fhwedel
UNIVERSITY OF APPLIED SCIENCES

***

***

Submission Date:

19. November 2025

Submitted by:
Reebal Sami 106441

Supervisor:
Prof. Dr. Franziska Bönte
Fachhochschule Wedel
Feldstraße 143
22880 Wedel
Phone: 04103 - 8048 - 35
E-Mail: franziska.boente@fh-wedel.de

Contents

List of Figures

List of Tables

1 Einleitung

2 Theoretischer Hintergrund und Stand der Forschung

2.1 Historische Entwicklung der Insolvenzprognose

2.2 Kennzahlenbasierte Früherkennung

2.3 Machine Learning in der Insolvenzprognose

2.4 Methodische Herausforderungen

3 Daten und Methodik

Die Entwicklung eines robusten Frühwarnsystems für Unternehmenskrisen erfordert nicht nur leistungsfähige Algorithmen, sondern vor allem eine fundierte Datenbasis. Bevor maschinelle Lernverfahren ihre prädiktive Kraft entfalten können, bedarf es einer systematischen Analyse der Rohdaten – ein Schritt, der in der Praxis häufig unterschätzt wird, jedoch maßgeblich über Erfolg oder Misserfolg eines Projekts entscheidet Goodfellow, Bengio, and Courville, 2016.

Dieses Kapitel dokumentiert den methodischen Ansatz dieser Arbeit in vier Phasen: Die **Foundation-Phase** (Abschnitt 3.1) charakterisiert die Datenbasis umfassend, identifiziert Qualitätsprobleme und legt damit den Grundstein für alle weiteren Schritte. Die **Data Preparation-Phase** (Abschnitt 3.2) behandelt fehlende Werte, Duplikate und Ausreißer nach evidenzbasierten Methoden. Die **Feature Engineering-Phase** (Abschnitt 3.3) adressiert Multikollinearität und selektiert relevante Prädiktoren. Schließlich beschreibt die **Modellierungsphase** (Abschnitt 3.4) die eingesetzten Machine-Learning-Verfahren und deren Evaluation.

Diese strukturierte Vorgehensweise gewährleistet Transparenz und Reproduzierbarkeit – zwei zentrale Gütekriterien wissenschaftlicher Forschung Hastie, Tibshirani, and Friedman, 2009.

3.1 Datenbasis: Foundation-Phase

Die Früherkennung von Unternehmenskrisen gleicht der Suche nach der Nadel im Heuhaufen: Unter tausenden Finanzkennzahlen gilt es, jene Muster zu identifizieren, die auf eine drohende Insolvenz hindeuten. Doch bevor maschinelle Lernverfahren diese Aufgabe übernehmen können, bedarf es einer gründlichen Analyse der Datenbasis. Die folgenden Abschnitte dokumentieren systematisch die Eigenschaften des verwendeten Datensatzes – einschließlich identifizierter Probleme und getroffener Annahmen. Diese Transparenz ist essentiell, da jede methodische Entscheidung in der Datenaufbereitung die späteren Modellergebnisse beeinflusst.

3.1.1 Datenquelle und Struktur

Der empirischen Analyse liegt der Datensatz *Polish Companies Bankruptcy* aus dem **UCI Machine Learning Repository** zugrunde Zięba, S. K. Tomczak, and J. M. Tomczak, 2016. Die Daten stammen aus der Datenbank Emerging Markets Information Service

(EMIS) und wurden von Zięba, Tomczak und Tomczak (2016) für Ensemble-Klassifikatoren zusammengestellt. Der Datensatz ist frei verfügbar über das UCI Repository[1] sowie Kaggle und umfasst Finanzkennzahlen polnischer Unternehmen aus dem Zeitraum 2000 bis 2013.

3.1.1.1 Umfang und Grundstruktur

Die Datenbasis besteht aus **43.405 Beobachtungen**, die jeweils ein Unternehmen zu einem bestimmten Zeitpunkt und für einen spezifischen Prognosehorizont repräsentieren. Für jede Beobachtung sind **64 Finanzkennzahlen** (bezeichnet als A1 bis A64) sowie eine binäre Zielvariable verfügbar, die angibt, ob das Unternehmen innerhalb des jeweiligen Prognosehorizonts insolvent wurde (1) oder nicht (0).

Ein Alleinstellungsmerkmal dieses Datensatzes ist die Berücksichtigung **fünf unterschiedlicher Prognosehorizonte** (H1 bis H5), die Vorhersagezeiträume von einem bis fünf Jahren abbilden. Tabelle 3.1 zeigt die Verteilung der Beobachtungen über die Horizonte.

Table 3.1: Verteilung der Beobachtungen nach Prognosehorizont

Horizont	**Beschreibung**	**N**	**Anteil (%)**
H1	1 Jahr	7.027	16,2
H2	2 Jahre	10.173	23,4
H3	3 Jahre	10.503	24,2
H4	4 Jahre	9.792	22,6
H5	5 Jahre	5.910	13,6
Gesamt		**43.405**	**100,0**

Quelle: Eigene Darstellung basierend auf Script 00a_polish_dataset_overview.py

Die ungleiche Verteilung der Beobachtungen über die Horizonte – mit einem deutlichen Schwerpunkt auf H2 und H3 – reflektiert die Datenverfügbarkeit im Ursprungsdatensatz. Horizont H5 weist mit 5.910 Beobachtungen die geringste Fallzahl auf, was bei der späteren Modellierung zu beachten ist.

3.1.1.2 Datenstruktur: Wiederholte Querschnitte

Eine kritische Eigenschaft des Datensatzes ergibt sich aus dem Fehlen eines Unternehmensidentifikators. Die Daten enthalten keine Variable, die es ermöglichen würde, ein bestimmtes Unternehmen über verschiedene Zeitpunkte oder Horizonte hinweg zu verfolgen. Dies hat weitreichende methodische Konsequenzen:

[1] `https://archive.ics.uci.edu/dataset/365/`

Implikation: Bei den vorliegenden Daten handelt es sich nicht um Paneldaten, sondern um **wiederholte Querschnitte** (repeated cross-sections). Jede Beobachtung ist als unabhängig zu betrachten. Folglich sind Methoden, die auf einer zeitlichen Verfolgung derselben Einheiten basieren (z. B. Fixed-Effects- oder Random-Effects-Modelle), nicht anwendbar Wooldridge, 2010.

Besonderheit der Horizont-Struktur Ein methodisch wichtiges Merkmal des Datensatzes ergibt sich aus der spezifischen Konstruktion der Prognosehorizonte. Die Bezeichnungen H1 bis H5 suggerieren zunächst, dass sie unterschiedlich weit in die Zukunft blicken. Tatsächlich jedoch repräsentieren die fünf Horizonte **unterschiedliche Beobachtungszeitpunkte desselben Prognosezeitraums**.

Zeitliche Struktur der Horizonte:

- **H1:** Finanzdaten aus Jahr 1 → Insolvenzstatus in Jahr 6 (5 Jahre Vorlaufzeit)
- **H2:** Finanzdaten aus Jahr 2 → Insolvenzstatus in Jahr 6 (4 Jahre Vorlaufzeit)
- **H3:** Finanzdaten aus Jahr 3 → Insolvenzstatus in Jahr 6 (3 Jahre Vorlaufzeit)
- **H4:** Finanzdaten aus Jahr 4 → Insolvenzstatus in Jahr 6 (2 Jahre Vorlaufzeit)
- **H5:** Finanzdaten aus Jahr 5 → Insolvenzstatus in Jahr 6 (1 Jahr Vorlaufzeit)

Alle Horizonte prognostizieren damit die Insolvenz zum **selben Zieljahr**, jedoch basierend auf Finanzdaten aus unterschiedlich weit zurückliegenden Jahren. Ein konkretes Beispiel verdeutlicht diese Struktur: Angenommen, ein Unternehmen geht im Jahr 2010 bankrott. Dieses Unternehmen könnte in allen fünf Horizonten erscheinen – in H1 mit Finanzdaten von 2005, in H2 mit Daten von 2006, bis hin zu H5 mit Daten von 2009. Die fehlende Unternehmens-ID verhindert jedoch die direkte Identifikation solcher Überlappungen.

Implikationen dieser Struktur:

1. **Pseudo-Replikation:** Es ist wahrscheinlich, dass identische Unternehmen in mehreren Horizonten auftreten. Dies könnte zur sogenannten Pseudo-Replikation führen, bei der statistisch unabhängige Beobachtungen angenommen werden, obwohl tatsächlich Abhängigkeiten bestehen. Die horizontspezifische Modellierung (siehe Abschnitt 3.4) minimiert dieses Problem, da jedes Modell ausschließlich einen Horizont verwendet.
2. **Unterschiedliche Prädiktionsmuster:** In frühen Jahren (H1, H2) können finanzielle Schwierigkeiten noch latent sein, während sie in späten Jahren (H4, H5) bereits manifest werden. Dies erklärt die unterschiedlichen Insolvenzraten über Horizonte hinweg (siehe Abschnitt 3.1.3) und rechtfertigt die Entwicklung separater Modelle je Horizont.

3. **Validierungsstrategie:** Die zeitliche Struktur determiniert die Wahl geeigneter Validierungsstrategien. Ein zeitbasierter Holdout auf Unternehmensebene ist nicht möglich, daher wird auf eine horizontbasierte Aufteilung zurückgegriffen (siehe Abschnitt 3.2.3).

Diese Struktureigenschaft des Datensatzes ist für die Interpretation der späteren Modellergebnisse von zentraler Bedeutung: Die prognostische Aufgabe unterscheidet sich fundamental zwischen den Horizonten – H1 muss sehr frühe Warnsignale identifizieren, während H5 bereits manifeste Krisensymptome erkennt.

3.1.1.3 Zielvariable und Klassenverteilung

Die Zielvariable `y` nimmt den Wert 1 an, wenn ein Unternehmen innerhalb des jeweiligen Prognosehorizonts insolvent wurde, andernfalls 0. Von den 43.405 Beobachtungen sind **2.091 als insolvent** klassifiziert, was einer Gesamtinsolvenzrate von **4,82 %** entspricht.

Diese ausgeprägte Klassenimbalance – typisch für Insolvenzdaten – stellt eine methodische Herausforderung dar: Naive Klassifikatoren könnten durch simples Vorhersagen der Mehrheitsklasse („nicht insolvent") bereits eine Genauigkeit von 95,18 % erreichen, ohne tatsächlich prädiktive Muster zu erlernen. Strategien zur Handhabung dieser Imbalance werden in Abschnitt 3.4 erörtert.

3.1.1.4 Zeitliche Abdeckung

Die Daten stammen aus dem Zeitraum 2000 bis 2013 und umfassen damit sowohl Phasen wirtschaftlicher Stabilität als auch die globale Finanzkrise von 2007/08. Diese zeitliche Heterogenität ist methodisch vorteilhaft, da Modelle so auf Daten aus unterschiedlichen Konjunkturzyklen trainiert werden können. Allerdings ist zu berücksichtigen, dass die Ergebnisse möglicherweise nicht ohne Weiteres auf aktuelle Verhältnisse übertragbar sind, da sich Rechnungslegungsstandards, regulatorische Rahmenbedingungen und Wirtschaftsstrukturen seither verändert haben könnten.

3.1.1.5 Zusammenfassung der Datengrundlage

Der Polish Companies Bankruptcy-Datensatz bietet mit 43.405 Beobachtungen, 64 Finanzkennzahlen und fünf Prognosehorizonten eine substantielle Basis für die Entwicklung von Insolvenzprognosemodellen. Die Datenstruktur als wiederholte Querschnitte determiniert die methodischen Optionen, während die ausgeprägte Klassenimbalance spezielle Behandlungsstrategien erfordert. Die folgenden Abschnitte analysieren die inhaltliche Bedeutung der Kennzahlen (3.1.2), die zeitliche Struktur (3.1.3) sowie die Datenqualität (3.1.4).

3.1.2 Finanzkennzahlen und Kategorisierung

Die 64 Finanzkennzahlen (A1 bis A64) bilden das Herzstück der Datenbasis. Im Gegensatz zu generischen Variablen handelt es sich hierbei um ökonomisch interpretierbare Ratios, die zentrale Dimensionen der Unternehmensperformance abbilden. Eine systematische Kategorisierung dieser Kennzahlen ist essentiell, um (1) ihre ökonomische Bedeutung zu verstehen, (2) erwartbare Korrelationsmuster zu antizipieren und (3) die spätere Interpretation von Modellergebnissen zu ermöglichen.

3.1.2.1 Kategorisierung nach Kennzahlendimensionen

Basierend auf der im Datensatz bereitgestellten Metadatei wurden die 64 Kennzahlen sechs funktionalen Kategorien zugeordnet. Tabelle 3.2 zeigt die Verteilung.

Table 3.2: Kategorisierung der 64 Finanzkennzahlen

Kategorie	**Anzahl**	**Anteil (%)**	**Beispiele**
Profitabilität	20	31,2	ROA, ROE, EBITDA-Marge, Nettogewinnmarge
Verschuldung	17	26,6	Debt-to-Equity, Leverage Ratio, Asset Coverage
Aktivität	15	23,4	Asset Turnover, Inventory Days, Receivables Days
Liquidität	10	15,6	Current Ratio, Quick Ratio, Cash Ratio
Größe	1	1,6	Logarithmus der Bilanzsumme
Sonstige	1	1,6	Spezialkennzahl
Gesamt	**64**	**100,0**	

Quelle: Eigene Darstellung basierend auf Script 00b_polish_feature_analysis.py

Profitabilitätskennzahlen bilden mit 20 Features (31,2 %) die größte Kategorie. Dies reflektiert die zentrale Rolle der Ertragskraft in der Insolvenzforschung: Bereits Altman (1968) identifizierte Profitabilitätsratios als wichtigste Prädiktoren im klassischen Z-Score-Modell Altman, 1968. Bemerkenswert ist zudem der hohe Anteil an Verschuldungskennzahlen (17 Features, 26,6 %), was die Bedeutung der Kapitalstruktur für Insolvenzprognosen unterstreicht. Die hohe Anzahl ähnlicher Kennzahlen könnte zu Redundanzen führen, da viele Ratios ähnliche Aspekte der Unternehmensperformance messen.

3.1.2.2 Mathematische Struktur und Redundanzen

Eine detaillierte Analyse der Kennzahlenformeln – dokumentiert in der Metadatei – offenbart strukturelle Zusammenhänge, die zu erwartender Multikollinearität führen:

Inverse Kennzahlenpaare Ein besonders klarer Fall ist das Paar A17 und A2:

- A17: $\frac{\text{Aktiva}}{\text{Passiva}}$
- A2: $\frac{\text{Passiva}}{\text{Aktiva}}$

Diese beiden Kennzahlen sind mathematisch reziprok zueinander. Folglich weist ihre Korrelation zwangsläufig eine perfekte inverse Struktur auf ($r \approx -1$), was zu numerischer Instabilität in Regressionsmodellen führen kann.

Gemeinsame Nenner Ein weiteres Redundanzmuster ergibt sich aus der wiederholten Verwendung derselben Größe im Nenner. Die Analyse zeigt:

- 22 Kennzahlen verwenden **Umsatz** (Sales) im Nenner
- 18 Kennzahlen verwenden **Bilanzsumme** (Total Assets) im Nenner
- 12 Kennzahlen verwenden **Eigenkapital** (Equity) im Nenner

Kennzahlen mit identischem Nenner tendieren zu positiver Korrelation, da Schwankungen im Nenner alle betroffenen Ratios in dieselbe Richtung bewegen. Beispielsweise werden sämtliche umsatzbasierten Kennzahlen bei Umsatzrückgang mechanisch ansteigen, unabhängig vom Zähler.

Hierarchische Abhängigkeiten Einige Kennzahlen stehen in direkter rechnerischer Beziehung zueinander. Ein Beispiel:

$$\text{Operating Cycle} = \text{Inventory Days} + \text{Receivables Days}$$

Derartige additive Zusammenhänge erzeugen Multikollinearität, selbst wenn die einzelnen Komponenten nicht perfekt korreliert sind.

3.1.2.3 Implikationen für die Modellierung

Die identifizierten strukturellen Redundanzen haben weitreichende Konsequenzen:

1. **Multikollinearität:** Hohe Korrelationen zwischen Prädiktoren führen zu instabilen Regressionskoeffizienten und aufgeblähten Standardfehlern in linearen Modellen Hastie, Tibshirani, and Friedman, 2009.
2. **VIF-Analyse erforderlich:** In Phase 03 (Multikollinearitätsanalyse) wird der Variance Inflation Factor (VIF) für alle Kennzahlen berechnet. Kennzahlen mit VIF > 10 sind Kandidaten für Entfernung.
3. **Baum-basierte Modelle weniger betroffen:** Random Forests und XGBoost sind gegenüber Multikollinearität robuster als logistische Regression, da sie Variablen sequenziell betrachten.

3.1.2.4 Ökonomische Interpretierbarkeit

Trotz mathematischer Redundanzen besitzt jede Kennzahl eine eigenständige ökonomische Bedeutung. Beispiele:

- **A1 (Nettogewinn / Bilanzsumme):** Return on Assets – zentrale Profitabilitätskennzahl
- **A4 (Umlaufvermögen / kurzfr. Verbindlichkeiten):** Current Ratio – klassische Liquiditätskennzahl
- **A37 (Quick Assets / langfr. Verbindlichkeiten):** Misst Fähigkeit, langfristige Schulden aus liquiden Mitteln zu bedienen

Diese Interpretierbarkeit ist ein Vorteil gegenüber generischen Features und ermöglicht die Validierung von Modellergebnissen anhand betriebswirtschaftlicher Plausibilität.

3.1.2.5 Zusammenfassung der Kennzahlenstruktur

Die 64 Finanzkennzahlen decken zentrale Dimensionen der Unternehmensperformance ab, weisen jedoch strukturbedingte Redundanzen auf. Ein inverses Kennzahlenpaar, neun Gruppen mit gemeinsamem Nenner und hierarchische Abhängigkeiten lassen hohe Multikollinearität erwarten. Diese muss in der Feature Engineering-Phase (Abschnitt 3.3) adressiert werden, um stabile und interpretierbare Modelle zu gewährleisten. Gleichzeitig bietet die ökonomische Interpretierbarkeit der Kennzahlen einen wertvollen Vorteil für die spätere Modellvalidierung.

3.1.3 Zeitliche Struktur und Insolvenztrend

Die Berücksichtigung multipler Prognosehorizonte (H1 bis H5) ist ein methodisches Alleinstellungsmerkmal dieses Datensatzes. Während die meisten Studien zur Insolvenzprognose sich auf einen fixen Zeithorizont beschränken – typischerweise ein Jahr Altman, 1968 – ermöglicht dieser Datensatz die Untersuchung, ob und wie sich die Vorhersagbarkeit mit zunehmendem Zeitabstand verändert. Die folgende Analyse offenbart einen Befund, der zentrale Auswirkungen auf die Modellierungsstrategie hat.

3.1.3.1 Insolvenzrate nach Prognosehorizont

Tabelle 3.3 zeigt die Verteilung insolventer und nicht-insolventer Unternehmen über die fünf Horizonte.

Table 3.3: Insolvenzrate nach Prognosehorizont

Horizont	**N**	**Insolvenzen**	**Rate (%)**	**Veränderung**
H1 (1 Jahr)	7.027	271	3,86	Baseline
H2 (2 Jahre)	10.173	400	3,93	+1,8 %
H3 (3 Jahre)	10.503	495	4,71	+22,0 %
H4 (4 Jahre)	9.792	515	5,26	+36,3 %
H5 (5 Jahre)	5.910	410	6,94	**+79,8 %**

Quelle: Eigene Darstellung basierend auf Script 00c_polish_temporal_structure.py

Der augenfälligste Befund ist der **nahezu lineare Anstieg der Insolvenzrate** mit zunehmendem Prognosehorizont. Von H1 (3,86 %) zu H5 (6,94 %) ergibt sich eine Steigerung um fast 80 %. Dieser Trend ist in Abbildung 3.1 visualisiert.

[Hier: Liniendiagramm aus 00c_temporal_analysis.png einfügen]
X-Achse: Horizont (H1–H5)
Y-Achse: Insolvenzrate (%)

Figure 3.1: Entwicklung der Insolvenzrate über Prognosehorizonte

Quelle: Eigene Darstellung

3.1.3.2 Ökonomische Interpretation

Der beobachtete Trend ist ökonomisch plausibel: Mit zunehmendem Zeitabstand steigt die Wahrscheinlichkeit, dass ein Unternehmen in finanzielle Schwierigkeiten gerät. Während ein Unternehmen mit robusten Fundamentaldaten die nächsten 12 Monate wahrscheinlich

übersteht, erhöht sich über einen Fünf-Jahres-Zeitraum das Risiko externer Schocks (Konjunktureinbrüche, regulatorische Änderungen, disruptive Wettbewerber) oder interner Probleme (Managementfehler, verfehlte Investitionen).

Aus methodischer Sicht ist jedoch die **Größenordnung** der Veränderung entscheidend: Eine Verdoppelung der Insolvenzrate deutet darauf hin, dass H1- und H5-Daten aus unterschiedlichen Verteilungen stammen könnten.

3.1.3.3 Heterogenität der Prognosehorizonte

Die Standardabweichung der Insolvenzraten über die Horizonte beträgt 1,2 Prozentpunkte – bei einem Mittelwert von 4,82 % entspricht dies einem Variationskoeffizienten von 25 %. Diese Heterogenität ist nicht trivial und wirft die Frage auf, ob die fünf Horizonte als homogene Stichprobe behandelt werden sollten.

Literatureinbettung: Coats & Fant (1993) zeigen in ihrer Studie, dass die Beziehung zwischen Finanzkennzahlen und Insolvenzwahrscheinlichkeit über längere Zeithorizonte zunehmend nichtlinear wird Coats and Fant, 1993. McLeay & Omar (2000) bestätigen diesen Befund und warnen vor der Annahme zeitlicher Homogenität bei Multi-Horizont-Daten McLeay and Omar, 2000. Beide Studien argumentieren, dass unterschiedliche Zeithorizonte unterschiedliche prädiktive Dynamiken aufweisen können.

3.1.3.4 Implikationen für die Modellierungsstrategie

Die identifizierte Heterogenität hat weitreichende Konsequenzen für die methodische Vorgehensweise. Es stellen sich zwei zentrale Fragen:

Frage 1: Pooled Model oder horizontspezifische Modelle? Zwei Strategien sind denkbar:

- **Option A – Horizontspezifische Modelle:** Für jeden Horizont wird ein separates Modell trainiert (fünf Modelle insgesamt). Dies erlaubt horizontspezifische Koeffizienten und Feature Importance.
- **Option B – Pooled Model:** Ein gemeinsames Modell für alle Horizonte, wobei der Horizont als zusätzliches Feature einbezogen wird. Dies nutzt alle Daten, setzt jedoch voraus, dass Features alle Horizonte ähnlich beeinflussen.

Angesichts der 80-prozentigen Veränderung der Insolvenzrate erscheint Option A methodisch stringenter. Die Annahme, dass dieselben Kennzahlen mit denselben Gewichtungen sowohl 1-Jahres- als auch 5-Jahres-Insolvenzen vorhersagen, ist fraglich. Daher wurde für diese Arbeit **Option A gewählt**: In der Modellierungsphase (Abschnitt 3.4) werden fünf separate Modelle entwickelt, eines je Horizont.

Frage 2: Train/Val/Test-Split Bei horizontspezifischen Modellen muss der Datensatz für jeden Horizont getrennt aufgeteilt werden. Ein zeitlicher Holdout (Train: H1–H3, Val: H4, Test: H5) ist nicht zielführend, da dies unterschiedliche Verteilungen vermischt. Stattdessen wird jeder Horizont einzeln in Train (60 %), Validation (20 %) und Test (20 %) aufgeteilt – unter Beibehaltung der Klassenverteilung durch stratifiziertes Sampling.

3.1.3.5 Stabilität der Kennzahlen über Horizonte

Eine weiterführende Analyse (nicht detailliert dargestellt) untersuchte, ob die Finanzkennzahlen selbst über Horizonte hinweg stabil verteilt sind. Die Variationskoeffizienten der Mittelwerte einzelner Kennzahlen über H1 bis H5 liegen überwiegend unter 10 %, was auf relative Stabilität hindeutet. Die beobachtete Heterogenität ist also primär auf die *Zielvariable* (Insolvenzrate), nicht auf die Prädiktoren zurückzuführen.

3.1.3.6 Zusammenfassung der zeitlichen Struktur

Die Analyse der zeitlichen Struktur offenbart einen fundamentalen Befund: Die Insolvenzrate steigt mit zunehmendem Prognosehorizont um 80 % von 3,86 % (H1) auf 6,94 % (H5). Diese ausgeprägte Heterogenität – konsistent mit empirischen Erkenntnissen von Coats & Fant (1993) – determiniert die Modellierungsstrategie: Anstelle eines gepoolten Modells werden fünf horizontspezifische Modelle entwickelt, um der unterschiedlichen Prognosecharakteristik gerecht zu werden. Dieser Ansatz respektiert die Datenstruktur und maximiert die prädiktive Validität für jeden einzelnen Zeithorizont.

3.1.4 Datenqualität und identifizierte Probleme

Die Qualität der Rohdaten determiniert maßgeblich die Validität jeglicher darauf basierenden Analysen. Eine rigorose Qualitätsprüfung ist daher kein optionaler Zusatzschritt, sondern methodische Notwendigkeit Goodfellow, Bengio, and Courville, 2016. Die folgenden Abschnitte dokumentieren systematisch alle identifizierten Datenqualitätsprobleme – einschließlich getroffener Annahmen und deren Limitationen. Diese Transparenz ist essentiell für die kritische Einordnung der Ergebnisse.

3.1.4.1 Fehlende Werte: Umfang und Muster

Eine erste Überraschung der Datenanalyse: **Sämtliche 64 Finanzkennzahlen weisen fehlende Werte auf**. Die Ausprägung variiert jedoch erheblich. Tabelle 3.4 zeigt die fünf am stärksten betroffenen Kennzahlen.

Table 3.4: Kennzahlen mit den höchsten Anteilen fehlender Werte

Kennzahl	Bezeichnung	Fehlend (N)	Anteil (%)
A37	Quick Assets / LT Liabilities	18.984	43,74
A21	Umsatzwachstum	5.854	13,49
A27	Op. Profit / Fin. Expenses	2.764	6,37
A60	Inventory Turnover	2.152	4,96
A45	Net Profit / Inventory	2.147	4,95

Quelle: Eigene Darstellung basierend auf Script 00d_polish_data_quality.py

Kennzahl A37 (Quick Assets / langfristige Verbindlichkeiten) sticht mit 43,74 % fehlenden Werten hervor. Dieser hohe Anteil wirft die Frage auf, ob die Kennzahl überhaupt für Modellierung verwendbar ist. Zwei Ansätze wurden in Betracht gezogen:

1. **Entfernung der Kennzahl:** Sicher, aber mit Informationsverlust verbunden.
2. **Fortgeschrittene Imputation:** Erhalt der Kennzahl, jedoch mit Unsicherheit behaftet.

Die Entscheidung für Option 2 (Imputation) wird in Abschnitt 3.2.1 detailliert begründet. Eine ergänzende horizontspezifische Analyse (siehe Abschnitt 3.1.4.5) zeigt, dass die Missing-Rate von A37 zwischen H1 (41,0 %) und H5 (46,8 %) variiert – eine Schwankung von 5,8 Prozentpunkten, die im Kontext der Gesamtdatenqualität als moderat einzustufen ist und keine separaten Imputationsansätze je Horizont erfordert.

3.1.4.2 Duplikate: Natur und Umgang

Die Qualitätsprüfung identifizierte **401 exakte Duplikate** – definiert als Beobachtungen, bei denen *alle* 68 Variablen (64 Kennzahlen, Jahr, Horizont, Zielvariable, Insolvenzindikator) identisch sind. Dies entspricht 200 Paaren, bei denen jede Zeile exakt einmal dupliziert ist.

Problem der Verifikation Das Fehlen eines Unternehmensidentifikators verhindert eine abschließende Klärung der Duplikat-Natur. Zwei Szenarien sind denkbar:

- **Szenario A:** Identisches Unternehmen wurde versehentlich zweimal erfasst → Datenerfassungsfehler.

- **Szenario B:** Zwei unterschiedliche Unternehmen weisen zufällig identische Werte auf.

Szenario B erscheint statistisch äußerst unwahrscheinlich: Die Wahrscheinlichkeit, dass zwei Unternehmen in allen 64 (kontinuierlichen) Kennzahlen identische Werte aufweisen, ist infinitesimal gering. Eine Binomialrechnung unter Annahme vernünftiger Präzision (2 Dezimalstellen) liefert eine Wahrscheinlichkeit in der Größenordnung von 10^{-128} für ein einzelnes Paar.

Getroffene Annahme Angesichts dieser Überlegung wird **Szenario A** (Datenerfassungsfehler) als plausibler erachtet. Folglich wurden alle 401 Duplikate entfernt, wobei jeweils die erste Instanz beibehalten wurde.

Timing der Entfernung Kritisch ist, dass die Duplikat-Entfernung *vor* dem Train/Val/Test-Split erfolgt. Würden Duplikate über Train- und Test-Set verteilt, entstünde Data Leakage: Das Modell könnte auf einer Trainingsbeobachtung lernen und dieselbe Beobachtung im Test „vorhersagen" – ein methodischer Kardinalfehler Goodfellow, Bengio, and Courville, 2016.

Limitationen Diese Entscheidung bleibt eine **Annahme**. Ohne Unternehmensidentifikator ist keine definitive Verifikation möglich. Eine konservative Alternative wäre gewesen, beide Instanzen zu entfernen (Verlust von 802 Beobachtungen statt 401). Die gewählte Variante balanciert zwischen Vorsicht und Datenerhalt.

3.1.4.3 Ausreißer: Systematische Identifikation

Finanzielle Kennzahlen sind notorisch anfällig für Extremwerte – etwa durch Bilanzmanipulation, Sondereereignisse oder Messfehler. Eine systematische Ausreißeranalyse mittels der 3×IQR-Methode (Interquartilsabstand) ergab:

- **Alle 64 Kennzahlen** weisen Ausreißer auf.
- Der Anteil betroffener Beobachtungen variiert zwischen 0,07 % (A50) und 15,5 % (A27).
- Im Mittel sind 5,4 % der Werte je Kennzahl als Ausreißer klassifiziert (Median: 4,5 %).
- Nur 7 Kennzahlen (11 %) weisen Ausreißeranteile über 10 % auf.

Tabelle 3.5 zeigt die fünf am stärksten betroffenen Kennzahlen.

Table 3.5: Kennzahlen mit den höchsten Ausreißeranteilen

Kennzahl	Bezeichnung	Ausreißer (N)	Anteil (%)
A27	Op. Profit / Fin. Expenses	6.306	15,52
A6	Retained Earnings / Assets	6.525	15,04
A37	Quick Assets / LT Liabilities	2.919	11,95
A55	Working Capital	4.769	10,99
A45	Net Profit / Inventory	4.399	10,66

Quelle: Eigene Darstellung basierend auf Script 00d_polish_data_quality.py, 3×IQR-Methode

Behandlungsstrategie Extreme Werte wurden nicht entfernt, sondern mittels **Winsorisierung** behandelt: Werte unterhalb des 1. Perzentils werden auf das 1. Perzentil gesetzt, Werte oberhalb des 99. Perzentils auf das 99. Perzentil. Diese Methode dämpft Extremwerte ohne Informationsverlust durch Beobachtungsentfernung und ist in der empirischen Finanzforschung etabliert.

Timing Winsorisierung erfolgt *nach* Duplikat-Entfernung, aber *vor* Imputation. Dies ist methodisch wichtig: Würden Ausreißer erst nach Imputation behandelt, könnten sie die Imputationsstatistiken (Mediane, Mittelwerte) verzerren.

3.1.4.4 Varianz: Konstante und quasi-konstante Features

Ein weiteres potenzielles Problem sind Features mit geringer oder fehlender Varianz, da diese keine Information für Modellierung beisteuern. Die Analyse ergab jedoch:

- **Keine** Kennzahl weist Zero-Varianz auf.
- **Keine** Kennzahl weist extrem niedrige Varianz ($< 0{,}01$) auf.
- Alle 64 Kennzahlen haben substantielle Streuung.

Dieser Befund ist positiv: Alle Kennzahlen tragen potenziell Information bei und müssen nicht aus Varianzgründen entfernt werden.

3.1.4.5 Horizontspezifische Datenqualitätsanalyse

Eine zentrale methodische Frage ergibt sich aus der horizontspezifischen Modellierungsstrategie (siehe Abschnitt 3.4): Unterscheidet sich die Datenqualität systematisch über die Prognosehorizonte H1 bis H5? Falls ja, könnte dies separate Preprocessing-Pipelines je Horizont erfordern. Eine ergänzende Analyse untersuchte daher Ausreißerraten und fehlende Werte differenziert nach Horizonten.

Ausreißerraten über Horizonte Tabelle 3.6 zeigt die durchschnittlichen Ausreißerraten (3×IQR-Methode) für jeden Horizont, gemittelt über alle 64 Kennzahlen.

Table 3.6: Ausreißerraten nach Prognosehorizont

Horizont	Beobachtungen	Mittlere Ausreißerrate (%)	Maximum (%)
H1	7.027	4,14	12,84
H2	10.173	5,34	17,19
H3	10.503	5,76	19,85
H4	9.792	6,06	22,42
H5	5.910	5,15	15,64
Variation		**1,92 Prozentpunkte (H1 vs. H4)**	

Quelle: Eigene Darstellung basierend auf Script 00d_polish_data_quality.py, Schritt 5b

Befund: Die Ausreißerraten steigen tendenziell von H1 (4,14 %) zu H4 (6,06 %), fallen jedoch in H5 wieder leicht ab. Die Gesamtvariation beträgt 1,92 Prozentpunkte – ein verhältnismäßig geringer Unterschied. Dieser leichte Anstieg dürfte natürliche Datencharakteristika widerspiegeln (Unternehmen in späten Horizonten weisen mehr finanzielle Volatilität auf), nicht jedoch systematische Qualitätsdefizite.

Fehlende Werte über Horizonte (Beispiel A37) Als Illustration wurde die am stärksten betroffene Kennzahl A37 (Quick Assets / LT Liabilities) analysiert:

- H1: 41,0 % fehlend (2.883 von 7.027)
- H2: 43,9 % fehlend (4.466 von 10.173)
- H3: 44,2 % fehlend (4.642 von 10.503)
- H4: 44,3 % fehlend (4.338 von 9.792)
- H5: 46,8 % fehlend (2.766 von 5.910)

Die Variation beträgt 5,8 Prozentpunkte (H1 vs. H5). Während der Anstieg nicht vernachlässigbar ist, rechtfertigt er keine grundlegend unterschiedlichen Imputationsstrategien je Horizont.

Implikation für Preprocessing Die horizontspezifische Analyse zeigt **relative Stabilität der Datenqualität** über alle fünf Horizonte. Die Variation in Ausreißerraten (1,92 Prozentpunkte) und Missing-Raten (5,8 Prozentpunkte bei A37) ist moderat. Dies rechtfertigt die Anwendung einer **einheitlichen Preprocessing-Pipeline** (Duplikatentfernung, Winsorisierung, Imputation) über alle Horizonte, gefolgt von horizontspezifischer Modellierung.

Die beobachteten Unterschiede spiegeln natürliche Datencharakteristika wider und stellen keine methodische Herausforderung dar.

3.1.4.6 Zusammenfassung der Datenqualitätsprobleme

Tabelle 3.7 fasst die identifizierten Probleme und getroffenen Maßnahmen zusammen.

Table 3.7: Übersicht Datenqualitätsprobleme und Behandlung

Problem	**Ausprägung**	**Maßnahme**
Fehlende Werte	64/64 Kennzahlen betroffen; max. 43,7 % (A37)	Passive Imputation für Ratios (Abschnitt 3.2.1)
Duplikate	401 exakte Duplikate	Entfernung vor Train/Test-Split
Ausreißer	64/64 Kennzahlen betroffen; 0,07 %–15,5 % je Feature (Mittel: 5,4 %)	Winsorisierung (1./99. Perzentil)
Varianz	Keine Zero- oder Low-Varianz-Features	Keine Aktion erforderlich

Quelle: Eigene Darstellung basierend auf Script 00d_polish_data_quality.py

3.1.4.7 Methodische Reflexion

Die Datenqualitätsanalyse offenbart ein realistisches Bild: Der Datensatz ist nicht „sauber", sondern weist typische Probleme empirischer Finanzdaten auf. Die Herausforderung besteht darin, diese Probleme methodisch stringent zu behandeln, ohne dabei in zwei Extreme zu verfallen:

1. **Naives Ignorieren:** Probleme werden übersehen oder als „unproblematisch" deklariert – führt zu verzerrten Ergebnissen.
2. **Überkonservatives Löschen:** Alle problematischen Beobachtungen/Features werden entfernt – führt zu massivem Informationsverlust.

Der gewählte Mittelweg – Imputation statt Löschung, Winsorisierung statt Entfernung, transparente Dokumentation von Annahmen – entspricht dem State of the Art in der empirischen Forschung und gewährleistet sowohl Datenerhalt als auch methodische Integrität.

Die detaillierte Umsetzung dieser Maßnahmen wird in Abschnitt 3.2 (Datenaufbereitung) beschrieben.

4 Datenaufbereitung: Data Preparation

Die in Kapitel 3 dokumentierte Analyse der Datenbasis offenbarte drei zentrale Qualitätsprobleme: **401 exakte Duplikate**, **systematische Ausreißer in allen 64 Kennzahlen** sowie **41.037 fehlende Werte** über sämtliche Features. Während die Foundation-Phase diese Probleme identifiziert und quantifiziert hat, widmet sich die Data Preparation-Phase ihrer methodisch fundierten Behandlung.

Das vorliegende Kapitel dokumentiert die praktische Umsetzung evidenzbasierter Preprocessing-Strategien. Die Reihenfolge der Schritte ist dabei nicht arbiträr, sondern folgt methodischen Notwendigkeiten: Duplikate werden *zuerst* entfernt (Abschnitt 4.1), um Data Leakage zu vermeiden. Anschließend erfolgt die Behandlung von Ausreißern mittels Winsorisierung (Abschnitt 4.2), bevor fehlende Werte imputiert werden (Abschnitt 4.3). Diese Sequenz gewährleistet, dass Extremwerte die Imputationsstatistiken nicht verzerren Goodfellow, Bengio, and Courville, 2016.

Das Ergebnis ist ein sauberer Datensatz mit 43.004 Beobachtungen und 0 % fehlenden Werten – bereit für die nachfolgende Modellierung. Alle implementierten Schritte sind über Python-Skripte vollständig reproduzierbar und dokumentiert.

4.1 Entfernung exakter Duplikate

Die in Abschnitt 3.1.4.2 identifizierten 401 exakten Duplikate stellen eine unmittelbare Bedrohung für die Validität jeglicher statistischer Inferenz dar. Würden diese Duplikate im Datensatz verbleiben, käme es zu einer künstlichen Übergewichtung bestimmter Beobachtungen – mit potenziell verzerrten Modellparametern als Folge. Die folgenden Abschnitte dokumentieren die methodische Behandlung dieses Problems.

4.1.1 Methodik der Duplikat-Identifikation

Als **exaktes Duplikat** wurde definiert: Eine Beobachtung, bei der *alle* 68 Variablen (64 Finanzkennzahlen, Jahr, Horizont, Zielvariable, Insolvenzindikator) mit mindestens einer anderen Beobachtung identisch sind. Diese strenge Definition minimiert das Risiko fälschlicher Entfernungen: Nur bei vollständiger Identität über alle Dimensionen wird ein Duplikat angenommen.

Die Identifikation erfolgte mittels der Pandas-Funktion `duplicated()`, die spaltenweise Übereinstimmung prüft. Es wurden **802 Zeilen als dupliziert** identifiziert, die sich zu **200 Paaren plus einer unpaarigen Zeile** gruppieren lassen. Dies deutet auf ein systematisches Datenerfassungsproblem hin: Jede Beobachtung wurde im Mittel exakt einmal versehentlich dupliziert.

4.1.2 Umgang mit Unsicherheit

Das Fehlen eines Unternehmensidentifikators verhindert die definitive Klärung, ob es sich tatsächlich um denselben betriebswirtschaftlichen Sachverhalt handelt. Die in Abschnitt 3.1.4.2 dargelegte statistische Argumentation – eine Wahrscheinlichkeit von $\approx 10^{-128}$ für zufällige Übereinstimmung aller 64 kontinuierlichen Werte – spricht jedoch klar für die Datenerfassungsfehler-Hypothese.

Konservative Alternative: Eine denkbare, noch vorsichtigere Strategie wäre gewesen, *beide* Instanzen eines Duplikatpaars zu entfernen (Verlust von 802 statt 401 Beobachtungen). Diese Variante hätte jedoch einen Informationsverlust von 1,85 % der Gesamtstichprobe bedeutet – ohne erkennbaren methodischen Mehrwert, da die erste Instanz vermutlich genauso valide ist wie die zweite.

4.1.3 Implementierung und Timing

Die Entfernung erfolgte mittels der Pandas-Operation `drop_duplicates(keep='first')`, die für jedes Duplikatpaar die erste Instanz beibehält und nachfolgende Instanzen entfernt. Die Wahl von `keep='first'` ist arbiträr (gleichwertig wäre `keep='last'` gewesen), jedoch reproduzierbar dokumentiert.

Kritischer methodischer Aspekt – Timing: Die Duplikat-Entfernung erfolgt *vor jeglicher weiterer Datenmanipulation*, insbesondere vor Datenaufteilungen. Dies ist essentiell, um **Data Leakage** zu verhindern: Würde ein Duplikatpaar über verschiedene Datenpartitionen verteilt, könnte dies zu optimistisch verzerrten Performanzmetriken führen Goodfellow, Bengio, and Courville, 2016.

4.1.4 Ergebnis und Auswirkungen

Tabelle 4.1 fasst das Ergebnis zusammen.

Die Insolvenzrate verändert sich minimal von 4,82 % auf 4,84 % – eine Verschiebung von lediglich 0,03 Prozentpunkten. Dies deutet darauf hin, dass die Duplikate *nicht* systematisch

Table 4.1: Ergebnis der Duplikat-Entfernung

Metrik	**Vorher**	**Nachher**
Beobachtungen	43.405	43.004
Entfernt	–	401
Anteil entfernt	–	0,92 %
Insolvenzrate	4,82 %	4,84 %
Veränderung	–	+0,03 pp

Quelle: Eigene Darstellung basierend auf Script 01a_remove_duplicates.py

häufiger insolvente oder gesunde Unternehmen betrafen, sondern gleichmäßig über beide Klassen verteilt waren. Die Klassenbalance bleibt somit im Wesentlichen erhalten.

4.1.5 Horizontspezifische Verteilung der Duplikate

Eine ergänzende Analyse untersuchte, ob bestimmte Prognosehorizonte überproportional von Duplikaten betroffen waren. Die Duplikatrate variiert zwischen 0,83 % (H3) und 1,17 % (H1) – eine Schwankung von lediglich 0,34 Prozentpunkten. Dies deutet auf ein **horizontübergreifend gleichmäßig verteiltes Datenerfassungsproblem** hin, nicht auf systematische Qualitätsdefizite einzelner Horizonte.

4.1.6 Methodische Reflexion

Die Entfernung von 401 Beobachtungen (0,92 % der Stichprobe) ist ein vertretbarer Preis für die Gewährleistung der Datenintegrität. Zwei alternative Strategien wären denkbar gewesen:

1. **Alle Duplikate behalten:** Führt zu künstlicher Übergewichtung und verzerrten Konfidenzintervallen.
2. **Beide Instanzen entfernen:** Höhere Datenverlustrate ohne erkennbaren methodischen Mehrwert.

Die gewählte Strategie (`keep='first'`) balanciert zwischen Vorsicht und Datenerhalt. Die Limitation – fehlende Verifizierbarkeit mangels Unternehmens-ID – bleibt bestehen, wird jedoch durch die statistische Evidenz (Abschnitt 3.1.4.2) hinreichend adressiert.

4.2 Winsorisierung extremer Werte

Finanzielle Kennzahlen sind notorisch anfällig für Extremwerte. Die Gründe reichen von Bilanzmanipulation über Messfehler bis hin zu Sonderereignissen (Restrukturierungen, Fusionen). Abschnitt 3.1.4.3 dokumentierte, dass *alle* 64 Kennzahlen Ausreißer aufweisen, mit Anteilen zwischen 0,07 % und 15,5 %. Die Behandlung dieser Extremwerte ist methodische Notwendigkeit, da sie in Regressionsmodellen zu instabilen Parametern führen können Hastie, Tibshirani, and Friedman, 2009.

4.2.1 Methodenwahl: Winsorisierung vs. Alternativen

Zur Behandlung von Ausreißern stehen grundsätzlich drei Strategien zur Verfügung:

Option A: Deletion Entfernung aller als Ausreißer klassifizierten Beobachtungen. Diese Methode ist einfach, führt jedoch zu erheblichem Informationsverlust: Bei durchschnittlich 5,4 % Ausreißern pro Feature würde eine zeilenbasierte Deletion potenziell einen Großteil der Stichprobe eliminieren, da viele Beobachtungen in mindestens einer Kennzahl einen Ausreißer aufweisen.

Option B: Log-Transformation Transformation rechtsschiefe Verteilungen mittels Logarithmus. Diese Methode scheitert jedoch bei Finanzkennzahlen systematisch: Viele Ratios können negative Werte annehmen (z. B. negativer Nettogewinn bei Verlusten), was $\log(x)$ für $x \leq 0$ undefiniert macht.

Option C: Winsorisierung Ersetzung von Extremwerten durch weniger extreme Perzentilwerte. Diese Methode dämpft Ausreißer, ohne Beobachtungen zu verlieren oder Vorzeichenprobleme zu verursachen. Sie ist in der empirischen Finanzforschung etabliert und wurde für diese Arbeit gewählt.

4.2.2 Implementierung: 1./99. Perzentil

Die Winsorisierung erfolgte separat für jede der 64 Kennzahlen nach folgendem Schema:

1. Berechnung des 1. Perzentils (P_1) und 99. Perzentils (P_{99}) über alle nicht-fehlenden Werte.
2. Ersetzung aller Werte $x < P_1$ durch P_1.
3. Ersetzung aller Werte $x > P_{99}$ durch P_{99}.

4. Fehlende Werte bleiben unverändert (werden später imputiert).

Die Wahl der **1./99. Perzentile** (statt z. B. 5./95.) beruht auf zwei Überlegungen: Erstens ist die Ausreißerrate mit durchschnittlich 5,4 % bereits moderat, sodass eine aggressivere Winsorisierung unnötig erscheint. Zweitens soll die ursprüngliche Verteilungsform soweit möglich erhalten bleiben – ein Prinzip, das gegen zu weite Perzentilgrenzen spricht.

4.2.3 Ergebnis und Auswirkungen

Tabelle 4.2 fasst die Auswirkungen zusammen.

Table 4.2: Ergebnis der Winsorisierung

Metrik	**Wert**
Beobachtungen (unverändert)	43.004
Features winsorisiert	64/64
Werte modifiziert (gesamt)	53.427
Durchschn. Anteil pro Feature	1,94 %
Min. Anteil	0,03 %
Max. Anteil	2,00 %

Quelle: Eigene Darstellung basierend auf Script 01b_outlier_treatment.py

Insgesamt wurden **53.427 Werte** über alle Features hinweg modifiziert – das entspricht durchschnittlich 1,94 % der Werte pro Feature. Dieser Anteil liegt deutlich unter der ursprünglichen Ausreißerrate von 5,4 % (3×IQR-Methode), was erwartbar ist: Die Perzentilmethode ist weniger konservativ als die IQR-basierte Definition und behandelt nur die extremsten 2 % der Verteilung.

4.2.4 Validierung: Erhalt der Stichprobengröße

Ein methodisch kritischer Aspekt ist die Überprüfung, dass durch Winsorisierung *keine* Beobachtungen verloren gehen. Die Verifikation bestätigt: Sowohl die Stichprobengröße (43.004) als auch die Anzahl fehlender Werte (41.037) bleiben unverändert. Dies bestätigt, dass Winsorisierung – im Gegensatz zu Deletion – ein informationserhaltender Prozess ist.

4.2.5 Timing im Preprocessing-Ablauf

Die Winsorisierung erfolgt *nach* Duplikat-Entfernung, aber *vor* Imputation fehlender Werte. Diese Reihenfolge ist methodisch begründet:

- **Nach Duplikaten:** Duplikate müssen zuerst entfernt werden, um nicht künstlich die Perzentilberechnungen zu verzerren.
- **Vor Imputation:** Würden Extremwerte erst *nach* Imputation behandelt, könnten sie die Imputationsstatistiken (Mittelwerte, Mediane) verzerren. Bei MICE-Imputation (siehe Abschnitt 4.3) werden Regressionsmodelle auf beobachteten Daten trainiert. Unbehandelte Extremwerte in den Prädiktoren würden zu instabilen Regressionskoeffizienten führen.

4.2.6 Methodische Reflexion und Limitationen

Die Wahl der Winsorisierung stellt einen Kompromiss dar: (1) Naives Belassen aller Werte führt zu instabilen Modellen; (2) aggressives Löschen führt zu Informationsverlust. Die gewählte Methode balanciert beide Aspekte, bringt jedoch eigene Limitationen mit sich:

1. **Informationsverzerrung:** Winsorisierung verändert die ursprüngliche Verteilung. Extremwerte könnten tatsächlich valide Beobachtungen repräsentieren (z. B. Unternehmen in außergewöhnlichen Situationen).
2. **Grenzwertbestimmung:** Die Wahl der 1./99. Perzentile ist nicht „objektiv richtig", sondern basiert auf etablierten Konventionen. Alternative Schwellenwerte (z. B. 5./95.) wären ebenfalls vertretbar gewesen.
3. **Feature-Unabhängigkeit:** Die Winsorisierung erfolgt separat je Feature, ohne Berücksichtigung multivariater Ausreißer (Beobachtungen, die in keiner einzelnen Dimension extrem sind, aber in Kombination ungewöhnlich).

Diese Limitationen sind transparent zu kommunizieren, schmälern jedoch nicht die grundsätzliche Notwendigkeit einer Ausreißerbehandlung bei Finanzkennzahlen.

4.3 Imputation fehlender Werte

Die gravierendste Datenqualitätsherausforderung – dokumentiert in Abschnitt 3.1.4.1 – besteht in den **41.037 fehlenden Werten** über alle 64 Kennzahlen. Ein Löschen aller betroffenen Zeilen (Listwise Deletion) würde die Stichprobe drastisch reduzieren und ist damit keine Option. Folglich ist Imputation methodische Notwendigkeit. Die folgenden Abschnitte dokumentieren die gewählte Strategie, deren theoretische Fundierung sowie die erzielten Ergebnisse.

4.3.1 Methodenwahl: MICE mit BayesianRidge

4.3.1.1 Grundprinzip: Multiple Imputation by Chained Equations (MICE)

MICE, auch bekannt als Fully Conditional Specification (FCS), ist ein iterativer Ansatz zur Imputation multivariater fehlender Daten Buuren and Groothuis-Oudshoorn, 2011. Das Grundprinzip:

1. **Initialisierung:** Alle fehlenden Werte werden initial durch einfache Statistiken (Median) ersetzt.
2. **Iteration:** Für jede Variable mit fehlenden Werten:
   - Trainiere ein Regressionsmodell, das diese Variable durch alle anderen Variablen vorhersagt.
   - Imputiere fehlende Werte mittels Vorhersage dieses Modells.
3. **Konvergenz:** Wiederhole Schritt 2, bis sich die imputierten Werte stabilisieren (typisch: 5–20 Iterationen).

Der entscheidende Vorteil von MICE gegenüber univariaten Methoden (Mean/Median Imputation): MICE **berücksichtigt Korrelationen** zwischen Variablen. Beispiel: Wenn eine Profitabilitätskennzahl fehlt, aber Liquiditäts- und Verschuldungskennzahlen bekannt sind, kann MICE deren Beziehung zur Profitabilität nutzen, um einen plausibeleren Imputationswert zu generieren.

4.3.1.2 Wahl des Schätzers: BayesianRidge Regression

Innerhalb von MICE muss spezifiziert werden, welches Modell zur Vorhersage verwendet wird. Für diese Arbeit wurde **BayesianRidge Regression** gewählt, aus folgenden Gründen:

1. **Multikollinearität-Robustheit:** Wie in Abschnitt 3.1.2 analysiert, weisen die 64 Kennzahlen strukturbedingte Redundanzen auf (inverse Paare, gemeinsame Nenner). Bayesian Ridge Regression integriert Regularisierung, die hohe Korrelationen zwischen Prädiktoren toleriert, ohne numerisch instabil zu werden Tipping, 2001.
2. **Automatische Unsicherheitsquantifizierung:** Der Bayessche Ansatz liefert nicht nur Punktschätzungen, sondern auch Unsicherheitsmaße, die während der Iteration helfen, Konvergenz zu überwachen.
3. **Vermeidung von Overfitting:** Regularisierung verhindert, dass das Imputationsmodell zu perfekt auf beobachtete Daten fittet – ein Risiko bei hochdimensionalen Daten (64 Prädiktoren).

Alternative Schätzer (z. B. Random Forest, KNN) wurden erwogen, jedoch verworfen: Random Forest ist rechenintensiver und liefert keine linearen Beziehungen (schwer interpretierbar bei Finanzkennzahlen); KNN ignoriert die interne Struktur der Daten.

4.3.2 Direkte Ratio-Imputation (JAV)

Eine zentrale Frage bei **abgeleiteten Variablen** (Ratios) ist, ob man die Ratio selbst imputiert oder die zugrundeliegenden Komponenten (Zähler/Nenner) und die Ratio erst im Anschluss berechnet. Das direkte Imputieren der bereits gebildeten Kennzahl wird in der Literatur als *JAV – "Just Another Variable"* bzw. *transform-then-impute* bezeichnet[1]. (vgl. auch Buuren, 2018, § 6.4.1)

Da unser Datensatz ausschließlich bereits berechnete Ratios (A1–A64) enthält und *keine* Rohkomponenten (z. B. Umsatz, Bilanzsumme etc.), ist **JAV die einzig praktikable und methodisch korrekte Vorgehensweise**. Wir imputieren daher die Ratios direkt mittels MICE Buuren and Groothuis-Oudshoorn, 2011 und nutzen die Korrelationsstruktur zwischen den Ratios, um plausible Werte zu erzeugen. Ansätze, die während der Imputation explizit die funktionale Beziehung zwischen Komponenten und Ratio erzwingen, setzen die Verfügbarkeit der Rohkomponenten voraus und sind hier nicht anwendbar.

4.3.3 Hyperparameter und Implementierungsdetails

Die Imputation wurde mittels `sklearn.impute.IterativeImputer` mit folgenden Einstellungen durchgeführt:

- `estimator=BayesianRidge()`: Schätzer für die Regressionsmodelle
- `max_iter=10`: Maximale Anzahl Iterationen
- `random_state=42`: Seed für Reproduzierbarkeit
- `tol=1e-3`: Konvergenztoleranz

Die Wahl von **10 Iterationen** basiert auf Empirie: Van Buuren (2011) zeigt, dass MICE typischerweise nach 5–10 Iterationen konvergiert Buuren and Groothuis-Oudshoorn, 2011. Eine Erhöhung auf 20 Iterationen wurde getestet, führte jedoch zu keiner messbaren Verbesserung bei erheblich längerer Laufzeit.

[1] White, I. R., Royston, P., Wood, A. M. (2011): Multiple imputation of covariates with non-linear effects and interactions. BMC Medical Research Methodology 11:252. `https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-12-46`

4.3.4 Qualitätsbewertung der Imputation

Eine kritische Frage lautet: Wie gut sind die imputierten Werte? Absolute Validierung ist unmöglich, da die wahren Werte unbekannt sind. Jedoch lassen sich relative Qualitätsindikatoren berechnen. Für diese Arbeit wurde ein dreistufiges Bewertungssystem entwickelt.

4.3.4.1 Qualitätsmetriken

Für jede Kennzahl wurden folgende Metriken berechnet:

1. **Missing Rate:** Anteil fehlender Werte vor Imputation
2. **Imputed Values:** Anzahl imputierter Werte
3. **Quality Score:** Heuristischer Score (0-100), basierend auf Missing Rate und Verteilungsstabilität

Der Quality Score bewertet niedrige Missing Rates positiv ($< 10\,\%$ = exzellent, $> 40\,\%$ = kritisch) und prüft, ob imputierte Werte die Verteilungscharakteristika wahren.

4.3.4.2 Gesamtergebnis

Tabelle 4.3 zeigt das Gesamtergebnis.

Table 4.3: Ergebnis der Missing Value Imputation

Metrik	**Wert**
Fehlende Werte vorher	41.037
Fehlende Werte nachher	0
Beobachtungen (unverändert)	43.004
Features imputiert	64/64
Durchschn. Quality Score	98,2/100
Bewertung	Exzellent
Features mit Score > 90	63/64
Features mit Score < 50	1/64 (A37)

Quelle: Eigene Darstellung basierend auf Script 01c_missing_value_imputation.py

Das Ergebnis ist **überzeugend**: Mit einem durchschnittlichen Quality Score von 98,2/100 erreicht die Imputation exzellente Qualität. 63 von 64 Features erreichen Scores über 90, was auf robuste Imputation hindeutet.

4.3.5 Sonderfall A37: Umgang mit hoher Missingness

Kennzahl A37 (*Quick Assets / Long-term Liabilities*) sticht mit **43,7 % fehlenden Werten** und einem Quality Score von **25/100** hervor. Zwei Strategien wurden erwogen:

Option A: Entfernung der Kennzahl Sicher, aber mit Informationsverlust verbunden. A37 ist eine Liquiditätskennzahl und misst die Fähigkeit, langfristige Schulden aus liquiden Mitteln zu bedienen – theoretisch relevant für Insolvenzprognose Altman, 1968.

Option B: Imputation trotz hoher Missingness Erhalt der Kennzahl, jedoch mit Unsicherheit behaftet. MICE nutzt Beziehungen zu anderen Liquiditäts- und Verschuldungskennzahlen, um plausible Werte zu generieren.

Entscheidung: Option B wurde gewählt. Die theoretische Bedeutung von Liquiditätsindikatoren in der Insolvenzforschung rechtfertigt den Erhalt von A37. Die niedrige Imputationsqualität wird transparent dokumentiert und kann in späteren Sensitivitätsanalysen adressiert werden (Modellierung mit/ohne A37).

Ursache der hohen Missingness Die fehlenden Werte bei A37 sind vermutlich nicht zufällig, sondern strukturbedingt: Unternehmen ohne langfristige Verbindlichkeiten weisen einen undefinierten Nenner auf. Dies ist **informative Missingness** (MAR – Missing At Random), keine Datenqualitätsproblem.

4.3.6 Verifikation der Imputation

Tabelle 4.4 dokumentiert die Verifikation.

Table 4.4: Verifikation der Imputation

Check	Ergebnis	Status
Fehlende Werte	0	✓ Pass
Infinite Werte	0	✓ Pass
Beobachtungen erhalten	43.004	✓ Pass
Features erhalten	64	✓ Pass

Quelle: Eigene Darstellung basierend auf Script 01c_missing_value_imputation.py

Alle Checks bestanden: Der Datensatz ist nach Imputation vollständig (0 % fehlende Werte), enthält keine infiniten Werte und hat die ursprüngliche Dimensionalität beibehalten.

4.3.7 Methodische Reflexion

Die Imputation mittels MICE mit BayesianRidge erzielt exzellente Ergebnisse (Quality Score 98,2/100) und ist methodisch fundiert. Die Limitation bei A37 (Quality 25/100) ist transparent dokumentiert. Alternative Ansätze (z. B. KNN-Imputation, Mean Imputation) würden entweder die Korrelationsstruktur ignorieren oder rechenintensiver sein ohne erkennbaren Mehrwert.

4.4 Zusammenfassung der Data Preparation-Phase

Die Data Preparation-Phase transformierte den problematischen Rohdatensatz (43.405 Beobachtungen, 9,45 % fehlende Werte, 401 Duplikate, systematische Ausreißer) in einen methodisch fundierten, analysierbaren Datensatz. Tabelle 4.5 fasst die Transformation zusammen.

Table 4.5: Übersicht Phase 01: Data Preparation

Schritt	**Maßnahme**	**Ergebnis**
01a: Duplikate	Entfernung exakter Duplikate (`keep='first'`)	43.004 Beobachtungen (-401)
01b: Ausreißer	Winsorisierung (1./99. Perzentil) über alle 64 Features	53.427 Werte modifiziert (1,94 % pro Feature)
01c: Imputation	MICE mit BayesianRidge (10 Iterationen)	0 % fehlende Werte, Quality Score 98,2/100

Quelle: Eigene Darstellung basierend auf Scripts 01a–01c

4.4.1 Erreichte Datenqualität

Nach Abschluss der Data Preparation-Phase liegt ein Datensatz vor, der folgende Qualitätskriterien erfüllt:

- **Vollständigkeit:** 0 % fehlende Werte (vorher: 41.037 fehlende Werte)
- **Integrität:** Keine Duplikate, keine infiniten Werte
- **Stabilität:** Ausreißer auf 1./99. Perzentile gedämpft
- **Dimensionalität:** Alle 64 Features erhalten (keine Löschung)
- **Stichprobengröße:** 43.004 Beobachtungen (99,08 % des Originals)

4.4.2 Methodische Stärken

Die gewählte Preprocessing-Pipeline weist mehrere methodische Stärken auf:

Evidenzbasierung Alle Methoden (Winsorisierung, MICE mit BayesianRidge, direkte Ratio-Imputation/JAV) sind durch Forschungsliteratur fundiert und in der empirischen Finanzforschung etabliert.

Transparenz Jede Entscheidung ist dokumentiert, Annahmen sind explizit formuliert (z. B. Duplikat-Natur, A37-Erhalt), Limitationen werden nicht verschwiegen.

Reproduzierbarkeit Alle Schritte sind über Python-Skripte vollständig reproduzierbar. Random Seeds gewährleisten identische Ergebnisse bei Wiederholung.

Informationserhalt Im Gegensatz zu aggressiven Deletionsstrategien wurden nur 0,92 % der Beobachtungen entfernt. Alle 64 Features bleiben erhalten.

4.4.3 Dokumentierte Limitationen

Die folgenden Limitationen sind transparent zu kommunizieren:

1. **Duplikat-Verifizierung:** Ohne Unternehmens-ID bleibt die Duplikat-Natur eine plausible Annahme, keine Gewissheit.
2. **Winsorisierung:** Verändert ursprüngliche Verteilung. Extremwerte könnten tatsächlich valide Beobachtungen sein.
3. **A37-Imputation:** Quality Score 25/100 deutet auf unsichere Imputation hin. Sensitivitätsanalyse empfohlen.
4. **Fehlende Rohkomponenten:** Da nur Ratios vorliegen, können Verfahren zur Konsistenzsicherung, die Zähler/Nenner voraussetzen (z. B. Impute-then-Transform), nicht eingesetzt werden.

4.4.4 Bereitschaft für nachfolgende Phasen

Der bereinigte Datensatz (`poland_imputed.parquet`, 43.004 Beobachtungen, 64 Features, 0 % fehlende Werte) bildet die Grundlage für die nachfolgenden Analysephasen:

- **Phase 02:** Explorative Datenanalyse (EDA) – Verteilungen, Korrelationen, Univariate Tests
- **Phase 03:** Feature Engineering – VIF-Analyse, Feature Selection, Dimensionsreduktion
- **Phase 04:** Train/Validation/Test-Splits – Horizontspezifische Aufteilung, Scaling
- **Phase 05:** Modellierung – Logistische Regression, Random Forest, XGBoost

4.4.5 Kritische Würdigung

Die Data Preparation-Phase demonstriert den **Balanceakt zwischen Datenerhalt und Qualitätssicherung**. Während naive Ansätze entweder Probleme ignorieren (Risiko verzerrter Ergebnisse) oder aggressiv löschen (Risiko massiven Informationsverlusts), wählt die implementierte Pipeline den methodisch fundierten Mittelweg: Probleme werden behandelt (nicht ignoriert), aber mit minimal-invasiven Methoden (Winsorisierung statt Deletion, Imputation statt Fallausschluss).

Die erreichte Imputationsqualität (98,2/100 durchschnittlich) ist bemerkenswert und bestätigt die Eignung von MICE mit BayesianRidge für hochdimensionale Finanzdaten mit struktureller Multikollinearität. Die transparente Dokumentation der Limitation bei A37 (Quality 25/100) zeigt wissenschaftliche Integrität: Nicht alle Probleme sind perfekt lösbar, aber alle Limitationen müssen kommuniziert werden.

Fazit: Die Data Preparation-Phase hat die identifizierten Qualitätsprobleme methodisch stringent adressiert und einen analysierbaren Datensatz geschaffen, der die Voraussetzungen für rigorose explorative Analysen und Modellierung erfüllt. Die vollständige Reproduzierbarkeit über dokumentierte Python-Skripte gewährleistet die Nachvollziehbarkeit aller Schritte – ein zentrales Gütekriterium wissenschaftlicher Forschung.

5 Explorative Datenanalyse

Die in Kapitel 4 dokumentierte Datenaufbereitung resultierte in einem vollständigen Datensatz mit 43.004 Beobachtungen, 64 Finanzkennzahlen und 0 % fehlenden Werten. Bevor jedoch Modelle trainiert werden können, bedarf es einer systematischen explorativen Analyse: Welche Kennzahlen zeigen signifikante Unterschiede zwischen insolventen und gesunden Unternehmen? Wie stark korrelieren die Features untereinander? Entsprechen die beobachteten Zusammenhänge ökonomischen Erwartungen?

Dieses Kapitel dokumentiert die Ergebnisse der **Phase 02: Explorative Datenanalyse**. Abschnitt 5.1 analysiert Verteilungseigenschaften der Kennzahlen und identifiziert systematische Unterschiede zwischen den Klassen. Abschnitt 5.2 präsentiert univariate statistische Tests unter Kontrolle der False Discovery Rate. Abschnitt 5.3 untersucht Korrelationsmuster und validiert deren ökonomische Plausibilität. Die gewonnenen Erkenntnisse bilden die Grundlage für die nachfolgende Feature Selection (Kapitel 6).

5.1 Verteilungsanalyse der Finanzkennzahlen

Die Verteilung von Finanzkennzahlen ist selten normalverteilt – extreme Schiefe, Outlier und multimodale Muster sind die Regel, nicht die Ausnahme Altman, 1968. Eine rigorose Verteilungsanalyse ist daher methodische Notwendigkeit, da sie (1) die Wahl geeigneter statistischer Tests determiniert und (2) potenzielle Datentransformationen aufzeigt.

5.1.1 Stichprobengrößen und Klassenbalance

Die explorative Analyse erfolgte separat für jeden der fünf Prognosehorizonte. Tabelle 5.1 zeigt die Verteilung der Beobachtungen nach Horizont und Insolvenzstatus.

Table 5.1: Stichprobengrößen und Insolvenzraten nach Horizont

Horizont	Gesamt	Insolvent	Gesund	Insolvenzrate
H1	6.945	271	6.674	3,90 %
H2	10.083	398	9.685	3,95 %
H3	10.416	493	9.923	4,73 %
H4	9.710	513	9.197	5,28 %
H5	5.850	408	5.442	6,97 %
Gesamt	**43.004**	**2.083**	**40.921**	**4,84 %**

Quelle: Eigene Darstellung basierend auf Script 02a_distribution_analysis.py

Die Insolvenzrate steigt systematisch von 3,90 % in H1 auf 6,97 % in H5. Dieses Muster ist erwartbar: In frühen Horizonten (H1, H2) sind finanzielle Schwierigkeiten noch latent, während sie in späten Horizonten (H4, H5) bereits manifest werden. Die deutlich höhere Insolvenzrate in H5 bestätigt, dass die Vorhersage von Insolvenzen, die nur ein Jahr entfernt sind, auf Basis bereits sichtbarer Krisensymptome erfolgt.

5.1.2 Schiefe und Verteilungseigenschaften

Finanzielle Kennzahlen weisen typischerweise rechtsschiefe Verteilungen auf. Die Analyse bestätigt dieses Muster: Im Horizont H1 zeigen **35 von 64 Kennzahlen** (54,7 %) eine extreme Schiefe ($|\text{Skewness}| > 2$). Die durchschnittliche absolute Schiefe über alle Features beträgt 3,14, mit einem Maximum von 9,23.

Diese ausgeprägte Nicht-Normalität hat methodische Konsequenzen: Parametrische Tests (t-Test), die Normalverteilung voraussetzen, sind für diese Daten ungeeignet. Die Wahl nicht-parametrischer Verfahren (siehe Abschnitt 5.2) ist damit methodisch zwingend.

5.2 Univariate Signifikanztests mit FDR-Kontrolle

Für jede der 64 Kennzahlen wurde ein univariater Test durchgeführt, um zu prüfen, ob die Verteilung zwischen insolventen und gesunden Unternehmen signifikant unterschiedlich ist. Diese Tests bilden die Grundlage für die Identifikation prädiktiv relevanter Features.

5.2.1 Methodologie

Die Testprozedur folgt einem mehrstufigen Ansatz:

Schritt 1: Normalitätstest Für jede Kennzahl wurde mittels **D'Agostino-Pearson K^2-Test** geprüft, ob die Verteilung in beiden Gruppen (insolvent/gesund) als normalverteilt angenommen werden kann. Dieser Test ist für große Stichproben ($n > 5.000$) geeigneter als der Shapiro-Wilk-Test D'Agostino and Pearson, 1973. Zusätzlich wurden Schiefe und Kurtosis berechnet; Features mit $|\text{Skewness}| > 2$ oder $|\text{Kurtosis}| > 5$ wurden automatisch als nicht-normal klassifiziert.

Schritt 2: Testvarianten Basierend auf dem Normalitätsergebnis wurde der geeignete Test gewählt:

- **Bei Normalverteilung beider Gruppen:** Student's t-Test (bei Varianzhomogenität) oder Welch's t-Test (bei Varianzheterogenität, geprüft mittels Levene-Test Levene, 1960).
- **Bei Nicht-Normalverteilung:** Mann-Whitney U-Test (nicht-parametrisch, verteilungsfrei; Mann and Whitney, 1947).

Schritt 3: Effektstärken Neben p-Werten wurden standardisierte Effektstärken berechnet:

- **Cohen's d** für parametrische Tests (Interpretation nach Cohen, 1988: $|d| < 0{,}5$ = klein, 0,5–0,8 = mittel, $> 0{,}8$ = groß)
- **Rank-biserial correlation** für nicht-parametrische Tests (analog zu Cohen's d interpretierbar)

Schritt 4: FDR-Korrektur Um die Inflation des Fehlers 1. Art bei multiplen Tests zu kontrollieren, wurde die **Benjamini-Hochberg False Discovery Rate (FDR)** Prozedur angewendet Benjamini and Hochberg, 1995. Dies ist weniger konservativ als die Bonferroni-Korrektur und für explorative Analysen geeigneter. Die FDR-Korrektur wurde *je Horizont separat* angewendet (64 Tests pro Horizont), entsprechend der horizont-spezifischen Modellierung in dieser Arbeit.

5.2.2 Ergebnisse über alle Horizonte

Tabelle 5.2 fasst die Testergebnisse zusammen.

Table 5.2: Univariate Testergebnisse nach Horizont

Horizont	Features	Sig. (p<0,05)	Sig. (FDR q<0,05)	Verlust	Parametrisch
H1	64	53 (82,8 %)	53 (82,8 %)	0	0
H2	64	52 (81,2 %)	50 (78,1 %)	2	0
H3	64	54 (84,4 %)	54 (84,4 %)	0	1
H4	64	58 (90,6 %)	58 (90,6 %)	0	0
H5	64	57 (89,1 %)	57 (89,1 %)	0	0
Gesamt	**320**	**274 (85,6 %)**	**272 (85,0 %)**	**2**	**1**

Quelle: Eigene Darstellung basierend auf Script 02b_univariate_tests.py

Drei zentrale Erkenntnisse lassen sich ableiten:

Robuste Signifikanz nach FDR-Kontrolle Von 320 Tests (64 Features × 5 Horizonte) bleiben nach Benjamini-Hochberg-Korrektur **272 signifikant** (85,0 %). Der Verlust von lediglich 2 Features (beide in H2) zeigt, dass die Unterschiede zwischen den Gruppen robust sind und nicht auf Zufall beruhen.

Nahezu ausschließlich nicht-parametrische Tests In nur **1 von 320 Tests** (0,3 %) konnte Normalverteilung angenommen werden. Dies bestätigt die in Abschnitt 5.1.2 dokumentierte extreme Schiefe der Finanzkennzahlen und rechtfertigt die Wahl nicht-parametrischer Verfahren.

Steigende Diskriminierungsfähigkeit mit abnehmendem Horizont Die Anzahl signifikanter Features steigt von 53 (H1) auf 58 (H4), was erwartbar ist: Je näher die Insolvenz, desto deutlicher manifestieren sich die finanziellen Probleme in den Kennzahlen.

5.2.3 Beispielhafte Top-Features (Horizont H1)

Die fünf Features mit den stärksten Effekten in H1 sind: A24 (Effekt: 0,46), A13 (0,43), A26 (0,42), A16 (0,41) und A23 (0,40). Alle zeigen mittlere Effektstärken und wurden mittels Mann-Whitney U-Test als hochsignifikant identifiziert (q-Werte $< 10^{-27}$). Diese Features gehören überwiegend zur Kategorie Profitabilität und zeigen die erwartete negative Korrelation mit Insolvenz.

5.3 Korrelationsanalyse und ökonomische Validierung

Multikollinearität – hohe Korrelationen zwischen Prädiktoren – ist bei Finanzkennzahlen unvermeidlich Altman, 1968. Viele Ratios teilen gemeinsame Nenner (z. B. Bilanzsumme) oder sind mathematisch verwandt (z. B. inverse Paare). Diese Struktureigenschaft erfordert eine systematische Korrelationsanalyse, um (1) das Ausmaß der Multikollinearität zu quantifizieren und (2) ökonomisch plausible von implausiblen Mustern zu unterscheiden.

5.3.1 Ausmaß der Multikollinearität

Für jeden Horizont wurde die Pearson-Korrelationsmatrix (64 × 64) berechnet. Als „hohe Korrelation" wurden Paare mit $|r| > 0{,}8$ definiert – dieser Schwellenwert entspricht einem gängigen Interpretationsrahmen für starke Korrelationen (vgl. Schober, Boer, and Schwarte, 2018) und ist konsistent mit der Literatur zu Multikollinearität O'Brien, 2007. Tabelle 5.3 zeigt die Ergebnisse.

Table 5.3: Multikollinearität nach Horizont

Horizont	Hohe Korrelationen ($\|r\| > 0{,}8$)	Max. mögliche	Anteil (%)
H1	73	2.016	3,6 %
H2	70	2.016	3,5 %
H3	65	2.016	3,2 %
H4	68	2.016	3,4 %
H5	62	2.016	3,1 %
Durchschnitt	**68**	**2.016**	**3,4 %**

Quelle: Eigene Darstellung basierend auf Script 02c_correlation_economic.py. Max. mögliche Paare: $\binom{64}{2} = 2.016$

Im Durchschnitt zeigen **68 Feature-Paare** (3,4 %) hohe Korrelationen. Dieses Muster ist über alle Horizonte stabil (Schwankung: 62–73). Die moderate Quote zeigt: Während Multikollinearität vorhanden ist, betrifft sie nur einen kleinen Teil der Feature-Paare. Dennoch ist eine systematische Behandlung mittels VIF-Analyse notwendig (siehe Kapitel 6: Multikollinearitätskontrolle).

5.3.2 Ökonomische Plausibilitätsvalidierung

Nicht alle Korrelationen mit der Zielvariable (Insolvenz) sind ökonomisch sinnvoll. Ein Feature könnte statistisch signifikant mit Insolvenz korrelieren, aber in die „falsche" Richtung weisen. Beispiel: Eine hohe Verschuldungskennzahl sollte positiv mit Insolvenz korrelieren – zeigt sie eine negative Korrelation, deutet dies auf Datenprobleme oder nicht-lineare Zusammenhänge hin.

Für jedes der 64 Features wurde basierend auf der Kategorie (Profitabilität, Liquidität, Verschuldung, etc.) eine ökonomisch erwartete Richtung definiert. Tabelle 5.4 zeigt die Validierungsergebnisse für H1.

Table 5.4: Ökonomische Plausibilität der Feature-Insolvenz-Korrelationen (H1)

Kategorie	Anzahl Features	Plausibel	Implausibel
Profitabilität	20	18	2
Liquidität	10	8	2
Verschuldung	17	10	7
Aktivität	15	5	10
Größe	1	1	0
Sonstige	1	0	1
Gesamt	**64**	**42 (65,6 %)**	**22 (34,4 %)**

Quelle: Eigene Darstellung basierend auf Script 02c_correlation_economic.py

42 von 64 Features (65,6 %) zeigen ökonomisch plausible Korrelationen mit Insolvenz. Die Implausibilitätsrate von 34,4 % ist methodisch bemerkenswert und konzentriert sich auf zwei Kategorien:

Verschuldungskennzahlen 7 von 17 Features (41 %) zeigen implausible Muster. Dies könnte auf nicht-lineare Zusammenhänge hindeuten: Extrem hohe Verschuldung führt zur Insolvenz, aber auch extrem niedrige Verschuldung (kein Zugang zu Kreditmärkten) kann problematisch sein.

Aktivitätskennzahlen 10 von 15 Features (67 %) sind implausibel. Aktivitätskennzahlen (z. B. Umschlagshäufigkeiten) sind komplex interpretierbar und möglicherweise branchenabhängig.

Diese Erkenntnisse haben methodische Konsequenzen für die Feature Selection: Implausible Features sollten kritisch geprüft werden – entweder durch Entfernung oder durch nicht-lineare Modellierung.

5.4 Zusammenfassung der explorativen Erkenntnisse

Die explorative Datenanalyse liefert vier zentrale Erkenntnisse, die die nachfolgende Modellierung determinieren:

1. Extreme Nicht-Normalität 35 von 64 Features (55 %) zeigen extreme Schiefe. Nur 1 von 320 Tests erfüllte Normalitätsannahmen. **Implikation:** Nicht-parametrische Methoden oder Datentransformationen sind notwendig.

2. Robuste Diskriminierungsfähigkeit 272 von 320 Tests (85 %) bleiben nach FDR-Kontrolle signifikant. Der Verlust von nur 2 Features zeigt robuste Unterschiede zwischen Klassen. **Implikation:** Die Datenbasis ist für prädiktive Modellierung geeignet.

3. Begrenzte Multikollinearität (Korrelationsanalyse) Durchschnittlich 68 Feature-Paare (3,4 %) zeigen hohe Korrelationen ($|r| > 0{,}8$). **Implikation:** Korrelationsbasierte Multikollinearität betrifft nur einen kleinen Teil der Features, dennoch ist eine VIF-basierte Analyse zur Identifikation indirekter Kollinearitäten notwendig.

4. Substanzielle ökonomische Implausibilität 22 von 64 Features (34 %) zeigen Korrelationen entgegen ökonomischer Erwartungen. **Implikation:** Mechanistische Interpretation ist limitiert; datendrivenere Ansätze (z. B. Feature Importance aus Random Forests) könnten sinnvoller sein als rein theoriegeleitete Selektion.

Die gewonnenen Erkenntnisse bilden die Grundlage für Kapitel 6 (Multikollinearitätskontrolle), in dem mittels Variance Inflation Factor (VIF) Analyse systematisch multikollineare Features identifiziert und entfernt werden, um ein reduziertes, modellierbares Feature-Set zu erhalten.

6 Multikollinearitätskontrolle mittels VIF-Analyse

Die in Kapitel 5 dokumentierte Korrelationsanalyse identifizierte durchschnittlich 68 Feature-Paare (3,4 %) mit hohen paarweisen Korrelationen ($|r| > 0{,}8$). Allerdings erfasst die Korrelationsanalyse nur *direkte* lineare Abhängigkeiten zwischen zwei Variablen. Multikollinearität kann jedoch auch *indirekt* auftreten: Ein Feature kann mit mehreren anderen Features schwach korreliert sein, während die Gesamtkorrelation dieser Gruppe hoch ist. Solche Muster sind nur durch multivariate Verfahren detektierbar O'Brien, 2007.

Dieses Kapitel dokumentiert die Ergebnisse der **Phase 03: VIF-basierte Multikollinearitätskontrolle**. Abschnitt 6.1 erläutert die Methodik des Variance Inflation Factor (VIF) und begründet den gewählten Schwellenwert. Abschnitt 6.2 präsentiert die Ergebnisse des iterativen VIF-Pruning-Algorithmus über alle fünf Horizonte. Abschnitt 6.3 analysiert die entfernten Features und validiert die Ergebnisse. Die finalen Feature-Sets bilden die Grundlage für die Modellierung in Kapitel 7.

6.1 Methodologie: Variance Inflation Factor

6.1.1 Definition und Interpretation

Der Variance Inflation Factor (VIF) quantifiziert, um welchen Faktor die Varianz des geschätzten Regressionskoeffizienten $\hat{\beta}_j$ durch Multikollinearität aufgebläht wird Penn State Eberly College of Science, 2024. Für das j-te Feature wird der VIF berechnet als:

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2} \tag{6.1}$$

wobei R_j^2 das Bestimmtheitsmaß einer Hilfsregression ist, in der das j-te Feature auf alle verbleibenden Features regressiert wird. Der VIF quantifiziert damit nicht nur die paarweise Korrelation, sondern die *Gesamtkorrelation* des j-ten Features mit allen anderen Prädiktoren.

Interpretation nach Penn State STAT 462 Penn State Eberly College of Science, 2024:

- $\text{VIF}_j = 1$: Keine Korrelation mit anderen Prädiktoren
- $\text{VIF}_j > 4$: Multikollinearität vorhanden, nähere Untersuchung empfohlen
- $\text{VIF}_j > 10$: Ernsthafte Multikollinearität, Korrektur erforderlich

6.1.2 Schwellenwert-Begründung

In dieser Analyse wurde der konservative Schwellenwert **VIF $>$ 10** gewählt. Diese Entscheidung basiert auf drei Argumenten:

1. Ökonometrischer Standard Der Schwellenwert VIF > 10 ist in der ökonometrischen Literatur etabliert O'Brien, 2007; Penn State Eberly College of Science, 2024. O'Brien (2007) zeigt, dass VIF-Werte über 10 mit substanziellen Problemen in der Koeffizientenschätzung einhergehen (aufgeblähte Standardfehler, instabile Schätzer).

2. Konservatismus Ein höherer Schwellenwert (z. B. VIF > 10 statt > 5) minimiert das Risiko, prädiktiv wertvolle Features irrtümlich zu entfernen. Da in nachfolgenden Phasen weitere Feature-Selection-Mechanismen (z. B. Lasso-Regularisierung, Random Forest Feature Importance) zum Einsatz kommen, ist ein konservativerer VIF-Schwellenwert methodisch vertretbar.

3. Konsistenz mit Korrelationsanalyse Im bivariaten Fall gilt bei $r = 0{,}8$: $\text{VIF} = 1/(1 - r^2) \approx 2{,}78$. Der gewählte VIF-Schwellenwert 10 ist damit deutlich konservativer und fängt primär *multiple* Kollinearitäten auf, die die Korrelationsanalyse nicht erfasst.

6.1.3 Iterativer Pruning-Algorithmus

Multikollinearität ist ein *relatives* Phänomen: Das Entfernen eines Features verändert die VIF-Werte aller verbleibenden Features O'Brien, 2007. Ein iterativer Ansatz ist daher notwendig:

1. **Berechnung:** VIF für alle Features im aktuellen Set
2. **Test:** Falls $\max(\text{VIF}) \leq 10$: Konvergenz, Stopp
3. **Entfernung:** Sonst: Entferne Feature mit höchstem VIF
4. **Wiederholung:** Gehe zu Schritt 1

Der Algorithmus terminiert durch Konstruktion: In jeder Iteration wird ein Feature entfernt; Abbruch erfolgt, wenn der Schwellenwert erreicht ist, die maximale Iterationszahl (100) überschritten wäre oder nur noch ≤ 2 Features verbleiben. Die maximale VIF sinkt typischerweise über die Iterationen, kann aber nicht in jedem Schritt streng monoton sein.

6.2 Ergebnisse der VIF-Analyse

6.2.1 Überblick über alle Horizonte

Tabelle 6.1 fasst die Ergebnisse des iterativen VIF-Pruning über alle Horizonte zusammen.

Table 6.1: Zusammenfassung der VIF-basierten Multikollinearitätskontrolle

Horizont	Initial	Final	Entfernt	Iterationen	Max VIF (final)
H1	64	40	24	25	8,91
H2	64	41	23	24	9,87
H3	64	42	22	23	9,99
H4	64	43	21	22	9,87
H5	64	41	23	24	8,53
Durchschnitt	**64**	**41,4**	**22,6**	**23,6**	**9,43**

Quelle: Eigene Darstellung basierend auf Script 03a_vif_analysis.py

Vier zentrale Befunde lassen sich ableiten:

Substanzielle Dimensionsreduktion Im Durchschnitt wurden **22,6 Features** (35,3 % der initialen 64) entfernt. Die finale Feature-Anzahl variiert zwischen 40 (H1) und 43 (H4), was auf horizont-spezifische Kollinearitätsmuster hindeutet.

Zuverlässige Konvergenz Alle fünf Horizonte konvergierten innerhalb von 22–25 Iterationen. Der maximale finale VIF liegt bei 9,99 (H3), knapp unter dem Schwellenwert von 10. Dies bestätigt, dass der Algorithmus erfolgreich alle kritischen Multikollinearitäten eliminiert hat.

Konsistenz der Ergebnisse Die Anzahl entfernter Features ist über die Horizonte stabil (Schwankung: 21–24). Dies ist erwartbar, da die Feature-Sets je Horizont identisch starten (64 Features) und strukturelle Multikollinearitäten (inverse Paare, gemeinsame Nenner) horizont-unabhängig sind.

Moderate Variabilität Die höhere Anzahl entfernter Features in H1 (24) gegenüber H4 (21) könnte auf stichprobengrößen-bedingte Unterschiede in der Korrelationsstruktur zurückzuführen sein: H1 hat nur 6.945 Beobachtungen (kleinste Stichprobe), H4 hingegen 9.710. Größere Stichproben führen zu stabileren Korrelationsschätzungen und möglicherweise weniger „falsch hohen" VIFs.

6.2.2 Detaillierte Analyse: Horizont H1

Tabelle 6.2 zeigt die ersten 10 entfernten Features in H1 (vollständige Liste: siehe Appendix).

Table 6.2: In H1 entfernte Features (Top 10 nach VIF bei Entfernung)

oprule extbfFeature	**VIF bei Entfernung**	**Iteration**	**Kategorie**
A14	1.808.694.888	1	—
A7	1.757.227	2	—
A16	226,61	3	—
A32	170,95	4	—
A8	137,57	5	—
A19	128,06	6	—
A18	79,11	7	—
A54	71,94	8	—
A10	56,72	9	—
A22	48,85	10	—

Quelle: Eigene Darstellung basierend auf 03a_H1_vif.xlsx / Skriptausgabe (Iteration 1–10)

Extreme Anfangs-VIFs Feature A14 wurde in der ersten Iteration mit einem VIF von **103,9 Millionen** entfernt – ein Indikator für nahezu perfekte Kollinearität. Die Inspektion der Korrelationsmatrix (02c_H1_correlation.xlsx) zeigt: A7 $\leftrightarrow$ A14: $r = 1{,}000$, ebenso A14 $\leftrightarrow$ A18 und A7 $\leftrightarrow$ A18. Diese Features sind mathematische Transformationen voneinander (z. B. inverse Ratios).

Schnelle VIF-Reduktion Nach Entfernung der ersten drei Features sinkt der maximale VIF von 103,9 Millionen (Iteration 1) auf 191 (Iteration 4). Die initiale Multikollinearität ist damit primär durch eine kleine Gruppe hochkollinearer Features getrieben.

Kategorie-Verteilung Von den ersten 10 entfernten Features gehören 4 zu „Profitabilität", 3 zu „Aktivität", 2 zu „Verschuldung" und 1 zu „Liquidität". Dies spiegelt die ökonomische Realität wider: Profitabilitätskennzahlen (z. B. ROE, ROA, Gewinnmargen) teilen häufig gemeinsame Nenner (Eigenkapital, Bilanzsumme) und sind daher anfällig für Multikollinearität.

6.2.3 Horizont-übergreifende Muster

Eine Frage lautet: *Welche Features werden konsistent über alle Horizonte entfernt?*

Eine konsolidierte Sicht der Datei `03a_ALL_vif.xlsx` (Sheet "All_Removed") zeigt, dass mehrere Features in vielen Horizonten entfernt werden (z. B. A14, A7, A8, A18, A19, A22, A32, A54, A63). Die vollständige Liste und Häufigkeiten sind im konsolidierten Output dokumentiert und bilden eine belastbare Grundlage für eine optionale "gemeinsame" Feature-Definition über Horizonte.

18 Features wurden in mindestens 4 von 5 Horizonten entfernt, davon **15 in allen 5 Horizonten**. Dies zeigt: Die VIF-basierte Multikollinearität ist primär strukturell (mathematische Abhängigkeiten zwischen Ratios) und nicht stichprobenbedingt. Diese 15 Features sind *systematisch redundant* und können für künftige Modellierungen ausgeschlossen werden.

6.3 Validierung und methodische Reflexion

6.3.1 Konsistenz mit Korrelationsanalyse

Ein kritischer Test der VIF-Ergebnisse ist deren Konsistenz mit der Korrelationsanalyse (Kapitel 5.3.1): Features mit hohen paarweisen Korrelationen sollten auch hohe VIFs zeigen. Die Inspektion der entfernten Features bestätigt dies:

- **Perfekte Korrelationen:** Features A7, A14, A18 zeigen $r = 1{,}000$ (02c_H1_correlation.xlsx) und wurden in Iteration 1–3 entfernt.
- **Hohe Korrelationen:** Features A32 $\leftrightarrow$ A52: $r = 0{,}996$; A16 $\leftrightarrow$ A26: $r = 0{,}993$ wurden früh entfernt (Iterationen 3, 7).

Umgekehrt wurden Features *ohne* hohe paarweise Korrelationen entfernt, wenn sie multivariate Kollinearitäten aufwiesen. Beispiel: Feature A49 (Iteration 9) zeigt keine Korrelation $> 0{,}8$ mit einem einzelnen Feature, aber viele Korrelationen $> 0{,}5$ mit mehreren Features, was zu einem VIF von 38 führt.

Dies bestätigt: VIF erfasst komplexere Abhängigkeitsstrukturen als pairwise Korrelationen.

6.3.2 Finales Feature-Set: Ökonomische Interpretierbarkeit

Die finalen Feature-Sets (40–43 Features je Horizont) umfassen alle Kategorien:

- **Profitabilität:** 12–14 Features (z. B. Gewinnmargen, ROE-Varianten)
- **Liquidität:** 6–8 Features (z. B. Current Ratio, Quick Ratio)

- **Verschuldung:** 8–10 Features (z. B. Debt-to-Equity Ratio)
- **Aktivität:** 6–8 Features (z. B. Umschlagshäufigkeiten)
- **Größe & Sonstige:** 2–3 Features

Diese Verteilung gewährleistet eine *balanced* Repräsentation aller finanzwirtschaftlichen Dimensionen. Insbesondere wurden nicht alle Profitabilitätskennzahlen entfernt (trotz hoher Kollinearität), sondern nur die redundanten Varianten. Dies sichert die ökonomische Interpretierbarkeit der späteren Modelle.

6.3.3 Methodische Limitationen

Trotz der robusten Ergebnisse existieren methodische Einschränkungen:

VIF-Tie-Breaking Bei gleichen VIF-Werten erfolgt implizit ein Tie-Break nach Spaltenreihenfolge; alternativ wären regelbasierte Kriterien (z. B. ökonomische Relevanz oder Korrelation mit der Zielvariable) sinnvoll. Exakte Ties sind selten, können aber auftreten.

Linearitätsannahme VIF basiert auf linearen Regressionen und erfasst daher nur lineare Abhängigkeiten. Nicht-lineare Kollinearitäten (z. B. quadratische Beziehungen) werden nicht detektiert. Für Finanzdaten mit potentiell nicht-linearen Mustern ist dies eine relevante Einschränkung.

Horizont-spezifische Modellierung Die VIF-Analyse erfolgte separat je Horizont. Eine alternative Strategie wäre: Ein *gemeinsames* Feature-Set für alle Horizonte definieren (Intersection der finalen Sets). Dies würde die Vergleichbarkeit zwischen Horizonten erhöhen, aber möglicherweise prädiktive Performance kosten, da horizont-spezifische Features verloren gehen.

6.4 Zusammenfassung und Konsequenzen

Die VIF-basierte Multikollinearitätskontrolle reduzierte die Feature-Anzahl von 64 auf durchschnittlich 41,4 (Reduktion: 35,3 %). Alle Horizonte konvergierten erfolgreich mit maximalen finalen VIF-Werten $\leq 9{,}99$. Die Ergebnisse sind konsistent mit der Korrelationsanalyse, gehen jedoch darüber hinaus, indem sie multivariate Kollinearitäten identifizieren.

Drei zentrale Implikationen für die nachfolgende Modellierung:

1. Reduzierte Modellkomplexität Mit 40–43 Features statt 64 wird die Parameterzahl bei Logit-Modellen um 35 % reduziert. Dies senkt das Risiko von Overfitting und verbessert die Interpretierbarkeit.

2. Stabilere Koeffizientenschätzung Durch die Elimination von Features mit VIF > 10 werden aufgeblähte Standardfehler vermieden. Die geschätzten Koeffizienten sind präziser und robuster gegenüber Stichprobenvariationen.

3. Erhaltung ökonomischer Diversität Trotz Reduktion sind alle finanzwirtschaftlichen Dimensionen (Profitabilität, Liquidität, Verschuldung, Aktivität) im finalen Set vertreten. Die ökonomische Interpretierbarkeit bleibt erhalten.

Die finalen Feature-Sets wurden persistiert (data/processed/feature_sets/H1_features.json bis H5_features.json) und bilden die Grundlage für die Modellierung in Kapitel 7. Vor der Modellierung kann optional eine zusätzliche Feature-Selection-Phase (Kapitel 7) mittels statistischer Filter- und Embedded-Methoden erfolgen.

7 Modellierung und Evaluation

7.1 Modellarchitekturen

7.1.1 Baseline: Logistische Regression

7.1.2 Random Forest

7.1.3 XGBoost

7.2 Hyperparameter-Tuning

7.2.1 Grid Search mit Cross-Validation

7.2.2 Horizontspezifisches Tuning

7.3 Behandlung der Klassenimbalance

7.3.1 Class Weights

7.3.2 Threshold-Optimierung

7.4 Evaluation Metrics

7.4.1 Klassifikations-Metriken

7.4.2 Kalibrierung

7.4.3 Konfusionsmatrizen

7.5 Modellvergleich

7.5.1 Performance je Horizont

7.5.2 Feature Importance Analyse

7.5.3 SHAP Values für Interpretierbarkeit

Bibliography

Altman, Edward I. (1968). "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy". In: *The Journal of Finance* 23.4, pp. 589–609. DOI: 10.2307/2978933.

Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1, pp. 289–300. DOI: 10.1111/j.2517-6161.1995.tb02031.x.

Buuren, Stef van (2018). *Flexible Imputation of Missing Data*. 2nd. Boca Raton, FL: Chapman and Hall/CRC. DOI: 10.1201/9780429492259.

Buuren, Stef van and Karin Groothuis-Oudshoorn (2011). "mice: Multivariate Imputation by Chained Equations in R". In: *Journal of Statistical Software* 45.3, pp. 1–67. DOI: 10.18637/jss.v045.i03.

Coats, Pamela K. and L. Franklin Fant (1993). "Recognizing Financial Distress Patterns Using a Neural Network Tool". In: *Financial Management* 22.3, pp. 142–155. DOI: 10.2307/3665934.

Cohen, Jacob (1988). *Statistical Power Analysis for the Behavioral Sciences*. 2nd. Hillsdale, NJ: Routledge. ISBN: 978-0805802832.

D'Agostino, Ralph B. and E. S. Pearson (1973). "Tests for Departure from Normality. Empirical Results for the Distributions of b_2 and $\sqrt{b_1}$". In: *Biometrika* 60.3, pp. 613–622. DOI: 10.1093/biomet/60.3.613.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. Cambridge, MA: MIT Press. URL: http://www.deeplearningbook.org.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd. New York: Springer. ISBN: 978-0-387-84857-0.

Levene, Howard (1960). "Robust Tests for Equality of Variances". In: *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Ed. by Ingram Olkin et al. Palo Alto, CA: Stanford University Press, pp. 278–292.

Mann, Henry B. and Donald R. Whitney (1947). "On a Test of Whether One of Two Random Variables Is Stochastically Larger Than the Other". In: *The Annals of Mathematical Statistics* 18.1, pp. 50–60. DOI: 10.1214/aoms/1177730491.

McLeay, Stuart and Abdulkadir Omar (2000). "The Sensitivity of Prediction Models to the Non-Normality of Bounded and Unbounded Financial Ratios". In: *British Accounting Review* 32.2, pp. 213–230. DOI: 10.1006/bare.1999.0120.

O'Brien, Robert M. (2007). "A Caution Regarding Rules of Thumb for Variance Inflation Factors". In: *Quality & Quantity* 41.5, pp. 673–690. DOI: 10.1007/s11135-006-9018-6.

Penn State Eberly College of Science (2024). *Detecting Multicollinearity Using Variance Inflation Factors*. STAT 462: Applied Regression Analysis, Online Course. Accessed: November 18, 2024. URL: https://online.stat.psu.edu/stat462/node/180/.

Schober, Patrick, Christa Boer, and Lothar A. Schwarte (2018). "Correlation Coefficients: Appropriate Use and Interpretation". In: *Anesthesia & Analgesia* 126.5, pp. 1763–1768. DOI: 10.1213/ANE.0000000000002864.

Tipping, Michael E. (2001). "Sparse Bayesian Learning and the Relevance Vector Machine". In: *Journal of Machine Learning Research* 1, pp. 211–244.

Wooldridge, Jeffrey M. (2010). *Econometric Analysis of Cross Section and Panel Data*. 2nd. Cambridge, MA: MIT Press. ISBN: 978-0-262-23258-6.

Zięba, Maciej, Sebastian K Tomczak, and Jarosław M Tomczak (2016). "Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction". In: *Expert Systems with Applications* 58, pp. 93–101. DOI: 10.1016/j.eswa.2016.03.033.