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 **goodfellow2016deep**.

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 **hastieelementsstatistical2009**.

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 **zieba2016ensemble**. 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 **wooldridge2010econometric**.

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 **altman1968financial**. 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 **hastieelementsstatis**
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 **altman1968financial** – 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 **coats1993recognizing**. McLeay & Omar (2000) bestätigen diesen Befund und warnen vor der Annahme zeitlicher Homogenität bei Multi-Horizont-Daten **mcleay2000prediction**. 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 **goodfellow2016deep**. 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 **goodfellow2016deep**.

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

4.1 Behandlung fehlender Werte

4.1.1 Passive Imputation für Finanzkennzahlen

4.1.2 Umgang mit Kennzahl A37

4.2 Behandlung von Duplikaten und Ausreißern

4.2.1 Duplikat-Entfernung

4.2.2 Winsorisierung extremer Werte

4.3 Train/Validation/Test-Split

4.3.1 Strategie bei horizontspezifischen Modellen

4.3.2 Stratifiziertes Sampling

4.4 Feature Scaling

4.4.1 Z-Score-Normalisierung

4.4.2 Horizontspezifische Skalierung

5 Feature Engineering und Selektion

5.1 Explorative Datenanalyse

5.1.1 Univariate Analyse

5.1.2 Bivariate Analyse

5.2 Multikollinearitätsanalyse

5.2.1 Variance Inflation Factor (VIF)

5.2.2 Korrelationsbasierte Filterung

5.3 Feature Selection

5.3.1 Filter-Methoden

5.3.2 Wrapper-Methoden

5.3.3 Embedded-Methoden

5.4 Finale Feature-Sets

5.4.1 Reduzierte Feature-Sets je Horizont

5.4.2 Validierung der Feature-Sets

6 Modellierung und Evaluation

6.1 Modellarchitekturen

6.1.1 Baseline: Logistische Regression

6.1.2 Random Forest

6.1.3 XGBoost

6.2 Hyperparameter-Tuning

6.2.1 Grid Search mit Cross-Validation

6.2.2 Horizontspezifisches Tuning

6.3 Behandlung der Klassenimbalance

6.3.1 Class Weights

6.3.2 Threshold-Optimierung

6.4 Evaluation Metrics

6.4.1 Klassifikations-Metriken

6.4.2 Kalibrierung

6.4.3 Konfusionsmatrizen

6.5 Modellvergleich

6.5.1 Performance je Horizont

6.5.2 Feature Importance Analyse

6.5.3 SHAP Values für Interpretierbarkeit