**Universität Ulm**

Fakultät für Mathematik und

Wirtschaftswissenschaften

Exploring the Use of a Random Forest and a Counterfactual Explanations Approach in Credit Scoring

Seminararbeit

Big (Social) Data Analytics

vorgelegt von

Denise Falk, Simon Hofer, Hannah Knehr

am 3. Juli 2023

**Gutachter**

Lars Moestue

Dr. Andreas Obermeier

**Inhaltsverzeichnis**

[Abbildungsverzeichnis III](#_Toc139285424)

[Abkürzungsverzeichnis IV](#_Toc139285425)

[1 Die Entschleierung des Kreditgenehmigungsprozesses mit (erklärbarer) künstlicher Intelligenz 1](#_Toc139285426)

[2 Transparenz im Finanzwesen: AI/XAI-Theorie und kontrafaktische Erklärungen 3](#_Toc139285427)

[2.1 AI mithilfe von XAI auf den Grund gehen 3](#_Toc139285428)

[2.2 AI und XAI in der Finanzbranche 4](#_Toc139285430)

[2.3](#_Toc139285431) *[Was wäre, wenn…?](#_Toc139285431)* [- Kontrafaktische Erklärungen 4](#_Toc139285431)

[3 Von Rohdaten zu Erkenntnissen: die Analyse 7](#_Toc139285432)

[3.1 Use Case 7](#_Toc139285433)

[3.2 Datenaufbereitung 7](#_Toc139285434)

[3.3 Auswahl AI-System von Optimierung seiner Ergebnisse 11](#_Toc139285435)

[3.4 Auswahl XAI-System und Modifikationen 13](#_Toc139285436)

[3.5 Auswertung 16](#_Toc139285437)

[4 Diskussion 18](#_Toc139285438)

[5 Fazit 20](#_Toc139285439)

[Literaturverzeichnis 21](#_Toc139285440)

# Abbildungsverzeichnis

[Abbildung 1: Ausschnitt aus der Verteilung der Werte der Attribute 8](#_Toc139233336)

[Abbildung 2: Plausibilitätsprüfung 9](#_Toc139233337)

[Abbildung 3: Entfernen irrelevanter Spalten 9](#_Toc139233338)

[Abbildung 4: Anpassung des Attributs "Type\_of\_Loan" 10](#_Toc139233339)

[Abbildung 5: Extraktion der Zielvariable 10](#_Toc139233340)

[Abbildung 6: Beste Parametereinstellungen nach Parametertuning 13](#_Toc139233341)

[Abbildung 7: Ausgewählte Restriktionen 15](#_Toc139233342)

[Abbildung 8: Ausschnitt von Fakt (hier „x“) und dazugehöriger Folie (hier „cf\_0“) 16](#_Toc139233343)

[Abbildung 9: Evaluationsmetriken um Unterschied von Fakt („x“) zu Folie („cf\_0“) fassbar zu machen 16](#_Toc139233344)

[Abbildung 10: Weiteres Beispiel 1 eines Fakts (hier „x“) und dazugehöriger Folie (hier „cf\_0“) mit ausgewählten Attributen 17](#_Toc139233345)

[Abbildung 11: Weiteres Beispiel 2 eines Fakts (hier „x“) und dazugehöriger Folie (hier „cf\_0“) mit ausgewählten Attributen 17](#_Toc139233346)

# 

# Abkürzungsverzeichnis

|  |  |
| --- | --- |
| Beziehungsweise | *bzw.* |
| Das heißt | *d.h.* |
| erklärbare AI-Techniken | *Explainable Artificial Intelligence*  *XAI* |
| Et cetera | *etc.* |
| Künstliche Intelligenz | *KI* |
| Sozialversicherungsnummer | *SSN* |
| Und so weiter | *usw.* |
| United States of America  (Vereinigte Staaten von America) | *USA* |
| Zum Beispiel | *z.B.* |

# Die Entschleierung des Kreditgenehmigungsprozesses mit (erklärbarer) künstlicher Intelligenz

Wenn man von künstlicher Intelligenz (Artificial Intelligence; AI) spricht, dann „bezieht [man] sich auf Systeme, die intelligentes Verhalten zeigen, indem sie ihre Umgebung analysieren und mit einem gewissen Maß an Autonomie handeln, um bestimmte Ziele zu erreichen“ (European Parliament, 2020). Diese Systeme haben mittlerweile zahlreiche Branchen erreicht und revolutioniert (Forbes Tech Council, 2021). Der Finanzsektor bildet hierbei keine Ausnahme, denn die Fähigkeit von AI-Systemen, große Datenmengen zu analysieren ist für Finanzinstitute zunehmend wertvoll geworden (McKinsey & Company, 2021). Die Integration von AI-Systemen in den Kreditvergabeprozess eröffnet vielfältige Möglichkeiten, um schnellere und effizientere Entscheidungen zu treffen. Dadurch können potenzielle KreditnehmerInnen besser bewertet und das Risiko von Kreditausfällen besser erkannt werden (Balogh, S. und Johnson, C., 2021). Die inhärente Komplexität von AI-Modellen gibt jedoch häufig Anlass zu Bedenken hinsichtlich ihrer Transparenz und Interpretierbarkeit. Diese Faktoren sind im Bankensektor jedoch von großer Bedeutung, da Entscheidungen über die Bewilligung von Krediten erhebliche Auswirkungen auf das Leben und das finanzielle Wohlergehen Einzelner haben können. In den USA gibt der Fair Credit Reporting Act of 1970 vor, dass KundInnen eine Erklärung zu den Modellen verlangen können, die BankberaterInnen dabei helfen, Entscheidungen über Kreditvergabeprozesse zu treffen (Emerj, 2019). Ein mangelndes Verständnis der Faktoren, die von AI-Modellen berücksichtigt werden, kann Unsicherheit bei KundInnen auslösen. Dies kann das Vertrauen in den Entscheidungsprozess der Bank beeinträchtigen und sich negativ auf die KundInnenzufriedenheit und -loyalität auswirken.

Daher greift man für die Lösung von diesem Problem zunehmend auf erklärbare AI-Techniken (E**x**plainable **A**rtificial **I**ntelligence; XAI) zurück. XAI zielt darauf ab, Einblicke in den Entscheidungsfindungsprozess von AI-Modellen zu geben und ihre Ergebnisse sowohl für Fachleute als auch für EndnutzerInnen verständlicher und interpretierbar zu machen. Durch den Einsatz von XAI kann die Kluft zwischen der „Blackbox“-Natur (fehlende Nachvollziehbarkeit für die Entscheidungsbasis von AI-Modellen) vereinzelter AI-Algorithmen (Random Forest, künstliche neuronale Netze, etc.) und dem Wunsch nach Transparenz überbrückt werden. Darüber hinaus können XAI-Techniken dabei helfen, mögliche Verzerrungen im Entscheidungsprozess des AI-Modells zu erkennen und abzumildern, um faire Kreditentscheidungen zu gewährleisten (Demajo, L. M., Vella, V. und Dingli, A. 2020).

Das Ziel dieser Seminararbeit besteht darin, ein geeignetes AI- und XAI-Modell zu entwickeln, um BankkundInnen mithilfe des XAI-Systems die Entscheidungen des AI-Modells bezüglich ihrer Kreditwürdigkeit erklärbar(er) zu gestalten. Dadurch soll für die KundInnen ein Mehrwert entstehen. Durch die Anwendung eines XAI-Modells auf den gewählten Klassifikationsalgorithmus (AI-Modell) sollen klare und verständliche Erklärungen für die getroffenen Entscheidungen kreiert werden. Ziel ist hierbei auch, dass KundInnen nicht nur nachvollziehen können, wieso Sie in ihre jeweilige Kreditwürdigkeitsstufe klassifiziert werden, sondern auch konstruktive Vorschläge zur Verbesserung ihres Finanzprofils abgegeben werden, damit die KundInnen in eine andere (bessere) Kreditwürdigkeitsstufe klassifiziert werden.

# Transparenz im Finanzwesen: AI/XAI-Theorie und kontrafaktische Erklärungen

## AI mithilfe von XAI auf den Grund gehen

Aufgrund ihrer Fähigkeit, mit großen Datenmengen umgehen zu können, hat sich die Nutzung von AI mittlerweile in vielen Sektoren, wie unter anderem dem Finanzsektor, dem Einzelhandel, Versorgungsunternehmen oder Medienlandschaften (Boobier, 2018) etabliert. Die Fähigkeit, mit vielen Daten umgehen zu können, führt zu einer erhöhten Perfomance bei Problemlösungen und dazu, dass Muster und Zusammenhänge innerhalb dieser Daten schneller erkannt werden, als es ein Mensch könnte (QAI, 2022). Der Einsatz von AI-Algorithmen, die durch die Datenverarbeitung diese Strukturen erlernen, Aufgaben erledigen und Entscheidungen treffen können, schreitet immer weiter voran. Dass es jedoch entscheidend ist, mit welchen Daten AIs trainiert werden, zeigt sich an Beispielen wie Unfällen durch selbstfahrende Fahrzeuge oder dem Microsoft Chatbot „Tay“, der durch NutzerInnen-Beiträge die Sprache erlernte und dadurch unter anderem vulgäre und rassistische Antworten gab. Da AI in Gebieten eingesetzt wird, in denen Fehler lebensgefährlich sein können und Rechenschaftspflicht, Vertrauen und ethische Überlegungen essenziell sind, wie zum Beispiel in der Medizin oder im Finanzsektor, ist ein entsprechendes Training der Systeme mit vollständigen, aktuellen und repräsentativen Datensätzen unabdingbar (QAI, 2022). Es ist wichtig, dass Menschen verstehen und auch nachvollziehen können, wie die angewandte AI funktioniert und ihre Entscheidungen trifft. Im Moment ist dies aufgrund des sogenannten „Blackbox“-Charakters noch nicht transparent genug (IBM, o.J.).

Um diesem Problem entgegenzuwirken, beschäftigen sich ForscherInnen aus verschiedenen Disziplinen immer wieder damit, Methoden und Ansätze zu entwickeln, die die Herausforderung der Erklärbarkeit von AI-Systemen behandeln (Samek, W., Wiegand, T. und Müller, K., 2017). Das Ziel von XAI ist es, Erklärungen für das „Wie und Warum“ hinter den Entscheidungen einer AI zu generieren (Chou, Y.‑L., et al., 2021). In der Literatur rund um XAI werden daher unter anderem folgende Charakteristika für XAI-Modelle genannt, die notwendig sind, um gute Ergebnisse zu erzeugen: Handlungsfähigkeit, algorithmische Transparenz, Kausalität, Kohärenz, Verständlichkeit, Vertrauen, Fairness, Treue, Informativität, Bewusstsein für Datenschutz, Übertragbarkeit, Vertrauenswürdigkeit und Verständlichkeit. Die Länge der Liste der Charakteristika zeigt jedoch, dass keine Einigung über die endgültige Anzahl der Merkmale gibt. Die Forschungen zu XAI und ihrer Akzeptanz dauern dementsprechend an. Letztlich geht es darum, ein Gleichgewicht zwischen der Vorhersagekraft komplexer AI-Modelle und der Fähigkeit zu finden, sinnvolle Erklärungen zu liefern, die es den NutzerInnen ermöglichen, AI-Systemen zu vertrauen, sie zu verstehen und effektiv mit diesen zu interagieren.

## AI und XAI in der Finanzbranche

Der Einsatz von AI und XAI in der Finanzbranche hat aufgrund seines Potenzials, verschiedene Prozesse und Dienstleistungen effektiver zu gestalten, erheblich an Beliebtheit gewonnen. Die Fähigkeit von AI-Algorithmen, große Datenmengen zu analysieren (darunter zum Beispiel Transaktionsaufzeichnungen, KundInnenverhaltensmuster und externe Faktoren), ermöglichen es zum Beispiel, verdächtige Aktivitäten und potenziellen Betrug zu erkennen. AI-Modelle können Ausfallwahrscheinlichkeiten vorhersagen und Kreditrisiken auf der Grundlage historischer Daten bewerten, was eine genauere Risikobewertung ermöglicht. AI-gestützte Chatbots und virtuelle Assistenten bieten personalisierten KundInnensupport, können Routineanfragen bearbeiten und Produktempfehlungen geben. Bei der Automatisierung von Compliance-Prozessen kann AI 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 Erklärungen für die von der AI getroffenen Entscheidungen und ermöglichen es Banken, Finanzinstituten, deren Angestellten und ErmittlerInnen, die Identifizierungsfaktoren zu verstehen.

Insgesamt ermöglicht die Integration von AI und XAI im Bankensektor es, die betriebliche Effizienz zu steigern, das KundInnenerlebnis zu verbessern, 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 AI- und XAI-Funktionen und der Gewährleistung einer ethischen und verantwortungsvollen Verarbeitung von KundInnendaten zu finden.

## *Was wäre, wenn…?* - Kontrafaktische Erklärungen

Um das Blackbox-Problem bei der Anwendung von AI-Algorithmen zu lösen, gibt es verschiedene Ansätze. Zur Bearbeitung der Fragestellung in diesem Seminar wurde der Ansatz der kontrafaktischen Erklärungen angewendet. Kontrafaktische Erklärung erforschen, welche Eingabevariablen sich in einem fiktiven Szenario ändern müssten, um zu einem unterschiedlichem Ergebnis zu kommen, als das, welches in der Realität passiert ist (Guidotti, R., Monreale, A., Giannotti, F., Pedreschi, D., Ruggieri, S., & Turini, F. 2019). Kontrafaktische Erklärungen beantworten die Frage: „Was hätte passieren können, wenn etwas anders gewesen wäre?“ Diese Erklärungen bieten Einblicke in die kausalen Beziehungen zwischen Inputs und Outputs eines maschinellen Lernmodells.

Es gibt verschiedene Ansätze zur Generierung kontrafaktischer Erklärungen. Diese Methoden können sich hinsichtlich der Auswahl des Optimierungsalgorithmuses, der Zielfunktionen im Optimierungsproblem bei der Suche nach den kontrafaktischen Ereignissen oder auch der Laufzeit unterscheiden. Ein XAI-Modell, das sich kontrafaktischer Erklärungen bedient, enthält Fakten (entspricht der reellen Ist-Situation) und Folien (auch „foil“ genannt; das synthetische Erklärungsereignis, welches in Realität nicht stattgefunden hat) (Stepin, I., Alonso, J. M., Catala, A. und Pereira-Fariña, M., 2021). Ein Fakt stellt einen Datenpunkt aus dem gegebenen, tatsächlichen Datensatz dar, der erklärt werden soll. 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 dem zugrundeliegenden Fakt unterscheidet. Es handelt sich dabei um eine modifizierte Version des Fakts, die dabei hilft zu verstehen, wie die Änderung bestimmter Attribute des Fakts 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 aus einem Fakt 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 Merkmale gewonnen werden, die zu dem Ergebnis beigetragen haben und Robustheit, Fairness und Vertrauenswürdigkeit von AI-Modellen beurteilt werden.

Um zu garantieren, dass die Erklärungen eines XAI-Models von den NutzerInnen akzeptiert werden, muss außerdem verstanden werden, 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.

Um effektive kontrafaktische Erklärungen zu erstellen, ist es entscheidend, den Kontext und die Einschränkungen des AI-Systems, als auch die spezifischen Anforderungen solcher Erklärungen zu berücksichtigen. Die Auswahl relevanter Merkmale, die geändert werden sollen, und die Festlegung geeigneter Änderungen bei gleichzeitiger Vermeidung von Verzerrungen oder unrealistischen Szenarien sind wichtige Überlegungen in diesem Prozess. Die Bewältigung dieser Herausforderungen wird dazu beitragen, die Qualität und Nützlichkeit kontrafaktischer Erklärungen im Bereich der erklärbaren künstlichen Intelligenz zu verbessern.

# Von Rohdaten zu Erkenntnissen: die Analyse

## Use Case

Der vorliegende Anwendungsfall behandelt die Implementierung eines AI-basierten Kreditwürdigkeitsvergabesystems in einem Bankinstitut. Der Einsatz eines AI-Klassifikators in Verbindung mit einem XAI-Modell soll die Entscheidung über Kreditwürdigkeiten an KundInnen erleichtern. Dabei soll das System die Daten von KundInnen analysieren und eine Einschätzung über deren Kreditwürdigkeit geben. Anschließend sollen durch den Einsatz eines XAI-Modells die Entscheidungen des AI-Modells erklärt und Empfehlungen sowie Vorschläge gegeben werden, die die Bonität der KundInnen verbessern würden. KundInnen erhalten Einblicke in die berücksichtigten Faktoren, was das Vertrauen in die Entscheidungsfindung der AI stärkt. Durch die Erläuterungen können sie notwendige Schritte einleiten, die ihre Chancen auf eine künftige Kreditgenehmigung erhöhen. Die genutzten Daten werden aus der Analyse eines entsprechenden KundInnenprofils geliefert. 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 eine genau Begründung, warum ein Kreditantrag eventuell abgelehnt wird und zudem konstruktive Hinweise wie sie ihr Verhalten eventuell anpassen können.

## Datenaufbereitung

Der genutzte Datensatz enthält 90.657 Einträge, die jeweils einen KundInneneintrag darstellen. Alle Einträge setzen sich aus 27 Attributen zusammen, die da sind: ID („ID“), KundInnen-ID („Customer\_ID“), Monat („Month“), Alter („Age“), Sozialversicherungsnummer („SSN“), Beruf („Occupation“), Jahreseinkommen („Annual\_Income“), Monatsgehalt („Monthly\_Inhand\_Salary“), Anzahl Bankkonten („Num\_Bank\_Accounts“), Anzahl Kreditkarten („Num\_Credit\_Card“), Zinssatz („Interest\_Rate“), Anzahl Darlehen („Num\_of\_Loan“), Art des Darlehens („Type\_of\_Loan“), Verspätung vom Fälligkeitsdatum („Delay\_from\_due\_date“), Anzahl verspäteter Zahlungen („Num\_of\_Delayed\_Payment“), Geändertes Kreditlimit („Changed\_Credit\_Limit“), Anzahl Kreditanfragen („Num\_Credit\_Inquiries“), Kreditmix („Credit\_Mix“), Ausstehende Schulden („Outstanding\_Debt“), Kreditauslastungsquote („Credit\_Utilization\_Ratio“), Alter Kredithistorie („Credit\_History\_Age“), Zahlung von Mindestbetrag („Payment\_of\_Min\_Amount“), Gesamte Rückzahlung pro Monat („Total\_EMI\_per\_month“), Monatlich angelegter Betrag („Amount\_invested\_monthly“), Zahlungsverhalten („Payment\_Behaviour“), Monatlicher Saldo („Monthly\_Balance“) und Bonität („Credit\_Score“).

Um ein Verständnis für die vorliegenden Daten zu bekommen, werden Minima, Maxima, sowie Mittelwerte für die Daten berechnet sowie eine Visualisierung aller Attribute und deren Wertverteilung generiert.

A picture containing text, diagram, plan, screenshot

Description automatically generated

Abbildung 1: Ausschnitt aus der Verteilung der Werte der Attribute

Bevor die Daten dem Klassifikationsmodel 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, deren Attribute fehlerhaft oder unrealistisch sind, aus dem Datensatz entfernt. Ebenso werden Einträge entfernt, bei denen nur unvollständige Informationen vorliegen. Im verbleibenden Datensatz ist somit jede Zeile vollständig, d.h. über jede verbleibende 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.

A screenshot of a computer

Description automatically generated with low confidence

Abbildung 2: Plausibilitätsprüfung

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

A screenshot of a computer

Description automatically generated with low confidence

Abbildung 3: Entfernen irrelevanter Spalten

Im dritten Schritt werden bestimmte Attribute weiter angepasst, um sinnvolle Ergebnisse zu erzielen. Beispielsweise enthält das Attribut Art des Darlehens („Type\_of\_Loan“), eine Auflistung verschiedener Kreditarten pro KundIn. 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 Art des Darlehens („Type\_of\_Loan“) gelöscht.

A picture containing text, screenshot, font, number

Description automatically generated

Abbildung 4: Anpassung des Attributs Art des Darlehens („Type\_of\_Loan“)

Im vierten Schritt wird die Zielvariable in Form des Attributs Bonität („Credit\_Score“) aus dem Datensatz extrahiert.

A picture containing text, screenshot, font, number

Description automatically generated

Abbildung 5: Extraktion der Zielvariable

Im fünften Schritt werden die Attribute in stetige und diskrete Attribute unterteilt. Die stetigen Attribute sind: Alter („Age“), Jahreseinkommen („Annual\_Income“), Monatsgehalt („Monthly\_Inhand\_Salary“), Anzahl Bankkonten („Num\_Bank\_Accounts“), Anzahl Kreditkarten („Num\_Credit\_Card“), Zinssatz („Interest\_Rate“), Anzahl Darlehen („Num\_of\_Loan“), Verspätung vom Fälligkeitsdatum („Delay\_from\_due\_date“), Anzahl verspäteter Zahlungen („Num\_of\_Delayed\_Payment“), Geändertes Kreditlimit („Changed\_Credit\_Limit“), Anzahl Kreditanfragen („Num\_Credit\_Inquiries“), Ausstehende Schulden („Outstanding\_Debt“), Kreditauslastungsquote („Credit\_Utilization\_Ratio“), Alter Kredithistorie („Credit\_History\_Age“), Gesamte Rückzahlung pro Monat („Total\_EMI\_per\_month“), Monatlich angelegter Betrag („Amount\_invested\_monthly“) und Monatlicher Saldo („Monthly\_Balance“).

Die diskreten Attribute lauten: Beruf („Occupation“), Kreditmix („Credit\_Mix“), Zahlung von Mindestbetrag („Payment\_of\_Min\_Amount“), Zahlungsverhalten („Payment\_Behaviour“), Auto-Darlehen (“Auto Loan”), Bauspar-Darlehen („Credit-Builder Loan'”) Privat-Darlehen („Personal Loan”), Schuldenkonsolidierungs-Darlehen („Debt Consolidation Loan”), Wohnungsbau-Darlehen (“Home Equity Loan”), Hypotheken-Darlehen („Mortgage Loan”), Zahltags-Darlehen („Payday Loan”), Studenten-Darlehen („Student Loan”) und Keine Angabe („Not Specified”).

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 mit Hinblick auf den Zusammenhang zur 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.

Nach der Datenvorverarbeitung setzt sich der Datensatz noch aus 84.498 KundInnen mit jeweils 30 Attributen zusammen. Insgesamt führen die beschriebenen Schritte der Datenvorverarbeitung dazu, dass die im folgenden durchgeführten Analysen durchführbar und interpretierbar sind.

## Auswahl AI-System von Optimierung seiner Ergebnisse

Im Bereich der Kreditwürdigkeitsprüfung, in dem eine genaue Vorhersage der Kreditwürdigkeit von größter Bedeutung ist, spielt die Auswahl eines geeigneten Klassifizierungsalgorithmus eine entscheidende Rolle. Unter den verschiedenen verfügbaren maschinellen Lerntechniken hat sich der Random-Forest-Algorithmus aufgrund seiner Stärken und Fähigkeiten als beliebte Wahl für die Kreditwürdigkeitsklassifizierung erwiesen (Demajo, L. M., Vella, V. und Dingli, A. 2020). Dabei spielt vor allem seine hohe Genauigkeit und Robustheit bei der Kreditrisikobewertung eine Rolle. Der Algorithmus kombiniert die Vorhersagekraft mehrerer Entscheidungsbäume, die jeweils unabhängig voneinander arbeiten, um zu einer endgültigen Klassifizierung zu gelangen. Der Ensemble-Charakter des Algorithmus verringert die Überanpassung (Overfitting), indem er mehrere Entscheidungsbäume auf verschiedenen Teilmengen der Daten erstellt und bei jeder Aufteilung zufällige Untergruppen von Merkmalen verwendet. Dieser Ansatz stellt sicher, dass sich das Modell auf die relevantesten Merkmale konzentriert, was zu verbesserten Generalisierungsfähigkeiten führt. Darüber hinaus verarbeitet er auch fehlende Daten ohne signifikanten Verlust an Vorhersagekraft, was ihn für die Kreditwürdigkeitsprüfung robust macht. Durch die Nutzung der kollektiven Weisheit dieser einzelnen Bäume erreicht Random Forest ein höheres Maß an Genauigkeit (Genauigkeit hinsichtlich Evaluationsmaß Accuracy beträgt ~77%) als ein einzelner Entscheidungsbaum.

Zudem erfasst er effektiv komplexe Beziehungen zwischen finanziellen und persönlichen Merkmalen. Lineare Modelle haben oft Schwierigkeiten, diese nichtlinearen Wechselwirkungen zu erfassen, was ihre Vorhersagekraft einschränkt. Eine inhärente Stärke von Random Forest liegt in seiner Fähigkeit, Einblicke in die Bedeutung von Merkmalen zu geben. Er hilft KreditgeberInnen, die wichtigsten Faktoren zu verstehen, die die Kreditwürdigkeit beeinflussen. Darüber hinaus werden unausgewogene Datensätze durch die Einbeziehung von Zufallsstichprobenverfahren berücksichtigt, was zu ausgewogenen Vorhersagen und einer genauen Identifizierung von säumigen Zahlern führt. Seine Parallelisierungsfähigkeit ermöglicht die effiziente Verarbeitung großer Kreditdatensätze und macht es für Kreditscoring-Aufgaben skalierbar.

Nach der Entscheidung für den Random Forest als Klassifizierungsalgorithmus wurde ein Parametertuning angewendet. Dies ermöglicht nicht nur eine Verbesserung der Performance, sondern auch einen zusätzlichen Gewinn an Robustheit. Dafür wurde der Random Forest Klassifikator des sklearn-Paketes in Python genutzt. Für das Parametertuning wurde für folgende Parameter verschiedene Werte getestet:

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

-max\_depth: Gibt die maximale Tiefe des Baumes an. Hierfür wurden 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 wurden die Werte 2, 5 und 10 getestet.

-min\_samples\_leaf: Bestimmt die Mindestanzahl an Trainingsdaten, die erforderlich ist, um einen Endknoten zu bilden. Hierfür wurden die Werte 1, 2 und 4 getestet.

Diese Parameterwerte wurden in allen Kombinationen getestet. Im Anschluss wurden 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 robuste und genauere Schätzungen erhalten werden. Resultierend aus dem Parametertuning/ 5-Folds Cross-Validation-Verfahren haben sich die Parametereinstellungen nach dem Parametertuning als bestmöglich ergeben, die in Abbildung 6 zu sehen sind.



Abbildung 6: Beste Parametereinstellungen nach Parametertuning

Durch das Parametertuning konnte eine Accuracy von 79.02% erreicht werden. Als zweites Gütemaß für die Leistung des Random Forest wurde das F1-Measure berechnet. Hier kann eine Genauigkeit von 78.98% vorgewiesen werden, was auf eine ausgewogene Leistung hinsichtlich der Klassifikation unserer Daten deutet.

## Auswahl XAI-System und Modifikationen

Für die post-hoc XAI-Methode wurde ein Algorithmus gewählt, der kontrafaktische Erklärungen erzeugt. Da das Ergebnis der XAI-Anwendung Erklärungen sein sollen, die sich an KundInnen richten, sind kontrafaktische Erklärungen ideal. Zum einen, weil NutzerInnen auf Basis der veränderten Attribute verstehen können, welche Werte wichtig für die Entscheidung des AI-Modells sind, zum anderen aber auch, weil die kontrafaktischen Ergebnisse im besten Fall auch umsetzbar für KundInnen sind. Im Idealfall könnten KundInnen die Empfehlungen direkt umsetzen, um das gewünschte Resultat zu erzielen.

Als Methode der kontrafaktischen Erklärungen wurde CARE nach Rasouli und Chien Yu (2022) genutzt. Die folgenden Informationen über den Aufbau sowie den Vorteil von CARE richten sich nach der Arbeit von Rasouli und Chieh Yu (2022). CARE ist ein Ansatz, welcher auf einer Paretooptimierung basiert, d.h. es fließen verschiedene Kriterien bei der Suche nach einer geeigneten Folie (der generierten, synthetischen kontrafaktischen Erklärungsinstanz) in die Optimierung ein. Dabei können bis zu sieben Zielfunktionen auf das Optimierungsproblem 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 die Folie in der richtigen Zielklasse landet. „distance“ ist eine Zielfunktion, die den Abstand zwischen Fakt und Folie berechnet. „sparsity“ zählt die Anzahl an Attributen, die sich von Fakt zu Folie unterscheiden.

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 die Folie nahe an den Instanzen des Datensatzes liegt, d.h, dass es sich tendenziell um eine typische Instanz handelt und nicht um einen Ausreißer. Des Weiteren sollen die Änderungen von Fakt zu Folie einem kohärenten, sinnvollen Pfad an Werten folgen.

Das dritte Modul „coherency“ besteht aus der gleichnamigen Zielfunktion. Hierbei soll durch Beachtung der Korrelationen zwischen den Variablen sichergestellt werden, dass die Änderungen zwischen Fakt und Folie kohärent sind. Das vierte und letzte Modul, „actionability“ soll gewährleisten, dass die Änderungen aus NutzerInnensicht machbar sind. Hierbei kann man Restriktionen für Wertebereiche einzelner Variablen setzen, um die Machbarkeit sicherzustellen.

Neben der vielschichtigen Beachtung verschiedener Konstrukte für die Generierung von kontrafaktischen Erklärungen, ist ein weiterer Vorteil von CARE die Verwendung eines gradientenfreien Optimierers. Dieser hat den Vorteil, auf natürlichem Wege mit kategorischen Variablen arbeiten zu können. Im gegebenen Datensatz waren einige kategorische Variablen enthalten (z.B. Beruf („Occupation“)), weshalb dieses Kriterium als wichtig erachtet wurde. Zusätzlich kann man CARE vorgeben, wie viele Folien aus einem Fakt entstehen sollen. Dies hilft besonders, wenn die kontrafaktischen Erklärungen Folien ausgeben, die zwar in der richtigen Zielklasse liegen, aber in der Realität nicht umsetzbar sind. Werden jedoch mehrere Folien generiert, steigt die Wahrscheinlichkeit, dass eine der Folien einen realistischen Vorschlag ausgibt. Anhand des Aufzeigens von alternativen Folien können weitere Einblicke in das AI-System gewonnen werden, da man die AI-Entscheidungsgrundlage ablesen kann. Zu guter Letzt ist CARE model-agnostisch ist, d.h. es kann unabhängig vom zugrundeliegenden AI-Modell verwendet werden.

Nachdem die Daten aufbereitet und der Random Forest trainiert wurde, wurde CARE noch an den Use Case angepasst. Hierfür wurde die Zielklassenzuweisung von CARE technisch angepasst. Im Originalansatz wird jene Klasse als Zielklasse verwendet, welche bei der Prognose des AI-Modells die zweithöchste Klassenwahrscheinlichkeitszuordnung annimmt. Für die Bearbeitung der Seminaraufgabe war es jedoch sinnvoller, die Zuordnung klar zu formulieren. So gelten folgende Zuordnungen für die Zielklasse:

Ist der Fakt in Klasse „Poor“ klassifiziert worden, so lautet die Zielklasse der Folie „Standard“. Lautet die Klassifikation des Fakts „Standard“, so ist die Zielklasse der Folie „Good“. Ist der Fakt in „Good“ klassifiziert worden, so ist die Zielklasse „Standard“.

Zusätzlich wurde ein „Probability Thresh” von 0.45 festgelegt. Die Folie muss mindestens diese Wahrscheinlichkeit vom zugrundeliegenden AI-Modell in der Zielklasse erhalten. Um die Ergebnisse zu speichern, wurde ein Loop implementiert, in welchem die Folien in einer pickle-Datei gespeichert werden.

Um die kontrafaktischen Erklärungen noch zu verbessern und alle Module zu aktivieren, wurden verschiedene Restriktionen festgelegt. In der Realität lassen sich manche Attribute leichter anpassen als andere. Manche Attribute können von KundInnen unmöglich geändert werden (beispielsweise ihr 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. Werden allerdings zu viele Attribute eingeschränkt, liefert der CARE-Algorithmus keine guten kontrafaktischen Erklärungen mehr. Deshalb empfiehlt es sich, einen 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 Restriktionen zu acht Attributen festgelegt.



Abbildung 7: Ausgewählte Restriktionen

Das Alter („Age“) und der Beruf („Occupation“) sind Attribute, die von KundInnen nur schwer geändert werden können. Die Zahlen am Ende der Klammern stehen dabei für die Wichtigkeit dieser Fixierung. Da das Alter („Age“) nicht geändert werden kann, der Beruf („Occupation“) jedoch mit viel Aufwand (bspw. einer Umschulung) eventuell angepasst werden könnte, wird die Restriktion des Alters mit einem hohen Strafterm ausgestattet, während das Attribut Beruf nur mit einem geringen Strafterm versehen wird. Der Zinssatz („Interest\_Rate“) und die Alter Kredithistorie („Credit\_History\_Age“), sind Attribute, die durch vergangenes Verhalten entstanden sind und somit ebenso nicht verändert werden können. Deshalb erhalten auch diese zwei Attribute den Strafterm 10. Die Attribute Jahreseinkommen („Annual\_Income“), Monatsgehalt („Monthly\_Inhand\_Salary“), Ausstehende Schulden („Outstanding\_Debt“) sowie Monatlich angelegter Betrag („Amount\_invested\_monthly“) 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.

Diese gewählten Restriktionen schränken den CARE-Algorithmus ausreichend ein, um sinnvolle und realitätsnahe kontrafaktischen Erklärungen zu erzeugen. Weitere Restriktionen, die die Realität noch besser repräsentieren würden, schränken den Algorithmus jedoch zu sehr ein und es werden keine sinnvollen kontrafaktischen Erklärungen mehr generiert. Um kundInnenfreundliche und konstruktive Erklärungen zu liefern, wurden deshalb diese Restriktionen und keine weiteren festgelegt.

## Auswertung

Da nun alle technischen, notwendigen Schritte implementiert wurden, konnten kontrafaktische Erklärungen generiert werden. Hierfür wurden zufällig ausgewählte Instanzen aus dem Testdatensatz herangezogen, welche allesamt in einem Lauf durch den geschriebenen Loop an den CARE-Algorithmus übergeben wurden. Im Folgenden werden drei Beispielerklärungen gezeigt: Die ersten zwei repräsentieren gelungene kontrafaktische Erklärungen, letztere wiederrum eine misslungene Erklärung. Die Interpretationen und Eindrücke werden im darauffolgenden Diskussionsteil ([Kapitel 5](#_Dikussion)) wiedergegeben.



Abbildung 8: Ausschnitt von Fakt (hier „x“) und dazugehöriger Folie (hier „cf\_0“)

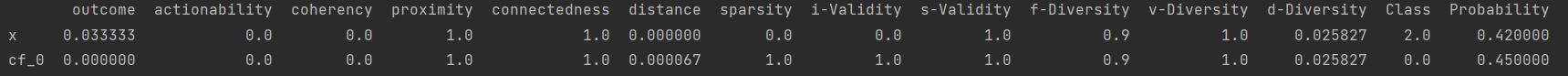


Abbildung 9: Evaluationsmetriken um Unterschied von Fakt („x“) zu Folie („cf\_0“) fassbar zu machen

In Abbildung 8 ist eine Ausgangsinstanz „x“ mit ihren Attributsausprägungen in den jeweiligen Variablen zu sehen, welche hier als Spaltenbezeichnungen angezeigt werden. Die Instanz wurde in die Kategorie „Standard“ von unserem AI-Modell klassifiziert (Abbildung 9: „Class“ bei „x“ = 2). Direkt unter den Attributsausprägungen von „x“ ist in Abbildung 8 zu sehen, welche Merkmale sich von Fakt zu Folie geändert haben. Liegt bei einer Folie keine Änderung bei einer Variable im Vergleich zu dem dazugehörigen Fakt vor, so wird es hier als „\_“ angezeigt. In der repräsentativen Folie sieht man, dass nur die Anzahl der Kreditkarten („Num\_Credit\_Card“) von fünf auf drei Karten sinkt. Diese Änderung reicht, um die Folie von „Standard“ auf „Good“ (Abbildung 9 unter „cf\_0“: „Class“ = 0) vom AI-Modell prognostizieren zu lassen.



Abbildung 10: Weiteres Beispiel 1 eines Fakts (hier „x“) und dazugehöriger Folie (hier „cf\_0“) mit ausgewählten Attributen

Im zweiten Beispiel gibt es ebenfalls eine Änderung zwischen Fakt und Folie. Hier ändert sich die Prognose des AI-Modells von „Poor“ (Klassifizierung des Fakts) zu „Standard“ (Klassifizierung der Folie) durch die Änderung an ausstehenden Schulden („Outstanding\_Debt“) von 1630.08 zu 703.21.



Abbildung 11: Weiteres Beispiel 2 eines Fakts (hier „x“) und dazugehöriger Folie (hier „cf\_0“) mit ausgewählten Attributen

Mit dem letzten Beispiel einer generierten Folie wird verdeutlicht, dass auch kontrafaktische Erklärungen entstehen können, welche keine optimalen oder brauchbaren Erklärungen liefern. So kann es sein, dass der Algorithmus bei der Suche nach einer geeigneten Folie auf keine geeignete Lösung stößt. Das vorliegende Beispiel aus Abbildung 11 zeigt, dass sich im Gegensatz zu den bisherigen zwei Beispielen viele Variablen in „cf\_0“ im Vergleich zu „x“ ändern. Insgesamt haben sich 19 Variablen zwischen Fakt und Folie geändert.

# Diskussion

In dieser Arbeit wurde auf Basis eines Datensatzes aus dem Bankenbereich sowohl ein AI-Modell als auch ein XAI-Modell implementiert, um die Kreditwürdigkeiten von KundInnen zu prognostizieren und die Entscheidung des AI-Modells zu erklären. Dafür wurden zunächst die Daten aufbereitet, indem Variablen gelöscht, bearbeitet und eingeschränkt wurden. Anschließend wurde ein Random Forest als AI-Modell trainiert und per Hyperparametertuning angepasst, um die Performance zu verbessern. Da zur Bearbeitung des Use Case die KundInnenperspektive angenommen wurde, wurden bezüglich des XAI-Systems ein Ansatz der kontrafaktischen Erklärungen gewählt. Mit dem Algorithmus CARE, welcher hinsichtlich verschiedener Punkte an den Use Case angepasst wurde, konnten kontrafaktische Erklärungen erzeugt werden. Anhand der ersten beiden Beispiele aus [Kapitel 3.5](#_Auswertung) kann die Wirksamkeit des Ansatzes, welcher die Erwartungen an das XAI-System erfüllt, bewiesen werden. KundInnen können zum einen verstehen, welche Variablen wichtig für die Entscheidungsbasis sind, zum anderen bekommen sie aber auch Informationen darüber, welche Schritte sie gehen können, um eine bessere Kreditwürdigkeit zu erhalten. So kann man in Abbildung 8 bzw. Abbildung 11 Erkenntnis darüber gewinnen, dass die Anzahl an Kreditkarten bzw. die ausstehenden Schulden ein wichtiges Kriterium für die Entscheidung des AI-Systems sind. Dadurch werden wichtige Informationen geliefert, die weiteres Vertrauen in das zugrundeliegende AI-Modell schaffen.

Obwohl der Random Forest in einigen Fällen eine gute Leistung zeigte, war er nicht durchgängig genau und erreichte eine Gesamtgenauigkeit (Accuracy) von 79.02%. Diese Schwankungen in der Leistung machen deutlich, dass weitere Verbesserungen erforderlich sind, um zuverlässigere Vorhersagen zu gewährleisten.

Darüber hinaus gab es Fälle, in denen das XAI-Modell Erklärungen lieferte, die nicht optimal waren, da es möglicherweise wichtige Faktoren wegrationalisierte oder entscheidende Variablen übersah bzw. diese im Vornherein bei der Bereinigung entfernt wurden. Auch kann es sein, dass das XAI-Modell Probleme damit hat, Ausreißer zu behandeln. So hätten sich in Abbildung 11 19 der 30 Variablen ändern müssen, um eine bessere Bonität zu erreichen. Dies wäre in der Realität kaum umsetzbar. Daher wird hier auch das Ziel verfehlt, entsprechenden KundInnen die Entscheidung des AI-Modells erklärbar zu machen.

Ein weiterer Nachteil des Projekts war die für die Auswertung des CARE-Algorithmus Modells erforderliche Laufzeit. Die Rechenkomplexität des Algorithmus kann zu längeren Verarbeitungszeiten führen, was bei Echtzeit-Kreditbewertungsanwendungen eine praktische Einschränkung darstellen kann.

Trotz dieser Einschränkungen sollten die Vorteile der Verwendung eines Random Forest und die des CARE-Algorithmus für die Kreditwürdigkeitsklassifizierung sowie der dazugehörigen Erklärungen nicht außer Acht gelassen werden. Der Random Forest konnte im Vergleich zu anderen Klassifikationsalgorithmen am besten abschneiden, zudem waren die Ergebnisse von einer höheren Robustheit geprägt. CARE verarbeiten erfolgreich kategorische Variablen und bieten Einblicke in wichtige Merkmale, die die Kreditwürdigkeit beeinflussen, und hilft KreditgeberInnen und KundInnen, die Faktoren zu verstehen, die die Entscheidungen des AI-Modells beeinflussen. Die vom XAI-System generierten kontrafaktischen Erklärungen durch den CARE-Algorithmus bieten den KundInnen wertvolle Informationen, um ihre Kreditwürdigkeit zu verbessern und ihr Verständnis für den Entscheidungsprozess zu erhöhen.

Es ist jedoch von entscheidender Bedeutung, die im Projekt hervorgehobenen negativen Aspekte, wie die fehlende Interpretation, die Variabilität der Leistung, die Laufzeit und die potenziell suboptimalen Erklärungen anzugehen. Zukünftige Forschungsarbeiten könnten sich darauf konzentrieren, die Interpretierbarkeit des AI-Modells zu verfeinern, seine Genauigkeit und Konsistenz zu verbessern, die Laufzeit zu optimieren und die Qualität der Erklärungen für die KundInnen zu verbessern.

# Fazit

Der Einsatz von AI und XAI bietet in vielen verschiedenen Feldern und Anwendungsbereichen diverse Vor- und Nachteile. In einem Finanzinstitut bietet die Verwendung von Klassifikationsalgorithmen und kontrafaktischen Erklärungsmethoden bei der Kreditwürdigkeitsprüfung viele Chancen durch genaue Kreditrisikobewertungen, der Fähigkeit zur Bereitstellung von Einblicken in einflussreiche Merkmale und Hinweisen zur Veränderung der eigenen Kreditwürdigkeit.

Ein solches Wissen ermöglicht präzisere Risikomanagementstrategien und hilft den Banken, ihre Kreditvergabepraxis zu optimieren und potenzielle Verluste zu minimieren. Die Verkürzung der Bearbeitungszeiten hilft außerdem, Kreditanträge effizient zu bewerten und zeitnah Antworten zu geben. Zusätzlich kann die Kundenzufriedenheit gesteigert werden, was sich wiederrum auch monetär positiv auswirken könnte.

Es ist aber auch wichtig, die Grenzen solcher Anwendungsfälle miteinzubeziehen. Eine Gesamtgenauigkeit von x% eines AI-Modells geht mit einer Fehlerquote von 100-x% einher. Diese KundInnen werden falsch eingeschätzt, was negative Konsequenzen mit sich ziehen kann. Die Anwendung von kontrafaktischen Erklärungen zeigt sich zwar als sehr hilfreich, ist aber nicht immer optimal. Das liegt auch daran, dass es ein junges Forschungsgebiet ist, welches viel Potential mit sich zieht, aber noch weiter ausgereift werden sollte.

In Anbetracht dieser Einschränkungen sollten Bankangestellte die Anwendbarkeit von kontrafaktischen Erklärungen bei der Kreditwürdigkeitsprüfung sorgfältig prüfen.

Im Rahmen des Seminars konnten erfolgreich kontrafaktische Erklärungen zu Entscheidungen eines AI-Modells mit Hilfe von KundInnendaten aus einem gegebenen Datensatz erstellt werden. Dabei wurden pro KundIn mehrere kontrafaktische Erklärungen generiert, welche durch Berücksichtigung von verschiedenen Zielfunktionen und Restriktionen gekennzeichnet sind. In der Realität könnten Bankangestellte ein passendes, zu änderndes Attribut identifizieren und die Erklärung der XAI im persönlichen Beratungsgespräch mit KundInnen einsetzen, um ihre angebotene Serviceleistung zu bereichern. Damit ist das Ziel des Seminars, erfolgreich ein XAI-Modell auf ein AI-Modell anzuwenden, erfüllt.

|  |
| --- |
| Literaturverzeichnis  Balogh, S. und Johnson, C. (2021): The Risks and Opportunities of AI in Credit Decisioning. Hg. v. Business Insider. Online verfügbar unter https://www.businessinsider.com/ai-lending-risks-opportunities-credit-decisioning-data-inequity-2021-6, zuletzt geprüft am 01.07.2023.  Boobier, T. (2018): Advanced analytics and AI. Impact implementation and the future of work. Wiley finance series: John Wiley & Sons Ltd.  Chou, Y.‑L., Moreira, C., Bruza, P., Ouyang, C., & Jorge, J. (2021): Counterfactuals and Causability in Explainable Artificial Intelligence. Theory, Algorithms, and Applications. Online verfügbar unter http://arxiv.org/pdf/2103.04244v2, zuletzt geprüft am 01.07.2023.  Demajo, L. M., Vella, V. und Dingli, A. (2020): Explainable AI for Interpretable Credit Scoring. Hg. v. AIRCC Publishing Corporation. Online verfügbar unter http://dx.doi.org/10.5121/csit.2020.101516, zuletzt geprüft am 01.07.2023.  Emerj (2019): AI Transparency in Finance – Understanding the Black Box. Online verfügbar unter https://emerj.com/partner-content/ai-transparency-in-finance/, zuletzt geprüft am 01.07.2023.  European Parliament (2020): The EU’s response to the COVID-19 outbreak. A European Green Deal perspective. Online verfügbar unter https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641547/EPRS\_STU%282020%29641547\_EN.pdf, zuletzt geprüft am 01.07.2023.  Forbes Tech Council (2021): Five Industries Reaping The Benefits Of Artificial Intelligence. Online verfügbar unter https://www.forbes.com/sites/forbestechcouncil/2021/06/02/five-industries-reaping-the-benefits-of-artificial-intelligence/, zuletzt geprüft am 01.07.2023.  Guidotti, R., Monreale, A., Giannotti, F., Pedreschi, D., Ruggieri, S., & Turini, F. (2019): Factual and Counterfactual Explanations for Black Box Decision Making (IEEE Intelligent Systems, 34(6)). Online verfügbar unter https://doi.org/10.1109/MIS.2019.2957223, zuletzt geprüft am 01.07.2023.  IBM (o.J.): What is Explainable AI (XAI)? Online verfügbar unter https://www.ibm.com/watson/explainable-ai, zuletzt geprüft am 01.07.2023.  McKinsey & Company (2021): Designing next-generation credit-decisioning models. Online verfügbar unter https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/designing-next-generation-credit-decisioning-models, zuletzt geprüft am 01.07.2023.  Miller, T. (2019): Explanation in artificial intelligence. Insights from the social sciences. Artificial Intelligence, 267. Online verfügbar unter https://doi.org/10.1016/j.artint.2018.07.007, zuletzt geprüft am 01.07.2023.  QAI (2022): The Pros And Cons Of Artificial Intelligence. Hg. v. Forbes. Online verfügbar unter https://www.forbes.com/sites/qai/2022/12/01/the-pros-and-cons-of-artificial-intelligence/, zuletzt geprüft am 01.07.2023.  Rasouli, P. und Chieh Yu, I. (2022): CARE: coherent actionable recourse based on sound counterfactual explanations. Hg. v. Springer Science and Business Media LLC. Online verfügbar unter http://dx.doi.org/10.1007/s41060-022-00365-6, zuletzt geprüft am 29.06.2023.  Samek, W., Wiegand, T. und Müller, K. (2017): Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. Online verfügbar unter https://arxiv.org/abs/1708.08296, zuletzt geprüft am 01.07.2023.  Stepin, I., Alonso, J. M., Catala, A. und Pereira-Fariña, M. (2021): A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence. Online verfügbar unter https://ieeexplore.ieee.org/abstract/document/9321372/metrics, zuletzt geprüft am 02.07.2023. |

Ehrenwörtliche Erklärung

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

Exploring the Use of a Random Forest and   
a Counterfactual Explanations Approach in Credit Scoring

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