

Konzeption und Anwendung eines Machine Learning-Modells für die praxisorientierte Bonitätseinschätzung von Unternehmen

Masterarbeit

IU Internationale Hochschule

Name: Leander Tripp

Studiengang: Master of Science Finance, Accounting, Taxation

Matrikelnr.: 32009913

Adresse: Düsseldorfer Str. 45, 45481 Mülheim an der Ruhr

E-Mail: leander.tripps@iu-study.org

Abgabe: 9.10.2023

Erstbetreuungsperson: Prof. Dr. rer. pol. Dirk Battenfeld

1. Einleitung	4
1.1 Problemstellung.....	4
1.2 Zielsetzung der Arbeit.....	6
1.3 Aufbau der Arbeit.....	7
2. Grundlegende Konzepte und Annahmen.....	8
2.1 Risikomanagement für Adressenausfallrisiken der Unternehmen	8
2.2 Kategoriale, Ordinale und Kardinale Modelle.....	10
2.3 Benchmarkmodelle zur Messung der Ausfallwahrscheinlichkeit.....	13
2.4 Einfluss finanzieller Kennzahlen auf Bonitätseinstufungen	19
2.5 Einfluss nicht-finanzieller Kennzahlen auf Bonitätseinstufungen.....	22
3. Entwicklung eines Bonitätseinstufungsmodells	24
3.1 Ziele und Anforderungen eines Modells.....	24
3.2 Die Nutzung von Machine Learning zur Vorhersage von Ratings	26
3.2.1 Testdatensatz zur Entwicklung der Machine Learning Modelle.....	27
3.2.2 Multinomiale Logistische Regression als Benchmarkmodell.....	31
3.2.3 K-Nearest Neighbors Algorithmus zur Vorhersage von Bonitätseinstufungen.....	33
3.2.4 Nutzung von Support Vector Machines zur Vorhersage von Bonitätseinstufungen	36
3.2.5 Extreme Gradient Boosting zur Vorhersage von Bonitätseinstufungen.....	39
3.2.6 Random Forest Klassifizierungsmodell zur Vorhersage von Bonitätseinstufungen.....	42
3.2.7 Extra-Trees Klassifizierungsmodelle zur Vorhersage von Bonitätseinstufungen.....	45
3.2.8 Vorhersage von Bonitätseinstufungen mit künstlichen neuronalen Netzen.....	47
3.3 Einbeziehung von ESG-Scores zur Verbesserung der Vorhersagequalität	59
4. Limitationen des Modells.....	67
4.1 Grenzen und Einschränkungen von ESG-Kennzahlen	67
4.2 Datensätze für die Entwicklung des neuronalen Netzes	70
4.3 Probleme mit Bonitätseinstufungen der Ratingagenturen	75
4.3.1 Bewertungskrise hypothekarisch besicherter Anleihen der 2000er Jahre	75
4.3.2 Verzögerungen bei der Auf- und Abstufung von Bonitätseinstufungen	77

4.4 Diskussion und Vergleich vorgestellter Benchmarkmodelle mit Machine-Learning Modellen zur Prognose von Bonitätseinstufungen	78
5. Fazit.....	83

1. Einleitung

1.1 Problemstellung

Kredite sind in der Wirtschaft allgegenwärtig. Ende Juni 2023 belief sich seitens der Banken an deutsche Unternehmen und Privatpersonen gewährte Kreditvolumen auf etwa 3,4 Billionen Euro (Statista, 2023). Es ist jedoch zu beachten, dass Kredite nicht ausschließlich durch Banken gewährt werden. So stellt beispielsweise jedes gegenüber einem Kunden gewährte Zahlungsziel einen Handelskredit dar. Damit ein Kredit gewährt werden kann, muss zunächst sichergestellt werden, ob ein Unternehmen in der Lage ist, den geliehenen Betrag zurückzuzahlen. Kann ein Unternehmen einen zuvor gewährten Kredit nicht wie vereinbart tilgen, ist es insolvent. Aus diesem Grund wird mithilfe verschiedener Modelle vor der Kreditvergabe die Insolvenzwahrscheinlichkeit eines Unternehmens prognostiziert.

Eine Möglichkeit zur Ermittlung der Ausfallwahrscheinlichkeit eines Unternehmens ist dabei die Ermittlung der Bonität, welche von zentraler Bedeutung für die Attraktivität und Vertrauenswürdigkeit eines Unternehmens innerhalb seines Geschäftsfelds ist. Die genaue Bestimmung der Bonität potenzieller Debitoren nimmt eine Schlüsselrolle im Risikomanagement beinahe jedes Unternehmens ein. Die Überprüfung und fortlaufende Überwachung der Debitoren verursacht gemäß der Prinzipal-Agent-Theorie Agenturkosten. Das Ziel eines Bonitätseinstufungsmodells ist die Reduzierung dieser Agenturkosten durch die Vereinfachung des Bewertungs- und Überwachungsverfahrens von Unternehmen.

Prominente Ratingagenturen wie Standard & Poor's oder Moody's nutzen zahlreiche finanzielle und nicht-finanzielle Kennzahlen sowie Expertenmeinungen, um die Bonität der zu bewertenden Unternehmen auf einer 17-stufigen Buchstabenskala einzutragen. Das aktuelle Geschäftsmodell der Ratingagenturen beruht dabei auf einem „Issuer Pays“-Ansatz, der besagt, dass Unternehmen, die Anleihen emittieren möchten, die Ratingagenturen beauftragen, eine Beurteilung der Bonität vorzunehmen. Obwohl die Bewertungen dieser Ratingagenturen als goldener Standard innerhalb der Industrie angesehen werden, besteht ein wesentliches Problem bei der Nutzung der Bonitätseinstufungen der Ratingagenturen darin, dass besonders größere Unternehmen durch die Emission von Anleihen ein Rating erhalten, während kleinere Unternehmen auf andere Methoden zur Kapitalbeschaffung zurückgreifen und daher nicht durch Ratingagenturen eingestuft werden.

Zur finalen Beurteilung der Bonität potenzieller Debitoren greifen viele Unternehmen auf Scorecard-Modelle zurück, da diese Modelle – vorausgesetzt die benötigten Kennzahlen sind verfügbar – eine Beurteilung jedes Unternehmens ermöglichen. Diese Modelle sind jedoch aufgrund ihrer technischen Konstruktion statisch und erfordern bei der Integration jeder neuen Kennzahl eine Neukalibrierung. Mit dem Aufkommen von Big Data und fortschrittlichen Technologien wie Machine Learning eröffnen sich Möglichkeiten, eine umfangreichere Palette an Datenpunkten in die

Bewertung der Bonität zu integrieren, was eine Steigerung der Adaptivität und Genauigkeit von Bonitätseinstufungsmodellen verspricht.

Obwohl große Ratingagenturen häufig betonen, dass die individuellen Meinungen hauseigener Experten einen maßgeblichen Beitrag zur finalen Bewertung beitragen, zeigten Studien bereits Anfang der 2000er Jahre Erfolg bei der Prognose von Bonitätseinstufungen, indem sie Machine Learning Modelle aus dem Bereich des überwachten Lernens mithilfe eigens ausgewählter Finanzkennzahlen trainierten. Seit dieser Zeit hat die Welt eine maßgebliche Transformation durchgemacht, die durch verschiedene Faktoren getrieben wurde. Die Veränderungen können im Kontext des VUCA-Modells (Volatility, Uncertainty, Complexity, Ambiguity) betrachtet werden, dass die Volatilität, Unsicherheit, Komplexität und Mehrdeutigkeit der globalen Geschäftsumgebung beschreibt. Da Ratingagenturen aus diesem Grund vermehrt nicht-finanzielle Kennzahlen berücksichtigen, sollten diese Kennzahlen auch in Machine Learning Modellen zur Prognose der Bonitätseinstufungen integriert werden. Darüber hinaus bieten die verbesserte Datenverfügbarkeit sowie moderne Data-Mining Technologien neue Möglichkeiten, Unternehmenskennzahlen im Internet zu ermitteln und diese für die Verwendung innerhalb der Machine-Learning-Modelle aufzubereiten. Zu diesem Zweck werden ebenfalls vermehrt Machine-Learning Verfahren eingesetzt, um die Richtigkeit der generierten Kennzahlen automatisiert zu prüfen. Dies unterstreicht, dass die Anwendung fortgeschrittener Machine-Learning-Techniken in einem Bereich zur Verbesserung der Machine-Learning-Modelle in anderen Bereichen beitragen kann, wodurch sich bereits heute Modelle künstlicher Intelligenz gegenseitig unterstützen und weiterentwickeln.

Durch die nun in Unternehmen verfügbaren Rechenleistungen und Datenhaushalten ist damit die Ablösung traditioneller Scorecard-Modelle durch die Entwicklung von Machine-Learning-Modellen denkbar. Im Vergleich zu den Scorecard-Modellen sind Machine-Learning-Modelle deutlich dynamischer und können nach erstmaliger Entwicklung schneller mithilfe neuer Kennzahlen erweitert werden, ohne dass eine vollständige Neukalibrierung erforderlich ist. Zusätzlich ist die Anwendung der mithilfe von Daten der Ratingagenturen trainierten Modelle in kleineren Unternehmen denkbar. Dies ermöglicht eine Beurteilung kleinerer Unternehmen auf Basis der Modelle führender Ratingagenturen, welche als Goldstandard im Bereich der Bonitätseinstufungen angesehen werden.

Ein inhärentes Problem im Bereich der Machine-Learning-Modelle liegt darin, dass nicht alle Modelle gleich aufgebaut sind. Es existiert eine Vielzahl unterschiedlicher Techniken im Bereich des Machine Learning, die für ein mehrklassiges Klassifizierungsproblem, wie die Prognose von Bonitätseinstufungen anhand von Finanz- und ESG-Kennzahlen, in Betracht gezogen werden können. Vor diesem Hintergrund widmet sich die vorliegende Arbeit der Frage, welche Art von Modell und Framework optimalerweise angewandt werden sollte. Zudem existieren weitere Herausforderungen im Kontext der nicht-finanziellen Kennzahlen. Da ESG-Kennzahlen ein neues

Forschungsgebiet darstellen, ist die Frage, wie diese die Insolvenz wahrscheinlichkeit eines Unternehmens tatsächlich beeinflussen, noch nicht umfassend geklärt.

1.2 Zielsetzung der Arbeit

Das Ziel der vorliegenden Masterarbeit besteht zunächst in der Identifikation der Anforderungen, die ein Bonitätseinstufungsmodell erfüllen sollte, sowie der technischen Anforderungen, die seitens des Modells berücksichtigt werden sollten. Im Einklang mit diesen Zielen soll eine Auswahl an repräsentativen Modellen etabliert werden, deren Leistung mithilfe einer zuvor etablierten Benchmark zu evaluieren ist.

Im weiteren Verlauf wird der Schwerpunkt auf die Konzeption eines eigenen Modells gelegt, das die bereits bestehende Forschung aus den Anfangsjahren des 21. Jahrhunderts zur Prognose von Bonitätseinstufungen mittels Machine-Learning-Verfahren mit den modernen Anforderungen an die Integration von ESG-Kennzahlen verknüpft. Durch die methodische Erprobung diverser Machine-Learning-Algorithmen ist zunächst das Modell zu identifizieren, das gemäß den zuvor formulierten Anforderungen die höchste Leistung zeigt. Da durch eine einfache Konstruktion das vollständige Potenzial der Modelle nicht extrahiert werden kann, ist eine kontinuierliche Optimierung bis zur Ausschöpfung des vollständigen Potenzials jedes einzelnen Modells erforderlich. Nachfolgend wird das Modell mit der höchsten Leistung um nicht-finanzielle Kennzahlen erweitert, um die potenziellen Vorteile der Einbeziehung von ESG-Kennzahlen zu evaluieren und zu ermitteln, ob durch diesen zusätzlichen Schritt eine Verbesserung der Vorhersagegenauigkeit erzielt werden kann.

Das im Rahmen dieser Arbeit entwickelte Modell sollte nach Abschluss der Trainingsphase durch weitere Kennzahlen ergänzbar sein, ohne eine Neuentwicklung des Modells erforderlich zu machen. Dies ist eine Notwendigkeit, damit das Modell zukünftigen Anforderungen der dynamischen Geschäftswelt standhält. Darüber hinaus ist über die Integration einer einfachen Benutzeroberfläche die Bedienbarkeit von Machine Learning Modellen im Vergleich zu Scorecard-Modellen zu demonstrieren. Abschließend ist über eine umfangreiche Analyse des Modells nachzuvollziehen, wie die einzelnen im Modell integrierten Kennzahlen zur Prognose tatsächlich beitragen. Hiermit können sowohl die Ergebnisse der vorangegangenen Literaturrecherche zur Auswirkung verschiedener Kennzahlen auf die Bonitätseinstufungen von Unternehmen überprüft werden sowie weitere Optimierungspotenziale innerhalb des Modells identifiziert werden. Dieser Schritt ist besonders für eine stetige Weiterentwicklung des Modells unerlässlich.

Abschließend sind die Limitationen des im Rahmen dieser Arbeit vorgestellten Modells zu identifizieren, indem auf die verschiedenen Aspekte von Machine-Learning Modellen zur Prognose von Bonitätseinstufungen einzugehen ist. Diese Analyse sollte sowohl für die im Rahmen dieser Arbeit verwendeten Datensätze und Annahmen als auch für Machine Learning Modelle, ESG-Kennzahlen und Insolvenzprognosemodelle allgemein ausgeführt werden. Durch die Identifikation dieser Schwachstellen sind allgemeine Verbesserungspotenziale herauszuarbeiten, welche Agenturkosten im Sinne der Prinzipal-Agent-Theorie weiter reduzieren können.

1.3 Aufbau der Arbeit

Die vorliegende Masterarbeit ist strukturiert in fünf Hauptteile, welche sich mit unterschiedlichen Aspekten der Bonitätseinschätzung von Unternehmen auseinandersetzen und einen schrittweisen Ansatz zur Entwicklung und Evaluierung eines Machine Learning-Modells für diese Aufgabe darstellen.

Der zweite Teil dieser Arbeit bietet eine Grundlage für das Verständnis der verschiedenen Insolvenzprognosemodelle. Zunächst werden die kategorialen, ordinalen und kardinalen Modelle vorgestellt, gefolgt von einer Präsentation finanzkennzahlenbasierter Modelle im zweiten Abschnitt. Der dritte Abschnitt erweitert diese Einführung um Modelle, die nicht-finanzielle, ökonomische und ökologische Kennzahlen berücksichtigen. Abschließend werden Benchmarkmodelle zur Messung der Schätzgüte von Insolvenzprognosemodellen vorgestellt.

Der anschließende Teil dieser Arbeit konzentriert sich auf die Entwicklung eines Modells zur Prognose von Bonitätseinstufungen. Dieser beginnt mit der Darstellung der Ziele und Anforderungen eines solchen Modells, gefolgt von einer Diskussion über die Nutzung von Machine Learning zur Vorhersage von Ratings. Innerhalb dieses Teils werden unterschiedliche Machine Learning-Modelle wie K-Nearest Neighbors, Support Vector Machines, Extreme Gradient Boosting, Random Forest, Extra-Trees Klassifizierungsmodelle und künstliche neuronale Netze anhand eines Testdatensatzes entwickelt und evaluiert. Im letzten Abschnitt des dritten Teils wird das Machine-Learning Modell mit der besten Prognoseleitung um ESG-Scores erweitert und die Auswirkung dieser auf die Prognosequalität evaluiert.

Der vierte Teil dieser Arbeit adressiert die Limitationen des entwickelten Modells. Hier werden die Probleme von ESG-Kennzahlen, Datensätze für die Entwicklung des neuronalen Netzes, die Probleme mit Bonitätseinstufungen der Ratingagenturen, die Verfügbarkeit von Finanzdaten, das Risikomanagement für Adressenausfallrisiken der Unternehmen und die Anwendung der Principal-Agent-Theorie erörtert. Zusätzlich wird ein Vergleich zwischen Scorecard-Modellen und dem Machine-Learning-Modell zur Beurteilung der Bonität vorgenommen.

Schließlich werden die im Rahmen dieser Arbeit gewonnenen Erkenntnisse innerhalb eines Fazits zusammengefasst. Im Ausblick werden zukünftige Forschungsmöglichkeiten skizziert, die sich insbesondere auf die Weiterentwicklung der Modelle und die Integration neuer Datenquellen beziehen. Dabei werden auch die Limitationen der Arbeit und Empfehlungen für die praktische Umsetzung der entwickelten Modelle in die Bonitätsbewertung hervorgehoben. Die Abschlussdiskussion bietet somit eine Basis für nachfolgende Forschungen in diesem innovativen und relevanten Bereich.

2. Grundlegende Konzepte und Annahmen

2.1 Risikomanagement für Adressenausfallrisiken der Unternehmen

In seiner grundlegendsten Form beschreibt das Adressenausfallrisiko den finanziellen Verlust, der entsteht, wenn ein Geschäftspartner seinen vertraglich vereinbarten Pflichten nicht nachkommt (Glaser, 2018, S. 17-18). Im Kontext des Risikomanagements bezieht sich das Adressenausfallrisiko primär auf den „Unexpected Loss“; den Verlust, der über den ursprünglich antizipierten Betrag hinausgeht. Zunächst ist das Adressenausfallrisiko in das Zahlungs- und Wiedereindeckungsrisiko zu unterteilen. Das Zahlungsrisiko beschreibt das Risiko, dass ein Schuldner den zuvor vereinbarten Kredit nicht bedienen kann. Das Wiedereindeckungsrisiko hingegen ist das Risiko, dass der Geschäftspartner seinen Liefer- oder Abnahmeverpflichtungen nicht nachkommen kann. Haben sich die Marktpreise in der Zwischenzeit nachteilig verändert, kann ein erneuter Vertragsabschluss zusätzlichen finanziellen Aufwand verursachen.

Besonders größere Unternehmen und Konzerne betreiben heute bereits eine Vielzahl von Systemen, die zur Erkennung eines möglichen Forderungsausfalls eingesetzt werden. Gemäß § 91 Abs. 2 Aktiengesetz trägt der Vorstand die Verantwortung für die Einrichtung eines effektiven Überwachungssystems, welches für den Fortbestand der Gesellschaft bedrohliche Entwicklungen frühzeitig erkennt. Die Existenz eines solchen Systems ist gemäß § 317 Abs. 4 HGB durch den Abschlussprüfer bei der Prüfung der Jahresabschlussunterlagen zu überprüfen.

Die BMW Group führt vor dem Abschluss sämtlicher Leasing- und Finanzierungsverträge eine umfangreiche Bonitätsüberprüfung durch (BMW AG, 2022, S. 214). Für Privatkunden wird für die Bonitätsbeurteilung ein validiertes Scoringssystem verwendet, während im Bereich der Händlerfinanzierung sowohl ein laufendes Kreditmonitoring als auch interne Ratingverfahren verwendet werden, da nicht nur die materielle Kreditwürdigkeit, sondern zusätzlich weitere Faktoren wie die Qualität der Geschäftsbeziehungen bei der Bewertung der Bonität berücksichtigt werden.

Diese internen Scoringssysteme funktionieren in ihrem Grundsatz wie die Ratingsysteme der großen Ratingagenturen (Glaser, 2018, S. 20). Die Anwendung von Scoringssystemen ist insbesondere bei Kreditprüfungen im Mengengeschäft weit verbreitet. Das Scoring basiert meistens auf einzelnen Scorecards, mithilfe denen auf Basis einer begrenzten Informationstiefe Entscheidungen über die Genehmigung der Geschäfte getroffen werden. Im Vergleich zum Rating umfasst das Scoringssystem deutlich weniger Kriterien und Informationen über den Status des potenziellen Geschäftspartners.

Jedoch ist für eine vollständige Risikoklassifizierung nicht nur die Eintrittswahrscheinlichkeit eines Forderungsausfalls, sondern zusätzlich der Nettobetrag der potenziellen Schadenshöhe zu betrachten. Während der Ausfall von Kredit- und Leasingnehmern bei einem Unternehmen mit einem Finanzierungsvolumen wie der BMW Group fast sicher ist, wird die Schadenshöhe eher gering ausfallen, da die durch die BMW Group an ihre Kunden ausgelieferten Fahrzeuge über verschiedene Handelsorganisationen in Zahlungsmittel eintauschbar sind (BMW AG, 2022, S. 214).

Abb. 1 Matrix zur Risikoklassifizierung. Potenzielle Schadenshöhe „kritisch“ entspricht $\geq 25\%$ des adjusted EBITDA des vergangenen Geschäftsjahres

Potenzielle Schadenshöhe*	Kritisch					
	Signifikant					
	Moderat					
	Niedrig					
	Marginal					
		Selten $< 10\%$	Unwahrscheinlich $\geq 10\% - < 25\%$	Möglich $\geq 25\% - < 50\%$	Wahrscheinlich $\geq 25\% - < 75\%$	Fast sicher $\geq 75\%$
Eintrittswahrscheinlichkeit						

Quelle: Übernommen aus DKV Mobility Group SE, 2023.

Gemäß IFRS 9 5.5.1 sowie § 252 Abs. 1 Nr. 4 HGB (Vorsichtsprinzip) sind für erwartete Verluste aus Leasingverhältnissen, Vertragsvermögenswerten oder einer Kreditzusage Wertberichtigungen zu erfassen. Die BMW Group ermittelt das Ausfallrisiko für Forderungen aus Lieferungen und Leistungen auf Basis der Überfälligkeit. Hierfür wird gemäß IFRS 9 das folgende Überfälligkeitssband angewandt: Eine Forderung ist entweder nicht überfällig, zwischen einem bis dreißig Tage überfällig, zwischen 31 und 60 Tagen überfällig, zwischen 61 und 90 Tagen überfällig oder über 90 Tage überfällig. Mit zunehmender Dauer der Fälligkeit erhöht sich die Wahrscheinlichkeit des Ausfalls.

Eine Möglichkeit der Absicherung des Kreditors gegenüber dem Ausfall von Forderungen aus Lieferungen und Leistungen ist die Anwendung eines Factorings (Perridon et al., 2016, S. 511). Factoring bezeichnet einen vertraglich geregelten, fortlaufenden Prozess, bei dem ein als Faktor bezeichnetes, spezialisiertes Finanz- oder Kreditinstitut, Forderungen aus erbrachten Lieferungen und Leistungen vor ihrer Fälligkeit ankaufst. Abhängig von der konkreten Ausgestaltung des Factoring-Vertrags übernimmt der Faktor bestimmte Servicefunktionen, darunter häufig auch das Ausfallrisiko. Ein Factoring kann somit sowohl Finanzierungsfunktionen über den Ankauf der Forderungen, Dienstleistungsfunktionen durch Verwaltung des Gesamtbestands aller Forderungen, als auch Kreditversicherungsfunktionen annehmen.

Zu den Vorteilen des Factorings zählen neben der Kosteneinsparung in den Bereichen der Debitorenbuchhaltung, Kreditprüfung und Mahnwesen auch die Einsparungen von Monitoring-Kosten für die Beschaffung von Informationen über den Debitor (Perridon et al., 2016, S. 513). Aus diesem Grund lohnt sich das Factoring eher für kleine und mittelgroße Unternehmen, besonders wenn aufgrund von Skaleneffekten der Faktor seine Debitorenbuchhaltung kosteneffizienter führen kann. Der Aufbau eines umfangreichen Scoring- und Überwachungssystems für Gläubiger stellt

somit eine Make-or-Buy-Entscheidung dar: Während größere Unternehmen wie die BMW Group über umfangreiche IT-Kapazitäten zum Aufbau eines Scoring-Systems verfügen, wäre der Aufbau derartiger Systeme mit großen Investitionskosten verbunden und damit auf absehbare Zeit nicht rentabel. Das Factoring stellt damit eine attraktive Möglichkeit dar, Dienstleistungen im Bereich der Kreditorenbuchhaltung auszulagern.

Eine weitere Möglichkeit der Absicherung gegenüber Adressenausfallrisiken ist die Forfaitierung einer Forderung. Analog zum Factoring kauft bei einer Forfaitierung der Forfaiteur die Forderung ohne Rückgriffsmöglichkeit auf (Perridon et al., 2016, S. 513). Mit Abschluss des Kaufvertrags zwischen dem Exporteur und dem Forderungskäufer gehen alle Rechte und Risiken aus der Forderung auf den Forfaiteur über (Grundmann, 2021, S. 143). Während der Exporteur lediglich für den Bestand der Forderungen haftet, trägt der Forfaiteur neben dem Adressenausfallrisiko jegliche mit der Forderung assoziierten Währungs-, Konvertierungs- und Transferrisiken. Anders als beim Factoring übernimmt der Forderungskäufer keine sonstigen Verwaltungsdienstleistungen für das Forderungsmanagement (Perridon et al., 2016, S. 513).

Die Veräußerung von Forderungen mittels einer Forfaitierung kommt in den meisten Fällen nur bei mittel- bis langfristigen Exportforderungen in Frage (Perridon et al., 2016, S. 513). Aufgrund mangelnder Rückgriffsmöglichkeiten müssen die zur Forfaitierung angebotenen Forderungen erstklassig und bankgarantiert sein, um eine aufgrund der Auslandsrisiken aufwändige Kreditwürdigkeitsanalyse des Importeurs zu vermeiden (Grundmann, 2021, S. 143). Sind die Voraussetzungen für die Forfaitierung erfüllt, kann der Forfaiteur seine Entscheidung zur Übernahme der Exportforderung gänzlich vom Länderrisiko abhängig machen.

2.2 Kategoriale, Ordinale und Kardinale Modelle

Insolvenzprognosemodelle generieren verfahrensbedingt Ergebnisse verschiedener Skalenniveaus (Bemann, 2005, S. 6-7). Sie lassen sich in kategoriale, ordinale und kardinale Modelle unterteilen. Kategoriale Insolvenzprognoseverfahren kennen lediglich zwei Ausprägungen des zu beurteilenden Unternehmens: Das Unternehmen wird innerhalb einer bestimmten Zeitperiode ausfallen, oder nicht. Ein Beispiel hierfür ist der 1968 von Altman entwickelte Z-Score, der im weiteren Verlauf dieser Arbeit vorgestellt wird (Altman, 1968, S. 594-596). Da jedoch auch kategoriale Insolvenzprognosemodelle meist einen Zahlenwert als Ergebnis zurückgeben, wird dieser oft ordinal interpretiert (Altman & Saunders, 1997, S. 1736-1737).

Eine Herausforderung bei kategorialen Insolvenzprognosemodellen besteht in der Trennschärfe der erstellten Prognose (Bemann, 2005, S. 9). Unternehmensinsolvenzen werden nicht nur durch prognostizierbare Ereignisse ausgelöst, die mit einem Blick auf den Jahresabschluss vorhersehbar sind, sondern auch durch Ereignisse, deren Eintrittswahrscheinlichkeit lediglich stochastisch erfassbar ist. Unternehmen, die durch kategoriale Insolvenzprognosemodelle falsch klassifiziert wurden, zeigen oft keinerlei Warnsignale, und zahlen im Vorjahr ihrer Insolvenz teilweise noch Dividenden aus (Ohlson, 1980, S. 129). In Ohlsons Untersuchungen wiesen 11 von 13

Unternehmen, die als „Nichtausfall“ klassifiziert wurden und anschließend ausfielen, im Vorjahr einen Gewinn aus. Diese Fehlprognosen werden in den Modellauswertungen als Fehler I. Art erfasst (Bemmamn, 2005, S. 9). Prognostiziert das Modell, dass ein Unternehmen innerhalb der kommenden Periode ausfällt, dieser Ausfall jedoch ausbleibt, wird diese Prognose den Fehlern II. Art zugeordnet. Hieraus lässt sich eine 2x2-Matrix ableiten (Abbildung 2).

Abb. 2 Kontingenztabelle für kategoriale Insolvenzprognosemodelle

	prognostizierte Nichtausfälle	prognostizierte Ausfälle	
tatsächliche Nichtausfälle	✓ korrekte Prognose	✗ Fehler II. Art	$\Sigma=100\%$
tatsächliche Ausfälle	✗ Fehler I. Art	✓ korrekte Prognose	$\Sigma=100\%$

Quelle: Übernommen aus Bemmamn, 2005, S. 9.

Die Kosten Fehler I. Art sind deutlich höher als die Kosten eines Fehler II. Art (Balcaen & Ooghe, 2006, S. 12). Darüber hinaus sind Unternehmen, die innerhalb der nächsten Periode nicht insolvent sein werden, deutlich häufiger in der Population vertreten als Unternehmen, die innerhalb der nächsten Periode tatsächlich ausfallen. Das Kostenverhältnis der Fehlklassifizierung sowie die Bevölkerungshäufigkeiten der untersuchten Unternehmen führen dazu, dass Insolvenzprognosemodelle allgemein hohe Fehlerraten II. Art ausweisen.

Im Gegensatz zu kategorialen Insolvenzprognosemodellen ermöglichen ordinale Insolvenzprognosemodelle einen Vergleich zwischen Unternehmen, ohne eine konkrete Ausfallwahrscheinlichkeit zu benennen. Damit sind ordinale Insolvenzprognosemodelle ideal geeignet für Ratingagenturen wie Moody's oder S&P, da der primäre Anspruch dieser Agenturen darin besteht, eine relative Ausfallwahrscheinlichkeit zu einem gewissen Zeitpunkt zu erfassen (Cantor & Mann, 2007, S. 6). Die meisten Ratingagenturen verwenden zur Veröffentlichung ihrer Bewertungen Buchstabenskalen (Estrella et al., 2000, S. 20-32). Hierfür hat sich in der Praxis die diskrete, 7- bzw. 17-stufige Skala durchgesetzt (Bemmamn, 2005, S. 6).

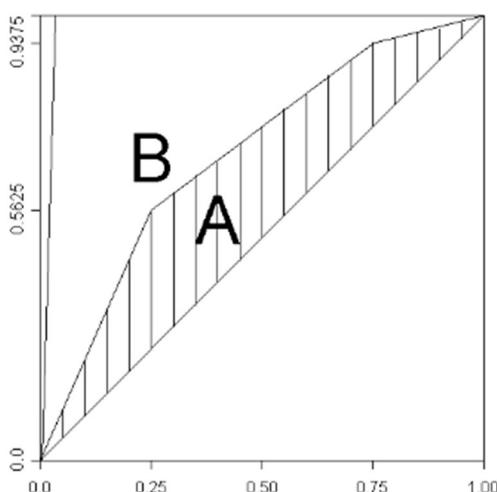
Mithilfe dieser Ratingskalen sind Ratingagenturen wie Moody's oder S&P in der Lage, Unternehmen mit unterschiedlichsten Ausfallwahrscheinlichkeiten zu identifizieren (Bemmamn, 2005, S. 13). Die Güte der Bewertung durch die jeweilige Agentur kann ex-post durch den monotonen Anstieg der Insolvenzhäufigkeit bewertet werden. Ein fehlerhafter Bewertungsprozess würde dann vorliegen, wenn Unternehmen einer besseren Abstufung (beispielsweise BBB) eine höhere Ausfallrate aufzeigen als Unternehmen einer niedrigeren Abstufung (beispielsweise CCC). Um jedoch die

Qualität eines ordinalen Insolvenzprognosemodells zu messen, muss auch die Verteilung der Unternehmen auf die jeweiligen Ratingklassen bekannt sein (Bemmann, 2005, S. 16).

Die Qualität eines ordinalen Insolvenzprognosemodells kann anhand der ROC-Kurve bestimmt werden (Krämer, 2003, S. 8). Die Datenvorbereitung besteht aus der Sortierung der zu untersuchenden Unternehmen von hoher Ausfallwahrscheinlichkeit nach geringer Ausfallwahrscheinlichkeit. Ein indifferentes Rating-System, das in jeder Klasse die gleichen, relativen Ausfallhäufigkeiten zeigt, hat eine diagonale ROC-Kurve. Die ROC-Kurve eines funktionierenden ordinalen Insolvenzprognosemodells zeigt durch die Sortierung der Ausfallwahrscheinlichkeiten von groß nach klein eine konkave Ausprägung (Abbildung 3). Liegt die ROC-Kurve eines Modells A bei der Bewertung derselben Unternehmen links oberhalb der ROC-Kurve eines anderen Modells B, liefert das Modell A für jeden Trennwert eine bessere Prognose (Bemmann, 2005, S. 18). Damit ermöglicht die ROC-Kurve einen grafischen Vergleich verschiedener Modelle.

Die mathematische Bestimmung des Flächeninhalts unterhalb der ROC-Kurve ermöglicht den quantitativen Vergleich verschiedener Prognosemodelle (Engelmann et al., 2003, S. 6). Das zufällige Modell ohne jegliche Unterscheidungskraft zwischen den Klassen weist einen Flächeninhalt von 0,5 auf, während ein perfektes Modell einen Flächeninhalt von 1 zeigen würde. Weist ein ordinales Insolvenzprognosemodell einen Flächeninhalt unterhalb der ROC-Kurve von 0,85 auf, läge die Wahrscheinlichkeit, dass ein zufällig ausgewähltes Unternehmen einer Ratingklasse eine geringere Insolvenzwahrscheinlichkeit aufweist als ein Unternehmen einer schlechteren Ratingklasse bei 85% (Engelmann et al., 2003, S. 9).

Abb. 3 Beispielhafte ROC-Kurve von Kreditausfällen



Quelle: Übernommen aus Krämer, 2003, S. 8.

Kardinale Insolvenzprognosemodelle ordnen jedem Unternehmen eine konkrete Ausfallwahrscheinlichkeit zu (Bemmann, 2005, S. 7-8). Diese werden meist als Grundlage für

quantitative, wirtschaftliche Entscheidungen verwendet. Die seit Ende 2006 geltenden neuen Eigenkapitalanforderungen des Basler Ausschusses für Bankenaufsicht schreiben aus diesem Grund die Nutzung kardinaler Insolvenzprognosen vor (Basler Ausschuss für Bankenaufsicht, 2006, S. 58-75).

Da auch Ausfallwahrscheinlichkeiten zunächst ordinal interpretiert werden können, sind die bereits für ordinale Insolvenzprognosemodelle vorgestellten Gütekriterien auch für kardinale Modelle anwendbar (Bemmamn, 2005, S. 32-34). Für die Bewertung kardinaler Insolvenzprognosemodelle muss zunächst zwischen der Trennfähigkeit und Kalibrierung unterschieden werden. Die zuvor vorgestellte ROC-Kurve (Abbildung 3) misst lediglich, inwiefern die Rangfolge der Ratings den tatsächlichen Ausfallquoten folgt, sagen jedoch nichts über die tatsächliche Treffsicherheit der Ausfallwahrscheinlichkeiten aus (Krämer, 2003, S. 9).

Zur Messung der Trennfähigkeit und Kalibrierung eines kardinalen Bewertungsmodells werden Modelle mit Straffunktionen verwendet, die Modelle bestrafen, die zwar gemäß der ROC-Kurve eine korrekte Kalibrierung aufweisen, jedoch nicht ausreichend trennfähige Ausfallwahrscheinlichkeiten bereitstellen (Bemmamn, 2005, S. 34). Ein Modell, bei dem 10% aller Unternehmen ausfallen, für die das Modell eine Ausfallwahrscheinlichkeit von 10% vorhergesagt hatte, sowie 20% aller Unternehmen ausfallen, für die das Unternehmen eine Ausfallwahrscheinlichkeit von 20% vorhergesagt hatte, würde gemäß der logarithmischen Straffunktion bestraft, da es zwar gemäß der ROC-Kurve eine perfekte Kalibrierung aufweist, jedoch keine trennfähigen Prognosen aufstellt. Eine perfekt trennfähige Prognose lässt lediglich zwei mögliche Ausprägungen zu: Entweder fällt ein Kredit sicher aus (100%) oder er fällt nicht aus (0%) (Krämer, 2003, S. 2). Fallen in einer Volkswirtschaft 2% der Unternehmen jährlich aus, würde ein gemäß der logarithmischen Straffunktion perfektes Modell für diese 2% der Unternehmen eine Ausfallwahrscheinlichkeit von 100% angeben, für alle anderen Unternehmen eine Ausfallwahrscheinlichkeit von 0% (Bemmamn, 2005, S. 34). Ein solches Modell würde sowohl eine perfekte Kalibrierung als auch eine perfekte Trennschärfe aufweisen.

2.3 Benchmarkmodelle zur Messung der Ausfallwahrscheinlichkeit

Eins der am häufigsten in der Praxis gewählten Instrumente zur Beurteilung der Ausfallwahrscheinlichkeit und des Kreditrisikos sind Kredit-Scorecards. Sie repräsentieren häufig eine Synthese diverser Techniken und Algorithmen, die zur Bewertung der Bonität sowohl von Unternehmen als auch Einzelpersonen herangezogen werden (Sadatrasoul, 2018, S. 93). Banken sind in den vergangenen Jahren dazu übergegangen, mehrere Scorecards gleichzeitig zu verwenden. Aus diesem Grund kann nicht von „der Scorecard“ als einheitliches Instrument gesprochen werden, sondern als Oberbegriff für die Kombination verschiedener Verfahren und Techniken. Einige Scorecard-Modelle kombinieren im Entscheidungsprozess bereits Machine Learning Verfahren mit traditionellen, statischen Verfahren. Im Rahmen dieser Arbeit wird lediglich

die Anwendung einer einzelnen Scorecard diskutiert, welche die Eignung potenzieller Debitoren mithilfe logistischer Regressionstechniken misst (Gao et al., 2016, S. 202-205).

In einer von Gao et al. veröffentlichten Untersuchung wird dargelegt, wie ein Portfoliomanager durch Anwendung einer Scorecard Entscheidungen bezüglich der Aufnahme von Klienten fällt (Gao et al., 2016, S. 202). Dabei repräsentiert die Ausgangsvariable Z zwei potenzielle Bewertungskategorien der künftigen Klienten: 'g' für solche Klienten, bei denen innerhalb eines Jahres kein Ausfall prognostiziert wird, und 'b' für diejenigen, bei denen ein solcher Ausfall innerhalb dieses Zeitraums wahrscheinlich erscheint. Der Vektor x aggregiert umfassende Informationen der potenziellen Klienten, einschließlich Verhaltensattributen und demographischen Daten. Aus dieser Matrix generiert die Scorecard einen kontinuierlichen Wert $s = s(x)$, welcher die prognostizierte Performance des Klienten für die nachfolgende Periode abbildet. Durch Anwendung des Bayesschen Theorems folgt, dass $p(Z|s)$ die bedingte Wahrscheinlichkeit des Ergebnisses Z in Abhängigkeit vom ermittelten Score s darstellt. Es zeigt sich, dass die Ausfallwahrscheinlichkeit $p(b|s)$ invers proportional zum Score s verläuft. Unter Einbeziehung dieser Scorecard ist der Portfoliomanager in der Position, jeden potenziellen Klienten adäquat zu evaluieren und fundierte Entscheidungen im Einklang mit seinen Zielsetzungen hinsichtlich der Aufnahme von Klienten zu treffen.

Bereits 1966 stellte William Beaver eine Insolvenzprognosestudie vor, in der anhand einzelner finanzieller Kennzahlen die Ausfallwahrscheinlichkeit eines Unternehmens ermittelt wird (Beaver, 1966, S. 100). In seinen abschließenden Bemerkungen stellte er dabei fest, dass in zukünftigen Untersuchungen die Leistung von multivariaten Modellen untersucht werden sollte, die mehrere Kennzahlen kombinieren. Bereits zwei Jahre später entwickelte Altman sein Z-Score Modell, welches mithilfe einer Gleichung aus finanziellen Kennzahlen eine Punktzahl ausgibt (Abbildung 4) (Altman, 1968, S. 594). Unternehmen, die einen Z-Score von über 2,99 erreichen, fallen in den „nicht-insolvent“ Bereich, Unternehmen mit einem Z-Score von unter 1,81 wird eine Insolvenz prognostiziert (Altman, 1968, S. 606). Der Bereich zwischen 1,81 und 2,99 wird als „Grauzone“ klassifiziert. Damit handelt es sich beim Z-Score um ein kategoriales Insolvenzprognosemodell, dessen Trennfähigkeit jedoch anhand ordinaler Gütekriterien gemessen wird (Bemann, 2005, S. 70).

Der besondere Vorteil des multivariaten Modells liegt im Verhältnis der Variablen zueinander. Eine Variable wie Umsatz/Bilanzsumme, die sich im univariaten Kontext als insignifikant herausgestellt hat, wurde aufgrund ihrer Verbindung zu den weiteren Variablen im Z-Score berücksichtigt (Altman, 1968, S. 595-596). Alle weiteren Variablen sind auf dem 0,001 Niveau signifikant (Altman & Hotchkiss, 2005, S. 243). Die starke Gewichtung der Kennzahl X_5 beruht darauf, dass die Werte der anderen Kennzahlen prozentual gemessen werden (Beinert et al., 2006, S. 3).

Abb. 4 Der Altmansche Z-Score

$$Z = 0,012 \cdot X_1 + 0,014 \cdot X_2 + 0,033 \cdot X_3 + 0,006 \cdot X_4 + 0,999 \cdot X_5$$

mit

- X_1 : Working Capital/Bilanzsumme
- X_2 : Gewinnrücklagen/Bilanzsumme
- X_3 : EBIT/Bilanzsumme
- X_4 : Marktwert des Eigenkapitals/Fremdkapital¹⁰
- X_5 : Umsatz/Bilanzsumme

Quelle: Übernommen aus Beinert et al., 2006, S. 3.

In einem Zeithorizont von einem Jahr zur Insolvenz erreichte das Modell eine Trefferquote von 95% (Altman, 1968, S. 604). Mit einem zunehmendem Zeithorizont fällt die Genauigkeit des Modells stark ab. Bei einem dreijährigen Zeithorizont wäre bereits ein Münzwurf treffsicherer (Beinert et al., 2006, S. 4). In späteren Werken Altman wurde das Modell weiterentwickelt, um Verzerrungen zu vermeiden und das Modell für Unternehmen verschiedener Industrien anwendbar zu machen (Altman & Saunders, 1997, S. 1736-1737). Der Zⁿ-Score eliminiert daher die Variable X_5 aus dem Modell und schätzt die übrigen Variablen neu.

Ein weiteres Modell zur Messung der Ausfallwahrscheinlichkeit eines Unternehmens ist das Merton Distance-to-Default Modell. Das Modell adaptiert das 1974 von Merton konzipierte Framework, welches die Aktie eines Unternehmens als Call-Option auf den inhärenten Unternehmenswert betrachtet, wobei der Ausübungspreis dem Nennwert des Fremdkapitals entspricht (Bharath & Shumway, 2008, S. 1340). Das Distance-to-Default Modell folgt dabei den drei grundlegenden Annahmen des Merton Modells (Jessen & Lando, 2015, S. 3). Zunächst folgt die Gesamtheit der Vermögenswerte des Unternehmens einer Brownschen Bewegung und zeigt daher eine konstante Volatilität. Darüber wird das gesamte Fremdkapital als eine einzelne Nullkupon-Anleihe konzeptualisiert, implizierend, dass ein potenzieller Ausfall ausschließlich zum Fälligkeitstermin eintreten kann. Innerhalb dieses Modellrahmens werden zudem jegliche Marktfriktionen außer Acht gelassen.

Abb. 5 "Distance to Default" nach Merton mit der erwarteten Vermögensabwanderung des Unternehmens μ , dem Unternehmenswert V und dem Ausübungspreis F

$$DD = \frac{\ln(V/F) + (\mu - 0.5\sigma_V^2)T}{\sigma_V \sqrt{T}},$$

Quelle: Übernommen aus Bharath & Shumway, 2008, S. 1344.

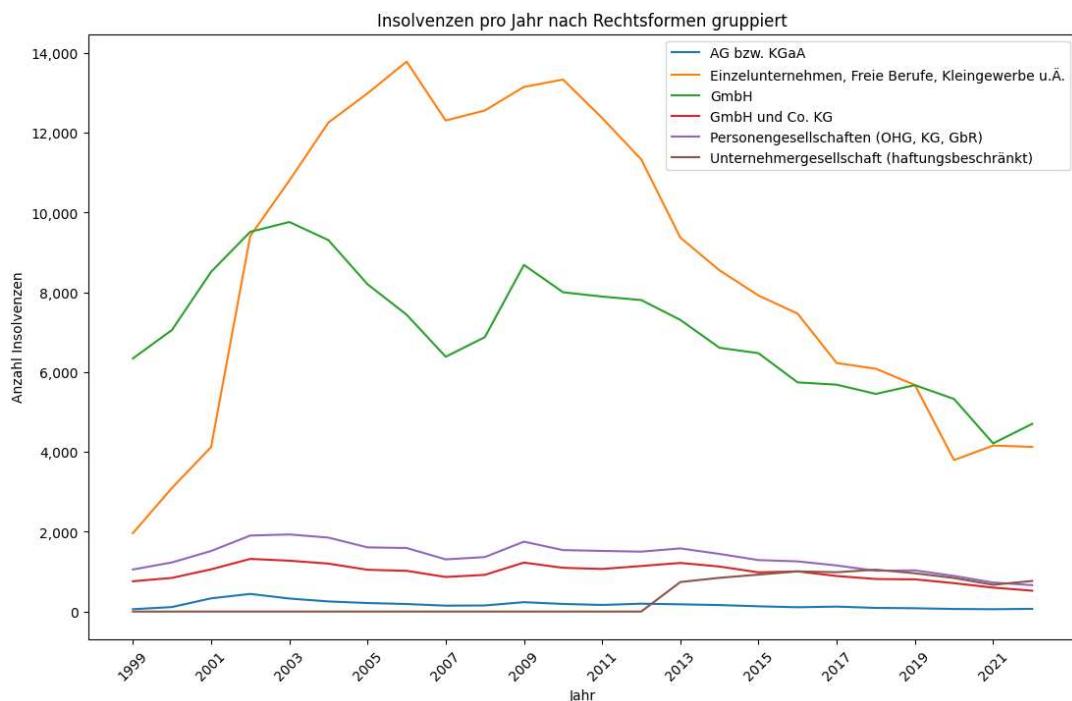
Zur Errechnung der Ausfallwahrscheinlichkeit nutzt das Distance-to-Default Modell nach Merton zwei nichtlineare Gleichungen, um den Wert und die Volatilität des Eigenkapitals in eine implizierte Ausfallwahrscheinlichkeit zu verwandeln (Bharath & Shumway, 2008, S. 1343-1344). Der Marktwert

des Eigenkapitals wird dabei zunächst mithilfe des Black-Scholes-Merton Modells errechnet. Hierfür wird der Marktwert des Fremdkapitals vom Gesamtwert des Unternehmens subtrahiert. Aufgrund der Annahmen des Merton Modells gleicht der Marktwert des Fremdkapitals dem Wert einer risikolosen Diskontanleihe abzüglich des Werts einer Put-Option mit dem Ausübungspreis des Fremdkapitalnennwerts und einer Fälligkeit in T (Bharath & Shumway, 2008, S. 1343). Die zweite Gleichung setzt dann die Volatilität des Gesamtwerts der Unternehmung ins Verhältnis zur Volatilität des Eigenkapitals. Der durch die Gleichung in Abbildung 5 ausgegebene „DD“-Wert ist analog zum Z-Score ordinal interpretierbar (Miller, 2009, S. 1-2). Ein höherer DD-Wert bedeutet eine niedrigere Distanz zum Ausfall und somit eine höhere Ausfallwahrscheinlichkeit.

Abschließend ist als weitere Benchmark eine auf Basis öffentlich verfügbarer Daten durchgeführte univariate Analyse der Rechtsform und Branchenzugehörigkeit zur Prognose der Ausfallwahrscheinlichkeit eines Unternehmens zu betrachten (Bemann, 2005, S. 51). Das statistische Bundesamt gibt Daten über Insolvenzhäufigkeiten und die Anzahl der Insolvenzen für Unternehmen verschiedener Wirtschaftsabschnitte, Rechtsformen sowie Bundesländer aus. Ein Blick auf die Insolvenzhäufigkeiten der verschiedenen Unternehmen offenbart, dass besonders Unternehmen aus dem Verkehr- und Lagereigewerbe hohe Insolvenzquoten aufweisen (in 2021 110 von 10.000 Unternehmen insolvent; damit eine Insolvenzquote von 1,1%).

Eine Analyse der jährlichen Insolvenzfälle nach Rechtsformen zeigt, dass insbesondere Kapitalgesellschaften in den zurückliegenden Jahren eine Zunahme bei den Insolvenzanmeldungen verzeichneten. Im Gegensatz dazu ist bei Einzelunternehmen ein signifikanter Rückgang und bei Personengesellschaften ein moderater Rückgang der Insolvenzfälle zu beobachten (Abbildung 6).

Abb. 6 Insolvenzen pro Jahr nach Rechtsform gruppiert



Quelle: Eigene Darstellung auf der Basis von Destatis, 2022.

Da das statistische Bundesamt für Einzelunternehmen, Personengesellschaften sowie Kapitalgesellschaften eine gruppierte Information über die Anzahl der Unternehmen herausgibt, können hierüber die Insolvenzquoten der Unternehmen nach Rechtsform für das Jahr 2021 ermittelt werden (Destatis, 2022a).

Auffällig ist hierbei, dass die unterschiedlichen Rechtsformen der Kapitalgesellschaften die höchsten relativen Insolvenzquoten aufweisen (GmbH, Unternehmergegesellschaft sowie Aktiengesellschaften). Dies ist konsistent mit den Auswertungen Bemmans, nach denen GmbHs mit Abstand die höchsten Insolvenzquoten aufweisen. Bemmans Auswertungen beruhen jedoch auf Daten von 2005, vor der Einführung von Unternehmergegesellschaften zum 1.11.2008 (Braun & Richter, 2010, S. 1). In den Daten des statistischen Bundesamtes werden die Insolvenzen von Unternehmergegesellschaften das erste Mal für das Jahr 2013 erfasst (Abbildung 6). Die Rechtsform der Unternehmergegesellschaft ermöglicht die Gründung einer Kapitalgesellschaft mit einem Stammkapital von einem Euro (Braun & Richter, 2010, S. 1). Damit wurde besonders für Kleinunternehmer und Gründer eine Eintrittsbarriere zur Gründung einer Kapitalgesellschaft abgebaut. Es ist jedoch erkennbar, dass Unternehmergegesellschaften eine hohe Insolvenzquote aufzeigen, die noch deutlich oberhalb der Insolvenzquoten der GmbHs bzw. Aktiengesellschaften liegt (Tabelle 1). Obwohl Unternehmergegesellschaften weniger als 1,5% aller Unternehmen ausmachen, werden ihnen 4,8% aller Insolvenzen im Jahr 2021 zugeschrieben.

Während Einzelunternehmen mit Abstand den größten Anteil an der Gesamtmenge an Unternehmen haben (66,56%) entfallen auf sie lediglich 40% aller Insolvenzen, während 53% der Insolvenzen auf Kapitalgesellschaften entfallen. Personengesellschaften wie OHGs und GbRs machen einen besonders geringen Anteil der Insolvenzen aus. Während knapp 13% aller untersuchten Unternehmen Personengesellschaften waren, entfielen auf sie lediglich 7% aller Insolvenzen im Jahr 2021 (Tabelle 1).

Tab. 1 Relative Insolvenzhäufigkeit nach zusammengefassten Rechtsformen

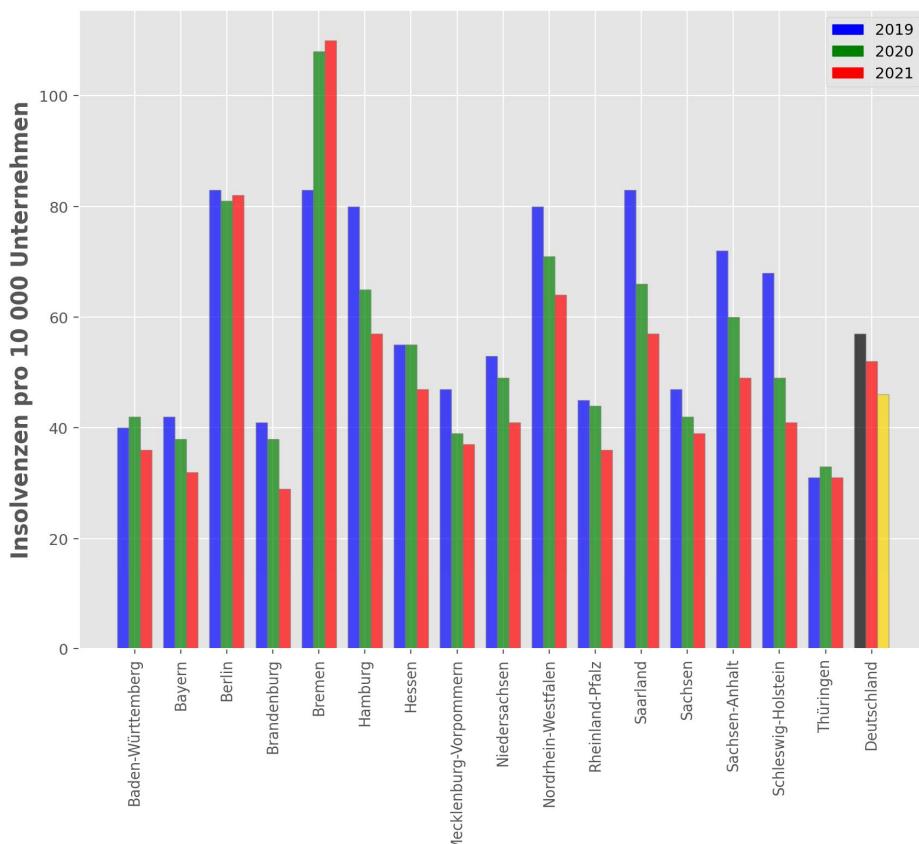
Jahr	Rechtsform	Anzahl Insolvenzen	Anteil an der Gesamtzahl Unternehmen	Insolvenzquote in Prozent
2021	Einzelunternehmen	4 161	62,56%	0,207%
2021	Personengesellschaften	728	12,82%	0,177%
2021	GmbH & Co. KG	600	4,93%	0,402%
2021	Unternehmergegesellschaft	671	1,48%	1,503%
2021	GmbH	4216	18,74%	0,744%

2021	AG bzw. KGaA	59	0,24%	0,796%
------	--------------	----	-------	--------

Quelle: Eigene Darstellung auf der Basis von Destatis, 2022.

Das letzte univariate Erklärungskriterium des Benchmarkmodells für Insolvenzwahrscheinlichkeiten ist die Verteilung der Unternehmensinsolvenzen auf die 16 Bundesländer der Bundesrepublik Deutschlands (Bemann, 2005, S. 54-55). Das statistische Bundesamt stellt hierfür eine tabellarische Auswertung der Jahre 2019 bis 2021 zur Verfügung (Destatis, 2022b). Die Hansestadt Bremen weist nicht nur die höchsten Insolvenzquoten der Bundesrepublik auf (1,1% in 2021), sondern zudem als einziges Bundesland eine höhere Insolvenzquote im Jahr 2021 als im Basisjahr 2019 aus (Abbildung 7). Weiterhin ist ein klares Nord-Süd-Gefälle der Insolvenzquoten erkennbar: Während Bundesländer wie Bremen, Hamburg und Schleswig-Holstein hohe Insolvenzquoten aufzeigen, schneiden Baden-Württemberg, Bayern und Rheinland-Pfalz deutlich besser ab (Abbildung 7). Dies stimmt mit den Auswertungen Bemanns überein und deutet auf geringe Veränderungen der Insolvenzquoten der Bundesländer hin (Bemann, 2005, S. 54). Abgesehen von den Daten der Hansestadt Bremen ist jedoch die Konstanz der nationalen Insolvenzquote bemerkenswert, insbesondere in Zeiten der COVID-19 Pandemie. Da die offizielle Datenreihe zur Zeit der Erstellung dieser Arbeit im Jahr 2021 endet ist die Anzahl der in die Folgejahre geschobenen Insolvenzen noch unbekannt.

Abb. 7 Insolvenzen pro 10 000 Unternehmen nach Bundesländern für 2019 bis 2021



Quelle: Eigene Darstellung auf der Basis von Destatis, 2022b.

Die drei Variablen Branche, Rechtsform und Bundesland eines Unternehmens können jeweils einzeln als univariate Variable zur Bestimmung der Insolvenz wahrscheinlichkeit verwendet werden. Zusätzlich zur univariaten Auswertung wären jedoch Erkenntnisse über die multivariate Insolvenzprognosegüte der Variablen hilfreich, um beispielsweise zu wissen, wie hoch die tatsächliche Insolvenzquote für GmbHs aus der Verkehrsbranche aus Nordrhein-Westfalen ist (Bemann, 2005, S. 55).

Diese Daten können jedoch nicht durch das statistische Bundesamt bereitgestellt werden, sondern müssten mithilfe der Daten der statistischen Landesämter zusammengestellt werden (Bemann, 2005, S. 55). Durch das statistische Bundesamt konnte jedoch eine Auswertung der Insolvenzen nach Rechtsform und Branche bereitgestellt werden, sodass eine multivariate Auswertung für diese Variablen erfolgen kann. Bemanns Auswertungen zeigen, dass Unternehmen mit gleicher Rechtsform abhängig von der Branche deutlich unterschiedliche Ausfallraten zeigen. Unternehmen aus dem Baugewerbe zeigen über alle Rechtsformen hinweg hohe Ausfallquoten, während für Unternehmen aus dem Energiesektor konstant geringe Ausfallquoten ausgewiesen werden. Darüber hinaus sind für alle Rechtsformen über die Zeit von 1998 bis 2003 die branchenspezifischen relativen Ausfallraten stabil.

2.4 Einfluss finanzieller Kennzahlen auf Bonitätseinstufungen

Der Ausbau eines finanzkennzahlenorientierten Insolvenzprognosemodells erfordert zunächst eine umfangreiche Sortierung der verwendeten Kennzahlen im Modell. Kaur et al. teilen hierfür alle in der Literatur verwendeten Kennzahlen in zwölf Oberkategorien auf: Profitabilität, Verschuldung, Liquidität, Unternehmensgröße, finanzielle Deckung, Vermögensstruktur, Wachstum, Risiko, Forschung & Entwicklung, Zahlungsströme, Börsenperformance und Andere (Kaur et al., 2023, S. 95-100). Die wichtigsten Profitabilitätskennzahlen sind Kapitalrentabilität, operative Marge, Netto-Umsatzrendite und Eigenkapitalrendite (Kaur et al., 2023, S. 94).

Eine höhere Verschuldung des Unternehmens bringt allgemein steuerliche Vorteile mit sich und diszipliniert die Geschäftsführung im Interesse der Gläubiger zu handeln (Jensen, 1986, S. 3-4). Allerdings hat eine zunehmende Verschuldung des Unternehmens einen negativen Einfluss auf die Bewertung des Unternehmens (Kaur et al., 2023, S. 100-101). Dies liegt am mit der Verschuldungsquote steigendem Risiko. Die für Bewertungen am häufigsten verwendete Finanzkennzahl ist die Fremdkapitalquote.

Die zur Messung der Liquidität eines Unternehmens am häufigsten verwendeten Kennzahlen sind die Liquidität ersten-, zweiten- und dritten Grades, die Umlaufintensität, sowie die Barreserven im Verhältnis zum Gesamtvermögen (Kaur et al., 2023, S. 101). Die Liquidität ersten, zweiten und dritten Grades steht laut Kaur et al. in einem positiven Zusammenhang mit dem Rating des Unternehmens, da überschüssige Liquidität das Unternehmen befähigt auf unvorhergesehene Ereignisse zu reagieren ohne Vermögenswerte zu einem diskontierten Wert zu verkaufen. Allerdings haben zu hohe Barreserven einen negativen Einfluss auf das Rating, da diese Unternehmen durch

eine konservative Unternehmenspolitik Chancen verpassen, in ihr eigenes Wachstum zu investieren (Baghai et al., 2014, S. 1963). Obwohl dies zunächst kontraintuitiv erscheint, sind die Auswirkungen konsistent mit den Untersuchungen Acharya et al., die ergaben, dass Unternehmen, die höhere Barreserven aufweisen, langfristig einem höheren Insolvenzrisiko ausgesetzt sind (Acharya et al., 2012, S. 25).

Die Größe des Unternehmens hat einen durchweg positiven Einfluss auf die Bewertung eines Unternehmens (Kaur et al., 2023, S. 101). Größere Unternehmen sind einem geringeren Insolvenzrisiko ausgesetzt, da sie ihre Risiken über mehrere Produkte und Märkte hinweg streuen können. Die aussagekräftigsten Kennzahlen bezüglich der Größe eines Unternehmens sind die Summe aller Vermögenswerte, Marktkapitalisierung, Gesamtumsätze, Summe des Eigenkapitals sowie die Summe aller Schulden.

Die für die finanzielle Abdeckung relevanten Kennzahlen sind der Zinsdeckungs- sowie Schuldendeckungsgrad (Kaur et al., 2023, S. 101). Der Zinsdeckungsgrad zeigt an, inwiefern das Unternehmen in der Lage ist, seine Zinsaufwendungen über die eigens generierten Zahlungsströme zu decken (Gupta, 2023, S. 1624). Ein höherer Zinsdeckungsgrad bedeutet ein niedrigeres Risiko für Gläubiger und führt somit zu einer besseren Unternehmensbewertung.

Die Kapitalstruktur des Unternehmens hat ebenfalls eine Auswirkung auf die Bewertung der Kreditwürdigkeit (Kaur et al., 2023, S. 102). Unternehmen, die einen großen Anteil ihres Vermögens in Sachanlagen investiert haben, sind durch die Flexibilität dieser Vermögenswerte einem geringeren Risiko ausgesetzt (Purda, 2008, S. 20). Darüber hinaus können materielle Vermögenswerte als Sicherheit hinterlegt werden und damit im Falle eines Ausfalls eine höhere Sicherheitsquote zu gewährleisten (Purda, 2008, S. 19). Darüber hinaus zeigen Unternehmen mit einer großen Menge an Vermögenswerten der Stufe 3 ein höheres Kreditrisiko, da diese Vermögenswerte mithilfe von theoretischen Modellen und nicht verifizierbaren Annahmen bewertet werden, welche ein hohes Informationsrisiko darstellen (Kaur et al., 2023, S. 102). Zur Messung der Kapitalstruktur hat sich die Kapitalintensitätsquote als wirksame Kennzahl herausgestellt.

Ein starkes Wachstum hat einen gemischten Einfluss auf die Bewertung eines Unternehmens (Kaur et al., 2023, S. 102). Während ein starkes Wachstum einerseits starke Zahlungsströme in der Zukunft ermöglicht, stellt es gleichzeitig ein Risiko für Kapitalgeber dar. Das Markt-zu-Buchwert-Verhältnis, welches traditionell ein Unternehmen in den Value- oder Wachstumsbereich einordnen lässt, zeigt eine inverse Beziehung zur Kreditwürdigkeit (Hirk et al., 2019, S. 525-526). Adams, Burton und Hardwick fanden jedoch einen positiven Zusammenhang zwischen Wachstum und Rating, da wachstumsstarke Unternehmen sich eher freiwillig einer Bewertung unterziehen (Adams et al., 2003, S. 546). Aufgrund der gemischten Evidenz im Zusammenhang mit der Kreditwürdigkeit eines Unternehmens findet das Wachstum eines Unternehmens im weiteren Verlauf dieser Arbeit keine Berücksichtigung.

Zur Messung des Risikos, dem ein Unternehmen ausgesetzt ist, werden zumeist marktorientierte Kennzahlen verwendet (Kaur et al., 2023, S. 102). Hierzu zählen das Beta der Aktie, diversifizierbare Risiko sowie die Standardabweichung der täglichen Aktienrenditen. Darüber hinaus finden strukturierte Modelle wie „Distance to Default“ im Rahmen des Black-Scholes-Merton Modells oder die mithilfe des KMV-Merton Modells ermittelte Insolvenzwahrscheinlichkeit Anwendung. Unternehmen können ihr Risiko mithilfe von funktionierenden Corporate Governance Mechanismen reduzieren, da diese potenzielle Interessenskonflikte zwischen dem Management und den Kapitalgebern verhindern, indem die Tätigkeiten des Managements besser überwacht werden (Bhojraj & Sengupta, 2003, S. 472).

Ähnlich zum Unternehmenswachstum zeigen Ausgaben für Forschung und Entwicklung gemischte Auswirkungen auf die Kreditwürdigkeit eines Unternehmens (Kaur et al., 2023, S. 102). Obwohl die Bereitschaft, innovative Produkte zu entwickeln grundsätzlich positiv angesehen wird, überwiegen die Risiken von Forschungs- und Entwicklungsausgaben meist ihrem Nutzen. Griffin et al. Zeigten in ihren Untersuchungen, dass die Auswirkungen von Forschungs- und Entwicklungskosten auf das Unternehmensrating oft um bis zu fünf Jahre verzögert sind, was besonders die kurzfristigen Kapitalkosten in die Höhe treibt (Griffin et al., 2018, S. 31). Aus diesem Grund finden Forschungs- und Entwicklungsausgaben im Rahmen dieser Arbeit keine weitere Berücksichtigung.

Zahlungsströme stehen ebenfalls in einem positiven Zusammenhang mit dem Unternehmensrating (Kaur et al., 2023, S. 102). Unternehmen, die höhere Zahlungsströme generieren, befinden sich in einer besseren Position ihre Schulden zu bedienen (Pittman & Fortin, 2004, S. 10). Die am häufigsten verwendete Kennzahl zur Messung der Zahlungsströme des Unternehmens sind die betrieblichen Zahlungsströme (Kaur et al., 2023, S. 102). Jedoch findet auch die Kennzahl der freien Zahlungsströme im Verhältnis zur Bilanzsumme Anwendung, da diese den Erfolg des Unternehmens bei der Generierung von Zahlungsströmen im Verhältnis zur Unternehmensgröße widerspiegelt.

Abhängig von der Ratingagentur, die ein Unternehmen bewertet, findet die Performance der Aktie eines Unternehmens ebenfalls Einfluss in die Unternehmensbewertung (Kozmenko & Plastun, 2012, S. 11). Kozmenko und Plastuns Untersuchungen zeigen, dass aus den vier größten Ratingagenturen (Moody's, Standard & Poor's, Fitch und Japan Rating and Investment) lediglich Standard & Poor's die Performance der Unternehmensaktie in ihren Bewertungen berücksichtigt. Hierbei finden jedoch die Kurse selbst keinerlei Berücksichtigung, sondern lediglich die Gründe hinter den Kursbewegungen. Der Aktienkurs eines Unternehmens ist auch für Standard & Poor's lediglich ein optionales Kriterium. Sind die Gründe der Kursschwankungen nicht offensichtlich, wird die Performance der Aktie gänzlich ignoriert. Insgesamt ist die Evidenz des Einflusses der Aktienperformance auf die Unternehmensbewertung gemischt (Kaur et al., 2023, S. 103). Dividendenzahlungen der Unternehmen deuten jedoch auf eine bessere Unternehmensperformance hin und haben einen durchweg positiven Einfluss auf die Unternehmensbewertung (Jiraporn et al., 2014, S. 520).

2.5 Einfluss nicht-finanzieller Kennzahlen auf Bonitätseinstufungen

Die in Abschnitt 2.4 vorgestellten Auswertungen analysieren den Einfluss finanzieller Kennzahlen auf Bonitätseinstufungen von Unternehmen. Ein grundlegendes Problem mit finanziellen Kennzahlen ist jedoch ihre die Rückwärtsorientierung, da Finanzergebnisse lediglich für bereits vergangene Perioden vorliegen und daraus nicht unbedingt hervorgeht, wie ein Unternehmen für die Zukunft aufgestellt ist (Grunert et al., 2005, S. 5). Bei einer Umfrage im Jahr 2000 fanden Günther und Grüning bereits heraus, dass 70 von 145 befragten Banken bereits qualitative Faktoren für ihre Insolvenzprognosen nutzten, wobei 77,6% behaupteten, die Qualität der Modellierungen habe sich durch den Einsatz nicht-finanzieller Faktoren verbessert (Günther & Grüning, 2000, S. 48-52). Hierbei spielte die Qualität des Personals und des Managements mit Abstand die wichtigste Rolle, gefolgt von den aktuellen Wettbewerbern sowie den potenziellen Wettbewerbern. Darüber hinaus konnten Grunert et al. zeigen, dass die kombinierte Nutzung von finanziellen und nicht-finanziellen Faktoren zu signifikant besseren Prognosen führten (Grunert et al., 2005, S. 29).

Umweltfaktoren spielen für Unternehmen der heutigen Zeit eine zunehmend wichtige Rolle und werden daher vermehrt bei der Einschätzung potenzieller Debitoren betrachtet. Unternehmen, deren Geschäftsmodell ein Risiko für die Umwelt darstellt, sehen sich mit großer Wahrscheinlichkeit in der Zukunft mit wirtschaftlichen Sanktionen konfrontiert. Aus diesem Grund zeigen Unternehmen mit einem höheren CO₂-Ausstoß einen geringeren Distance-to-Default-Wert als umweltfreundlichere Unternehmen und gelten damit als insolvenzgefährdet (Capasso et al., 2020, S. 10-14). Darüber hinaus konnten Capasso et al. zeigen, dass die Vereinbarungen des Pariser Klimaabkommens einen signifikanten Einfluss auf das Insolvenzrisiko umweltschädlicher Unternehmen hatten, indem sie eine Dummy-Variable namens „Post-Event“ in ihre Gleichung aufnahmen, die sie für die Jahre nach dem Pariser Klimaabkommen auf eins setzten (Abbildung 8). Aufsichtsbehörden sollten Unternehmen dazu verpflichten, Berichte über ihr Emissionsverhalten und Klimarisiken ihrer Geschäftspraktiken offenzulegen (Capasso et al., 2020, S. 19).

Abb. 8 Distance to Default Gleichung zur Messung des Insolvenzrisikos unter Gewichtung der Jahre nach dem Pariser Klimaabkommen

$$DD_{it} = \alpha + \beta_1 Carbon\ Intensity_i + \beta_2 Post\ Event + \beta_3 Carbon\ Intensity \times Post\ Event + \gamma' Y_{it} + \varepsilon_{it} \quad (9)$$

Quelle: Übernommen aus Capasso et al., 2020, S. 10.

Zeidan und Onabolu schlagen daher vor, dass Insolvenzprognosemodelle nicht nur potenzielle Umweltrisiken eines Unternehmens berücksichtigen, sondern zusätzlich Unternehmen belohnen, deren Geschäftsmodelle als umweltfreundlich angesehen werden (Zeidan & Onabolu, 2023, S. 22).

Zeidan et al. entwickelten bereits 2015 das Sustainability Credit Score System (SCSS), welches die Messung der Nachhaltigkeit eines Unternehmens in sechs Stufen aufteilt: Wirtschaftliches Wachstum, Umweltschutz, soziale Fortschrittlichkeit, sozio-ökonomische Entwicklung, öko-effizienz

und sozio-umwelttechnische Entwicklung (Zeidan et al., 2015, S. 2). Das Endresultat ist ein Bewertungssystem bestehend aus sechs Matrizen für die sechs unterschiedlichen Nachhaltigkeitsdimensionen, das mithilfe von 30 Fragen ausgefüllt wird. Das Endergebnis eines Unternehmens liegt zwischen 0 und 1. Das Sona Sustainability Credit Score System erweitert die Funktionalität des SCSS indem es Endnutzern ermöglicht, Berichte zu generieren (Zeidan & Onabolu, 2023, S. 30).

Entscheidend ist jedoch, wie Nachhaltigkeitsfaktoren bestmöglich in bestehende Bonitätseinstufungsmodelle integriert werden können. Es gibt bereits Anzeichen, dass die größten bestehenden Ratingagenturen Moody's und Standard & Poor's Umweltfaktoren nun stärker gewichten. Moody's entzog Exxon Mobile seine AAA-Bewertung im November 2019 und führte als Begründung eine mangelnde Anpassung zu einer emissionsarmen Wirtschaft an, während Standard & Poor's die Bewertung der Deutschen Bahn (DB) nach seitens der Bundesregierung angekündigten klimapolitischen Maßnahmen im Oktober 2019 erhöhte (Capasso et al., 2020, S. 18). In einer Untersuchung US-amerikanischer Unternehmen fanden Bhattacharya und Sharma einen signifikanten positiven Zusammenhang zwischen dem ESG-Rating eines Unternehmens und seinem Rating (Bhattacharya & Sharma, 2019, S. 475).

Neben der Umweltbelastung eines Unternehmens fließen weitere nicht-finanzielle Faktoren in das Rating eines Unternehmens ein. Der Aufbau der Unternehmensleitung hat einen Einfluss auf die Bewertung eines Unternehmens. Treffen mehrere Personen Entscheidungen innerhalb eines Unternehmens hat dies einen positiven Einfluss auf die Unternehmensbewertung, da diese bessere, weniger risikoreiche Entscheidungen treffen (Dwivedi & Jain, 2005, S. 171). Darüber hinaus erhalten Unternehmen mit einem größeren Anteil an unabhängigen Direktoren ein besseres Rating, da diese unvoreingenommene Entscheidungen treffen (Kaur et al., 2023, S. 103). Das Alter der Vorstandsmitglieder hat ebenfalls einen Einfluss auf das Rating eines Unternehmens, da jüngere Geschäftsführer als risikobereiter gelten.

Die Rolle des CEOs innerhalb eines Unternehmens hat ebenfalls Einfluss auf die Bewertung eines Unternehmens. Die hierfür meist genutzte Kennzahl ist „CEO Dualität“ und misst ob der CEO gleichzeitig die Rolle des Verwaltungsratsvorsitzenden innehat (Kaur et al., 2023, S. 105).

Die Eigentümerstruktur eines Unternehmens hat ebenfalls Auswirkungen auf das Rating eines Unternehmens. „Aktive Investoren“ nach Jensen halten große Schuld- oder Besitzanteile an einem Unternehmen und sind aktiv an der Lenkung und strategischen Ausrichtung eines Unternehmens beteiligt (Jensen, 1993, S. 867). Die Existenz aktiver Investoren hat einen positiven Einfluss auf das Rating eines Unternehmens, da ihnen ein finanzielles Interesse bei gleichzeitiger Unabhängigkeit innewohnt. Dies befähigt sie, die Unternehmensführung und ihre Handlungen auf eine unvoreingenommene Art und Weise zu evaluieren. Andererseits haben diese aktiven Investoren möglicherweise ein gegenläufiges Interesse zu anderen Investoren. Ist ein aktiver Investor Anteilseigner des Unternehmens, zeigt er möglicherweise eine höhere Risikobereitschaft, da er von

höheren Gewinnen profitiert, während Fremdkapitalgeber des Unternehmens die Verluste im Falle eines Versagens tragen müssen (Shleifer & Vishny, 1997, S. 760). Ist der aktive Investor ein Fremdkapitalgeber, kann er möglicherweise seinen Einfluss auf das Management des Unternehmens ausnutzen, um vielversprechende Projekte zu verhindern, da er lediglich die Kosten eines solchen Projekts tragen müsste, ohne die Gewinne einzuziehen zu können. Die in der Literatur meistgenutzte Kennzahl für die Eigentümerstruktur im Rahmen eines Ratingmodells ist der Prozentsatz von Anteilen, die von institutionellen Investoren gehalten werden, übereinstimmend mit der Hypothese von Shleifer und Vishny (Kaur et al., 2023, S. 106). Eine große Anzahl von Anteilseignern, die 5% oder mehr der Anteil eines Unternehmens halten, hat ebenfalls eine negative Auswirkung auf das Unternehmensrating, da eine stärkere Konzentration von Anteilen als ein größeres Risiko angesehen wird.

Darüber hinaus beeinflusst die Qualität der finanziellen Berichterstattung eines Unternehmens die Bewertung eines Unternehmens. Hierbei ist besonders eine korrekte Abgrenzung der Erträge über ein periodengerechtes Ertragsmanagement sowie „echtes“ Ertragsmanagement wichtig (Kaur et al., 2023, S. 106). „Echtes“ Ertragsmanagement bezieht sich auf Handlungen der Geschäftsführung, die darauf abzielen, Stakeholder zu überzeugen, dass Unternehmen habe seine finanziellen Ziele erreicht (Zamri et al., 2013, S. 87-88). Dies geschieht durch Einflussnahme auf das Timing von Finanztransaktionen, Investments und Strukturierungen der Operationen nehmen. Da Finanzberichte geringer Qualität ein Informationsrisiko für alle Stakeholder des Unternehmens darstellen, steht die finanzielle Berichtsqualität in einem positiven Verhältnis zur Unternehmensbewertung.

Neben diesen unternehmensspezifischen qualitativen Faktoren haben allgemeine makroökonomische Faktoren Einfluss auf das Kredit-Rating eines Unternehmens. Gute wirtschaftliche Bedingungen am Standort des Unternehmens bringen verbesserte Möglichkeiten für das Wachstum und den Ausbau eines Unternehmens (Kaur et al., 2023, S. 108). Die zur Messung der wirtschaftlichen Lage eingesetzten Kennzahlen sind das Bruttoinlandsprodukt, Wachstum des Bruttoinlandsprodukts, Pro-Kopf-BIP sowie die Volatilität des BIP. Ein höheres BIP signalisiert eine höhere Produktion und steht im Einklang mit einem höheren Konsum der Bevölkerung (Mamilla et al., 2019, S. 330). Ein positiver Zusammenhang zwischen BIP-Wachstum und Unternehmensrating besteht daher besonders bei Unternehmen des Fertigungssektors erkennbar.

3. Entwicklung eines Bonitätseinstufungsmodells

3.1 Ziele und Anforderungen eines Modells

Ein Insolvenzprognosemodell sollte einige grundlegende Anforderungen erfüllen, um sicherzustellen, dass das Modell mit minimalen Anpassungen über viele Jahre hinweg verwendet werden kann und dabei konstante Ergebnisse hoher Prognosegüte liefert. Insolvenzprognosemodelle sollten in der Lage sein, sowohl aktuelle, als auch frühere und zukünftige

Kunden zu bewerten (Krahnen & Weber, 2001, S. 9). Dies stellt besonders in der heutigen VUCA-Umgebung eine besondere Herausforderung dar, da Unternehmen nicht wissen können, welche neuen Risikokriterien in der Zukunft eine besondere Rolle spielen (siehe insbesondere Abschnitt 2.3: Wichtigkeit der Umweltverträglichkeit von Unternehmen nach Pariser Klimaabkommen). Hat ein Modell Schwierigkeiten mit der Bewertung ausländischer Firmen oder Unternehmen einer bestimmten Branche, kann dies zu potenziell dauerhaft geschäftsschädigenden Verzögerungen beim Abschluss von Verträgen führen. Darüber hinaus sollten Kunden, die in der Vergangenheit ausgefallen sind oder nicht mehr Kunde des Unternehmens sind, fortlaufend bewertet werden, da das Entfernen dieser Kunden aus der Datenbank zu einem Bias der Daten führen kann (Krahnen & Weber, 2001, S. 9). Ein wichtiger Teil der Ausfallstatistik stellt die Quote der Kunden dar, die sich aus einer Situation der Zahlungsunfähigkeit erholen können. Zudem kann ausgewertet werden, wie hoch die Ausfallquoten der Kunden verschiedener Branchen sind.

Venkiteswaran stellte in seinen Auswertungen über die Auswirkungen des Bias der Ratings auf die Yield Spreads von Anleihen fest, dass Ratingagenturen über die Zeit strengere Standards zu Vergabe von Ratings verwenden (Venkiteswaran, 2013, S. 82). Diese Änderungen zeigten reale Auswirkungen auf den Anleihenmarkt, da eine stark negative Korrelation der Yield Spreads mit dem erhaltenen Rating besteht (Venkiteswaran, 2013, S. 88). Anleihen, die bereits unterbewertet waren, zeigten durch den Rating Bias eine weitere Reduktion des Yield Spreads. Deflationäre Tendenzen der Ratings führen durch erhöhte Yield Spreads zu erhöhten Fremdkapitalkosten für die Unternehmen, während Investoren und Portfolio Manager profitieren, da sie nun eine höhere Rendite erhalten, als durch die Ausfallwahrscheinlichkeit des Unternehmens gerechtfertigt wäre.

Darüber hinaus sollten vergangene und aktuelle Ratingdaten verfügbar und zugänglich sein, um das Testen des Systems zu vereinfachen (Krahnen & Weber, 2001, S. 13). Zur Messung der Performance von Ratingmodellen wird üblicherweise ein Backtesting durchgeführt, welches die ex-ante Ratings mit den tatsächlichen ex-post Ausfällen vergleicht. Die ex-post Ausfallquoten jeder Kategorie sollten höher sein als die Ausfallquoten der nächstbesseren Kategorie. Darüber hinaus sollten Ausfallquoten über einen Zeithorizont von fünf Jahren höher sein als die Ausfallquoten über einen Zeithorizont von zwei Jahren. Backtesting ist jedoch nicht die einzige Methode, das Ratingsystem zu verbessern. Hat das Management ex-ante Wissen über die Veränderung der Strukturen und Gewichtung der Variablen zur Beurteilung der Kreditwürdigkeit, so sollte nicht auf das Backtesting gewartet werden (Krahnen & Weber, 2001, S. 13). Diese Arten von Änderung sollten jedoch mit Vorsicht durchgeführt werden, da sie das zukünftige Backtesting erschweren.

Verlässlichkeit ist darüber hinaus ein wichtiges Kriterium beim Aufbau eines Insolvenzprognosemodells. Angenommen, jedes Unternehmen hat eine „echte“ Ausfallwahrscheinlichkeit, so sollte das Rating für Unternehmen mit der gleichen Ausfallwahrscheinlichkeit identisch sein (Krahnen & Weber, 2001, S. 11). Das Rating kann sich mit

der Zeit ändern, falls sich die Ausfallwahrscheinlichkeit ändert oder sich die wirtschaftlichen Bedingungen geändert haben.

Die finale Anforderung an ein funktionierendes Ratingsystem ist die kontinuierliche oder randomisierte Überwachung durch externe Compliance-Beauftragte (Krahnen & Weber, 2001, S. 15). Die Untersuchung der vergangenen Bewertungen darf hierbei keinerlei Bias oder vorsätzliche Falschdarstellung aufweisen. Externe Compliance stellt nicht den Informationsgehalt der Ratings selbst sicher, sondern lediglich ihre Konstanz. Externe Überwachung ist normalerweise für interne Ratingsysteme irrelevant, da das Unternehmen keinerlei Anreize hat, eigene Bewertungen zu manipulieren, da dies die eigene Geschäftsposition schwächen würde. Im Bankengeschäft hat sich dies jedoch mit den Baseler Abkommen geändert. Diese hängen einen „versteckten Preis“ an die Ratings der Unternehmen an (Krahnen & Weber, 2001, S. 16). Die Menge an Eigenkapital, die eine Bank gegenüber ihrer Bilanzsumme zu halten hat, ist direkt von den internen Ratings abhängig. Haben Schuldner der Bank ein schlechteres internes Rating, beeinflusst dies die risikogewichteten Aktiva der Bank. Eine höhere Summe risikogewichteter Aktiva erhöht gleichzeitig das regulatorische Eigenkapital der Bank (Basler Ausschuss für Bankenaufsicht, 2006, S. 14-15). Damit haben die internen Ratings für Banken gegenläufige Anreizeffekte: einerseits muss das Risiko, dem die Bank ausgesetzt ist, korrekt gemessen werden, andererseits beeinflussen die Ratings interne Eigenkapitalanforderungen.

3.2 Die Nutzung von Machine Learning zur Vorhersage von Ratings

Bonitätsbewertungen von Unternehmen sind kostspielig, da die großen Ratingagenturen wie Standard & Poor's und Moody's mit hohem Zeitaufwand und Ressourcenverbrauch das Risikoprofil eines Unternehmens anhand vieler Parameter ermitteln (Huang et al., 2004, S. 543). Aus diesem Grund können nicht alle Unternehmen sich ein jährlich aktualisiertes Rating von diesen Agenturen leisten. Besonders in aufstrebenden Märkten wie Indien sind die Finanzdaten der Unternehmen meist unstrukturiert (Pol et al., 2022, S. 39). Dies führt dazu, dass die Daten zunächst aufwendig strukturiert werden müssen, bevor Analysten mit ihrer Bewertung anfangen können.

Obwohl die Ratingagenturen und ihre Analysten die Wichtigkeit von subjektiven Einschätzungen bei der Erstellung ihrer Ratings betonen, hatten in der Vergangenheit bereits Untersuchungen Erfolg bei der Schätzung von Bonitätseinstufungen mithilfe von Machine Learning und Deep Learning Modellen (Huang et al., 2004, S. 543-544). Die unterliegende Annahme ist hierbei, dass einige Finanzkennzahlen einen derartig signifikanten Einfluss auf die Bonitätseinstufung haben, dass ein Modell mit hinreichender Genauigkeit auf der Basis veröffentlichter Finanzinformationen entwickelt werden kann.

Bereits in den 1960er Jahren entwickelte statistische Methoden nutzten Regressionstechniken um das Rating von Anleihen vorherzusagen (Horrigan, 1966, S. 51-52). Diese Modelle konnten mit einer kleinen Menge an Finanzkennzahlen ungefähr zwei Drittel einer Stichprobe korrekt klassifizieren (Huang et al., 2004, S. 545). Ein Problem dieser Modelle ist jedoch, dass für multivariate Datensätze

von Finanzkennzahlen die Normalitätsannahme meist verletzt wird, was sie für die Anwendung in Stichproben ungültig macht. Im Gegensatz zu traditionellen statistischen Techniken extrahieren Machine Learning Modelle Beziehungen zwischen einzelnen Datenpunkten (Huang et al., 2004, S. 545). Während traditionelle, statistische Regressionstechniken von einem linearen Aufbau eines Datensatzes abhängig sind, können Machine Learning Modelle nicht-lineare Strukturen des Modells durch die Daten erfassen.

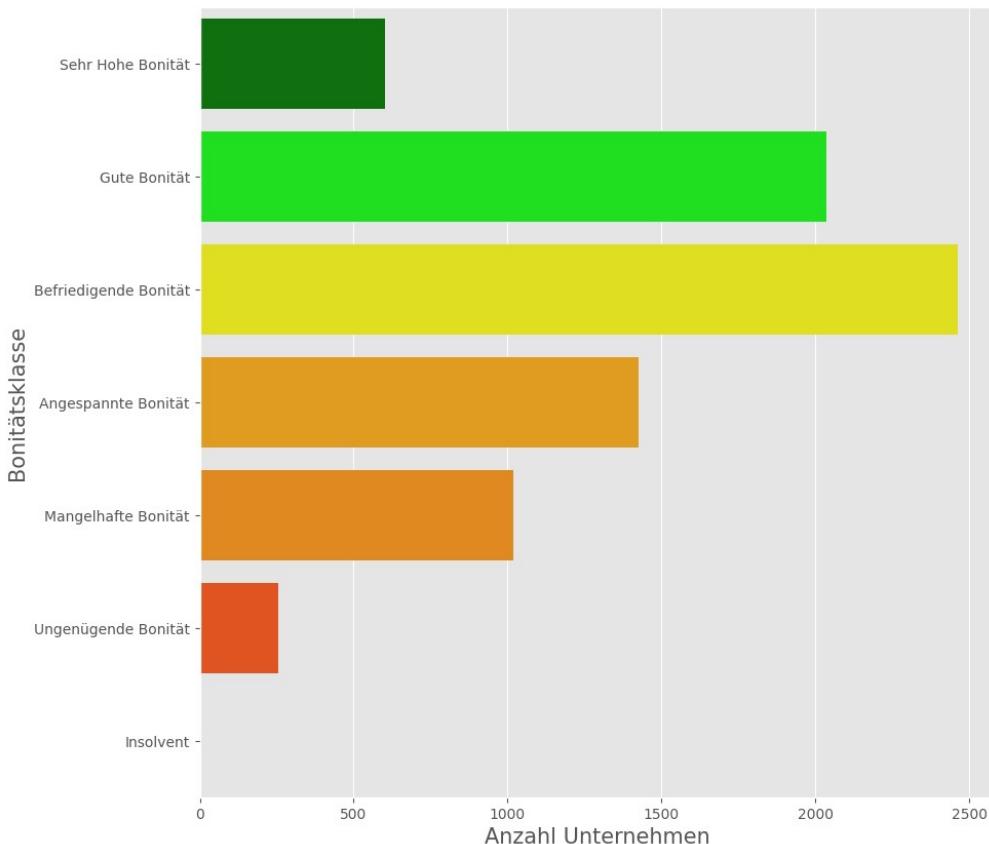
Die im Rahmen dieser Arbeit vorgestellten Machine-Learning Techniken fallen in die Kategorie des überwachten Lernens. Hierbei wird das Ergebnis jedes Datensatzes als Zielvariable für das Training des Modells verwendet (Andrae, 2023, S. 8). Ein Algorithmus des überwachten Machine Learnings zielt darauf ab, anhand der in das Modell eingeführten Kohortendaten deren Output vorhersagen zu können. Hierzu wird zunächst der Algorithmus des überwachten Lernens mit der Kohorten-Matrix und der Zielvariable gefüttert. Der Algorithmus versucht, basierend auf dem ihm bereits bekannten Kategorienset, den soeben eingespeisten, unbekannten Datensatz einer entsprechenden Kategorie zuzuordnen (Andrae, 2023, S. 10).

Zur Erstellung der Machine Learning Modelle im Rahmen dieser Arbeit wird die Python Programmiersprache verwendet. Python ist eine hochentwickelte, interpretierte Programmiersprache, welche die Anwendung verschiedener Programmierparadigmen ermöglicht und daher häufig für die Erstellung von serverseitigen Anwendungen verwendet wird (Sheetal & Partibha, 2014, S. 78). Eine der wichtigsten Aspekte der Sprache ist ihre vielseitige Anwendbarkeit im Bereich von Machine Learning und künstlicher Intelligenz, da Bibliotheken von Drittanbietern nicht nur Machine Learning Frameworks bereitstellen, sondern zusätzlich Werkzeuge zur Datenaufbereitung, Datenverarbeitung und -visualisierung.

3.2.1 Testdatensatz zur Entwicklung der Machine Learning Modelle

Für die Entwicklung der Machine Learning Modelle wird ein Datensatz der Website Kaggle verwendet, der 7805 Ratingdatensätze inklusive entscheidender Finanzkennzahlen zum Zeitpunkt der Erstellung der Bewertung enthält.

Abb. 9 Verteilung der Bonitätseinstufungen im Testdatensatz



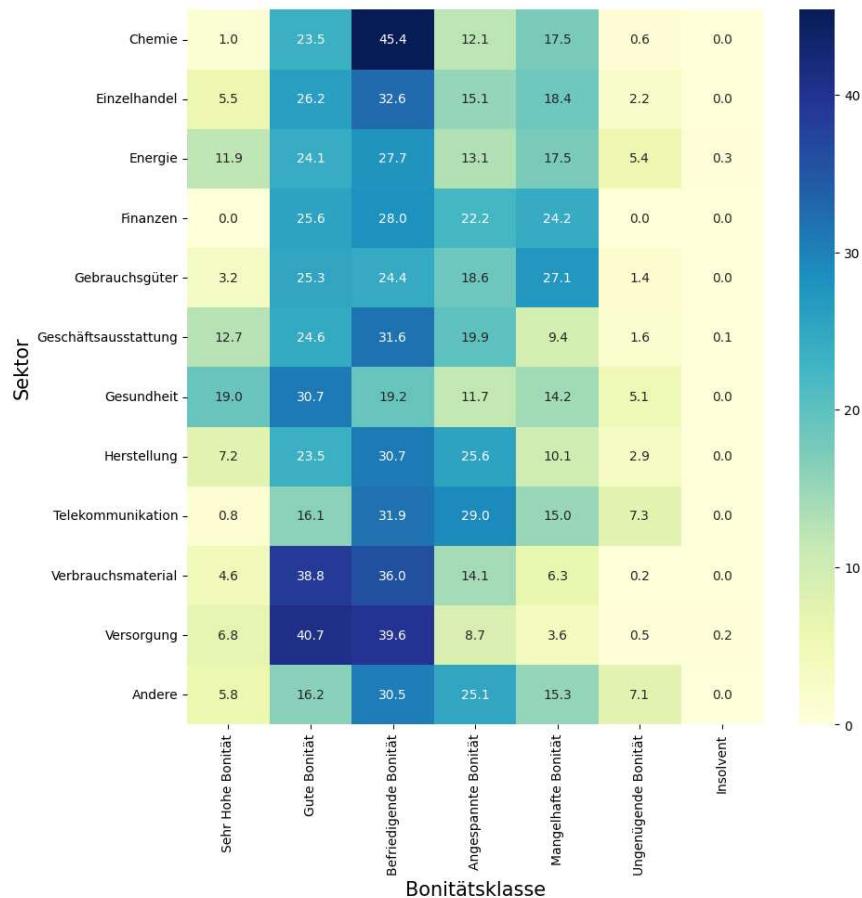
Quelle: Eigene Darstellung.

Da Machine Learning Modelle für einen optimalen Lernzprozess möglichst viele Datensätze in jeder Klasse benötigen, werden die Ratings in sieben Kategorien zusammengefasst (Abbildung 9). Unternehmen mit einer Bewertung zwischen AAA und AA- fallen in die Kategorie „Sehr Hohe Bonität“. Ratings zwischen A+ und A- fallen in die Kategorie „Gute Bonität“, während Ratings zwischen BBB+ und BBB- in die Klasse „Befriedigende Bonität“ fallen. „Angespannte Bonität“ gilt für die Ratingklassen von BB+ bis BB-, während Unternehmen, die ein Rating zwischen B+ und B- erhalten haben, eine mangelhafte Bonität aufweisen. Unternehmen, deren Bonität mit CCC+ oder schlechter bewertet wurde, fallen in die Kategorie „Ungenügende Bonität“. „Insolvent“ sind Unternehmen der Ratingklasse D, die jedoch aufgrund ihres äußerst niedrigen Bestands im Datensatz nicht für das Training berücksichtigt werden können.

Der verwendete Datensatz umfasst Bonitätseinstufungen von sieben Ratingagenturen: Egan-Jones Ratings Company, S&P, Moody's, Fitch, DBRS, Japan Credit Rating Agency und HR Ratings de Mexico. Obwohl zwischen Bewertungen einzelner Ratingagenturen minimale Differenzen auftreten können, zeigen die Bonitätseinstufungen unterschiedlicher Ratingagenturen einen Korrelationskoeffizient von 99% (Berg et al., 2022, S. 1320). Da die Ratings zusätzlich in sechs Klassen zusammengefasst werden, ist der Effekt dieser minimalen Abweichungen vernachlässigbar.

Darüber hinaus sind im Datensatz die Sektoren der einzelnen Unternehmen erfasst, was eine genaue Analyse der prozentualen Ratingklassenverteilung innerhalb der Sektoren ermöglicht. Die Erstellung einer Heatmap ermöglicht hierbei die zweidimensionale Datenanalyse (Abbildung 10).

Abb. 10 Prozentuale Verteilung der Bonitätsklassen innerhalb einzelner Sektoren



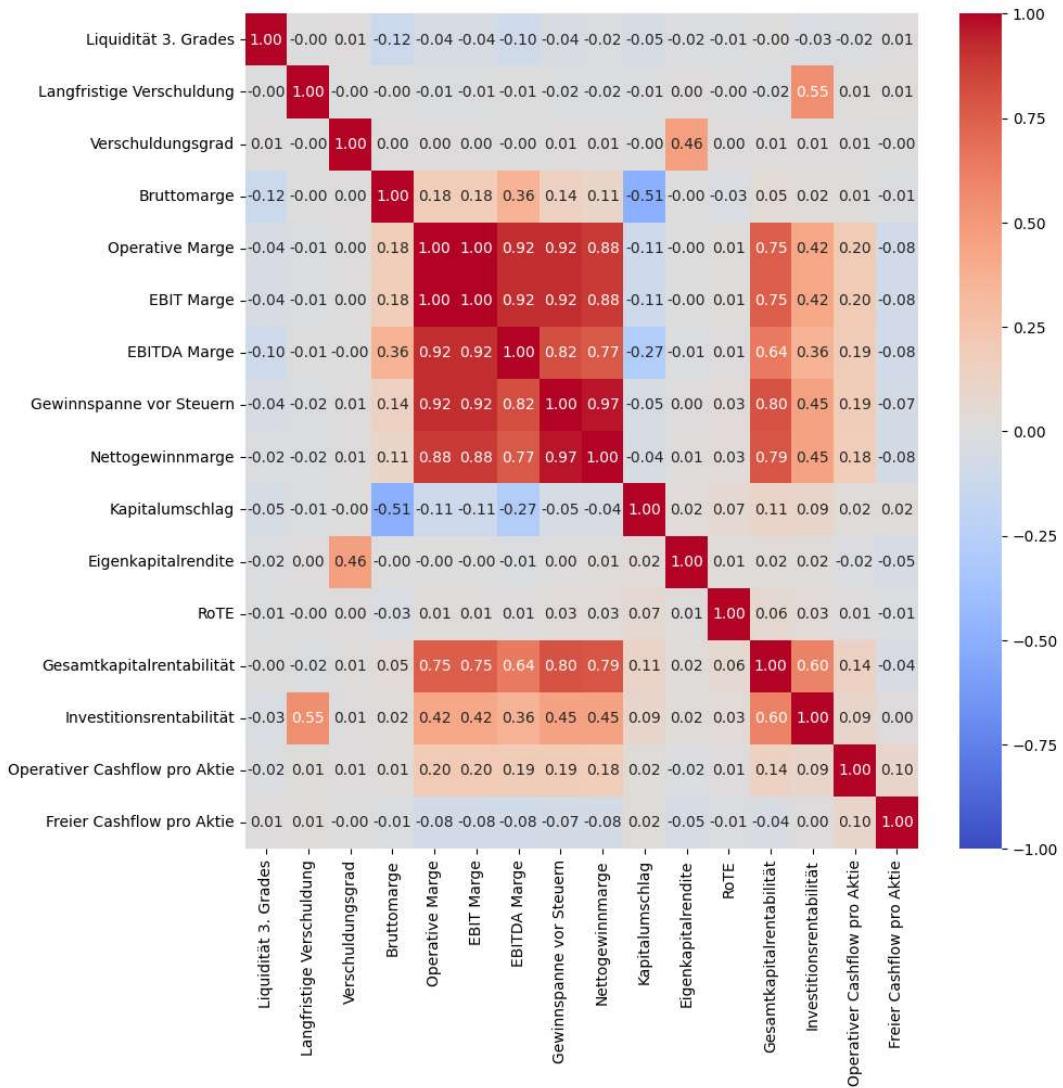
Quelle: Eigene Darstellung.

Auffällig ist hierbei zunächst, dass mit Ausnahme der Sektoren Geschäftsausstattung, Herstellung, Verbrauchsmaterial und Versorgung ein signifikanter Teil der Ratings mangelhafte Bonität ausweisen. Der Datensatz enthält zudem fast keine Unternehmen, die bereits insolvent sind und damit in die Kategorie D fallen, während die beiden höchsten Ratingkategorien „Höchste Bonität“ sowie „Sehr Hohe Bonität“ ebenfalls unterrepräsentiert sind.

Der Testdatensatz enthält neben dem Rating und Sektor insgesamt 16 Finanzkennzahlen für die gelisteten Unternehmen. Da gemäß Kaur nicht alle der Kennzahlen Einfluss auf die Bonitätseinstufung haben, müssen zunächst die für Bonitätseinstufungen relevanten Kennzahlen ausgewählt werden (Kaur et al., 2023, S. 94-103). Als Profitabilitätskennzahlen werden zunächst die Kennzahlen Gesamtkapitalrentabilität, operative Marge, Nettogewinnmarge, Eigenkapitalrendite sowie Bruttomarge verwendet. Die Verschuldung des Unternehmens wird durch den langfristigen Verschuldungsgrad sowie dem allgemeinen Verschuldungsgrad innerhalb des Modells repräsentiert. Die Liquidität wird durch die Liquidität 3. Grades repräsentiert. Cashflows werden durch den

operativen Cashflow je Aktie sowie dem freien Cashflow je Aktie im Modell berücksichtigt. Der Kapitalumschlag ist eine Aktivitätskennzahl, die angibt, wie effizient die Vermögenswerte im Unternehmen zur Generierung von Umsatz eingesetzt werden, indem die Besitzwerte des Unternehmens in Relation zum Umsatz aus Verkäufen gesetzt werden (Utami et al., 2017, S. 29). Höhere Aktivitätskennzahlen bedeuten ein geringeres Risiko und wirken sich damit direkt auf die Bonitätseinstufung des Unternehmens aus. Aus diesem Grund findet der Kapitalumschlag Berücksichtigung.

Abb. 11 Korrelationsmatrix der Finanzkennzahlen im Testdatensatz



Quelle: Eigene Darstellung.

Darüber hinaus ist die Korrelation der einzelnen Finanzkennzahlen zueinander zu berücksichtigen (Abbildung 11). Sind zwei Kennzahlen miteinander hoch korreliert, sollten nicht beide berücksichtigt werden, da dies sowohl zu unnötigem Ressourcenverbrauch als auch zu Overfitting im Machine Learning Modell führen kann.

Hierbei ist auffällig, dass die Kennzahlen Gesamtkapitalrentabilität, Operative Marge, EBIT Marge, EBITDA Marge, die Vorsteuer-Gewinnspanne sowie die Nettogewinnmarge einen

Korrelationskoeffizienten nahe eins ausweisen. Da diese Kennzahlen alle in die Kategorie der Profitabilitätskennzahlen fallen, ist dies zunächst zu erwarten. Beim Training des Modells wird jedoch mit der EBIT-Marge lediglich eine der Kennzahlen berücksichtigt.

3.2.2 Multinomiale Logistische Regression als Benchmarkmodell

Ein multinomiales logistisches Regressionsmodell ist eine Erweiterung des traditionellen logistischen Regressionsmodells (El-Habil, 2012, S. 272-273). Logistische Regression wird eingesetzt, wenn es sich bei der Zielvariable des Modells um eine kategoriale Kennzahl handelt, die ein binäres Messniveau verwendet (Bender et al., 2007, S. 33). Da die Bonität eines beliebigen Unternehmens im Rahmen dieser Auswertung sechs Ausprägungen annehmen kann, ist das einfache logistische Regressionsmodell in einer Art und Weise zu erweitern, dass Logit-Wahrscheinlichkeiten über sechs verschiedene Klassen ausgegeben werden können. Hierfür wird die Zielvariable Bonitätseinstufung, die sechs verschiedene Ausprägungen annehmen kann, in fünf verschiedene Dummy-Variablen kodiert (Bayaga, 2010, S. 289). Jede Dummy-Variable nimmt einen Wert von 1 für ihre eigene Kategorie an, alle anderen Variablen nehmen den Wert 0 an. Mithilfe dieses Aufbaus kann anschließend für jede Ausprägung der Zielvariable ein eigenes binäres logistisches Regressionsmodell entwickelt werden.

Ähnlich zu Machine-Learning Modellen kann ein multinomiales logistisches Regressionsmodell mithilfe der Bibliothek SciKit-Learn in Python entwickelt werden. Hierzu wird der Datensatz zunächst in einen Trainings- und einen Testdatensatz unterteilt. Mithilfe des One-Hot-Encoders kann die Zielvariable *y* in die verschiedenen Dummy-Variablen kodiert werden. Zur Entwicklung des Regressionsmodells kann anschließend durch Anwendung der Python-Funktion „Argmax“ auf der horizontalen Achse die Kategorie der Zielvariablen ermittelt werden. Ist das Regressionsmodell entwickelt, wird die Präzision des Modells mithilfe der zuvor ungesiehten Testdaten ermittelt.

Abb. 12 Entwicklung des multinomialen logistischen Regressionsmodells als Benchmark

```
regression = LogisticRegression(max_iter=10000, multi_class='multinomial', solver='lbfgs')

regression.fit(X_train, y_train.argmax(axis=1))

y_pred = regression.predict(X_test)

accuracy = accuracy_score(y_test.argmax(axis=1), y_pred)

print(f'Die Testgenauigkeit des Modells ist: {round(accuracy*100, 4)}%')
✓ 5.7s
```

Quelle: Eigene Darstellung.

Das Benchmarkmodell erreicht für die Prognose der Bonitätseinstufungen eine gewichtete Präzision von 35,23% (Abbildung 13). Dies bedeutet, dass 35,23% der prognostizierten positiven Fälle korrekt identifiziert werden. Die Präzision einer Ratingklasse wird über die Formel $TP/(TP+FP)$ ermittelt (Gray et al., 2011, S. 2). Im Rahmen der in dieser Arbeit vorgestellten Modelle wird die Präzision als wichtigste Bewertungsmautrik verwendet, da die Richtigkeit positiver Vorhersagen für jede Klasse

besonders wichtig ist. Die Sensitivität, auch Recall genannt, setzt alle korrekt identifizierten, positiven Fälle ins Verhältnis zu allen tatsächlich positiven Fällen und wird daher über die Formel $TP/(TP+FN)$ ermittelt. Im Fall der Kategorie „Befriedigende Bonität“ bedeutet dies, dass 67,71% der Unternehmen, die im Testdatensatz tatsächlich in die Kategorie „Befriedigende Bonität“ fallen, korrekt klassifiziert wurden. Ein Blick auf den Klassifikationsbericht offenbart hierbei, dass das Regressionsmodell stark dazu tendiert, Datensätze aller Klassen in die mittleren Bonitätsabstufungen einzuordnen. Damit erreicht es für die Klasse „Befriedigende Bonität“ eine Sensitivität von über 67%, jedoch über alle Klassen hinweg eine Präzision von 44% oder schlechter. In die Klasse „Sehr Hohe Bonität“ wurden durch das Regressionsmodell keine Datensätze eingeordnet.

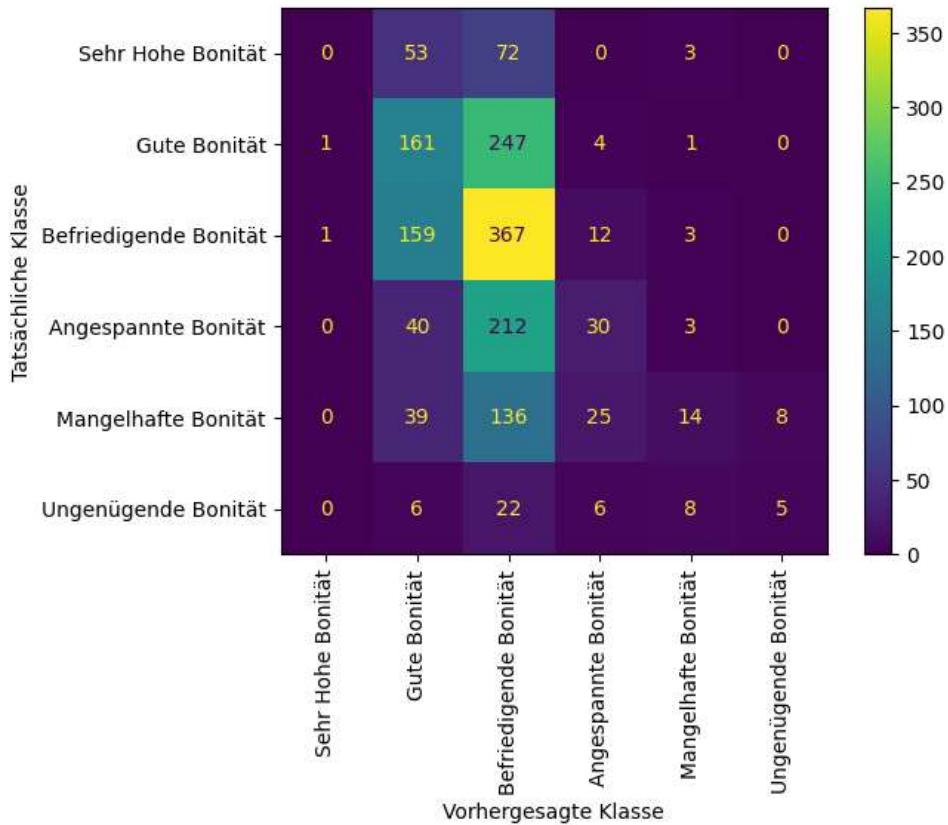
Abb. 13 Klassifikationsbericht des multinomialen Regressionsmodells

	Präzision	Sensitivität	F1-Score	Testdatensätze
Sehr Hohe Bonität	0.0000	0.0000	0.0000	128
Gute Bonität	0.3515	0.3889	0.3693	414
Befriedigende Bonität	0.3475	0.6771	0.4593	542
Angespannte Bonität	0.3896	0.1053	0.1657	285
Mangelhafte Bonität	0.4375	0.0631	0.1102	222
Ungenügende Bonität	0.3846	0.1064	0.1667	47
Genauigkeit	0.3523	0.3523	0.3523	0
Ungewichteter Durchschnitt	0.3185	0.2235	0.2119	1638
Gewichteter Durchschnitt	0.3420	0.3523	0.2939	1638

Quelle: Eigene Darstellung.

Der F1-Score kombiniert die Sensitivität und die Präzision und wird über die Formel $(2 * \text{Präzision} * \text{Sensitivität}) / (\text{Präzision} + \text{Sensitivität})$ errechnet (Gray et al., 2011, S. 2). Aus diesem Grund wird er häufig für eine vereinfachte Quantifizierung der Performance von Klassifizierungsmodellen verwendet. Darüber hinaus gibt der Klassifikationsbericht eine Zeile mit der Beschreibung „Genauigkeit“ aus. Die Genauigkeit ist eine separate Kennzahl, die alle korrekten Vorhersagen des Modells ($TP+TN$) ins Verhältnis zu der Gesamtanzahl der Vorhersagen ($TP+TN+FP+FN$) setzt. Die Genauigkeit ist jedoch eine Kennzahl, die bei mehrklassigen Klassifizierungsproblemen nur beachtet werden sollte, wenn die Anzahl der Testdatensätze über die Klassen gleichmäßig verteilt ist. Die Kennzahl findet im Rahmen der Bonitätseinstufungsmodelle daher keine weitere Beachtung, da ein Modell, welches für jeden Datensatz die Kategorie „Befriedigende Bonität“ vorhersagt, durch eine schwache Prognoseleistung eine vergleichsweise hohe Genauigkeit erreichen könnte.

Abb. 14 Konfusionsmatrix des multinomialen Regressionsmodells



Quelle: Eigene Darstellung.

Die Klassifikationsmatrix (Abbildung 14) zeigt, wie die Unternehmen des Testdatensatzes durch die Vorhersagefunktion des Modells eingestuft wurden und bietet in Kombination mit der Klassifikationstabelle (Abbildung 13) einen tiefen Einblick in die Prognosefähigkeiten des Modells. Von den 542 Testdatensätzen der Klasse „Befriedigende Bonität“ wurden 367 korrekt eingestuft, jedoch wurde ein Datensatz der Klasse „Sehr Hohe Bonität“, 159 Datensätze der Klasse „Gute Bonität“, 12 Einträge der Klasse „Angespannte Bonität“ und drei Datensätze der Kategorie „Mangelhaft Bonität“ zugeordnet.

Ein möglicher Faktor der suboptimalen Leistung des Regressionsmodells ist, dass das Modell aufgrund seines linearen Aufbaus nicht die komplexen Zusammenhänge der Kovariaten erfassen kann. In traditionellen Regressionsmodellen werden die verschiedenen Klassen durch gerade Ebenen getrennt (Dreiseitl & Ohno-Machado, 2002, S. 354). Dies ist jedoch nur dann effektiv, wenn die Daten linear trennbar sind.

3.2.3 K-Nearest Neighbors Algorithmus zur Vorhersage von Bonitätseinstufungen

Das einfachste Machine Learning Modell ist der K-Nearest Neighbors (*Deutsch: nächste-Nachbarn*) Algorithmus, der neue Datenpunkte auf Basis ihrer Ähnlichkeit zu bereits existierenden Datenpunkten klassifiziert (Melisah & Muhathir, 2023, S. 26). Zunächst wird die Anzahl K der Nachbarn festgelegt (Rakshna et al., 2023, S. 899). Anschließend wird die euklidische Distanz zu den Nachbarn berechnet. Für jeden Datensatz wird die euklidische Distanz in aufsteigender

Reihenfolge geordnet und anschließend der Nachbar K mit der geringsten euklidischen Distanz gewählt. Der K-Nearest Neighbors Algorithmus wird im Zusammenhang mit der Saving-Matrix-Methode häufig für die optimale Berechnung von Lieferrouten verwendet (Lidiawati et al., 2023, S. 958-959).

Zunächst ist der Datensatz in eine Feature-Matrix sowie eine Zielvariable aufzuteilen. Die Feature-Matrix enthält alle relevanten Finanzkennzahlen, die innerhalb des Modells berücksichtigt werden (siehe Abschnitt 3.2.1). Darüber hinaus werden sowohl die Datensätze der Feature-Matrix als auch der Zielvariablen in einen Trainingsdatensatz für den Aufbau des Modells und einen Testdatensatz zur Evaluierung des Modells unterteilt. Über den Parameter „Random Seed“ kann sichergestellt werden, dass für jeden Anlauf die Daten gleich geteilt werden. Der letzte Schritt der Datenverarbeitung umfasst die Kodierung bzw. Skalierung der Datensätze. Die Daten der Zielvariable „Ratingkategorie“ werden mithilfe einer Label-Kodierung zu Zahlenwerten transformiert. Die Finanzkennzahlen durchlaufen eine standardisierte Skalierung, indem der Mittelwert der jeweiligen Spalte subtrahiert und anschließend durch die Standardabweichung dividiert wird.

Abb. 15 Testen von einem bis elf Nachbarn in Python

```
training_accuracy = []
test_accuracy = []

# Anzahl der zu testenden Nachbarn
neighbors_settings = range(1, 12)

for neighbors_number in neighbors_settings:
    # Nutzung der euklidischen Distanz zur Messung
    knn = KNeighborsClassifier(n_neighbors=neighbors_number, weights='distance', p=2)
    knn.fit(X_train, y_train)
    training_accuracy.append(knn.score(X_train, y_train))
    test_accuracy.append(knn.score(X_test, y_test))
```

Quelle: Eigene Darstellung.

Zum Aufbau des Modells wird die „KNeighborsClassifier“ Klasse der Bibliothek „scikit-learn“ verwendet. Um die optimale Anzahl der Nachbarn zu ermitteln, wird zunächst die Performance des Modells für ein bis elf Nachbarn getestet (Abbildung 15). Die Anzahl von Nachbarn, die beim Test die höchste Genauigkeit erzielen kann, wird zur Evaluierung des Modells verwendet.

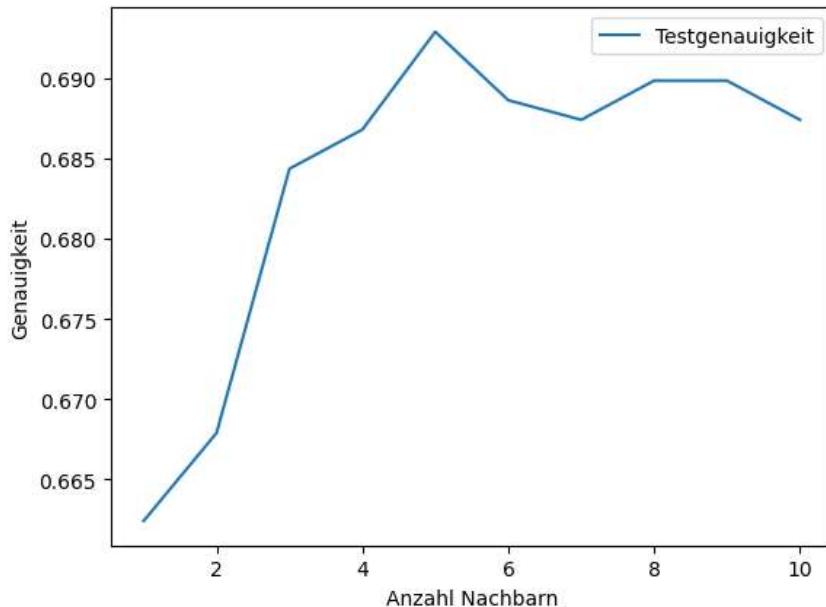
Abb. 16 Klassifikationsbericht des überarbeiteten KNN-Modells

	Präzision	Sensitivität	F1-Score	Testdatensätze
Sehr Hohe Bonität	0.5852	0.6172	0.6008	128
Gute Bonität	0.6719	0.7222	0.6962	414
Befriedigende Bonität	0.7450	0.7491	0.7470	542
Angespannte Bonität	0.6857	0.6737	0.6796	285
Mangelhafte Bonität	0.7000	0.6306	0.6635	222
Ungenügende Bonität	0.5758	0.4043	0.4750	47
Genauigkeit	0.6929	0.6929	0.6929	0
Ungewichteter Durchschnitt	0.6606	0.6328	0.6437	1638
Gewichteter Durchschnitt	0.6928	0.6929	0.6919	1638

Quelle: Eigene Darstellung.

Das Klassifizierungsmodell erreicht mit fünf Nachbarn für den Testdatensatz eine gewichtete Präzision von 69,28% (Abbildung 16). Der ungewichtete Durchschnitt der Präzision aller Ratingklassen liegt bei 66,06%.

Abb. 17 Genauigkeit der Klassifizierung des Testdatensatzes



Quelle: Eigene Darstellung.

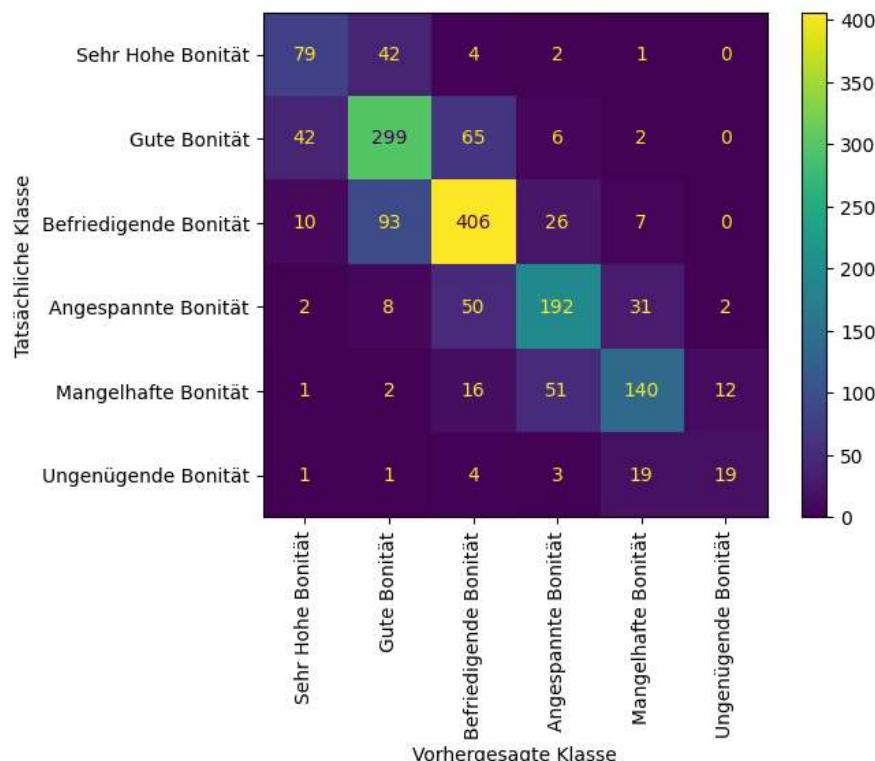
Wie bereits im Benchmarkmodell erreicht das KNN-Klassifizierungsmodell für die Ratingkategorie „Befriedigende Bonität“ die höchste Präzision (Tabelle 2). Das Modell konnte für alle Ratingkategorien eine Präzision von über 50% erzielen.

Auffällig ist, dass die besseren Ratingkategorien eine höhere Sensitivität als Präzision aufweisen, während bei den niedrigeren Kategorien die Präzision höher als die Sensitivität ist. Besonders die Ratingkategorie „Ungenügende Bonität“ zeigt mit 40,43% eine schwache Sensitivität. In diesem Fall ist es wichtig zu wissen, in welche Kategorien die falsch Klassifizierten Unternehmen der Kategorie

„Ungenügende Bonität“ eingestuft wurden, da eine Einordnung ausfallender Unternehmen in Investment-Grade Kategorien besonders kostspielig ist. Die Konfusionsmatrix zeigt, dass im Modell 81% der Unternehmen, die der Kategorie „Ungenügende Bonität“ angehören, in die Kategorien „Mangelhafte Bonität“ oder „Ungenügende Bonität“ eingeordnet wurden (38 von 47), jedoch auch sechs der 47 Unternehmen Investment-Grade Kategorien zugeordnet wurden (Abbildung 18).

Von den Unternehmen, die im Testdatensatz in die Kategorie „Sehr Hohe Bonität“ fallen, stuft das Modell 95% der Datensätze in die Kategorien „Sehr Hohe Bonität“ oder „Gute Bonität“ ein. Dies zeigt, dass dieses K-Nearest-Neighbors Modell bei finanziell gesünderen Unternehmen eine bessere Klassifikationsperformance zeigt als bei angeschlagenen Unternehmen.

Abb. 18 Konfusionsmatrix des Klassifikationsmodells



Quelle: Eigene Darstellung.

Auffällig ist zudem, dass mit Ausnahme der Kategorie „Gute Bonität“ das Modell dazu tendiert, Unternehmen besser zu bewerten, wenn die richtige Kategorie nicht vorhergesagt werden konnte.

Da das Modell für die Ratingkategorie „Befriedigende Bonität“ bereits die höchste Präzision und Sensitivität aufzeigt, erreicht es für diese Kategorie den höchsten F1-Score, während das Klassifizierungsmodell für die Kategorie „Ungenügende Bonität“ als einzige Kategorie unterhalb der 50%-Marke fällt und damit nicht zufriedenstellende Ergebnisse erreicht.

3.2.4 Nutzung von Support Vector Machines zur Vorhersage von Bonitätseinstufungen

Support Vector Machines (SVM) wurden erstmals in den neunziger Jahren vorgestellt und basieren auf dem Prinzip der strukturellen Risikominimierung (Vapnik, 1999, S. 138). Diese erstellen zunächst

ein nicht-lineares Mapping der Eingangsvektoren X in einen hochdimensionalen Raum Z. In diesem Raum wird anschließend durch nicht-lineare Optimierung eine optimale Hyperebene konstruiert, welche die Eingangsvektoren X optimal vom hochdimensionalen Raum Z trennt (Abbildung 19).

Abb. 19 Optimierung der Hyperebene im mehrdimensionalen Raum

$$\begin{aligned} \min_{w,b} & \langle w, w \rangle \\ \text{s.t. } & y_i(\langle w, \phi(x_i) \rangle + b) \geq 1, \\ & i = 1, \dots, l \end{aligned}$$

Quelle: Übernommen aus Huang et al., 2004, S. 550.

Die grundlegende Annahme von Support Vector Machines liegt in der Existenz einer Funktion $Y=f(X)$ (Huang et al., 2004, S. 549). Die Klassifikationsaufgabe besteht aus der Konstruktion einer heuristischen Funktion $h(X)$, sodass die Funktion h mithilfe einer erfolgreichen Vorhersage von Y zur Funktion f führt. Support Vector Machines sind, wie bereits der K-Nearest-Neighbor Algorithmus, für binäre Klassifizierungsprobleme (Ausfallwahrscheinlichkeit) und mehrklassige Klassifizierungsprobleme einsetzbar.

Ein erstes Support Vector Machine Modell, welches sowohl die gleichen Daten als auch dieselbe Standardskalierung der Feature-Matrix verwendet, erreicht für den Testdatensatz eine Genauigkeit von lediglich 41% (Abbildung 20). Ähnlich zum K-Neighbors-Classifier ermöglicht die Bibliothek „sklearn“ zahlreiche Methoden, um die Performance verschiedener Modelle zu testen. Support Vector Machines kennen vier unterschiedliche Kernel-Funktionen, um die Datenpunkte X in den hochdimensionalen Raum Z zu projizieren: Linear, Radiale Basis, Polynomiell und Sigmoid (Markowetz, 2003, S. 28).

Abb. 20 Erster Lauf mit einfacher SVM

```
svc1 = svm.SVC(random_state=42)
svc1.fit(X_train, y_train)

print(f"Genauigkeit des Trainingsdatensatzes: {svc1.score(X_train, y_train).round(2)}")
print(f"Genauigkeit des Testdatensatzes: {svc1.score(X_test, y_test).round(2)}")
✓ 3.7s

Genauigkeit des Trainingsdatensatzes: 0.44
Genauigkeit des Testdatensatzes: 0.41
```

Quelle: Eigene Darstellung.

In einem Test der vier Kernelfunktionen erreichte der Radiale Basis-Kernel die höchste Genauigkeit mit 65,07% (Tabelle 3). Der Strafparameter C wurde für alle Support Vector Machines auf eins festgelegt. Zur Entscheidung wurde eine „one-vs-one“ Funktion gewählt (Sklearn.Svm.SVC, n. d.).

Tab. 2 Ergebnisse des Genauigkeitstests der verschiedenen Kernelfunktionen

Kernelfunktion	Genauigkeit
Linear	34,07%
Polynomiell	34,86%
Radial Basis	65,07%
Sigmoid	24,36%

Quelle: Eigene Darstellung.

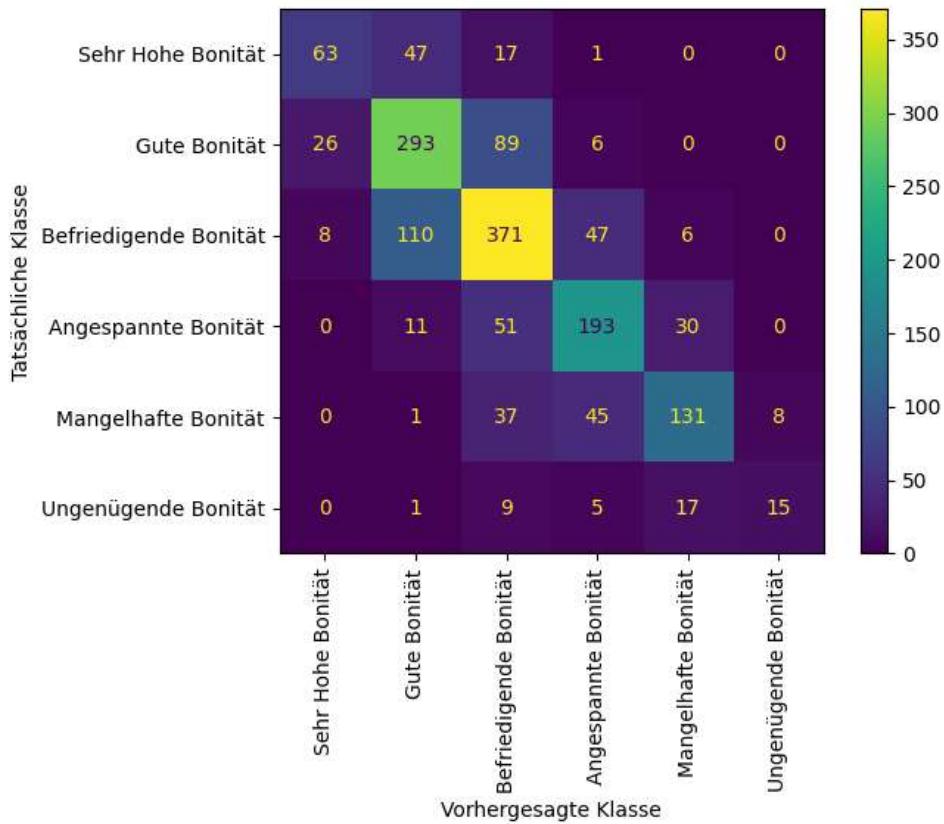
Obwohl der explizite Einsatz einer Radial-Basis Kernelfunktion die Genauigkeit des SVM-Klassifizierungsmodells erhöht hat, bleibt die Performance hinter dem K-Nearest-Neighbors Klassifizierungsmodell zurück. Der Klassifizierungsbericht offenbart, dass das Modell über alle Ratingkategorien hinweg eine gute Präzision zeigt, die Sensitivität jedoch besonders in Klassen mit wenigen Testdatensätzen mangelhaft ist (Abbildung 21).

Abb. 21 Klassifizierungsbericht der Radial-Basis Support Vector Machine

	Präzision	Sensitivität	F1-Score	Testdatensätze
Sehr Hohe Bonität	0.6495	0.4922	0.5600	128
Gute Bonität	0.6328	0.7077	0.6682	414
Befriedigende Bonität	0.6463	0.6845	0.6649	542
Angespannte Bonität	0.6498	0.6772	0.6632	285
Mangelhafte Bonität	0.7120	0.5901	0.6453	222
Ungenügende Bonität	0.6522	0.3191	0.4286	47
Genauigkeit	0.6508	0.6508	0.6508	0
Ungewichteter Durchschnitt	0.6571	0.5785	0.6050	1638
Gewichteter Durchschnitt	0.6528	0.6508	0.6478	1638

Quelle: Eigene Darstellung.

Abb. 22 Konfusionsmatrix der Radial-Basis Support Vector Machine



Quelle: Eigene Darstellung.

3.2.5 Extreme Gradient Boosting zur Vorhersage von Bonitätseinstufungen

In den vergangenen Jahren hat die Nutzung von Gradient Boosting in der praktischen Anwendung von Machine Learning Modellen stark zugenommen (Chen & Guestrin, 2016, S. 785-786). XGBoost ist eine open-source Bibliothek mit dem Ziel, eine effiziente und skalierbare Implementierung von Gradient Boosting Algorithmen zu ermöglichen. XGBoost kann für Klassifikations- und Vorhersagemodelle verwendet werden, alleinstehend oder in Kombination mit Deep Learning Netzen.

Entscheidungsbäume sind eine der einfacheren Machine Learning Algorithmen, da sie lediglich mit Bedingungen arbeiten. XGBoost verwendet Regressionsbäume mit dem Spitznamen „CART“, die Klassifikation- und Entscheidung zugleich treffen (Chen & Guestrin, 2016, S. 786-787). Die CART-Blätter des Regressionsbaums enthalten nicht nur die finalen Entscheidungswerte, sondern gleichzeitig einen kontinuierlichen Score, der zu ihrer Evaluierung verwendet wird. Die Evaluierung der einzelnen Modelle wird mit Boosting durchgeführt. Boosting ist eine sequenzielle Art des Ensemble Learnings, bei der das Ergebnis des vorherigen Modells als Input des Folgemodells verwendet wird. Dies führt zu einer erheblichen Beschleunigung des Trainings, da die einzelnen Modelle nicht getrennt voneinander trainiert werden, sondern jedes Modell lediglich die Fehler des vorherigen Modells korrigiert. Das Modell verwendet Gewichte, um durch die Iterationen falsche Vorhersagen des Modells stärker zu gewichten als korrekte Vorhersagen.

Gradient Boosting verwendet einen iterativen Gradienten-Abstiegs-Optimierungsalgorithmus, der kontinuierlich eine vorab definierte Verlustfunktion, beispielsweise die Wurzel der mittleren quadratische Abweichung (RMSE), optimiert (Wisesa et al., 2021, S. 104-105). Im Fall des vorliegenden Modells quantifiziert die Verlustfunktion die Distanz der Vorhersage von der tatsächlichen Ratingkategorie der Unternehmen.

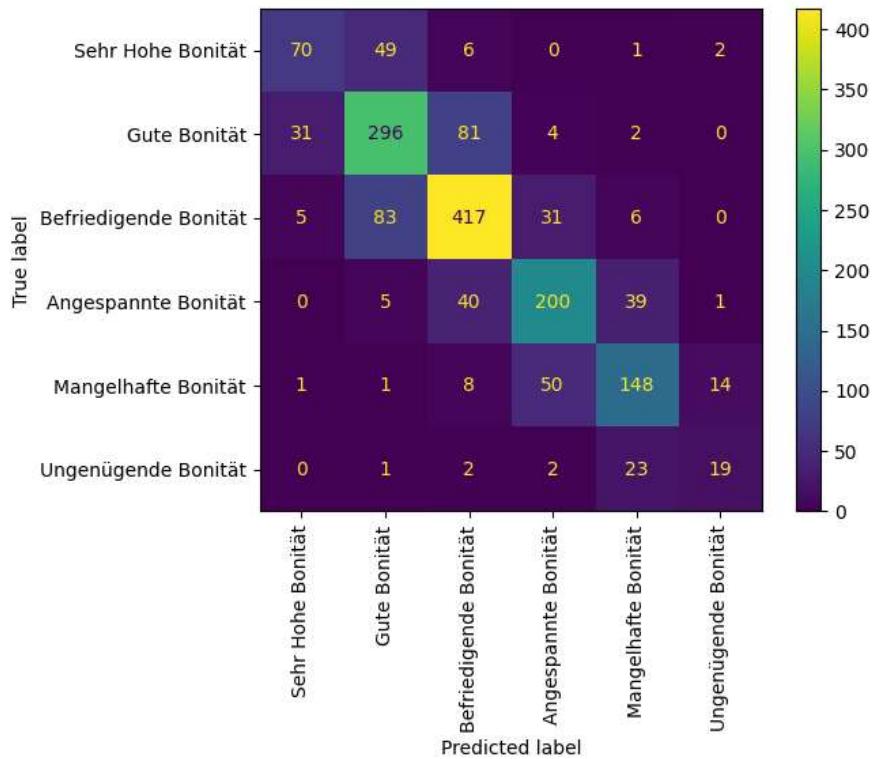
Ein weiterer Vorteil des Extreme Gradient Boosting Modells ist die Möglichkeit, Modelle auf einer NVIDIA Grafikkarte zu berechnen sowie andere Python-Bibliotheken in das Modell einzubinden (Abbildung 25). Dies ist insbesondere für die Optimierung der Hyperparameter des Modells hilfreich, da hierbei das Klassifizierungsmodell tausendfach ausgeführt wird, um die optimale Abstimmung zu finden. NVIDIA Grafikkarten können mithilfe ihrer CUDA-Treiber die Ausführung von Machine Learning Modellen um den Faktor zehn beschleunigen (Chollet, 2021, S. 65-567).

Abb. 23 Klassifikationsbericht des XGBoost-Modells

	Präzision	Sensitivität	F1-Score	Testdatensätze
Sehr Hohe Bonität	0.6542	0.5469	0.5957	128
Gute Bonität	0.6805	0.7150	0.6973	414
Befriedigende Bonität	0.7527	0.7694	0.7609	542
Angespannte Bonität	0.6969	0.7018	0.6993	285
Mangelhafte Bonität	0.6758	0.6667	0.6712	222
Ungenügende Bonität	0.5278	0.4043	0.4578	47
Genauigkeit	0.7021	0.7021	0.7021	0
Ungewichteter Durchschnitt	0.6646	0.6340	0.6471	1638
Gewichteter Durchschnitt	0.7002	0.7021	0.7004	1638

Quelle: Eigene Darstellung.

Abb. 24 Konfusionsmatrix des XGBoost-Modells



Quelle: Eigene Darstellung.

Das XGBoost-Klassifizierungsmodell erreicht eine Genauigkeit von 68,82%. Die Bibliothek ermöglicht zudem eine grafische Darstellung der Variablen der Feature-Matrix, die den höchsten Einfluss auf das Modell haben (Abbildung 26). Hierfür zählt das Modell, wie oft jede der Variablen der Feature-Matrix in den Entscheidungsbäumen vorkommt. Das Modell erstellt den F-Score (Frequency Score) indem es das Vorkommen einer Variablen auf den verschiedenen Entscheidungsbäumen zählt. Das Modell zeigt, dass die Liquidität dritten Grades der wichtigste Indikator für die Klassifizierung der Unternehmen im Gradient Tree Boosting Modell ist, dicht gefolgt von der Rohertragsmarge. Der Industriesektor, indem das Unternehmen operiert, zeigt mit Abstand die geringste Wichtigkeit.

Abb. 25 XGBoost Machine Learning Modell mit Hyperparameter-Tuning der Bibliothek „sklearn“ und Ausführung auf der GPU

```

params = {
    'max_depth': [3, 5, 7, 10],
    'learning_rate': [0.01, 0.1, 0.2, 0.3],
    'n_estimators': [50, 100, 200],
    'gamma': [0, 0.5, 1],
    'subsample': [0.8, 1],
    'colsample_bytree': [0.3, 0.5, 0.8, 1]
}

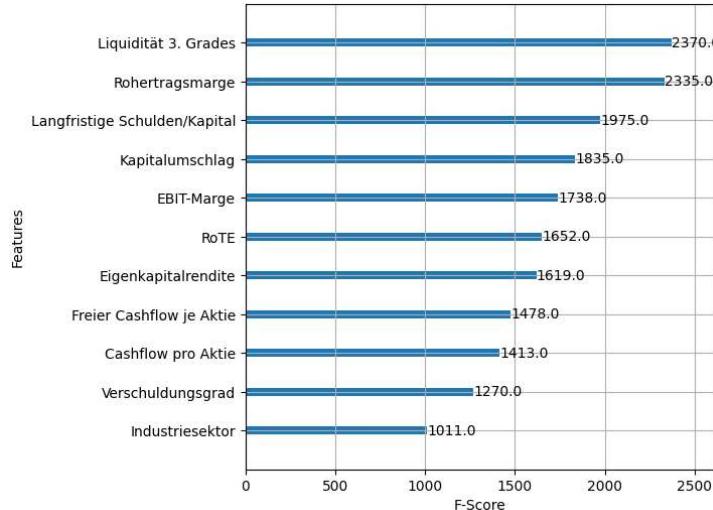
XGB_Model = xgb.XGBClassifier(objective='multi:softmax', num_class=6, random_state=42, gpu_id=0, tree_method='gpu_hist')

clf = GridSearchCV(estimator=XGB_Model,
                    param_grid=params,
                    cv=3,
                    scoring='accuracy',
                    verbose=3)
clf.fit(X_train, y_train)

```

Quelle: Eigene Darstellung.

Abb. 26 Wichtigkeit der einzelnen Features nach ihrer Nutzungs frequenz



Quelle: Eigene Darstellung.

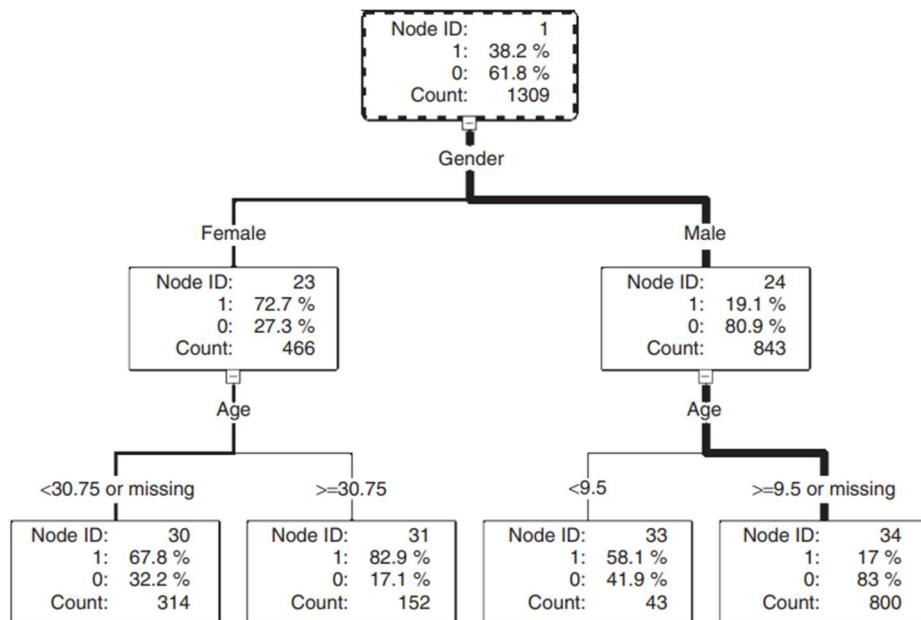
Auffällig ist, dass das Modell zwar eine höhere gewichtete Präzision über alle Ratingkategorien hinweg erreicht als das Support Vector Machine Modell, die Einstufungen in Kategorien mit weniger Testdatensätzen jedoch eine geringere Präzision zeigen. So wurden zwei Unternehmen des Testdatensatzes, die in die Ratingkategorie „Sehr Hohe Bonität“ fallen, im Klassifizierungsmodell der Kategorie „Ungenügende Bonität“ zugeordnet. Aufgrund dieser erhöhten Quote an falschen, positiven Einordnungen fällt das Gradient Boosting Modell in seiner ungewichteten Präzision hinter das Support Vector Machine Modell zurück. Die ungewichtete Sensitivität des Modells liegt zwischen dem K-Nearest-Neighbor und Support Vector Machine Modell.

3.2.6 Random Forest Klassifizierungsmodell zur Vorhersage von Bonitätseinstufungen

Random Forest Klassifizierungsmodelle sind ein grundlegender Teil der Extreme Gradient Boosting (XGBoost) Modelle, jedoch einfacher aufgebaut und damit in der Anwendung leicht zu optimieren,

da weniger Hyperparameter für die Optimierung zur Verfügung stehen (Rodriguez-Galiano et al., 2012, S. 95-96). Der Klassifikator wurde ursprünglich von Breiman entworfen und bot viele Vorteile gegenüber bestehenden Klassifizierungsmodellen, wie die Möglichkeit, tausende Variablen zu verarbeiten sowie effizient große Datenmengen klassifizieren zu können. Darüber hinaus reagiert das Modell unempfindlich gegenüber Ausreißern und ist strukturiert aufgebaut, wodurch Parallelisierung ermöglicht wird (Breiman, 2001, S. 10).

Abb. 27 Analyse der Sterberate des Titanic Unfalls mithilfe eines Entscheidungsbaums



Quelle: Übernommen aus De Ville, 2013, S. 449.

Entscheidungsbäume sind der Grundbaustein des Random Forest Klassifizierungsmodells. Abbildung 27 zeigt einen Klassifizierungs-Entscheidungsbaum zur Sterberate von Passagieren während des Titanic Unglücks (De Ville, 2013, S. 449). Der oberste Knoten ist der Wurzelknoten, der die globale Verteilung ausgibt. Dieser zeigt, dass von 1309 Passagieren des Schiffs lediglich 38,2% (500 Passagiere) überlebt haben. Dieser Wurzelknoten entfaltet sich anschließend in die beiden absteigenden Knoten „Männlich“ und „Weiblich“, indem die Grundgesamtheit aller Passagiere nach Geschlecht aufgeteilt wird. Die erste Stufe des Entscheidungsbaums zeigt, dass 72,7% aller weiblichen Passagiere das Unglück überlebt haben, sich jedoch lediglich 19,1% der männlichen Passagiere sich retten konnten. Bei der Erstellung eines Entscheidungsbaums ist es üblich, das Feature zuerst auszuwählen, dass die größte Unterteilung in der Variabilität der absteigenden Knoten erstellt.

Der unterste Knotenpunkt erstellt eine weitere Unterteilung männlicher und weiblicher Passagiere nach Alter. Hierbei ist auffällig, dass ältere, weibliche Passagiere eine höhere Überlebensrate zeigen als jüngere, weibliche Personen, unter männlichen Passagieren jedoch das Gegenteil zu beobachten ist (Abbildung 27), ein direktes Resultat der sozialen Dynamik während eines Seeunglücks („Frauen und Kinder zuerst!“). Weiterhin ist zu beobachten, dass die

Altersschnittpunkte für männliche und weibliche Passagiere nicht gleich sind. Entscheidungsbaum-Modelle verwenden Suchalgorithmen, um die drastischsten Schnittpunkte zwischen den jeweiligen Zweigen zu finden (De Ville, 2013, S. 451).

Das Random Forest Klassifizierungsmodell nutzt eine große Anzahl dieser Entscheidungsbäume, die innerhalb des Modells als Ensemble fungieren (Rodriguez-Galiano et al., 2012, S. 96). Jeder Baum des Modells gibt eine Vorhersage aus, in welche Klasse der Datensatz einzuordnen ist. Die Klasse mit den meisten Vorhersagen ist die Vorhersage des Modells. Die Performance der Random Forest Modelle ist im Wesentlichen von zwei Faktoren abhängig: Die Genauigkeit der Komponenten sowie die Diversität (Korrelation) dieser Komponenten zueinander (Bharathidason & Jothi Venkataeswaran, 2014, S. 27).

Das finale Random Forest Klassifizierungsmodell nutzt eine maximale Tiefe der Entscheidungsbäume von 20 und benötigt ein Minimum von zehn Datensätzen für das Teilen eines Knotens. Der Wald des Modells besteht insgesamt aus 250 Bäumen. Das Modell verwendet zur Optimierung die Kennzahl „Gini-Impurity“, welche die Güte der Teilung der Entscheidungsknoten misst (Disha & Waheed, 2022, S. 12). Darüber hinaus wird die Funktion „Warmstart“ verwendet, Bootstrap-Stichproben jedoch nicht.

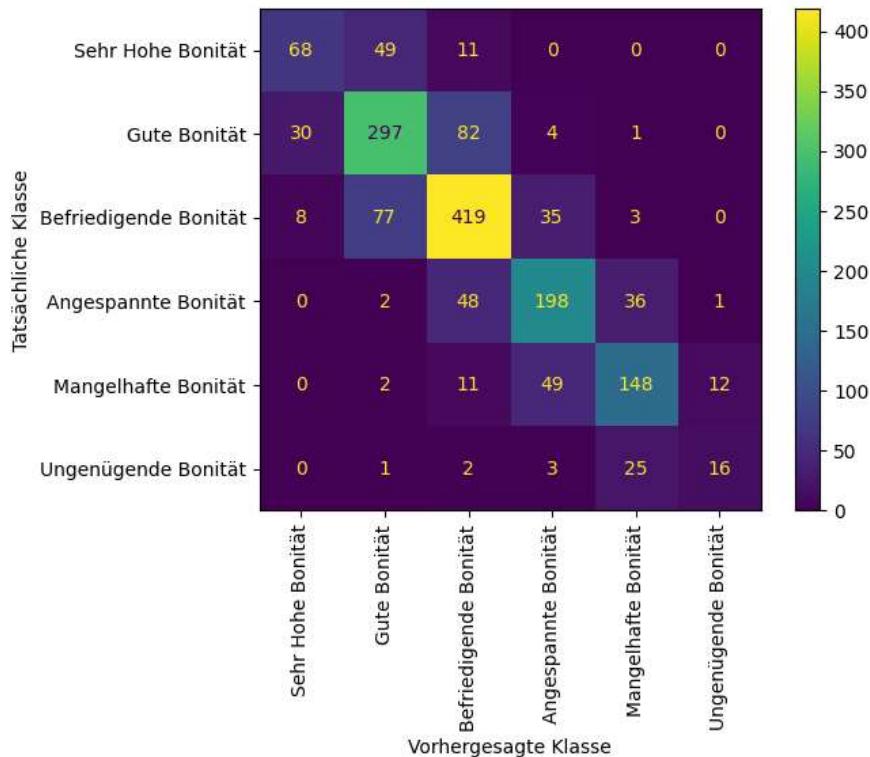
Das Modell zeigt mit einer Genauigkeit von 69,96% den bisher höchsten Wert aller getesteten Machine Learning Modelle (Abbildung 28). Ähnlich zum Gradient Boosting Modell hat das Modell Probleme mit der geringen Menge an Testdaten in der obersten und untersten Ratingkategorie. An der Konfusionsmatrix ist jedoch abzulesen, dass falsch-positive Werte in der überwiegenden Mehrheit der Fälle lediglich ein bis zwei Klassen abweichen.

Abb. 28 Klassifizierungsbericht des Random Forest Klassifizierungsmodells

	Präzision	Sensitivität	F1-Score	Testdatensätze
Sehr Hohe Bonität	0.6415	0.5312	0.5812	128
Gute Bonität	0.6939	0.7174	0.7055	414
Befriedigende Bonität	0.7312	0.7731	0.7516	542
Angespannte Bonität	0.6851	0.6947	0.6899	285
Mangelhafte Bonität	0.6948	0.6667	0.6805	222
Ungenügende Bonität	0.5517	0.3404	0.4211	47
Genauigkeit	0.6996	0.6996	0.6996	0
Ungewichteter Durchschnitt	0.6664	0.6206	0.6383	1638
Gewichteter Durchschnitt	0.6967	0.6996	0.6968	1638

Quelle: Eigene Darstellung.

Abb. 29 Konfusionsmatrix des Random Forest Klassifizierungsmodells



3.2.7 Extra-Trees Klassifizierungsmodelle zur Vorhersage von Bonitätseinstufungen

Der Extra-Trees Algorithmus wurde 2006 als eine Weiterentwicklung des traditionellen Random-Forest Klassifizierungsmodells vorgestellt (Geurts et al., 2006, S. 5). Die zwei wesentlichen Weiterentwicklungen des Extra-Trees Algorithmus bestehen in der rein zufälligen Auswahl der Knotenschnittpunkte sowie der Nutzung des gesamten Datensatzes zur Entwicklung der Entscheidungsbäume. Der Extra-Trees Klassifizierungsalgorithmus unterscheidet sich in zwei Aspekten gegenüber dem Random-Forest Klassifizierungsmodell. Einerseits wählt das Extra-Trees Modell die Werte und Attribute der Knotenschnittpunkte rein zufällig, während das Random-Forest-Klassifizierungsmodell den „Greedy Search“ Algorithmus zur Ermittlung des bestmöglichen Schnitts verwendet (siehe Abschnitt 3.3.4). Darüber hinaus nutzen Random-Forest Klassifizierungsmodelle die „Bootstrap Aggregation“-Methode, um den Lerndatensatz in viele Teilmengen zu unterteilen, um mit unterschiedlichen Variation der Trainingsdaten möglichst unterschiedliche Bäume zu konstruieren (Bharathidason & Jothi Venkataeswaran, 2014, S. 26). Das Extra-Trees Klassifizierungsmodell nutzt hingegen den gesamten verfügbaren Datensatz zur Konstruktion der Entscheidungsbäume (Geurts et al., 2006, S. 5-6).

Es ist nicht pauschal beantwortbar, welcher der beiden Algorithmen geeigneter ist, da dies von den Eigenschaften der Trainingsdaten sowie dem Problem selbst abhängig ist. Ein wichtiges Problem bei der Anwendung von Machine Learning Algorithmen ist das Finden der bestmöglichen Balance zwischen Underfitting und Overfitting (Belkin et al., 2019, S. 15849). Bias führt zu Underfitting und kann durch die Trainingsdaten in das Modell übernommen werden oder durch das Modell selbst eingeführt werden, indem nicht alle Informationen der Trainingsdaten verarbeitet werden.

(Chakraborty et al., 2021, S. 4-5). Am anderen Ende des Spektrums besteht die Möglichkeit, dass Modelle sich zu stark auf die Trainingsdaten anpassen, ohne dass daraus eine nennenswerte Verbesserung in der Vorhersage der Testdaten resultiert (Dietterich, 1995, S. 326-327). Ein derartiges Modell hat eine hohe Varianz, da auch das Rauschen der Trainingsdaten in das Modell aufgenommen wird. Das Modell ist „overfitted“. Während Extra-Trees Modelle einen höheren Bias als andere entscheidungsbaumbasierten Modelle zeigen, wird die Varianz stark reduziert (Geurts et al., 2006, S. 18-19).

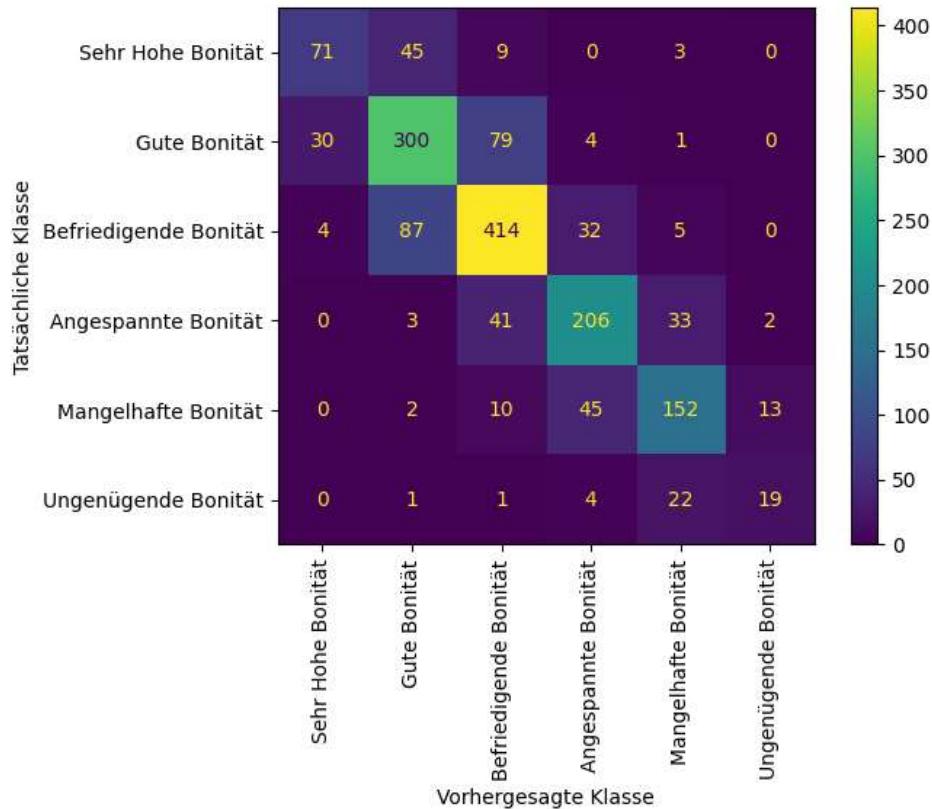
Das vollständig optimierte Extra-Trees Klassifizierungsmodell zeigt mit 150 Entscheidungsbäumen und einer maximalen Baumtiefe von 40 Knoten eine Vorhersagepräzision der Testdaten von 70,82%, der bislang höchste gemessene Wert. Im Vergleich zum Klassifizierungsbericht des Random-Forest Modells ist auffällig, dass die Sensitivität in der Ratingklasse „Befriedigende Bonität“ mit 542 Testdatensätzen unterhalb des Sensitivitätswerts des Random-Forest Modells liegt, das Extra-Trees Klassifizierungsmodell jedoch besonders in Ratingklassen mit wenigen Trainings- und Testdatensätzen mit hohen Sensitivitätswerten überzeugen kann (Abbildung 31).

Abb. 30 Klassifizierungsbericht des Extra-Trees Modells

	Präzision	Sensitivität	F1-Score	Testdatensätze
Sehr Hohe Bonität	0.6762	0.5547	0.6094	128
Gute Bonität	0.6849	0.7246	0.7042	414
Befriedigende Bonität	0.7473	0.7638	0.7555	542
Angespannte Bonität	0.7079	0.7228	0.7153	285
Mangelhafte Bonität	0.7037	0.6847	0.6941	222
Ungenügende Bonität	0.5588	0.4043	0.4691	47
Genauigkeit	0.7094	0.7094	0.7094	0
Ungewichteter Durchschnitt	0.6798	0.6425	0.6579	1638
Gewichteter Durchschnitt	0.7078	0.7094	0.7076	1638

Quelle: Eigene Darstellung.

Abb. 31: Konfusionsmatrix des ExtraTrees-Modells



Quelle: Eigene Darstellung.

3.2.8 Vorhersage von Bonitätseinstufungen mit künstlichen neuronalen Netzen

Ein künstliches neuronales Netz ist ein Machine Learning Algorithmus, dessen Ziel es ist, durch das Lernen von Mustern der Eingangsdaten möglichst präzise Ergebnisprognosen aufstellen zu können (Novac et al., 2022, S. 4). Es wurde von der menschlichen Biologie inspiriert; seine Funktionsweise wurde abgeleitet von der Art und Weise wie das menschliche Gehirn Zufuhr über seine Sinne verarbeitet (Daanouni et al., 2019, S. 2-3).

Ein künstliches neuronales Netz besteht aus verschiedenen Schichten von Funktionen zwischen der Einfuhr der Feature-Matrix und dem Ausgang, der die Vorhersage des Modells finalisiert (Daanouni et al., 2019, S. 3). Die Anzahl der Neuronen der Eingangsschicht des künstlichen neuronalen Netzes muss mit der Anzahl der Features übereinstimmen, während die Anzahl der Neuronen der Ausgangsschicht im vorliegenden mehrklassigen Klassifizierungsproblem auf die Anzahl der möglichen Ratingkategorien (sechs) abgestimmt sein muss. Liegt zwischen der Eingangs- und Ausgangsschicht mehr als eine Zwischenschicht, sogenannte „Hidden Layer“, handelt es sich um ein Deep Neural Network (Daanouni et al., 2019, S. 4). Diese Zwischenschichten identifizieren nicht nur die vorhandenen Features der eingeführten Feature-Matrix, sondern erstellen auf Basis der eingeführten Daten neue Features, analog zum Verhalten des menschlichen Gehirns. Dies ermöglicht eine Modellierung von komplexeren Strukturen innerhalb der Daten, die mit traditionellen neuronalen Netzen nicht möglich ist (Längkvist et al., 2014, S. 2-3).

Der Lernprozess eines neuronalen Netzes beginnt durch den Vergleich des Netzwerk-Outputs mit der Zielvariable (Novac et al., 2022, S. 4). Die Differenz wird in einer Verlustfunktion festgehalten, mit dem Abgleich der Daten werden die Gewichte des Netzwerks angepasst. Durch Minimierung der Verlustfunktion über mehrere Epochen hinweg lernt das Netzwerk, die Features der Inputdaten ideal zu extrahieren, sodass der gewünschte Output erreicht werden kann.

Während neuronale Netze, die für die binäre Klassifizierung eingesetzt werden, eine Log-Sigmoid Ausgangsfunktion verwenden ($f(x) = 1/(1+\exp(-x))$), welche einen Wert zwischen 0 und 1 herausgibt, verwenden neuronale Netze, die für eine mehrklassige Klassifizierung eingesetzt werden, eine Softmax Ausgangsfunktion. (Sharma et al., 2020, S. 312-314). Die Softmax Ausgangsfunktion kombiniert mehrere Log-Sigmoid Ausgangsfunktionen, welche anschließend als Wahrscheinlichkeiten der Zugehörigkeit der einzelnen Klassen interpretiert werden können.

Abb. 32 Softmax Aktivierungsfunktion der Ausgangsschicht

$$\sigma(\mathbf{z})_j = \frac{e^{z_j}}{\sum_{k=1}^K e^{z_k}} \quad \text{for } j = 1, \dots, K.$$

Quelle: Übernommen aus Sharma et al., 2020, S. 314.

Um eine optimale Lernperformance des künstlichen neuronalen Netzwerks zu erreichen, müssen die Daten der Feature-Matrix zunächst standardisiert werden (Raschka, 2019, S. 125). Zwar ist die Skalierung der Daten mithilfe der Min-Max-Methode ein beliebtes Verfahren, um die Features-Werte in ein gebundenes Intervall zu bringen, jedoch ist die Standardisierung der Daten für den Gradientenabstieg-Optimierungsalgorithmus eines neuronalen Netzwerks geeigneter. Nach der Standardisierung hat jede Spalte der Feature-Matrix die gleichen Eigenschaften wie die Standardnormalverteilung (Mittelwert von 0 und Standardabweichung von 1), wodurch das Erlernen der Gewichte vereinfacht wird. Darüber hinaus bleiben, anders als bei einer Skalierung der Daten, Informationen über Ausreißer erhalten.

Darüber hinaus ist für die Kreuzentropie-Verlustfunktion des PyTorch-Frameworks bei mehrklassigen Klassifizierungsproblemen das One-Hot-Kodieren der Zielvariablen erforderlich. Hierbei wird für jede mögliche Klasse der Zielvariablen ein Dummy-Feature erstellt (Raschka, 2019, S. 119). Anschließend werden Binärwerte verwendet, um anzusehen, welcher Kategorie das Unternehmen zugeordnet wird (Abbildung 33). Ein Unternehmen, welches der zweitbesten Bonitätsklasse „Gute Bonität“ zugeordnet wird, enthält somit den Vektor 0, 1, 0, 0, 0, 0.

Abb. 33 Kodieren der Zielvariablen mithilfe des One-Hot-Kodierers der SciKit-Learn Bibliothek

```
encoder = OneHotEncoder(sparse_output=False)
y = encoder.fit_transform(y)
print(f'Form von y: {y.shape[0]} Zeilen, \
{y.shape[1]} Spalten')
print(y)

✓ 0.0s

Form von y: 7799 Zeilen, 6 Spalten
[[0. 1. 0. 0. 0.]
 [1. 0. 0. 0. 0.]
 [0. 0. 1. 0. 0. 0.]
```

Quelle: Eigene Darstellung.

Die erste Version des neuronalen Netzes enthält lediglich eine Zwischenschicht mit 16 Neuronen, die mit einer Rectified-Linear-Unit (ReLU) Funktion aktiviert wird (Abbildung 34). Diese ReLU Funktion ist definiert als $f(x) = \max(0, x)$ und setzt damit alle negativen Werte auf 0, während positive Werte unverändert weitergegeben werden. Da die Ableitung dieser Funktion damit immer entweder 0 oder 1 ist, ist die ReLU-Aktivierungsfunktion berechnungseffizient.

Abb. 34 Einfaches künstliches neuronales Netz in PyTorch

```
class RatingsNet(nn.Module):
    def __init__(self):
        super().__init__()
        self.hidden = nn.Linear(11, 16)
        self.output = nn.Linear(16, 6)
        self.relu = nn.ReLU()
        self.softmax = nn.Softmax(dim=1)

    def forward(self, x):
        x = self.relu(self.hidden(x))
        x = self.softmax(self.output(x))
        return x
```

Quelle: Eigene Darstellung.

Das Modell nutzt zur Evaluierung und Optimierung eine Kreuzentropie-Verlustfunktion, die durch den Lernprozess stetig minimiert wird. Der ideale Wert der Verlustfunktion ist damit 0. Die Kreuzentropie-Verlustfunktion gibt die durchschnittliche Differenz aller Klassen zwischen den vorhergesagten Wahrscheinlichkeiten und den tatsächlichen Werten wieder (Jacobson, 2021, S. 10). Für den ersten Durchlauf des Modells wird die Anzahl der Trainingsepochen auf 200 gesetzt, die Batch-Größe auf 8. Die Lernrate des „Adam“ Optimierers wird auf den Standardwert 0.001 festgelegt.

Im ersten Durchlauf erreicht das künstliche neuronale Netz eine Testgenauigkeit von lediglich 46,09%, der bisher niedrigste gemessene Wert der getesteten Machine Learning Modelle. Bei der Analyse der ausgegebenen Konfusionsmatrix fällt zudem auf, dass die beiden seltener auftretenden Ratingkategorien „Sehr Hohe Bonität“ und „Ungenügende Bonität“ keine Vorhersagen des Modells

erhalten haben. Eine mögliche Ursache hierfür ist, dass das einfache neuronale Netz mit nur einer Zwischenschicht und 16 Neuronen die Komplexität dieser Randkategorien aufgrund von Underfitting nicht erfassen kann. Außerhalb des Aufbaus eines komplexeren neuronalen Netzes bietet PyTorch zudem verschiedene Optionen, um den Lernprozess des Modells selbst zu optimieren, damit ein für die Daten optimales Modell mit den bestmöglichen Hyperparametern entwickelt werden kann.

Dauer und Aufwand der Optimierung bleiben jedoch Nachteile in der Anwendung künstlicher neuronaler Netze, besonders im Vergleich zu den bereits vorgestellten Ensemble-Lernmethoden. Die zu optimierende Zielfunktion der neuronalen Netze ist nicht-konvex und auch die beliebtesten Algorithmen zeigen ohne aufwendige Optimierung in den meisten Fällen zunächst eine schwache Performance (Goodfellow et al., 2015, S. 1). Zum Finden der bestmöglichen Parameter wird meistens eine Rastersuche durchgeführt, die abhängig von der Anzahl der gesuchten Hyperparameter sowie der Anzahl von Variationen pro Hyperparameter eine Zehntausendfache Ausführung des Modells bedeuten kann, insbesondere weil jede Konfiguration für eine Kreuzvalidierung der Ergebnisse mehrfach ausgeführt wird.

Abb. 35 Parameter-Raster zur Rastersuche der idealen Hyperparameter mit dreifacher Kreuzvalidierung für SciKit-Learns MLP-Klassifizierungsmodell

```
param_grid = {
    'hidden_layer_sizes': [(100, 100, 100, 100), (150, 150, 150)],
    'activation': ['tanh', 'relu'],
    'solver': ['sgd'],
    'alpha': [0.0001, 0.05, 0.001, 0.00001],
    'learning_rate': ['constant', 'adaptive'],
    'learning_rate_init' : [.1, .01, .001, .0001],
    'power_t' : [0.5, 0.3, 0.7],
    'shuffle': [True, False],
    'momentum': [0, 0.5, 0.9],
    'batch_size': ['auto', 64, 128]
}

Fitting 3 folds for each of 6912 candidates, totalling 20736 fits
```

Quelle: Eigene Darstellung.

Um die schnell aufeinander folgenden Ausführungen zu optimieren, wird häufig die Technik des „Early Stoppings“ (*Deutsch: frühzeitiges Abbrechen*) eingesetzt. Hierbei handelt es sich um ein Objekt außerhalb des künstlichen neuronalen Netzes, welches den Trainingsfortschritt fortlaufend überwacht. Dies geschieht durch den Abgleich der Epochewerte der Präzision mit dem bisher höchsten aufgezeichneten Wert Präzisionswert. Zeigt die Präzision über eine festgelegte Anzahl von Epochen hinweg keine Verbesserung, wird die Ausführung des Modells angehalten; im Falle einer Rastersuche wird zur nächsten Konfiguration übergegangen. Die für das frühzeitige Abbrechen gewählte Kennzahl kann je nach Einsatz und Art des Modells variieren. Da die für das Training eingesetzten Grafikkarten oftmals eine Leistung von mehreren Hundert Watt angeben (NVIDIA Corporation, n. d.), und für das Training großer Modelle meist mehrere Grafikkarten in einem Raster

verwendet werden, kann die Implementierung einer frühzeitigen Abbruchfunktion nicht nur Overfitting verhindern, sondern zusätzlich eine erhebliche Zeit- und Kosteneinsparung bedeuten.

Eine weitere Technik zum Verhindern von Overfitting der Trainingsdaten ist die Einführung eines Dropouts in das künstliche neuronale Netzwerk (Anh et al., 2023, S. 641-645). Dropout entfernt zufällig ausgewählte Knoten mit deren Verbindungen in den verschiedenen Schichten des neuronalen Netzes. Dropout bewirkt, dass einige Neuronen während der Einführung der Feature-Matrix in das Modell (Forward-Pass) ignoriert werden. Dies führt dazu, dass die Neuronen, die flussabwärts mit dem durch Dropout wegfallenden Neuron verbunden sind, ebenfalls ignoriert werden. Zusätzlich werden während der Optimierung des Modells (Backward-Pass) die Gewichte der wegfallenden Neuronen nicht optimiert. Die einzelnen Neuronen innerhalb des Netzwerks nehmen mit ihren Gewichten eine spezifische, Feature-abhängige Rolle für die Optimierung des Netzwerks ein. Dies bewirkt eine Abhängigkeit angrenzender Neuronen zueinander. Fällt ein Neuron während des Trainingsprozesses weg, müssen angrenzende Neuronen die Rolle des ausgefallenen Neurons übernehmen. Das Resultat der Implementierung des Dropouts ist, dass das Modell weniger different auf die exakten Gewichte der Neuronen reagiert. Dies führt zu einem Modell, das weniger anfällig für Overfitting ist.

Neben den Parametern des Modells selbst bestimmen die Hyperparameter wie der Trainingsprozess selbst zu erfolgen hat. Ein wichtiger Parameter ist dabei die Lernrate, die festlegt, wie schnell das Netzwerk seine Gewichte anpasst (Jacobson, 2021, S. 8-9). Ein Vorteil einer niedrigen Lernrate ist das gleichmäßige Konvergieren zum Optimum, jedoch führt eine zu niedrige Lernrate zu Overfitting sowie langsamerer Trainingsperformance (Smith, 2018, S. 6). Obwohl höhere Lernraten Overfitting verhindern können, kann ihr Einsatz zur frühzeitigen Divergenz des Trainings führen, wodurch im Trainingsprozess nicht das optimale Modell gefunden wird. Eine mögliche Lösung zum Finden der optimalen Lernrate außerhalb der Rastersuche ist der Einsatz einer zyklischen Lernrate. Das in dieser Arbeit verwendete Vorhersagemodell für Bonitätseinstufungen nutzt hierfür die PyTorch-Optimizer Funktion „Reduce Learning Rate on Plateau“ (*Deutsch: Verringerung der Lernrate auf Plateau*). Die Optimierung des neuronalen Netzes startet in der ersten Epoche mit einer verhältnismäßig hohen Lernrate von 0,1. Stagniert der Präzisionswert des Modells über zehn Epochen hinweg und zeigt keine Verbesserung, wird die Lernrate halbiert. Das Modell mit der höchsten Präzision der Lernrate 0,1 wird nun mit einer Lernrate von 0,05 weiterentwickelt. Zeigt der Präzisionswert des Modells bis zum Ende des Trainings durch die Epochen stetige Verbesserungen, bleibt die Lernrate unberührt. Um ein extremes Verringern der Lernrate gegen 0 zu verhindern, wird vorab eine minimale Lernrate von 0,001 festgelegt.

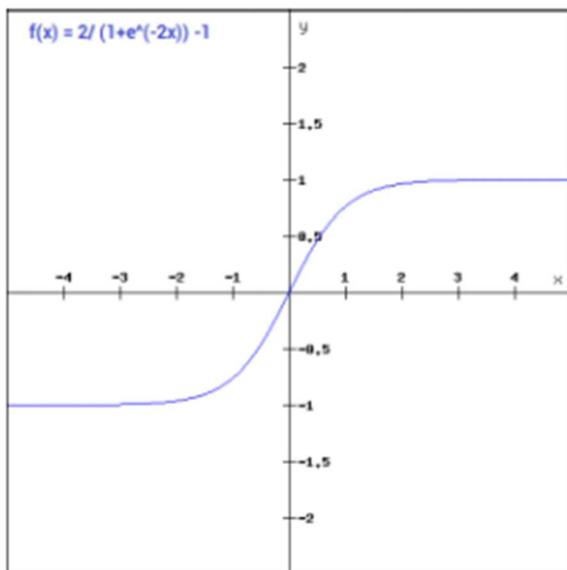
Das erste gezeigte Modell verwendet zur Optimierung des neuronalen Netzes während des Trainingsprozesses den Adam-Optimierer (*Adaptive Moment Estimation*). Dieser kombiniert den adaptiven Gradient-Algorithmus AdaGrad mit der Wurzel-Mittelwert-Vermehrungsalgorithmus RMSProp (Bock & Weis, 2019, S. 2). Das optimierte Vorhersagemodell für Bonitätseinstufungen

verwendet hingegen im Trainingsprozess ein traditionelleres stochastisches Gradientenverfahren (SGD), da dies eine höhere Testgenauigkeit aufwies. Eine mögliche Ursache hierfür ist die Kombination des SGD-Optimierers mit einer kleineren Batch-Größe, da diese Kombination Rauschen in den Optimierungsprozess einführt (Ziyin et al., 2022, S. 1). Dieses Rauschen kann, in einem kontrollierten Ausmaß, Overfitting verhindern und damit die Testperformance des Modells verbessern.

Abschließend ist die Architektur des künstlichen neuronalen Netzes selbst zu optimieren. Das erste Modell mit lediglich 16 Neuronen und nur einer Zwischenschicht hatte besonders bei den Randkategorien „Sehr Hohe Bonität“ und „Ungenügende Bonität“ Probleme, die Komplexität der Daten zu erfassen. Zur Lösung dieses Problems können dem Modell Schichten hinzugefügt werden, durch eine Erhöhung der Anzahl der Neuronen in jeder Schicht kann die Kapazität des Netzwerks erhöht werden. Das finale Bonitätseinstufungsmodell hat vier Zwischenschichten mit jeweils 256, 512, 256 und 64 Neuronen.

Die Aktivierungsfunktionen der Zwischenschichten sind Tanh-Funktionen, die Ausgangsschicht nutzt eine Log-Softmax Aktivierungsfunktion. Obwohl die ReLU-Aktivierungsfunktionen des ersten Modells effizient sind, zeigten Tanh-Funktionen im direkten Vergleich für dieses mehrklassige Klassifikationsmodell eine bessere Leistung. Die Tanh-Aktivierungsfunktion ist kontinuierlich und differenzierbar, die Ausgangswerte liegen im Bereich von -1 bis 1 (Sharma et al., 2020, S. 313). Durch ihre Nullzentrierung hat die Tanh-Aktivierungsfunktion eine stärkere Steigung (Abbildung 36).

Abb. 36 Die Tanh-Aktivierungsfunktion

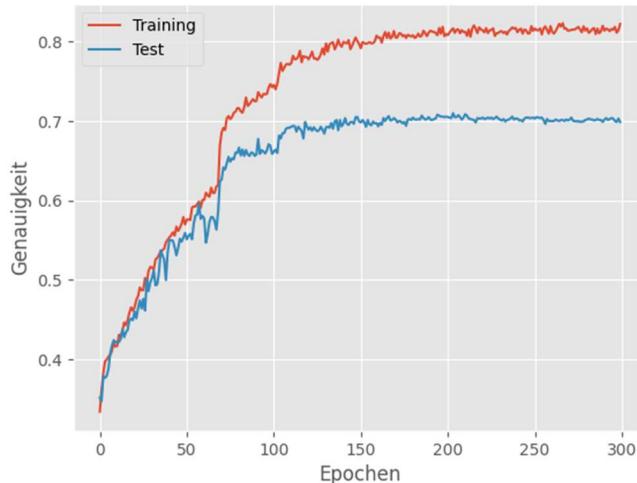


Quelle: Übernommen aus Sharma et al., 2020, S. 313.

Die Log-Softmax Ausgangsfunktion nutzt das gleiche Prinzip der Softmax-Funktion, gibt jedoch anstelle der relativen Wahrscheinlichkeiten Log-Wahrscheinlichkeiten aus. Dies führt im Falle einer Fehlklassifizierung während des Trainingsprozesses zu einer stärkeren Bestrafung des Modells.

Das überarbeitete künstliche neuronale Netz für Bonitätseinstufungen erreicht durch die Änderungen eine Testgenauigkeit von 70,9%, und erreicht damit eine 25 Prozentpunkte bessere Performance als das nicht optimierte Modell. Auf dem Verlauf der Genauigkeitskurve über die Trainingsepochen ist zu erkennen, dass sowohl in der Epoche 70 als auch in der Epoche 104 die Genauigkeit sprunghaft ansteigt, da in diesen Epochen aufgrund eines Plateaus eine Verringerung der Lernrate initiiert wurde.

Abb. 37 Trainings- und Testgenauigkeit über 300 Epochen



Quelle: Eigene Darstellung.

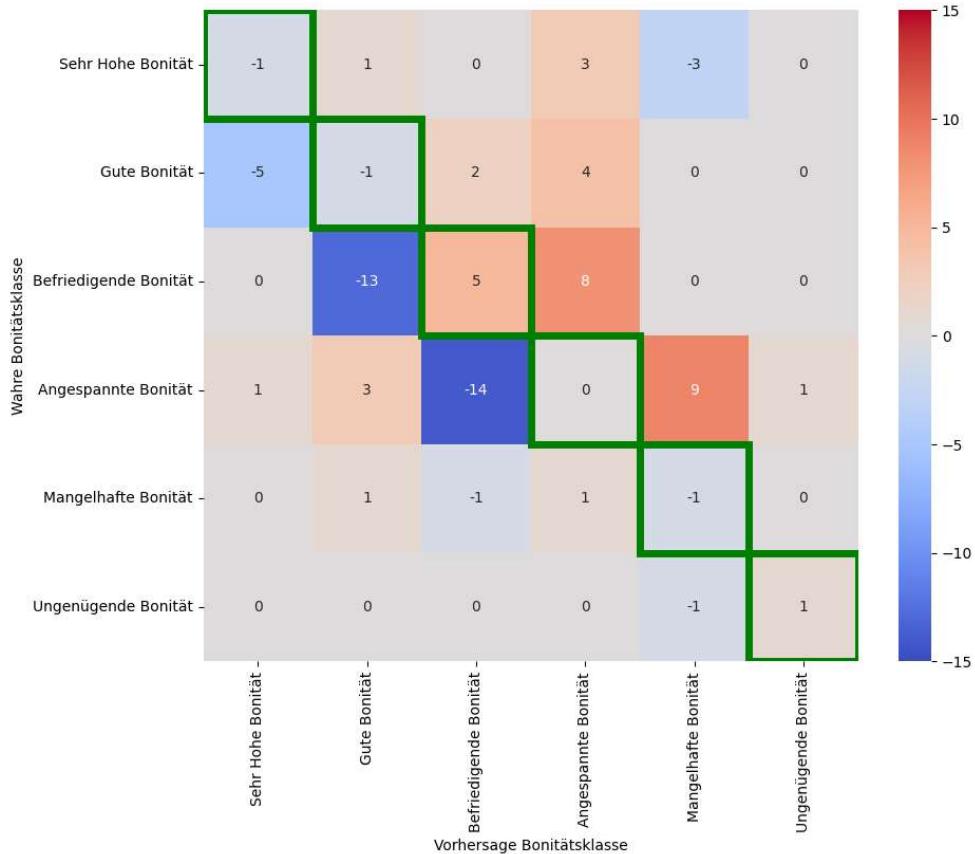
Das neuronale Netz erreicht mit seiner Testgenauigkeit von über 71% einen Bestwert aller Modelle. Ein Vergleich der Konfusionsmatrix mit dem ExtraTrees-Klassifizierungsmodell zeigt, dass das neuronale Netz durch sein hochoptimiertes Training besonders in Klassen mit vielen Testdatensätzen besser abschneidet (Abbildung 38). Durch die erhöhte Komplexität des überarbeiteten neuronalen Netzes können nun auch Datensätzen den Bonitätsklassen „Sehr Hohe Bonität“ sowie „Ungenügende Bonität“ zugeordnet werden. In der Bonitätsklasse „Sehr Hohe Bonität“ zeigt das neuronale Netz zudem im Fall einer Fehlklassifizierung eine geringere Abweichung. Während das ExtraTrees-Klassifizierungsmodell drei Unternehmen mit tatsächlichem Label „Sehr Hohe Bonität“ in die Klasse „Mangelhafte Bonität“ einordnet, werden durch das neuronale Netz diese drei Unternehmen lediglich in die Klasse „Angespannte Bonität“ eingeordnet.

Abb. 38 Klassifikationsbericht des neuronalen Netzes

	Präzision	Sensitivität	F1-Score	Testdatensätze
Sehr Hohe Bonität	0.6786	0.5938	0.6333	128
Gute Bonität	0.6918	0.7319	0.7113	414
Befriedigende Bonität	0.7661	0.7675	0.7668	542
Angespannte Bonität	0.6689	0.6947	0.6816	285
Mangelhafte Bonität	0.6991	0.6802	0.6895	222
Ungenügende Bonität	0.5758	0.4043	0.4750	47
Genauigkeit	0.7100	0.7100	0.7100	0
Ungewichteter Durchschnitt	0.6800	0.6454	0.6596	1638
Gewichteter Durchschnitt	0.7090	0.7100	0.7087	1638

Quelle: Eigene Darstellung.

Abb. 39 Differenz der Konfusionsmatrizen des neuronalen Netzes und des ExtraTrees-Klassifizierungsmodells

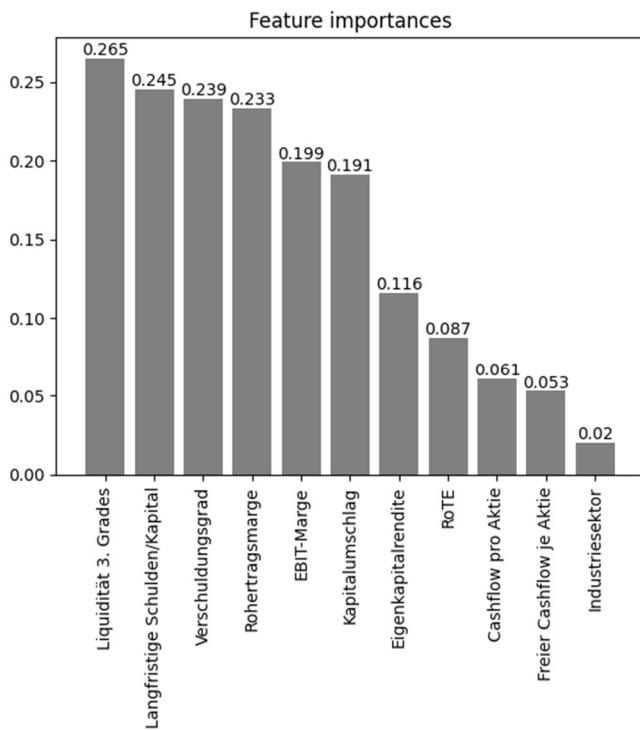


Quelle: Eigene Darstellung.

Abschließend ist die Wichtigkeit der einzelnen Features im neuronalen Netz zu analysieren. Die Permutationswichtigkeit ermittelt hierbei, wie die Genauigkeit des Modells negativ beeinflusst wird, wenn ein Feature zufällig gemischt wird und damit keinen Mehrwert für das Training des Modells darstellt (Adams & Collyer, 2015, S. 826). Der Score von 0,265 der Liquidität dritten Grades ist so zu interpretieren, dass die Genauigkeit des Modells um 26,5% sinkt, wenn die Werte dieses Features

zufällig durchmischt werden (Abbildung 40). Im Vergleich zur zuvor ausgeführten Auswertung für das XGBoost-Klassifizierungsmodell fällt auf, dass die mit wenigen Ausnahmen die meisten Features einen fast identischen Stellenwert für das Training des Modells einnehmen. Ein wesentlicher Unterschied zwischen beiden Modellen ist bei der Verschuldung zu erkennen: Während für das neuronale Netz der Verschuldungsgrad an dritter Stelle steht, ist diese für das XGBoost-Klassifizierungsmodell die zweit-unwichtigste Kennzahl. Eine ähnliche Diskrepanz ist ebenfalls für den RoTE zu erkennen: Dieses Feature ist für das Training des neuronalen Netzes deutlich unwichtiger, als es für das XGBoost-Klassifizierungsmodell ist. Auffällig ist zudem, dass auch bei einer Klassifizierung durch ein neuronales Netz laut der Permutationswichtigkeit der Industriesektor keinen besonderen Einfluss hat.

Abb. 40 Permutationswichtigkeit der Features im neuronalen Netz



Quelle: Eigene Darstellung.

Das Ranking der Featurewichtigkeit innerhalb eines Modells ist jedoch nicht ausreichend zur Erklärung einzelner Vorhersagen (Meng et al., 2020, S. 477). Eine Möglichkeit zur Erklärung einer individuellen Vorhersage ist die Messung des Beitrags eines jeden Features zur Vorhersage, indem durch Berücksichtigung aller möglicher Kombinationen von Features die wechselseitigen Abhängigkeiten der Features zueinander für die Generierung der Vorhersage analysiert werden.

Aus spieltheoretischer Sicht lässt sich ein mehrklassiges Klassifizierungsmodell rationalisieren als kollaboratives Spiel, in welchem die Agenten (Features) auf strategische Art und Weise miteinander agieren, um das gemeinsame Ziel der Vorhersage zu erreichen (Padarian et al., 2020, S. 390). Spieltheorie ermöglicht die mathematische Beobachtung der Interaktionen und Strategien der

involvierten Agenten. Jeder Agent erhält als Resultat eine Auszahlung, die im direkten Verhältnis zur erbrachten Leistung steht. Im Falle eines Klassifizierungsmodell ist der erreichte Gewinn bzw. Verlust die Abweichung einer Vorhersage vom Mittelwert aller Vorhersagen.

Der marginale Beitrag eines Features kann mithilfe des Shapley Werts gemessen werden (Shapley, 1953, S. 4-9). Um die Shapley Werte zu errechnen, ist es erforderlich jeweils ein Modell $f_{S \cup \{i\}}$ mit dem zu beobachteten Feature und ein Modell f_S ohne das zu beobachtende Feature zu entwickeln (Padarian et al., 2020, S. 390). Dieser Vorgang ist für alle Features zu wiederholen. Die Vorhersagendifferenz für den Input x ist der Beitrag des Features i . Da das Bonitätseinstufungsmodell elf Features verwendet, ist der Beitrag jedes Features abhängig von der Interaktion mit allen anderen Features.

Abb. 41 Gleichung zur Ermittlung des Shapley-Werts für Kovariate i

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|! (|F| - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)]. \quad (1)$$

Quelle: Übernommen aus Padarian et al., 2020, S. 390.

Da die Ermittlung des Shapley Werts mithilfe der ursprünglichen Gleichung im Falle des künstlichen neuronalen Netzes für Bonitätseinstufungen das Training von $2^{11}=2048$ Modellen bedeuten würde, verwendet die von Lundberg und Lee entwickelte Python-Bibliothek „shap“ eine abgewandelte Version der Shapley-Kennzahl mit dem Namen „Shapley additive explanations“ (Lundberg & Lee, 2017, S. 4-8). „Deep SHAP“ ermittelt die SHAP-Werte jeder Komponente eines Deep Neural Networks mithilfe von linearen Approximationen, die anschließend über eine Kompositionsregel aggregiert werden (Padarian et al., 2020, S. 390). Dies ermöglicht eine effiziente Approximation der SHAP-Kennzahl für das gesamte Modell.

Abb. 42 Bildung von 100 Clustern mithilfe des Trainingsdatensatzes

```
kmeans_background = shap.kmeans(X_train.detach().cpu().numpy(), 100)
```

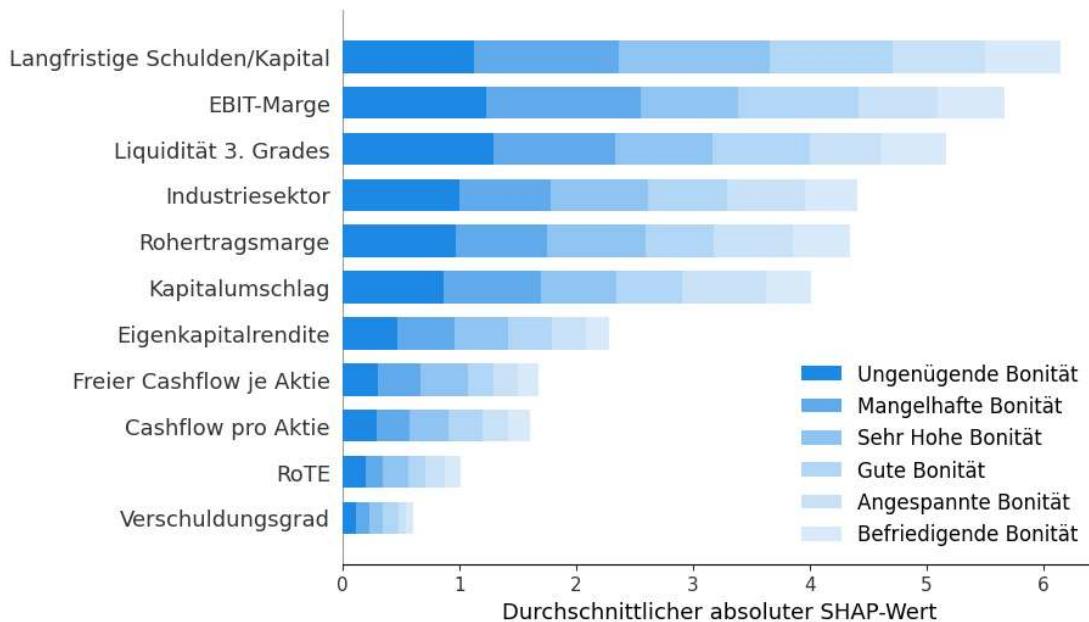
Quelle: Eigene Darstellung.

Zur Ermittlung der SHAP-Werte des Bonitätseinstufungsmodell wird der Trainingsdatensatz mit 6161 Datensätzen mithilfe der K-Means Clustering Methode zunächst in 100 verschiedene Cluster aufgeteilt (Abbildung 42). Anschließend wird der repräsentativste Datenpunkt eines jeden Clusters, der sogenannte Zentroid, als Probe zur Ermittlung der SHAP-Werte ausgewählt. Die 100 Zentroide repräsentieren die verschiedenen Muster und Strukturen des Trainingsdatensatzes.

Die SHAP-Zusammenfassung zeigt, dass das Feature „Langfristige Schulden/Kapital“ mit dem höchsten durchschnittlichen absoluten SHAP-Wert das einflussreichste Feature der Feature-Matrix ist (Abbildung 43). Damit trägt es am meisten zur Vorhersage einer Prognose bei. Auffällig ist darüber

hinaus, dass der Industriesektor über alle verschiedenen Klassen hinweg einen konstanten und insgesamt signifikanten Beitrag zur Prognose des Modells leistet, anders als durch die Permutationsauswertung angedeutet. Damit hat das Feature „Industriesektor“ beim zufälligen Durchmischen nur einen geringen Einfluss auf die Präzision des Modells, ändert die Vorhersage des Modells jedoch signifikant. Dies offenbart, dass das Feature komplexe Beziehungen zu anderen Features pflegt, die durch den Permutationsscore nicht erfasst werden.

Abb. 43 Zusammenfassung der durchschnittlichen absoluten SHAP-Werte der Kovariaten



Quelle: Eigene Darstellung.

Die Klassen der SHAP-Zusammenfassung sind nach der kumulierten absoluten Summe der SHAP-Werte sortiert. Auffällig ist hierbei, dass besonders die außenliegenden Klassen („Ungenügende Bonität“, „Mangelhafte Bonität“ und „Sehr Hohe Bonität“) hohe SHAP-Werte zeigen und die Wahrscheinlichkeit deren Prognose besonders durch die drei Features mit den höchsten mittleren absoluten SHAP-Werten stark beeinflusst werden. Der Einfluss der einzelnen Features auf die Klassen „Angespannte Bonität“ sowie „Befriedigende Bonität“ ist weniger stark, da die Kennzahlwerte dieser Unternehmen weniger auffällig sind.

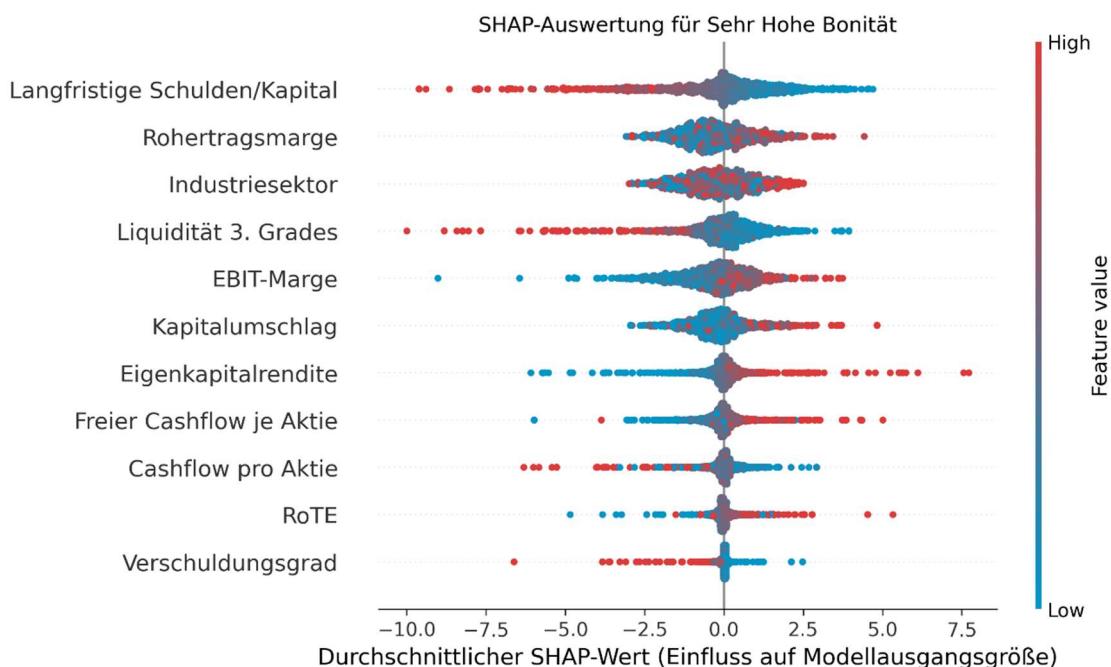
Da die SHAP-Zusammenfassung lediglich die Mittelwerte der absoluten SHAP-Werte erfasst, ist an ihr nicht abzulesen, in welche Richtung ein bestimmtes Feature die Vorhersage einer Klasse beeinflusst. Für die Vorhersage der Klasse „Sehr Hohe Bonität“ hat eine hohe EBIT-Marge eine andere wahrscheinlichkeitsbedingte Auswirkung als für eine schlechtere Bonitätseinstufungsklasse.

Erwartungsgemäß führt ein hoher Wert der Kennzahl „Langfristige Schulden/Kapital“ zu einer drastisch geringeren Wahrscheinlichkeit, dass das Unternehmen in die Kategorie „Sehr Hohe Bonität“ eingestuft wird (Abbildung 44). Ein SHAP-Wert von -8 heißt jedoch nicht, dass ein hoher Wert bei diesem Feature zu einer acht Prozent geringeren Wahrscheinlichkeit der Prognose für diese Klasse führt, sondern bezieht sich lediglich auf die Logit-Werte des Modells, die anschließend durch

die Log-Softmax Aktivierungsfunktion in Log-Wahrscheinlichkeiten für diese Klasse umgewandelt werden. Während ein Datensatz lediglich entweder einer Klasse zugeordnet werden kann oder nicht (0 oder 1), kann die Wahrscheinlichkeit p , die durch die Softmax-Ausgangsfunktion für die Klasse ausgegeben wird, jeden beliebigen Wert zwischen 0 und 1 annehmen (Bender et al., 2007, S. 33). Die Gleichung $p/(1-p)$ setzt dabei die Wahrscheinlichkeit des Eintreffens eines Ereignisses ins Verhältnis zum Nichteintreffen desselben Ereignisses. Der Logarithmus dieser Gleichung ist der Logit, der die Wahrscheinlichkeit p im gesamten Raum der reellen Zahlen abbildet.

Überraschend ist, dass ein höherer Wert für die Liquidität 3. Grades zu einer geringeren Wahrscheinlichkeit der Prognose für die Klasse „Sehr Hohe Bonität“ führt (Abbildung 44). Aus betriebswirtschaftlicher Sicht gibt es hierfür verschiedene Erklärungen. Eine Möglichkeit ist, dass Unternehmen mit einer hohen Liquidität dritten Grades eine erhöhte Menge an Vorräten halten, was in vielen Fällen auf den Nichtverkauf von Produkten im Lager zurückzuführen ist. Darüber hinaus ist diese Beobachtung konsistent mit den Auswertungen Baghai et al., da Unternehmen mit einer konservativen Liquiditätspolitik Chancen verpassen, in wachstumsfördernde Projekte zu investieren (Baghai et al., 2014, s. auch Abschnitt 2.2).

Abb. 44 SHAP-Auswertung "Sehr Hohe Bonität"



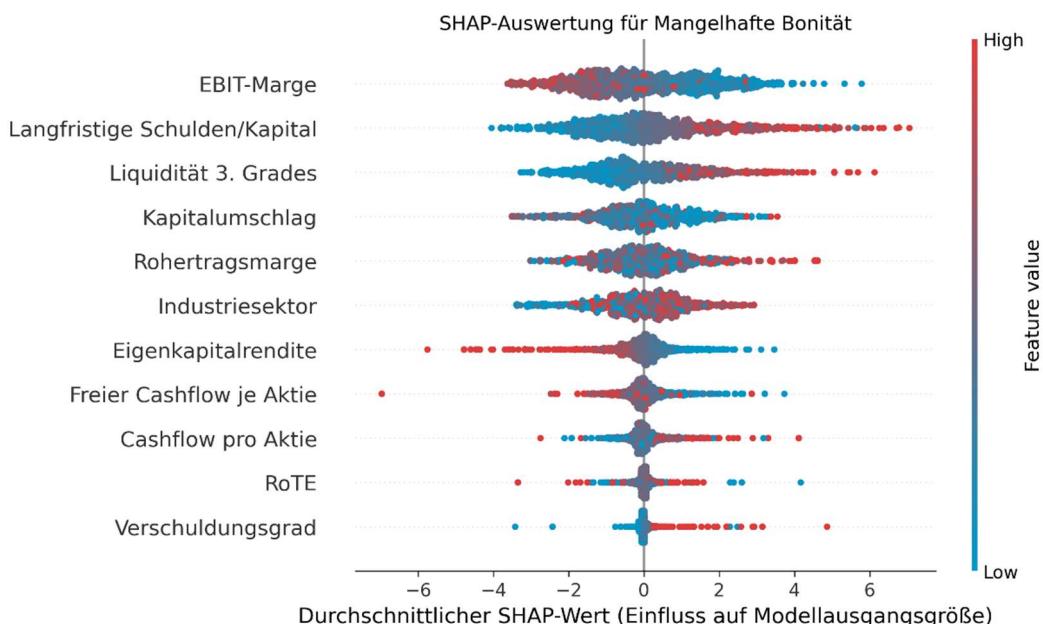
Quelle: Eigene Darstellung.

Die Features Freier Cashflow je Aktie, Cashflow pro Aktie, RoTE und Verschuldungsgrad zeigen zwar vereinzelt hohe SHAP-Beträge, jedoch stauen sich die meisten Werte um den Nullpunkt und führen damit zu einer geringeren durchschnittlichen Bedeutung dieser Features. Eine hohe Rohertragsmarge, eine hohe EBIT-Marge und einer Kapitalumschlag trägt signifikant zur Einstufung des Unternehmens in die Klasse „Sehr Hohe Bonität“ bei (Abbildung 44).

Zu einer Einstufung in die Bonitätsklasse „Mangelhafte Bonität“ führen besonders niedrige Werte bei den Features EBIT-Marge und Kapitalumschlag, sowie eine besonders hohe langfristige Verschuldung im Verhältnis zum Gesamtvermögen sowie eine hohe Liquidität 3. Grades (Abbildung 45). Während die EBIT-Marge für die Bonitätsklasse „Sehr Hohe Bonität“ lediglich das fünfwichtigste Feature ist, gehört es für alle anderen Bonitätsklassen zu den drei wichtigsten Features. Mögliche Auslöser hierfür ist erneut die Interaktion der Features untereinander, besonders in einer Klasse mit einer geringeren Anzahl an Trainings- und Testdatensätzen sowie potenziellen Ausreißern.

Im Vergleich der beiden Klassen ist zudem auffällig, dass die SHAP-Werte für die Klasse „Sehr Hohe Bonität“ in der Verteilung höhere Werte erreichen als es bei der Klasse „Mangelhafte Bonität“ der Fall ist. Damit haben einzelne Features der Unternehmen, die in die Klasse „Sehr Hohe Bonität“ eingestuft wurden eine höhere Auswirkung auf die Modellausgangsgröße als es für Unternehmen der Klasse „Mangelhafte Bonität“ der Fall ist. Eine Prognose der Klasse „Mangelhafte Bonität“ wird somit gleichmäßiger durch die verschiedenen Features getragen.

Abb. 45 SHAP-Auswertung "Mangelhafte Bonität"



Quelle: Eigene Darstellung.

3.3 Einbeziehung von ESG-Scores zur Verbesserung der Vorhersagequalität

Unternehmen sehen sich seit Beginn der digitalen Transformation in den 1990er Jahren zunehmend neuen Herausforderungen gegenübergestellt. Diese neue Geschäftswelt wird häufig mithilfe des Akryoms „VUCA“ charakterisiert: Volatilität, Ungewissheit, Komplexität und Mehrdeutigkeit. Allein im Jahr 2020 gab es mehrere „Black Swan“ Ereignisse: Die globale Ausweitung der COVID-19 Pandemie, deutliche Einbrüche am Aktienmarkt innerhalb von zwei Wochen, Heuschreckenplagen in Afrika und einige mehr (Li et al., 2021, S. 1).

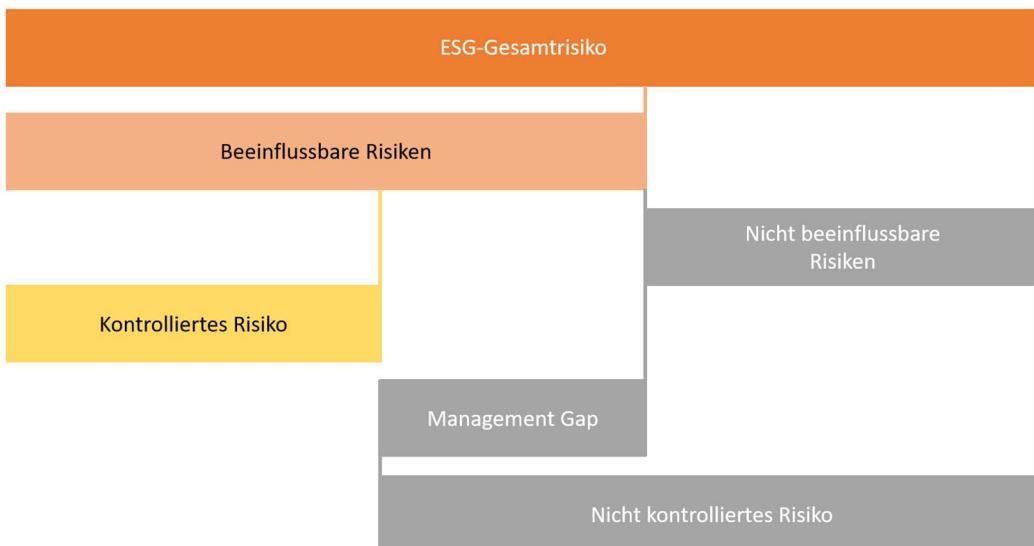
Als Antwort auf die entstehenden Probleme im 21. Jahrhundert wurden neue Kriterien entwickelt um Unternehmen nicht nur auf Basis ihrer finanziellen Performance, sondern ebenfalls über die Nachhaltigkeit ihrer Geschäftsmodelle, ihren Arbeitsbedingungen und Gleichstellungsmaßnahmen sowie ihrer Corporate Governance zu beurteilen. Das ESG-System wurde 2004 eingeführt und berücksichtigt als Teil der Responsible-Investment-Bewegung 28 Faktoren der Bereiche Umwelt, Soziales und Governance (Li et al., 2021, S. 1-2). Im Zuge der ESG-Bewegung werden Unternehmen zunehmend in die Pflicht genommen, sich ihrer sozialen Verantwortung anzunehmen und ihre Geschäftsmodelle umweltfreundlicher zu gestalten. Darüber hinaus trägt die Beachtung von ESG-Kriterien zu einem effektiven Risikomanagement bei. Unternehmen, deren Geschäftsmodelle potenziell umweltschädlich sind, werden in der Zukunft mit einer erhöhten Wahrscheinlichkeit durch politische Maßnahmen in ihren Geschäftstätigkeiten beeinflusst.

Untersuchungen zeigen, dass im Vergleich zur Umweltdimension, die soziale Dimension einen noch höheren, positiven Einfluss auf die Performance eines Unternehmens hat (Li et al., 2021, S. 15). Flammer et al. Konnten zusätzlich zeigen, dass die Aufnahme von ESG KPIs in die Verträge von Vorstandsmitgliedern nicht nur zu effektiven Umsetzung von Umwelt- und CSR-Zielen führt, sondern zusätzlich die finanzielle Performance als auch die Unternehmensbewertung steigert (Flammer et al., 2019, S. 24-29).

Zur Messung des Einflusses der ESG-Bewertungen auf die Bonitätseinstufungen der Unternehmen wird das „ESG Risk Rating“ der Morningstar Sustainalytics Website verwendet. Dies kombiniert verschiedene ESG-Faktoren in einem Rating, das einen Vergleich verschiedener Unternehmen auf der Basis einer einzelnen Kennzahl ermöglicht. Sustainalytics analysiert hierfür zunächst, welchen ESG-Risiken das Unternehmen insgesamt ausgesetzt ist (Morningstar Sustainalytics, 2020). Ähnlich zur Portfoliotheorie werden diese Risiken in beeinflussbare und nicht beeinflussbare Risiken aufgeteilt, wobei der Industriesektor eine wichtige Rolle spielt (Abbildung 46). Eine Fluggesellschaft ist typischerweise einem höheren nicht beeinflussbaren ESG-Risiko ausgesetzt als ein Softwareunternehmen.

Anschließend werden die beeinflussbaren Risiken erneut aufgeteilt in Risiken, die vom Unternehmen kontrolliert werden, und dem „Management Gap“, welches die Risiken darstellt, denen das Unternehmen aktuell ausgesetzt ist, obwohl diese kontrolliert werden könnten (Morningstar Sustainalytics, 2020). Ein Beispiel einer Management Gap ist die Absenz einer einschlägigen Diskriminierungspolice des Unternehmens, die eine Management Gap im Bereich Humankapital darstellt. Das finale ESG-Risikorating setzt sich aus dem gesamten nicht kontrollierten Risiko zusammen: Dem sektorbedingten, nicht beeinflussbaren Risiko und dem „Management Gap“.

Abb. 46 Morningstar Sustainalytics ESG Risikobewertungsprozess



Quelle: Eigene Darstellung in Anlehnung an Morningstar Sustainalytics, 2020.

Sustainalytics teilt die ESG-Scores je nach Höhe in fünf unterschiedliche Klassen auf: Unternehmen mit einem Risikoscore von Null bis Zehn sind einem vernachlässigbaren ESG-Risiko ausgesetzt, ein Score von Zehn bis Zwanzig bedeutet ein geringes Risiko, Unternehmen mit einem Score von 20 bis 30 sind einem mittleren Risiko ausgesetzt, während Unternehmen mit einem Score zwischen 30 und 40 einem hohen, darüber hinaus einem schwerwiegendem ESG-Risiko ausgesetzt sind.

Ein entscheidender Vorteil der Morningstar Sustainalytics Bewertungen gegenüber anderen ESG-Bewertungen ist die Verfügbarkeit von Bewertungen für über 20.000 Unternehmen weltweit (Morningstar Sustainalytics, 2023). Dies ermöglicht die Erhaltung von über 95% der Daten aus dem zuvor verwendeten Datensatz, wodurch ein direkter Vergleich zu vorherigen Vorhersagemodellen ermöglicht wird.

Morningstar bietet für seinen Sustainalytics Service keine kostenfreie API an, daher werden die Daten mithilfe eines selbstentwickelten Data Mining Skripts von der Website abgegriffen.¹ Von den 7805 Datensätzen des ursprünglichen Testdatensatzes konnten so fast 96% der Datensätze (7476) für die ESG-Auswertung erhalten bleiben. Aufgrund der sektortypischen, nicht kontrollierbaren Risiken, wird in der Datenvorbereitung neben dem ESG-Score selbst noch ein weiteres Feature „ESG-Score zu Sektordurchschnitt“ erstellt, welches das Verhältnis des ESG-Ratings eines Unternehmens dem jeweiligen Durchschnittsscore seines Sektors misst.

Tab. 3 Durchschnitt und Standardabweichung der ESG-Ratings nach Sektoren

Sektor	\varnothing ESG-Rating	σ ESG-Rating
Geschäftsausstattung	17,6967	5,8177
Chemie	29,1706	6,9162

¹ Selbstentwickelter Sustainalytics ESG Scraper: https://github.com/leander-ms/esg_scrap

Gebrauchsgüter	17,0454	5,1453
Energie	36,4658	7,9030
Gesundheit	25,5598	5,3982
Herstellung	25,7510	8,0230
Finanzen	18,8109	7,4269
Verbrauchsmaterial	23,4926	9,1342
Andere	24,3417	7,2395
Einzelhandel	19,5547	8,3249
Telekommunikation	23,1987	5,5203
Versorgung	28,4934	5,6604

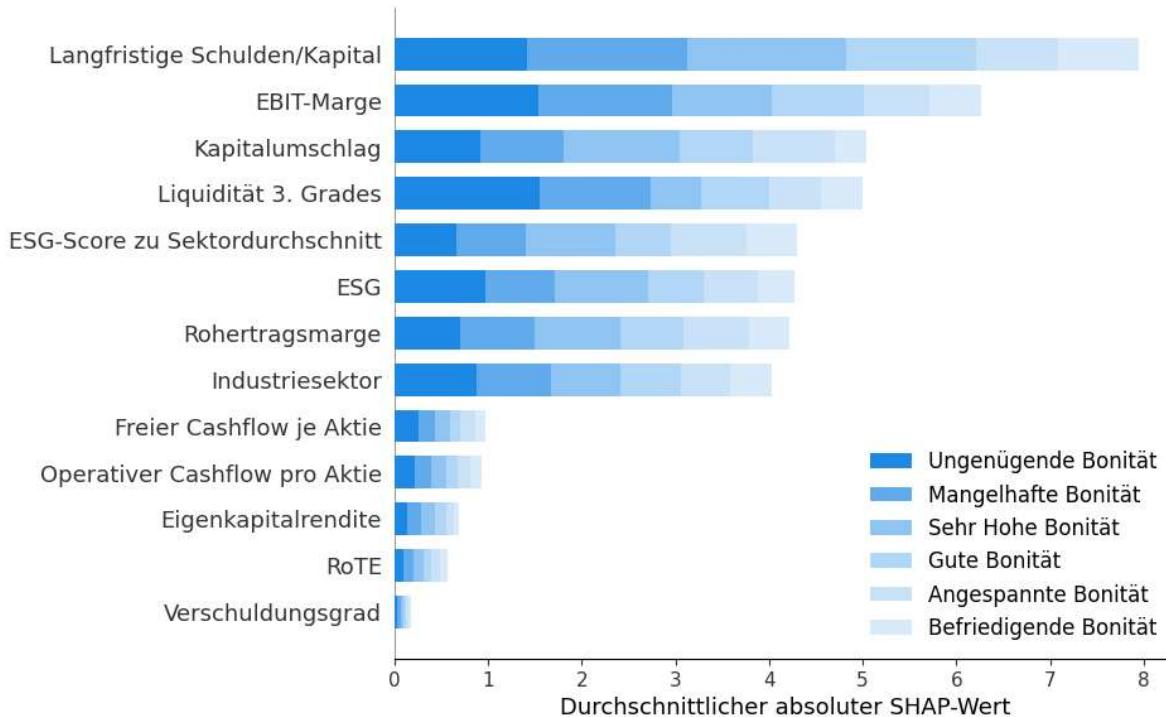
Quelle: Eigene Darstellung.

Der Durchschnitt aller ESG-Ratingscores beträgt 24,55. Erwartungsgemäß zeigen Unternehmen des Energiesektors mit 36,4658 den höchsten durchschnittlichen ESG-Ratingscore (Tabelle 4). Darüber hinaus ist die Standardabweichung der Unternehmen des Energiesektors mit 7,9030 im Vergleich zu anderen Sektoren hoch, möglicherweise verursacht durch umweltfreundlichere Energieunternehmen, die bereits zu einem großen Teil auf erneuerbare Energien setzen und damit ein niedrigeres ESG-Rating erhalten, sowie besonders umweltschädlichen Energieunternehmen mit einem sehr hohen Rating. Die Sektoren der Hersteller von Gebrauchsgütern, Geschäftsausstattung sowie Finanzunternehmen weisen durchschnittlich die niedrigsten ESG-Ratings auf.

Aufgrund des verkleinerten Datensatzes erreicht das PyTorch-Modell ohne die ESG-Features „ESG-Score“ und „ESG-Score zu Sektor durchschnitt“ eine Genauigkeit von lediglich 69%. Die Größe des Testdatensatzes wurde auf 18% des Gesamtdatensatzes reduziert, um einen Verlust an Trainingsdatensätzen zu verhindern. Das finale Modell mit den zuvor genannten Features erreicht eine Genauigkeit von 71,71%. Damit verbessert sich die Genauigkeit des Modells um knapp zwei Prozentpunkte gegenüber dem Modell ohne ESG-Features.

Aufgrund der erhöhten Komplexität durch die zwei zusätzlichen Features wurde das neuronale Netz selbst angepasst. Die angepasste Version besteht nun aus vier Zwischenschichten mit einer Ausgangsschicht.

Abb. 47 Durchschnittliche absolute SHAP-Werte der Kovariaten



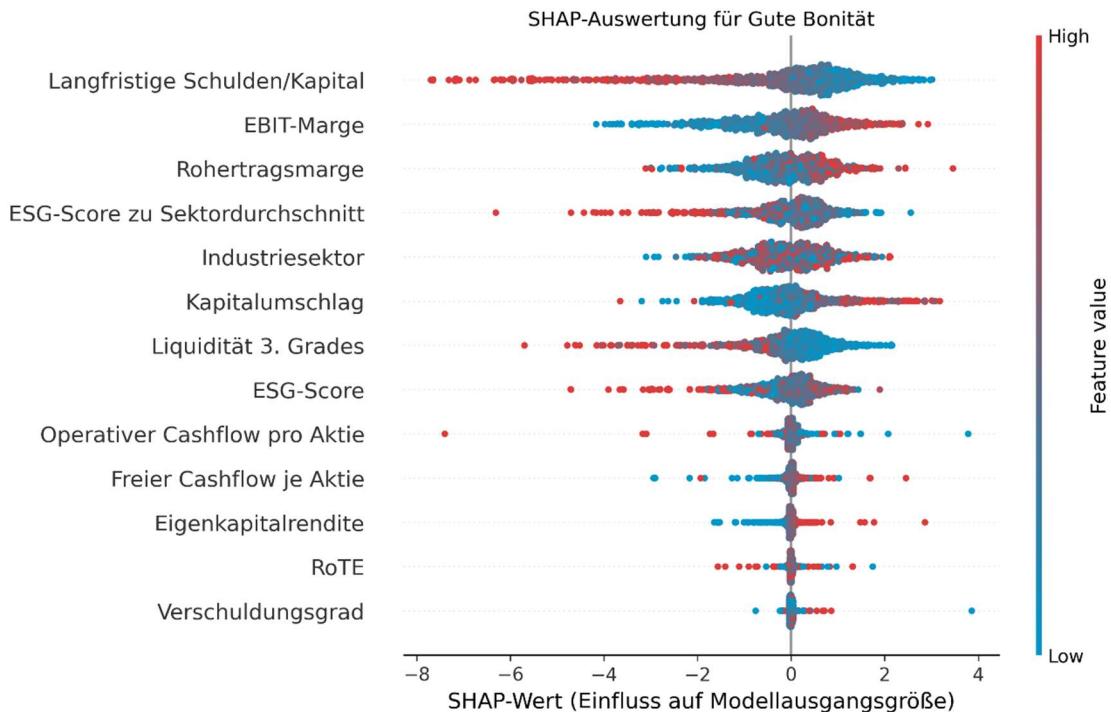
Quelle: Eigene Darstellung.

Anhand der SHAP-Auswertung ist zu erkennen, dass die beiden Features im Zusammenspiel mit anderen Variablen einen wichtigen Beitrag zur Modellausgangsgröße beitragen (Abbildung 47). Auffällig ist hierbei, dass der SHAP-Wert für die Variable „ESG-Score zu Sektordurchschnitt“ über die Klassen hinweg einen gleichmäßig hohen absoluten SHAP-Wert zeigt. Der ESG-Score selbst ist – ähnlich zu bereits anderen Features – für die außenliegenden Klassen deutlich aussagekräftiger, besonders für die Klasse „Sehr Hohe Bonität“. Darüber hinaus ist bemerkenswert, wie sich die SHAP-Werte der bestehenden Variablen im Vergleich zum künstlichen neuronalen Netz ohne den ESG-Werten entwickelt haben. Die weiterhin wichtigste Kennzahl „Langfristige Schulden/Kapital“ hat durch die Interaktion mit den neuen Features ihren durchschnittlichen absoluten SHAP-Wert von sechs im ersten Modell auf acht erhöht. Die Kennzahl Kapitalumschlag, die zuvor im Mittelfeld aller Kennzahlen lag, ist nun die dritt wichtigste Kennzahl. Dies ist auf die Interaktion zwischen den addierten ESG-Kennzahlen und den bestehenden Kennzahlen zurückzuführen.

Die detaillierte Auswertung der SHAP-Werte für die Klasse „Gute Bonität“ zeigt, dass der ESG-Score im Verhältnis zum Sektordurchschnitt bei Unternehmen mit vergleichsweise hoher Bonität eine größere Auswirkung auf die Modellausgangsgröße hat als der ESG-Score selbst (Abbildung 48). Ein hoher ESG-Score im Verhältnis zum Sektordurchschnitt verringert die Logit-Wahrscheinlichkeit einer Einstufung in die Klasse „Gute Bonität“. Die SHAP-Werte des absoluten ESG-Scores stehen für die Klasse „Gute Bonität“ lediglich an achter Stelle, ihr Einfluss auf die Modellausgangsgröße ist weniger eindeutig. Es ist jedoch erkennbar, dass ein besonders hoher ESG-Score einen negativen Einfluss auf die Logit-Wahrscheinlichkeit für die Zuordnung zur Klasse „Gute Bonität“ hat. Die Beobachtungen zum absoluten ESG-Score sind jedoch nicht konstant. Während in der Klasse „Gute

„Bonität“ der durchschnittliche SHAP-Wert für Unternehmen mit einem überdurchschnittlich hohem ESG-Score (nach der Standardkalierung ESG-Score > 0) bei -0,03 liegt, wird der Klasse „Sehr Hohe Bonität“ für dieses Kriterium ein durchschnittlicher SHAP-Wert von 0,80 zugeordnet.

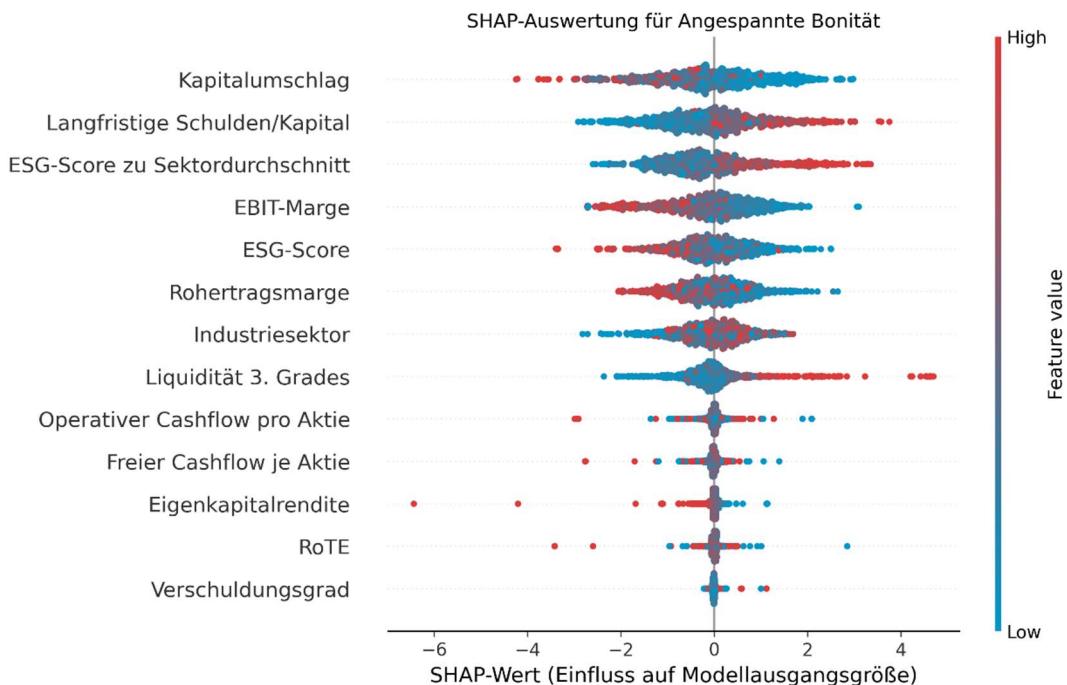
Abb. 48 SHAP-Auswertung unter Einbeziehung von ESG-Kennzahlen für die Klasse "Gute Bonität"



Quelle: Eigene Darstellung.

Das Feature „ESG-Score zu Sektordurchschnitt“ hilft dem Modell besonders dabei, Unternehmen der Bonitätsklasse „Befriedigende Bonität“ von der Bonitätsklasse „Angespannte Bonität“ zu unterscheiden. Diese Klassenunterscheidung ist besonders wichtig, da Unternehmen der Klasse „Befriedigende Bonität“ grundsätzlich als Investment Grade eingestuft sind, während Unternehmen der Klasse „Angespannte Bonität“ kein Investment Grade Rating erhalten haben. Während der durchschnittliche SHAP-Wert in der Klasse „Befriedigende Bonität“ für Unternehmen mit überdurchschnittlich hohem ESG-Score im Verhältnis zu ihrem Sektordurchschnitt bei -0,67 liegt, zeigt die Klasse „Angespannte Bonität“ unter Anwendung dieser Kriterien einen durchschnittlichen SHAP-Wert von 0,504.

Abb. 49 SHAP-Auswertung unter Einbeziehung von ESG-Kennzahlen für die Klasse "Angespannte Bonität"



Quelle: Eigene Darstellung.

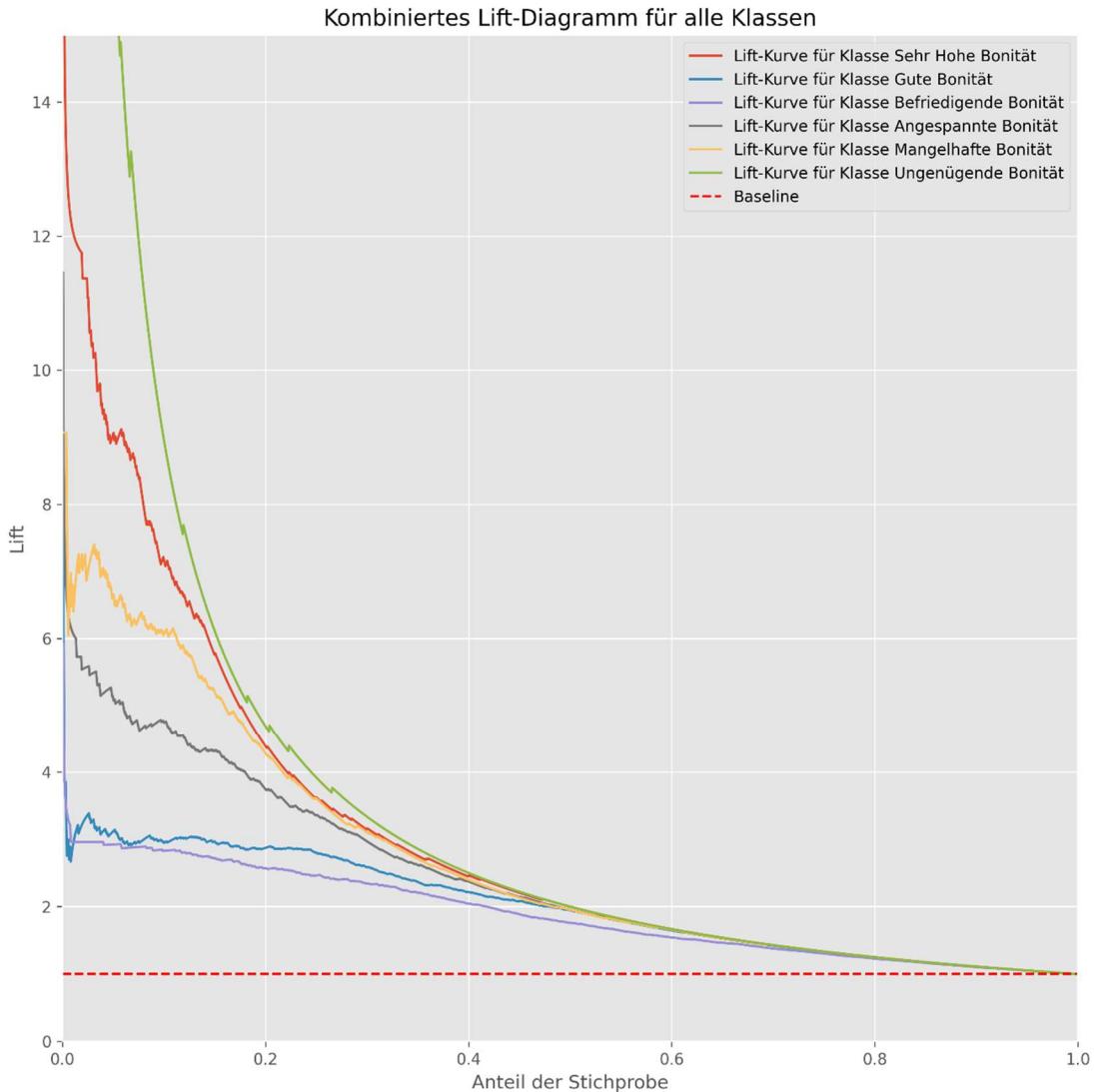
Insgesamt zeigt die finale Iteration des künstlichen neuronalen Netzes eine starke Vorhersagekraft für die Einstufung der Unternehmen in die Bonitätsklassen. Um die Vorhersagekraft des Modells gegenüber einer zufälligen Zuordnung in die Klassen zu messen, werden Lift-Diagramme eingesetzt. Ein Lift-Wert von zwei bedeutet, dass das Modell doppelt so effektiv positive Datensätze für die Klasse auswählt wie ein Zufallsmodell.

Das Lift-Diagramm der verschiedenen Klassen zeigt über alle Klassen hinweg für den gesamten Datensatz einen Lift-Wert von über 1 (Abbildung 50). Das zufällige Vorhersagemodell ist durch die rot-gestrichelte Linie dargestellt.

Alle Lift-Diagramme zeigen eine über die Stichprobe abfallende Kurve, da Vorhersagemodelle ihre Vorhersagen in einer nach Konfidenz abfallenden Reihenfolge sortieren. Hieraus resultiert, dass das Modell in seinen ersten Vorhersagen der Stichprobe am sichersten ist. Ein Lift-Wert von fünf für die Klasse „Mangelhafte Bonität“ bei einem Anteil der Stichprobe von 20% bedeutet, dass das neuronale Netz innerhalb der obersten 20% der Stichprobe, die das Modell als am wahrscheinlichsten für diese Klasse einstuft, circa fünfmal so viele Unternehmen in die Klasse „Mangelhafte Bonität“ korrekt einstuft als die zufällige Vorhersage.

Damit zeigt das künstliche neuronale Netz für Bonitätseinstufungen über alle Klassen hinweg eine bessere Vorhersagequalität als das zufällige Vorhersagemodell.

Abb. 50 Lift-Diagramm aller Klassen

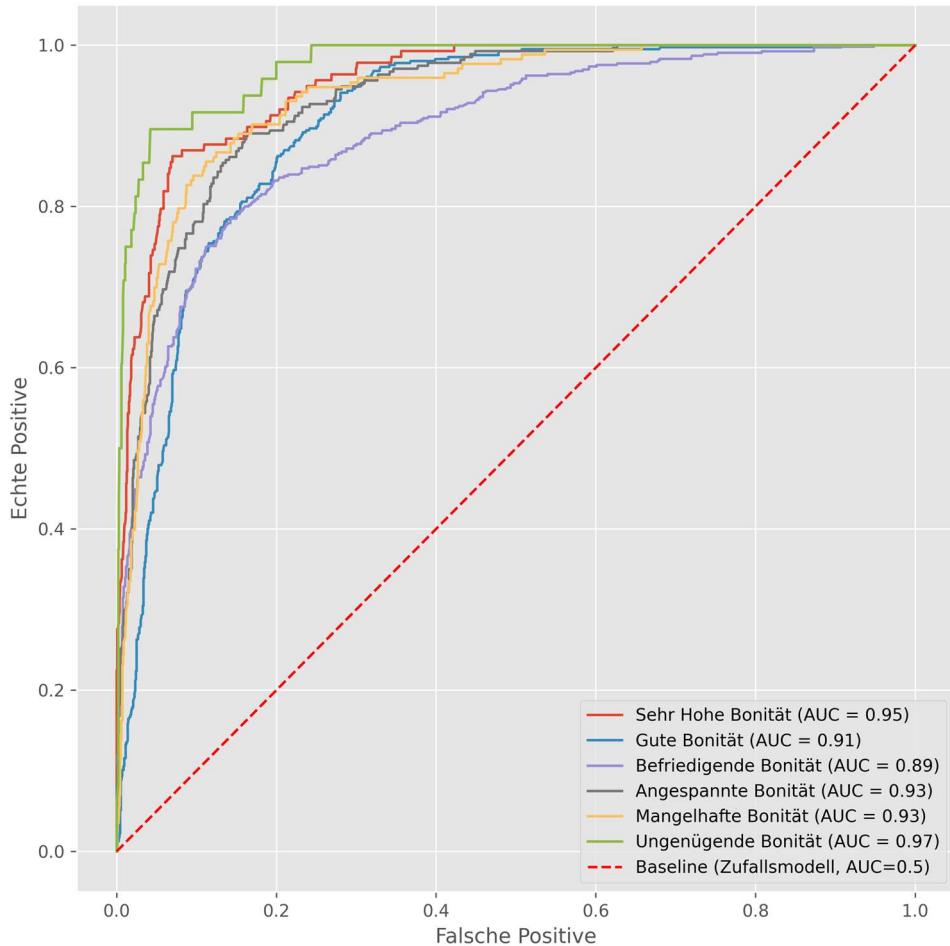


Quelle: Eigene Darstellung.

Die ROC-Kurve setzt die Sensitivität (True Positive Rate) ins Verhältnis zur Spezifität (False Positive Rate, siehe auch Abschnitt 2.1). Auch bei der Analyse der ROC-Kurven ist erkennbar, dass das Modell für alle Klassen eine höhere Vorhersagequalität bietet als ein Zufallsmodell. Der AUC-Wert von 0,95 für die Klasse „Sehr Hohe Bonität“ bedeutet, dass ein zufällig ausgewähltes Unternehmen, welches dieser Klasse zugeordnet wurde, mit einer 95-prozentigen Wahrscheinlichkeit eine höheren Bonitätsklasse angehört als Unternehmen, die schlechteren Ratingklassen zugeordnet wurden.

Auch in der ROC-Auswertung ist zu erkennen, dass das Modell für die Klasse „Befriedigende Bonität“ eine geringere Performance im Vergleich zu den extremeren Klassen zeigt. Dies ist darauf zurückzuführen, dass einzelne Kovariaten in mittleren Klassen eher überlappen als in außenliegenden Klassen, in denen einzelne Merkmale teilweise extreme Ausprägungen annehmen.

Abb. 51 ROC-Kurven aller Klassen



Quelle: Eigene Darstellung.

4. Limitationen des Modells

4.1 Grenzen und Einschränkungen von ESG-Kennzahlen

Ein grundlegendes Problem von ESG-Kennzahlen ist die zahlentechnische Erfassung der sozialen- bzw. Umweltauswirkungen, sowie der Corporate Governance eines Unternehmens (Pérez et al., 2022, S. 3). Obwohl die erforderlichen Daten über die zuvor genannten Dimensionen erfasst werden können, haben die aggregierten ESG-Score nur eine geringe Aussagekraft und sind lediglich ordinal interpretierbar: Texas Instruments hat mit einem ESG-Score von 20,6 einen ca. 50% niedrigeren Score als Exxon Mobil (41,6), jedoch kann nicht quantifiziert werden, wie viel besser Texas Instruments hinsichtlich der ESG-Kriterien tatsächlich aufgestellt ist.

Ein weiteres Problem besteht in der Nichterfassung bestimmter Emissionen, die durch das Unternehmen verursacht werden. Zur Messung von Unternehmensemissionen wird zwischen drei verschiedenen Scopes unterschieden. Scope 1 beinhaltet alle Emissionen, die direkt durch das Unternehmen ausgestoßen werden (Kaplan & Ramanna, 2021, S. 5). Beispiele hierfür sind sowohl die Fahrzeugflotte des Unternehmens als auch das Produktionsequipment. Scope 2 Emissionen sind jene, die nicht direkt beim Unternehmen entstehen, jedoch einfach zuordenbar sind, wie

beispielsweise die Emissionen der Elektrizität, die das Unternehmen verbraucht. Diese sind in der Regel durch Daten des Energielieferanten ermittelbar. Scope 3 beinhaltet die Emissionen, die indirekt von anderen Unternehmen der Wertschöpfungskette oder durch Kunden verursacht werden. Diese Emissionen stellen den größten Teil der durch das Unternehmen emittierten Treibhausgase dar, ihre genaue Messbarkeit ist jedoch eingeschränkt. Scope 3 Emissionen eines Unternehmens sind Scope 1 und 2 Emissionen seiner Zulieferer und Kunden (Wittevrongel, 2022, S. 1-2). Damit ist die Erfassung von Emissionen des dritten Scopes nicht nur deutlich aufwendiger, sondern führt zusätzlich zur Doppelerfassung von Emissionen (Kaplan & Ramanna, 2021, S. 5-6). Die meisten Unternehmen vermeiden daher jegliche Berichterstattung über Scope 3 Emissionen, besonders weil ihr direkter Einfluss auf diese oft begrenzt ist.

Zur Lösung dieses Problems schlagen Kaplan und Ramanna ein System der Emissionsbuchhaltung vor („E-Liability“), welches Blockchain-Technologie verwendet, um alle während der Produktherstellung anfallenden Emissionen ohne eine Doppelerfassung zu verbuchen (Kaplan & Ramanna, 2021, S. 6-8). Nach diesem System erfasst jedes Unternehmen der Wertschöpfungskette lediglich seine Scope 1 Emissionen und verbucht diese auf den hergestellten Wertgegenstand. Anschließend addiert es über die Scope 2 Emissionen jegliche „indirekten Kosten“ für Elektrizität oder Ähnliches und bucht diese auf das Produkt. Das fertige Produkt erhält am Ende der Wertschöpfungskette einen vollständigen Bericht über alle Treibhausgasemissionen, die durch seine Herstellung verursacht wurden. Ähnlich zu den Nährstoffberichten auf der Rückseite von Nahrungsmitteln würde ein solcher Bericht die Kaufentscheidung potenzieller Kunden beeinflussen.

Ein Vorteil der E-Liability Lösung ist die Erfassung sämtlicher Emissionen der Wertschöpfungskette, sodass Unternehmen, die ihre Scope 1 Emissionen mithilfe von Outsourcing verringern möchten, diese Emissionen nun nicht mehr aus ihrem Bericht verlagern können. (Kaplan & Ramanna, 2021, S. 8). Alle Treibhausgasemissionen, die durch einen Outsourcing-Zulieferer bei der Herstellung eines Produkts entstanden sind, werden beim Kauf des Produkts auf das Unternehmen übertragen. Darüber hinaus ist das E-Liability-System vollständig anreizkompatibel. Ein Zulieferer profitiert nicht von einer Unterbewertung der Treibhausemissionen, die es an Kunden weitergibt, weil dies seine eigene Bilanz verschlechtern würde. Eine Überbewertung der Emissionen wird auf Widerstand der potenziellen Käufer des Zulieferers treffen.

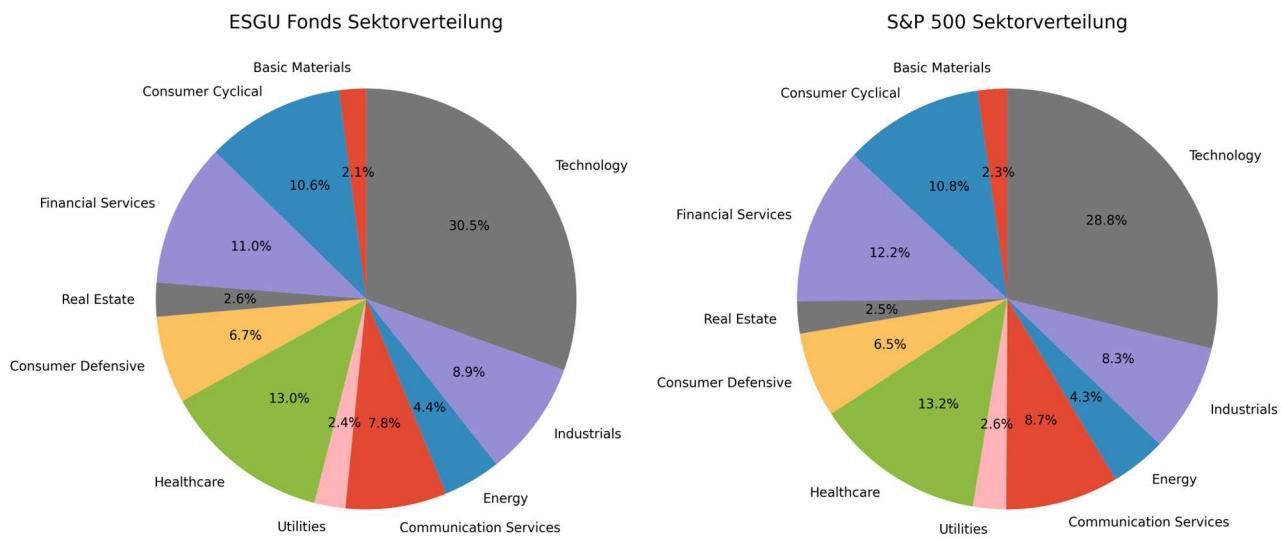
Die E-Liability-Bilanz eines Unternehmens kann ähnlich zu veröffentlichten Finanzberichten von unabhängigen Dritten überprüft werden (Kaplan & Ramanna, 2021, S. 9). Durch Echtzeit-Buchungen auf der Blockchain können Experten jegliche Unstimmigkeiten in der E-Liability-Bilanz umgehend identifizieren.

Darüber hinaus variieren die ESG-Bewertungen der unterschiedlichen Anbieter stark. Während für die Bonitätseinstufungen zwischen Anbietern wie Standard & Poor's und Moody's eine Bewertungskorrelation von 99% besteht, zeigen die ESG-Bewertungen der sechs größten Anbieter eine durchschnittliche Korrelation von lediglich 54% (Berg et al., 2022, S. 1321). Zwischen den

Anbietern Sustainalytics, S&P, Moody's, KLD, Refinitiv und MSCI korrelieren die Sustainalytics und Moody's Bewertungen am stärksten mit 0,71. Während die Bewertungen der Umweltdimension am stärksten korrelieren, zeigen die Untersuchungen Berg et al. für die Governance-Dimension mit 0,3 die geringste Korrelation. Dies zeigt das größte Problem mit ESG-Kennzahlen in ihrer aktuellen Form auf: Anders als TQM oder JIT ist ESG kein einheitliches Konzept (Kaplan & Ramanna, 2021, S. 2). Jede der drei ESG-Dimensionen stellt eine eigene Herausforderung hinsichtlich der Messung und Auswertung der Unternehmensdaten dar. Darüber hinaus können ESG-Kennzahlen, abhängig von den angewandten Auswertungsmethoden, negative Anreize schaffen: Ein Unternehmen, das aufgrund vieler Unfälle am Arbeitsplatz einen schlechten ESG-Score hat, könnte diesen Score verbessern, indem es zunehmend auf Automatisierung und Outsourcing setzt. Diese Verbesserung des ESG-Scores geschieht jedoch zu Lasten lokaler Arbeitsplätze und Zulieferer.

Große Investmentshäuser entwickeln hinsichtlich der ESG-Kriterien meist ihre eigenen, proprietären Auswertungsmethoden, die aus einer Bandbreite von ESG-Kennzahlen eine finale Bewertung zusammenstellen (Pérez et al., 2022, S. 3). Hiermit werden jedoch die Probleme der ESG-Kennzahlen nicht umgegangen, sondern durch verringerte Transparenz befördert. Die vermeintlich ESG-konformen BlackRock ETFs ESGU (4,8%) und SUSA (3,8%) halten Anteile an emissionsstarken und ESG-schwachen Unternehmen des Energiesektors wie Exxon Mobil, Chevron und Valero Energy (Schmidt, 2022). Damit weicht die Gewichtung von Unternehmen des Energiesektors innerhalb der ESG-Fonds nicht bedeutend vom S&P 500 (4,28%) ab (Abbildung 52).

Abb. 52 Vergleich Sektorverteilung ESGU und S&P 500



Quelle: Eigene Darstellung auf der Basis von Yahoo Finance, 2023.

Laut Todd Rosenbluth, Forschungsleiter bei VettaFi, enthalten die ESG-Fonds Anteile an potenziell umweltschädlichen Unternehmen zu Diversifizierungszwecken (Schmidt, 2022). Somit sollte der Sinn von Investitionen in ESG-Fonds hinterfragt werden. Im Bereich der Wirtschaftswissenschaften herrscht Einigkeit, dass es für Investoren mit sozialen Beweggründen unmöglich ist, Politik und

Produktion eines Unternehmens hinsichtlich der ESG-Kriterien durch den Kauf seiner Aktie positiv zu beeinflussen (Brest & Born, 2013).

4.2 Datensätze für die Entwicklung des neuronalen Netzes

Ein Einschränkung des im Rahmen dieser Arbeit entwickelten künstlichen neuronalen Netzes zur Vorhersage der Ratingklasse eines Unternehmens liegt in der zeitlichen Streuung der Finanzdatensätze, die für das Training des Modells verwendet wurden. Der älteste Datensatz ist die Bewertung des Unternehmens „Southern Copper Corp.“ vom 6. April 2010, während der neueste Datensatz, ein Rating des Unternehmens „POSCO“, vom 27. Dezember 2016 stammt. Dies ist problematisch, da die Ratingagenturen die Gewichtung der in die Bewertung einfließenden Kennzahlen anpassen, sowie neue Metriken hinzufügen. Capasso et al. zeigten, dass nach dem 2015 verabschiedeten Pariser Klimaabkommen seitens der Ratingagenturen ein erhöhter Fokus auf die CO₂-Emissionen der Unternehmen gelegt wurde (Capasso et al., 2020, S. 18-19). Da das Abkommen in den Zeitrahmen des Ratingdatensatzes fällt, ist es wahrscheinlich, dass Ratings ab Dezember 2015 mit aktualisierten Parametern erstellt wurden, während vor 2015 veröffentlichte Bonitätseinstufungen keine Umweltkriterien berücksichtigten. Da lediglich 13% der veröffentlichten Bonitätseinstufungen innerhalb des Datensatzes in den Zeitraum nach Abschluss des Pariser Klimaabkommens fallen, wäre die Entwicklung eines neuronalen Netzes auf Grundlage der 1017 verbleibenden Datensätze nicht praktikabel.

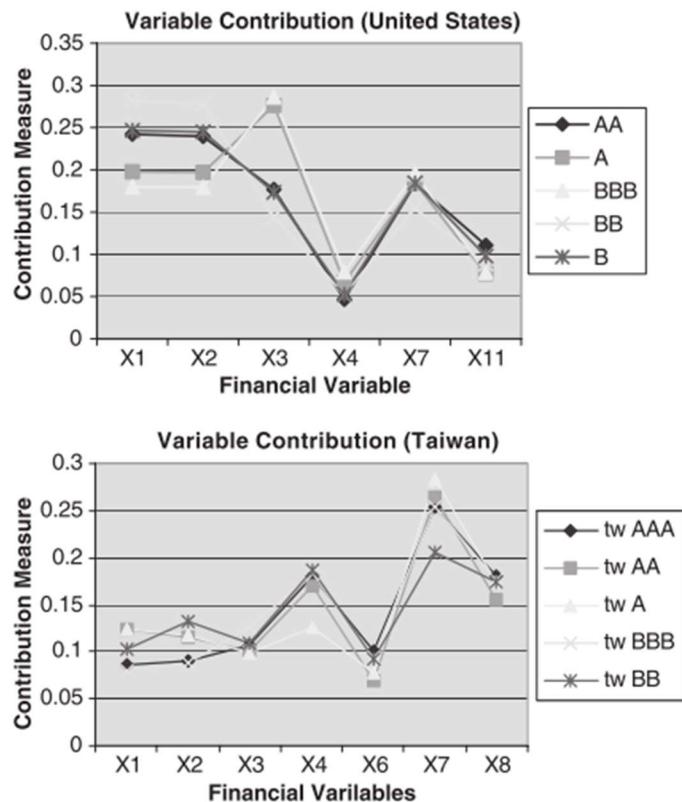
Eine Nebenauswirkung der teilweise veralteten Finanzkennzahlen ist die zeitliche Diskrepanz zwischen Finanz- und ESG-Daten. Während ältesten Finanzdaten des Datensatzes bis zu 13 Jahre alt sind, werden die ESG-Scores auf der Morningstar Sustainalytics Website jährlich aktualisiert (Morningstar Sustainalytics, 2020). Ältere ESG-Scores sind nicht öffentlich abrufbar. Während eine geringfügige zeitliche Diskrepanz zwischen Finanz- und ESG-Kennzahlen aufgrund von abweichenden Veröffentlichungszeiträumen unumgänglich ist, kann bei einer zeitlichen Abweichung von mehreren Jahren eine materielle Differenz zwischen aktuellem ESG-Score und dem ESG-Score zum Veröffentlichungszeitpunkt der Bonitätseinstufung entstehen. Daher sind die zur Entwicklung des neuronalen Netzes verwendeten ESG-Scores lediglich als Schätzwerte zu betrachten.

Eine weitere Möglichkeit zur Optimierung des Modells ist die Verwendung einer Datenquelle, welche die ESG-Kennzahlen in ihre drei Dimensionen Umwelt, Soziales und Governance unterteilt. Dies würde eine tiefere Analyse des Einflusses der verschiedenen ESG-Dimensionen auf die Bonitätseinstufungen ermöglichen. Anbieter, die in ihre Dimensionen getrennte ESG-Kennzahlen ausgeben, machen diese jedoch nicht ohne kostenpflichtigen API-Zugang abrufbar.

Weiterhin sind die über die gesamte Welt verstreuten Standorte der Unternehmen des Datensatzes suboptimal. In einem Vergleich zwischen US-amerikanischen Ratings und taiwanesischen Ratings zeigten Huang et al., dass Finanzkennzahlen von Unternehmen verschiedener Länder unterschiedlich signifikant sind (Huang et al., 2004, S. 555). Während in den USA über die beiden Features Gesamtvermögen und Gesamtverschuldung die Unternehmensgröße eine besonders

wichtige Rolle spielt, ist bei taiwanesischen Unternehmen die Profitabilität ausschlaggebend (Abbildung 53). Dies ist auf die unterschiedlichen Kulturen zurückzuführen: Während amerikanische Unternehmen aufgrund zeitgenössischer Wirtschaftstheorie einen höheren Verschuldungsgrad aufzeigen, sind ähnlich wie in anderen asiatischen Ländern taiwanesische Unternehmen aufgrund konservativer Unternehmenspolitik weniger stark verschuldet.

Abb. 53 Garson-Maß der Finanzkennzahlen der USA und Taiwan



Quelle: Übernommen aus Huang et al., 2004, S. 555.

Damit die höchstmögliche Präzision eines Machine-Learning Modells erreicht werden kann, sollten daher länderspezifische Daten in das Modell eingeführt werden. Dies könnte durch die Einführung einer weiteren kategorialen Kennzahl für die Kennzeichnung von Unternehmensdaten verschiedener Länder erfolgen. Um ein Modell zu entwickeln, dass für die Einstufung deutscher Unternehmen die höchstmögliche Präzision erreicht, sollte auf Daten zurückgegriffen werden, die bereits von deutschen Unternehmen veröffentlicht wurden. Hierbei entstehen jedoch einige Schwierigkeiten.

Gemäß § 325 Abs. 1 HGB haben Kapitalgesellschaften mit Sitz in Deutschland den Jahresabschluss, einen Lagebericht sowie den Bericht des Aufsichtsrats gegenüber dem Betreiber des Bundesanzeigers offenzulegen. Kleine Kapitalgesellschaften im Sinne des § 267 Abs. 1 HGB sind jedoch lediglich verpflichtet, eine Bilanz sowie den Anhang des Abschlusses einzureichen. Auch für mittelgroße Kapitalgesellschaften im Sinne des § 267 Abs. 2 HGB gibt es hinsichtlich der Publizitätspflichten Erleichterungen: Gemäß § 327 Nr. 1 HGB ist das Einreichen einer vereinfachten

Bilanz für diese Gesellschaften ausreichend. Für Kleinstkapitalgesellschaften, die mindestens zwei aus den drei in § 267a HGB genannten Kriterien erfüllen (Bilanzsumme bis zu 350 000 Euro, 700 000 Euro Jahresumsatz, durchschnittlich zehn Mitarbeiter oder weniger), ist eine dauerhafte elektronische Hinterlegung der Bilanz ausreichend. Für börsennotierte Kapitalgesellschaften gilt unabhängig der Rechtsform eine allgemeine Publizitätspflicht (§ 264d HGB i. V. m. § 1 PublG).

Obwohl für größere Kapitalgesellschaften durch die Publizitätspflicht eine Vielzahl von Informationen im Bundesanzeiger bereitgestellt werden, stellen diese Informationen im Rahmen der Anwendung des Machine Learning Modells nicht zwangsläufig einen Mehrwert dar, da die Bonität größerer Kapitalgesellschaften in der Regel bereits einer umfangreichen Analyse durch die Ratingagenturen unterliegt. Mittelständische Unternehmen hingegen, deren Bonität noch nicht durch eine Ratingagentur bewertet wurde, werden durch die Regelungen des HGB in ihren Publizitätspflichten erleichtert, sodass für diese Unternehmen weniger Finanzinformationen öffentlich bereitgestellt werden. Während die Entwicklung eines Bonitätseinstufungsmodells auf Basis von einfachen, durch alle Kapitalgesellschaften veröffentlichten Bilanzkennzahlen möglich ist, wird ein auf tiefergreifenden Finanzkennzahlen fußendes Modell (Gewinn- und Verlustrechnung, Kapitalflussrechnung, Eigenkapitalveränderungsrechnung) stets eine bessere Trennschärfe und damit eine höhere Prognoseleistung zeigen (siehe auch Abschnitt 4.2). Alle Kaufleute und Betriebe, die zur doppelten Buchführung verpflichtet sind, haben im Rahmen ihres Jahresabschlusses gemäß § 242 HGB sowohl eine vollständige Bilanz als auch eine Gewinn- und Verlustrechnung aufzustellen. Da diese Daten jedoch für Personengesellschaften nicht und für Kapitalgesellschaften nicht immer zu veröffentlichen sind, kann aufgrund von Monitoring-Kosten auf der Kreditorenseite ein Wohlfahrtsverlust entstehen, falls die Einholung der Informationen mit zusätzlichem Aufwand verbunden ist.

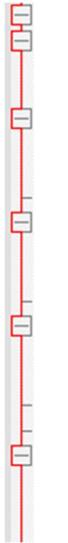
Eine weitere Verbesserungsmöglichkeit besteht in der Darstellung der im Bundesanzeiger veröffentlichten Daten. Während die Jahres- und Konzernabschlüsse durch die Suchfunktion auf der Internetseite Bundesanzeiger.de gefunden werden können, werden diese nicht in einer zur elektronischen Verarbeitung geeigneten Form dargestellt. Da die entsprechenden Abschlüsse sowohl Fließtexte als auch Darstellungen in tabellarischer Form beinhalten, wird eine isolierte Deduktion der Bilanz, Gewinn- und Verlustrechnung sowie anderen relevanten Finanzinformationen aus den Berichten erschwert. Darüber hinaus wird für den Bundesanzeiger keine offizielle API für die Entwicklung automatisierter Datenströme bereitgestellt. Mitglieder des deutschen CCC haben eine unabhängige Python-API entwickelt, welche ein Machine Learning-Modell verwendet, um CAPTCHA-Überprüfungen der Seite zu beantworten². Durch Modifikationen dieser API können die erforderlichen Finanzdaten innerhalb des Berichts isoliert und anschließend in tabellarischer Form heruntergeladen werden³. Da insbesondere Bilanzdaten hierarchisch angeordnet sind, sollten die

² GitHub-Seite: <https://github.com/bundesAPI/deutschland>

³ Modifizierte Version der Bundesanzeiger API für den isolierten Download von Finanzdaten: <https://github.com/leander-ms/deutschland/blob/main/src/deutschland/bundesanzeiger/bundesanzeiger.py>

Daten nicht innerhalb einer klassischen, relationalen Datenbank gespeichert werden. Stattdessen wird eine NoSQL Datenbank verwendet. Eine MongoDB-Datenbank speichert alle Daten im JSON-Dateiformat, um ihre hierarchische Struktur beizubehalten (Abbildung 54). Darüber hinaus besteht die Möglichkeit, verlorene Hierarchieebenen durch Skripte mit regulären Ausdrücken wiederherzustellen.

Abb. 54 Speicherung der Daten im JSON-Dateiformat



```
"1.2.3 Schadefrei GmbH": {
    "A. Anlagevermögen": {
        "31.12.2021": 310332.0,
        "31.12.2020": 110453.17,
        "I. Immaterielle Vermögensgegenstände": {
            "31.12.2021": 626.0,
            "31.12.2020": 1663.0
        },
        "II. Sachanlagen": {
            "31.12.2021": 309606.0,
            "31.12.2020": 108690.17
        },
        "III. Finanzanlagen": {
            "31.12.2021": 100.0,
            "31.12.2020": 100.0
        }
    },
    "B. Umlaufvermögen": {
        "31.12.2021": 636333.98,
        "31.12.2020": 365605.38,
        "I. Vorräte": {}
    }
}
```

Quelle: Eigene Darstellung.

Da die CCC-API jedoch keine offiziell unterstützte Lösung ist, besteht jederzeit die Möglichkeit, dass die Funktionalität des Skripts durch minimale Veränderungen an der Internetseite des Bundesanzeigers gebrochen wird. Daher sollte der Bundesanzeiger selbst eine API bereitstellen, die neben der Möglichkeit ganze Berichte in Textform herunterzuladen außerdem eine direkte Verbindung zu einer JSON-Datenbank bereitstellt, über die Finanzdaten direkt zugänglich sind.

Aufgrund der aktuellen Situation des Bundesanzeigers ist eine Verbesserung des öffentlich zugänglichen digitalen Angebots jedoch unwahrscheinlich. Aufgrund der Privatisierungspolitik des Bundes hält seit 1. Januar 2006 der Kölner Verlag DuMont Schauberg 100% der Anteile am Bundesanzeiger-Verlag (Spiegel, 2006). Dieser vermarktet eine elektronische Datenschnittstelle zu den im Bundesanzeiger durch die Unternehmen kostenpflichtig veröffentlichten Daten unter dem Namen „Validatis“ (NorthData, 2023). Aufgrund der Privatisierung und dem Aufbau einer kostenpflichtigen API fehlen Anreize, die Finanzdaten strukturiert über eine verbesserte Schnittstelle bereitzustellen. Im Zuge der Digitalisierung sollte der Gesetzgeber die privaten Betreiber digitaler Plattform verpflichten, neben der Veröffentlichung der Daten umfangreiche digitale Schnittstellen zur Verfügung zu stellen.

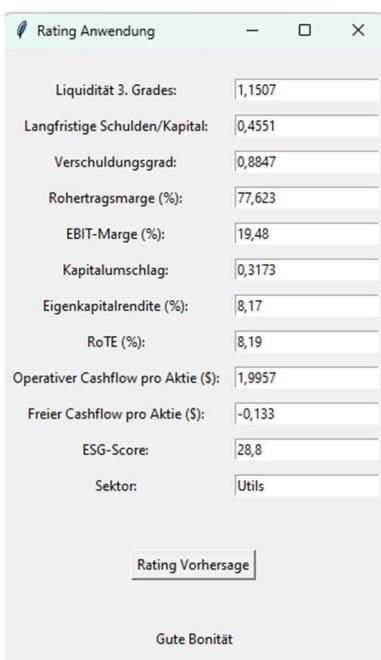
Abschließend ist zu beachten, dass das im Rahmen dieser Arbeit vorgestellte künstliche neuronale Netz nicht uneingeschränkt das bestmögliche Modell zur Vorhersage von Bonitätseinstufungen darstellt. Die Forschung an Machine Learning Modellen und neuronalen Netzen hat in den letzten Jahren neue Modellarten hervorgebracht. Hierzu zählen faltende neuronale Netze, ein Derivat der

bereits vorgestellten mehrschichtigen Perzeptron-Modelle (Golbayani et al., 2020, S. 4-5). Faltende neuronale Netze eignen sich besonders für die Verarbeitung von hierarchischen Daten und Rastern, wie sie in Bildern zu finden sind. Da die Verarbeitung hochauflösender Bilder eine hohe Menge an Eingangsvariablen bedeutet, sind faltende neuronale Netze auf eine effiziente Berechnung ausgelegt. Ein einfaches Bilderkennungsmodell wie das CIFAR-10 Modell verwendet Graustufenbilder mit $32 \times 32 = 1024$ Pixeln, welche anschließend als Eingangsvariablen für das Modell verwendet werden.

Golbayani et al. zeigten, dass zur Vorhersage von Bonitätseinstufungen keine pauschalen Aussagen über die Vorteilhaftigkeit verschiedener Netzwerkarchitekturen getroffen werden können (Golbayani et al., 2020, S. 12). Bei einem Vergleich der Performance von drei Modellen (mehrschichtiges Perzeptron-Modell, faltendes Netz und LSTM) in den Sektoren Finanzen, Energie und Gesundheit übertraf kein Modell durchweg alle anderen Modelle. Aufgrund des erhöhten Aufwands wurde im Rahmen dieser Arbeit von der Entwicklung weiterer Netzwerkarchitekturen abgesehen.

Darüber hinaus sollte eine Weiterentwicklung des vorgestellten Machine Learning Modells weitere Kennzahlen in seine Feature-Matrix aufnehmen, um die Präzision weiter zu steigern. Hierzu zählen weitere Finanzkennzahlen, wie die Gesamtgröße des Unternehmens gemessen am Gesamtvermögen, sowie nicht-finanzielle Kennzahlen wie die Qualität der Finanzberichterstattung (Abschnitt 2.2). Den Idealfall stellt ein Machine Learning Modell dar, welches alle Kennzahlen verwendet, die seitens der Ratingagenturen verwendet werden, sowie mithilfe der Bonitätseinstufungen der Ratingagenturen trainiert wird. Nach dem Training kann das Modell von den Daten der Ratingagenturen vollständig abgekoppelt werden und über eine Benutzeroberfläche für die Einstufung bislang unbewerteter Unternehmen verwendet werden (Abbildung 55).

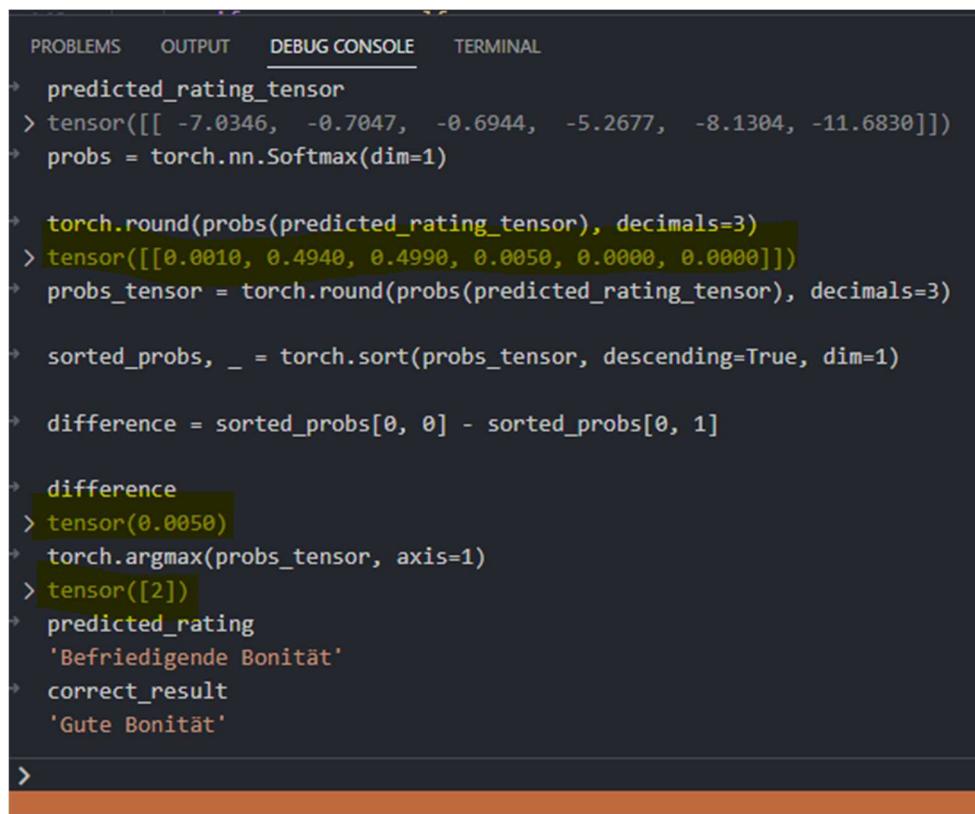
Abb. 55 Einfache Benutzeroberfläche des Bonitätseinstufungsmodells mit Tkinter in Python



Quelle: Eigene Darstellung.

Die Einzelauswertung der fehlklassifizierten Datensätze zeigt ebenfalls, dass die Trennschärfe des Modells durch die Einführung zusätzlicher Features profitieren würde, auch wenn diese Features nur einen geringfügigen Beitrag zur Vorhersagequalität leisten würden. Das Unternehmen Devon Energy Corporation, ein amerikanisches Energieunternehmen mit einer A-Bonitätseinstufung (Gute Bonität), wird vom neuronalen Netz als Unternehmen mit befriedigender Bonität (Index 2 aufgrund der nullbasierten Nummerierung in Python) eingestuft. Die Einführung der Ausgangsmatrix in eine klassische Softmax-Aktivierungsfunktion ermöglicht einen Überblick der vorhergesagten klassenbedingten Wahrscheinlichkeiten, bevor PyTorch mithilfe der „Argmax“-Funktion den Index mit der höchsten Wahrscheinlichkeit und damit die Vorhersage des Modells ausgibt. Der Blick auf diese Wahrscheinlichkeitsverteilung der Ausgangsmatrix zeigt, dass das Modell der Klasse „Befriedigende Bonität“ mit 49,9% lediglich eine 0,5% höhere Wahrscheinlichkeit zugeordnet hat als der Klasse „Gute Bonität“ (Abbildung 56).

Abb. 56 SoftMax-Output des PyTorch Modells zeigt für die Klasse mit Index 2 (Befriedigende Bonität) eine 0,5% höhere Wahrscheinlichkeit als für die korrekte Klasse mit Index 1 (Gute Bonität)



```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL
↳ predicted_rating_tensor
> tensor([[-7.0346, -0.7047, -0.6944, -5.2677, -8.1304, -11.6830]])
↳ probs = torch.nn.Softmax(dim=1)

↳ torch.round(probs(predicted_rating_tensor), decimals=3)
> tensor([[0.0010, 0.4940, 0.4990, 0.0050, 0.0000, 0.0000]])
↳ probs_tensor = torch.round(probs(predicted_rating_tensor), decimals=3)

↳ sorted_probs, _ = torch.sort(probs_tensor, descending=True, dim=1)

↳ difference = sorted_probs[0, 0] - sorted_probs[0, 1]

↳ difference
> tensor(0.0050)
↳ torch.argmax(probs_tensor, axis=1)
> tensor([2])
↳ predicted_rating
'Befriedigende Bonität'
↳ correct_result
'Gute Bonität'

>

```

Quelle: Eigene Darstellung.

4.3 Probleme mit Bonitätseinstufungen der Ratingagenturen

4.3.1 Bewertungskrise hypothekarisch besicherter Anleihen der 2000er Jahre

Das ursprüngliche Geschäftsmodell der Ratingagenturen folgte der Vision John Moodys aus 1909 und bestand aus dem direkten Verkauf von Ratings an Investoren (White, 2013, S. 102). In den späten 1960er und -70er Jahren änderten die großen Ratingagenturen ihr Geschäftsmodell: nicht die Konsumenten der Bonitätseinstufungen zahlen für die Dienste der Agenturen, sondern die

Herausgeber der Anleihen selbst. Die ausschlaggebendsten Gründe für diesen Wechsel waren die stetig steigende Komplexität der zu bewertenden Finanzprodukte sowie exponentielle Fortschritte der Informationstechnik, insbesondere im Bereich der Fotokopie (Medvedev, 2011, S. 6). Darüber hinaus erhöhte 1970 der prominente Konkurs der Eisenbahngesellschaft Penn Central, der zum Ausfall des Handelspapiers des Unternehmens führte, die Bereitschaft der Anleiheemittenten, ihre Anleihen bewerten zu lassen, um ihre Anleger zu beruhigen.

Dieses geänderte Geschäftsmodell stellt einen potenziellen Interessenskonflikt dar, da nun die Emittenten der Anleihen versuchen könnten, unter Androhung der Beauftragung von Mitbewerbern für zukünftige Emissionsvorhaben eine bessere Bewertung durch die Ratingagentur zu forcieren. Dieses Verhalten der Emittenten wird „Rating Shopping“ genannt (Benmelech & Dlugosz, 2010, S. 162). Trotz dieses Interessenskonflikts stellte das geänderte Geschäftsmodell für die Bewertung von Plain-Vanilla-Anleihen kein Problem dar (White, 2013, S. 107-108). Der Grund hierfür ist, dass einer Ratingagentur ihre langfristige Reputation wichtiger ist als die Aufträge eines einzelnen Emittenten. Aufgrund der tausenden Emittenten innerhalb des Plain-Vanilla-Anleihemarktes repräsentieren die Aufträge eines einzelnen Emittenten nur einen minimalen Anteil des Gesamtumsatzes einer Ratingagentur. Sollte jedoch die Ergebnisliste einer Ratingagentur gegenüber einem Mandanten aufgedeckt werden, würde dies dazu führen, dass zukünftige Bewertungen der Ratingagentur durch den Markt ignoriert werden, sowie zukünftige Emittenten die Beauftragung der Agentur nicht in Betracht ziehen.

Eine Ausnahme dieser Beobachtung stellen jedoch hypothekarisch besicherte Anleihen (MBS) dar. Das explosive Wachstum von nicht staatlich geförderten, hypothekarisch gesicherten Wertpapieren war einer der Hauptauslöser der globalen Finanzkrise von 2008 (He et al., 2011, S. 131). Die drei größten Ratingagenturen Moody's, Standard & Poor's und Fitch nahmen hierbei eine entscheidende Rolle ein (White, 2013, S. 105). Da die MBS deutlich undurchsichtiger und komplexer als Plain-Vanilla Anleihen waren, verließen sich Investoren zunehmend auf die Bewertungen der Ratingagenturen. Darüber hinaus konnten aufsichtlich regulierte Investoren wie Pensionskassen nur Anleihen erwerben, die zuvor durch eine staatliche anerkannte Ratingagentur bewertet wurden. Dies verringerte auf Seite der staatlich regulierten Investoren den Anreiz, eigene Untersuchungen durchzuführen.

Aufgrund des boomenden Immobilienmarktes Anfang der 2000er Jahre sowie stark ansteigenden Immobilienpreisen seit Anfang der 1990er Jahre herrschte am Finanzmarkt weitgehende Einigkeit über die Verlässlichkeit hypothekarisch gesicherter Wertpapiere (White, 2013, S. 105). Im Falle des Ausfalls eines Kreditnehmers konnte dieser das Haus noch immer mit einem Profit verkaufen, da Häuserpreise über einen langen Zeitraum stetig anstiegen, und kraft dessen die Hypothek abbezahlt werden. Da Hypotheken nur in seltenen Fällen nicht abbezahlt wurden, wurden die hypothekarisch gesicherten Wertpapiere, die aus diesen Hypotheken bestanden, als sichere Investitionen angesehen.

Unter diesen Umständen waren die Ratingagenturen bereit, mehrere hunderte Milliarden Dollar von MBS mit hohen Bewertungen zu versehen (White, 2013, S. 106). Diese Bewertungen waren jedoch äußerst optimistisch, besonders für Wertpapiere, die zwischen 2005 und 2007 ausgestellt wurden, als die Steigung der Immobilienpreiskurve bereits abnahm und die Qualität der ausgestellten Hypotheken einen Tiefpunkt erreichte.

Anders als bei Plain-Vanilla Anleihen hat das „Issuer Pays“ Geschäftsmodell der Ratingagenturen bei der Bewertung hypothekarisch gesicherter Wertpapiere eine entscheidende Rolle gespielt (US Senate Perm. Subcomm. Investig., 2011, S. 287). Moody's Chief Credit Officer Andy Kimball sagte in seiner Anhörung vor dem Kongress aus, dass die Existenz von Rating Shopping vor 2008 innerhalb der Industrie wohlbekannt war. Bereits in einer 2007 geführten Townhall sprachen hochrangige Moody's Mitarbeiter gemeinsam mit ihrem CEO offen über die gesunkenen Bewertungsstandards ihrer Mitbewerber. „*What happened in '04 and '05 with respect to subordinated tranches is that our competition, Fitch and S&P, went nuts. Everything was investment grade. It didn't really matter. It's all going into CBO.*“ (McDaniel et al., 2007, S. 63)

Ein Grund für das Zusammenbrechen des Bollwerks gegen Rating Shopping war der Aufbau des Markts für hypothekarisch besicherte Anleihen. Während der Plain-Vanilla-Anleihenmarkt aus einer Vielzahl von Emittenten besteht, gaben im MBS-Markt 12 Emittenten 80-90% der bewerteten Wertpapiere aus (White, 2013, S. 109). Darüber hinaus waren sowohl die Margen als auch die aggregierten Ankäufe wesentlich höher als im Plain-Vanilla-Anleihengeschäft. Dies resultierte darin, dass eine Drohung eines MBS-Emittenten, aktuelle und zukünftig emittierte Anleihen durch Mitbewerber bewerten zu lassen, ein wesentlich höheres Risiko darstellte. Zusätzlich machte die undurchsichtige Struktur der hypothekarisch gesicherten Wertpapiere die Aufdeckung des Rating Shoppings unwahrscheinlich.

4.3.2 Verzögerungen bei der Auf- und Abstufung von Bonitätseinstufungen

Eine häufige Beobachtung im Zuge der Insolvenz eines Unternehmens ist die zeitliche Verzögerung der Bewertungen der Ratingagenturen. So wiesen alle der großen Ratingagenturen Investment Grade Bewertungen auf die Anleihen der Investmentbank Lehman Brothers am Morgen der Konkurserklärung des Unternehmens aus (White, 2013, S. 109). Derartige Beobachtungen gehen jedoch bis in die 1930er Jahre zurück, zu einer Zeit, in der noch die Investoren für den Erhalt der Bewertungen zahlen mussten (Flandreau et al., 2009, S. 24).

Die Erklärung dieses Phänomens liegt jedoch in dem Ansatz, mit dem die Ratingagenturen Unternehmen bewerten. Die Ratingagenturen bewerten die Unternehmen im Einklang mit dem Wirtschaftszyklus und ignorieren kurzfristige Störeinflüsse, die nach Auffassung der Ratingagenturen mit hoher Wahrscheinlichkeit in der nächsten Zeit verschwinden (White, 2013, S. 109). Laut den Agenturen profitieren Anleiheinvestoren von diesem Vorgehen, da diese von genauen Bewertungen profitieren, jedoch gleichzeitig Transaktionskosten durch Verkaufen und Rückkaufen aufgrund von kurzfristigen Störfaktoren vermeiden möchten. Die Konsequenz dieses Ansatzes ist

jedoch, dass im Falle negativer Veränderungen in der finanziellen Performance eines Anleiheemittenten die Ratingagenturen abwarten müssen, ob diese Veränderung temporär oder anhaltend ist und damit eine Veränderung der Bonitätseinstufung des Emittenten rechtfertigt.

4.4 Diskussion und Vergleich vorgestellter Benchmarkmodelle mit Machine-Learning Modellen zur Prognose von Bonitätseinstufungen

Bislang verwenden viele Unternehmen Scorecards, die eine Reihe von mathematischen und statistischen Methoden nutzen, um einen Bonitätsscore für das Unternehmen zu ermitteln (Ahmed & Rajaleximi, 2019, S. 276-278). Diese sind häufig zunächst einfach zu implementieren und in ihrer Methodik verständlich, sodass diese Modelle über die Jahre fortlaufend angepasst und skaliert werden (siehe Abschnitt 4.5). Ein Nachteil der Scorecard ist jedoch, dass sie komplexe, nichtlineare Beziehungen zwischen den berücksichtigten Variablen nur unzureichend oder gar nicht abbilden kann.

Hierin liegt ein entscheidender Vorteil der im Rahmen dieser Arbeit vorgestellten Machine Learning Modelle. Das in Abschnitt 3.4 vorgestellte neuronale Netz ist ein 280 Tausend Parameter Modell, welches mithilfe seiner vielschichtigen Architektur komplexe Beziehungen zwischen den Variablen erfassen kann. Eine entscheidende Rolle spielt hierbei die Anwendung der nichtlinearen Tanh-Aktivierungsfunktion. Unabhängig von der Anzahl der Schichten des neuronalen Netzes würde ein Modell ohne Aktivierungsfunktion nur lineare Zusammenhänge erfassen können, da die Hintereinanderschaltung von linearen Funktionen stets in einer linearen Funktion resultiert. Die Ausführung der Tanh-Funktionen im Modell ermöglicht die Erfassung einfacherer Zusammenhänge in den Eingangsschichten, während die hinteren Schichten komplexe, nichtlineare Beziehungen zwischen den Merkmalen darstellen, indem sie auf den Ausgaben der Eingangsschichten aufbauen.

Darüber hinaus ist es für Machine Learning Modell einfacher, neue Features zu adaptieren und zu lernen. Wird ein neues Feature in die Feature-Matrix eingefügt, wird es automatisch in den Trainingsprozesses eingebunden. Das Modell lernt anschließend im Trainingsprozess die Beziehungen des neuen Features zur Zielvariable sowie die Verbindungen zu anderen Features, ohne das eine erneute manuelle Kalibrierung erforderlich ist. Im Falle eines neuronalen Netzes ist lediglich eine Anpassung der Eingangsschicht erforderlich, jedoch ist eine Automatisierung dieser Anpassung möglich (Abbildung 57).

Abb. 57 Die Variable "input_shape" der "__init__"-Methode der Modellklasse sorgt für eine automatische Erfassung der Anzahl von Features

```
class RatingsNet(nn.Module):
    def __init__(self, input_shape: int):
        super().__init__()
        self.hidden1 = nn.Linear(input_shape, 256)
        self.hidden2 = nn.Linear(256, 512)
        self.hidden3 = nn.Linear(512, 256)
        self.hidden4 = nn.Linear(256, 64)
        self.output = nn.Linear(64, 6)
        self.tanh = nn.Tanh()
        self.softmax = nn.LogSoftmax(dim=1)
        self.dropout = nn.Dropout(0.08, inplace=False)

model = RatingsNet(input_shape=X_train.shape[1]).to(device)
```

Quelle: Eigene Darstellung.

Damit liegt der größte Aufwand in der Erstellung und dem Aufbau des neuronalen Netzes. Ist diese Aufgabe abgeschlossen, kann es stetig durch die Einführung neuer Variablen verbessert werden. Im Gegensatz zu Machine Learning Modellen setzen traditionelle Scorecards häufig auf OLS-Regressionstechniken, die Interaktionsfunktionen nutzen, um die Beziehungen einzelner Variablen zueinander abzubilden (Malagueño et al., 2018, S. 14). In derartigen Modellen ist das Hinzufügen eines neuen Merkmals mit erheblichem Kalibrierungsaufwand verbunden.

Ein entscheidender Nachteil der Machine Learning Modelle im Vergleich zu traditionellen Scorecards liegt jedoch in der Interpretierbarkeit der Modelle. Obwohl durch neue Auswertungstechniken die Interaktion zwischen Features innerhalb der Modelle besser verstanden werden kann, bleiben die Modelle häufig eine „Black Box“, besonders für Mitarbeiter, die nicht an der Entwicklung des Modells mitgewirkt haben (Watson et al., 2019, S. 2). Aufgrund dieser Eigenschaften sind Techniken wie neuronale Netze oder Support Vector Machines im stark überwachten Bankgeschäft nicht zulässig (Thomas, 2009, zitiert nach Sadatrasoul, 2018, S. 93).

Darüber hinaus benötigen Machine Learning Modelle für ihr Training deutlich mehr Daten als traditionelle Scorecard Modelle. Für eine vollständige Entwicklung des Modells müssen die Daten in einen Trainings- und einen Testdatensatz aufgeteilt werden. Die Aufteilung erfolgt in der Regel nach einem 80/20-Split, sodass 80% der Daten für das Training und 20% der Daten für die anschließende Validierung verwendet werden. Mithilfe des Testdatensatzes kann die Performance des Modells anhand der Einführung bisher ungesiehener Daten gemessen werden. Daher sollte der Ursprungsdatensatz keine Duplikate enthalten, um ein Leck der Trainingsdaten in den Testdatensatz zu verhindern.

Neben dem Scorecard-Modell zur Ermittlung der Bonität potenzieller Debitoren wurde im Rahmen dieser Arbeit das Z-Score Modell von Altman vorgestellt (siehe Abschnitt 2.1). Dies ermöglicht die Ermittlung der Bonität anhand des Einsetzens fünf finanzieller Kennzahlen in die Z-Gleichung.

Unternehmen, deren Z-Wert unter 1,81 fällt, fallen voraussichtlich aus, während Unternehmen mit einem Z-Wert von über 2,99 laut des Z-Score Modells nicht ausfallen werden (Altman, 1968, S. 606).

Ein Vorteil des Z-Score Modells gegenüber dem im Rahmen dieser Arbeit entwickelten neuronalen Netzes zur Vorhersage von Bonitätseinstufungen ist jedoch die einfache Interpretierbarkeit des Modells. Durch die festgelegten Koeffizienten innerhalb des Z-Score Modells ist nachzuvollziehen, wie die einzelnen Koeffizienten zur Vorhersage des Modells beitragen. Darüber hinaus verwendet das Modell Kennzahlen, die nicht nur für Kapitalgesellschaften, sondern für jedes Unternehmen ermittelt werden können. Auch das Resultat des Modells (der Z-Wert) ist einfach zu interpretieren, da durch die Kategoriale Natur des Modells lediglich zwischen insolvent und nicht-insolvent unterschieden wird. Dies ist ein wesentlicher Vorteil gegenüber dem im Rahmen dieser Arbeit entwickelten Vorhersagemodell für Bonitätseinstufungen, da dies über die Einstufung der Bonität eine ordinal interpretierbare, prognostizierte Ausfallwahrscheinlichkeit ausgibt. Diese kann jedoch, ähnlich zum Z-Score, auch kategorial interpretiert werden.

Ein grundlegendes Problem des Z-Score Modells liegt jedoch in der Kalibrierung der Koeffizienten (Bemmann, 2005, S. 74). Die für den Aufbau der Studie verwendeten Unternehmen wurden nicht zufällig ausgewählt (Shumway, 2001, S. 19). So fand Shumway in seinen Untersuchungen, dass aufgrund einer Selektionsverzerrung in Altmans Studien lediglich die beiden Variablen EBIT/Bilanzsumme sowie Marktwert des Eigenkapitals/Fremdkapital statistisch signifikante Kennzahlen sind.

Dies führt dazu, dass das Z-Score Modell als multivariates Modell keine besseren Prognosen trifft als univariate Modelle, die lediglich eine Kennzahl verwenden (Bemmann, 2005, S. 74). Bemmann führt die Popularität des Altmanschen Z-Score Modells als Vergleichsmodell darauf zurück, dass es vergleichsweise „einfach zu schlagen“ sei. Darüber hinaus sind Versuche gescheitert, das Z“-Score Modell auf weitere Märkte wie den deutschen Markt zu übertragen, da auch nach einer Kalibrierung kein linearer Zusammenhang zwischen der Ausfallwahrscheinlichkeit deutscher Unternehmen und deren Z“-Score festgestellt werden konnte (Beinert et al., 2006, S. 8-14).

Dies ist ein wesentlicher Vorteil des im Rahmen dieser Arbeit entwickelten Machine Learning Modells. Sind Trainingsdaten für ein Land verfügbar, ist es möglich, das Modell anhand dieser Daten spezifisch für dieses Land zu trainieren und zu optimieren. Das länderspezifische Training kann durchgeführt werden, ohne, dass eine vollständige Neuprogrammierung erforderlich ist. Dies ist besonders wichtig aufgrund der in Abschnitt 4.1 erwähnten Gesichtspunkte, dass diverse Finanzkennzahlen in verschiedenen Kulturen einen unterschiedlichen Stellenwert einnehmen.

Ein Vorteil des Altmanschen Z-Score bleibt die einfache Anwendung und Handhabung des Modells. Allerdings gilt zu beachten, dass das Modell lediglich fünf Finanzkennzahlen (keine nicht-finanziellen Kennzahlen) integriert und seine Kalibrierung einige Jahrzehnte zurückliegt. Ein weiteres Problem des Modells liegt in der Interpretation des Z-Werts. Fällt dieser zwischen 1,81 und 2,99, befindet sich

das Unternehmen in der „Grauzone“, wodurch das Modell keine Aussage über diese Unternehmen treffen kann.

Ein weiteres im Rahmen dieser Arbeit vorgestellte Benchmarkmodell ist die Auswertung von Insolvenzhäufigkeiten von Unternehmen in Deutschland. Mithilfe von Daten des statistischen Bundesamtes konnte Bemann eine multivariate Auswertung der Insolvenzhäufigkeiten verschiedener Rechtsformen innerhalb der verschiedenen Industrien durchführen (Bemann, 2005, S. 51-60). Durch diese Auswertungen konnte eine Präzision der Ausfallprognosen von 45%-55% erzielt werden (Bemann, 2005, S. 59). Ein Vorteil dieses Modells liegt in der Genauigkeit der Insolvenzprognosen. Anhand der Auswertungen historischer Daten lässt sich exakt quantifizieren, wie viel höher oder niedriger die Ausfallwahrscheinlichkeit eines Unternehmens mit bestimmter Rechtsform im Vergleich zu anderen Unternehmen einer anderen Industrie mit anderer Rechtsform ist. Damit ist das Modell sowohl höchst interpretierbar als auch in seiner Errechnung der Ausfallwahrscheinlichkeiten nachvollziehbar. Die für das Modell erforderlichen Daten können jedes Jahr erneut beim statistischen Bundesamt angefragt werden.

Ein entscheidender Nachteil dieses Modells ist jedoch, dass für seine Ausfallwahrscheinlichkeiten lediglich makro-ökonomische Daten heranzieht, ohne unternehmensspezifische Kennzahlen zu verwenden. Damit nimmt das Modell an, dass Unternehmen gleicher Rechtsform innerhalb eines Industriesektors die gleiche Ausfallwahrscheinlichkeit besitzen, eine unrealistische Annahme. Fallen gemäß der Vorhersage eines solchen Modells 3,5% aller GmbHs im Baugewerbe aus, würde jeder GmbH im Baugewerbe eine prognostizierte Ausfallwahrscheinlichkeit von 3,5% zugeschrieben werden. Damit zeigt das Modell zwar eine perfekte Kalibrierung, gibt jedoch keine trennscharfen Prognosen aus (s. Abschnitt 2.1).

Denkbar ist jedoch die Aufnahme der Rechtsform eines Unternehmens in das neuronale Netz für Bonitätseinstufungen als zweite kategoriale Kennzahl neben dem Industriesektor. Die Auswertungen Bemanns zeigen, dass Unternehmen verschiedener Rechtsformen in Kombination mit dem Industriesektor deutlich unterschiedliche Ausfallraten zeigen. Damit könnte die Rechtsform besonders bei der Entwicklung eines Machine Learning Modells für die Prognose der Bonität deutscher Unternehmen eine wichtige Rolle spielen.

Das letzte im Rahmen dieser Arbeit vorgestellte Benchmarkmodell ist das Merton Distance-to-Default Modell. Dieses errechnet mithilfe nichtlinearer Gleichungen eine Ausfallwahrscheinlichkeit des Unternehmens auf der Basis des Werts und der Volatilität des Eigenkapitals (Bharath & Shumway, 2008, S. 1344). Ein Vorteil des Distance-to-Default Modells gegenüber dem im Rahmen dieser Arbeit entwickelten neuronalen Netz zur Vorhersage von Bonitätseinstufungen liegt in der Geschwindigkeit der Verarbeitung von Informationen. Während das entwickelte Machine-Learning Modell auf der Basis von Bonitätseinstufungen der großen Ratingagenturen entwickelt wurde, welche Verschlechterungen von Finanzinformationen absichtlich mit einer zeitlichen Verzögerung berücksichtigen (s. Abschnitt 4.3.2), kann das Distance-to-Default Modell durch die Verarbeitung von

Aktienmarktinformationen neue Informationen zeitnah reflektieren (Bharath & Shumway, 2008, S. 1346).

Jedoch hat auch die Anwendung des Merton Distance-to-Default Modell einige Nachteile. Bharath und Shumway stellen in ihrer Analyse des Modells heraus, dass es sich bei dem Modell um ein äußerst untypisches Modell handelt, dass nicht den klassischen ökonometrischen Techniken folgt (2008, S. 1345). Aus diesem Grund ist unklar, wie das Modell um zusätzliche Kennzahlen erweitert werden könnte. Zusätzlich gilt zu beachten, dass der Marktwert des Eigenkapitals im Distance-to-Default Modell einen signifikanten Stellenwert einnimmt (Bharath & Shumway, 2008, S. 1345). Sinkt der Marktwert des Eigenkapitals, nimmt die seitens des Modells ausgegebene Ausfallwahrscheinlichkeit zu. Kurz vor der Insolvenz des Energiekonzerns Enron fiel der Kurs der Aktie deutlich, nachdem dem Unternehmen erstmalig Bilanzfälschung in großem Ausmaß vorgeworfen wurde (Bharath & Shumway, 2008, S. 1346). Folglich stieg mit dem fallenden Aktienkurs des Unternehmens die Ausfallwahrscheinlichkeit. Die vor der Veröffentlichung der Accounting-Probleme durch das Distance-to-Default ausgegebene Ausfallwahrscheinlichkeit war jedoch deutlich geringer, als es die Bonitätseinstufungen der Ratingagenturen signalisiert hatten. Dieser Vorfall illustriert sowohl die Vorteile als auch die Nachteile des Distance-to-Default Modells: Sind alle Voraussetzungen erfüllt, können präzise Ausfallwahrscheinlichkeiten ausgegeben werden, die aktuelle Ereignisse einpreisen. Damit das Modell jedoch funktionieren kann, müssen alle Annahmen des Merton Modells selbst sowie die Markteffizienzhypothese vollständig erfüllt sein (Bharath & Shumway, 2008, S. 1346).

Die im Rahmen dieses Abschnitts analysierten Modelle unterscheiden sich grundsätzlich in ihrer Anwendung und Handhabung. Obwohl sie unterschiedliche Kennzahlen kombinieren und auf unterschiedliche Art und Weise entwickelt wurden, können alle Modelle eingesetzt werden, um die Eignung potentieller Geschäftspartner und Debitoren festzustellen. Nichtsdestotrotz bleibt anzumerken, dass die Bonitätsbewertungen der Ratingagenturen in der Unternehmenswelt nach wie vor als maßgeblicher Referenzpunkt für die Evaluierung des Ausfallrisikos externer Unternehmen angesehen werden. „Um das Ausfallrisiko auf ein Minimum zu begrenzen, schließen wir im Finanzierungsbereich Geschäfte grundsätzlich nur mit Kontrahenten ab, deren Kredit-Rating mindestens BBB+/Baa1 ist und betreiben zudem ein aktives Limit-Management“ (Deutsche Telekom, 2023, S. 161). Das im Rahmen dieser Arbeit entwickelte Modell vereint die Stärken traditioneller Bonitätseinstufungen von Ratingagenturen mit den innovativen Möglichkeiten des Machine Learning. Aufgrund seiner strukturierten Architektur ermöglicht es die nahtlose Integration zusätzlicher Kennzahlen und kann nicht-lineare Beziehungen zwischen diesen Variablen effizient erfassen. Bei Vorliegen der relevanten Unternehmensdaten ermöglicht es eine Ermittlung der Bonität jedes betrachteten Unternehmens. Die Vorgaben der Telekom (mindestens BBB+) zeigen, dass Bonitätseinstufungen in der Praxis trotz ihrer ordinalen Natur meist kategorial interpretiert werden, indem eine Schwellenkategorie identifiziert wird. Dies kann in der individuellen

Ausgestaltung eines Machine Learning Modells berücksichtigt werden, da bei Skalenniveaus stets die Möglichkeit besteht, auf geringere Niveaus zurückzustufen.

5. Fazit

Ziel der vorliegenden Arbeit war aufzuzeigen, welchen Beitrag intelligente Analyseverfahren mithilfe der Anwendung von Machine-Learning Modellen bei der Ermittlung von Forderungsausfallwahrscheinlichkeiten leisten können. Hierzu wurde zunächst eine umfangreiche Literaturrecherche zu Insolvenzprognosemodellen unterschiedlicher Skalenniveaus und Kennzahlen durchgeführt. Frühe Insolvenzprognosemodelle, wie der Altmansche Z-Score, nutzten ausschließlich Finanzkennzahlen, um eine binäre Vorhersage über die Insolvenzwahrscheinlichkeit eines Unternehmens zu treffen (Unternehmen X fällt aus oder nicht). Durch umfangreiche Forschung im Bereich der Insolvenzprognosemodelle konnten in den vergangenen Jahrzehnten Modelle entwickelt werden, die eine Vielzahl von Kennzahlen und Eigenschaften der Unternehmen aufgreifen sowie detailliertere Ergebnisse über die aktuelle Situation des Unternehmens liefern, sodass ein direkter Vergleich verschiedener Unternehmen ermöglicht wird.

Ratingagenturen verwenden ordinale Insolvenzprognosemodelle, um mithilfe von multivariaten Verfahren Unternehmen auf einer 7- bzw. 17-stufigen Skala einzurordnen. Diese Modelle enthalten dabei nicht nur Finanzkennzahlen, sondern greifen auf Kennzahlen aus Umwelt-, Sozial- und Governance-Bereichen zurück. Hierfür werden seit einiger Zeit von verschiedenen Anbietern ESG-Kennzahlen herausgegeben, welche die Unternehmen hinsichtlich dieser Kriterien bewerten. Aktuelle Studien legen nahe, dass diese ESG-Metriken zunehmend Berücksichtigung in den Bonitätsbewertungen der Ratingagenturen finden, da Unternehmen, die in Bezug auf ESG-Kriterien schlecht abschneiden, ein erhöhtes Risiko für weitere operative und finanzielle Probleme aufweisen.

Jedoch sind die von Ratingagenturen vergebenen Bonitätseinstufungen kritisch zu betrachten. Ein entscheidendes Problem ist das „Issuer Pays“ Geschäftsmodell der Ratingagenturen. Das bedeutet, dass lediglich Unternehmen bewertet werden, welche die Ratingagentur zuvor für die Erstellung eines Ratings beauftragt haben. Daraus resultiert, dass besonders kleinere Unternehmen seltener von einer Ratingagentur bewertet werden, entweder weil sie keine Anleihen ausgeben oder weil die Kosten für ein solches Rating unerschwinglich sind. Da jedoch auch kleinere Unternehmen Kredite und Zahlungsziele in Anspruch nehmen, muss auch die Bonität dieser Unternehmen bewertet werden.

Zu diesem Zweck werden bislang meist Scorecard-Modelle verwendet, welche OLS-Regressionstechniken und eine feste Kalibrierung der einzelnen Faktoren verwenden, um die Bonität der Debitoren zu erfassen.

Bereits in den frühen 2000er Jahren zeigten Untersuchungen, dass Machine Learning-Modelle in Kombination mit ausgewählten Finanzkennzahlen die Bonitätseinstufungen von Ratingagenturen mit hinreichender Genauigkeit vorhersagen konnten. Da Ratingagenturen vermehrt nicht-finanzielle

Kennzahlen berücksichtigen, sollten aktuelle Machine Learning-Modelle zur Vorhersage von Bonitätseinstufungen Kennzahlen aus dem ESG-Bereich berücksichtigen. Für das im Rahmen dieser Arbeit entwickelte Machine Learning-Modell wurde zunächst der Datensatz, der für das Training aller getesteten Modelle verwendet wird, untersucht. Hierbei wurden zunächst die 17 Ratingstufen in sieben unterschiedliche Ratingkategorien aufgeteilt, um das Training der Modelle zu erleichtern. Darüber hinaus wurden umfangreiche Analysen zur prozentualen Verteilung der Unternehmen auf die Bonitätsklassen innerhalb der einzelnen Sektoren sowie eine Korrelationsmatrix zur Analyse aller im Datensatz enthaltenen Kennzahlen aufgebaut.

Nach Durchführung der Korrelationsanalyse wurden zehn Finanzkennzahlen aus dem ursprünglichen Datensatz für das Training des Modells ausgewählt. Des Weiteren wurde der Sektor einbezogen, um mögliche Interaktionen zwischen den Kennzahlen, der Zielvariable und dem Sektor zu berücksichtigen. Da der Testdatensatz sowohl den Aufbau einer Feature-Matrix mithilfe der Finanzkennzahlen als auch die Zielvariable (die Ratingkategorie für jeden Datensatz) enthielt, kann auf Techniken des überwachten Lernens zurückgegriffen werden.

Unter den verschiedenen eingesetzten Techniken des überwachten Lernens weist das neuronale Netz die höchste Genauigkeit auf. Dieses Netz wurde mit dem Framework PyTorch implementiert und verknüpft nach Eingabe der Features über 280.000 Parameter mit der Zielvariable. Ein signifikanter Nachteil neuronaler Netze ist der notwendige Optimierungsaufwand, um optimale Ergebnisse zu erzielen. Aufgrund der hohen Anzahl an zu optimierenden Parametern im Modell sind häufig mehrere Tausend Durchläufe erforderlich, um die bestmöglichen Hyperparameter zu bestimmen. Nach der erfolgreichen Optimierung des Modells zeigte es jedoch eine nochmals verbesserte Performance gegenüber dem zuvor leistungsstärksten Modell, dem ExtraTrees-Modell. Insbesondere in Bonitätsklassen mit einer hohen Anzahl an Trainingsdatensätzen erzielte das neuronale Netz bessere Ergebnisse als die zuvor evaluierten Machine Learning-Methoden.

Durch eine umfangreiche Analyse konnte mithilfe der SHAP-Technik herausgestellt werden, welche Features in welchen Klassen besonders zur Ausgabe der Klassenwahrscheinlichkeiten beitrugen. Hierbei konnte festgestellt werden, dass besonders der langfristige Verschuldungsgrad eine entscheidende Rolle bei der Klassifizierung der Unternehmen spielt. Sowohl in der Permutationsanalyse als auch in der SHAP-Auswertung wurde diese Kennzahl konstant als eine der aussagekräftigsten Kennzahlen bewertet. Auch die Liquidität dritten Grades und die Rohertragsmarge wurden in den verschiedenen Auswertungsmethoden als aussagekräftige Kennzahlen identifiziert.

Um herauszustellen, inwiefern diese Kennzahlen die Logit-Wahrscheinlichkeit einer Einstufung in die einzelnen Klassen beeinflussen, wurde zudem eine klassenbasierte SHAP-Auswertung der Kennzahlen vorgenommen.

Die klassenbasierte Auswertung der Kovariaten zeigt, dass besonders bei Kennzahlen zur Verschuldung hohe Werte zu einer geringeren Wahrscheinlichkeit der Einstufung des Unternehmens

in die Bonitätsklasse „Sehr Hohe Bonität“ führt. Unerwartet war dagegen, dass ein hoher Wert der Liquidität 3. Grades ebenfalls zu einer geringeren Logit-Wahrscheinlichkeit für die Einstufung in die Klasse „Sehr Hohe Bonität“ führt. Obwohl eine hohe Liquidität typischerweise positiv angesehen wird, kann eine zu hohe Liquidität auf große Lagerbestände aufgrund ausbleibender Verkäufe zurückgeführt werden. Ein Blick auf den Testdatensatz zeigt, dass obwohl alle Unternehmen oberhalb des untersten Terzils die goldene Bilanzregel einer Liquidität 3. Grades von mindestens 120% erfüllen, bei Unternehmen mit einem schlechten Rating oft eine besonders hohe Liquidität 3. Grades festzustellen ist.

Auf der Gegenseite bedeuten eine hohe Rohertragsmarge, hohe EBIT-Marge, ein hoher Kapitalumschlag und eine hohe Eigenkapitalrendite eine höhere Logit-Wahrscheinlichkeit der Einordnung eines Unternehmens in die Ratingkategorie „Sehr Hohe Bonität“. Für alle anderen Klassen kann ebenfalls eine SHAP-Auswertung vorgenommen werden, jedoch fallen insbesondere für die Klassen im mittleren Bereich (Angespannte Bonität und Befriedigende Bonität) die absoluten SHAP-Werte geringer aus. Darüber hinaus sind auch Änderungen in der Wichtigkeit einzelner Kovariaten in den einzelnen Klassen zu erkennen. Während in vier der sechs Klassen die langfristige Verschuldung die ausschlaggebendste Kennzahl ist, ist in der Klasse Mangelhafte Bonität die EBIT-Marge am aussagekräftigsten, während in der Klasse Ungenügende Bonität die Liquidität 3. Grades am bedeutendsten ist.

Anschließend konnten mithilfe eines eigens entwickelten Scraping-Tools für die Website Morningstar Sustainalytics die ESG-Kennzahlen der im Trainingsdatensatz enthaltenen Unternehmen abgefragt werden. Obwohl aufgrund des Fehlens vergangener ESG-Kennzahlen keine exakten Auswertungen vorgenommen werden können, da einige Daten aus dem Trainingsdatensatz bereits mehrere Jahre alt sind, kann dennoch eine Analyse über Schätzwerte unter der Annahme der geringfügigen Änderung von ESG-Kennzahlen über die Jahre vorgenommen werden. Mithilfe der Daten der Sustainalytics Website konnten zwei zusätzliche Kennzahlen dem Trainingsdatensatz hinzugefügt werden: Der ESG-Score selbst, sowie der ESG-Score im Verhältnis zum Sektordurchschnitt. Gemäß der Logik der Morningstar Sustainalytics Website bedeutet ein höherer ESG-Score ein schlechteres ESG-Rating und somit eine größere Aussetzung des Unternehmens gegenüber ESG-Risiken. Damit ist auch ein hoher ESG-Score im Verhältnis zum Sektordurchschnitt negativ einzuordnen.

Eine erneute SHAP-Auswertung zeigte, dass beide Kennzahlen bei der Einstufung eines Unternehmens mithilfe des neuronalen Netzes beitragen können. Während jedoch der Einfluss des reinen ESG-Scores nicht immer eindeutig ist, zeigt besonders die Addition der Kennzahl „ESG-Score im Verhältnis zum Sektordurchschnitt“ positive Ergebnisse. Obwohl einige prominente Unternehmen des Trainingsdatensatzes wie Exxon Mobil mit einem hohen ESG-Score eine überdurchschnittlich gute Bonitätseinstufung erreichen, zeigt die SHAP-Auswertung dass ein im Sektorverhältnis hoher ESG-Score zu einer geringeren Wahrscheinlichkeit einer guten Bonitätseinstufung führt. Besonders bei der Abstufung zwischen Investment-Grade- (mindestens Kategorie „Befriedigende Bonität“) und

Nicht-Investment-Grade-Unternehmen trägt die Kovariate ESG-Score zu Sektordurchschnitt einen entscheidenden Teil zur Trennung bei.

Im Kontext der Bonitätsvorhersage für Unternehmen weist das im Rahmen dieser Arbeit entwickelte Modell einige Limitationen auf. Eine solche Limitation betrifft den für das Training des Modells verwendeten Datensatz. Dieser umfasst nicht nur Daten von deutschen Unternehmen, sondern von Firmen aus aller Welt. Frühere Studien haben allerdings aufgezeigt, dass die Übertragbarkeit zwischen verschiedenen Ländern und, noch bedeutsamer, unterschiedlichen Kulturen limitiert ist. Während in westlichen Kulturen ein gewisser Grad an Unternehmensverschuldung als akzeptabel angesehen wird, betrachten Kulturen im asiatischen Raum einen erhöhten Verschuldungsgrad deutlich kritischer.

Eine zusätzliche Limitation geht aus den für das Training verwendeten Kennzahlen hervor. Diese wurden auf Grundlage einer Korrelationsanalyse sowie einer umfassenden Literaturrecherche von Kaur et al. aus dem ursprünglichen Kennzahlensatz ausgewählt. Allerdings beschränkte sich der Trainingsdatensatz ausschließlich auf Bewertungen für kapitalmarktorientierte Unternehmen, weshalb einige der verwendeten Kennzahlen ausschließlich für kapitalmarktorientierte Unternehmen ermittelt werden können. Hieraus ergeht eine weitere Einschränkung der Übertragbarkeit des Modells auf nicht-kapitalmarktorientierte Unternehmen.

Das im Rahmen dieser Arbeit vorgestellte Modell stellt einen bedeutenden Schritt in der Anwendung neuronaler Netze zur Vorhersage von Bonitätseinstufungen von Unternehmen dar. Es offenbaren sich im Rahmen dieser Arbeit bestimmte Potenziale für zukünftige Forschungsvorhaben, sodass im Hinblick auf weitergehende Untersuchungen einige Empfehlungen und Aspekte zu nennen sind, die zur Verbesserung und Erweiterung des entwickelten Modells beitragen könnten. Zunächst kann die Präzision des Modells durch das Hinzufügen zusätzlicher Kovariaten weiter erhöht werden. Dies betrifft insbesondere die nicht-finanziellen Kennzahlen, da das im Rahmen dieser Arbeit vorgestellte Modell mit dem ESG-Score lediglich eine einzelne nicht-finanzielle Kennzahl umfasst, deren Ziel es ist, verschiedene Kennzahlen zusammenzufassen. Darüber hinaus sollte der Trainingsdatensatz Daten von nicht-kapitalmarktorientierten Unternehmen erfassen, um die Generalisierbarkeit des Modells zu erhöhen. Zur Extrapolation von Kennzahlen, die für nicht-kapitalmarktorientierte Unternehmen nicht verfügbar sind, können unterstützende Datenaufbereitungsmodelle entwickelt werden.

Um die Bonität deutscher Unternehmen präzise zu bestimmen, sollte ein Trainingsdatensatz erstellt werden, der ausschließlich Unternehmensdaten aus Deutschland umfasst. Hierzu sollten zunächst die in Abschnitt 4.4 vorgeschlagenen Verbesserungen umgesetzt werden, um die allgemeine Verfügbarkeit von Finanzinformationen zu erhöhen. Aktuell setzen die meisten Unternehmen zur Bonitätsbewertung potenzieller Debitoren Scorecard-Modelle ein. Im Rahmen einer fortgeschrittenen Analyse sollte die Präzision des entwickelten Modells mit der eines traditionellen Scorecard-Modells verglichen werden.

Darüber hinaus sollten die genauen Wirkungsbeziehungen der finanziellen- und nicht-finanziellen Kennzahlen durch Aufbau eines Zeitreihendatensatzes berücksichtigt werden. Die im Rahmen dieser Arbeit vorgestellte Literatur zeigt, dass beim Einfluss von Finanzkennzahlen auf das Rating des Unternehmens eine Zeitverzögerung zu betrachten ist, nicht-finanzielle Kennzahlen jedoch bei ihrer Veröffentlichung das Rating direkt beeinflussen können. Die Entwicklung eines Zeitreihenmodells kann hierbei helfen, besser zu verstehen in welchen Zeitachsen die verschiedenen Arten von Kennzahlen das Bonitätsrating beeinflussen.

Abschließend lässt sich feststellen, dass es sich bei dem in diesem Modell vorgestellten Insolvenzprognosemodell lediglich um ein Bonitätsprognosemodell handelt, welche die Bonität der verwendeten Unternehmen prognostiziert und dabei eine ordinal interpretierbare Bonitätskategorie herausgibt. Durch die Exploration verschiedener Modelle, gemeinsam mit der Einbindung weiterer finanzieller- und nicht-finanzieller Kennzahlen, sollten zukünftige Ansätze die Entwicklung eines Modells verfolgen, das Ausfallwahrscheinlichkeiten noch genauer vorhersagen kann.