschinelles Lernen sind nicht in der Lage, kategorische Variablen zu verarbeiten. Daher ist es wichtig, die Daten in einer geeigneten Form zu kodieren, damit Sie diese Variablen vorverarbeiten können. Die kategoriale Kodierung ist eine Technik, bei der Variablensätze mit kategorischen Daten, die Beschriftungsvariablen anstelle von numerischen Werten enthalten, entsprechend kodiert werden. Je nach Art des Attributes werden unterschiedliche Kodierungsverfahren angewendet, abhängig von ihrer Bedeutung. Dabei kommen der Ordinalkodierer, der One-Hot-Kodierer und der Label-Kodierer zum Einsatz. Der Ordinalkodierer wird für Attribute verwendet, die zwar nominell sind, aber über ein Ordnungselement und eine sinnvolle Reihenfolge verfügen. Der One-Hot-Kodierer wird verwendet, wenn die einzelnen Ausprägungen eines Attributes keine Beziehungen zueinander haben. Der Label-Kodierer wird für die Zielvariable verwendet. Durch diese Kodierungsverfahren wird sichergestellt, dass die Attribute angemessen in numerischer Form dargestellt werden, um sie in anschließenden Analysen oder Modellierungen verwenden zu können.

7 Schritt Indexing  ?

Insgesamt führen die beschriebenen Schritte der Datenvorverarbeitung dazu, dass die im folgenden durchgeführten Analysen durchführbar und interpretierbar sind. Ein weiterer Effekt wird bei der Accuracy des Random Forest deutlich. Diese konnte durch die beschriebenen Maßnahmen auf 0.XX erhöht werden.

## 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 Mindestzahl 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. Des Weiteren 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, dass 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-Modells die 2.höchste Klassenwahrscheinlichkeitszurordnung annimmt. Für unsere Zwecke ist es aber sinnvoller, diese Zuordnung klar zu formulieren. So gelten folgende Zuordnungen 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.

## Auswahl XAI System mit Modifikationen (Restriktionen etc)

Um die kontrafaktischen Erklärungen noch zu verbessern, wurden verschiedene Restriktionen festgelegt. In der Realität lassen sich manche Attribute leichter anpassen als andere. Manche Attribute können von KundInnen überhaupt nicht geändert werden (bspw. das Alter). Um ein möglichst konstruktives Ergebnis zu liefern, sollen die kontrafaktischen Erklärungen also beschränkt werden.

Zunächst war es wichtig die einzelnen Attribute im Detail zu verstehen, um diese dann einschränken zu können. Werden allerdings zu viele Attribute eingeschränkt, liefert der CARE-Algorithmus keine guten der kontrafaktischen Erklärungen mehr. Deshalb empfiehlt es sich einen guten Mittelweg zwischen ausreichender Freiheit und den wichtigsten Beschränkungen, die in der Realität zu Geltung kommen zu gehen.

Auf Basis dessen und durch Ausprobieren von verschiedenen Möglichkeiten wurden folgende Restriktionen festgelegt:

constraints = {'Age': ('fix', 10),   
 'Occupation': ('fix', 2),   
 'Annual\_Income': ([0, x\_org[1]\*1.5],2),   
 'Monthly\_Inhand\_Salary': ([0, x\_org[2]\*1.5],2),   
 'Interest\_Rate': ('fix', 10),

'Outstanding\_Debt': ([0, x\_org[11]\*1.5],2),

'Credit\_History\_Age': ('fix', 10),

'Amount\_invested\_monthly': ([0, x\_org[14]\*5],2)

}

Das Alter und der Beruf sind Attribute, die von KundInnen nur schwer geändert werden können. Die Zahlen stehen dabei für die Wichtigkeit dieser Fixierung. Da das Alter überhaupt nicht geändert werden kann, der Beruf jedoch mit viel Aufwand (bspw. einer Umschulung), wird die Restriktion des Alters mit einem hohen Strafterm ausgestattet, während das Attribut Beruf nur einen geringen Strafterm verpasst bekommt. Die Interest Rate und die Credit History sind Attribute, die durch vergangenes Verhalten festgelegt wurden und somit ebenso nicht verändert werden können. Deshalb erhalten auch diese zwei Attribute den Strafterm 10. Das jährliche und monatliche Einkommen, die ausstehenden Schulden sowie der Betrag an investiertem Vermögen können zwar geändert werden, in der Realität ist dies jedoch meistens nur zu einem gewissem Prozentsatz möglich. Deshalb wird festgelegt, dass diese Attribute sich jeweils nur um maximal 50% erhöhen dürfen. Außerdem wird durch den Strafterm zwei ebenfalls berücksichtigt, dass es einen gewissen Aufwand benötigt, um diese Attribute zu verändern.

Es hat sich gezeigt, dass diese Restriktionen den CARE-Algorithmus genug einschränken, dass sinnvolle kontrafaktischen Erklärungen entstehen. Bei weiteren, in der Realität sicher auch sinnvollen Einschränkungen, ist der Algorithmus zu sehr limitiert und liefert keine guten Erklärungen mehr. Deshalb ist es aus Kundenperspektive am sinnvollsten sich hier auf die wichtigsten Attribute zu beschränken.

## 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

1. Dieser Begriff beschreibt eine Situation, die sich oftmals bei der Nutzung von KI ergibt: Man füttert die KI mit Daten (Input), weiß nicht, wie diese die Daten genau verarbeitet (Blackbox) und ein fertiges Ergebnis wird ausgegeben (Output). [↑](#footnote-ref-2)