Erklärung algorithmischer Entscheidungssysteme im Kreditvergabeprozess – ein Ansatz auf Basis emergenter Statistik

2. Projektarbeit

Jahrgang 2017

Kurs B

Fakultät für Wirtschaft

Studiengang Betriebswirtschaftslehre Bank

DUALE HOCHSCHULE BADEN-WÜRTTEMBERG

VILLINGEN SCHWENNINGEN

Bearbeiter: Patrik Reichelt Betreuer: Prof. Dr. André Kuck

Ausbildungsbetrieb: Sparkasse Hegau-Bodensee

Inhaltsverzeichnis

Seite

Abkürzungsverzeichnis

Abb.	Abbildung
Abs.	Absatz
ADM	Algorithmic Decision Making
a.F.	alte Fassung
AT	Allgemeiner Teil
Art.	Artikel
BaFin	Bundesanstalt für Finanzdienstleistungsaufsicht
BDSG	Bundesdatenschutzgesetz
BGH	Bundesgerichtshof
BSDG	Bundesdatenschutzgesetz
bspw.	beispielsweise
bzw.	beziehungsweise
d.h.	das heißt
DSGVO	Datenschutz-Grundverordnung
e.V.	eingetragener Verein
ErwG	Erwägungsgrund
EU	Europäische Union
f.	folgende
ff.	fortfolgende
FICO	Fair Isaac Corporation
Gem.	Gemäß
i.d.R.	in der Regel
i.H.v.	in Höhe von
i.S.d.	im Sinne des/der

KI	Künstliche Intelligenz
KWG	Kreditwesengesetz
LIME	Local Interpretable Model-Agnostic Explanations
MaRisk	Mindestanforderungen an das Risikomanagement
ML	Machine Learning
Nr.	Nummer
o.ä.	oder ähnlichem
o.g.	oben genannte(n)
o.V.	ohne Verfasser
S.	Seite
s.	siehe
s.o.	siehe oben
SCHUFA	Schutzgemeinschaft für allgemeine Kreditsicherung
sog.	so genannt(e)
TGAMT	Transparent Generalized Additive Model Tree
u.a.	unter anderem
u.U.	unter Umständen
Vgl.	Vergleiche
z.B.	zum Beispiel
z.T.	zum Teil

Symbolverzeichnis

T	=	Emergenzmenge
TU	=	Universelle Emergenzmenge
DiV	=	Degree of inductive Verification

Abbildungsverzeichnis

Anhangverzeichnis

Hinweis:

Der Anhang liegt der gedruckten Fassung dieser Arbeit als Speichermedium bei.

1 Einleitung

1.1 Problemstellung

Künstliche Intelligenz ist in der heutigen Zeit längst nicht mehr nur auf dem Vormarsch. In vielen Bereichen ist künstliche Intelligenz aus dem Alltag nicht mehr wegzudenken. Digitale, intelligente Helfer wie Alexa (Amazon) oder Siri (Apple) koordinieren den Kalender, erinnern uns auf Wunsch an alles, was man in der Hektik des Alltags vergessen könnte und steuern gerne auch Licht, Strom und Heizung in den eigenen vier Wänden. Künstliche Intelligenz ist außerdem in weiteren Bereichen im Einsatz, in denen man sie trotz täglicher Nutzung kaum wahrnimmt. So wählen Algorithmen anhand des Nutzungsverhaltens aus, welche Serien für den Nutzer beim Fernsehabend als nächstes in Frage kommen könnten, oder welches Produkt ihm beim Online-Einkauf als nächstes zusagen könnte. Dabei lernen die Algorithmen aus ihren gemachten Erfahrungen, d.h. aus den Daten, die sie gesammelt haben.

Wenngleich der praktische Nutzen derartiger Systeme gerne in Anspruch genommen wird, so wissen die Nutzer in der Regel dennoch nichts darüber, wie die Entscheidungen der Systeme zustande kommen und welche Technologie dahintersteckt. Laut einer Umfrage der Bertelsmann Stiftung haben 72% der Befragten den Begriff „Algorithmus“ schon einmal gehört, 87% davon gaben jedoch an, kaum etwas über den Begriff zu wissen bzw. lediglich eine ungefähre Vorstellung von diesem Begriff zu haben.[1] Für den Nutzer stellt sich stets die Frage, warum er der Empfehlung bzw. der Entscheidung des Algorithmus vertrauen sollte. In den bislang beschriebenen Anwendungsgebieten künstlicher Intelligenz ist davon auszugehen, dass der Nutzer bei mangelndem Vertrauen das System nicht weiter benutzen wird. Doch was, wenn sich die nächste Empfehlung nicht mehr mit alltäglichen Situationen beschäftigt, sondern mit der Entscheidung über den gestellten Kreditantrag, oder mit der Auswahl von Bewerbern im beruflichen Bewerbungsverfahren? Derartige Fälle sind für die Betroffenen häufig von großer Bedeutung und sie können sich der automatisierten Entscheidung oftmals nicht entziehen. Umso wichtiger ist es

[1] Vgl. Bertelsmann Stiftung, 2018a, S. 14.

hierbei für die Betroffenen, die getroffenen Entscheidungen und entsprechende Begründungen nachvollziehen zu können.

Die Nachvollziehbarkeit automatisierter Entscheidungen ist auch aus regulatorischer Sichtweise von großer Relevanz. Nur wenn die Systeme durch Dritte nachvollzogen werden können, können sie von unabhängigen Aufsichtsbehörden überwacht werden, um Probleme wie z.B. Diskriminierung zu verhindern und das Vertrauen der Betroffenen zu stärken. Diesem Umstand ist sich die Gesetzgebung bewusst und hat mit Erlass der Datenschutz-Grundverordnung (DSGVO)[2] Standards auf europäischer Ebene geschaffen, welche die Verantwortlichen für den Einsatz von automatisierten Entscheidungssystemen mit personenbezogenen Daten dazu verpflichten, die betroffenen Personen über die Hintergründe der getroffenen Entscheidungen aufzuklären.

Komplexe Algorithmen sind dazu in der Lage, äußerst genaue Ergebnisse für verschiedenste Anwendungszwecke zu erreichen. Mit steigender Komplexität der Systeme werden die Ergebnisse jedoch immer weniger nachvollziehbar.[3] Um trotzdem eine Nachvollziehbarkeit der Ergebnisse für Dritte gewährleisten zu können bedarf es eines Erklärungsmodelles, um komplexe künstliche Intelligenz möglichst einfach erklären zu können.[4]

1.2 Zielsetzung und Aufbau

Ziel der vorliegenden Arbeit ist die Erörterung der Notwendigkeit einer erklärbaren künstlichen Intelligenz im Kreditvergabeprozess sowie die Präsentation verschiedener Ansätze zur Erklärung algorithmischer Entscheidungssysteme.

Im ersten Abschnitt werden zunächst mögliche Risiken und Fehlerquellen für den Einsatz von automatisierten Entscheidungssystemen beschrieben. Darauf aufbauend erfolgt eine Darstellung der regulatorischen Behandlung derartiger Systeme. Der Schwerpunkt hierbei liegt insbesondere auf den Vorschriften der Datenschutz-

[2] Verordnung (EU) 2016/679 des Europäischen Parlaments und des Rates vom 27. April 2016 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95/46/EG (Datenschutz-Grundverordnung)
[3] Vgl. Hornung/Herfort in König/Schröder/Wiegand, 2018, S. 170.
[4] Vgl. Molnar, 2019, S. 15.

grundverordnung zur Zulässigkeit und Transparenz beim Einsatz der Systeme. Abschließend folgt eine Beurteilung der Eignung der getroffenen Regelungen zur Vermeidung der beschriebenen Risiken.

Im darauffolgenden Kapitel wird zunächst eine Kategorisierung für Erklärungsansätze von algorithmischen Entscheidungssystemen hinsichtlich ihres Umfangs und der Qualität der Erklärungen eingeführt. Anschließend erfolgt eine Einordnung in diese Kategorien sowie die Präsentation der Grundlagen zweier bekannter Ansätze, welche zur Erklärung automatisierter Kreditentscheidungen Anwendung finden können.

Im vierten Kapitel wird ein weiterer Ansatz zur Erklärung automatisierter Kreditentscheidungen auf Basis emergenter Statistik präsentiert. Dazu werden zunächst die Grundlagen und die Terminologie der emergenten Statistik erläutert, ehe ein konkretes Modell zur Bewertung von Kreditanträgen und zur Erklärung der Ergebnisse beschrieben wird.

Im abschließenden fünften Kapitel erfolgt eine kritische Würdigung der in vorgegangenen Kapiteln beschriebenen Erklärungsansätze sowie eine Einordnung derselben in den Gesamtkontext.

2 Notwendigkeit erklärbarer KI

2.1 Risiken durch ADM

Der Einsatz von künstlicher Intelligenz (KI) in Form von algorithmic-decision-making-Systemen (ADM-Systemen) bietet den Anwendern eine Reihe von Vorteilen und Chancen. Algorithmen können deutlich größere Datenmengen verarbeiten und ihre Erfahrungen daraus gewinnen als ein menschlicher Entscheidungsträger. Sie lassen sich dabei nicht von Gefühlen o.ä. leiten und treffen die Entscheidungen völlig rational. Trotzdem beinhalten ADM-Systeme auch Fehlerpotentiale, die insbesondere dann, wenn die Systeme zur Verarbeitung personenbezogener Entscheidungen (z.B. Kreditentscheidungen, Bewerbungsverfahren) eingesetzt werden, Risiken für Einzelpersonen, Gruppen sowie für die gesamte Gesellschaft darstellen können. Die Fehlerpotentiale bestehen dabei in sämtlichen Phasen des Entwicklungs- und Anwendungsprozesses von ADM-Systemen.[5]

[5] Vgl. Bertelsmann Stiftung, 2018, S. 14.

Bereits bei der Entwicklung der Algorithmen kann ein fehlerhaftes mathematisch-statistisches Modell von vorne herein dazu führen, dass die Ergebnisse, die das System liefert, falsch sein werden.[6] Gleichsam fehlerhafte Ergebnisse können entstehen, wenn die zu analysierenden Sachverhalte fehlerhaft, bzw. für den Anwendungszweck nicht hinreichend quantifiziert und in Variablen dargestellt werden.[7] Modelle, die sich zu stark auf kurzfristige Effizienzmaximierung konzentrieren, können sich negativ auf die Erreichung langfristiger sozialer Interessen auswirken, was ebenfalls als Risiko bereits in der Entwicklungsphase angeführt werden kann.[8] Ein weiteres Risikopotential kann entstehen, wenn Algorithmen und ADM-Systeme für andere als die bei der Entwicklung vorgesehenen Zwecke eingesetzt werden. Selbst wenn die Einsatzgebiete ähnlich gelagert zu sein scheinen, kann die Qualität des verwendeten Systems für solche Einsatzgebiete unter Umständen unzureichend sein.[9]

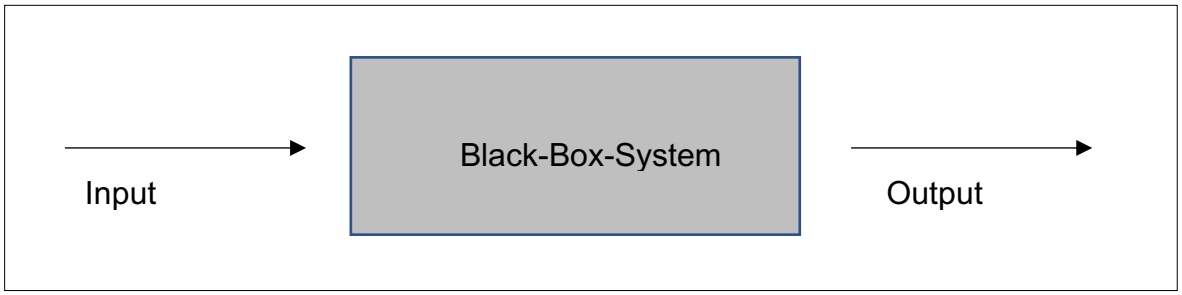
Weitere Risiken können bei der Implementierung eines ADM-Systems in den Praxiseinsatz auftreten. Wird das System bspw. auf Grundlage einer zu geringen Datenmenge betrieben, sind statistisch verwertbare Ergebnisse im Lernprozess des Systems praktisch ausgeschlossen.[10] Weitere Probleme für die statistische Qualität des Systems können auftreten, wenn bereits in den vorgegebenen Datensätzen Verzerrungen („Biases“) enthalten sind. Solche Biases führen dazu, dass der Algorithmus anhand fehlerhafter Daten lernt, und diese für die Zukunft fehlerhaft weiterverwendet, wodurch die Verzerrungen weiter verstärkt werden.[11]

Ist für Außenstehende (z.B. betroffene Personen) nicht ersichtlich, wie eine von einem ADM-System getroffene Entscheidung zustande gekommen ist, spricht man von einer „Blackbox“.[12] Die Daten, die dem System vorgegeben werden, sowie das Ergebnis der Datenverarbeitung sind zwar bekannt, hinsichtlich der Verarbeitung der Daten im Inneren des Systems besteht jedoch keine Transparenz (s. Abb. 1).

---

[6] Vgl. Konrad-Adenauer-Stiftung e.V., 2019, S. 6.
[7] Vgl. Bertelsmann Stiftung, 2018, S. 14.
[8] Vgl. Bertelsmann Stiftung, 2018, S. 14.
[9] Vgl. Konrad-Adenauer-Stiftung e.V., 2019, S. 7.
[10] Vgl. Konrad-Adenauer-Stiftung e.V., 2019, S. 6.
[11] Vgl. Bertelsmann Stiftung, 2018, S. 14.
[12] Vgl. Vogl/Waltl, 2018, S. 613.

Abb. 1: Black-Box-System



Quelle: Eigene Darstellung, in Anlehnung an Samek/Wiegand/Müller, 2017, S. 4.

Sämtliche oben beschriebenen Problemfelder stellen Risiken für Einzelpersonen, Gruppen von betroffenen Personen sowie die Gesellschaft als Ganzes dar, da sie zu Benachteiligungen derselben führen können. Rein automatisiert getroffene Entscheidungen können zu sozialer Ungleichbehandlung, fehlender Fairness sowie Diskriminierung führen. Dies kann dadurch bedingt sein, dass das zugrunde gelegte Modell als solches fehlerhaft ist, dass entsprechende Tendenzen bereits in den Trainingsdaten des Systems vorhanden waren und das System auf diese Weise falsch trainiert wurde, oder dass das System selbständig Gründe für eine systematische Benachteiligung bestimmter Personengruppen gefunden hat.

Als Beispiel für eine solche Diskriminierung dient eine Begebenheit aus dem Jahr 2015. Bei einem US-amerikanischen Unternehmen für Onlinehandel wurden Bewerber im Bewerbungsprozess vorab durch einen Algorithmus bewertet. Da dieser anhand der bisher im Unternehmen angestellten Mitarbeiter trainiert wurde, und der Großteil der Mitarbeiter im Unternehmen männlichen Geschlechts waren, wurden weibliche Bewerber bei der Bewertung durch den Algorithmus systematisch nachteilig beurteilt.[13]

2.2 Regulatorische Behandlung von ADM-Systemen

Um die in 2.1 aufgeführten Risiken, welche durch den Einsatz von ADM-Systemen auftreten können, begrenzen zu können, wurden aufsichtsrechtliche Vorschriften im Bereich Datenschutz geschaffen, um die Zulässigkeit von ADM-Systemen zu regeln und eine hohe Transparenz der getroffenen Entscheidungen zu gewährleisten. Insbesondere die in 2018 erlassene Datenschutzgrundverordnung soll diesbezüglich grenzübergreifend Standards setzen. Im Folgenden werden die wichtigsten Bereiche des Datenschutzrechts mit Bezug zu ADM-Systemen dargestellt. Dabei erfolgt

[13] Vgl. Dastin, 2018.

stets eine Bezugnahme zur Behandlung von ADM-Systemen im Kreditvergabeprozess.

2.2.1 Zielsetzungen der Regulatorik

Die DSGVO soll grundsätzlich den Schutz von personenbezogenen Daten gewährleisten.[14] Betroffenen Personen soll die Kontrolle über ihre eigenen Daten ermöglicht werden.[15] Dabei sollen die festgelegten Standards ein auf EU-Ebene länderübergreifend gleichwertiges Datenschutzniveau schaffen.[16] Ziel der DSGVO im Hinblick auf die in 2.1 aufgeführten Risiken von ADM-Systemen ist es, eine größtmögliche Transparenz der betroffenen Personen insbesondere bei automatisierter Datenverarbeitung zu schaffen, jedoch gleichzeitig die Interessen der verantwortlichen Personen (insbesondere Betriebsgeheimnisse) zu schützen.[17]

Zusätzlich zu den Transparenzzielen soll der Einsatz von ADM-Systemen im Allgemeinen reguliert werden, um betroffene Personen vor auftretenden rechtlichen Konsequenzen oder ähnlichen Beeinträchtigungen zu schützen, die aus rein automatisierten Entscheidungen resultieren, und vor denen die betroffene Person sich sonst nicht entziehen könnte. Als Beispiel hierfür werden automatisiert abgelehnte Online-Kreditanträge oder Bewerbungsverfahren aufgeführt.[18] Außerdem soll für eine automatisierte Verarbeitung personenbezogener Daten sichergestellt werden, dass für die Verarbeitung „geeignete mathematische oder statistische Verfahren“[19] angewendet werden, um zu verhindern, dass eine Benachteiligung einzelner betroffener Personen aufgrund bestimmter personenbezogener Daten wie z.B. Rasse, politischer Einstellung oder Religion benachteiligt werden.[20]

2.2.2 Zulässigkeit von ADM-Systemen nach Art. 22 DSGVO

Gemäß Art. 22 Abs. 1 DSGVO haben natürliche Personen das Recht, dass ihnen gegenüber zu treffende Entscheidungen, welche entweder rechtliche Konsequen-

[14] Vgl. ErwG 1 DSGVO.
[15] Vgl. ErwG 7 DSGVO.
[16] Vgl. ErwG 10 DSGVO.
[17] Vgl. ErwG 63 DSGVO.
[18] Vgl. ErwG 71 DSGVO.
[19] ErwG 71 DSGVO.
[20] Vgl. ErwG 71 DSGVO.

zen beinhalten, oder die betroffenen Personen „in ähnlicher Weise erheblich beeinträchtigen“[21], nicht ausschließlich auf einer automatisierten Datenverarbeitung (einschließlich Profiling) beruhen. Unter Profiling ist gem. Art. 4 Nr. 4 DSGVO jede automatisierte Verarbeitung personenbezogener Daten zu verstehen, welche zum Zweck hat, die betroffene Person hinsichtlich persönlicher Aspekte wie Zuverlässigkeit oder wirtschaftlicher Lage der Person zu analysieren bzw. prognostizieren.[22]

Damit sind rein automatisiert getroffene Entscheidungen sowie Entscheidungen rein auf Basis von Profiling, die die betroffene Person benachteiligen, vom Grundsatz her verboten. In Art. 22 Abs. 2 DSGVO werden jedoch drei Ausnahmen von diesem Verbot ermöglicht. Nach Buchstabe a) des Art. 22 Abs. 2 DSGVO ist eine rein automatisierte Entscheidung i.S.d. Abs. 1 zulässig, wenn die Entscheidung für Abschluss oder Erfüllung eines Vertrages zwischen der betroffenen und der verantwortlichen Person erforderlich ist. Dies ist insbesondere dann der Fall, wenn der Vertragsabschluss ohne eine automatisierte Datenverarbeitung dem Verantwortlichen aufgrund eines erheblichen Arbeitsaufwandes nicht mehr zugemutet werden kann.[23] Gemäß Buchstabe b) des Art. 22 Abs. 2 DSGVO können rein automatisierte Entscheidungen dann zulässig sein, wenn sie aufgrund von Gesetzen der EU bzw. des betroffenen Mitgliedsstaates als zulässig erklärt werden. Letztlich sind nach Buchstabe c) des Art. 22 Abs. 2 DSGVO automatisierte Entscheidungen dann zulässig, wenn die betroffene Person ausdrücklich ihre Zustimmung dem Verantwortlichen gegenüber erklärt hat.

Ist eine automatisierte Entscheidung zulässig, weil sie für den Vertragsabschluss bzw. für Vertragserfüllung notwendig ist, oder weil die betroffene Person ausdrücklich ihre Einverständnis erklärt hat, hat der Verantwortliche nach Art. 22 Abs. 3 DSGVO „angemessene Maßnahmen“[24] zu treffen, um die Interessen der betroffenen Person zu schützen. Als Mindeststandard hierfür werden das Recht auf Eingreifen einer dritten Person in den Entscheidungsprozess, die Darlegung des eigenen Standpunktes sowie eine Anfechtung der Entscheidung aufgeführt.[25]

---

[21] Art. 22 Abs. 1 DSGVO
[22] Konkretisierungen zum Zulässigen Umfang von Profiling enthalten die §§ 31 ff. BDSG
[23] Vgl. Vgl. Taeger in Taeger/Gabel, 2019, S. 544.
[24] Art. 22 Abs. 3 DSGVO
[25] Vgl. Art. 22 Abs. 3 DSGVO

Der Einsatz von ADM-Systemen im Kreditvergabeprozess fällt vom Grundsatz her unter das Verbot rein automatisierter Entscheidungen nach Art. 22 Abs. 1 DSGVO. Es kommen jedoch verschiedene Ausnahmetatbestände in Betracht. Zunächst kann angeführt werden, dass sich die Anzahl gestellter Kreditanträge ohne Zuhilfenahme von ADM-Systemen nicht bewältigen lässt, und eine rein manuell getroffene Entscheidung für jeden Einzelfall daher unzumutbar wäre. Damit wäre eine automatisierte Entscheidung notwendig für die Vertragserfüllung und somit zulässig nach Art. 22 Abs. 2 DSGVO.

Die BaFin hat mit den MaRisk ihre Auslegung zu diversen gesetzlichen Anforderungen an das Risikomanagement von Banken (z.B. KWG) ausgegeben, welche für Banken von großer Relevanz ist.[26] Gemäß AT 4.3 ff. der MaRisk haben Banken entsprechende Systeme zur Identifikation, Beurteilung und Überwachung von wesentlichen Risiken einzurichten. Anhand dieser Bestimmung könnten Banken den Einsatz von Profiling-Systemen, die Kunden anhand ihrer persönlichen Daten hinsichtlich ihres Kreditausfallrisikos beurteilen, als legitimiert betrachten. Banken können weiterhin damit argumentieren, dass eine automatisierte Kreditzusage lediglich einen rechtlichen Vorteil für den Verbraucher darstellt und im Falle einer Absage die Entscheidung von Mitarbeitern manuell überprüft werden kann.

Im Zweifelsfall könnten sich die Kreditinstitute bereits bei der Antragstellung durch den Verbraucher die ausdrückliche Einwilligung für den Einsatz von ADM-Systemen erteilen lassen, bevor der Antrag weiterbearbeitet werden kann. Die Zulässigkeit von ADM-Systemen im Kreditvergabeprozess kann damit als gegeben angenommen werden.

2.2.3 Informationspflichten nach Art. 13 f. DSGVO

Um die Transparenz von automatisierten Entscheidungen zu gewährleisten, schreiben die Art. 13 f. DSGVO den Verantwortlichen vor, die Verbraucher über die „involvierte Logik“[27] der Entscheidungsautomatismen zu informieren. Über das Ausmaß dieser Informationspflicht besteht aufgrund aktuell noch fehlender Urteile bzw. Konkretisierungen noch Unklarheit. Die Offenlegung von Quellcodes dürfte davon

[26] Vgl. Hans-Böckler-Stiftung, 2014, S. 9.
[27] Art. 13 DSGVO.

nicht eingeschlossen sein, da diese i.d.R. hochkompliziert sind und für den Verbraucher daher keine Erkenntnisse bringen dürften.[28] Außerdem stellen die Systeme und deren genaue Funktionsweise u.U. Betriebsgeheimnisse dar, deren Schutz für den Verantwortlichen ebenfalls gewährleistet werden sollte.[29]

Als Präzedenzfall könnte in diesem Zusammenhang die in der Vergangenheit mehrfach hervorgebrachte Forderung an die SCHUFA dienen, diese müsse aufgrund datenschutzrechtlicher Bestimmungen ihren Quellcode für die Ermittlung des Schufa-Scores offenlegen.[30] Diesbezüglich gibt es ein Urteil des BGH vom 28. Januar 2014, in dem unmissverständlich dargelegt wird, dass die Methode der Scorewertberechnung ein Betriebsgeheimnis darstellt und daher nicht der damaligen datenschutzrechtlichen Auskunftspflicht (§ 34 Abs. 4 Satz 1 Nr. 4 BDSG a.F.) unterliegt.[31]

2.2.4 Eignung regulatorischer Vorschriften zur Zielerreichung

Artikel 22 der DSGVO soll betroffene Personen vor automatisiert getroffenen Entscheidungen schützen, die sich nachteilig auf die Person auswirken. Welche Entscheidungen dabei im Einzelfall als nachteilig i.S.d. DSGVO anzusehen sind, wird z.T. kontrovers diskutiert.[32] Im Hinblick auf den Kreditvergabeprozess ist zunächst festzuhalten, dass eine positive Kreditentscheidung der verantwortlichen Person zwar eine rechtliche Konsequenz für die betroffene Person herbeiführt, welche sich jedoch zunächst nicht negativ für diese auswirkt. Fraglich ist dabei, in wie weit eine automatisiert getroffene Entscheidung über eine Kreditkondition in Form des zu zahlenden Sollzinssatzes, z.B. anhand verschiedener Bonitätsmerkmale, eine negative Beeinträchtigung der betroffenen Person darstellt. Obwohl negative automatisierte Kreditentscheidungen die Rechtsposition der betroffenen Person zunächst nicht verändern, und damit keine negative rechtliche Konsequenz darstellen, wird die Ablehnung eines Online-Kreditantrages in ErwG 71 der DSGVO ausdrücklich als Beispiel für eine negative Beeinträchtigung der betroffenen Person aufgeführt. Die ebenfalls in Artikel 22 DSGVO ermöglichten Ausnahmetatbestände für den Ein-

[28] Vgl. Konrad-Adenauer-Stiftung e.V., 2019, S. 8f.
[29] Vgl. ErwG 63 DSGVO.
[30] Vgl. o.V., 2018.
[31] Vgl. BGH v. 28.01.2014, BGHZ 200, S. 38.
[32] Vgl. Vgl. Taeger in Taeger/Gabel, 2019, S. 541-543.

satz von ADM-Systemen bieten den verantwortlichen Personen ausreichend Möglichkeiten, um den Einsatz der Systeme regulatorisch zu legitimieren. Selbst wenn eine Notwendigkeit des Systemeinsatzes für die Vertragserfüllung nicht gegeben ist, und eine Zulässigkeit des Einsatzes sich auch aus nationalen Regelungen nicht ableiten lässt, kann der Verantwortliche sich bei entsprechender Verhandlungsmacht vor Antragstellung durch den Betroffenen dessen ausdrückliche Zustimmung zum Einsatz von ADM-Systemen geben lassen, womit eine Unzulässigkeit des Systemeinsatzes nach Artikel 22 DSGVO abschließend nicht in Frage kommt.

Hinsichtlich der Transparenzvorschriften der Artikel 13 und 14 DSGVO lässt sich festhalten, dass hierbei die Interessenlagen der verantwortlichen und der betroffenen Person gänzlich voneinander abweichen. Die Regelung in den o.g. Artikeln ist daher ein Kompromiss, der Auslegungsspielraum zu Gunsten beider Parteien zulässt. Zum aktuellen Zeitpunkt erhält der Betroffene lediglich kurz gehaltene Informationen über die Systematik der Datenverarbeitung, die über die genauen Hintergründe und Logiken keine Erkenntnisse vermittelt (s. Anhang Nr. 1 Abschnitt Nr. 6, Datenschutzinformation des S-Kreditpartners).

3 Ansätze erklärbarer KI

3.1 Kategorisierung

Wenngleich die Entwicklung erklärbarer künstlicher Intelligenz in den vergangenen Jahren an Bedeutung gewonnen hat, liegt deren Beginn dennoch genau so weit zurück, wie die Entwicklung der künstlichen Intelligenz selbst. Seither wurden zahlreiche Modelle entwickelt, die zur Erklärbarkeit automatischer Algorithmen beitragen sollen und sich z.T. deutlich in ihrer Funktionsweise unterscheiden.[33]

Sheh und Monteath argumentieren in ihrer Arbeit, dass verschiedene Anwendungsbereiche unterschiedliche Anforderungen haben und unterschiedliche Informationen benötigen, um entsprechende eingesetzte ADM-Systeme hinreichend erklären zu können.[34] Um die verschiedenen Anforderungen der einzelnen Anwendungsgebiete an die Erklärungssysteme und deren Erfüllung darstellen zu können, präsentieren sie eine Kategorisierung, welche im Folgenden näher erläutert wird.

[33] Vgl. Sheh/Monteath, 2018, S. 261.
[34] Vgl. Sheh/Monteath, 2018, S. 261.

Die erste Einteilung beschäftigt sich mit der Quelle („Source“[35]), aus der die Erklärung über das System entsteht. Diesbezüglich lassen sich „introspektive“[36] Erklärungen unterscheiden von „Post-Hoc-Rationalisation“[37]. Erklärungen, die auf einer Post-Hoc-Rationalisierung basieren, werden durch außenstehende Modelle erstellt, die das zu erklärende System von außen als Blackbox betrachten. Die Erklärung wird durch den Vergleich der vorgegebenen Daten mit dem Ergebnis der Blackbox generiert.[38] Dabei ist es möglich, dass die jeweiligen Erklärungen nur für einzelne, lokale Entscheidungen funktionieren und sich auf andere Fälle nicht anwenden lassen. Aus diesem Grund ist die Aussagekraft derartiger Erklärungen für Zwecke der rechtlichen Aufsicht oder auch der Qualitätssicherung begrenzt.[39] Introspektive Erklärungen hingegen werden von dem zu erklärenden Modell selbst generiert. Daher werden sie oftmals auch als die „wahre Erklärung“ für eine Entscheidung bezeichnet.[40] Jedes ADM-System, unabhängig von dessen Komplexität, könnte eine introspektive Erklärung bereitstellen. Bei entsprechend komplexen ADM-Systemen, bspw. neuronalen Netzen, kann die zugehörige Erklärung jedoch derart komplex sein, dass sie für den Anwender keinerlei Aussagekraft bzw. Nutzen darstellt.[41]

Eine weitere Einteilung vorhandener Erklärungsansätze wird vorgenommen anhand der Tiefe („Depth“[42]) der Erklärungen. Erklärungsmodelle, die lediglich Informationen darüber bereitstellen, auf welche Art verschiedene Attribute vom zu erklärenden Modell zum Treffen einer Entscheidung herangezogen werden, werden als Attribut-Erklärung bezeichnet.[43] Hierbei wird weiter unterschieden zwischen solchen Attribut-Erklärungen, die lediglich die verwendeten Attribute nennen, jedoch nicht erklären können (sog. „Attribute Identity explanations“[44]), und solchen, die bezüglich der Verwendung von Attributen einen logischen Zusammenhang darstellen können (sog. „Attribute Use explanations“[45]).

[35] Sheh/Monteath, 2018, S. 263.
[36] Sheh/Monteath, 2018, S. 263.
[37] Sheh/Monteath, 2018, S. 263.
[38] Vgl. Sheh/Monteath, 2018, S. 263.
[39] Vgl. Sheh/Monteath, 2018, S. 263.
[40] Vgl. Sheh/Monteath, 2018, S. 263.
[41] Vgl. Sheh/Monteath, 2018, S. 263.
[42] Sheh/Monteath, 2018, S. 263.
[43] Vgl. Sheh/Monteath, 2018, S. 263.
[44] Sheh/Monteath, 2018, S. 264.
[45] Sheh/Monteath, 2018, S. 264.

Liefert die Erklärung zusätzlich zu den verwendeten Attributen Erkenntnisse über Hintergründe der Modellbildung aus den Trainingsdaten, bezeichnet man sie als „Modell-Erklärung“[46]. Insbesondere für Zwecke der Qualitätssicherung sind solche Erklärungsansätze von Bedeutung, da sie Aufschlüsse über bestehende Fehlerpotentiale und deren Lösungsmöglichkeiten bieten.[47]

Eine abschließende Einteilung von Erklärungsansätzen erfolgt anhand der Zielsetzung, die mit der Erklärung verfolgt wird.[48] Erklärungen einzelner von ADM-Systemen getroffener Entscheidungen im Sinne einer Rechtfertigung der Entscheidung werden als „Justification“[49] bezeichnet. Soll die Erklärung hingegen Aufschlüsse über die grundsätzliche Funktionsweise des Systems geben, insbesondere um ex-ante das Verhalten des Systems bei bisher nicht aufgetretenen Einzelfällen abschätzen zu können, bezeichnen Sheh und Monteath die Erklärung als „Teaching“[50].

Im Folgenden sollen zwei verschiedene Ansätze zur Erklärung von algorithmischen Entscheidungen präsentiert werden. Dazu erfolgen stets eine Beschreibung der grundsätzlichen Funktionsweise sowie eine Einordnung in o.g. Kategorien, um die Eignung des jeweiligen Ansatzes zur Erklärung automatisiert getroffener Kreditentscheidungen beurteilen zu können.

3.2 Local Interpretable Model-Agnostic Explanations (LIME)

Local Interpretable Model-agnostic Explanations (LIME) ist ein universell einsetzbarer Erklärungsansatz, der in 2016 von Guestrin, Ribeiro und Singh in ihrer gemeinsamen Arbeit „‘Why Should I Trust You?‘ Explaining the Predictions of Any Classifier“ vorgestellt wurde.[51]

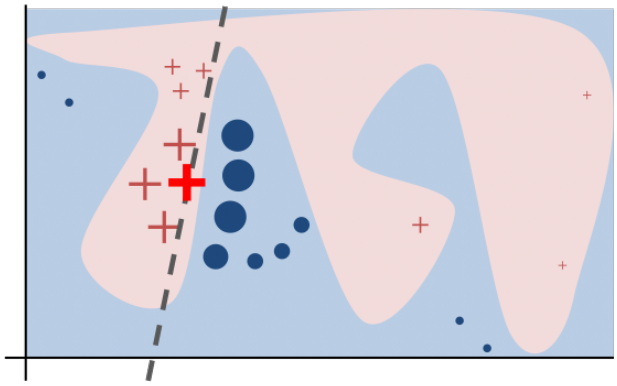
Da es sich bei LIME um einen universellen Ansatz handelt, der auf jedes zu erklärende Modell angewendet werden kann[52], ist er als Erklärung auf Basis von Post-Hoc-Rationalisierung (s. 3.1) einzuordnen. In Kapitel 3.1 wurde bereits auf die Unterscheidung von Erklärungsmodellen hinsichtlich ihrer Zielsetzung eingegangen

[46] Sheh/Monteath, 2018, S. 263.
[47] Vgl. Sheh/Monteath, 2018, S. 264.
[48] Vgl. Sheh/Monteath, 2018, S. 264.
[49] Sheh/Monteath, 2018, S. 264.
[50] Sheh/Monteath, 2018, S. 264.
[51] Vgl. Ribeiro/Singh/Guestrin, 2016, S. 1135-1144.
[52] Vgl. Ribeiro/Singh/Guestrin, 2016, S. 1136.

(Justification bzw. Teaching). Eine sinngemäße Kategorisierung wird in LIME ebenfalls vorgenommen. Unterschieden werden hierbei globale Erklärungen, die Aufschluss über die Funktionsweise des Algorithmus als Ganzes geben sollen, sowie lokale Erklärungen, welche sich auf einzelne Entscheidungen bzw. Prognosen des Algorithmus beschränken.[53]

Der Grundsatz von LIME besteht in der Annahme, dass selbst komplexe Algorithmen, deren Klassifizierung auf globaler Ebene nicht durch einfache (z.B. lineare) Modelle erklärbar ist, auf lokaler Ebene durch einfache Modelle durchaus erklärt werden können (s. Abb. 2).[54]

Abb. 2: Lineares Modell zur lokalen Erklärung komplexer Modelle



Quelle: Ribeiro/Singh/Guestrin, 2016, S. 1139.

Um eine bestimmte Prognose bzw. Klassifizierung zu erklären, werden zunächst die Ausgangsdaten, anhand derer das Modell die Prognose bzw. die Klassifizierung vorgenommen hat, zufällig verändert. Anschließend wird das Ausmaß der Veränderung gemessen (similarity score)[55]. Das Prognosemodell erstellt daraufhin anhand der veränderten Ausgangsdaten jeweils neue Prognosen. Die erstellten Prognosen verwendet LIME, um diejenigen erklärenden Variablen herauszustellen, die auf lokaler Ebene den größten Einfluss auf das Prognoseergebnis zeigen. Dabei wird die Anzahl der vorhandenen erklärenden Variablen erheblich reduziert. Auf dieser Grundlage wird ein einfaches (z.B. lineares) Erklärungsmodell gebildet, wobei die einzelnen gebildeten Datenpunkte (Samples) anhand der ermittelten similarity

[53] Vgl. Ribeiro/Singh/Guestrin, 2016, S. 1138.
[54] Vgl. Ribeiro/Singh/Guestrin, 2016, S. 1138.
[55] Vgl. Ribeiro/Singh/Guestrin, 2016, S. 1138.

scores (s.o.) gewichtet werden.[56] Die Gewichtung der Samples wird in Abb. 2 durch die Größe der Datenpunkte (Kreuze bzw. Kreise) symbolisiert.

Aufbauend auf LIME präsentieren Guestrin, Ribeiro und Singh einen weiteren Ansatz „SP-LIME“[57], mit dessen Hilfe ein globales Verständnis über zugrundeliegende Algorithmen erlangt werden kann.[58] Die Grundlage hierbei besteht darin, viele sinnvoll ausgewählte einzelne Erklärungen, welche möglichst viele Prognosen bzw. Klassifizierungen umfassen, als Basis einer globalen Erklärung zusammenzufassen.[59]

3.3 Transparent Generalized Additive Model Tree (TGAMT)

Ein weiterer Ansatz zur Erklärung automatisiert getroffener Kreditentscheidungen wurde von der Fair Isaac Corporation (FICO) entwickelt. Ziel dieses Ansatzes ist es, vorhandenes Fachwissen im Kreditbereich mit Hilfe von leistungsstarken Algorithmen zu verbinden, um ein System zu schaffen, welches eine gute Prognoseperformance zeigt, genügend Flexibilität bietet, um auf zukünftige Entwicklungen reagieren zu können, sowie stets einfach erklärt werden kann.[60]

Die Grundlage für den Ansatz der FICO ist die Scorecard-Technologie[61]. Dieses Verfahren ist in der Finanzbranche weit verbreitet und wird für unterschiedliche Anwendungsbereiche (z.B. Marketing, Betrugsprävention sowie Kreditentscheidungen) angewandt.[62] Bei Anwendung von Scorecards zur Lösung von Entscheidungs- oder Prognoseproblematiken werden zunächst die erklärenden Variablen (sog. „Merkmale“[63]) sowie deren mögliche Ausprägungen (sog. „Attribute“[64]) definiert. Anschließend wird jedem Attribut eine Punktzahl zugewiesen. Dabei wird stets berücksichtigt, als wie wünschenswert die jeweilige Merkmalsausprägung angesehen wird. Die Zuordnung von Punkten ist nach einem erklärbaren Muster vorzunehmen.[65]

---

[56] Vgl. Ribeiro/Singh/Guestrin, 2016, S. 1139.
[57] Ribeiro/Singh/Guestrin, 2016, S. 1136.
[58] Vgl. Ribeiro/Singh/Guestrin, 2016, S. 1138.
[59] Vgl. Ribeiro/Singh/Guestrin, 2016, S. 1140-42.
[60] Vgl. Fahner, 2018, S. 7.
[61] Vgl. Fahner, 2018, S. 7.
[62] Vgl. Fahner, 2018, S. 7.
[63] Vgl. Fahner, 2018, S. 7.
[64] Vgl. Fahner, 2018, S. 7.
[65] Vgl. Fahner, 2018, S. 8.

Abschließend werden die Punktzahlen der betroffenen Person über alle Merkmale hinweg addiert und zu einer Gesamtpunktzahl zusammengefasst.[66] Dabei erfolgt außerdem eine Gewichtung der einzelnen Merkmale. Durch die Gewichtung kann der Ersteller der Scorecard Beschränkungen vornehmen, um z.B. sicherzustellen, dass rechtliche Vorschriften im Scorecard-Modell berücksichtigt werden.[67]

Die reine Summierung der einzelnen Punktwerte birgt die Problematik, dass Wechselwirkungen zwischen den einzelnen Merkmalen nicht berücksichtigt werden können. Liegen solche Wechselwirkungen vor, führt dies zu Verzerrungen im Modell.[68] Dieser Problematik kann begegnet werden durch die Anwendung segmentierter Scorecards. Dabei werden für unterschiedliche Zielgruppen der Antragsteller unterschiedliche Scorecards verwendet.[69] Die Herausforderung, die sich den Entwicklern segmentierter Scorecards stellt, ist die Wahl eines geeigneten Segmentierungsschemas. Die FICO verwendet zur Bestimmung eines „optimalen" Segmentierungsschemas eine rekursive Suchheuristik.[70] Dabei teilt ein Algorithmus die Grundgesamtheit der Trainingsdaten anhand einer Vielzahl verschiedener Kriterien in zwei Teile, passt jeweils entsprechende Scorecards an die Segmente an und misst den Leistungszuwachs, der durch die Segmentierung erreicht wurde. Die Aufteilung, die den größten Leistungszuwachs erreicht, wird letztlich beibehalten. Dieser Vorgang wird so lange wiederholt, bis es keine weitere Aufteilung der Daten mehr gibt, die einen ausreichend großen Leistungszuwachs generiert. Schritt für Schritt entsteht auf diese Weise ein baumartiges Segmentierungsschema mit jeweils passenden Scorecards.[71]

Die Erklärung automatisiert getroffener Kreditentscheidungen nach oben beschriebenem Modell entsteht durch die klar definierten Merkmale, Attribute sowie Punktzahlen derselben, und damit aus dem Modell selbst, weshalb die Erklärung als introspektiv eingeordnet werden kann. Dabei umfasst die Erklärung sowohl die einzelnen prognoserelevanten Merkmale sowie durch die bekannten Punktzahlen auch deren Auswirkung, was auf eine attribute-use-explanation schließen lässt. Ziel des

[66] Vgl. Fahner, 2018, S. 8.
[67] Vgl. Fahner, 2018, S. 9.
[68] Vgl. Fahner, 2018, S. 9.
[69] Vgl. Fahner, 2018, S. 9.
[70] Vgl. Fahner, 2018, S. 10-11.
[71] Vgl. Fahner, 2018, S. 10-11.

Erklärungsansatzes ist sowohl die Erklärung des Einzelfalles als auch die Beurteilung der Funktionsweise des Modells als Ganzes. Das zu erwartende Ergebnis bei bisher noch nicht vorgekommenen Datenkonstellationen kann ohne Weiteres bestimmt werden. Damit kann als Ziel sowohl die Kategorie Justification als auch Teaching ausgemacht werden.

4 Ansatz auf Basis emergenter Statistik

In folgendem Kapitel wird ein Ansatz erklärbarer künstlicher Intelligenz vorgestellt, mit dessen Hilfe ein vollständiges Profiling von Kreditnehmern und Antragstellern vorgenommen werden kann. Zunächst werden hierzu die Grundlagen der sog. emergenten Statistik erläutert, ehe die Ergebnisse des Profilingmodells dargestellt werden.

4.1 Grundlagen emergenzbasierter Statistik

4.1.1 Motivation zur Suche nach emergenten empirischen Gesetzen

„Wirtschaftswissenschaft: das einzige Fach, in dem jedes Jahr auf dieselben Fragen andere Antworten richtig sind.“

- Danny Kaye[72]

Verlangt man von einer empirischen Wissenschaft, dass sie Gesetzmäßigkeiten vorweisen kann, welche in der Vergangenheit immer zutreffend waren, so müsste man den Wirtschaftswissenschaften unterstellen, keine echte empirische Wissenschaft zu sein.[73] Kaye deutet in seiner Aussage treffend die Problematik an, dass Auffassungen, Vorgehensweisen und andere Regeln im Bereich der Wirtschaftswissenschaften von einer kontinuierlichen Schwankung geprägt sind. Außerdem können Vermutungen in den Wirtschaftswissenschaften, im Gegensatz zu den Naturwissenschaften, nicht durch Experimente im Labor, sondern lediglich anhand von Beobachtungen in der Realität überprüft werden.[74] Dies wirft die Frage auf, ob es für wirtschaftswissenschaftliche Problemstellungen überhaupt möglich ist, eindeutig überprüfbare empirische Gesetze zu finden.[75]

---

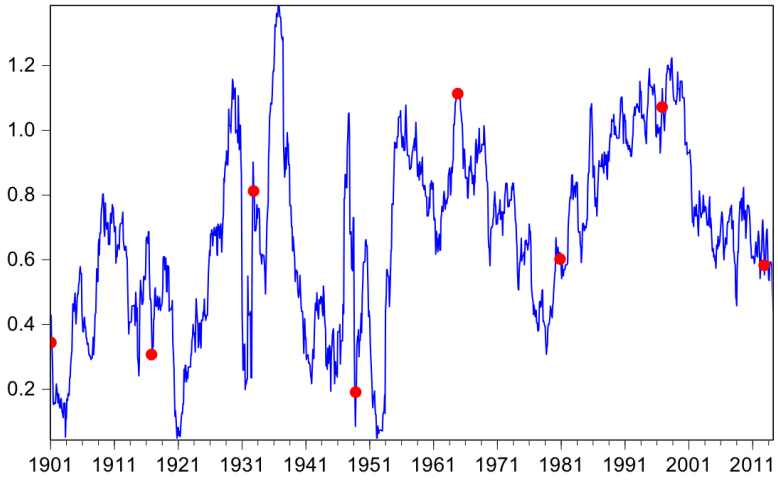
[72] Kaye, 1976, S. 41.
[73] Vgl. Kuck/Harries/Kuck, 2015, S. 3.
[74] Vgl. o.V., 2010, S. 53.
[75] Vgl. Kuck/Harries/Kuck, 2015, S. 3.

4.1.2 Einführung

Wie Kuck in seiner Arbeit darlegt, ist es sehr wohl möglich, eindeutig überprüfbare Gesetze selbst in Themenfeldern zu finden, in denen man sie kaum vermuten würde. Als Beispiel hierfür führt er unter anderem die Kapitalmärkte an.[76]

Beispiele zeigen plastisch, dass es in einer Vielzahl von Problemstellungen möglich ist, Aussagen zu treffen, die bisher immer wahr gewesen sind, sobald man keine einzelnen Beobachtungen, sondern Gruppen von Beobachtungen betrachtet. So hätte beispielsweise ein Investor, der immer dann für einen Monat in den S&P 500 Index investiert, wenn dieser im Vormonat eine positive Rendite aufgewiesen hat, nach 16 Jahren bisher immer eine positive Gesamtrendite erzielt. Entscheidend ist, dass diese Gesetzmäßigkeit erst ab einer Anzahl von 192 Monaten (16 Jahren) auftritt, nicht jedoch für die Betrachtung kürzerer Perioden. Dieses Gesetz wurde bis zum Jahr 2011 siebenmal bestätigt (s. Abb. 3).[77]

Abb. 3: Rendite der Anlagestrategie seit 1901



Quelle: Kuck/Harries/Kuck, 2015, S. 4.

Die Kenntnis solcher Gesetzmäßigkeiten aus Beobachtungen in der Vergangenheit kann für Prognosen in der Zukunft verwendet werden. Was in der Vergangenheit immer so war, neigt dazu, auch in der Zukunft wieder so zu sein.[78] Es ist jedoch ausdrücklich festzuhalten, dass nicht davon ausgegangen werden kann, dass ein einmal gefundenes emergentes Gesetz bis in alle Ewigkeit gültig sein wird. Ist ein

[76] Vgl. Kuck/Harries/Kuck, 2015, S. 3.
[77] Vgl. Kuck/Harries/Kuck, 2015, S. 4.
[78] Vgl. Kuck/Harries/Kuck, 2015, S. 27.

Gesetz einmal falsifiziert worden, gilt es für alle Zeit als falsifiziert und kann nie wieder wahr werden.[79] Bestenfalls kann das Gesetz in einer größeren Gruppe von Beobachtungen bestehen bleiben.

4.1.2 Terminologie

Im Folgenden werden einige Begrifflichkeiten erläutert, die für die Suche, Beschreibung und Anwendung emergenter Gesetze notwendig sind.

Der Begriff Emergenz im allgemeinen Sprachgebrauch bedeutet, dass die Eigenschaften einer Gesamtheit nicht aus den Eigenschaften ihrer Einzelteile abgeleitet werden können. Im Kontext dieser Arbeit bedeutet der Begriff der Emergenz, dass Vorhersagen über Problemstellungen und deren Gesetzmäßigkeit erst in einem autonomen Suchprozess und durch Betrachtung von Sequenzen möglich sind.[80] Die Anzahl der Beobachtungen, nach denen zum ersten Mal eine Gesetzmäßigkeit emergiert, wird als Emergenzmenge (T) bezeichnet. Davon ist die universelle Emergenzmenge (TU) zu unterscheiden. TU wird erreicht, sobald es keine größere Sequenz mehr gibt, für die das gefundene Gesetz nicht gilt.[81]

Mit dem „Degree of inductive Verification" (DiV) wird ausgedrückt, wie oft ein gefundenes Gesetz seit seiner ersten Beobachtung bisher bestätigt ist.[82] Die Metrik ist definiert als:

$$DiV = \frac{Anzahl\ der\ Beobachtungen}{T} - 1 \qquad (1)$$

Ein DiV von 23,3 bedeutet demnach, dass das Gesetz seit seinem ersten Auftreten bereits 23-mal bestätigt wurde, und es noch 0,7xT Beobachtungen benötigt, bis es sich ein weiteres Mal bestätigen kann.

4.1.3 Meta-Gesetze und Reliability

In Kapitel 4.1.1 wurde bereits auf den Induktionsschluss hingewiesen, der anhand emergenter Gesetzmäßigkeiten getroffen werden kann. Von Zusammenhängen, die in der Vergangenheit immer so waren, kann man erwarten, dass sie auch in der Zukunft wieder so sein werden. Wie bereits erwähnt wurde, wird nicht behauptet,

[79] Vgl. Kuck/Frischhut, 2015, S. 7.
[80] Kuck/Harries/Kuck, 2015, S. 8.
[81] Vgl. Kuck/Harries/Kuck, 2015, S. 32.
[82] Vgl. Kuck/Harries/Kuck, 2015, S. 32 – 33.

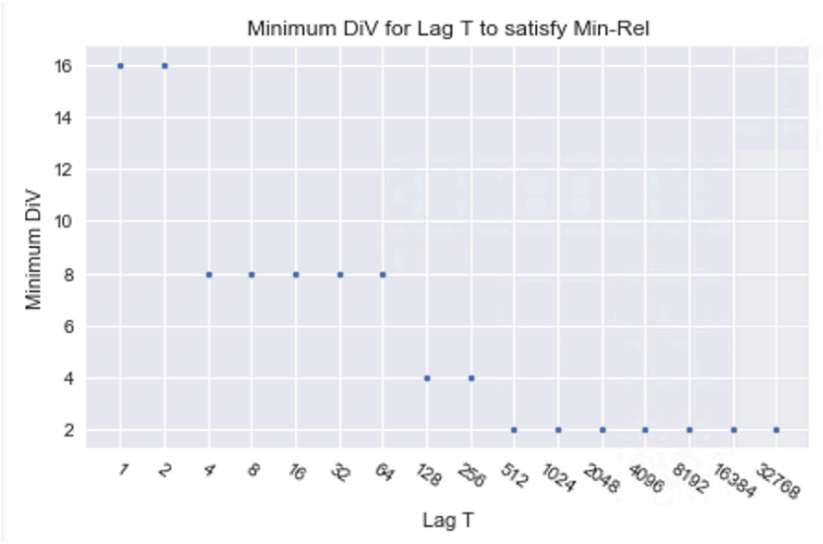
dass der Induktionsschluss aus emergenten Gesetzen für zukünftige Beobachtungen immer zutreffen wird, da Gesetze im Laufe der Zeit falsifiziert werden können. Dennoch ist der Induktionsschluss nur logisch, da seine Aussage bei einem emergenten Gesetz in der Vergangenheit immer richtig gewesen wäre.[83]

Neben der logischen Rechtfertigung des Induktionsschlusses kann außerdem eine empirische Betrachtung Aufschlüsse geben. Betrachtet man verschiedene Gesetze mit jeweils verschiedenem Bestätigungsgrad (DiV) stellt man zunächst fest, dass die Anzahl gefundener Gesetzeshypothesen mit steigendem DiV abnimmt.[84] Betrachtet man nun die Prognosen verschiedener Gesetze, die auf dem Induktionsschluss basieren, lässt sich feststellen, dass Prognosen von Gesetzen mit hohem DiV einen höheren Anteil von Bestätigungen aufweisen als Gesetze mit niedrigerem DiV.[85]

Über diesen Zusammenhang lassen sich wiederum emergente Gesetzmäßigkeiten finden. Dabei hat sich gezeigt, dass Gesetze über die Mindest-Reliability verschiedener Kombinationen von T und DiV existieren. Solche Gesetze werden als „Meta-Gesetze“ bezeichnet.[86] Durch besagte Meta-Gesetze lässt sich eine Matrix aufstellen, wie oft ein Gesetz der Emergenzmenge T in der Vergangenheit bestätigt sein muss (DiV), damit der Induktionsschluss auf Basis dieses Gesetzes einen Mindestanteil bestätigter Prognosen (Reliability) aufweist. Abbildung 4 zeigt Kombinationen von T und DiV, zu denen eine Mindest-Reliability von 0,8 gegeben ist. Gesetze der Länge T=4 müssen demnach achtmal bestätigt werden, um eine Reliability von 0,8 zu erreichen, während ein Gesetz der Länge T=512 lediglich zweimal bestätigt sein muss.

[83] Vgl. Kuck/Frischhut, 2015, S. 10.
[84] Vgl. Kuck/Frischhut, 2015, S. 9.
[85] Vgl. Kuck/Frischhut, 2015, S. 8.
[86] Vgl. Kuck, 2019.

Abb. 4: Emergenzmenge und DiV für Reliability 0.8



Quelle: Methodenserver des ZES

4.1.4 KnowledgeNets

Eine Anwendungsmöglichkeit emergenter Statistik (und gleichzeitig ein erster Schritt zur Entwicklung emergenzbasierter Prognosemodelle) sind sogenannte „KnowledgeNets“[87]. Zur Erstellung eines KnowledgeNets wird der zugrunde gelegte Datensatz in mehreren Iterationen auf Gruppen von Beobachtungen („Objekte“[88]) untersucht, die in Bezug auf eine beliebige Zielvariable oder eine bestimmte Zielvariable bislang gesetzmäßig immer unterschiedliche Mittelwerte aufgewiesen haben.[89] Damit eignen sich KnowledgeNets insbesondere zur Ursachenforschung und zum Erlangen von Erkenntnissen über bestimmte Sachverhalte durch die bloße Analyse der beobachteten Daten. Abbildung 5 zeigt exemplarisch ein Objekt eines KnowledgeNet.

Abb. 5: Beispiel für Objekte in KnowledgeNets

Object	Rel	Mean	Mean (oos)	TU (Mean)	DiV (Mean)
+ (JobLevel==1)&(OverTime==&...	0.8	0.5824	0.5	64	17.375
(JobLevel==1)&(OverTime=='Yes')&~(TrainingTimesLastYear==3)			0.4118	256	3.5938

Quelle: Methodenserver des ZES

Das KnowledgeNet in Abbildung 5 betrifft die Kündigungsrate von Mitarbeitern in einem Unternehmen. Das markierte Objekt zeigt die Auswahlregel der Mitarbeiter,

[87] Vgl. Kuck/Kratz/Frischhut, 2018.
[88] Vgl. Kuck/Kratz/Frischhut, 2018.
[89] Vgl. Kuck/Kratz/Frischhut, 2018.

welche im Mittel die höchste Kündigungsrate aufweisen. Dabei handelt es sich um Mitarbeiter, die sich in der untersten Hierarchiestufe befinden (Joblevel==1), die außerdem Überstunden machen (Overtime==‘Yes‘) und im vergangenen Jahr weniger als drei Fortbildungsstunden absolviert haben (~TrainingTimesLastYear==3).

4.1.5 Modellbildungsprozess

Primäres Ziel der Modellbildung auf Basis emergenter Statistik ist zunächst nicht die Prognose der endogenen Variable selbst, sondern die rekursive Suche nach Schätzheuristiken, die die Gesamtprognose bisher immer verbessert hätten.[90] Hierfür kommen z.B. Null-Schätzer (Prognose, dass die endogene Variable immer gleich null ist), Mittelwert-Schätzer (Prognose der endogenen Variable anhand ihres rollierenden Mittelwerts) oder kleinste-Quadrate-Schätzer (Prognose anhand der Methode der kleinsten quadratischen Abweichung) zum Einsatz. Die abschließende Prognose der endogenen Variablen ergibt sich aus der Summe der einzelnen gefundenen Heuristiken.

Werden dabei ausschließlich solche Heuristiken gesucht, die auf den gesamten Datensatz angewendet werden und die Prognose bisher immer verbessert hätten, so spricht man von einem Null-Modell. Unter Umständen gibt es Heuristiken, die die Gesamtschätzung ebenfalls weiter verbessern, jedoch nur dann, wenn sie lediglich für bestimmte Objekte eines zuvor erstellten KnowledgeNets angewendet werden. In vollständigen Modellen werden somit auch gewonnene Erkenntnisse aus KnowledgeNets in die Prognose mit einbezogen.[91]

Zur Beurteilung der Vorteilhaftigkeit einzelner Heuristiken im relativen Bezug zueinander werden verschiedene Fehlermetriken eingesetzt (z.B. mittlerer absoluter Prognosefehler, Korrelation und Bestimmtheitsmaß). Eine Heuristik, die eine andere Heuristik bzw. die Gesamtschätzung hinsichtlich einer bestimmten Metrik gesetzmäßig immer verbessert hat, wird als T-dominant bezeichnet. War eine Heuristik dabei für alle eingesetzten Metriken bisher immer besser, wird diese als universell T-dominant bezeichnet.[92] Abb. 6 zeigt exemplarisch Fehlermetriken für die Schätzung von Tordifferenzen bei Fußballspielen anhand einer kleinste-Quadrate-Heuristik im Vergleich zu einer Schätzung anhand des expandierenden Mittelwerts.

---

[90] Vgl. Kuck/Kratz/Frischhut, 2018.
[91] Vgl. Kuck/Kratz/Frischhut, 2018.
[92] Vgl. Kuck/Harries/Kuck, 2015, S. 51.

Abb. 6: Fehlermetriken zur Beurteilung von Heuristiken

	Metric	Value	Law	T	Div
0	Bias	0.002375	Bias always in same direction after T=	None	None
1	mean absolute prediction error	1.221265	mean always lower as bench after T=	512	27.3184
2	mean squared prediction error	2.459328	mean always lower as bench after T=	256	55.6367
3	root of mean squared prediction error	1.568224	mean always lower as bench after T=	256	55.6367
4	Correlation	0.452128	mean always higher as bench after T=	64	225.547
5	Coefficient of determination	0.204420	mean always higher as bench after T=	256	55.6367
6	mean for yd>0	0.772621	mean for yd>0 always >0 after T=	64	159.609
7	mean for yd<0	-0.513385	mean for yd<0 always <0 after T=	128	31.9844
8	Slope of regression y=f(yd)	1.029039	Slop y=f(yd) always >0 after T=	128	112.273
9	NaN	NaN	mean(y,t,T) always increasing after T=	2048	6.07959

Quelle: Methodenserver des ZES

4.2 Scoring-Modell für Kreditanträge

Im Folgenden wird dargelegt, wie sich auf Basis emergenter Statistik ein Kreditscoring-Modell erstellen lässt, welches sowohl eine hohe Prognosequalität aufweist als auch klar verständliche Erklärungen vorweisen kann.

4.2.1 Datenherkunft

Der zur Modellbildung verwendete Datensatz stammt von der FICO, welche eigens einen Forschungswettbewerb zur Erstellung erklärbarer automatisierter Kreditanalysen durchgeführt hat.[93] Er enthält ca. 10.400 einzelne Baufinanzierungen mit den jeweiligen Antragsdaten sowie einer Variable „RiskPerformance“, die für den jeweiligen Kredit angibt, ob dieser ausgefallen (RiskPerformance = „Bad“) oder nicht ausgefallen (RiskPerformance = „Good“) ist. Insgesamt enthält der Datensatz 24 Features. Der vollständige Datensatz sowie eine Beschreibung der einzelnen enthaltenen Features können in den Anhängen 2 und 3 eingesehen werden.

4.2.2 Modellbildungsprozess

Für die Suche nach emergenten Gesetzmäßigkeiten wurde zunächst eine Mindestprognosegüte (Reliability) von 0,85 bestimmt. Außerdem wurden die Ausprägungen des Merkmals RiskPerformance faktorisiert (Good = 0, Bad = 1), um ein berechenbares Skalenniveau sicherzustellen.

[93] Vgl. Fair Isaac Corporation, 2019.

In einem ersten Schritt wird ein KnowledgeNet erzeugt, welches Gruppen von Kreditanträgen identifiziert und sortiert, die in Bezug auf die Rate von Kreditausfällen der jeweiligen Gruppen immer unterschiedliche Werte aufgewiesen haben. Dies dient lediglich dazu, im folgenden Modellbildungsprozess eine möglichst gute Prognoseperformance zu erreichen. Aufbauend auf dieses KnowledgeNet erfolgt die Bildung eines Modells zur Prognose der Ausfallwahrscheinlichkeit (RiskPerformance). Die prognostizierte Ausfallwahrscheinlichkeit wird bezeichnet als „yd_ges_RiskPerformance“ und könnte zum besseren Verständnis in der Kommunikation nach außen auch als „Score“ bezeichnet werden. Anschließend wird ein weiteres KnowledgeNet aufgebaut. Diesmal erfolgt die Identifikation einzelner Objekte des Datensatzes lediglich anhand der prognostizierten Ausfallwahrscheinlichkeiten (yd_ges_RiskPerformance), die gesetzmäßig immer unterschiedliche tatsächliche Ausfallraten (RiskPerformance) aufgewiesen haben. Die herausgestellten Objekte sind dabei disjunkt, d.h., dass ein Kreditnehmer auf Basis der für ihn prognostizierten Ausfallwahrscheinlichkeit (yd_ges_RiskPerformance) genau einem Objekt zugeordnet werden kann. Dieses KnowledgeNet könnte nach außen als „Ratingklassen“ bezeichnet werden.

Bei der Modellbildung zur Schätzung der Ausfallwahrscheinlichkeit einzelner Kreditanträge wurden 11 Heuristiken gefunden, welche die Gesamtprognose bisher immer verbessert hätten. Alle dieser Heuristiken werden angewendet auf sämtliche Antragsteller, weshalb man das hier zugrunde gelegte Modell als Null-Modell bezeichnet. Ein vollständiges Modell unter Zuhilfenahme des zuvor erstellten KnowledgeNets und mit einer geringeren Reliability (z.B. i.H.v. 0,8) würde u.U. mehrere Heuristiken finden.

Anhand der Trainingsdaten konnte das Modell eine Korrelation von 0,4872 erreichen, während in der Evaluation eine leicht höhere Korrelation von 0,5186 erreicht wurde.

Das darauf aufbauende KnowledgeNet (die Ratingklassen) umfasst 12 Auswahlregeln von Kreditanträgen, die bisher immer unterschiedliche tatsächliche Ausfallraten aufgewiesen haben.

4.2.3 Erklärbarkeit der Ergebnisse

Die Berechnung von Scores anhand von Merkmalsausprägungen zur Darstellung prognostizierter Ausfallwahrscheinlichkeiten sowie die darauf aufbauende Eingruppierung in ein System von Ratingklassen ist eine gängige Vorgehensweise und in der Bankpraxis weit verbreitet.[94] Die Besonderheit der Vorgehensweise auf Basis emergenter Gesetze sind u.a. die Art der Berechnung sowie die Möglichkeit, die Klassifizierung für sämtliche denkbaren Adressaten (Antragsteller, Aufsichtsbehörden, Qualitätssicherung) in einfacher Form erklären zu können.

Eine globale Erklärung des Modells bzw. der Klassifizierung, die z.B. für eine Revision seitens Aufsichtsbehörden oder zu Zwecken der Qualitätssicherung von Bedeutung ist, ergibt sich durch die reine Betrachtung des Modells bzw. der Klassifizierung unter Berücksichtigung der in 4.1 erläuterten Grundlagen der emergenten Statistik. Die Betrachtung der Schätzheuristiken erklärt vollständig das Zustandekommen der zu ermittelnden Scores. Dabei lässt sich jedes relevante Merkmal und die damit vorgenommenen Berechnungen nachvollziehen. Abbildung 7 zeigt exemplarisch eine Schätzheuristik des Kreditscoring-Modells.

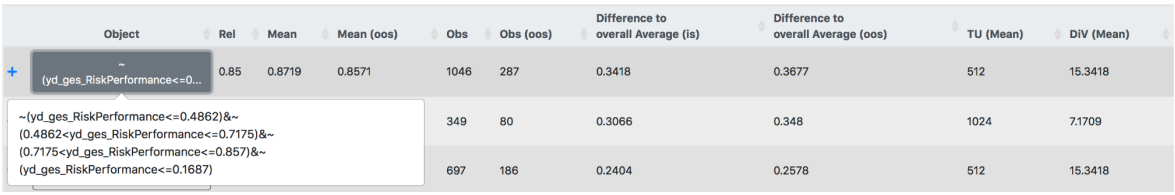
Abb. 7: Exemplarische Schätzheuristik

Quelle: Methodenserver des ZES

Die markierte Schätzheuristik korrigiert die Gesamtprognose um einen Wert in Abhängigkeit der Differenz aus dem Anteil offener Kredite mit Restschuld („PercentTradesWBalance“) und der externen Risikoeinschätzung („ExternalRiskEstimate“). Diese Korrektur hat die Gesamtprognose bisher in jeder Sequenz von 2789 Krediten verbessert, was bislang dreimal bestätigt wurde. Die Prognose, dass die Heuristik die Schätzung auch in den nächsten 2789 Krediten verbessern wird, gehört zu einer Klasse von Prognosen, die bislang immer zu mindestens 85% richtig gewesen sind.

[94] Vgl. Wingendorf in Brunner/Seeger/Turturica, 2010, S. 129.

Abb. 8: Exemplarisches Objekt (Ratingklasse)



	Object	Rel	Mean	Mean (oos)	Obs	Obs (oos)	Difference to overall Average (is)	Difference to overall Average (oos)	TU (Mean)	DiV (Mean)
+	~ (yd_ges_RiskPerformance<=0...	0.85	0.8719	0.8571	1046	287	0.3418	0.3677	512	15.3418
	~(yd_ges_RiskPerformance<=0.4862)&~ (0.4862<yd_ges_RiskPerformance<=0.7175)&~ (0.7175<yd_ges_RiskPerformance<=0.857)&~ (yd_ges_RiskPerformance<=0.1687)				349	80	0.3066	0.348	1024	7.1709
					697	186	0.2404	0.2578	512	15.3418

Quelle: Methodenserver des ZES

Abbildung 8 zeigt ein Objekt des KnowledgeNets (Ratingklasse). Kreditanträge mit einem Score von über 0,857 haben bislang in Sequenzen von TU = 512 Krediten im Durchschnitt immer höhere tatsächliche Ausfallraten (durchschnittlich 87,19% innerhalb der Trainingsdaten bzw. 85,71% innerhalb der Evaluierungsdaten) aufgewiesen als alle anderen Ratingklassen. Diese Gesetzmäßigkeit hat sich bislang DiV = 15,3418 Mal bestätigt. Damit gehört die Prognose, dass Kredite dieser Ratingklasse auch zukünftig die höchsten Ausfallraten aufweisen werden zu einer Klasse von Prognosen, die bislang immer zu mindestens 85% richtig waren.

Neben globalen Erklärungen zur Funktionsweise des Modells bzw. der Klassifizierung als solches sind ebenfalls Erklärungen einzelner Ergebnisse möglich. Um die Ermittlung des Scores vollständig erklären zu können, bedarf es einer schrittweisen Erläuterung der einzelnen Heuristiken des Prognosemodells unter Bezugnahme der jeweiligen Merkmalsausprägungen des betroffenen Antrags. Nachdem der ermittelte Score des Antrags durch die einzelnen Schätzheuristiken erklärt wurde, kann die Erklärung der Klassifizierung über das KnowledgeNet erfolgen. Nachfolgend wird auszugsweise eine exemplarische verbale Erklärung der Ergebnisse für einen Antragsteller präsentiert.

1.: „Weil Ihre externe Risikoeinschätzung („External RiskEstimate“, Variable x_1) einen Wert von 70 aufweist, Ihre Verschuldungsquote 40% beträgt („PercentTradesWBalance“, Variable x_2), die Verwendung der Schätzfunktion:

$$0{,}6834 - 0{,}1766 * \left(\frac{x_1}{22{,}01} - \frac{x_2}{28{,}24}\right) \qquad (2)$$

bisher in jeder Sequenz von 2789 Anträgen zu einer Verbesserung der Gesamtschätzung geführt hat und die Prognose, dass diese Schätzung auch in der Zukunft zu einer Verbesserung der Prognose führen wird zu einer Klasse von Prognosen gehört, die bisher immer in mindestens 85% der Fälle korrekt war, wird Ihr Score auf 0,37189 gesetzt.“

2.: Da Ihr Score nach Addition sämtlicher Heuristiken insgesamt 1,0 beträgt, gehören Sie zu der Ratingklasse, welche in Sequenzen von 512 Kreditanträgen bisher immer die höchsten tatsächlichen Kreditausfallraten aufgewiesen hat. Dieses Gesetz ist bisher 15,3418 Mal bestätigt worden und gehört somit ebenfalls zu einer Klasse von Prognosen, welche bislang immer zu mindestens 85% zugetroffen hat.

Sowohl globale Erklärungen als auch Erklärungen für einzelne Beobachtungen entstehen aus dem Modellbildungsprozess selbst heraus, weshalb sie hinsichtlich der Quelle (s. 3.1) als introspektive Erklärung einzustufen sind. Bezüglich der Tiefe der Erklärung lässt sich festhalten, dass die Erklärungen sowohl die Identität und Verwendung der Attribute als auch das Zustandekommen des Modells selbst umfassen. Hinsichtlich der Zielsetzung der Erklärungen auf Basis emergenter Statistik kann sowohl die lokale Einzelfallerklärung („Justification“) als auch eine globale Erklärung des Systems als Ganzes („Teaching“) verfolgt werden.

5 Fazit und Ausblick

Die Bedeutung von künstlicher Intelligenz in der Finanzbranche sowie im Bankenumfeld nimmt stetig zu. Bereits heute sind verschiedene Anwendungen künstlicher Intelligenz in Bereichen wie Marketing und Kundenkommunikation, Datenanalyse oder Kredit- und Anlagegeschäft im Einsatz. Experten sehen dabei noch große vorhandene Potentiale. Künstliche Intelligenz kann dabei zur Reduzierung von Kosten und der Erschließung neuer Ertragsfelder beitragen.[95]

In der Vergangenheit wurde jedoch durch verschiedene Beispiele regelmäßig deutlich, dass der Einsatz von künstlicher Intelligenz diverse Risiken mit sich bringt, weswegen die Ergebnisse automatischer Entscheidungsalgorithmen mit der notwendigen Sorgfalt und Vorsicht zu bewerten sind. Diesem Umstand ist sich die Gesetzgebung bewusst und hat sich mit den Regelungen der Datenschutzgrundverordnung unter anderem zum Ziel gesetzt, den besagten Risiken entgegenzuwirken. Dabei ist für den Gesetzgeber jedoch neben dem Schutz betroffener Personen (Schutz bei der Verarbeitung personenbezogener Daten) auch der Schutz der verantwortlichen Personen (insbesondere Schutz von Betriebsgeheimnissen) von Bedeutung. Bei den getroffenen Regelungen der DSGVO (insbesondere Art. 13 ff.

[95] Vgl. Schneider, 2019, S. 4.

sowie Art. 22 DSGVO) handelt es sich daher um Kompromisslösungen, die Auslegungsspielraum sowohl zu Gunsten der betroffenen als auch zu Gunsten der verantwortlichen Personen zulassen. Um das Vertrauen der betroffenen Personen in ein automatisiertes Entscheidungssystem zu stärken sowie eine Qualitätssicherung durch externe Aufsichtsbehörden zu ermöglichen, bedarf es einer größtmöglichen Transparenz bei der Entscheidungsfindung. Die Erklärung muss dabei für Außenstehende nachvollziehbar sein und gleichzeitig das Betriebsgeheimnis des Verantwortlichen schützen.

In Kapitel drei der vorliegenden Arbeit wurden zwei bekannte Erklärungsansätze präsentiert, welche zur Erklärung von automatisiert getroffenen Kreditentscheidungen eingesetzt werden können.

Bei dem ersten vorgestellten Ansatz (LIME) handelt es sich um eine universelle Methode, mit der verschiedene Modelle, unabhängig von ihrer Komplexität, einfach erklärt werden können. Da die Erklärung hierbei durch eine Betrachtung der Blackbox von außen entsteht, können lediglich näherungsweise Erkenntnisse gewonnen werden. Das System selbst bleibt opaque. Daher sind Erklärungen auf Basis von LIME nur sehr bedingt zur Qualitätssicherung und zur Erklärung von Kreditentscheidungen geeignet.

Beim Ansatz der FICO (TGAMT) hingegen stammt die Erklärung der getroffenen Entscheidungen aus dem Entscheidungssystem selbst. Die Anwendung von Scorecards mit einem nachvollziehbaren Punktsystem macht die Entscheidung für Außenstehende leicht erklärbar. Außerdem können die Auswirkungen bislang unbekannter Konstellationen auf die Entscheidung des Modells leicht abgeschätzt werden. Wenngleich die Scorecards und deren Segmentierung durch rekursive Algorithmen angepasst und optimiert werden, sind diese dennoch in hohem Maße von vorhandenem Fachwissen der Entwickler im Bereich Kreditgeschäft beeinflusst. Dies kann sich für das System sowohl vor- als auch nachteilig auswirken. Die Vorgehensweise impliziert, dass über das Kreditgeschäft bereits ein umfassendes Wissen vorhanden ist, welches für die Entscheidungsfindung wertvoller ist als mögliche neue Erkenntnisse, die aus der Analyse der gesammelten Daten gewonnen werden könnten. Diese Implikation verhindert u.U. die Gewinnung neuen Wissens.

Die Besonderheit bei dem in Kapitel vier vorgestellten Ansatz auf Basis emergenter Statistik besteht zunächst in dessen flexibler Einsatzmöglichkeit. Der zugehörige

Algorithmus kann zur Analyse verschiedenster Problemstellungen (bspw. Beurteilung von Kreditanträgen) eingesetzt werden. Dabei erhält der Anwender automatisch eine introspektive Erklärung der Ergebnisse durch die Interpretation der Objekte bzw. der Schätzheuristiken in Verbindung mit den Kennzahlen wie z.B. T bzw. TU sowie DiV. Die Erklärung kann für Laien, z.B. für den Kreditantragsteller, in wenigen Sätzen leicht verständlich vermittelt werden. Während unterschiedliche statistische Verfahren bzw. ML-Verfahren in beliebigen Anwendungsgebieten zu unterschiedlichen, z.T. widersprüchlichen Ergebnissen führen[96], sind die gewonnenen Erkenntnisse, welche auf Basis emergenter Statistik gewonnen wurden, völlig frei von Widersprüchen. Zusammenhänge, die in der Vergangenheit in einem gleichen Zeitraum immer zu beobachten waren, können sich nicht widersprechen. Die Existenz der in 4.1.3 beschriebenen Meta-Gesetze ermöglicht es zudem, die gewonnenen Erkenntnisse solange für die Zukunft als Wissen zu verwenden, bis einzelne Zusammenhänge durch neu auftretende Beobachtungen falsifiziert werden.

Die Ergebnisse des vorgestellten Scoring-Modells zeigen eine Möglichkeit erklärbarer automatischer Kreditprüfungen, die sämtlichen Qualitätsanforderungen sowohl aus datenschutzrechtlicher Sicht, als auch aus Sicht von Aufsichtsbehörden und Kunden entspricht.

Die zunehmende Transparenz bei automatischen Kreditentscheidungen trägt dabei zu einem besseren Verständnis der Antragsteller über Gründe von Kreditablehnungen bei. Sie werden hierdurch u.U. dazu in die Lage versetzt, durch Manipulation ihrer Angaben eine positive Kreditentscheidung herbeizuführen. Diesem Risiko werden die Kreditinstitute durch eine sorgfältige Überprüfung der Angaben anhand von Nachweisen begegnen müssen.

[96] Vgl. Kuck, 2019.

Literaturverzeichnis

Bertelsmann Stiftung	(2018) Impuls Algorithmenethik, Nr. 5, Was bringt die Datenschutz-Grundverordnung für automatisierte Entscheidungssysteme?, bearb. V. Stephan Dreyer und Wolfgang Schulz, Gütersloh 2018.
Bertelsmann Stiftung	(2018a) Impuls Algorithmenethik, Nr. 7, Was Deutschland über Algorithmen weiß und denkt, bearb. V. Sarah Fischer und Thomas Petersen, Gütersloh 2018.
Brunner, Wolfgang L. Seeger, Jürgen Turturica, Willi	(2010) Fremdfinanzierung von Gebrauchsgütern, Das alltägliche Risiko, 1. Auflage, Wiesbaden 2010.
Dastin, Jeffrey	(2018) Amazon scraps secret AI recruiting tool that showed bias against women, in https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G, Oktober 2018, Zugriff 04.10.2019.
Fahner, Gerald	(2018) Developing Transparent Credit Risk Scorecards More Effectively: An Explainable Artificial Intelligence Approach, in: Data Analytics 2018: The Seventh International Conference on Data Analitics 2018, S. 7-14.
Fair Isaac Corporation	(2019) Explainable Machine Learning Challenge, https://community.fico.com/s/explainable-machine-learning-challenge, Zugriff: 05.09.2019.
Hans-Böckler-Stiftung	(2014) Betriebliche Mitbestimmung und betriebliche Handlungshilfen, Nr. 285, Die Mindestanforderungen an das Risikomanagement bei Kreditinstituten (MaRisk), bearb. V. Achim Sollanek und Jeanette Klessig, Düsseldorf 2014.
Kaye, Danny	(1976) Capital, 15. Ausgabe, Capital Verlag, 1976.
Konrad-Adenauer-Stiftung e.V.	(2019) Analysen & Argumente Digitale Gesellschaft, Nr. 338, Algorithmische Entscheidungen: Transparenz und Kontrolle, bearb. V. Katharina A. Zweig, Berlin 2019.
König, Christian Schröder, Jette Wiegand, Erich	(2018) Big Data Chancen, Risiken, Entwicklungstendenzen, Wiesbaden 2018.

Kuck, André Harries, Jan Philipp Kuck, Elsbeth	(2015) Der Weg zu wahren empirischen Gesetzen und rationale Handeln, durch Emergenz statt der Illusion metaphysischer Wahrheit zu empirischer Kenntnis, https://www.udpl.info/wp-content/uploads/2017/03/Kuck_Harries_WahreGWahreG.pdf, Zugriff: 01.09.2019.
Kuck, André Frischhut, Hans	(2015) Die Suche nach emergenten deterministischen empirischen Gesetzen als Alternative zur wahrscheinlichkeitsbasierten Statistik, https://www.udpl.info/wp-content/uploads/2017/03/Kuck_Frischut_.pdf, Zugriff: 01.09.2019.
Kuck, André Kratz, Norbert Frischhut, Hans	(2018) Objektives Wissen und dominante Entscheidungsregeln auf Basis emergenzbasierten Maschinenlernens, https://www.udpl.info/wp-content/uploads/2018/03/Vortrag_Fachtagung_DD_ZES.html, Zugriff: 10.11.2019.
Kuck, André	(2019) Digitale Intelligenz – emergenzbasierte Statistik, https://www.udpl.info/wp-content/uploads/2019/11/Vortrag_EUFH_HandouH.html, Zugriff: 13.11.2019.
Molnar, Christoph	(2019) Interpretable Machine Learning, A Guide for Making Black Box Models Explainable, https://christophm.github.io/interpretable-ml-book/, 2019, Zugriff: 13.11.2019.
o.V.	(2018) Stellungnahme zur aktuellen Medienberichterstattung, in https://www.schufa.de/de/ueber-uns/themenportal/detailseite/themenportal-detailseite.12544.jsp, November 2018, Zugriff 02.11.2019.
o.V.	(2010) Ökonomie ist eigentlich keine Wissenschaft, Der Mathematiker Claus Peter Ortlieb über griechische Krisen, große Zahlen und die Lebenslügen der Volkswirte, in: Frankfurter Allgemeine Sonntagszeitung, Nr. 18 v. 09.05.2010, S. 53.
Ribeiro, Marco Tulio Singh, Sameer Guestrin, Carlos	(2016) “Why Should I Trust You?”, Explaining the Predictions of Any Classifier, in: KDD’16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2016, S. 1135-1144.
Samek, Wojciech Wiegand, Thomas Müller, Klaus-Robert	(2017) Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models, Saarbrücken 2017.

Schneider, Katharina	(2019) KI erobert die Finanzbranche, in: Handelsblatt News am Abend Deutsche Leasing Edition, 27.08.2019, S. 4.
Sheh, Raymond Monteath, Isaac	(2018) Defining Explainable AI for Requirements Analysis, in: KI - Künstliche Intelligenz 2018, S. 261-266.
Taeger, Jürgen Gabel, Detlev	(2019) Kommentar DSGVO – BDSG, 3. Auflage, Frankfurt am Main 2019.
Vogl, Roland Waltl, Bernhard	(2018) Increasing Transparency in Algorithmic-Decision-Making with Explainable AI, in: Datenschutz und Datensicherheit (DuD) 2018, S. 613 – 617.

Erklärung

Ich versichere hiermit, dass ich meine Projektarbeit mit dem Titel

Erklärung algorithmischer Entscheidungssysteme

im Kreditvergabeprozess –

ein Ansatz auf Basis emergenter Statistik

selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Ich versichere zudem, dass die eingereichte elektronische Fassung mit der gedruckten Fassung übereinstimmt.

.....

Ort

....

Datum Unterschrift