Herausforderung des maschinellen Lernens in der praktischen Prüfarbeit

D I S S E R T A T I O N
der Universität St.Gallen,
Hochschule für Wirtschafts-,
Rechts- und Sozialwissenschaften,
Internationale Beziehungen und Informatik (HSG),
zur Erlangung der Würde eines
Doktors der Wirtschaftswissenschaften

vorgelegt von

Sven Bösiger

von

Langenthal (Bern)

Genehmigt auf Antrag der Herren

Prof. Dr. Peter Leibfried

und

Prof. Dr. Klaus Möller

Dissertation Nr. 5267

Difo-Druck GmbH, Untersiemau 2023

Die Universität St.Gallen, Hochschule für Wirtschafts-, Rechts- und Sozialwissenschaften, Internationale Beziehungen und Informatik (HSG), gestattet hiermit die Drucklegung der vorliegenden Dissertation, ohne damit zu den darin ausgesprochenen Anschauungen Stellung zu nehmen.

St.Gallen, 24. Oktober 2022

Der Rektor:

Prof. Dr. Bernhard Ehrenzeller

Vorwort

Mit dem Abschluss meiner Arbeit möchte ich an dieser Stelle allen beteiligten Personen, die mich bei der Anfertigung meiner Doktorarbeit unterstützt haben, meinen grossen Dank aussprechen.

Insbesondere gilt dies meinem Doktorvater Prof. Dr. Peter Leibfried sowie dem Korreferenten Prof. Dr. Klaus Möller für die ausgezeichnete Betreuung und akademische Freiheit, die sie mir bei der Umsetzung der Arbeit zuteilwerden liessen.

Auch möchte ich Dr. Philippe Wenk und Dr. Robert Gwadera meinen Dank aussprechen, die mir im Rahmen der Anwendung maschinellen Lernens mit Rat, kritischen Anmerkungen und produktiven Gesprächen zur Seite gestanden haben.

Ohne den Zugang zu spannenden und anspruchsvollen Kunden wäre diese Arbeit undenkbar gewesen, womit mein besonderer Dank der Prüfgesellschaft gilt, die sich massgeblich für die Zusammenarbeit eingesetzt hat und stark darum bemüht war, Türen bei Kunden zu öffnen. Dem Erfolg dieser Bemühungen ist es zu verdanken, dass eine intensive Kooperation mit Kunden sowie der Prüfgesellschaft aufgebaut und gepflegt werden konnte, die entscheidend zum Erkenntnisgewinn beigetragen hat.

Ausserdem möchte ich mich bei all meinen Kolleginnen und Kollegen bedanken, die mir im Rahmen meiner Doktoratszeit an der Universität St. Gallen nicht nur eine unvergessliche Zeit bereitet haben, sondern mich stets auch mit bestem Rat und hilfreichen Anregungen begleitet haben.

Meiner Familie und meinen Freunden danke ich für ihre Geduld und mannigfaltige Unterstützung während der Ausarbeitung dieser Doktorarbeit.

Zürich, Januar 2023 Sven Bösiger

Inhaltsverzeichnis

1 Einleitung **12**

1.1 Übersicht 12

1.2 Problemstellung 14

1.3 Aufbau der Arbeit 16

2 Theorie **18**

2.1 Maschinelles Lernen 18

2.1.1 Aktualität 18

2.1.2 Begriff des maschinellen Lernens 20

2.1.3 Ziele des maschinellen Lernens 21

2.1.4 Methoden zur Schätzung von f 23

2.1.5 Abwägung von Genauigkeit und Interpretierbarkeit 25

2.1.6 Überwachtes versus nicht überwachtes Lernen 27

2.1.7 Regressions- versus Klassifikationsprobleme 28

2.1.8 Evaluation der Modellgenauigkeit 29

2.1.9 Kreuzvalidierung 30

2.1.10 Multiples lineares Modell 33

2.1.11 Entscheidungsbäume 34

2.1.12 Support-Vektor-Maschine (SVM) 37

2.1.13 Künstliche neuronale Netze 39

2.1.14 Assoziationsanalyse 44

2.1.15 Clustering mit k-Means 46

2.1.16 Maschinelles Lernen im Prozess 48

2.2 Wirtschaftsprüfungsprozess 54

2.2.1 Ziele und Vorgehensweise während der Abschlussprüfung . . 54

2.2.2 Rahmenbedingungen 55

2.2.3 Aussagen im Abschluss 56

2.2.4 Aufteilung der Prüfarbeit in Phasen 57

2.2.5 Prüfungsplanung 58

2.2.6 Prüfungsdurchführung 65

2.2.7 Prüfabschluss 71

2.3 Maschinelles Lernen in der Prüfung 74

3 Methodologie **80**

3.1 Charakterisierung des Ansatzes 80

3.2 Vorgehen im konstruktiven Forschungsansatz 81

4 Fallstudien **87**

4.1 Gewinnung des Studienpartners 87

4.1.1 Aufnahme Forschungskooperation 87

4.1.2 Eckdaten der Zusammenarbeit 88

4.1.3 Mandatsauswahl mit dem Team 89

4.1.4 Prüfgesellschaft 89

4.2 Vorgehen in den Fallstudien 90

4.2.1 Review der einschlägigen Literatur 90

4.2.2 Akquisition der benötigten Daten 91

4.2.3 Anwendung der ML-Methoden 91

4.2.4 Diskussion der Implikationen 92

4.3 Fallstudie 1 92

4.3.1 Rahmenbedingungen 92

4.3.2 ML-unterstützte Risikobeurteilung 95

4.3.3 ML-unterstützte Aufdeckung doloser Handlungen 104

4.3.4 ML-unterstützte analytische Prüfungshandlungen 121

4.3.5 ML-unterstützte Analyse der Fortführungsfähigkeit 133

4.3.6 ML-unterstützte Berichterstattung 147

4.4 Fallstudie 2 157

4.4.1 Rahmenbedingungen 157

4.4.2 Vorgehen in der Fallstudie 159

4.4.3 ML-Ansätze aufgrund externer Beobachtungen 160

4.4.4 ML-Ansätze aufgrund unternehmensinterner Beobachtungen 164

5 Analyse **170**

5.1 Angewandte Konstrukte 170

5.1.1 Zwei Konstrukte 170

5.1.2 Notwendige Infrastruktur 173

5.2 Exogener Anwendungsfall 176
5.2.1 Übersicht 176
5.2.2 Unterprozessschritte der Modellbildung 176
5.2.3 Unterprozesschritte der Prognosegenerierung 181
5.2.4 Herausforderungen 182
5.3 Endogener Anwendungsfall 185
5.3.1 Übersicht 185
5.3.2 Unterprozessschritte 185
5.3.3 Herausforderungen 192
5.4 Generalisierbarkeit 194
5.5 Einfluss von SARS-CoV-2 auf Ergebnisse der Fallstudien 195

6 Konklusion **197**
6.1 Gewonnene Erkenntnisse 197
6.1.1 Integration des maschinellen Lernens in den Prüfprozess . . 197
6.1.2 Herausforderungen bei der Anwendung 201
6.2 Limitierungen 203
6.3 Ausblick für weiterführende Forschung 204

7 Literaturverzeichnis **206**

Abbildungsverzeichnis

1 Verortung des Begriffs des maschinellen Lernens – eignene Darstellung nach Goodfellow et al. (2016) 21
2 Zielkonflikt zwischen Genauigkeit und Interpretierbarkeit der Methoden des maschinellen Lernens nach James et al. (2013) 26
3 Validierungsset-Ansatz nach Friedman et al. (2001) 31
4 Schematische Darstellung des *Leave-one-out*-Kreuzvalidierungsansatzes nach Friedman et al. (2001) 32
5 k-Fold Kreuzvalidierung nach Friedman et al. (2001) 33
6 Erste Aufteilung im Prozess zur Modellierung eines Entscheidungsbaums (Lantz, 2015) 35
7 Zweite Aufteilung im Prozess zur Modellierung eines Entscheidungsbaums (Lantz, 2015) 36
8 Mögliche Hyperebene im Klassifikationsfall einer Support-Vektor-Maschine (Lantz, 2015) 38
9 Einzug der Maximum Margin Hyperplane zur optimalen Separierung der Klassen (Lantz, 2015) 39
10 MMH mit unvollständiger Trennung der Klassen (Lantz, 2015) . . . 40
11 Illustration des Kerneltricks nach Lantz (2015) 40
12 Funktionsweise eines Neurons nach Lantz (2015) 41
13 Plot der Sigmoidenfunktion (eigene Darstellung) 42
14 Beispiel eines neuronalen Netzes (Lantz, 2015) 43
15 Initiale Zuordnung der Beobachtungen zu Clustermittelpunkten im k-Means Clustering 48
16 Erste Aktualisierungsphase im k-Means Clustering 49
17 Abschluss des k-Means Clusterings 50
18 Phasen des CRISP-DM-Prozessmodells für die Datenanalyse nach Wirth und Hipp (2000) 51
19 Phasen des CRISP-DM-Prozessmodells für das Data Mining 53
20 Phasen im Prüfungsprozess nach Arens et al. (2016) 58
21 Phasen im *konstruktiven Forschungsansatz* nach Labro und Tuomela (2003) 86

22 Vorgehen in der Fallstudie (eigene Darstellung) 90
23 Ellenbogenkurve zur Bestimmung der optimalen Clusterzahl 116
24 Prinzip der One-Step-Ahead-Prognose nach Bishop et al. (1995) . . 122
25 Training aufgrund externer Daten mit Prognose eines Vektors (eigene Darstellung) 170
26 Training aufgrund interner Daten mit der Prognose eines Vektors (eigene Darstellung) 172
27 Unterprozessschritte im exogenen Anwendungsfall (eigene Darstellung) 177
28 Sphären der Datensammlung (eigene Darstellung) 178
29 Training aufgrund externer Daten mit Prognose eines Vektors (eigene Darstellung) 186

Tabellenverzeichnis

1 Übersicht über Methoden im maschinellen Lernen und die zugehörigen Anwendungsfälle nach Lantz (2015) 29
2 Paradigmatische Konfusionsmatrix in Anlehnung an Hastie et al. (2009) 30
3 Geschätzte Koeffizienten für die unabhängigen Variablen nach Dikmen und Küçükkocaoğlu (2010) 96
4 Konfusionsmatrix zu erwartender und tatsächlich manipulierter sowie nicht manipulierter Abschlüsse nach Dikmen und Küçükkocaoğlu (2010) 96
5 Konfusionsmatrix zu erwartender und tatsächlich manipulierter sowie nicht manipulierter Abschlüsse nach Dbouk und Zaarour (2017) 97
6 Konfusionsmatrix zu erwartender und tatsächlich manipulierter sowie nicht manipulierter Abschlüsse nach Dechow et al. (2011) . . . 98
7 Konfusionsmatrix zu erwartender und tatsächlich manipulierter sowie nicht manipulierter Abschlüsse mit nicht finanziellen Werten nach Dechow et al. (2011) 98
8 Konfusionsmatrix zu erwartender und tatsächlich manipulierter sowie nicht manipulierter Abschlüsse mit nicht finanziellen und marktbasierten Werten Dechow et al. (2011) 98
9 Features, die im Auszug aus den Umsatztransaktionen enthalten waren 112
10 k-Means Cluster und die Anzahl der enthaltenen Transaktionen . . 117
11 Übersicht über verwendete unabhängige Variablen im Modell von Koskivaara (2000a) 122
12 Ergebnisse des künstlichen neuronalen Netzes im Artikel von Koskivaara (2000a) 124
13 Ergebnisse des künstlichen neuronalen Netzes im Artikel nach Koskivaara (2000a) 129
14 Ergebnisse des künstlichen neuronalen Netzes im Artikel nach Koskivaara (2000a) 130
15 Prognosegüte des Modells von Jones (2017) 136

16 Prognosegüte des klassischen neuronalen Netzes mit *backpropagation-of-error* nach Anandarajan et al. (2001) 138
17 Prognosegüte des neuronalen Netzes mit genetischem Lernalgorithmus nach Anandarajan et al. (2001) 138
18 Features und ihre Definitionen, wie sie von Martens et al. (2008) verwendet wurden 139
19 Genauigkeit der Prognose in Prozent nach Martens et al. (2008) . . 140
20 Going-Concern Konfusionsmatrix für das Random-Forest Modell . . 144
21 Going-Concern Konfusionsmatrix eines künstlichen neuronalen Netzes 145
22 Going-Concern Konfusionsmatrix einer Support-Vektor-Maschine . 145
23 Going-Concern Konfusionsmatrix einer logistischen Regression . . . 146
24 Miteinbezogene Features im Modell von Gaganis et al. (2007) 148
25 Genauigkeiten des logistischen Regressionsmodells von Gaganis et al. (2007) 150
26 Genauigkeit in der Klassifizierung der einzuschränkenden Abschlüsse nach Kirkos, Spathis, Nanopoulos et al. (2007) 151
27 Konfusionsmatrix für die Qualifizierung des Prüfurteils mithilfe des logistischen Regressionsmodells 154
28 Konfusionsmatrix für die Qualifizierung von Prüfurteilen mithilfe des Entscheidungsbaum 154
29 Konfusionsmatrix für die Qualifizierung des Prüfurteils mithilfe eines künstlichen neuronalen Netzes 154
30 Übersicht über Autoren und zugehörige Methoden 161
31 Übersicht über Autoren und zugehörige Methoden 165
32 Angewandte Methoden bei analytischen Prüfungshandlungen 190
33 Angewandte Methoden bei unüberwachten Lernansätzen 190

Zusammenfassung

Obschon maschinelles Lernen bereits breitflächig eingesetzt wird, spielt es in der praktischen Wirtschaftsprüfung bislang nur eine untergeordnete Rolle. Dieser Technologie wird zwar ein hohes Potenzial zugeschrieben, um das Audit effizienter und effektiver zu gestalten, es fehlt jedoch eine systematische Untersuchung zu der Frage, wie die Anwendung maschinellen Lernens und der Prüfprozess ineinandergreifen und welche Herausforderungen sich daraus für die Prüfpraxis ergeben. Diese Arbeit schliesst diese Lücke mithilfe zweier Fallstudien und eines konstruktiven Forschungsansatzes bezüglich der Frage, wie der Prüfprozess und der Datenanalyseprozess in der Praxis angepasst und erweitert werden müssen, um die gewünschten Ergebnisse zu erhalten. Darüber hinaus zeigt die Untersuchung, welchen Herausforderungen sich der Prüfer gegenübersieht und wie er diesen begegnen kann. Dabei werden zwei fundamental unterschiedliche Anwendungsfälle identifiziert, welche eine differenzierte Handhabung erfordern: ein auf exogenen Labels (abhängigen Variablen) beruhender, der umweltgestützte Prognosen erlaubt, und ein auf endogenen Labels abstellender, der interne Prognosen zulässt, wobei beide in unterschiedlicher Weise in der Prüfungsvorbereitung, -durchführung und -beendigung berücksichtigt werden müssen.

Abstract

Although machine learning is being applied broadly, it plays only a minor role in the conduct of audits of financial statements. The technology is thought to have a great potential to make the audit more insightful and more efficient. Nevertheless, there is a lack of systematic research into how the audit process and the application of machine learning could be combined as well as the associated challenges. This research closes this gap. By providing two separate field studies, using a constructive research approach, it shows how the audit process and the data analytics process need to be amended, to retrieve the desired results. Moreover, it demonstrates the challenges faced by auditors and how to deal with them. Two fundamentally different use cases are identified: one relying on exogenous labels (dependend variables), that allows for externally based predictions, and another based on endogenous labels, that is used for internally based predictions. Both impact the preparation of an audit, its performance and its completion.

1 Einleitung

1.1 Übersicht

Seit 2015 ist die Wirtschaftsprüfungsbranche optimistisch, dass maschinelles Lernen die Wirtschaftsprüfung „smarter, more insightful and more efficient“ gestalten wird. „This is the future of the audit profession and the users of financial statements deserve it.“ (Raphael, 2015).

Auch Susskind und Susskind (2015) sehen in ihrem Werk über die Zukunft der Berufe die Wirtschaftsprüfung als eine der Branchen, die in erheblichem Umfang durch maschinelles Lernen verändert wird.

Diese Einschätzungen sind dem Umstand geschuldet, dass in den letzten zehn Jahren maschinelles Lernen – bewusst und unbewusst – langsam Einzug in das Leben der meisten Konsumenten[1] gehalten hat. So wird die Technologie von Unternehmen in ihren Produkten verwendet, um eine akkurate Bildsuche zu ermöglichen, menschliche Stimmen zu erkennen und Autos selbstständig durch den Verkehr zu lenken.

Nach Goodfellow et al. (2016) wird die Anwendung des maschinellen Lernens massgeblich von drei Entwicklungen begünstigt oder gar erst ermöglicht:

- Neue *theoretisch-mathematische* Erkenntnisse, insbesondere durch Hinton et al. (2006) und ihre Ausführungen zum Training eines *Deep Believe Network*.
- Dramatischer *Anstieg der Datenmengen* im Zuge der Digitalisierung der Gesellschaft, weil stetig mehr Arbeit am Computer verrichtet wird. Die gestiegene Vernetzung ermöglicht darüber hinaus eine vereinfachte Zentralisierung und Pflege der Daten.
- *Exponentieller Anstieg der Rechenleistung* mit der Einführung neuer Prozessorgenerationen und der Weiterentwicklung von GPUs, die es erlauben, zunehmend komplexere Probleme zu berechnen und aufwendigere Modelle zu schätzen.

[1] In dieser Arbeit wird aus Gründen der besseren Lesbarkeit das generische Maskulinum verwendet. Weibliche und anderweitige Geschlechteridentitäten werden dabei ausdrücklich mitgemeint, soweit es für die Aussage erforderlich ist.

Neben den bekannten Anwendungsfeldern hat maschinelles Lernen nach Lantz (2015) in weiteren Bereichen subtil Einzug gehalten: So wird die Technologie unter anderem bei der Identifizierung von Spam-Nachrichten eingesetzt, sie hilft bei der verhaltensorientierten Segmentierung für eine bessere und gezieltere Kundenansprache, unterstützt die langfristigen Wetter- und Klimaprognosen, deckt Kreditkartenmissbrauch auf, verbessert aktuarische Schätzungen zu Unwetterschäden und erlaubt genauere Wahlprognosen.

Im Kern ermöglicht maschinelles Lernen eine automatisierte Aufdeckung von Mustern in gegebenenfalls grossen Datenmengen, auf deren Grundlage Vorhersagen erstellt sowie Interdependenzen ermittelt und quantifiziert werden können (Goodfellow et al., 2016). Viele Ansätze und Konzepte, die unter dem Begriff maschinelles Lernen subsumiert werden, reichen nach Friedman et al. (2001) historisch bereits weit zurück und sehen ihren Anfang in einem Vorläufer der linearen Regression zu Beginn des 19. Jahrhunderts durch Gauss und Legendre. Von Fisher (1936) wurde die lineare Diskriminanzanalyse vorgeschlagen, um den Anwendungsbereich für qualitative Merkmale zu erweitern und eine Prognose für derartige Merkmale zu ermöglichen – beispielsweise ob ein Patient überlebt oder nicht. Bis Ende der 1970er-Jahre sind weitere Methoden wie die logistische Regression hinzugekommen.

Wie Friedman et al. (2001) bemerken, handelt es sich bei diesen Ansätzen fast ausschliesslich um lineare Modelle, weil die Schätzung von nicht linearen Modellen als zu rechenintensiv galt. In den 1980er-Jahren wurden Computer schliesslich genügend leistungsstark, sodass Leo et al. (1984) Klassifikations- und Regressionsbäume anwendeten. Hastie und Tibshirani (1987) prägten etwa zur selben Zeit den Begriff der Variablenselektionsmodelle und boten Software für Schätzmodelle an (James et al., 2013). Schliesslich liefern neue Microchip-Generationen seit 2006 künstlichen neuronalen Netze, deren Erforschung bis auf Rosenblatt (1958) zurückgeführt werden kann, genügend performante Ergebnisse für die Schätzung von komplexen Modellen.

Als Ergebnis dieser Entwicklung stehen den Unternehmen heute fertige, teilweise integrierte Lösungen – die Hardware und Software miteinander kombinieren – zur Verfügung (Kokina & Davenport, 2017). Diese Anwendungen können auch von der Wirtschaftsprüfung genutzt werden (Westermann & Spindler, 2017). Ent-

sprechend haben die grossen vier Prüfgesellschaften in diesem Bereich Investitionen getätigt: So hat KPMG verlauten lassen, *IBM Watson* einzusetzen, um Anomalien in den Finanzdaten aufzuspüren (Melendez, 2016), während PricewaterhouseCoopers (PwC) AI-Techniken wie *DeNovo* operativ verwendet (PricewaterhouseCoopers, 2016). Ernst & Young hat ein System entwickelt, das menschliches Verhalten modelliert (Ernst&Young, 2016) und Deloitte nutzt *Argus for AI* in der Wirtschaftsprüfung (Raphael, 2015).

1.2 Problemstellung

Nicht neu ist die Idee, künstliche Intelligenz (KI) oder das maschinelle Lernen, welches ebenfalls unter dem Begriff KI eingeordnet werden kann (Goodfellow et al., 2016), im Audit einzusetzen, sondern lässt sich bis zu Keenoy (1958) zurückverfolgen. Infolge des umfangreichen öffentlichen Interesses am Thema ist es nicht überraschend, dass die Möglichkeiten für die Einführung des maschinellen Lernens auch unter Wirtschaftsprüfern diskutiert werden, verfügen deren Prüfgesellschaften doch einerseits über genügend Mittel, um die erforderliche Infrastruktur und die Mitarbeiter bereitzustellen, und besitzen sie andererseits Zugriff auf grosse Mengen an betrieblichen Daten ihrer Mandanten. Das Audit bietet sich besonders für maschinelles Lernen an, hat sich die Zunahme an strukturierten und unstrukturierten Daten bei der Prüfung in den letzten Jahren doch als grosse Herausforderung erwiesen (Kokina & Davenport, 2017).

Auf nationaler wie auch auf internationaler Ebene sind die Regulierungsbehörden aktiv geworden. In der Schweiz hat die Eidgenössische Revisionsaufsichtsbehörde (RAB) bereits 2016 damit begonnen, den Einfluss der Datenanalyse und des maschinellen Lernens, das sie der Datenanalyse zurechnet, auf die Prüfung und insbesondere deren Qualität zu untersuchen (Revisionsaufsichtsbehörde, 2016). Sie hat zu diesem Zeitpunkt erkannt, dass Unternehmen sich noch in der Einführungsphase befinden und die angewandten schulungsintensiven Methoden eher die Effektivität des Audits beeinflussen als dessen Effizienz. Ein Jahr später berichtete die RAB, dass die Analyseverfahren ergänzend zum ordentlichen Prüfansatz eingesetzt werden, es jedoch noch offene Fragen für den Einsatz der Analysewerkzeuge (Revisionsaufsichtsbehörde, 2017) gibt, die derzeit vom International Auditing and

Assurance Standards Board (IAASB) – dem internationalen Standardsetter für die Wirtschaftsprüfung – gesammelt und in nächster Zukunft beantwortet werden sollen (Eimers, 2016). Grundsätzlich werden die Anwender derzeit motiviert, die Datenanalysewerkzeuge, deren Anwendung *per se* nicht ausgeschlossen wird, in der Praxis zu übernehmen (Schäfer et al., 2017).

Die Voraussetzungen für eine Einführung des maschinellen Lernens in der Wirtschaftsprüfung sind damit überwiegend erfüllt, dennoch finden sie im tatsächlichen praktischen Prüfalltag bisher keine wesentliche Anwendung (Westermann & Spindler, 2017). Gierbl et al. (2020) weisen für die Einführung bereits exemplarisch auf eine Reihe von Herausforderungen im Zusammenhang mit KI in der Prüfungspraxis hin. Wird dazu die einschlägige Literatur zum Thema bemüht, so zeigt sich, dass lediglich eine geringe Anzahl Publikationen die operativen Herausforderungen für den Wirtschaftsprüfer bei der Datenanalyse aufgreifen (Tackett, 2013; Jans et al., 2011) bzw. als Vorstufe dazu systematisch die Interaktion des Prozesses der Abschlussprüfung und der Datenanalyse mithilfe maschinellen Lernens analysieren – wo Fallstricke auf den Prüfer warten können. In den Fällen, in welchen der Prozess diskutiert wird, bleiben Vorgehen und Herausforderungen abstrakt (beispielsweise bei Sun, 2019).

Die erste in der vorliegenden Studie zu beantwortende Forschungsfrage (RQ) lautet demnach:

> *RQ1: Wie lässt sich in der Praxis der Wirtschaftsprüfungsprozess mit dem archetypischen Vorgehen des maschinellen Lernens verbinden?*

Daneben ergeben sich durch den Einsatz des maschinellen Lernens in der Wirtschaftsprüfung spezifische Herausforderungen. So zeichnen sich, wie J. L. Perols et al. (2017) analysieren konnten, beispielsweise verschiedene Wirtschaftsprüfungsprozesse durch unausgewogene Datensätze bzw. Probleme mit der Datenqualität aus.

Entsprechend lautet eine zweite Forschungsfrage:

> *RQ2: Welchen operativen Herausforderungen begegnet der Prüfer in der praktischen Anwendung maschinellen Lernens im Wirtschaftsprüfungsprozess und wie kann er diese überwinden?*

Diese Forschungsfragen werden mithilfe von zwei Fallstudien mit der Methodologie des *konstruktiven Forschungsansatzes* nach Kasanen et al. (1993) analysiert.

Primär stehen dabei der Prüfprozess und die Herausforderungen der Schweizer Prüfgesellschaften bzw. deren Mitarbeiter im Mittelpunkt, zumal in den verschiedenen Ländern die regulatorischen Auflagen bzw. deren Interpretationen anders ausfallen können (EXPERTsuisse, 2016). Zudem liegt der Fokus auf den grossen und mittelgrossen Prüfgesellschaften, die über die notwendigen finanziellen Ressourcen und Mitarbeiter verfügen, die die Anwendung des maschinellen Lernens gegenwärtig überhaupt erst ermöglichen. Prüfunternehmen verwenden maschinelles Lernen derzeit in der Datenaufbereitung, d.h. bei der Strukturierung und Verbesserung der Datenqualität. Dieser Anwendungsfall wird in der vorliegenden Untersuchung jedoch ausgeschlossen, weil derartige Arbeiten nicht zur tatsächlichen praktischen Tätigkeit eines Prüfers gehören. Überdies wird auf aktuelle Techniken, wie sie in der Literatur beispielsweise bei James et al. (2013) oder Lantz (2015) über das maschinelle Lernen analysiert werden, zurückgegriffen, weil deren Potenzial bezüglich Effektivität und Effizienz für die Prüfgesellschaften besonders hoch ist.

Ziel der Forschung ist es, die Herausforderungen offenzulegen, die mit der praktischen Anwendung des maschinellen Lernens einhergehen. Gleichzeitig sollen mögliche Lösungsansätze für die aufgedeckten Probleme dargelegt werden.

1.3 Aufbau der Arbeit

Zur Beantwortung der Forschungsfragen wird wie folgt vorgegangen.

In Kapitel 2 werden die theoretischen Grundlagen für die Arbeit vorgestellt. Abschnitt 2.1 widmet sich der Theorie zum maschinellen Lernen, seinen Methoden und Prozessen. Anschliessend werden in Abschnitt 2.2 der Wirtschaftsprüfungsprozess und die einzelnen Schritte mit ihren Unterprozessen genauer beleuchtet. Es folgt in Abschnitt 2.3 eine Ausführung zur Literatur, die maschinelles Lernen und die Wirtschaftsprüfung miteinander verbindet.

Kapitel 3 enthält Ausführungen zum methodologischen Hintergrund, dem konstruktiven Forschungsansatz nach Kasanen et al., 1993, der für die Durchführung dieser Fallstudie herangezogen wurde. Es geht dabei um den vorgesehenen Studi-

enablauf und die Möglichkeiten zur Einschätzung bzw. Bewertung des Konstrukts.

Kapitel 4 widmet sich der Darstellung sowohl der Abläufe der Fallstudien von ihrer Aufsetzung bis hin zum Abschluss, wie auch des Einsatzes des maschinellen Lernens in den einzelnen Prüfprozessen und -phasen.

Kapitel 5 befasst sich anschliessend mit einer Gegenüberstellung der Erkenntnisse aus den Fallstudien und der theoretischen Ausführungen in Kapitel 2, wobei als Synthese die Konstrukte hervorgehen sowie Herausforderungen und Lösungen für den Einsatz des maschinellen Lernens in der Wirtschaftsprüfung offengelegt werden. Zudem wird die Generalisierbarkeit der Konstrukte diskutiert.

Kapitel 6 fasst die herausgearbeiteten Erkenntnisse zusammen, zeigt Limitierungen auf und gibt einen Ausblick für weiterführende Forschungsfragen.

2 Theorie

2.1 Maschinelles Lernen

2.1.1 Aktualität

Entwicklung der Theorie Obwohl maschinelles Lernen (ML) als Begriff relativ neu ist, reichen viele seiner Konzepte historisch weit zurück (James et al., 2013). So verfassten Legendre (1806) und Gauss (1809) bereits anfangs des 19. Jahrhunderts Aufsätze über die Methode der kleinsten Quadrate, die heute als früheste Form der linearen Regression verstanden wird. Für die Prognose gleichfalls qualitativer Werte, beispielsweise zu der Frage, ob ein Patient überlebt oder stirbt oder ob ein Börsenkurs steigt oder fällt, führte Fisher (1936) die lineare Diskriminanzanalyse ein. Schliesslich wurde in den 1940er-Jahren die logistische Regression entwickelt (James et al., 2013).

Zu diesen Ansätzen kamen in den 1940er-Jahren rechnergestützte Methoden hinzu, die biologisches Lernen abbildeten. Insbesondere das McCulloch-Pitts-Neuron war ein frühes Modell, mit dem die Funktionsweise des Gehirns nachgebildet werden sollte. Dieses lineare Modell konnte zwei Kategorien von Eingangsstimuli unterscheiden (Goodfellow et al., 2016). In den 1950er-Jahren wurde mit dem Perzeptron ein erstes Modell von Rosenblatt (1958) geschaffen, das auf der Grundlage einer Liste von Beispieldaten für Inputs und Kategorien Vorhersagen ermöglichte. Als dahinterstehender Trainingsalgorithmus kam eine modifizierte Version des Gradientenabschlagsverfahren zum Einsatz, das noch heute im Bereich des maschinellen Lernens verwendet wird (Goodfellow et al., 2016). Bei diesen Ansätzen wurden jedoch eine Reihe von Anwendungsfällen vom Training ausgeschlossen, weil sie auf lineare Modelle limitiert waren: So war es beispielsweise nicht möglich, eine XOR-Funktion zu approximieren. Dieses Problem führte zunächst dazu, dass das Interesse an neuronalen Netzen zurückging (Goodfellow et al., 2016).

Bis in die 1970er-Jahre kamen viele weitere Methoden hinzu, die das Lernen auf der Grundlage der vorgegebenen Datensätze ermöglichten. Es handelte sich jedoch fast ausschliesslich um lineare Methoden, weil die damalige Rechenleistung die Berechnung von nicht linearen Beziehungen noch nicht zuliess. In den 1980er-Jahren war die Computertechnologie schliesslich so weit vorangeschritten, dass auch nicht

lineare Modelle berechnet werden konnten. In den 1980er-Jahren bot sich Breiman et al. (1984) damit die Möglichkeit, Klassifikations- und Regressionsbäume zu entwickeln. Gleichfalls waren sie die Ersten, die Techniken wie *Kreuzvalidierung* und *Modellselektionen* in der Praxis implementierten (James et al., 2013). Gleichzeitig wurden in diesen Jahren erstmals künstliche neuronale Netze gebildet, die mithilfe des dazu notwendigen *Back-Propagation Algorithmus* berechnet werden konnten. Die spezifischen Algorithmen waren aber immer noch zu rechenintensiv, daher setzten sich bis 2007 vorerst konkurrierende Ansätze wie *Kernelmethoden* durch.

In den letzten Jahren haben sich die Fortschritte im maschinellen Lernen weiter beschleunigt. Einen grossen Anteil daran hat auf theoretischer Ebene Geoffrey Hinton, der seit 2006 effiziente Ansätze entwickelt, die sämtliche anderen maschinellen Lernmethoden übertreffen (Goodfellow et al., 2016). Hinzu kamen zahlreiche leistungsstarke und relativ benutzerfreundliche Softwarepakete wie *R* und *Python*, die als Open-Source-Produkte frei verfügbar sind. Beide Entwicklungen führen dazu, dass diese Methoden immer häufiger zur Anwendung kommen (James et al., 2013).

Entwicklung der Datenmengen Obschon bereits seit Langem viele theoretische Grundlagen vorlagen, haben sich die Ansätze des maschinellen Lernens erst in den letzten Jahren durchgesetzt (Goodfellow et al., 2016). Das hat damit zu tun, dass maschinelles Lernen zum Training der Modelle eine grosse Menge an Daten benötigt, die aber erst seit neuester Zeit in entsprechender Zahl verfügbar sind. Massgeblich beeinflusst wird dieser Trend von der zunehmenden Digitalisierung der Gesellschaft. Mit der wachsenden Verbreitung und Vernetzung von Computern werden mehr und mehr Daten aufgezeichnet. Zudem ist es einfacher geworden, diese Daten zu zentralisieren und zu verwalten, wodurch mehr Datensätze für das maschinelle Lernen zur Verfügung stehen. Das Zeitalter von *Big Data* hat das maschinelle Lernen sehr vereinfacht, weil sich die Bürde statistischer Inferenz aufgrund kleiner Datenmengen wesentlich gelichtet hat (Goodfellow et al., 2016).

Entwicklung der Rechenleistung Ein weiterer Grund für den Erfolg des maschinellen Lernens sind die massiv verbesserten Rechenleistungen, die es zulassen,

dass die Grösse und Komplexität der Modelle wesentlich gesteigert werden konnten (Goodfellow et al., 2016).

So war die Zahl der Neuronen, die in künstlichen neuronalen Netzen zum Einsatz kommen, noch bis vor Kurzem sehr klein. Die Netze konnten sich, seit der Einführung von versteckten Schichten, in ihrer Grösse alle 2.4 Jahre verdoppeln.

Die gesteigerte Modellkomplexität und -grösse ist aufgrund der schnelleren CPUs, des Aufkommens von GPUs, der schnelleren Netzwerkverbindungen und der besseren Infrastruktur die wichtigsten Trends in der Geschichte des maschinellen Lernens (Goodfellow et al., 2016).

2.1.2 Begriff des maschinellen Lernens

Der Begriff des maschinellen Lernens (ML) wird im Folgenden in Anlehnung an Samuel (1969) verwendet und steht für einen Forschungsbereich, der sich mit Computern auseinandersetzt, denen die Fähigkeit gegeben wird, zu lernen, ohne explizit programmiert zu werden.

Damit kann maschinelles Lernen als Teilgebiet der künstlichen Intelligenz (KI) gesehen werden. Maschinelles Lernen ist im Gegensatz zu alternativen, konkurrierenden Techniken fähig, sich selbst Wissen über die Extraktion von Mustern in Daten anzueignen (Goodfellow et al., 2016). Als besonders mächtig und flexibel hat sich *Deep Learning* herausgestellt, das die Welt in Form von hierarchisierenden Konzepten repräsentiert, wobei jedes Konzept durch seine Beziehung zu einfacheren Konzepten definiert wird (Goodfellow et al., 2016).

Abbildung 1 stellt die Verortung des Begriffs des maschinellen Lernens grafisch dar.

Es existieren eine Reihe von Begriffen, die das Phänomen *maschinelles Lernen* kennzeichnen. So wird in der Literatur auch des *Statistischen Lernens* (James et al. 2013) verwendet, während andere Autoren die Bezeichnung *Computational Lerning* (Samuel, 1969) oder *Predictive Analytics* (Nyce & Cpcu, 2007) wählen.

Eine nahe verwandte Methode des maschinellen Lernens ist das *Data Mining*. Obwohl noch Unklarheiten über die Abgrenzung bestehen, wird in der Literatur üblicherweise davon ausgegangen, dass maschinelles Lernen sich mit der Ausführung bekannter Aufgaben auseinandersetzt, während das *Data Mining* neue

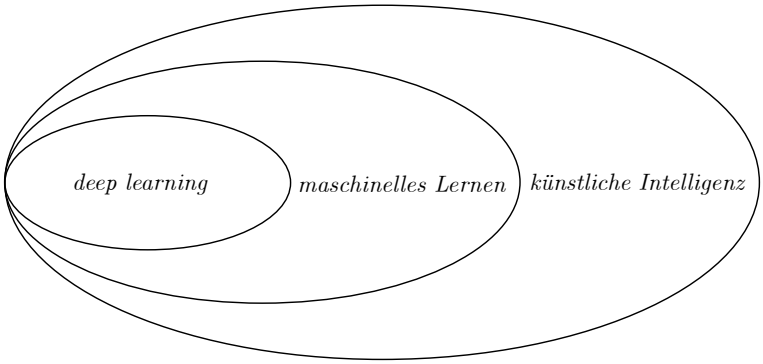


Abbildung 1: Verortung des Begriffs des maschinellen Lernens – eignene Darstellung nach Goodfellow et al. (2016)

Einsichten (*hidden nuggets*) aus einer umfangreichen Datenmenge gewinnen möchte. Beispielsweise eignet sich maschinelles Lernen dazu, einem Kraftfahrzeug zu vermitteln, sich selbst zu lenken, während *Data Mining* dazu verwendet werden kann, herauszufinden, welches Kraftfahrzeug besonders sicher ist. Grundsätzlich bildet maschinelles Lernen eine Voraussetzung für das *Data Mining*, während diese Voraussetzung in umgekehrter Richtung nicht zutrifft (Lantz, 2015).

2.1.3 Ziele des maschinellen Lernens

Schätzung einer Funktion am Beispiel der Wirtschaftsprüfung Während seiner Tätigkeit trifft der Wirtschaftsprüfer üblicherweise auf Informationen für die externe, aber auch für die interne Berichterstattung. Diese Informationen können potenzielle Wirkungszusammenhänge zwischen mehreren Grössen enthalten, die er sich zunutze machen kann (Arens et al., 2016).

So ist für den Prüfer beispielsweise erkennbar, dass die Absatzmengen seines Mandanten, der qualitativ hochwertigen Stahl produziert, massgeblich von der konjunkturellen Lage und dem Euro-CHF-Wechselkurs abhängen. Falls der Prüfer darlegen kann, dass dieser Zusammenhang besteht und gleichfalls eine Quantifizierung möglich ist, sind Prognosen über die Absatzmengen möglich.

In diesem Beispiel sind die konjunkturelle Lage und der Wechselkurs Eingabevariablen, während die Absatzmenge die *Ausgabevariable* ist. Typischerweise werden die *Eingabevariablen* mit den Symbolen X_1 bis X_p gekennzeichnet, während die *Ausgabevariable* mit Y dargestellt wird. Synonyme für *Eingabevariablen* sind

die Begriffe unabhängige Variable, *Prediktoren* oder der im Datascience-Umfeld gängige Begriff *Features*, während sich in der Literatur die Begriffe *abhängige Variable* oder *Response* für *Ausgangsvariable* finden (James et al., 2013).

Allgemein nimmt der Prüfer an, dass eine Beziehung zwischen der abhängigen Variable Y und den unabhängigen Variablen $X_1, X_2, ...X_p$ besteht, die wie folgt beschrieben werden kann (Hastie et al., 2009):

$$Y = f(X) + e$$

Hier ist f eine fixe, aber unbekannte Funktion der Variablen X_1 bis X_p und e ein Zufallsterm, der unabhängig von X ist und einen Erwartungswert von 0 aufweist.

Grundsätzlich lässt sich maschinelles Lernen auf die Schätzung der Funktion f reduzieren (James et al., 2013).

Nutzen der Schätzung von f Es gibt zwei wesentliche Gründe, die es interessant machen, eine Funktion f zu schätzen (James et al., 2013):

- *Prognosefunktion*

  Wie im Beispiel der Schätzung von Absatzmengen sind die Inputvariablen X gut verfügbar, weil Konjunktur- und Wechselkursdaten einfach zu beschaffen sind. Da der Fehlerterm einen Erwartungswert von 0 hat, kann eine Prognose mit

  $$\hat{Y} = \hat{f}(X)$$

  erstellt werden, wobei $\hat{f}(X)$ die Schätzung für die tatsächliche Funktion f ist und $\hat{Y}$ die resultierende Prognose darstellt. In einem solchen Gebilde wird $\hat{f}(X)$ häufig als *Blackbox* bezeichnet, zumal sie nicht direkt von Interesse ist, sondern lediglich akkurate Schätzungen für Y zulassen soll.

- *Inferenz*

  Ein weiterer Grund, warum die Funktion $\hat{f}(X)$ von Interesse ist, liegt in dem Verständnis begründet, wie sich eine Veränderung der unabhängigen Variablen auswirkt. Anders als bei der Prognose kann $\hat{f}(X)$ nicht mehr als Blackbox verstanden werden, sondern ihre exakte Form muss bekannt sein.

Im Beispiel der Absatzmengen könnte sich der Prüfer fragen, welcher Treiber diese Kennzahl massgeblich beeinflusst – die Wechselkurse oder die konjunkturelle Lage? Zudem steht die Frage im Raum, in welcher Weise die Wechselkurse mit den Absatzmengen in Verbindung stehen: Steigt oder sinkt die Absatzmenge mit fallender Konjunktur? Des Weiteren stellt sich die Frage, ob eine lineare Beziehung zwischen jeder einzelnen unabhängigen und der abhängigen Variablen besteht oder ob die Beziehung doch komplizierter ist.

Abhängig von der Frage, ob das endgültige Ziel entweder die Prognose, die Inferenz oder eine Kombination ist, kommen unterschiedliche Methoden des Schätzens von f zur Anwendung. Beispielsweise sind es lineare Modelle, die eine relativ simple und einfach zu interpretierende Inferenz liefern, mit ihrer Prognosegüte aber hinter anderen Modellen zurückbleiben. Umgekehrt gibt es hochgradig nicht lineare Modelle, die sehr genaue Prognosen ermöglichen, bei denen die Inferenz jedoch zur Herausforderung wird.

2.1.4 Methoden zur Schätzung von f

Lineare und nicht lineare Ansätze verfügen teilweise über gemeinsame Merkmale. So wird für die Schätzung der Funktion f bei beiden Gruppen eine Anzahl an n Beobachtung herangezogen, die die Funktion approximieren, was auch als *Training* bezeichnet wird. Die dafür verwendeten Daten werden entsprechend Trainingsdaten oder *Trainingset* genannt (James et al., 2013).

Im Beispiel der Absatzmengen könnten 30 Beobachtungen von abgesetztem Stahl und die dazugehörigen Konjunktur- und Wechselkursdaten vorliegen. Sei x_{ij} die j-te unabhängige Variable für Beobachtung i, wobei $i = 1, 2, ...n$ und $j = 1, 2, ...p$. Gleichzeitig sei y_i die i-te abhängige Variable. Dann bestehen die Trainingsdaten formal aus $\{(x_1, y_1), (x_2, y_2), ...(x_n, y_n)\}$ wobei $x_i = (x_{i1}, x_{i2}, ...x_{ip})^T$.

Das Ziel besteht darin, Methoden des maschinellen Lernens in einer Art und Weise anzuwenden, dass sie die unbekannte Funktion f schätzen. Dies bedeutet, dass eine Funktion f gefunden werden soll, sodass $Y \approx \hat{f}(X)$ für sämtliche Beobachtungen (X, Y) gilt. Grundsätzlich lassen sich alle Methoden in parametrische und nicht parametrische Funktionen einteilen (Hastie et al., 2009).

Parametrische Funktionen Die Schätzung der parametrischen Funktionen erfordert einen zweistufigen Modellansatz (James et al., 2013):

- Zuerst muss eine Annahme zur funktionellen Form von f getroffen werden. In einem sehr einfachen Fall ist f linear in X: $f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_P X_P$. Dieses Vorgehen vereinfacht die Schätzung massgeblich, zumal die Schätzung einer komplett losgelösten Funktion f durch eine lineare Funktion mit $p + 1$ Koeffizienten ersetzt wird.

- Nach der Wahl des Modells ist anschliessend ein Ansatz notwendig, der das Modell mithilfe der Trainingsdaten trainiert. Im Falle des linearen Modells ist eine Schätzung der Parameter β_0, β_1,...β_P erforderlich. Es sollen Parameter gefunden werden, für die gilt: $Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_P X_P$. Der am häufigsten verwendete Ansatz, der das Modell schätzt, ist die *Kleinste-Quadrat Methode*. Es existieren jedoch weitere Methoden, um ein lineares Modell zu trainieren.

  Der dargestellte Ansatz wird als *parametrisch* bezeichnet. Er reduziert das Problem der Schätzung von f auf die Schätzung einer Reihe von Parametern. Wird eine parametrische Form für f angenommen, so ist es grundsätzlich einfach, Parameter wie $\beta_0, \beta_1, ..., \beta_P$ zu schätzen. Ein potenzieller Nachteil eines solchen Ansatzes besteht darin, dass das ausgewählte Modell selten der tatsächlichen unbekannten Form entspricht. Daher können die Prognosewerte erheblich abweichen, wenn die tatsächliche Funktion ebenfalls erheblich davon abweicht. Dem Problem kann begegnet werden, indem ein flexibleres Modell gewählt wird, das verschiedene funktionale Formen für f zulässt. Dies bedeutet aber üblicherweise, dass eine grössere Anzahl von Parametern zu schätzen ist.

Nicht parametrische Methoden Nicht parametrische Methoden treffen keine expliziten Annahmen über die funktionale Form von f. Sie suchen vielmehr eine Schätzung von f, welche sich so nahe an die Datenpunkte wie möglich annähert, ohne dabei zu uneben oder zu verwackelt zu sein (James et al., 2013).

Damit haben solche Ansätze wesentliche Vorteile gegenüber ihren parametrischen Pendants: So gewährleistet das Abrücken von einer angenommenen funk-

tionalen Form die Möglichkeit, dass die Funktion akkurat eine grosse Anzahl von möglichen Ausprägungen annimmt. Denn jede parametrische Methode birgt die Gefahr in sich, dass die angenommene funktionale Form, welche zur Schätzung von f verwendet wird, tatsächlich sehr weit von der wirklichen Form von f entfernt ist, wobei in diesem Fall das resultierende Modell nur sehr schlecht die Daten approximiert. Demgegenüber stehen die nicht parametrischen Ansätze, welche dieses Problem komplett umgehen, da keine Annahme über die Form von f gemacht wird. Ein gewaltiger Nachteil begleitet die nicht parametrische Ansätze dennoch: Da bei ihnen eben nicht nur eine kleine Anzahl von Parametern geschätzt werden muss, sind sie auf eine grosse Anzahl von Beobachtungen (wesentlich mehr als bei parametrischen) angewiesen, um eine Schätzung von f zu erhalten.

2.1.5 Abwägung von Genauigkeit und Interpretierbarkeit

Das maschinelle Lernen kennt eine Reihe unterschiedlicher Methoden in der Gruppe der parametrischen und nicht parametrischen Ansätze: Gewisse Ansätze sind weniger flexibler oder restriktiver in dem Sinne, dass sie zur Schätzung von f lediglich eine geringe Anzahl verschiedener Formen annehmen können. Als Beispiel lässt sich die lineare Regression anführen, die als relativ unflexibel gilt, weil sie nur lineare Funktionen abbilden kann. Andere Methoden wie die *künstliche neuronale Netze* sind hingegen wesentlich flexibler und fähig, eine viel grössere Anzahl verschiedener Formen für f zu schätzen.

Aufgrund dieser Ausführungen könnte der Verdacht entstehen, dass flexiblere Ansätze den restriktiveren Ansätzen stets vorzuziehen sind. Es gibt jedoch gute Gründe, restriktivere Modelle zu wählen. Insbesondere wenn es um die Inferenz geht, lassen sich die restriktiveren Modelle wesentlich einfacher interpretieren. So kann ein lineares Modell mit dem Ziel der Inferenz eine gute Wahl sein, weil der Zusammenhang zwischen Y und $X_1, X_2, ..., X_p$ sehr einfach verständlich ist. Umgekehrt können flexiblere Ansätze wie *Spline-Kurven* oder *Boosting-Methoden* zu wesentlich komplizierteren Schätzungen von f führen, für die es sehr schwer ist zu verstehen, wie ein individuelles *Feature* mit der abhängigen Variablen interagiert.

Abbildung 2 zeigt die Abwägung zwischen Flexibilität und Interpretierbarkeit für eine Auswahl von Methoden. So gilt die Methode der Schätzung eines linearen

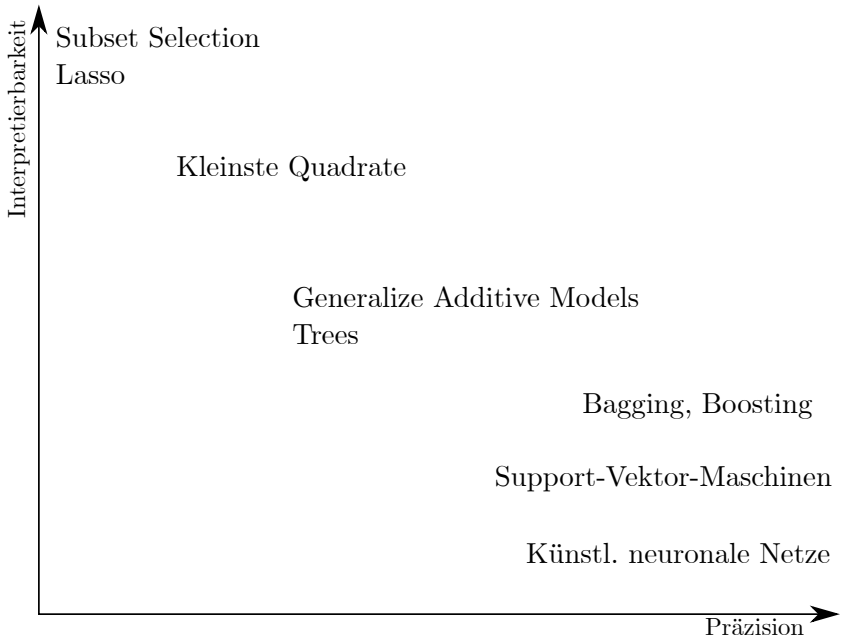


Abbildung 2: Zielkonflikt zwischen Genauigkeit und Interpretierbarkeit der Methoden des maschinellen Lernens nach James et al. (2013)

Modells mit KQ-Schätzer als relativ unflexibel, aber sehr einfach zu interpretieren. Der Lasso-Ansatz – ein Variablenselektionsverfahren – nimmt gleichfalls Rückgriff auf das lineare Modell, nutzt aber einen anderen Prozess für die Schätzung der Koeffizienten $\beta_1, \beta_2, ..., \beta_p$. Dieser alternative Prozess ist restriktiver und setzt gewisse *Features* genau auf null, womit sie eliminiert werden. In diesem Sinne ist *Lasso* im Vergleich zur linearen Regression weniger flexibler, gleichzeitig erhöht dieser Prozess aber die Interpretierbarkeit, weil im finalen Modell nur eine Untergruppe der unabhängigen Variablen enthalten ist, d.h. diejenigen, die nicht null werden. Generalisierte Additive Modelle (GAMs) erweitern das lineare Modell, sodass es auch gewisse nicht lineare Beziehungen erlaubt. Damit geht einher, dass GAMs zwar flexibler als das lineare Modell sind, jedoch gleichzeitig weniger gut interpretierbar, weil die Beziehung zwischen unabhängiger und abhängiger Variable nun mithilfe einer Kurve modelliert wird. Daneben sind vollständig nicht lineare Modelle wie *Bagging, Boosting, Support-Vektor-Maschinen* und *künstliche neuronale Netze* flexible Ansätze, die gleichfalls sehr schwer zu interpretieren sind.

Für den Anwender von Datascience-Methoden bestehen ein Trade-off und gros-

se Vorteile mit nicht parametrischen Ansätzen: In vielen praktischen Fällen sind einzig Prognosen von Interesse, und die Interpretierbarkeit wird zweitrangig (James et al., 2013) – beispielsweise in der Wirtschaftsprüfung, wenn es darum geht, im Rahmen von *analytischen* Prüfungshandlungen Abschlussbestände vorherzusagen.

2.1.6 Überwachtes versus nicht überwachtes Lernen

Die meisten Probleme des maschinellen Lernens lassen sich in *überwachte* und *unüberwachte* kategorisieren (James et al., 2013). Das Beispiel der Approximation von Absatzmengen fällt in die Domäne des überwachten Lernens, zumal für jede Beobachtung der unabhängigen Variablen $x_i, i = 1, ..., n$ eine damit in Verbindung stehende abhängige Variable y_i vorliegt. Ziel ist es, ein Modell zu finden, das einen Zusammenhang zwischen abhängiger Variable und den unabhängigen Variablen aufzeigt, um damit Prognosen zu generieren und die Beziehung zwischen den unabhängigen Variablen und der abhängigen Variable zu verstehen, d.h. daraus die Inferenz abzuleiten. Viele klassische statistische Methoden, wie die lineare Regression oder die logistische Regression, sowie neuere Ansätze, wie die *Support-Vektor-Maschinen* und *Deep Neural Networks*, operieren in der Sphäre des überwachten Lernens.

Demgegenüber stehen die Ansätze des unüberwachten Lernens: Für die Beobachtungen mit dem Vektor x_i sind bei dieser Methode keine unabhängigen Variablen y_i vorhanden. Es kann entsprechend kein lineares oder parametrisches Modell geschätzt werden, weil keine abhängige Variable vorliegt.

Bei Problemen dieser Art verbleibt allein die Möglichkeit, Beziehungen zwischen den Beobachtungen untereinander zu identifizieren. Eine Methode, die in diesem Zusammenhang zum Einsatz kommt, ist das *Clustering* bzw. die *Clusteranalyse*. Das Ziel dieses Ansatzes ist es, bei gegebener Variable $x_i, ..., x_n$ festzustellen, ob sich eine Beobachtung einer relativ abgegrenzten Gruppe zuordnen lässt.

Eine Intuition für das Audit besteht in Buchungen, die im Laufe der Berichtsperiode anfallen und typischerweise verschiedene Merkmale aufweisen wie Buchungstage, Buchungsvermerke, Uhrzeiten oder Beträge. Es kann angenommen werden, dass es einen Unterschied zwischen regulären Buchungen sowie Buchun-

gen anlässlich des Abschlusses – sogenannte Abschlussbuchungen – gibt. Wenn für jede Buchung bekannt ist, ob es sich um eine reguläre oder Abschlussbuchung handelt, entspräche das Problem einem überwachten Lernproblem. Liegt dies jedoch nicht vor, liesse sich das Clustering anwenden. Sollten mehrere Cluster identifiziert werden, so könnten potenzielle Gruppen verschiedener Buchungen mit speziellen Eigenschaften identifiziert werden, wie die Zugehörigkeit zu Abschlussbuchungen.

2.1.7 Regressions- versus Klassifikationsprobleme

Die Menge der Probleme des überwachten Lernens lässt sich wiederum in zwei Anwendungsfälle unterteilen: Klassifikations- und Regressionsprobleme (Hastie et al., 2009).

Dabei können eingesetzte unabhängige Variablen bzw. *Features* entweder in quantitativer oder qualitativer Form vorliegen. Quantitative Variablen nehmen numerische Werte wie Preise, Absatzmengen, Aktienkurse u.a. an, während es bei qualitativen Variablen Werte in einer Klasse bzw. Kategorie K sein können. Beispiele sind das Geschlecht (männlich, weiblich, divers), Marke eines Produkts (Marke A, B oder C) oder Zugehörigkeit zur Menge der Abschlussbuchungen (Abschlussbuchung ja/nein).

Ein Regressionsproblem liegt vor, wenn eine quantitative Variable als abhängige Variable definiert wird. Demgegenüber stehen Probleme, bei denen die abhängige Variable qualitativer Natur ist – also ein Klassifikationsproblem vorliegt.

Die Auswahl des Ansatzes für maschinelles Lernen wird auf der Grundlage getroffen, welcher Anwendungsfall vorliegt: Bei quantitativen abhängigen Variablen kann ein lineares Regressionsmodell angewandt werden während bei qualitativen abhängigen Variablen der Einsatz eines logistischen Regressionsmodells möglich ist (Lantz, 2015).

Weiter weniger entscheidend ist die Frage, ob die unabhängigen Variablen quantitativer oder qualitativer Natur sind: Die meisten Ansätze des maschinellen Lernens machen keinen Unterschied, sofern die Variablen korrekt kodiert sind, beispielsweise als Dummy-Variablen (Lantz, 2015).

In Tabelle 1 sind die verschiedenen Methoden des maschinellen Lernens und ihre Anwendungsfälle aufgeführt.

ML-Methoden	**Anwendungsfall**
Überwachtes Lernen	
k-nächste Nachbarn	Klassifikation
naive Bayes	Klassifikation
Entscheidungsbäume	Klassifikation
Assoziationsanalyse	Klassifikation
Lineare Regressionen	Regressionen
Regressionsbäume	Regressionen
Neuronale Netze	Beides
Support-Vektor-Maschinen	Beides
Nicht überwachtes Lernen	
k-means Clustering	Clustering
DBSCAN-Clustering	Clustering

Tabelle 1: Übersicht über Methoden im maschinellen Lernen und die zugehörigen Anwendungsfälle nach Lantz (2015)

.

2.1.8 Evaluation der Modellgenauigkeit

Eine Messgrösse ist erforderlich, die darüber Auskunft gibt, wie gut eine Vorhersage den beobachteten Daten entspricht, um die Leistungsfähigkeit der Ansätze des maschinellen Lernens bei einem gegebenen Datensatz zu evaluieren (James et al., 2013). Mit anderen Worten braucht es die Quantifizierung des Masses, in welchem der prognostizierte Wert für eine Beobachtung ihrem wirklichen Wert nahe kommt.

Für die Analyse der Regressionsprobleme wird üblicherweise der mittlere quadrierte Fehler oder auch *Mean Square Error (MSE)* berechnet:

$$MSE = (1/n)\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

Es gilt, dass $\hat{f}(x_i)$ die Prognose für die i-te Beobachtung ist. Der *MSE* wird folglich klein sein, wenn die Prognosen nahe beim beobachteten Wert sind, und gross für den Fall, dass sie erheblich davon abweichen.

Der *MSE* kann auf Grundlage eines Datensatzes mit den Trainingsdaten berechnet werden und wird daher als *Trainings-MSE* bezeichnet. Wesentlich entscheidender für die Evaluation der Leistungsfähigkeit ist jedoch die Grösse der

durchschnittlich quadrierten Abweichungen für noch *nicht genutzten* Beobachtungen. Werden derartige Beobachtungen für die Berechnung des *MSE* benutzt, die vormals noch nicht als Trainingsdaten zur Schätzung des Modells verwendet wurden, wird von *Test-MSE* gesprochen. Spielt die Inferenz keine Rolle und stehen konkurrierende Modelle zur Auswahl, wird folglich das Modell gewählt, das den tiefsten *Test-MSE* aufweist.

Die Qualität der Prognosen bei Klassifikationsproblemen lässt sich nicht über einen *Test-MSE* errechnen, sondern wird im einfachsten Fall mithilfe der Fehlerrate (entweder als *Trainigsfehlerrate* oder als *Testfehlerrate*) bestimmt, die den Anteil der falsch klassifizierten Vorhersagen im Verhältnis zu allen Vorhersagen misst:

$$Fehlerrate = (1/n) \sum_{i=1}^{n} I(y_i \neq \hat{f}(x_i))$$

Eine detailliertere Aufschlüsselung wird mit einer Konfusionsmatrix erreicht (siehe Beispiel in Tabelle 2)

	Erwart. Fraud	Erwart. Nicht-Fraud
Tatsächlich Fraud	51	3 (*Typ-1 Fehler*)
Tatsächlich Nicht-Fraud	5 (*Typ-2 Fehler*)	61

Tabelle 2: Paradigmatische Konfusionsmatrix in Anlehnung an Hastie et al. (2009)

Die Matrix stellt den tatsächlichen Klassen (dolose Handlungen vorhanden bzw. nicht vorhanden) die prognostizierten Klassen gegenüber – damit lassen sich einfach *Typ-1*- und *Typ-2 Fehler* ablesen.

2.1.9 Kreuzvalidierung

Bereits oben wurden die *Testfehlerrate* wie auch die *Trainingsfehlerrate* dargestellt. Unter der *Testfehlerrate* wird hierbei der durchschnittliche Fehler verstanden, welcher aus der Anwendung von maschinellen Lernmethoden zur Prognose von neuen Beobachtungen genutzt wird – also Stichproben, welche zuvor nicht für das Training des Modells benutzt wurden.

Fehlt ein solches dediziertes Testset, können Techniken angewandt werden, welche es erlauben, den Testfehler aus den Trainingsdaten zu berechnen. Dazu zählen die folgenden drei Methoden (James et al., 2013):

- *Validierungsset Ansatz*

  Für den Fall, dass der Testfehler für ein spezifisches Modell geschätzt werden soll, kann im einfachsten Fall auf einen Ansatz zurückgegriffen werden, wie er in Abbildung 3 dargestellt wird. Dieser Ansatz sieht vor, dass die Menge der vorhandenen Beobachtungen in zwei Gruppen unterteilt wird: ein Trainingsset und ein Validierungsset.

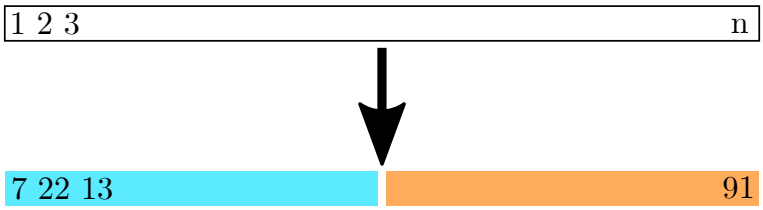


Abbildung 3: Validierungsset-Ansatz nach Friedman et al. (2001)

Das Modell kann anschliessend mit den Trainingsdaten trainiert und dazu benutzt werden, Vorhersagen für die Beobachtungen im Validierungsset zu treffen. Die daraus resultierende Fehlerrate des Validierungssets – oder im Falle von quantitativen abhängigen Variablen der MSE – bietet eine Schätzung der *Testfehlerrate*.

Der Validierungsset-Ansatz verfügt aufgrund seiner Einfachheit über zwei Nachteile:

1. Die geschätzte Fehlerrate kann äusserst variabel ausfallen und massgeblich davon abhängen, wie das Test- und Validierungsset voneinander getrennt werden.
2. Im *Trainingsset* wird nur ein Teil der Beobachtungen für die Schätzung des Modells herangezogen, wobei statistische Methoden die Eigenheit besitzen, bei einer kleinen Menge an Beobachtungen schlechter zu funktionieren.

- *Leave-one-out-Kreuzvalidierungsansatz*

  Der *Leave-one-out*-Kreuzvalidierungsansatz ($LOOCV$) ist dem Validierungsset-Ansatz zwar sehr ähnlich, versucht aber dessen Nachteile zu überwinden.

Gleich dem Validierungsset-Ansatz werden beim *Leave-one-out*-Kreuzvalidierungsansatz die Beobachtungen in zwei Gruppen eingeteilt. Der Unterschied besteht jedoch darin, dass stets nur eine einzige Beobachtung für die Validierung herangezogen wird, während die übrigen Beobachtungen der Schätzung des Modells dienen. Die Fehler, die für eine Beobachtung berechnet werden können, entsprechen einer unverzerrten Schätzung für den Testfehler, der als einzige Schätzung jedoch hoch variabel ist. Es ist daher üblich, dass der Vorgang n-mal wiederholt und der durchschnittliche *MSE* anschliessend errechnet wird. Figur 4 illustriert das Vorgehen grafisch:

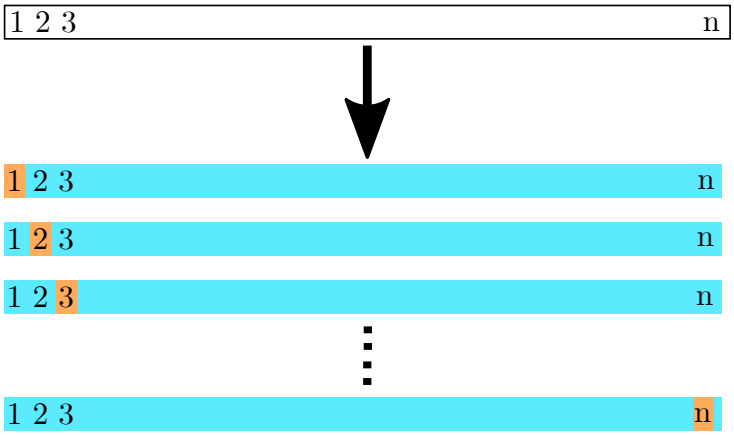


Abbildung 4: Schematische Darstellung des *Leave-one-out*-Kreuzvalidierungsansatzes nach Friedman et al. (2001)

Der LOOCV verfügt über wesentliche Vorteile gegenüber dem Validierungsset-Ansatz. Der Bias ist geringer, weil mit diesem Ansatz fast alle Beobachtungen herangezogen werden können. Im Gegensatz dazu wird beim Validierungsset-Ansatz üblicherweise lediglich die Hälfte der Beobachtungen für das Training des Modells genutzt. Entsprechend neigt der LOOCV-Ansatz weniger dazu, den Testfehler zu überschätzen. Zweitens unterscheidet sich der LOOCV-Ansatz vom Validierungsset-Ansatz dadurch, dass Ersterer auch bei einer wiederholten Anwendung stets das gleiche Resultat ergeben wird, zumal es keine Zufallskomponente – wie eine Aufteilung in Trainings- und Testset

– gibt.

- *k-Fold Kreuzvalidierung*

  Als Alternative zu LOOCV kann die *k-Fold-Kreuzvalidierung* eingesetzt werden. Dieser Ansatz sieht vor, dass die Beobachtungen zufällig in k-Gruppen bzw. *Folds* von approximativ gleicher Grösse aufgeteilt werden (vgl. Abbildung 5). Die erste Gruppe wird als Validierungsset genutzt, während die restlichen Folds für das Training des Modells verwendet werden. Der MSE wird anschliessend für das Validierungsset berechnet und verschiedene Male wiederholt, indem jedes Mal eine neue Gruppe als Validierungsset verwendet wird. Daraus resultieren verschiedene Testfehler, deren Durchschnitt berechnet werden kann. Es ist leicht zu erkennen, dass der LOOCV-Ansatz ein Spezialfall der k-Fold-Kreuzvalidierung ist. Der Rückgriff auf grössere Validierungssets hat hierbei den grossen Vorteil, dass die benötigte Rechenleistung massiv gesenkt werden kann.

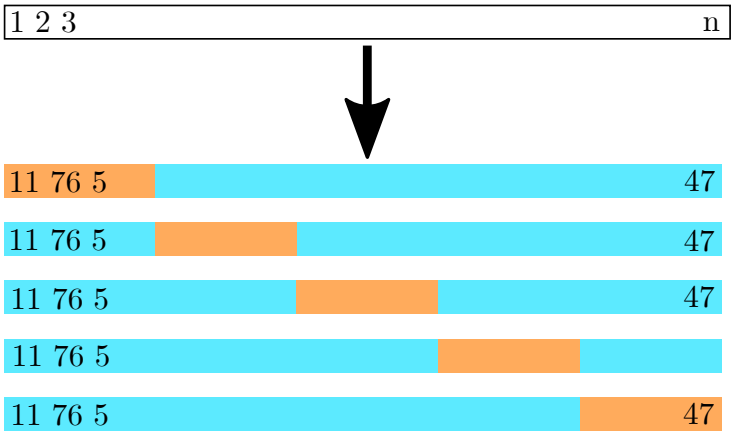


Abbildung 5: k-Fold Kreuzvalidierung nach Friedman et al. (2001)

2.1.10 Multiples lineares Modell

Funktionsweise: Das multiple lineare Regressionsmodell als Vertreter einer parametrischen Methode des maschinellen Lernens enthält, im Gegensatz zum einfa-

chen linearen Modell, für jede unabhängige Variable einen *Steigungsfaktor* (James et al., 2013):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon$$

wobei X_j die j-te unabhängige Variable und β_j die Beziehung von unabhängiger und abhängiger Variable ist.

Die Parameter werden über die Minimierung der Summe der quadrierten Abweichungen berechnet:

$$(RSS) = \sum_{i=1}^{n}(y_i - \hat{y_i})^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \quad (1)$$

Geschätzt werden können die Parameter wie folgt:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Vor- und Nachteile: Nach Lantz (2015) handelt es sich beim linearen Regressionsmodell um einen verbreiteten Ansatz zur Modellierung von numerischen Daten. Es lässt sich so adaptieren, dass es fast alle Modellierungsaufgaben bewältigen kann und sowohl die Grösse als auch die Beziehung, in der die unabhängigen mit der abhängigen Variablen stehen, zeigt. In seiner Anwendung macht das Modell in Bezug auf die Daten plausible Annahmen und muss bereits vorab durch den Benutzer spezifiziert werden, während es gleichzeitig keinen Umgang für fehlende Daten kennt. Darüber hinaus funktioniert es nur mit numerischen Variablen. Für die Nutzung kategorischer Variablen ist ein zusätzlicher Schritt, die Bildung von *Dummy-Variablen*, erforderlich. Abschliessend wird beim linearen Regressionsmodell statistisches Wissen benötigt, um ein resultierendes Modell bewerten zu können.

2.1.11 Entscheidungsbäume

Funktionsweise: Die Modellierung und Verwendung von Entscheidungsbäumen (*Decision Trees*) gilt für Klassifikationsprobleme bei gegebenen unabhängigen Variablen als sehr mächtig (Lantz, 2015). Entscheidungsbäume bedienen sich der

Heuristik der *rekursiven Partitionierung*. Dieser Ansatz, der auch als *Divide et Impera* bezeichnet wird, teilt die Daten in Untergruppen, die anschliessend in weitere Untergruppen gegliedert werden. Der Prozess ist erst abgeschlossen, wenn die resultierenden Gruppen homogen genug sind oder ein anderes Abbruchkriterium erreicht wird.

Der Prozess sieht vor, dass ausgehend von allen Beobachtungen eine unabhängige Variable gewählt wird, mit welcher der Datensatz geteilt wird, woraus neue Äste entstehen. Durchgehend durch alle Äste teilt der Algorithmus die Daten und wählt jedes Mal die beste Variable, bis entweder alle Beobachtungen an einem Ast die gleiche Klasse haben oder keine unabhängigen Variablen mehr vorhanden sind bzw. der Baum eine vordefinierte Grösse erreicht hat.

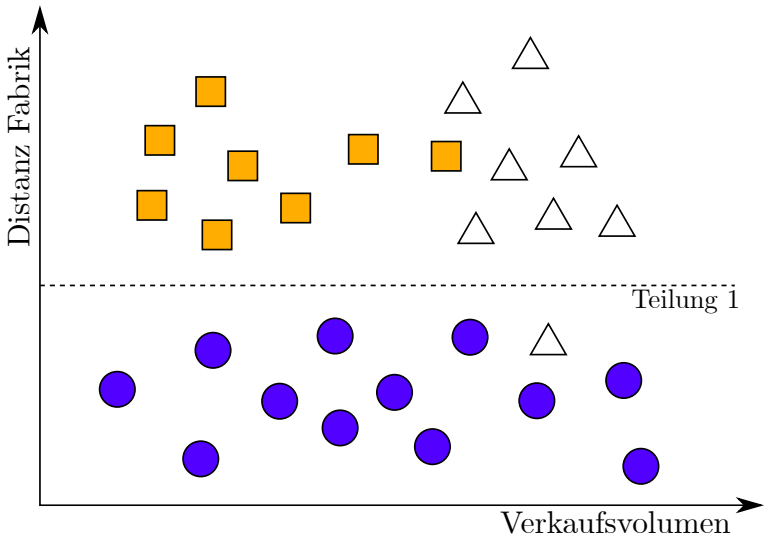


Abbildung 6: Erste Aufteilung im Prozess zur Modellierung eines Entscheidungsbaums (Lantz, 2015)

Abbildung 6 skizziert diesen Prozess: Für drei Klassen von Verkaufstransaktionen, die eine Unternehmung kennt und mit denen sie unterschiedliche Umsätze auf dem Markt erzielt, erkennt der Algorithmus in einem ersten Schritt, dass eine Trennung in Produkte, die eine Nähe zur Fabrik aufweisen, und solche, die weiter von ihrer Produktionsstätte abgesetzt werden, die homogensten Gruppen entstehen lässt.

In einem zweiten Schritt (vgl. Abbildung 7) trennt der Algorithmus die Da-

ten entlang der Verkaufsvolumina, die die Transaktionen mit grosser räumlicher Distanz zur Fabrik erneut unterteilen.

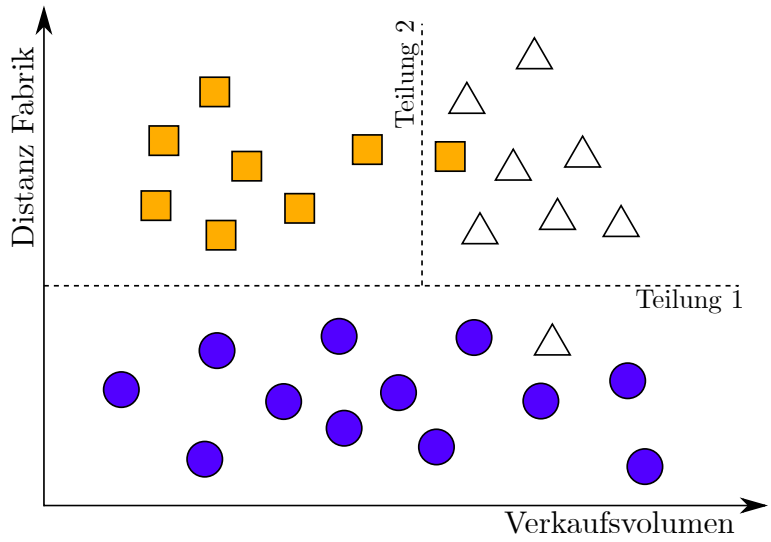


Abbildung 7: Zweite Aufteilung im Prozess zur Modellierung eines Entscheidungsbaums (Lantz, 2015)

Für die Entscheidung, welche unabhängige Variable für die Unterteilung heranzuziehen ist, wird häufig auf die Entropie abgestellt.

$$Entropy(S) = \sum_{i}^{c} = 1 - p_i log$$

Für jede Spaltung errechnet der Algorithmus die veränderte Homogenität, die durch eine Auftrennung für jede unabhängige Variable entstünde, wobei die Veränderung der Homogenität auch als Informationsgewinn (*Information Gain*) definiert ist.

Vor- und Nachteile: Es handelt sich bei den Entscheidungsbäumen um einen allseits einsetzbaren Klassifizierer, der sich für die meisten Probleme als funktionsfähig erwiesen hat, hoch automatisiert ist und sowohl für numerische als auch für kategoriale Variablen benutzt werden kann (Kelleher & Tierney, 2018), während er gleichzeitig einen Umgang mit fehlenden Daten kennt. Er lässt sich dabei sowohl auf kleine wie auch grosse Datensätze anwenden und resultiert in einem Modell,

das auf einfache Weise – auch ohne mathematisches Hintergrundwissen – interpretierbar ist. Die Berechnung von Entscheidungsbäumen läuft zudem effizienter als bei komplexeren Ansätzen ab.

Als Nachteil muss angeführt werden, dass Entscheidungsbäume öfters eine hohe Anzahl von Auftrennungen verursachen und in Bezug auf die Daten die Tendenz zum Over- bzw. Underfitting besteht. Zudem können gewisse Beziehungen nur unzureichend abgebildet werden, weil der Algorithmus auf die Aufspaltung parallel zu den Achsen angewiesen ist und eine kleine Veränderung in den Trainingsdaten erhebliche Auswirkung auf den Baum hat. Letztlich führen grosse Entscheidungsbäume zu Schwierigkeiten bei ihrer Interpretation und zu wenig intuitiven Resultaten.

2.1.12 Support-Vektor-Maschine (SVM)

Funktionsweise: Eine Support-Vektor-Maschine (SVM) kann als Oberfläche dargestellt werden, die eine Grenze zwischen Trainingsdaten als Punkte im multidimensionalen Raum aufstellt. Das Ziel der SVM ist es, eine flache Ebene bzw. eine *Hyperebene* einzuziehen, die diesen Raum in homogene Zonen aufteilt (Lantz, 2015).

Obschon die Support-Vektor-Maschine bereits vor einiger Zeit entwickelt wurde, hat ihre Popularität insbesondere in den letzten Jahren erheblich zugenommen, weil ihre Leistungsfähigkeit anderen Ansätzen lange Zeit – beispielsweise bei der optischen Erkennung von Postleitzahlen – weit überlegen war und sie gleichfalls in mehreren Statistikpaketen implementiert wurde.

Support-Vektor-Maschinen können für Regressions- wie auch Klassifikationsprobleme eingesetzt werden. Im Klassifikationsfall wird, wie in Abbildung 8 dargestellt, eine Ebene zwischen den zwei Klassen von Objekten (Kreise und Quadrate) eingezogen. Dabei gibt es, wie aus Abbildung 8 zu entnehmen ist, mehrere Möglichkeiten der Datentrennung. Die bestmögliche Trennung, die die Daten korrekt klassifiziert, besteht in der Ebene, die die grösste Separation zwischen beiden Klassen ermöglicht und *Maximum Margin Hyperplane* (MMH) genannt wird (vgl. Abbildung 9). Die Ortsvektoren eines jeden Punktes, die der MMH am nächsten sind, werden als *Support-Vektoren* bezeichnet und verorten die MMH eindeutig.

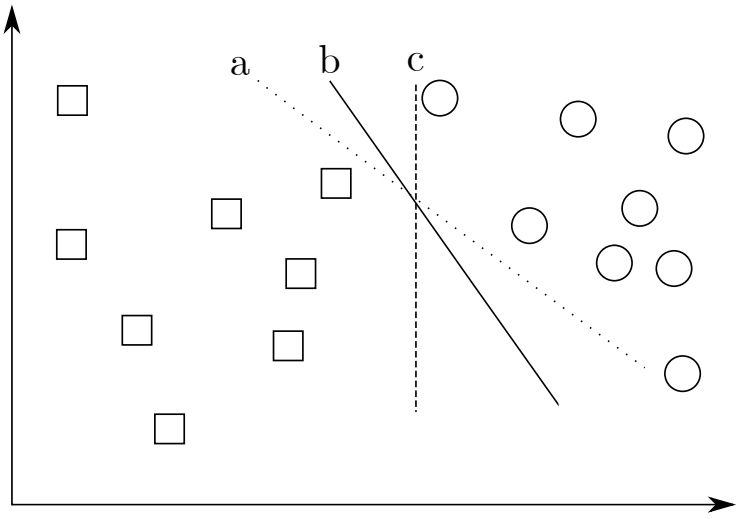


Abbildung 8: Mögliche Hyperebene im Klassifikationsfall einer Support-Vektor-Maschine (Lantz, 2015)

Für den Fall, dass die Daten nicht linear mit einer Hyperebene zu separieren sind, wie Abbildung 10 zeigt, wird ein *Kosten-Faktor* eingeführt, der die Fehlallokation aufnimmt und im Rahmen der Optimierung der Hyperebene minimiert wird.

Bei vielen Problemen sind die Beziehungen zwischen den Variablen nicht linear bzw. sind diese nicht linear zu trennen. Ein Beispiel zeigt Abbildung 11.

Mittels des sogenannten *Kernel-Tricks* – einer Funktion – können Trainingsdaten in einen höherdimensionalen Raum projiziert werden, wo sie in der Folge wieder linear separierbar werden (Friedman et al., 2001).

Vor- und Nachteile: Die Methode lässt sich sowohl bei Klassifikations- als auch bei Regressionsproblemen nutzen (Lantz, 2015). Der Einfluss von Störungen auf die Daten ist gering, wobei die Methode wenig anfällig für ein *Overfitting* ist. Daneben ist die Methode einfacher im Einsatz als neuronale Netzwerke. Das Training erfordert jedoch das Testen von mehreren Kombinationen aus Kernel und Modellparameter, was lange dauern kann, insbesondere wenn der Datensatz sehr umfangreich ist. Das Ergebnis ist ein Blackbox-Modell, das sich einer Interpretation entzieht.

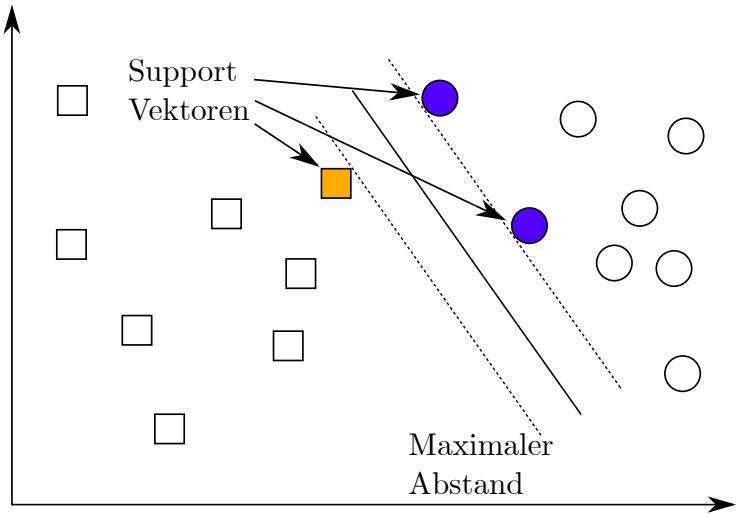


Abbildung 9: Einzug der Maximum Margin Hyperplane zur optimalen Separierung der Klassen (Lantz, 2015)

2.1.13 Künstliche neuronale Netze

Funktionsweise: Ein künstliches neuronales Netz modelliert die Beziehung von Eingangs- und Ausgangssignalen (Lantz, 2015). Das Modell basiert auf der Funktionsweise eines biologischen Gehirns und der Verarbeitung von Stimuli. Anwendung finden künstliche neuronale Netze bei der Lösung von Klassifikations- und Regressionsproblemen.

Die Funktionsweise eines einzelnen Neurons beruht auf einem Eingangssignal (x-Variable), das mit Gewichten multipliziert (w-Werte) wird. Aufsummiert werden diese Werte anschliessend mit einer Aktivierungsfunktion f verrechnet. Der sich ergebende Ausgangswert y dient in einem neuronalen Netz potenziell als Eingang (beispielsweise x_1) eines neuen Neurons (vgl. dazu Abbildung 12).

$$y(x) = f(\sum_{i=1}^{n} w_i x_i)$$

Künstliche neuronale Netze nutzen Neuronen, welche in dieser Art definiert wurden, zur Konstruktion komplexer Modelle. Obwohl verschiedenste Varianten von neuronalen Netzen entwickelt wurden, lassen sich diese durch drei Charakte-

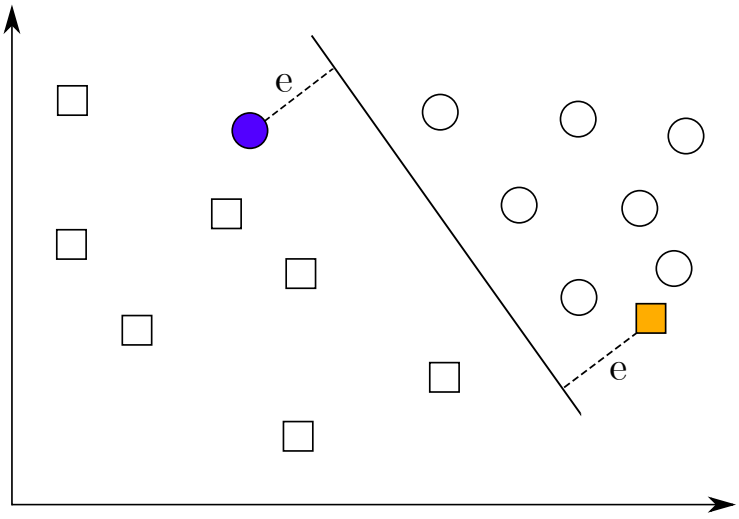


Abbildung 10: MMH mit unvollständiger Trennung der Klassen (Lantz, 2015)

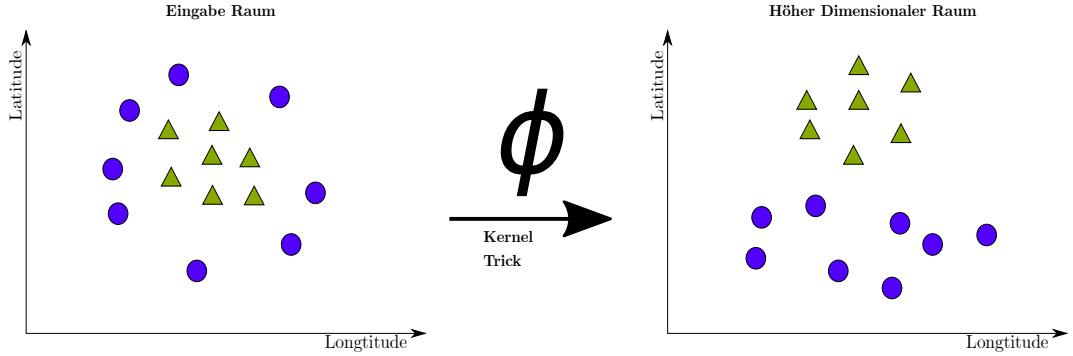


Abbildung 11: Illustration des Kerneltricks nach Lantz (2015)

ristiken definieren (Lantz, 2015):

- eine spezifische *Aktivierungsfunktion*, die die Eingangssignale in ein Ausgangssignal überführt und im neuronalen Netz weitergibt
- eine *Netzwerktopologie* (Architektur), welche die Anzahl der Neuronen und ihre Anordnung auf verschiedenen Ebenen wie auch ihre Verbindungen untereinander bestimmt
- ein *Trainingsalgorithmus*, der verwendet wird, um die optimalen Gewichte zu ermitteln.

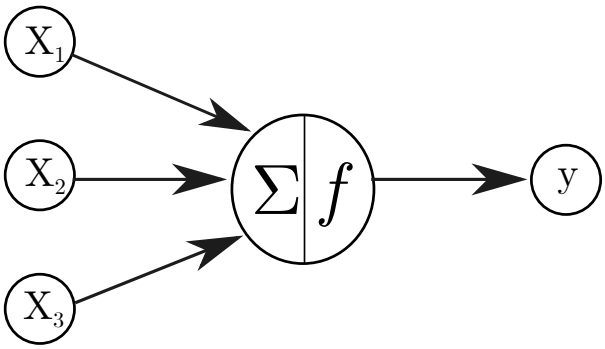


Abbildung 12: Funktionsweise eines Neurons nach Lantz (2015)

Unter den möglichen Aktivierungsfunktionen ist die Sigmoide hervorzuheben, die in Abbildung 13 dargestellt ist.

$$f(x) = \frac{1}{1 + e^{-x}}$$

Die Sigmoide zeichnet sich durch eine einfache Ableitung aus (Friedman et al., 2001). Ihr Wertebereich liegt zwischen null und eins, sodass auch grosse Eingangswerte in diesen Bereich komprimiert werden können; dies hat eine Häufung der Werte an den Enden zur Folge. Diesem Umstand wird dadurch Rechnung getragen, dass die Eingangswerte standardisiert bzw. normalisiert werden. Daneben hat sich die RELU-Aktivierungsfunktion etabliert.

Die Möglichkeit zu lernen, liegt in der Netzwerktopologie eines neuronalen Netzwerks begründet (Kelleher & Tierney, 2018). Diese wird von der Anzahl der benutzten Ebenen, der Anzahl der Neuronen pro Ebene sowie von den Typen der erlaubten Verbindungen bestimmt. Abbildung 14 zeigt ein beispielhaftes neuronales Netzwerk. Es besteht aus drei Eingangsneuronen, zwei Neuronen in einer verdeckten Schicht und einem Ausgangsneuron. Sämtliche Neuronen haben eine Verbindung zu allen vor ihnen liegenden Schichten. Es wird darum von einem vollständig verbundenen, vorwärts gerichteten neuronalen Netz gesprochen. Kein Neuron ist mit einer vorhergehenden Ebene verbunden, was einen Informations-

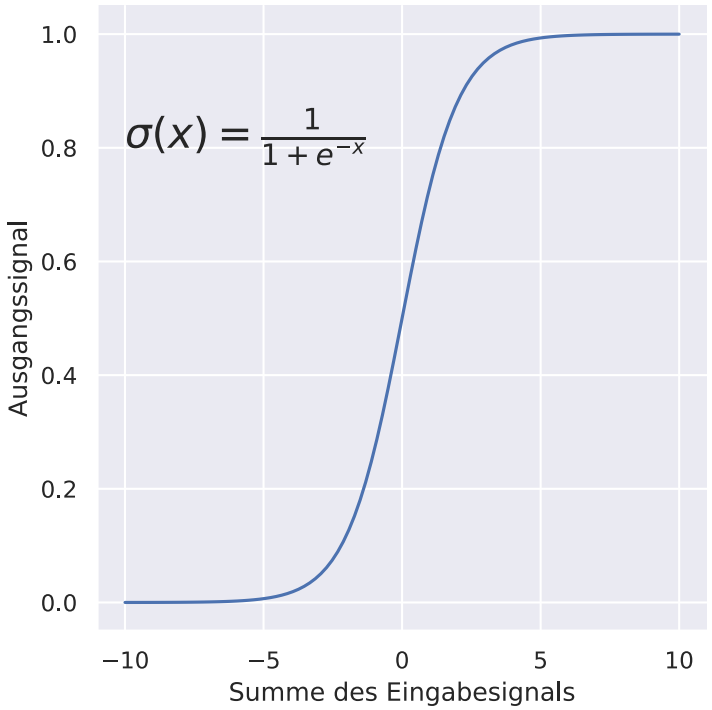


Abbildung 13: Plot der Sigmoidenfunktion (eigene Darstellung)

fluss vom Eingang hin zum Ausgang impliziert.

Für die Wahl der Anzahl der Neuronen auf den verdeckten Ebenen liegen keine gesicherten Heuristiken vor (Lantz, 2015). Die angemessene Zahl dieser Neuronen ist von der Anzahl der Eingangsneuronen, der Anzahl der Trainingsdaten, der Komplexität der Problemstellung und von weiteren Faktoren abhängig, während die Anzahl der Eingangsneuronen den Features bzw. unabhängigen Variablen entspricht. Generell gilt, dass ein grösseres Netz Prognosen für komplexere Problemstellungen bewältigen kann. Demgegenüber bedeutet eine höhere Anzahl von Neuronen aber auch, dass das Eingangssignal sehr genau abgebildet werden kann, unter diesen Bedingungen gleichzeitig jedoch die Generalisierbarkeit potenziell leidet, während sich das Training derartiger Netze als sehr rechenintensiv herausgestellt hat. Als *Best Practice* erweist sich der Grundsatz, so viele Neuronen wie nötig und so wenige wie möglich zu verwenden.

Das Training von neuronalen Netzen sowie die Optimierung ihrer Gewichtun-

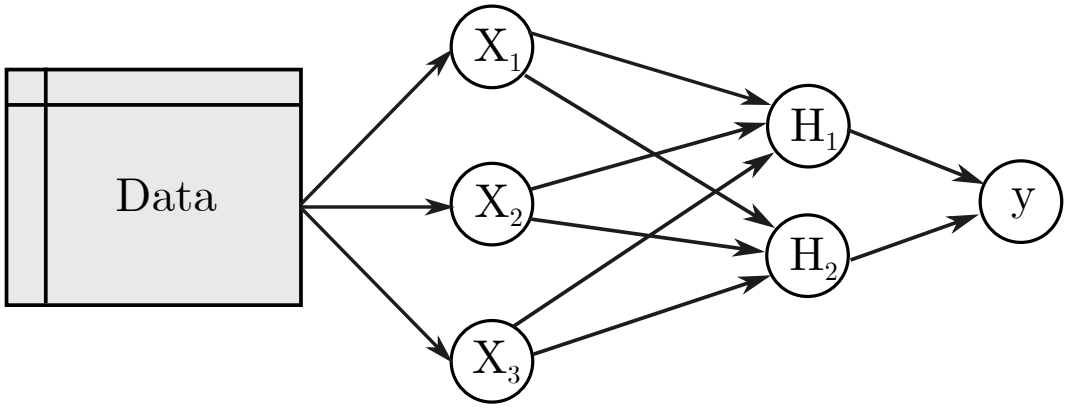


Abbildung 14: Beispiel eines neuronalen Netzes (Lantz, 2015)

gen sind äusserst rechenintensiv und werden mithilfe eines *backprogation-of-error*-Ansatzes durchgeführt (Abbasi et al., 2012). Dieser zweistufige Ansatz sieht im ersten Schritt vor, die Neuronen nacheinander zu aktivieren und die Gewichtungen auf die Inputdaten anzuwenden. Das Resultat ist ein Outputsignal, das mit dem wahren Wert der Beobachtung verglichen wird. Die Differenz lässt sich danach als Fehler zurück ins Netz propagieren und die Gewichtungen können entsprechend angepasst werden.

Über die Zeit werden die angepassten Gewichte den Fehler inkrementell verringern, was mithilfe eines Gradientenabstiegs erreicht wird, der die optimalen Anpassungen an den Gewichten vorschlägt (Goodfellow et al., 2016).

Vor- und Nachteile: Neuronale Netze kommen gleichfalls bei Klassifikations- als auch bei Regressionsproblemen zur Anwendung. Dabei können neuronale Netze viel komplexere Muster als nahezu alle konkurrierenden Methoden abbilden, während die Annahmen für die darunterliegenden Daten äusserst gering sind. Demgegenüber steht die enorme Rechenleistung, die für das Training der Netze notwendig ist. Zudem neigen neuronale Netze dazu, die Daten zu *overfitten* und resultieren in einem *Blackbox-Modell*, das kaum oder gar nicht zu interpretieren ist (Lantz, 2015).

2.1.14 Assoziationsanalyse

Funktionsweise: Die Assoziationsanalyse hat sich aus der Untersuchung von Kaufmustern entwickelt, erwies sich aber auch in anderen Kontexten als überaus nützlich (Kelleher & Tierney, 2018).

Sie stellt auf Elemente – wie beispielsweise Brot oder Erdnussbutter – ab, die zusammen gekauft werden. Zusammenhängende Gruppen von Elementen werden für die Analyse in einer Elementgruppe zusammengeführt und erscheinen üblicherweise als solche mit einer Regelmässigkeit in den Daten.

Transaktionen werden als eine Elementgruppe spezifiziert. Folgendes Beispiel könnte für einen Supermarkt gelten:

$$\{Brot, Ernussbutter, Konfitüre\}$$

Das Resultat der Assoziationsanalyse ist eine Sammlung von Assoziationsregeln, die Beziehungen zwischen Elementen für die Elementgruppen wiedergeben. Dabei bestehen sämtliche Assoziationsregeln aus einer Untermenge an Elementgruppen und werden mit einer linkshändigen Elementgruppe dargestellt, die in Verbindung mit einer rechtshändigen Elementgruppe steht. Dabei gilt nach der Logik der Assoziationsanalyse, dass wenn die Regel der linken Hand erfüllt ist, die Elementgruppe rechts erwartet wird. Beispielsweise könnte gelten:

$$\{Ernussbutter, Konfitüre\} \rightarrow \{Brot\}$$

Nicht formell ausgedrückt besagt die Regel, dass für Brot gleichfalls eine hohe Wahrscheinlichkeit besteht, mit gekauft zu werden, wenn Erdnussbutter und Konfitüre zusammen gekauft werden.

Assoziationsregel-Algorithmen sind unbeaufsichtigt und müssen daher nicht trainiert werden, womit das *Labeling* der Daten entfällt (Lantz, 2015). Der Algorithmus wird vielmehr direkt auf die Daten angewandt, beispielsweise in der Hoffnung, verkaufsfördernde Zusammenhänge aufzudecken.

Trotz ihres Ursprungs in der Warenkorbanalyse und in Empfehlungssystemen findet der Ansatz weitere Anwendungsfälle (Lantz, 2015):

- Aufdeckung von Käufen, die mit illegaler Kreditkartennutzung oder Versi-

cherungsbetrügen verbunden sind

- identifizierte Kombinationen des Kundenverhaltens, dem eine Kündigung oder ein Upgrade der Festnetz- oder Kabel-TV-Verträge vorausgeht

Insoweit decken die Assoziationsregeln bedeutende Verbindungen zwischen einer umfangreichen Masse von Elementen auf. Während Menschen in dieser Hinsicht sehr intuitiv Einsichten generieren, ist Expertenwissen oder ein grosser Erfahrungsschatz notwendig, um Einsichten zu gewinnen, die eine Assoziationsanalyse auf Knopfdruck bereitstellen. Daneben können die Algorithmen Datenmengen bearbeiten, die gänzlich zu umfangreich oder zu komplex sind und der Suche nach der Nadel im Heuhaufen gleichen.

Um der hohen Anzahl von möglichen Kombinationen zu begegnen, die sich beispielsweise bereits bei hundert verschiedenen Artikeln einstellen, hat Agrawal (1994) eine Heuristik *a priori* für eine Reduktion vorgeschlagen: Jede Untergruppe von öfters auftauchenden Elementgruppen muss auch häufig erscheinen und damit relevant sein. Für das oben angeführte Beispiel zu Lebensmitteln bedeutet diese Erkenntnis, dass die Gruppe $\{Butter, Brot\}$ nur dann häufiger auftaucht, wenn auch $\{Butter\}$ und $\{Brot\}$ häufiger auftreten. Entsprechend gilt umgekehrt, dass Elementgruppen, die diese Elemente enthalten, ausgeschlossen werden können, wenn entweder Butter oder Brot nur selten auftreten.

Ein Mass für das Interesse ist der *Support*, der angibt, wie oft eine Elementgruppe im Datensatz vorkommt:

$$support(X) = \frac{count(X)}{N}$$

Daneben wird auch die *Konfidenz* herangezogen, die ein Mass für die Prognosegüte darstellt bzw. die Genauigkeit abbildet. Sie ist definiert als der Support einer Elementgruppe, die sowohl X als auch Y enthält, geteilt durch den Support der Elementgruppe, die lediglich X enthält:

$$confidence(X \rightarrow Y) = \frac{support(X, Y)}{support(X)}$$

Nach Massgabe eines minimalen *Supportniveaus* identifiziert der Algorithmus die Elementgruppen, die von Interesse sind, und leitet daraus Regeln für diejenigen

ab, die eine minimale Schwelle für die Konfidenz erreichen (Tackett, 2013).

Vor- und Nachteile: Nach Lantz (2015) liegen die Vorteile des Ansatzes darin begründet, dass mit ihm eine grosse Anzahl von Transaktionen bearbeitet werden kann. Gleichzeitig sind die Resultate, die der Ansatz liefert, sehr einfach zu interpretieren, während er äusserst nützlich im Datamining ist, insbesondere für die Aufdeckung von neuem Wissen. Demgegenüber ist als Nachteil anzuführen, dass der Ansatz nur schlecht mit kleinen Datenmengen umgehen kann. Gleichwohl erfordert er eine kognitive Leistung des Anwenders, um tatsächliche Einsichten in den Datensatz zu erhalten. Diesbezüglich besteht zuweilen die Gefahr, aus zufälligen Mustern keine verwertbaren Einsichten zu erhalten.

2.1.15 Clustering mit k-Means

Funktionsweise: Buchhaltungsdaten und insbesondere die Buchung von Geschäftsvorfällen weisen wiederkehrende Muster aus. So können Teile der Einträge, die anhand des Datums identifiziert werden, über das Jahr verteilt sein, während ein anderer Teil insbesondere an Tagen des Buchhaltungsabschlusses auftritt (z.B. monatlich, quartalsweise oder jährlich) – also Abschlussbuchungen darstellt. Je nach Geschäftsmodell können diese Abschlussbuchungen auch weitere Spezifika aufweisen, wie einen überdurchschnittlich hohen Betrag oder spezialisierte Buchhalter, die beispielsweise Buchungen für das Anlagevermögen durchführen.

Clustering ist nicht sehr viel mehr als die Aufdeckung dieser Muster in Buchungen bzw. Datensätzen. Die Methode wird als nicht überwachter Ansatz des maschinellen Lernens bezeichnet, der automatisch Daten in Gruppen oder eben *Cluster* einteilt, die sich ähnlich sind, ohne jedoch vorher zu wissen, wie die Gruppen auszusehen haben. Da vorab kein Wissen zu dem *Cluster* vorliegt, handelt es sich bei dieser Methode mehr um einen Ansatz zur Aufdeckung von Zusammenhängen als um Prognosen (Kelleher & Tierney, 2018).

Clustering weicht in seinem Ansatz erheblich von anderen Ansätzen ab. Für jede der bereits besprochenen Methoden – mit Ausnahme der Assoziationsanalyse – war das Resultat ein Modell, das die Daten, d.h. unabhängige Variablen mit einer abhängigen Variablen verbunden hat. Konzeptuell hat das Modell daher existierende Muster abgebildet. Im Gegensatz dazu führt die Anwendung eines *Clusterings*

zu neuen Daten. Beobachtungen ohne eine vorliegende abhängige Variable werden verwertet und einem *Cluster* zugeordnet, wobei sich diese Zuordnung gänzlich aus den Beziehungen der Beobachtungen untereinander ergibt.

Die Herausforderung beim *Clustering* ist dem Umstand geschuldet, dass die so entwickelten Gruppen ohne inhärente Bedeutung sind. Es obliegt dem Spezialisten, eine Typeneinteilung vorzunehmen (Lantz, 2015), wenn er wie beschrieben in den Buchungen eine grosse Anzahl betraglich grösserer Buchungen zum Abschluss des Jahres neben regelmässig kleinen Buchungen während des Jahres feststellt.

Als sehr verbreiteter Algorithmus für das *Clustering* hat sich *k-Means* etabliert. Er bildet auch die Basis für zahlreiche weiterführende Algorithmen (Lantz, 2015).

Sämtliche Beobachtungen werden in einem multidimensionalen Raum verortet, um die Cluster aus den Daten abzuleiten (Hastie et al., 2009). Im Beispiel sollen Buchungen in einem zweidimensionalen Raum mit einer Achse für den Tag im Jahr und mit einer anderen Achse die Beträge einem Cluster zugeordnet werden, die mit der Buchung im Zusammenhang stehen.

Sollen drei Cluster aufgedeckt werden, so beginnt der Algorithmus, indem er k Beobachtungen – k ist die Anzahl der Cluster – zufällig auswählt, die als initiale Clustermittelpunkte dienen. Nachdem die Cluster-Mittelpunkte ausgewählt wurden, werden sämtliche Beobachtungen dem Clustermittelpunkt zugeordnet, der am nächsten liegt, wobei die Distanz häufig als euklidische Distanz verstanden wird. Es könnte jedoch auch die Manhattan- oder Minkowski-Distanz verwendet werden. Damit werden sämtliche Beobachtungen initial einer Gruppe zugeordnet (vgl. Abbildung 15).

Dieser initialen Zuordnung schliesst sich eine Aktualisierungsphase an: In einem ersten Schritt wird der Mittelpunkt der gebildeten Cluster berechnet. Es folgt eine neue Zuordnung der Beobachtungen zum nächstgelegenen Clustermittelpunkt (siehe Abbildung 16).

Diese Aktualisierungsphase wird mehrere Male wiederholt, bis eine Abbruchbedingung gegeben ist, wenn beispielsweise genügend Durchläufe getätigt wurden oder es keine Verschiebungen mehr in den Clustern gibt (vgl. Abbildung 17). Abschliessend können die finalen Cluster auf zwei Arten ausgegeben werden: Entweder wird für jede Beobachtung eine entsprechende Clusterzuordnung mitgegeben oder die Koordinaten der finalen Clustermittelpunkte werden erfasst.

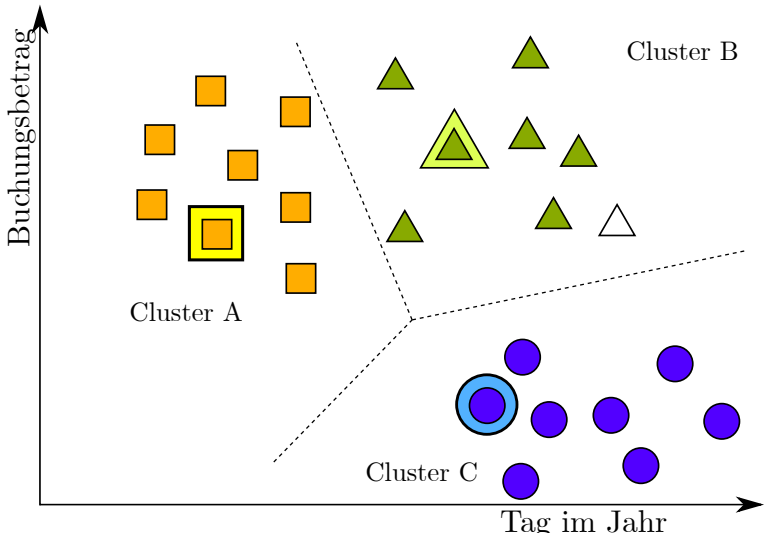


Abbildung 15: Initiale Zuordnung der Beobachtungen zu Clustermittelpunkten im k-Means Clustering

Vor- und Nachteile: Ein wesentlicher Vorteil des Algorithmus ist, dass er relativ einfach und ohne ein spezielles statistisches Wissen verständlich ist (Lantz, 2015). Zugleich ist er äusserst flexibel und lässt sich mit einfachen Mittel in einer Art und Weise anpassen, dass seine negativen Aspekte überkommen werden können. Nicht zuletzt funktioniert der Algorithmus sehr gut bei praktischen Herausforderungen. Demgegenüber als Nachteil erweist sich der Umstand, dass der Algorithmus nicht in der Form sophistiziert ist wie neuere Clustering-Ansätze. Das k-Means Clustering enthält daneben auch ein Element des Zufalls, das dazu führen kann, dass nicht immer die optimalen Cluster gefunden werden. *A priori* setzt der Algorithmus auch ein Wissen um die optimale Anzahl der Cluster voraus, die anhand der Daten erwartet werden. Ferner eignet sich der Ansatz nicht für nicht sphärische Cluster oder für Cluster mit einer ausgeprägten variierenden Dichte.

2.1.16 Maschinelles Lernen im Prozess

Das maschinelle Lernen kennt Standardprozesse, die dabei unterstützen, Herausforderungen des Geschäftsumfeldes in Probleme des maschinellen Lernens zu übersetzen, die angemessene Datentransformationen und methodische Ansätze empfeh-

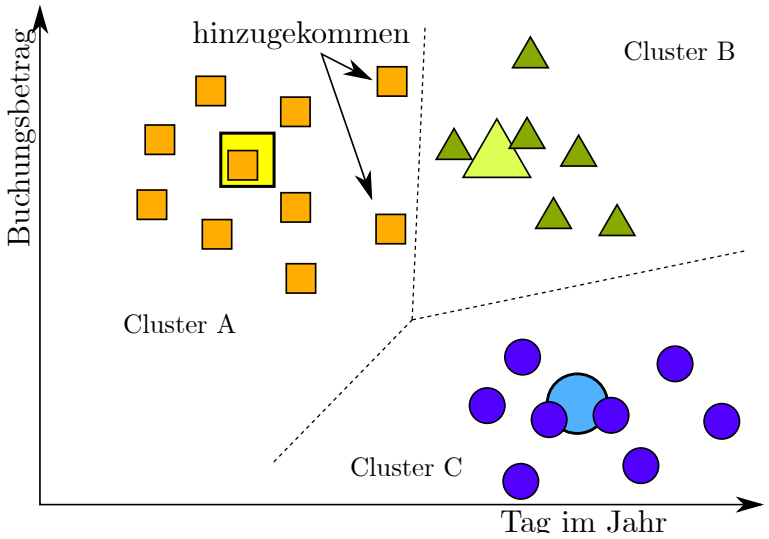


Abbildung 16: Erste Aktualisierungsphase im k-Means Clustering

len und Evaluationsverfahren zur Überprüfung der Effektivität der Modelle aufzeigen.

Dabei hat sich insbesondere der *Cross Industry Standard Process for Data Mining-Prozess (CRISP-DM)* etabliert, der ein Framework vorgibt, wie derartige Projekte unabhängig von Industrie oder Technologie bearbeitet werden können, und nach Kelleher und Tierney (2018) der am weitesten verbreitete ist. Das Prozessmodell zielt darauf ab, bei umfangreichen Projekten die Kosten zu reduzieren, die Verlässlichkeit zu steigern und die Reproduzierbarkeit sicherzustellen, die praktische Umsetzung zu erreichen und die Generierung von Modellen zu beschleunigen. Daneben kann das Modell als Referenzpunkt für Diskussionen hinzugezogen werden. Das Modell soll das Verständnis gängiger Datascience-Probleme für alle Teilnehmer, insbesondere aber für interne Diskussionen, verbessern. Ausserdem kann das Modell darlegen, dass es sich bei Datascience um eine strukturierte Ingenieursleistung handelt und auch unter Einsatz anderer Tools und anderer Anwender gleiche Abläufe stattfinden, sodass die Vergleichbarkeit erheblich verbessert wird.

Für Anwender bietet das Modell vielfache Vorteile: Für Anfänger stellt der Prozess geeignete Vorgaben zur Verfügung, unterstützt die Strukturierung der Projekte und leistet Hilfestellung für jede Teilaufgabe. Erfahrene Analysten können

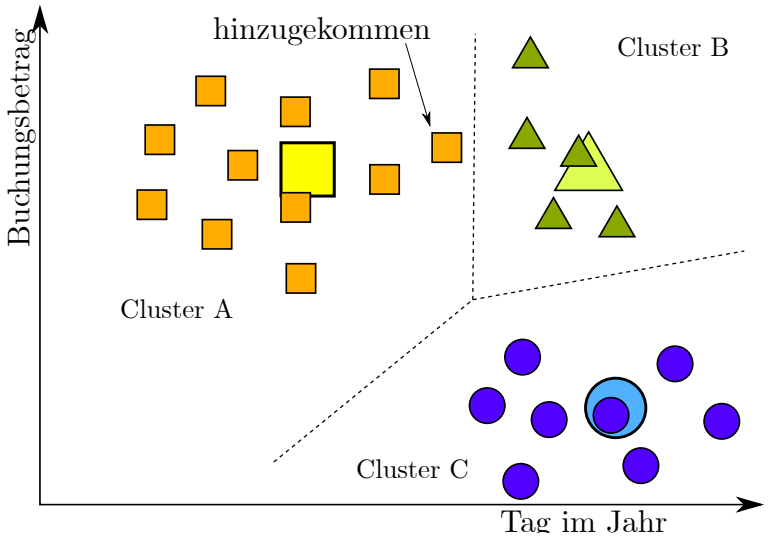


Abbildung 17: Abschluss des k-Means Clusterings

auf Checklisten zurückgreifen, sodass Wissen erhalten bleibt. Genauso wichtig ist, dass das Modell die Kommunikation und Dokumentation unterstützt. Es hilft, dass verschiedene Tools und unterschiedliche Personen mit einer breiten Palette an Fähigkeiten und Hintergründen effizient und effektiv an gemeinsamen Projekten arbeiten können (Wirth & Hipp, 2000).

Das CRISP-DM-Modell bezieht sich auf vorhergehende Versuche wissensgenerierender Methodologien (Reinartz und Wirth (1995), Brachman et al. (1996)).

Als Referenzmodell bietet CRISP-DM eine Übersicht über den Lebenszyklus eines Datenanalyse-Projektes. Es umfasst Projektphasen mit den dazugehörigen Aufgaben und ihren Resultaten (Wirth & Hipp, 2000).

Der Lebenszyklus wird in sechs Phasen unterteilt, die in Abbildung 18 dargestellt sind. Die Sequenz der Phasen muss jedoch nicht strikt einhalten werden. Die Pfeile indizieren jeweils nur die wichtigsten und häufigsten Abhängigkeiten zwischen den Phasen. Dabei geht es um die Frage, welche konkrete Aufgabe im nächsten Schritt zu erledigen ist.

Der äussere Kreis in Abbildung 18 weist auf die zyklische Natur der Analysen hin und deutet an, dass mit der Bereitstellung einer Lösung das maschinelle Lernen potenziell noch nicht abgeschlossen ist. Die Erfahrung aus dem Prozess und aus der Bereitstellung der Lösung kann weitere Fragen aufwerfen, die sich sodann

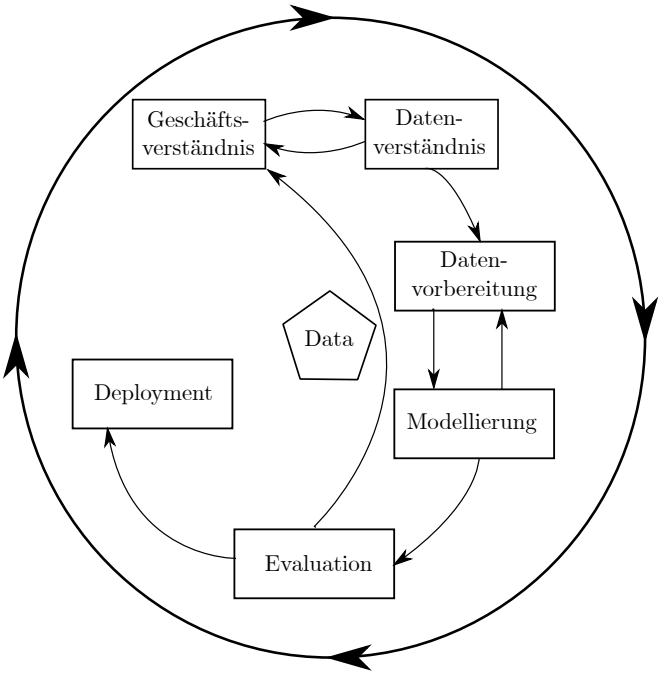


Abbildung 18: Phasen des CRISP-DM-Prozessmodells für die Datenanalyse nach Wirth und Hipp (2000)

beantworten lassen.

Folgende Aspekte sind für die einzelnen Phasen typisch (Wirth & Hipp, 2000):

1. *Erlangung des Geschäftsverständnisses*

   Am Anfang eines Projekts muss ein Verständnis für den Geschäftszweck aufgebaut werden, um auf dieser Grundlage die Problemstellung in eine für maschinelles Lernen geeignete Form übersetzen zu können. Daneben muss ein initialer Projektplan aufgestellt werden, der es ermöglicht, die Projektziele zu erreichen.

2. *Erlangung Datenveständnis*

   In einem weiteren Schritt werden die Daten initial geladen und Aktivitäten gestartet, um sich in den Daten zurechtzufinden, Qualitätsprobleme in den Daten zu identifizieren, gegebenenfalls erste Einsichten zu gewinnen und potenziell interessante Teilmengen zu erkennen, um daraus Hypothesen für verborgene Information abzuleiten. Das Erlangen eines Verständnisses der Da-

ten ist eng verbunden mit der Erlangung des Geschäftsverständnisses: Denn um die Problemstellung zu formulieren, ist ein grundlegendes Verständnis der Daten unablässig.

3. *Datenvorbereitung*

   Die Vorbereitung der Daten schliesst alle weiteren Schritte mit ein, die notwendig sind, um über einen finalen Datensatz für die Analyse zu verfügen. Dies umfasst die Selektion von Tabellen, Beobachtungen und Attributen, die Datenbereinigung, die Erstellung von neuen Attributen und die Transformation der Daten. Zudem müssen die Daten aus mehreren Quellen integriert werden.

4. *Modellierung der Daten*

   In dieser Phase werden Techniken des maschinellen Lernens ausgewählt und auf die Daten angewandt. Anschliessend werden die daraus resultierenden Modelle durch eine Anpassung der Parameter optimiert. Üblicherweise können verschiedene Techniken bzw. Ansätze auf die Daten angewandt werden, wobei gegebenenfalls spezifische Datenformate erforderlich sind. Es besteht eine Abhängigkeit zur Datenvorbereitung: So können auch in dieser Phase Datenprobleme identifiziert oder neue Ideen für zusätzliche Datenfelder entwickelt werden.

5. *Evaluation der Modelle*

   Zum Zeitpunkt der Evaluation sind bereits eines oder mehrere Modelle trainiert worden, die eine hohe Güte versprechen. Diese Modelle müssen final getestet und die Schritte, die zu seiner Konstruktion geführt haben, erneut geprüft werden. Weiterhin gilt es herauszufinden, ob alle Geschäftsziele damit abgedeckt werden können oder es in Bezug auf die Geschäftstätigkeit begründete Problemstellungen gibt, die bislang noch nicht adäquat berücksichtigt wurden. Am Ende dieser Phase wird die Entscheidung getroffen, ob die Ergebnisse des Modells angewendet werden sollen.

6. *Bereitstellung der Modelle*

Die Aufstellung eines Modells markiert selten den Abschluss des Prozesses. Normalerweise gilt es, das erarbeitete Wissen abschliessend zu organisieren und in einer Weise aufzubereiten, dass es internen Stakeholder vermittelbar ist. Abhängig von den Voraussetzungen kann die Bereitstellungsphase in einen simplen Bericht münden oder zu einer komplexen Lösung führen, die eine wiederholte Prognose ermöglicht.

Neben den vorgestellten sechs Phasen macht CRISP-DM im Referenzmodell weitergehende Ausführungen zu generischen und spezialisierten Aufgaben, die potenziell auszuführen sind (siehe dazu Abbildung 19). Damit legt das Modell grundsätzlich das *Was* dar, das innerhalb eines Projekts auszuführen ist. Das dazugehörige Anwenderhandbuch gibt weiterführende und detaillierte Tipps und Hinweise für jede Phase und jede Aufgabe einer Phase sowie eine Hilfestellung, wie ein Datascience-Projekt auszuführen ist.

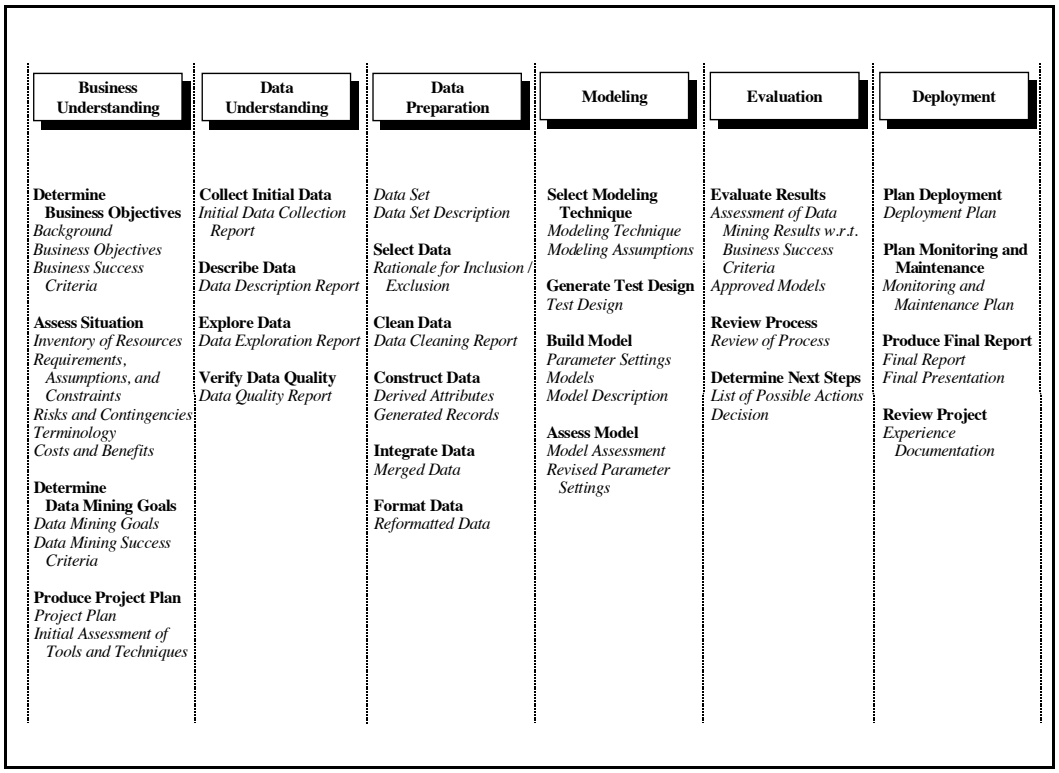


Abbildung 19: Phasen des CRISP-DM-Prozessmodells für das Data Mining

2.2 Wirtschaftsprüfungsprozess

2.2.1 Ziele und Vorgehensweise während der Abschlussprüfung

Für den Zweck einer Abschlussprüfung hält die *International Federation of Accountants (IFAC)* als regulierende Institution zur Einhaltung von Standards für die Wirtschaftsprüfung in ISA 200.3 fest:

> „Der Zweck einer Abschlussprüfung besteht darin, das Mass an Vertrauen der vorgesehenen Nutzer in den Abschluss zu erhöhen. Dies wird dadurch erreicht, dass der Abschlussprüfer ein Prüfurteil darüber abgibt, ob der Abschluss in allen wesentlichen Belangen in Übereinstimmung mit einem massgebenden Regelwerk der Rechnungslegung aufgestellt wurde.“ (IAASB, 2009a)

Das Prüfurteil, das den Abschluss und das Resultat der Prüfung markiert, wird üblicherweise in Form eines Berichts bzw. eines Testats abgegeben. In den einzelnen Rechtsgebieten können den Wirtschaftsprüfern im Rahmen ihrer Abschlussprüfung weitere Aufgaben übertragen werden, wie beispielsweise für die Prüfung des internen Kontrollsystems in den USA als Teil der Sektion 404 des *Sarbanes-Oxley Acts* oder in der Schweiz als Teil des Obligationenrechts (EXPERTsuisse, 2016).

Wirtschaftsprüfer sammeln nach Massgabe einschlägiger Vorgaben, namentlich ISA 500.6 (IAASB, 2009g), Prüfnachweise im Rahmen ihres Audits, um zu einer Einschätzung darüber zu gelangen, ob der Abschluss frei von wesentlichen Fehldarstellungen ist. Diese Prüfnachweise bestehen nach ISA 500.5c aus „Informationen sowohl in den Aufzeichnungen der Buchhaltung, die dem Abschluss zugrunde liegen, wie auch anderen Informationen“.

Fehldarstellungen werden nach ISA 320.2 (IAASB, 2009f) üblicherweise als wesentlich bezeichnet, wenn die Summe der nicht korrigierten Abweichungen in der Darstellung von Abschlüssen aufgrund von Fehlern oder deliktischen Handlungen dazu führte, dass sie die Entscheidung von vernünftigen Personen beeinflussen oder ändern kann.

Der Wirtschaftsprüfer stellt sein Urteil mit hinreichender Sicherheit aus. Seine Zusicherung ist ein Mass für das Sicherheitsniveau, das der Wirtschaftsprüfer bei

seiner Abschlussprüfung erreicht hat. Die Standards für die Abschlussprüfung sehen vor, dass dieser Grad hoch ist, jedoch nicht vollständig bzw. absolut. Dementsprechend kann eine Abschlussprüfung, die nach den Vorgaben der IFAC korrekt durchgeführt wurde, wesentliche Fehldarstellungen enthalten (IAASB, 2009a).

Die Auditstandards unterscheiden zwischen zwei Formen von Fehldarstellungen: Fehler können aufgrund von Irrtümern auftreten oder aber ihre Ursache in deliktischen Handlungen haben, die sich wiederum in Veruntreuung von Vermögensgegenständen der Unternehmung und eine deliktische Fehldarstellung des Managements („Management Fraud“) nach ISA 240.3 (IAASB, 2009c) unterscheiden.

Die Prüfungsstandards machen jedoch keinen Unterschied in der Verantwortung des Prüfers, diese Fehler aufzudecken: In beiden Fällen muss der Auditor genügend geeignete Prüfnachweise sammeln, um eine Einschätzung über die Absenz von wesentlichen Fehldarstellungen im Abschluss abgeben zu können. Der Standard weist jedoch darauf hin, dass es meistens weitaus schwieriger ist, dolose Handlungen aufzudecken, zumal das Management oder die Mitarbeiter, die die deliktischen Handlungen durchführen, versuchen werden, diese Verstösse zu verbergen – entsprechende Hinweise sind in ISA 200.5 (IAASB, 2009a) und ISA 240 (IAASB, 2009c) aufgeführt.

2.2.2 Rahmenbedingungen

In der Schweiz sind die Rechnungslegung und die Prüfung von Abschlüssen von Unternehmen im Obligationenrecht oder im Zivilgesetzbuch relativ knapp geregelt. Etwaige Konkretisierungen bleiben der Lehre und Praxis überlassen (EXPERTsuisse, 2016). Neben gesetzlichen Bestimmungen zur Rechnungslegung sind insbesondere die Internationale Rechnungslegungsstandards (IFRS) und das US-amerikanischen US-GAAP sowie nationale Rechnungslegungsstandards wie die Swiss GAAP FER relevant.

Über die im Gesetz festgehaltenen Anforderungen hinaus ist die Tätigkeit des Wirtschaftsprüfers durch vom Berufsstand erlassene Grundsätze auf nationaler und internationaler Ebene geregelt. Die Standes- und Berufsregeln machen Vorgaben zum Verhalten der Berufsangehörigen (EXPERTsuisse, 2016).

Auf fachlicher Seite bestehen in der Schweiz und auf internationaler Ebene

verbindliche Pflichten, die in der Schweiz von EXPERTsuisse erlassen und um weiter gehende Prüfungsgrundsätze für staatlich beaufsichtigte Revisionsunternehmen durch die Eidgenössische Revisionsaufsicht (RAB) ergänzt werden. Auf internationaler Ebene erlässt das International Audit and Assurance Standards Board (IAASB) Standards, die die EXPERTsuisse bei Beachtung der schweizerischen Gesetzesbestimmungen sowie anderer rechtlicher Anforderungen umsetzt und als Prüfungsstandards (PS) publiziert (EXPERTsuisse, 2013).

Die PS von 2013 setzen gemäss den fachlichen Verlautbarungen die International Standards on Auditing (ISA 2009) um. Die EXPERTsuisse hat nur minimale Ergänzungen aufgrund der schweizerischen Gesetzesbestimmungen vorgenommen und gewisse zusätzliche Standards erlassen, wie PS 290, PS 701, PS 870, PS 880, PS 890 und PS 950 EXPERTsuisse (2013).

2.2.3 Aussagen im Abschluss

Mit der Abschlusserstellung macht das Management mit den Finanzkennzahlen und den sonstigen Offenlegungen sowohl explizit als auch implizit verschiedene Aussagen. Die Prüfungsstandards der US-amerikanischen Revisionsaufsichtsbehörde, *Public Company Accounting Oversight Board (PCAOB)*, verwenden fünf Kategorien zur Unterteilung (PCAOB, 2010):

- *Existenz bzw. Eintritt:* Vermögensgegenstände oder Verbindlichkeiten existieren zu einem bestimmten Zeitpunkt und die vermerkten Transaktionen haben während der Periode stattgefunden.
- *Vollständigkeit:* Sämtliche Transaktionen und Konten, die in den Abschluss einbezogen werden sollten, sind darin enthalten.
- *Bewertung und Allokation:* Vermögensgegenstände, Verbindlichkeiten, Eigenkapital, Umsätze und Aufwendungen sind mit dem angemessenen Wert im Abschluss enthalten.
- *Rechte und Pflichten:* Die Organisation hält oder kontrolliert die Rechte zu den Vermögensgegenständen, während die Verpflichtungen den tatsächlichen Pflichten der Unternehmung entsprechen.

- *Präsentation und Offenlegung:* Die Komponenten des Finanzabschlusses sind korrekt klassifiziert, beschrieben und offengelegt worden.

Darüber hinaus kennen die internationalen Prüfungsstandards eine noch differenziertere Systematisierung der Aussagen im Abschluss, die sie nach ISA 315.A111 (IAASB, 2009e) in drei Kategorien einteilt: 1. Aussagen über die Klassen der Transaktionen, 2. Aussagen zu Abschlusssaldos und 3. Aussagen zur Präsentation und Offenlegung. Für die Zwecke dieser Arbeit ist die PCAOB-Klassifizierung jedoch ausreichend.

Aus den Aussagen im Abschluss, die das Management macht, lassen sich die Prüfziele ableiten, die sehr eng miteinander verbunden sind. Diese Verbundenheit ist folgerichtig, liegt es doch im Auftrag des Abschlussprüfers, festzustellen, ob die vom Management getätigten Aussagen gerechtfertigt sind.

2.2.4 Aufteilung der Prüfarbeit in Phasen

Der Wirtschaftsprüfer muss aufgrund seiner Arbeit nach ISA 200.3 (IAASB, 2009a) ein Prüfurteil abgeben und zu diesem Zweck genügend angemessene Prüfnachweise identifizieren, die sämtliche vom Management explizit und implizite getätigte Aussagen im Abschluss unterstützen.

Um dies zu erreichen, sehen die Prüfungsstandards für das Vorgehen drei grundlegende Phasen vor (Arens et al., 2016):

1. Prüfungsplanung
2. Prüfungsdurchführung
3. Prüfungsabschluss

Diese drei Prüfphasen werden stets von zwei begleitenden Prozessen unterstützt: die fortwährende Dokumentation des Prüfvorgangs wie auch die Qualitätskontrolle (IAASB, 2009b). Es handelt es sich um einen iterativen und nicht um einen strikt linearen Prozess, in dem Rückkoppelungseffekte beispielsweise durch eine veränderte Risikobeurteilung, wie in ISA 300.10 dargelegt wird (IAASB, 2009d), auftreten. Abbildung 20 zeigt die Phasen des Prüfungsprozesses.

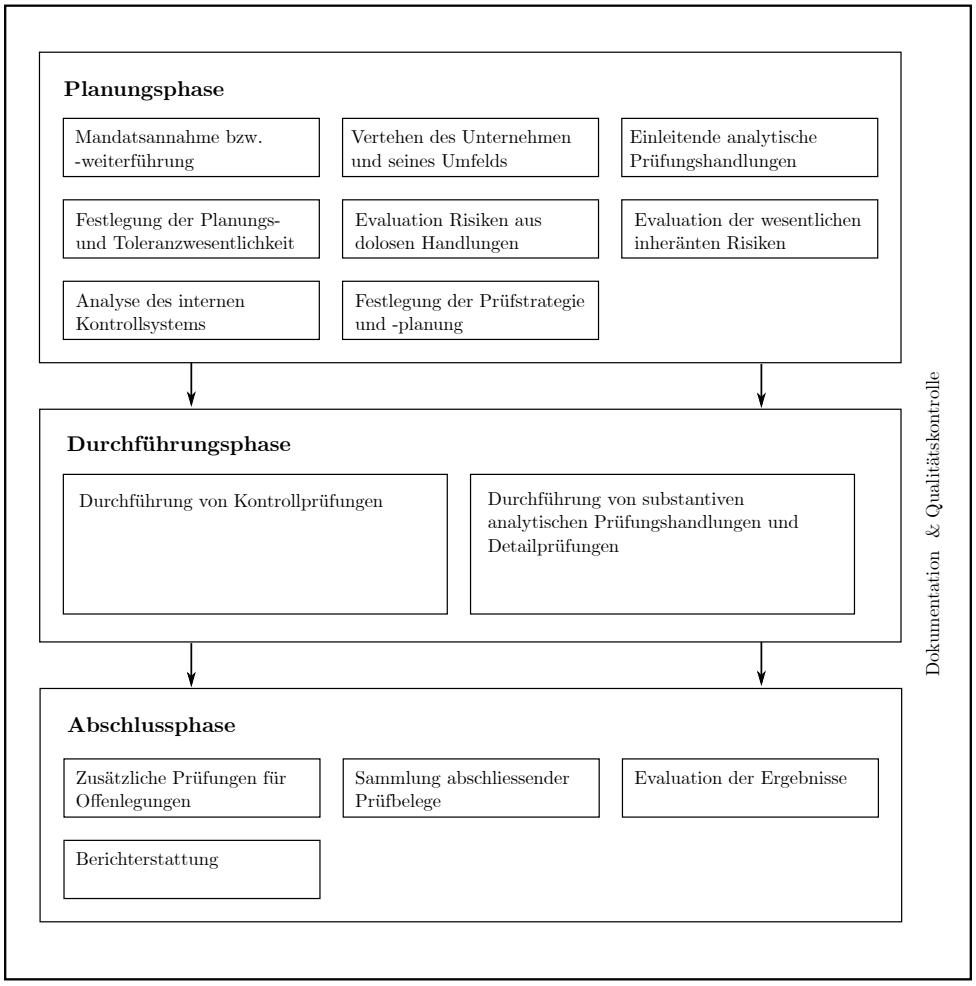


Abbildung 20: Phasen im Prüfungsprozess nach Arens et al. (2016)

2.2.5 Prüfungsplanung

Das Ziel der ersten Phase einer Prüfung besteht nach ISA 300.4 (IAASB, 2009d) stets darin, einen Prüfplan zu entwickeln, der es zulässt, effektiv vorzugehen und bis zu einem vereinbarten Termin zu einem Prüfurteil zu gelangen.

Die Prüfungsplanung kann nach Arens et al. (2016) in acht Teilschritte eingeteilt werden:

1. Entscheid über Mandatsannahme und -weiterführung
2. Verständnis des Kunden und seines Umfelds gewinnen

3. Durchführung einleitender analytischer Prüfungshandlung
4. Festlegung der Planungs- und Toleranzwesentlichkeit
5. Evaluation der wesentlichen Risiken aus dolosen Handlungen
6. Evaluation der wesentlichen inhärenten Risiken
7. Analyse des internen Kontrollsystems
8. Festlegung der Prüfstrategie und des Prüfplans.

Mandatsannahme und -weiterführung Am Anfang eines jeden Audits steht der Entscheidung des Wirtschaftsprüfers, das Mandat anzunehmen bzw. abzulehnen oder eine bereits bestehende Mandatsbeziehung weiterzuführen oder zu beenden (ISA 210.6).

Dabei gilt es, die Frage zu klären, ob bei einem Wirtschaftsprüfer die notwendigen formellen Voraussetzungen für die Wahrnehmung des Mandats wie Wählbarkeit, Zulassung und Unabhängigkeit gegeben sind. Sind diese Voraussetzungen erfüllt, wird eine Auftragsbestätigung aufgesetzt, die durch eine eindeutige und umfassende schriftliche Definition des Prüfauftrags bei den Auftraggebern und Adressaten der Berichte der Revisionsstelle eine Klarheit über Art und Umfang des Auftrags sowie über die Aussagekraft der Berichte schafft (Arens et al., 2016).

Im Folgejahr prüft der Abschlussprüfer, ob die bestehenden Auftragsbedingungen gegenüber dem Auftraggeber noch gültig sind oder eine neue Auftragsbestätigung zu erstellen ist (Treuhand-Kammer, 2009).

Verständnis über Kunden und Branche gewinnen Die Erarbeitung eines umfassenden Verständnisses über den Prüfungsmandanten und sein Umfeld ist eine der zentralen Voraussetzungen für ein ordentliches Audit. Darum ist als Pflicht für den Prüfer festgehalten:

> *Der Wirtschaftsprüfer identifiziert und evaluiert Risiken einer wesentlichen Fehldarstellung aufgrund von Irrtümern oder doloser Handlungen, basierend auf einem Verständnis des Unternehmens und seinem Umfeld, inklusive seines internen Kontrollsystems.* (IAASB, 2009e)

Der Prüfer muss über ein Verständnis zu folgenden Themengebieten verfügen:

- Industrie und externes Umfeld
- Operatives Geschäft und Prozesse
- Management und Governance
- Ziele und Strategie
- Management und Performance

Der Prüfer eignet sich diese Kenntnisse üblicherweise mithilfe von Managementinterviews und analytischen Prüfungshandlungen an (Arens et al., 2016).

Initiale analytische Prüfungshandlungen Für den Zweck der Prüfungsplanung und für die Risikobeurteilung können analytische Prüfungshandlungen eingesetzt werden, die in den Prüfungsstandards nach ISA 520.4 als Evaluation von finanziellen Informationen definiert werden und eine Analyse von plausiblen Beziehungen zwischen finanziellen und nicht finanziellen Daten ermöglichen (IAASB, 2009h).

Im Rahmen der Planung können die noch ungeprüften Abschlusswerte mit denen des Vorjahres oder gar mit einer längeren Zeitreihe verglichen und bedeutende Abweichungen erkannt werden. Diese Abweichungen können bestimmte Trends oder Einzelereignisse repräsentieren, die zusammen das Potenzial haben, den Prüfansatz zu beeinflussen (Treuhand-Kammer, 2009).

Festlegung der Wesentlichkeit bzw. Toleranzwesentlichkeit Darüber hinaus erfordern die Prüfungsstandards, dass der Auditor eine quantitative Grenze festlegt, die bestimmt, wann Fehldarstellungen einzeln oder zusammengenommen als materiell gelten. Diese Grenze bildet das Maximum, bei dem der Auditor die berechtigte Annahme hat, dass Entscheidungen von vernünftigen Adressaten nicht beeinflusst werden. Sie wird nach ISA 320.2 als Gesamtwesentlichkeit bzw. *Wesentlichkeit* bezeichnet (IAASB, 2009f).

Wirtschaftsprüfer sind gleichfalls in der Verantwortung, die Entscheidung zu treffen, ob ein Finanzabschluss frei von wesentlichen Fehlern ist, und müssen im

Rahmen ihrer Prüfarbeit das Management über festgestellte Fehldarstellungen in Kenntnis setzen und zu einer Korrektur auffordern. Im Falle, dass das Management sich aber weigert, die Fehler zu beseitigen, ist vonseiten des Wirtschaftsprüfers eine Qualifikation oder gar Versagung des Prüfurteils notwendig (Arens et al., 2016).

Es wird eine Referenzgrösse herangezogen, die bei gewinnorientierten Unternehmen in der Regel der Nettogewinn vor Steuern ist, um zu einem quantitativen Wert der Wesentlichkeit zu gelangen. Diese Kennzahl ist ein kritischer Faktor für die Adressaten der Finanzinformationen. Darüber hinaus werden aber auch andere Referenzgrössen – beispielsweise bei erheblich fluktuierenden Gewinnen, Start-ups oder NPOs – wie der Nettoumsatz, der Bruttogewinn oder das Gesamtkapital herangezogen (EXPERTsuisse, 2016).

Aufbauend auf der ausgewählten Referenzgrösse wird die Wesentlichkeit als ein Prozentsatz dieser berechnet. Obgleich die fachlichen Vorgaben keine Aussagen zu den anzuwendenden Prozentsätzen enthalten, weil befürchtet wird, sie könnten von den Nutzern ohne weitere Erwägungen mechanisch angewandt werden, gibt es in der einschlägigen Literatur einige Empfehlungen für akzeptable Bandbreiten: So können für die operativen Gewinne zwischen 3 % und 6 %, für das Gesamtkapital zwischen 1 % und 3 % und für das Umlaufvermögen zwischen 3 % und 6 % angesetzt werden (EXPERTsuisse, 2016).

Als weitere quantitative Grösse wird für die Prüfung die vorgegebene *Toleranzwesentlichkeit* vom Prüfer berechnet. Sie ist stets kleiner als die Wesentlichkeit für den Abschluss als Ganzes und wird berechnet, um das Risiko zu senken, dass aggregierte nicht korrigierte oder nicht entdeckte Fehler die *Wesentlichkeit* übersteigen (ISA 320.9). Die Festlegung der *Toleranzwesentlichkeit* ist notwendig, weil der Prüfer die Prüfnachweise in einzelnen Teilbereichen sammelt und dieser Richtwert ihn in der Entscheidung unterstützt, welche Prüfnachweise angemessen sind. Die *Toleranzwesentlichkeit* wird dabei meist als ein fester Prozentsatz der *Wesentlich* für den Abschluss als Ganzes berechnet und bewegt sich üblicherweise in einem Wertebereich von 50 % bis 75 %.

Evaluation der inhärenten Risiken Unter dem inhärenten bzw. innewohnenden Risiko wird dasjenige Risiko verstanden, dass ein bestimmtes Konto oder eine bestimmte Art von Geschäftsvorfällen falsche Darstellungen enthält, die ein-

zeln oder zusammen mit falschen Darstellungen anderer Kontensalden oder Arten von Geschäftsvorfällen wesentlich sein können, und zwar in Absicht der Existenz von etwaigen Kontrollen. Als Beispiel können flüssige Mittel genannt werden, die ein relativ hohes Mass an inhärenten Risiken bezüglich der Aussage über das tatsächliche Vorhandensein aufweisen und grundsätzlich anfällig für Eigentumsdelikte sind (EXPERTsuisse, 2016).

Die Evaluation der inhärenten Risiken beginnt in der Planungsphase und wird während des gesamten Audits stets angepasst. Relevante Informationen über die Beurteilung dieser Risikokategorie entstammen beispielsweise der Analyse der Unternehmung und ihrem Umfeld, früheren Abschlussprüfungen beim Mandanten, Werksbesichtigungen oder analytischen Prüfungshandlungen (Arens et al., 2016).

Verständnis des internen Kontrollsystems Für die Prüfung von Finanzabschlüssen sind die Wirtschaftsprüfer ferner angehalten, ein Verständnis über das vorhandene interne Kontrollsystem zu erlangen, um eine umfassendere Einschätzung über die Risiken von wesentlichen Fehlaussagen treffen zu können (ISA 315.12 IAASB, 2009e)). Für *integrierte Abschlussprüfungen* von grossen in den USA kotierten Unternehmen gilt es gar, die Effektivität der Kontrollen zu bestätigen, während in der Schweiz lediglich eine Beurteilung über die Existenz abgegeben werden muss (EXPERTsuisse, 2016).

Für die Erlangung eines besseren Verständnisses werden Prüfungshandlungen angewandt, um Prüfnachweise über die Ausgestaltung und Anwendung zu erhalten. Dazu gehören Inspektionen, Personalbefragungen, Beobachtungen und der Nachvollzug von Kontrolltätigkeiten. Üblicherweise benutzt der Prüfer drei Formen von Dokumenten, um das Verständnis zu erlangen und zu dokumentieren: schriftliche Kontrolldokumentationen, Flussdiagramme und interne Kontrollumfragen (Arens et al., 2016).

Konnte sich der Prüfer eine Übersicht und ein Verständnis über die Ausgestaltung des internen Kontrollsystems verschaffen, muss er in einem weiteren Schritt seine Umsetzung bzw. Implementierung beurteilen. Es ist üblich, dass dieser Schritt simultan mit der Beurteilung der Ausgestaltung erhoben wird – beispielsweise anlässlich eines Prozessdurchlaufs (Walkthrough), einer Beobachtung, einer Inspektion von Dokumenten oder anlässlich von Befragungen (EXPERTsuis-

se, 2016).

Im nächsten Schritt trifft der Prüfer unter Berücksichtigung der Erkenntnisse aus der Prüfung über die Ausgestaltung und Implementierung des internen Kontrollsystems eine Einschätzung zum Kontrollrisiko, das angibt, inwieweit die Annahme des Prüfers zutrifft, dass interne Kontrollen wesentliche Fehlaussagen im Abschluss verhindern bzw. korrigieren können, wenn sie bereits aufgetreten sind. Der Startpunkt für diese Beurteilung besteht häufig in *unternehmensweiten Kontrollen*, die eine Wirkung für mehrere transaktionsorientierte Kontrollen entfalten. Gleichfalls ist die Beurteilung über die Effektivität von *generellen IT-Kontrollen* wichtig, bevor es an die Evaluation automatischer oder manueller Kontrollen geht, die von der IT abhängen. Probleme bzw. ineffektive *generelle IT-Kontrollen* können potenzielle wesentliche Fehldarstellungen über alle Systeme hinweg implizieren. Hat sich der Wirtschaftsprüfer davon überzeugen können, dass die unternehmensweiten und die generellen IT-Kontrollen Kontrollen adäquat ausgestaltet sind und tatsächlich angewandt werden, kann er eine vorübergehende Einschätzung über transaktionsorientierte Kontrollen vornehmen (EXPERTsuisse, 2016).

Einfluss von dolosen Handlungen Obgleich die Oberaufsicht über eine Einheit – beispielsweise der Verwaltungsrat bei einer Aktiengesellschaft – letztlich die Verantwortung dafür trägt, dass dolose Handlungen verhindert oder zumindest aufgedeckt werden, nimmt ISA 240.5 den Prüfer in die Pflicht, sicherzustellen, dass der Abschluss frei von wesentlichen Fehlaussagen ist, auch von beabsichtigten dolosen Handlungen.

Der Prüfer muss nach ISA 240.10 (IAASB, 2009c) eine Risikoeinschätzung zum Abschluss für die Anfälligkeit für dolose Handlungen treffen und diese Ergebnisse in der Planung und der Umsetzung berücksichtigen.

Die Standards sehen zwei exponierte Bereiche der finanziellen Berichterstattung vor, die in Bezug auf dolose Handlungen einer hohen Aufmerksamkeit bedürfen:

- *Dolose Umsatzerfassung:* Bei der Umsatzerfassung muss der Wirtschaftsprüfer evaluieren, welche Arten von Umsatz, Umsatztransaktionen oder Aussagen potenziell betroffen sind. Die Ergebnisse müssen anschliessend adäquat adressiert werden (ISA240.26, IAASB, 2009c).

- *Ausserkraftsetzen von Kontrollen durch das Management:* Diese Ausserkraftsetzung bezieht sich insbesondere auf Buchungen im Hauptbuch, weil das Management die Möglichkeit hat, Kontrollen ausser Kraft zu setzen. In der Folge ist es notwendig, dass der Prüfer die Angemessenheit von Journaleinträgen abklärt (ISA 240.32, IAASB, 2009c).

ISA 240.6 (IAASB, 2009c) warnt, dass das Risiko für wesentliche Fehldarstellungen im Abschluss als Folge von dolosen Handlungen ungleich höher als das Risiko für Irrtümer ist, weil potenziell eine wohldurchdachte und sorgfältig organisierte Vorgehensweise für eine Verschleierung vorliegt.

Durchführung der Risikobeurteilung Prüfer akzeptieren ein gewisses Mass an Risiko oder Unsicherheit in Bezug auf die Prüfung von Abschlüssen. So weiss der Prüfer, dass inhärente Unsicherheiten bei der Angemessenheit von Prüfnachweisen, Unsicherheiten bei der Effektivität von Kontrollen des Kunden und Unsicherheiten bei den daraus resultierenden Finanzinformationen bestehen bleiben (Arens et al., 2016).

Das Risiko wird von wesentlichen Fehlern im Abschluss von Irrtümern oder deliktischen Handlungen bestimmt. Das Risiko wesentlicher Fehler kann nach ISA 315.25 (IAASB, 2009e) grundsätzlich in zwei Richtungen auftreten: Einerseits kann es auf der Stufe des Abschlusses als Ganzes – wenn z.B. Kontrollen auf der Unternehmensebene nicht greifen – und andererseits auf der Stufe der Aussagen im Abschluss bestehen (IAASB, 2009e).

Auf Aussageebene wird das Risiko für Fehldarstellungen von zwei Komponenten determiniert: dem inhärenten Risiko (IR) und dem Kontrollrisiko (KR). In beiden Fällen handelt es sich um Risiken der Unternehmung, die unabhängig von der Abschlussprüfung bestehen. Das *Prüfrisikomodell* wird herangezogen, um zu entscheiden, wie viel Prüfevidenz nötig ist, um Risiken auf der Aussageebene zu begegnen. Das Prüfrisikomodell besagt, dass sich das Prüfrisiko aus der Multiplikation des inhärenten Risikos (IR), des Kontrollrisikos (KR) und des Aufdeckungsrisikos (AR) ergibt (EXPERTsuisse, 2016):

$$AR * IR * KR = Prüfrisiko$$

Dabei wird unter Aufdeckungsrisiko das Risiko verstanden, dass ein materieller Fehler trotz Prüfungshandlungen des Abschlussprüfers unbemerkt bleibt. Das inhärente Risiko bemisst die Anfälligkeit einer Aussage für wesentliche Fehldarstellungen, bevor die Effektivität des internen Kontrollsystems berücksichtigt worden ist, während das Kontrollrisiko, als letzter Teil des Modells, die vom Prüfer gemachte Einschätzung darstellt, dass das Kontrollsystem ungeeignet ist, Fehler in der Buchhaltung zu entdecken und zu korrigieren. Das daraus resultierende Prüfrisiko misst das Risiko, das der Prüfer tragen muss und die Anfälligkeit angibt, dass in den Finanzabschlüssen wesentliche Fehler unentdeckt bleiben (EXPERTsuisse, 2016).

Festlegung der Prüfstrategie und des Prüfungsplans Auf Grundlage der Risikobeurteilung und der Einschätzung über die Fehlerrisiken kann der Prüfer das Entdeckungsrisiko über die Auswahl von Art, Umfang und zeitlichem Ablauf der aussagebezogenen Prüfungshandlungen in einer Art und Weise festlegen, dass er sein Prüfungsurteil mit hinreichender Sicherheit fällen kann. Als Regel gilt, dass mit abnehmenden Fehlerrisiken das Entdeckungsrisiko wächst (EXPERTsuisse, 2016).

Mit einer Prüfungsstrategie bestimmt der Wirtschaftsprüfer die angemessene Mischung aus Kontroll- und Detailprüfungen sowie aus analytischen Prüfungshandlungen. Danach entwift er auf Grundlage der Prüfstrategie eine Liste der konkreten Prüfungshandlungen für das Audit bzw. stellt das Prüfprogramm zusammen (Arens et al., 2016).

2.2.6 Prüfungsdurchführung

Die Prüfungsdurchführung setzt schliesslich den in der Planungsphase entwickelten Prüfungsplan um. Der Schwerpunkt dieser Phase liegt auf der Sammlung von Prüfnachweisen, die die Grundlage für das Prüfurteil bilden. Diese Phase gliedert sich in zwei Teile: Einerseits wird das interne Kontrollsystem und seine Funktionsfähigkeit geprüft, während die Detailprüfungen andererseits die Prüfung der Kontensalden und Transaktionen zum Gegenstand haben (Arens et al., 2016).

Prüfnachweise Es gilt in dieser Phase für den Abschlussprüfer, Prüfungshandlungen nach ISA 500 zu planen und durchzuführen, die ausreichende und geeignete Prüfnachweise zutage fördern (IAASB, 2009g).

Prüfnachweise sind Informationen, die dem Prüfer bei der Entscheidung helfen, ob die gegenwärtig geprüften Informationen den definierten Kriterien wie einer Rechnungslegungsnorm entsprechen, und zu der Einschätzung zu gelangen, dass der Abschluss frei von wesentlichen Fehlaussagen ist. Dabei variiert die Eignung der Prüfbelege und reicht von sehr gut für Kontoauszüge bis hin zu eher schwach für Aussagen von Mitarbeitern (Arens et al., 2016).

Den Prüfer begleitet stets die Frage, welche Prüfnachweise geeignet oder zumindest noch ausreichend sind. Vier Entscheidungen zu Art und Umfang der Prüfbelege sind zu treffen (Arens et al., 2016):

- Welche Prüfungshandlungen sind durchzuführen?
- Welche Stichprobengrössen sind anzuwenden?
- Welches Element ist aus der Grundgesamtheit auszuwählen?
- Zu welchem Zeitpunkt sind die Prüfungshandlungen durchzuführen?

In den Prüfungsstandards (ISA 500 (IAASB, 2009g) wird gefordert, dass der Wirtschaftsprüfer ausreichend angemessene Prüfbelege sammelt, die die Grundlage für sein Prüfurteil bilden. Es liegt in der Natur der Prüfnachweise und der Kostenerwägungen, die für eine Abschlussprüfung zu berücksichtigen sind, dass der Prüfer sein Urteil nicht mit absoluter Sicherheit fällen kann. Unter Berücksichtigung aller Prüfnachweise, die während eines Audits gesammelt werden, muss der Prüfer sein Urteil jedoch mit einem hohen Mass an Sicherheit fällen und darüber in einem Prüfurteil berichten. Eine Beurteilung der Prüfbelege ist massgeblich durch ihre *Eignung* und ihren (quantitativen) *Umfang* determiniert (Arens et al., 2016).

In Rahmen einer Prüfung versteht sich die *Eignung* nach ISA 500.5b als Massgabe für die Qualität der Belege, die wiederum durch die *Relevanz* und die *Verlässlichkeit* bestimmt wird. Relevant ist ein Prüfnachweis, wenn er zum Ziel des Audits bzw. zu neuen Erkenntnissen über das Prüfgebiet beiträgt (IAASB, 2009g).

Die *Verlässlichkeit* ist der Grad, mit dem einem Prüfbeleg Vertrauen geschenkt werden kann. Auch für die *Verlässlichkeit* gilt, dass sie den Prüfer darin unterstützt, zu der Überzeugung zu gelangen, dass der Abschluss frei von wesentlichen Fehldarstellungen ist. Die *Verlässlichkeit* hängt nach Arens et al. (2016) dabei von sechs wesentlichen Faktoren ab:

- *Unabhängigkeit:* Eine Prüfevidenz, die ausserhalb der geprüften Organisation erhoben wurde, gilt grundsätzlich als verlässlicher als eine Prüfevidenz, die die Unternehmung selbst zur Verfügung stellt. So sind beispielsweise die Schriftwechsel und Dokumente von Banken, Anwälten und Kunden verlässlicher als Interviews, die mit dem Betriebspersonal entstanden.

- *Effektivität des internen Kontrollsystems:* Ist das interne Kontrollsystem des Prüfkunden effektiv, so sind auch die von ihm erhaltenen Dokumente von höherer Verlässlichkeit.

- *Durch Prüfer erhoben:* Prüfbelege, die direkt vom Prüfer beschafft wurden, wie bei einer Inventurbeobachtung, Inaugenscheinnahme, Nachrechnung und Inspektion, sind verlässlicher als indirekt erhaltene Belege.

- *Qualifikation der Informationsquellen:* Auch wenn die Informationsquelle unabhängig ist, können Prüfbelege als unzuverlässig gelten, insbesondere dann, wenn die Quelle nicht über eine geeignete Qualifikation verfügt. So kann beispielsweise die Prüfevidenz eines Prüfers im Zuge einer Inventurbeobachtung bei einem Diamanthändler unzuverlässig sein, wenn er nicht über das spezifische Fachwissen im Bereich Edelsteine verfügt.

- *Grad der Objektivität:* Prüfbelege gelten als verlässlicher, wenn ihre Objektivität hoch ist und nur geringe Ermessensspielräume bestehen. So sind Anwaltsschreiben mit Einschätzungen zum Ausgang eines Gerichtsverfahrens weniger verlässlich als die physische Anzahl von Aktien oder Geräten eines Mandanten.

- *Zeitnähe:* Die Zeitnähe bezieht sich auf den Zeitpunkt der Erhebung von Prüfbelegen und auf die Periode, die sie abdeckt. Für Abschlussposten gilt, dass die Nähe zum Bilanzstichtag die Verlässlichkeit grundsätzlich erhöht.

Erfolgsrechnungsposten gelten als verlässlich, wenn sie die gesamte zu prüfende Periode abdecken.

Die Quantität der Prüfbelege bestimmt, ob diese ausreichend sind. Für gewisse Prüfungshandlungen bestimmt die Stichprobengrösse, ob hinreichend Prüfbelege gesammelt wurden. Für andere Prüfungshandlungen sind es primär die Anzahl und Qualität der Prüfungshandlungen, die zum Prüfziel beitragen. Mehrere Faktoren bestimmen, wie gross eine Stichprobe sein sollte: Die zwei wichtigsten sind die vom Prüfer gebildete Erwartung von wesentlichen Fehlern sowie die Effektivität der internen Kontrollen (Arens et al., 2016).

Prüfungshandlungen Wenn es darum geht, welche Prüfungshandlungen angewandt werden sollen, kann der Wirtschaftsprüfer nach den Prüfungsstandards im Rahmen der Prüfungsdurchführung (Arens et al., 2016) aus acht Möglichkeiten auswählen:

- *Einsichtnahme/Inaugenscheinnahme:* Bei dieser Möglichkeit zählt und inspiziert der Prüfer physische Anlagegüter. Diese Form der Prüfnachweise wird häufig mit der Prüfung von Lagerbeständen und Bargeld in Verbindung gebracht, findet aber auch eine Anwendung bei Aktien oder Sachanlagen.

- *Externe Bestätigung:* Diese Möglichkeit bezieht sich auf den Empfang einer direkten schriftlichen Antwort einer Drittpartei, die die Genauigkeit einer Prüferanfrage bestätigt. Da die Antwort von einer Drittpartei gegeben wird, gilt dieser Prüfnachweis als äusserst verlässlich und teuer bzw. ressourcenintensiv.

- *Inspektionen:* Bei der Inspektion werden Dokumente des Kunden gesichtet, mit denen die Informationen substantiviert werden können, die in den Abschlüssen enthalten sind bzw. darin enthalten sein sollten. Die Dokumente, die vom Prüfer inspiziert werden, sind Aufzeichnungen, die Informationen darüber enthalten, wie ein Kunde sein Geschäft führt. Da jeder Buchung grundsätzlich ein Beleg beiliegt, können grössere Mengen an Dokumenten zur Verfügung stehen.

- *Analytische Prüfungshandlungen:* Bei den analytischen Prüfungshandlungen handelt es sich um Evaluationen von Finanzinformationen durch die Analyse von plausiblen Abhängigkeiten zwischen finanziellen und nicht finanziellen Daten. So kann der Auditor möglicherweise die Bruttogewinnmargen des Berichtsjahres mit der vergangenen Periode vergleichen und erkennen, dass Abschlussbuchungen fehlen. Generell wird diese Möglichkeit ausgiebig genutzt, zumal diese Form der Beschaffung von Prüfnachweisen durch die Standards in der Planungs- und in der Schlussphase verbindlich vorgeschrieben ist.

- *Befragungen:* Eine weitere Möglichkeit des Nachweises resultiert aus dem Ergebnis von Befragungen, bei denen der Prüfer schriftliche oder mündliche Informationen des Prüfkunden als Antwort auf seine Fragen erhält. Trotz des Umstands, dass häufig sehr viele Prüfnachweise auf diese Art erhoben werden, reichen Befragungen allein nicht aus, weil sie nicht von einer unabhängigen Partei stammen und möglicherweise eine Verzerrung zugunsten des Unternehmens aufweisen.

- *Nachrechnungen:* Üblich ist die Nachprüfung von Berechnungen des Kunden, indem insbesondere die arithmetische Korrektheit geprüft wird, beispielsweise bei einem Abzug von Abschreibungen vom Buchwert. Hilfreich sind spezialisierte Softwarepakete für die Wirtschaftsprüfung oder *Excel*.

- *Nachvollzug:* Mit dieser Möglichkeit testet der Prüfer die Prozesse in der Buchhaltung oder die Kontrollen, die als Teil des internen Kontrollsystems vorhanden sind.

- *Beobachtung:* Der Prüfer kann die Prozesse und Abläufe beobachten, die vom Management oder den verantwortlichen Mitarbeitern ausgeführt werden. Er kann sich Prozesse und Verantwortungsbereiche vergegenwärtigen, die der Prüfmandant ausführt, ist jedoch mit seiner Beobachtung auf die Zeit während seiner Anwesenheit beschränkt. Deshalb reicht auch die Beobachtung im Rahmen der Kontrollprüfung allein nicht aus, sondern muss durch ergänzende Prüfbelege unterstützt werden.

Analytische Prüfungshandlungen Analytische Prüfungshandlungen sind in den Prüfungsstandards unter ISA 520.4 (IAASB, 2009h) als

> *„Beurteilung von Finanzinformationen durch die Analyse von plausiblen Beziehungen zwischen sowohl finanziellen als auch nicht finanziellen Daten“*

definiert.

Diese Art der Prüfung kommt grundsätzlich während der drei Phasen des Audits zum Einsatz und wird in einzelnen Phasen vorgeschrieben (Arens et al., 2016):

- *Planungsphase:* Analytische Prüfungshandlungen sind zwingend während der Planungsphase als Teil der Risikoeinschätzung durchzuführen. Damit wird ein Verständnis des Kunden für sein Geschäft und der Branche, in der sich dieser bewegt, erreicht und gleichfalls eine Einschätzung über Art, Umfang und Zeitpunkt der Prüfungshandlungen ermöglicht. Dem Prüfer wird geholfen, wesentliche Risiken zu identifizieren, die eine besondere Behandlung erfordern und im weiteren Verlauf der Prüfung zu berücksichtigen sind. So kann ein verlangsamter Debitorenumschlag eine spezielle Prüfungsbehandlung – beispielsweise mit externen Bestätigungen – nach sich ziehen. Ausmass, Tiefe und Zeitpunkt der analytischen Prüfungshandlungen variieren je nach Kunde und werden in der Prüfungsplanung üblicherweise auf einem hohen Datenaggregationsniveau wie Abschlussposten durchgeführt. Für manche Prüfkunden kann es genügen, Veränderungen gegenüber dem Vorjahr für die noch nicht geprüften Abschlüsse zu analysieren.

- *Prüfungsdurchführung:* Oft werden analytische Prüfungshandlungen während der Prüfungsdurchführung als substantive Prüfungshandlungen angewandt, um Prüfevidenz für die Kontenbestände zu erhalten. Diese substantiven analytischen Prüfungshandlungen werden häufig im Zusammenspiel mit anderen Prüfungshandlungen durchgeführt. So kann der Bestand an vorausbezahlten Versicherungsprämien im vorangegangenen Jahr mit den Vorauszahlungen im gegenwärtigen Jahr verglichen werden und damit eine Prüfevidenz für diesen Abschlussposten bzw. Kontensaldo erlangt werden.

Die Verlässlichkeit der Prüfbelege, die analytische Prüfungshandlungen hervorbringen, hängt von der Prognosefähigkeit der Beziehungen, der genauen Vorhersage des Prüfers und der Verlässlichkeit der zugrunde liegenden Daten ab, die für die Vorhersage benutzt wurden. Die Prüfungsstandards verlangen im Hinblick auf die angewendeten substantiven Prüfungshandlungen, dass die Arbeitspapiere einen Erwartungswert und die Faktoren enthalten, die für die Bildung des Erwartungswerts benötigt werden. Daneben wird gleichfalls verlangt, dass der Wirtschaftsprüfer die Verlässlichkeit der Daten evaluiert, die für die Erarbeitung des Erwartungswerts herangezogen wurden, inklusive der Datenquellen und der Kontrollen, die für die Datenvorbereitung durchgeführt wurden.

- *Prüfabschluss:* Abschliessend sind analytische Prüfungshandlungen als Teil des Mandatsabschlusses zwingend vorgesehen. Damit soll sichergestellt werden, dass der Wirtschaftsprüfer einen finalen objektiven Blick auf die finanzielle Berichterstattung wirft und den geprüften Abschluss einem letzten Review für materielle Fehldarstellungen oder finanzielle Probleme unterzieht. Typischerweise ist es ein Mandatspartner mit umfangreichem Erfahrungsschatz, der die finale analytischen Prüfungshandlung als Teil des Mandatsabschlusses durchführt.

Ziel und Durchführung der analytischen Prüfungshandlungen unterscheiden sich hinsichtlich der Phasen, in denen sie zur Anwendung kommen. Analytische Prüfungshandlungen, die während der Planungsphase durchgeführt werden, verwenden generell aggregierte Daten, um damit den Kunden und sein Geschäft zu verstehen bzw. Prüfgebiete zu identifizieren, die für materielle Fehler anfällig sind. Demgegenüber stehen die substantiven analytischen Prüfungshandlungen, die Prüfevidenz produzieren und verlässlichere Prüfbelege erfordern, die wiederum oftmals auf disaggregierten Daten beruhen (EXPERTsuisse, 2016).

2.2.7 Prüfabschluss

Neben der Sammlung der letzten Prüfnachweise wechselt in der letzten Phase der Fokus der Abschlussprüfung auf die vorsichtige Durchsicht der Resultate und zu

weiteren kritischen Punkten wie die Offenlegung der Finanzinformationen und der Berichterstattung.

Prüfung der Offenlegung In Bezug auf die Prüfung der offengelegten Finanzinformationen muss der Prüfer entscheiden, ob die Prüfsicherheit in den vorangegangenen Phasen erreicht wurde oder ob noch weitere Prüfungshandlungen notwendig sind. Diesbezüglich steht die Vollständigkeit der Informationen im Vordergrund und die Frage, ob das Management alle notwendigen Angaben gemacht hat. Aufgrund der speziellen Eigenschaften von Eventualverpflichtungen und Ereignissen nach dem Bilanzstichtag sind üblicherweise signifikante Risiken für Fehldarstellungen vorhanden (Arens et al., 2016).

Sammlung abschliessender Prüfbelege Daneben hat der Prüfer im Rahmen seiner Prüfungsbeendigung noch fünf weitere Formen der abschliessenden Prüfevidenz zu sammeln.

Es bedarf einer finalen analytischen Prüfungshandlung, die sich dafür eignet, potenzielle materielle Fehldarstellungen zu identifizieren oder andere Probleme aufzudecken, die bislang verborgen geblieben sind, und einen letzten Blick auf den Abschluss zu tätigen (Arens et al., 2016).

Daneben sehen es die Prüfungsstandards vor, dass mit Abschluss des Mandats erneut geprüft wird, ob wesentliche Zweifel daran bestehen, dass die Fortführungsfähigkeit der Unternehmung eingeschränkt sein könnte (für die Zeit bis frühestens ein Jahr nach Bilanzstichtag). Diese Prüfungshandlung bietet sich zu diesem Zeitpunkt an, werden doch sehr oft während eines Mandats Anpassungen an den Finanzabschlüssen notwendig (Arens et al., 2016).

Den Prüfstandards folgend ist zudem während des Abschlusses zwingend eine *Vollständigkeitserklärung* seitens des (Top-)Managements notwendig, die dazu dient, das Management daran zu erinnern, dass es dafür verantwortlich zeichnet, dass die Finanzinformationen wahr und fair dargestellt werden. Zudem soll das Management erneut eine Erinnerungsstütze erhalten, was eine vollständige Finanzberichterstattung ausmacht, beispielsweise indem es aufgefordert wird zu bestätigen, dass keine verpfändeten Anlagegüter vorliegen. Zudem sind in der Vollständigkeitserklärung Antworten zu Managementbefragungen aufgeführt, wo-

mit der Gefahr von Missverständnissen begegnet werden soll (Arens et al., 2016).

Ferner können Unternehmen über das geforderte Mass der Rechnungslegungsstandards hinaus berichten, wobei der Prüfer entsprechende Vorkehrungen zu treffen hat, die diesen Umstand berücksichtigen. Zudem sind die Finanzkennzahlen häufig Teil eines grösseren Berichts, den der Auditor abschliessend zu lesen hat und auf seine Konsistenz hin prüfen muss (Arens et al., 2016).

Evaluation der Ergebnisse Sind die abschliessenden, vorgegebenen Prüfungshandlungen durchgeführt, so hat der Wirtschaftsprüfer die Resultate zu bewerten und zu einem übergreifenden Fazit bezüglich des Abschlusses zu kommen (Arens et al., 2016).

Er muss entscheiden, ob genügend geeignete Prüfbelege vorliegen, um zu dem Ergebnis zu gelangen, dass der Finanzabschluss im Einklang mit einem vorgegebenen Regelwerk entstanden ist. Zu diesem Zweck muss er sämtliche Prüfdokumentationen durchsehen, um zu entscheiden, ob alle wesentlichen Transaktionen, Konti und Offenlegungen adäquat geprüft wurden. Wesentlich ist die Kontrolle, dass alle Bereiche des Audits vollständig abgearbeitet und dokumentiert sind. Daneben geht es um die Entscheidung, ob der Finanzabschluss korrekt dargestellt und die wesentlichen Fehldarstellungen hinreichend aufgedeckt wurden. Gleichfalls gilt es, vor Mandatsabschluss die Offenlegungen im Finanzabschluss zu prüfen und dabei sicherzustellen, dass die Anforderungen der Präsentations- und Offenlegungsziele abschliessend erreicht wurden (Arens et al., 2016).

Berichterstattung Die Berichterstattung bildet den Abschluss des Mandats und umfasst die Kommunikation mit dem *Auditcommittee* oder anderen für die Unternehmungsführung verantwortlichen Personen. Die aufgedeckten deliktischen und anderen illegalen Handlungen seitens des Unternehmens sowie das Versagen der internen Kontrollen sind gleichfalls an Kommunikationspflichten des Abschlussprüfers geknüpft. Hier sind häufig auch die mit der Unternehmensführung betrauten Personen oder das obere Management betroffen, die über den Sachstand informiert werden müssen (Arens et al., 2016).

2.3 Maschinelles Lernen in der Prüfung

Nicht erst seit der Publikation von Susskind und Susskind (2015) und ihren Ausführungen zur Zukunft der Berufe wurde die Frage evaluiert, in welchem Masse maschinelles Lernen für das Audit eingesetzt werden kann. Vielmehr lässt sich der Versuch, maschinelles Lernen und Aufgaben der Wirtschaftsprüfung miteinander zu verbinden, bis in die 1950er-Jahre zurückverfolgen (Keenoy, 1958).

Heute führen Verfechter der Technik sogar an, dass mit der steigenden Verbreitung und Nutzung von Onlinesystemen, der Durchdringung von *Smart Devices* im Berufsalltag und dem „*Internet of Things*“ Buchhaltungstransaktionen zunehmend komplexer und anfälliger für Manipulationen geworden sind. Diese neuen Technologien erforderten einen versierteren Wirtschaftsprüferstand und den Einsatz immer komplexerer Techniken – speziell maschinelles Lernen (Amani & Fadlalla, 2017).

Aufbauend auf dem Framework von Louwers et al. (2013) haben Issa et al. (2016) theoretisch gezeigt, dass maschinelles Lernen in jeder Phase der Abschlussprüfung miteinbezogen werden kann:

1. *(Vor-)Planungsphase:* Maschinelles Lernen kann die Erhebung, Aggregation und Analyse von Unternehmensdaten – vorgehalten als *Big Data* – unterstützen und die initialen Risiken des Unternehmens identifizieren helfen.

2. *Verhandlungsphase:* In dieser Phase kann maschinelles Lernen einen Beitrag für die Schätzung der aufzuwendenden Stunden und des Audithonorars leisten. Darüber hinaus liesse sich die Aufbereitung von *Engagement-Letters* aufgrund der bereits ausgestellten Vorlagen automatisieren.

3. *Verständnis des internen Kontrollsystems und Identifikation von Risikofaktoren:* Dieser Schritt kann dahingehend mithilfe des maschinellen Lernens unterstützt werden, dass Computer dazu eingesetzt werden, Flowcharts, Prozesse und Fragebögen zu analysieren. Zudem lassen sich mithilfe bestimmter Muster in den Finanzdaten die Risikofaktoren identifizieren.

4. *Bestimmung des Kontrollrisikos:* Maschinelles Lernen macht es grundsätzlich möglich, sämtliche Durchführungen einer Kontrolle auf ihre ordnungsgemässe Abwicklung zu testen und etwaige Verstösse oder Unregelmässigkeiten zu

melden. Issa und Kogan (2014) haben vorgeschlagen, ein Rankingsystem zu implementieren, das Vorkommnisse priorisiert und mithilfe des maschinellen Lernens eine Risikoeinschätzung vorbereitet. Weiter ist es durch maschinelles Lernen möglich, gestützt auf Prozessmining-Techniken nicht nur zu bestimmen, ob ein internes Kontrollsystem adäquat entworfen wurde, sondern auch, ob es angemessen konfiguriert und implementiert worden ist (Jans et al., 2014).

5. *Aussagebezogene Prüfungshandlungen:* Der grösste Einfluss auf die Prüfung wird bei den aussagebezogenen Detailprüfungen erwartet. Datenherkunft und -qualität können mit der neuen Technologie bestenfalls gleich bei ihrer Erhebung evaluiert werden. Im Gegensatz zum gegenwärtig auf Stichproben gestützten Ansatz können die Wirtschaftsprüfungsgesellschaften bei Flussgrössen – wie beispielsweise Umsatz oder Aufwandsposten – kontinuierlich und vollständig die Daten untersuchen. Diese Idee lässt sich jedoch auch auf Abschlussbestände anwenden. Die kontinuierliche und umfassende Erfassung sowie Prüfung von Buchungen reduzieren die Anfälligkeit einzelner Kontensalden und des Abschlusses als Ganzes für wesentliche Fehler. Die Anwendung von Mustererkennung, Visualisierung, Benchmarks und Ausreisseridentifikation kann zusätzlich zu den analytischen Prüfungshandlungen die Effektivität der Prüfung massgeblich steigern.

6. *Prüfungsabschluss:* Am Ende eines Prüfauftrags stehen die abschliessende Urteilsbildung, die auf den zuvor gesammelten Prüfnachweisen basiert, und die Ausarbeitung eines Abschlussberichts. Prädiktive analytische Modelle können helfen, das Prüfurteil mithilfe von differenzierten Wahrscheinlichkeiten auszudrücken, und damit die gegenwärtig verpflichtend anzuwendende dichotome Einschätzung ablösen, dass ein Abschluss entweder frei oder nicht frei von wesentlichen Fehlaussagen ist.

Gierbl (2021) untersucht gleichfalls die Datenanalysemethoden, die in der Wirtschaftsprüfung einsetzbar sind, und nennt als Einsatzgebiet für die Methoden des maschinellen Lernens folgende Prüfgebiete:

1. Risikoeinschätzung

2. Dolose Handlungen

3. Analytische Prüfungshandlungen

4. Einschätzung Unternehmensfortführung

5. Berichterstattung

Für das Prüfgebiet der *Risikoeinschätzung* zeigten beispielsweise Amani und Fadlalla (2017) sowie Gierbl (2021), dass gerade mit der Entwicklung hin zu grossen Datenmengen die Risikoanalyse entscheidend von der Erkennung von Mustern wie beispielsweise komplexen Anomalien und Diskrepanzen abhängt, die potenzielle Gefahren in sich bergen (Ramamoorti et al., 1999). In Bezug auf das interne Audit konnten die Autoren wie auch Calderon (1999) die Nützlichkeit neuronaler Netze vor allem bei der Identifikation risikobehafteter Prüfgebiete aufzeigen.

Für die *Prüfungsdurchführung* wurde von Argyrou und Andreev (2011) ein Kontenclustering vorgestellt, das als interne Kontrolle unter Anwendung von selbst organisierenden Karten die Abarbeitung von Transaktionen verifizieren und die Richtigkeit der Finanzinformationen sicherstellen kann. Empirische Daten zeigen, dass dieser Ansatz eine umfangreiche Menge an Buchhaltungstransaktionen verarbeiten kann und homogene, klar abgegrenzte und interpretierbare Cluster liefert. Für die aussagebezogenen Prüfungshandlungen verwendeten Coakley und Brown (1993) sowie Koskivaara (2000a) neuronale Netze, um Muster in den Saldi von Monatsabschlüssen vorherzusagen. Sie konnten zeigen, dass neuronale Netze Muster in Abschlüssen wie auch Dynamiken und Beziehungen zwischen Konten effektiver aufgedeckt haben, als dies bei der Anwendung von Regressions- oder Ratio-Analysen der Fall gewesen war. Coakley (1995) empfahl die Anwendung von neuronalen Netzen für die Aufdeckung von Mustern bei analytischen Prüfungshandlungen. Er konnte zeigen, dass neuronale Netze verlässlicher materielle Fehler identifizierten als Revisionsexperten. Ein neuronales Netz wurde gleichfalls im Rahmen eines kontinuierlichen Audits empfohlen. Koskivaara und Back (2007) konnten belegen, dass dieses im Audit bessere Vorhersagen für Umsätze und Ausgaben lieferte, als dies mit traditionellen Methoden möglich gewesen wäre.

Im Rahmen der Beurteilung der *Unternehmensfortführungsfähigkeit* haben einige Autoren unterschiedliche Schwerpunkte gelegt: So untersuchten Etheridge et

al. (2000) die Eignung maschinellen Lernens bei der Einschätzung der Solvenz eines Unternehmens. Eine Reihe weiterer Autoren befasste sich mit der Frage, wie sich Insolvenzen mithilfe des maschinellen Lernens prognostizieren lassen (Ahn et al., 2000; Chakraborty und Sharma, 2007; Tang und Chi, 2005; S.-M. Huang et al., 2008; Youn und Gu, 2010; Benhayoun et al., 2013; M.-Y. Chen, 2013; Y. Chen et al., 2013; Li et al., 2013). Auf der anderen Seite wurde auch die Prognose einer Überschuldung intensiv und durchgehend erforscht. Viele der Autoren setzten maschinelles Lernen für solche Vorhersagen ein (Jo et al., 1997; O'leary, 1998; Yang et al., 1999; Zhang et al., 1999; Charalambous et al., 2000; Lee et al., 2005; Min und Lee, 2005; McKee, 2007; Tsai und Wu, 2008; H.-J. Chen, Huang et al., 2009; H.-J. Chen, Huang et al., 2009; Rafiei et al., 2011; Shirata et al., 2011; Olson et al., 2012; Fedorova et al., 2013; Kasgari et al., 2013; Korol, 2013; Serrano-Cinca und Gutiérez-Nieto, 2013; Tinoco und Wilson, 2013). Darüber hinaus haben Shirata et al., 2011 gezeigt, dass sich auch der Einsatz von Textmining – eine spezifische Anwendung des maschinellen Lernens – effektiv zur Erkennung von Überschuldungen eignet. Insbesondere beim Abschluss eines Mandats hat der Wirtschaftsprüfer die Annahme der Unternehmensfortführung zu prüfen. Diese Einschätzung kann sich nach ISA 570 (IAASB, 2009i) gelegentlich als komplex erweisen. Um dennoch eine Vorhersage zu machen, hat Peel (1989) mithilfe einer logistischen Regression gezeigt, dass ein hoher Verschuldungsgrad, eine geringe Profitabilität und eine Aktionariatskonzentration konsistent mit der Entscheidung der Wirtschaftsprüfer zusammengefallen sind, eine Going-Concern-Einschränkung anzubringen. Weitere Autoren setzten logistische Regressionen und neuronale Netze wie auch Entscheidungsbäume ein, um eine Going-Concern-Unsicherheit zu identifizieren (Koh und Tan, 1999, Lenard et al., 1995 und Chye Koh und Kee Low, 2004). Das Ergebnis der beiden zuerst genannten Studien lautet, dass neuronale Netze robuste Modelle liefern und den Prüfer in seiner Einschätzung unterstützen. Chye Koh und Kee Low (2004) Studie führt zu dem Ergebnis, dass Entscheidungsbäume als prädiktive Methode für die Vorhersage eines Problems bei der Unternehmensfortführung den logistischen Regressionen überlegen sind.

Neben der Going-Concern-Problematik hat sich der Wirtschaftsprüfer nach ISA 240 (IAASB, 2009c) stets der *deliktischen Handlung* bzw. *dolosen Handlungen (kurz Fraud)* anzunehmen. Die Aufdeckung von manipulierten Finanzberich-

ten unter Anwendung der etablierten und empfohlenen Prüfungshandlungen ist in den letzten Jahren zunehmend schwieriger geworden (Dikmen & Küçükkocaoğlu, 2010). Die parallel verlaufende Forschung, die auf die Aufdeckung von Fraud abzielt, konnte jedoch in der gleichen Zeit wesentliche Fortschritte erzielen. Infolgedessen liessen sich Erfolge auf der Makroebene ermitteln (Comunale et al., 2010), während andere Autoren Fraud auf der Mikroebene bzw. Transaktionsebene untersuchten (Debreceny und Gray, 2010; Bella et al., 2009; Tackett, 2013). Während Debreceny und Gray (2010) Nummern, Buchungsbeträge und deren Verteilung analysierten, untersuchten Bella et al. (2009) elektronische Zahlungsdaten mit vierstufigen selbstorganisierenden Karten. Tackett (2013) empfiehlt die Anwendung von assoziativen Regeln für die Erkennung von Fraud im Transaktionsdatenbestand einer Unternehmung. Bay et al. (2006) deckten Irregularitäten auf der Hauptbuchebene auf, während Jans et al. (2010), Jans et al. (2011) und Owusu-Ansah et al. (2002) Fraud auf der Prozessebene analysierten. Neben deliktischen Handlungen, die eine Vermögensentschädigung des Unternehmens bewirkten, widmet sich die Forschung auch Fraud auf der Managementebene, zumal diese Art der deliktischen Handlungen durch bewusste Falschdarstellungen im Finanzreporting die Anleger besonders betrifft. Zu diesem Zweck wurde maschinelles Lernen eingesetzt, um zu erkennen, ob das Topmanagement deliktische Handlungen begeht (Fanning und Cogger, 1998; Pai et al., 2011). Die durch maschinelles Lernen unterstützte Identifikation von Fraud basiert beispielsweise auf dem Forecast der Unternehmensperformance (Virdhagriswaran & Dakin, 2006) und der Identifikation von Fraud direkt in Finanzberichten (Kirkos, Spathis und Manolopoulos, 2007; Ata und Seyrek, 2009; Deng, 2009; Zhou und Kapoor, 2011). Weitere Forschungsgruppen verwendeten maschinelles Lernen, um Fraud zu prognostizieren (Cerullo und Cerullo, 2006; Kirkos et al. 2000; Green und Choi, 1997; Hoogs et al., 2007; S.-Y. Huang et al., 2012; Liang und Lv, 2009; Krambia-Kapardis et al., 2010; Öğüt et al., 2009; Ravisankar et al., 2011; Tsaih et al., 2009; Yue et al., 2009; Zouboulidis und Kotsiantis, 2012; Kotsiantis et al., 2006; J. Perols, 2011). Wesentliche Erkenntnis aus dieser umfangreichen Forschung ist die Genauigkeit der Vorhersage für eine SEC-Untersuchung *ex-post*. Eine Erklärung für den Erfolg insbesondere der neuronalen Netze liegt in ihrem Vermögen, während des adaptiven Lernens zu entscheiden, welche Signale wichtig sind und wie das Rauschen zu verstehen

ist. Die Suche nach Indikatoren für Management-Fraud ist jedoch nicht auf den quantitativen Teil der Finanzberichte beschränkt, sondern wurde auch auf den Anhang der Abschlüsse ausgedehnt. So entwickelten Yu et al. (2013) maschinelle Lernmodelle, die Verstösse in den Offenlegungspflichten erkennen. Gupta und Gill (2012) verwendeten Support-Vektor-Maschinen für die Erkennung von Fraud im qualitativen Teil der Finanzberichterstattung.

3 Methodologie

Methodologisch wird für die Beantwortung der eingangs gestellten Forschungsfragen auf den *Constructive Research Approach (CRA)* nach Kasanen et al. (1993) zurückgegriffen. Die angewandte konstruktive Forschung ist in den technischen Wissenschaften, der klinischen Medizin und im *Operations Research* zu finden, bisweilen aber auch in Bereichen, die sich mit Bilanzierung und Rechnungslegung befassen (Kasanen et al., 1993).

Kasanen et al. (1993) haben die Methode mit dem Ziel vorgestellt, Wissenschaftler in Forschungsgebieten zur Bilanzierung und Rechnungslegung zu unterstützen, eine aktivere Rolle einzunehmen, um die gegenwärtige Praxis zu verbessern. Im Idealfall ist es mithilfe dieses Ansatzes möglich, ein praktisches Problem zu lösen und gleichzeitig wichtige wissenschaftliche Einsichten zu gewinnen. Dabei wird das vorliegende real existierende Problem mit neuen Konstrukten wie einem Managementwerkzeug, Technikanwendungen, Prozessanalysen oder Organisationscharts bewältigt (Lukka, 2000). Durch diese Kombination von Theorie und dem real existierenden Problem kommt dem *konstruktiven Forschungsansatz* ein hohes Potenzial zu, eine Brücke zwischen Wissenschaft und Praxis zu schlagen (Lehtiranta et al., 2015).

3.1 Charakterisierung des Ansatzes

Dem *konstruktiven Forschungsansatz* liegt die Annahme zugrunde, dass der Wert des Wissens in Anlehnung an einen Pragmatismus (Lukka, 2000) durch seine praktischen Konsequenzen determiniert ist (Hammersley, 2004). Es wird die Lösung eines Problems gesucht, das einen wesentlichen Einfluss auf gegenwärtige Entwicklungen hat, womit dem gewonnenen Wissen eine instrumentale Bedeutung zukommt (Lehtiranta et al., 2015).

Die konstruktive Forschung grenzt sich in zweifacher Hinsicht von der *Aktionsforschung* ab: Einerseits richtet sie sich auf das Konstrukt als das Ergebnis, während die Aktionsforschung möglicherweise andere Ziele in den Blick nimmt. Daneben ist die Interaktion mit der Praxis sowie mit Praktikern in der konstruktiven Forschung lediglich üblich, während diese in der Aktionsforschung verbindlich

ist (Lehtiranta et al., 2015).

Konstruktive Forschung sollte die Identifikation eines praktisch relevanten Forschungsproblems als Ausgangspunkt haben (Kasanen et al., 1993). Nachdem ein solches Problem gefunden wurde, richtet sich die Aufmerksamkeit des Forschers auf die Problemlösung, indem er sich ein umfassendes Verständnis über die Problemsituation und die relevanten Theorien verschafft, die sich potenziell für die Konstruktion einer Problemlösung eignen.

Es gilt, dass nicht jegliche Problemlösungsaktivität der konstruktiven Forschung anzurechnen ist, obwohl die Lösung eines praktischen Problems im Zentrum des konstruktiven Forschungsansatzes steht. Kasanen et al. (1993) identifizierten folgende vier Elemente, die stets in der konstruktiven Forschung enthalten sein müssen:

- praktische Relevanz,
- theoretische Verbindung,
- praktische Einsatzmöglichkeit,
- theoretischer Beitrag.

3.2 Vorgehen im konstruktiven Forschungsansatz

Basierend auf der Arbeit von Kasanen et al. (1993) und Lukka (2000, 2002) werden sieben massgebliche Schritte beim konstruktiven Forschungsansatz identifiziert:

Schritt 1: Identifikation eines praktischen Problems mit akademischem Potenzial Die in einem konstruktiven Forschungsprojekt vorgesehene Problemlösung sollte für Manager oder andere Entscheidungsträger im Unternehmen unmittelbar relevant sein (Kasanen et al., 1993). Das Problem muss zudem ein *signifikantes theoretisches Potenzial* besitzen. Es sollte also keine offensichtliche Lösung in der Literatur zu finden sein (Lukka, 2000).

Die konstruktive Forschung unterscheidet sich von der Beratung, die sich ausschliesslich am praktischen Nutzen orientiert, während sich Erstere beiden Ziele

verschreibt: Augenblickliche praktische Fortschritte und ein direkter theoretisch-wissenschaftlicher Beitrag sind die beiden zentralen Ziele der konstruktiven Forschung (Labro & Tuomela, 2003).

Schritt 2: Evaluation des Potenzials für eine längerfristige Forschungskooperation mit einer Zielorganisation Des Weiteren wird mit dem *CRA* das Ziel verfolgt, langfristig mit Unternehmen zu kooperieren. Die Sicherung und Pflege eines Commitments sind von zentraler Bedeutung (Lukka, 2000). Eine monetäre Verrechnung der erbrachten Leistungen kann sich dabei als problematisch erweisen (Westbrook, 1995).

Zur Vermeidung von Abhängigkeiten einer Person oder Führungskraft (Kasurinen, 2002) sollten die Verhandlung und die Analyse der Probleme mit Personen aus verschiedenen Funktionsbereichen und hierarchischen Ebenen geführt werden.

Das Potenzial für eine Kooperation ist aber nicht nur von der Frage abhängig, ob die Manager im Zielunternehmen das Vorhaben unterstützen. Vielmehr sollte im Falle einer angestrebten gemeinsamen Entwicklung der Wertbeitrag aller Teilnehmer detailliert analysiert werden. Wenn der Wertbeitrag des Forschenden in starkem Kontrast zu dem des Managements steht, sollte auf eine Kooperation verzichtet werden (Puolamäki, 2000). Immer sollten der Wert des Forschenden und der des Managements explizit gemacht werden, um die Kredibilität der konstruktiven Forschung zu garantieren (Mattessich, 1995).

Abschliessend muss der Beitrag des Forschenden einem grösseren Publikum zugänglich gemacht werden, um sicherzustellen, dass die Forschung ihren Zweck für die Praxis und die Wissenschaft erfüllt sowie weitere Forschung initiiert (Kaplan, 1998). Im Kontext des konstruktiven Forschungsansatzes wird auf die Wichtigkeit des Commitments hingewiesen, dass die Forschung publiziert werden darf (Lukka, 2000). So sollten sämtliche Problemstellungen mit der Publikation der Ergebnisse, die später während der Forschungsarbeit auftauchen könnten, bereits in diesem Schritt adressiert werden.

Schritt 3: Ein generelles und vertieftes Verständnis für das Themas erlangen Die theoretische und praktische Untermauerung des Themas charakterisiert den dritten Schritt des konstruktiven Forschungsprozesses (Lukka, 2000).

Er unterscheidet sich von der Beratung, weil der Forscher für informierte Interventionen während der Forschungsarbeit mehr theoretisches Vorwissen benötigt (Dickens & Watkins, 1999).

Obwohl die Innovationsphase sehr von einer Heuristik geprägt sein kann (Kasanen et al., 1993), wird eine umfassende Kenntnis über vorhandene Theorien benötigt, um theoretische Beiträge der Studie identifizieren und analysieren zu können.

Mit Hinblick auf die empirische Arbeit ähnelt diese Phase anderen Methodologien zu Fallstudien, bei denen ein tiefes Verständnis mithilfe von Beobachtungen, Interviews und dem Studium von Archivdaten erlangt wird (Lukka, 2000). Unerlässlich erweist sich die Analyse von Validitäts- und Verlässlichkeitsproblemen, weil Konklusionen des CRA teilweise auf diesem Material beruhen (siehe Schritt 7). Klar sein muss, dass die Untersuchung eines sich verändernden Studienobjekts – also der Organisation – mit einem einzigen Instrument – dem Forscher – erfolgt, wobei sich Validitäts- und Verlässlichkeitsprobleme vermischen. Wesentlich ist die Einsicht, vier mögliche Risiken für die Verlässlichkeit und Validität der Feldforschung zu berücksichtigen: beobachterinduzierter Effekt, Beobachterverzerrung, Datenlimitierungsfaktor sowie Komplexität und Limitierung des menschlichen Verstands. McKinnon (1988) schlagen drei komplementäre Strategien vor, um diese vier Gefahren zu überwinden. So sollten genügend Zeit für die Fallstudien eingeplant werden, mehrere Methoden und Beobachtungsverfahren zur Anwendung kommen und ein angemessenes soziales Verhalten gezeigt werden.

Labro und Tuomela (2003) sehen fünf kritische Faktoren, die während der Vorbereitungsphase berücksichtigt werden müssen:

1. Sicherstellung, dass das untersuchte Problem von mehreren Personen innerhalb der Organisation, in der die Fallstudie durchgeführt wird, anerkannt ist. Zugleich sollten diese Personen motiviert sein, das Problem zu lösen.

2. Versichern, dass die Organisation, in der die Fallstudie durchgeführt wird, genügend Ressourcen zur Verfügung stellen kann – insbesondere Zeit des Managements und der weiteren Mitarbeiter.

3. Die evaluierten Wertvorstellungen des Managements sollten nicht grundlegend von denen des Forschenden abweichen.

4. Der Fallstudienorganisation genügend Informationen zum Forschungsansatz zur Verfügung stellen, damit die fundamentalen Ziele und Eigenheiten allgemein bekannt sind.

5. Verständigung auf die Publikation der erhobenen Erkenntnisse.

Schritt 4: Konstruktion einer innovativen und theoretisch verankerten Idee Die intensive Suche nach einer innovativen Lösung – mit theoretischen und praktischen Implikationen – ist das herausragende Merkmal des konstruktiven Forschungsansatzes. Dieses Merkmal unterscheidet ihn von den anderen Formen der Aktionsforschung sowie von einem Grossteil der Beratungsarbeit. Daher entspricht ein einfaches Modell, das einem Lehrbuch entnommen wurde, nicht den Kriterien der konstruktiven Forschung, obwohl diese Sichtweise möglicherweise neu für die Zielorganisation sein kann. Die Innovationsphase ist eine kreative und möglicherweise heuristische Abfolge von Ereignissen, für die kooperatives Teamwork zwischen dem Forscher und dem Management des Unternehmens eminent wichtig ist (Lukka, 2000).

Schritt 5: Implementierung der Lösung und Test in der Praxis Die konstruktive Forschung folgt einem pragmatischen Ansatz der Wahrheitsfindung. Aus diesem Grund nimmt die Implementierungsphase eine herausragende Rolle für den Forschungsprozess ein (Lukka, 2000). Ausserdem ist die Implementierung wichtig, weil auch ein fehlgeschlagener Test wichtige theoretische Ergebnisse liefern kann (Lukka & Tuomela, 1998). Die Implementierung erhält vor diesem Hintergrund eine zweifache Bedeutung: Eine erfolgreiche Implementierung bedeutet, dass der Forschungsprozess erfolgreich war – zumindest im Hinblick auf die wichtigsten Faktoren – und das Konstrukt technisch realisierbar ist (Lukka, 2000).

Schritt 6: Evaluation des Geltungsbereichs der Lösung Für den Abschluss des Forschungsprozesses sieht die Methodologie vor, dass sich der Forscher von den empirischen Details löst und sich mit weiterführenden Implikationen, wie mit der externen Validität, befasst. In diesem sechsten Schritt sollte folglich eine Diskussion stattfinden über die Aspekte des Konstrukts, die sich auf andere Unternehmen übertragen lassen. Für den Fall, dass die Implementierung fehlschlägt, ist

es möglich, dass die Implementierungsprobleme auch in anderen Organisationen auftreten (Lukka, 2000).

Kasanen et al. (1993) legen eine marktbasierte Validierung des Konstrukts nahe. Ein schwacher Markttest gilt als bestanden, wenn ein Manager dazu bereit ist, das Konstrukt in die eigenen Entscheidungsprozesse einzubeziehen. Lukka (2000, 2002) geht weiter und spezifiziert für den schwachen Markttest, dass er erst nach einem tatsächlichen Einsatz in der Praxis als bestanden gilt – und nicht bereits bei seiner beabsichtigten Berücksichtigung. Der halbstrenge Markttest ist bestanden, wenn das Konstrukt eine weitere Verbreitung in der Organisation findet, während der strenge Markttest es erfordert, dass das Konstrukt systematisch bessere Resultate hervorbringt, als es ohne seinen Einsatz der Fall gewesen wäre. Der dritte Test gilt als äusserst streng, weil bereits für weit verbreitete Techniken wie das *Activity Based Costing (ABC)*, *Just-in-time* und *Total-Quality* Probleme für die Quantifizierung der finanziellen Vorteile bei einer Einführung vorliegen.

Es wurde dargelegt, dass es für Fallstudien sehr schwierig ist, über einen schwachen Markttest hinaus zu gelangen. Die Problemstellung besteht darin, zu identifizieren, ob die Organisation das Konstrukt anwendet oder nicht (Lukka, 2000). Nichtsdestotrotz kann bzw. muss eine Überprüfung der Transferierbarkeit des Konstrukts in einem eingeschränkten Rahmen gegeben sein, obwohl ein abschliessendes Urteil darüber beim Adressaten verbleibt, der ein Spezialist seiner spezifischen Umstände ist. Mithilfe einer genauen Beschreibung des Forschungsprozesses und des dazugehörigen Kontextes können dem Adressaten wichtige Anregungen für den Beurteilsprozess, beispielsweise in Bezug auf die Transferierbarkeit, gegeben werden (Lukka and Kasenen, 1995).

Schritt 7: Aufzeigen einer theoretischen Verbindung und Beitrag der Forschungsarbeit für die Problemlösung Nach Lukka (2000) gibt es grundsätzlich zwei Hauptwege, um mithilfe des konstruktiven Forschungsansatzes neue Erkenntnisse für die Theorie beizusteuern.

Zunächst ist es möglich, dass das Konstrukt selbst neu ist, sodass die Einführung die Bekanntheit eines komplett neuen Ansatzes für die Erreichung von Zielen erlaubt. Dies entspricht der fundamentalen Zielsetzung angewandter Forschung (Mattessich, 1995).

Ausserdem kann die konstruktive Forschung helfen, eine Theorie zu entwickeln, zu illustrieren, zu verfeinern oder zu testen (Keating, 1995). Mithilfe des pragmatischen Implementierungstests werden die darunterliegenden positiven Beziehungen gleichfalls getestet. Damit ist sowohl eine erfolgreiche als auch eine fehlgeschlagene Implementierung möglicherweise von Interesse. Solche Fehlschläge können zu Verfeinerungen oder gar zur Zurückweisung von alten Theorien führen oder neue Theorien begründen (Lukka, 2000, 2002).

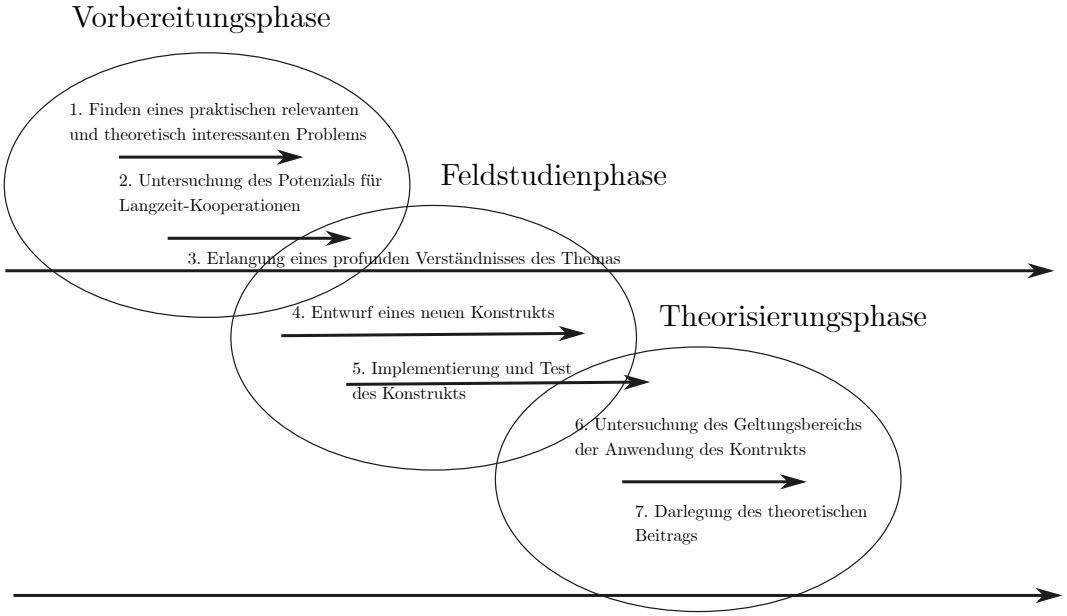


Abbildung 21: Phasen im *konstruktiven Forschungsansatz* nach Labro und Tuomela (2003)

Abbildung 21 zeigt die Phasen (Vorbereitungs-, Fallstudien- und Theorisierungsphase) und die dazugehörenden sieben Schritte in ihrer zeitlichen Abfolge. Die Darlegung des theoretischen Beitrags und die Erlangung eines Verständnisses für das Thema erfassen den gesamten Forschungsprozess, während die weiteren Schritte für sich genommen abgeschlossene Teilabschnitte darstellen.

4 Fallstudien

4.1 Gewinnung des Studienpartners

Die praktischen Arbeiten an den beiden Fallstudien begannen im Februar 2017. Zunächst war es notwendig, eine Forschungskooperation mit einer Prüfgesellschaft aufzubauen.

Zu diesem Zweck wurden grosse und mittelgrosse Wirtschaftsprüfungsgesellschaften der Schweiz mithilfe direkter Kontakte oder elektronischer Kontaktaufnahme über die Website angeschrieben und über Ziel, Inhalte und Herangehensweise der Forschungsarbeit informiert. Vier angeschriebene Unternehmen zeigten unmittelbar Interesse für eine Teilnahme an Fallstudien.

Mit den vier Unternehmen wurde daraufhin der Dialog aufgenommen, sodass ein detaillierter Austausch möglich war. In diesen Gesprächen, die der Studienleiter mit jeweils einem Partner und einem (Senior-)Manager führte, wurden neben den Eckdaten der Fallstudien, den Zielen, dem Ablauf, personellen Fragen und dem wissenschaftlichen Beitrag auch Datenschutzfragen, der Nutzen für die Prüfgesellschaft und die Prüfkunden sowie die interne redaktionelle Verarbeitung besprochen.

Von den vier Unternehmen, bei denen die Fallstudie teilweise mehrmals vorgestellt wurde, blieb eine Gesellschaft übrig, mit der eine Zusammenarbeit vereinbart wurde. Die drei anderen Unternehmen hatten Vorbehalte im Hinblick auf den Datenschutz sowie den Einbezug einer Drittpartei (die Prüfkunden) und entschieden sich aus diesem Grund gegen eine Teilnahme.

4.1.1 Aufnahme Forschungskooperation

Der erste Kontakt mit dem *Studienpartner* fand mit einem Mitarbeiter auf der Managerstufe statt und umfasste eine Erläuterung des möglichen Ablaufs der Studie, der Ansprüche an die potenziellen Prüfkunden, der Ziele, der möglichen Methoden sowie der dieser Arbeit zugrunde liegenden Thesen. Dabei war für den Manager die Möglichkeit von speziellem Interesse, lediglich aufgrund der Daten eines Kunden Einsichten zu generieren bzw. Teile der Prüfung zu automatisieren. Aufgrund des Gesprächs trug dieser Manager die Studie sodann ins Führungsteam der Gesell-

schaft.

Die Projektpräsentation am Hauptsitz der Prüfgesellschaft fand danach später im Jahr statt. Das Projekt stiess dabei beim leitenden IT-Partner, welcher für die IT-unterstützte Wirtschaftsprüfung und die IT-Infrastruktur verantwortlich zeichnet, sowie dem operativen Leiter der IT auf grosses Interesse. Ihnen wurden zunächst die Ziele der Fallstudie sowie ihre angewandte Forschungsmethodologie nähergebracht, um im Anschluss Fragen im Zusammenhang mit maschinellem Lernen zu diskutieren. Zudem wurden mit ihnen mögliche Problemstellungen – wie die Sicherheit der Daten des Mandats – analysiert sowie eventuelle Sicherheitskonzepte und der Verlass auf die Infrastruktur der Prüfgesellschaft als Ausweg erörtert.

Wenige Wochen später bestätigte die Prüfgesellschaft dem Studienleiter, dass eine Kooperation vonseiten der Prüfgesellschaft gewünscht wird und dass nun mit der Planung, der Ausarbeitung eines Non-Disclousure-Agreements (NDA) sowie der Abstimmung mit Prüfteams begonnen werden kann. Zu diesem Zwecke wandte sich der Leiter der IT-Prüfung an sämtliche Schweizer Standorte mit der Frage, ob sie über ein geeignetes Mandat oder gar Mandate für eine Kooperation verfügten – der Studienleiter wurde dabei stets in den E-Mail-Verkehr einbezogen.

4.1.2 Eckdaten der Zusammenarbeit

Gleichzeitig mit der Suche nach Prüfteams wurde dem juristischen Dienst der Prüfgesellschaft ein Standard-NDA der Universität St. Gallen übergegeben. Punktuell fügte die Rechtsabteilung Ergänzungen an, wobei nachfolgend die materiellsten aufgeführt sind:

- Keine Daten dürfen je die Infrastruktur der Prüfgesellschaft verlassen.
- Eine Konventionalstrafe von 10'000 Schweizer Franken bei Übertretung.
- Anonymität der Prüfgesellschaft und des Prüfkunden.
- Der Name der Prüfgesellschaft wie auch der Prüfkunden dürfen nur mit Einverständnis der jeweiligen Gesellschaften genannt werden. Ansonsten sind sämtliche Daten und Erkenntnisse zu anonymisieren.

Das NDA wurde schliesslich wenige Monate später unterzeichnet und der Empfang beim juristischen Dienst einen Tag später bestätigt.

Kurze Zeit danach fand sich zudem eine Vertretung, die geeignete Mandate, Interesse und ein verfügbares Team anmeldete.

4.1.3 Mandatsauswahl mit dem Team

In einem späteren Gespräch mit den Verantwortlichen der Vertretung ging es darum, mögliche Ansprüche an einen Prüfkunden aus Sicht des Studienleiters zu definieren.

Die Prüfgesellschaft verfügte über ein breites Portfolio von Prüfmandaten, die von öffentlich-rechtlichen Anstalten bis hin zu international tätigen Unternehmen mit Umsätzen zwischen 10'000 und 100 Mio. Schweizer Franken reichten. Der Studienleiter erklärte den Verantwortlichen, dass ein optimaler Prüfkunde zum Zwecke der Fallstudien sich insbesondere durch eine hohe Anzahl an Transaktionsdaten und gute Datenverfügbarkeit auszeichnete. Derartige Transaktionen würden auf Konten erfolgen, deren Höhe und zeitlicher Anfall weitgehend ausserhalb des Einflusses des Prüfkunden stünden.

Daneben war die Frage wichtig, wie sich ein Prüfkunde für ein derartiges Projekt gewinnen liesse, der von der Fallstudie nicht kurzfristig profitierte. Vom Studienleiter wurde das Argument angeführt, dass der Kunde gegebenenfalls neue Einsichten in die eigenen Daten erhielte und ihm neue Methoden der modernen Prüfung vermittelt werden könnten.

Die Sitzung wurde mit einer generellen Interessensbekundung vonseiten der Prüfgesellschaft bzw. der Vertretung beendet. Wenig später konnte die Prüfgesellschaft zwei Mandate identifizieren, die die Voraussetzungen für eine Teilnahme an der Studie erfüllten. Es handelte sich um eine international tätige Unternehmung in der Güterproduktion und um eine grosse öffentlich-rechtliche Betriebsstätte, die neben der administrativen Tätigkeit auch ein Werk unterhält.

4.1.4 Prüfgesellschaft

Bei der Prüfgesellschaft handelt es sich um eine mittelgrosse Schweizer Wirtschaftsprüfungs-, Treuhand- und Beratungsgesellschaft, die einem internationalen

Netzwerk angehört.

4.2 Vorgehen in den Fallstudien

Entsprechend der von Gierbl (2021) aufgezeigten fünf Prüfgebiete, für die der Einsatz von maschinellem Lernen in der Literatur beschrieben wird, ergibt sich folgender Studienablauf (siehe Abbildung 22).

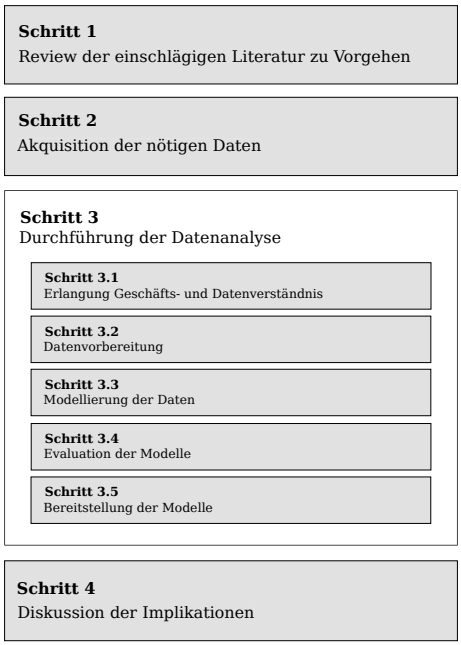


Abbildung 22: Vorgehen in der Fallstudie (eigene Darstellung)

Das Vorgehen in den Fallstudien lehnt sich, ergänzt um einen Schritt *Akquise der benötigten Daten* (Schritt zwei), an CRISP-DM an und wird umrahmt durch einen vertieften theoretischen Exkurs über anzuwendende Methoden in Schritt eins sowie eine abschliessende Diskussion über die Implikationen für die Wirtschaftsprüfung in Schritt vier.

4.2.1 Review der einschlägigen Literatur

Für jedes der fünf *Prüfgebiete* müssen im ersten Schritt die in der Literatur erfolgreich getesteten Methoden identifiziert und diskutiert werden. Zu diesem Zweck wird auf die von Gierbl (2021) analysierten und einem Peer-Review unterzogenen

Methoden zurückgegriffen. Bei der Auswahl der konkreten Methoden soll eine breite Palette an verschiedenen Ansätzen und verwendeter *Features* erreicht werden, wobei ein Rückgriff auf bereits vorhandene Methoden mögliche methodologische Fehler verhindert.

4.2.2 Akquisition der benötigten Daten

Nach Identifizierung der geeigneten Ansätze werden im zweiten Schritt die für die Modellierung und Prognose benötigten Daten beschafft. Hierbei kann es sich um unternehmensinterne Daten handeln, die beim Revisionsmandanten erhoben werden (z.B. Umsatzzahlen, Buchungsdaten, Mitarbeiterzahlen aus Buchhaltungs-, ERP- und weiteren Systemen). Ausserdem müssen für die Modellierung potenziell auch unternehmensexterne Daten wie Wechselkurse, Konjunkturdaten und mitunter sogar Witterungsdaten akquiriert werden.

4.2.3 Anwendung der ML-Methoden

Auf der Grundlage der theoretisch fundierten Ansätze aus Schritt eins, die mit den erhobenen Daten aus Schritt zwei kombiniert werden, können im nächsten Schritt die Analyse und Modellierung der Daten beginnen. Diese Analyse stützt sich in ihrem Ablauf auf das CRISP-DM-Modell (Wirth & Hipp, 2000) und umfasst folgende Aspekte:

- Erlangung des Geschäfts- und Datenverständnisses
- Datenvorbereitung
- Modellierung
- Evaluation
- Bereitstellung

Es muss beachtet werden, dass es sich beim CRISP-DM-Modell nicht um einen strikt linearen, vorwärts gerichteten Prozess handelt, sondern grundsätzlich um ein iteratives Modell, das die Möglichkeit umfasst, dass nach dem Abschluss der Modellierung aufgrund von praxisrelevanten Einsichten eine neue Iteration angestossen werden kann.

4.2.4 Diskussion der Implikationen

Ziel des vierten Schrittes ist es, die Einsichten, die durch den Einsatz des maschinellen Lernens in der Wirtschaftsprüfung für den Prozess der Prüfarbeit aufgedeckt wurden, festzuhalten. Ferner sollen die Herausforderungen in diesem Prozess sowie die Art und Weise, wie diesen begegnet wurde, aufgezeigt werden.

4.3 Fallstudie 1

4.3.1 Rahmenbedingungen

Eckdaten zum Prüfkunden Die erste Fallstudie wurde mit der Prüfgesellschaft und einem mittelgrossen, weltweit tätigen Schweizer Unternehmen der verarbeitenden Industrie durchgeführt. Aufgrund der Grösse mit rund 100 Mio. Schweizer Franken Umsatz und 200 Mitarbeitern ist die etablierte Unternehmung zu einer ordentlichen Revision nach *Art. 727 des Schweizer Obligationenrechts (OR)* verpflichtet.

Die Unternehmung produziert im Bereich Maschinenbau für zwei Segmente:

1. *Maschinenbau:* Für den konzerninternen Gebrauch werden grossindustrielle Produktionsmaschinen hergestellt, die wiederum für die Produktion im Konzernverbund eingesetzt werden können. Daneben bietet die Unternehmung aber auch spezialisierte Produktionsmaschinen für Drittparteien an.

2. *Maschinen- und Apparatebau in Lohnarbeit:* Als Outsourcing-Partner bietet die Gesellschaft in einer zweiten Sparte den Maschinen- und Apparatebau in Lohnarbeit an.

Vor rund 20 Jahren wurde die Unternehmung von einer US-amerikanischen Grossunternehmung aufgekauft und in deren Konzernstrukturen eingebunden. Die Muttergesellschaft ist dabei in den Vereinigten Staaten am *New York Stockexchange (NYSE)* kotiert und erstellt den Jahresabschluss nach den US-amerikanischen Rechnungslegungsstandards *(US-GAAP)* und den Auflagen der PCAOB für die Sicherstellung funktionierender Kontrollen nach dem *Sarbanes-Oxley Act*.

Folglich erstellt die Unternehmung anlässlich des Jahresabschlusses ein lokales Reporting *(Reporting Package)* und quartalsweise Zwischenabschlüsse. Als eine in

der Schweiz ansässige Gesellschaft muss die Unternehmung einen Steuerabschluss nach Schweizer Obligationenrecht vorlegen.

Prüfteam Für ein Mandat in dieser Grössenordnung stellte die Prüfungsgesellschaft ein aus mehreren Prüfern bestehendes Team zusammen:

- *Mandatsleiter bzw. -partner:* Aufgabe des Mandatsleiter war es, an der Gestaltung der Kundenbeziehung, in der Risikobeurteilung und der damit zusammenhängenden Ressourcenallokation mitzuwirken und zudem die Qualität der Prüfungsdurchführung sicherzustellen.
- *Mandatsmanager:* Der Mandatsmanager, der zugleich als Revisionsexperte tätig war, verantwortete die Detailplanung und spezielle Fragestellungen in der Rechnungslegung sowie im Prüfansatz.
- *Mandats-Incharge:* Beim Mandats-Incharge handelte sich um einen diplomierten Wirtschaftsprüfer, der die Prüfung vor Ort operativ leitete, Rückfragen der Assistenten und des Kunden beantwortete sowie gleichzeitig herausfordernde Prüfaufgaben übernahm.
- *Mandats-Assistenten:* Zwei Mandatsassistenten waren in die Mandatsabwicklung eingebunden, deren Aufgabe es war, die Prüfung der weniger komplexen Sachverhalte durchzuführen.

Diesem Team durfte sich der Studienleiter während der Prüfarbeiten anschliessen.

Infrastruktur Für die Fallstudie wurde dem Studienleiter von der Prüfgesellschaft ein Windows Notebook der aktuellen Produktgeneration zur Verfügung gestellt. Das Notebook enthielt keine systemtechnischen Einschränkungen, sodass zusätzliche Software frei installiert werden konnte.

Zudem erhielt der Studienleiter einen Zugang zur internen Plattform, die dem Datenaustausch zwischen der Prüfgesellschaft und dem Mandanten diente.

Als Analyseplattform zur Umsetzung der nötigen Transformations- und Modellierungsschritte wurde initial ein Softwarestack, aufbauend auf *Python*, ausgewählt. Er umfasste:

- *Jupyter Notebooks:* zur Programmierung und gleichzeitigen Dokumentation der Programmierschritte.
- *Scikit-Learn:* elaubte, verschiedenste Modelle zu berechnen.
- *Pandas:* ermöglichte Datentransformationen und -integration sowie einfache univariate Datenanalysen.
- *Matplotlib:* zur Visualisierung der Datenreihen.

Ausschlaggebend für die Wahl dieses *Stack* war die grosse Verbreitung von *Python*, welches gleichzeitig ein Open-Source-Projekt ist, das eine lokale Installation – also direkt auf dem PC – erlaubt.

Prüfungsdurchführung Die Prüfung der Gesellschaft vor Ort war für zwei Zeitpunkte vorgesehen:

1. *Zwischenrevision:* Es handelt sich bei der ersten Unternehmung um eine grössere Gesellschaft, die ein Teil eines global tätigen Konzerns ist. Aus diesem Grund führt die Prüfgesellschaft jährlich eine Zwischenprüfung durch. Während dieser Zwischenprüfung kann die Prüfgesellschaft bereits einen Teil des Prüfauftrags abschliessen: So prüft sie beim Mandanten das Vorhandensein eines internen Kontrollsystems. Daneben sammelt sie Prüfnachweise, die für eine Aktualisierung der Risikoanalyse wichtig sind, sie prüft aber auch bereits Belege, die für die Schlussrevision bzw. die abschliessende Beurteilung des Abschlusses herangezogen werden können. Die Prüfevidenz bezieht sich beispielsweise auf Belege für die Anschaffung neuer Maschinen und auf die Funktionsfähigkeit der Kontrollen hinsichtlich der Effektivität.

   Anlässlich der Zwischenrevision fanden sich im November 2019 das Prüfteam wie auch der Studienleiter ein erstes Mal beim Mandanten ein.

2. *Schlussrevision:* Der zweite Teil der Abschlussprüfung fand schliesslich im April 2020 statt. Zu diesem Zeitpunkt wurde der Jahresabschluss massgeblich substantiv geprüft und ein Urteil dazu gebildet.

4.3.2 ML-unterstützte Risikobeurteilung

Für die Risikoanalyse werden Ansätze des maschinellen Lernens eingesetzt: Es wird versucht, Earnings Management mithilfe *logistischer Regression* und von Clustering zu identifizieren (Dikmen & Küçükkocaoğlu, 2010), während Dbouk und Zaarour (2017) *naive Bayes-Klassifizierer* verwenden. Zudem wurden mit logistischer Regression und erweiterten *Features* potenzielle Buchhaltungsfehler prognostiziert (Dechow et al., 2011).

Ansätze in der Literatur In einem ersten Modell zur Prognose von *Earnings Management* wird auf bekannte unabhängige Variablen bzw. *Features* von Beneish (1997) zurückgegriffen und ein logistisches Regressionsmodell durch Dikmen und Küçükkocaoğlu (2010) geschätzt. Folgende *Features* finden Verwendung:

- Leverage Index (LVGI): $\frac{Fremdkapital_t/Gesamtkapital_t}{Fremdkapital_{t-1}/Gesamtkapital_{t-1}}$
- Inventory Index (DINV): $\frac{Lagerwert_t/Umsatz_t}{Lagerwert_{t-1}/Umsatz_{t-1}}$
- Financial Expenses Index (FEI): $\frac{Finanzaufwand_t/Umsatz_t}{Finanzaufwand_{t-1}/Umsatz_{t-1}}$
- Days' Sales in Receivables Index (DSRI): $\frac{Forderungen_t/Umsatz_t}{Forderungen_{t-1}/Umsatz_{t-1}}$
- Gross Margin Index (GMI): $\frac{Umsatz_{t-1}-Umsatzkosten_{t-1}/Umsatz_{t-1}}{Umsatz_t-Umsatzkosten_t/Umsatz_t}$
- Asset Quality Index (AQI): $\frac{1-\frac{Umlaufvermögen_t/Sachanlagevermögen_t}{Gesamtkapital_t}}{1-\frac{Umlaufvermögen_{t-1}/Sachanlagevermögen_{t-1}}{Gesamtkapital_{t-1}}}$
- Sales Growth Index (SGI): $\frac{Umsatz_t}{Umsatz_{t-1}}$
- Depreciation Index (DEPI):
  $\frac{Abschreibungen_{t-1}/(Abschreibungen_{t-1}+Sachanlagevermögen_{t-1})}{Abschreibungen_t/(Abschreibungen_t+Sachanlagevermögen_t)}$
- S&G Expense Index (SGAI): $\frac{(Vertriebsaufwendungen_t+Verwaltungsaufwand_t)/Umsatz_t}{(Verwaltungs-undVertriebsaufwendungen_{t-1})/Umsatz_{t-1}}$
- Leverage Index (LVGI): $\frac{(Langristiges+KurzfristgesFremdkapital_t/Gesamtvermögen_t)}{(Langristiges+KurzfristgesFremdkapital_{t-1}/Gesamtvermögen_{t-1})}$
- Total Accrual to Total Asset Index (TATA) :
  $$\frac{\begin{array}{r}\Delta Umlaufvermögen - \Delta CashundWertschriften \\ -\Delta Kurzfr.Fremdkapital - \Delta Langfr.Fremdkapitals \\ - \Delta Rueckstellungen - Abschreibungen\end{array}}{Gesamtkapital_t}$$

Für die genannten *Features*, die ins Modell aufgenommen werden, kann eine Intuition beigefügt werden. So fängt beispielsweise der Abschreibungsindex abnehmende Abschreibungen ein, die eine Form des Earnings Managements darstellen können.

In ihrem Artikel stellen Dikmen und Küçükkocaoğlu (2010) die geschätzten Parameter für das logistische Modell bereit (siehe Tabelle 3).

Unabhängige Variable	Koeffizient
Days' Sales in Receivable Index (DSRI)	16.4952
Gross Margin Index (GMI)	-13.1584
Total Accrual to Total Asset Index (TATA)	8.3547
Inventory Index (DINV)	8.1473
Leverage Index (LVGI)	-2.5999
Depreciation Index (DEPI)	1.8112
S&G Expense Index (SGAI)	1.6927
Sales Growth Index (SGI)	-1.559
Asset Quality Index(AQI)	1.2698
Financial Expense Index (FEI)	0.6528

Tabelle 3: Geschätzte Koeffizienten für die unabhängigen Variablen nach Dikmen und Küçükkocaoğlu (2010)

Wird anhand der Daten im Abschluss (als Eingabevektor) eine Prognose für eine Organisation berechnet, so bedeutet ein erhöhter resultierender Wert – bei Verrechnung der Indizes mit dem oben angeführten Parameter – eine erhöhte prognostizierte Wahrscheinlichkeit, dass ein Earnings Management stattgefunden hat.

	Erwartete M	Erwartete NM	%	Total %
Tatsächliche M	139	32	81.29	66.99
Tatsächliche NM	375	687	64.69	

Tabelle 4: Konfusionsmatrix zu erwartender und tatsächlich manipulierter sowie nicht manipulierter Abschlüsse nach Dikmen und Küçükkocaoğlu (2010)

Die Verlässlichkeit des Modells publizieren Dikmen und Küçükkocaoğlu (2010) in der für Klassifikationsaufgaben üblichen Form einer Konfusionsmatrix (siehe Tabelle 4). Die Tabelle stellt die Abschlüsse mit Manipulationen (M) und Nichtmanipulationen (NM) den erwarteten bzw. prognostizierten Manipulationen und Nichtmanipulationen gegenüber. Aus der Tabelle lässt sich entnehmen, dass rund

81 % der tatsächlichen Manipulationen durch das Modell korrekt klassifiziert werden, während die tatsächlichen Nichtmanipulationen zu 64 % korrekt identifiziert werden. Aus der Risikoperspektive eines Wirtschaftsprüfers ist der erste Wert – auch *Typ-1-Fehler* genannt – entscheidend, weil damit das Risiko eingefangen wird, im Risikoeinschätzungsverfahren potenziell manipulierte Abschlüsse zu übersehen.

Weitere Autoren untersuchen die Modellierbarkeit von Manipulationen bzw. des Earnings Managements und verwenden die unabhängigen Variablen von Beneish (1997). Anstelle des *logistischen Regressionsmodells* stellen sie auf einen *naiven Bayes-Klassifzierer* ab (Dbouk & Zaarour, 2017). Die Parameter, die das Modell bestimmen, werden im Artikel jedoch nicht offenlegt. Zudem, wie in Tabelle 5 dargelegt, liefert ihr Modell eine Genauigkeit von 87 %. Der Typ-1-Fehler liegt dagegen bei 67 %.

	Erwartete M	Erwartete NM	%	Total %
Tatsächliche M	2	4	33.33	86.86
Tatsächliche NM	1	31	96.88	

Tabelle 5: Konfusionsmatrix zu erwartender und tatsächlich manipulierter sowie nicht manipulierter Abschlüsse nach Dbouk und Zaarour (2017)

Dechow et al. (2011) schlagen drei *logistische Regressionsmodelle* für die Identifikation von wesentlichen Fehldarstellungen in Unternehmensabschlüssen vor, die sich für die Risikoeinschätzung heranziehen lassen. Die unabhängigen Variablen, die zu diesem Zweck ins Modell aufgenommen werden, beziehen sich auf die Qualität der Rechnungsabgrenzung, die Unternehmensleistung (im Sinne der Performance) und berücksichtigen auch nicht finanzielle Kennzahlen.

Im Vergleich zum Modell von Dikmen und Küçükkocaoğlu (2010) werden im vollen Ausbau mit insgesamt 28 Variablen wesentlich mehr Daten berücksichtigt. Daneben sind die Variablen gleichfalls auch aufwendiger zu berechnen, weil verschiedene Variablen wie Rechnungsabgrenzungen mit anderen Unternehmen in derselben Branche in Relation gesetzt werden müssen.

Insgesamt handelt es sich um 28 Variablen, die in unterschiedlichen Kombinationen in drei Modellen verwendet werden. In dem ersten Modell, das lediglich Werte aus den Finanzabschlüssen berücksichtigt, wird folgende Konfusionsmatrix in Tabelle 6 ermittelt.

	Erwartete M	Erwartete NM	%	Total %
Tatsächliche M	339	155	68.6	63.71
Tatsächliche NM	48'282	84'685	63.7	

Tabelle 6: Konfusionsmatrix zu erwartender und tatsächlich manipulierter sowie nicht manipulierter Abschlüsse nach Dechow et al. (2011)

Werden dazu noch die Ausserbilanzwerte sowie nicht finanzielle Variablen – wie beispielsweise eine ausserordentliche Mitarbeiterfluktuation – hinzugezogen, so ergibt sich folgende Leistungsfähigkeit in Tabelle 7.

	Erwartete M	Erwartete NM	%	Total %
Tatsächliche M	305	144	67.9	64.06
Tatsächliche NM	43'996	78'370	64.0	

Tabelle 7: Konfusionsmatrix zu erwartender und tatsächlich manipulierter sowie nicht manipulierter Abschlüsse mit nicht finanziellen Werten nach Dechow et al. (2011)

Dechow et al. (2011) nehmen noch marktbasierte Werte in das Modell auf und erhalten folgende Prognosegüte in Tabelle 8.

	Erwartete M	Erwartete NM	%	Total %
Tatsächliche M	238	116	67.2	63.74
Tatsächliche NM	41'934	46'098	63.7	

Tabelle 8: Konfusionsmatrix zu erwartender und tatsächlich manipulierter sowie nicht manipulierter Abschlüsse mit nicht finanziellen und marktbasierten Werten Dechow et al. (2011)

Erhebung der benötigten Daten Zur Reproduktion der Modelle war es notwendig, Daten zu erheben bzw. zu akquirieren, die die Berechnung von *Features* zuliessen, die einerseits auf unternehmensinternen Daten (z.B. Bilanzpositionen), aber auch auf unternehmensexternen Daten wie Erfolgskennzahlen von Unternehmen in der gleichen Branche beruhten.

Zum Zeitpunkt der Zwischenrevision war die Mandatsplanung noch im Aufbau. Es wurde festgestellt, dass eine Jahresrechnung in den Aggregationenformen, wie

sie Dikmen und Küçükkocaoğlu (2010), Dechow et al. (2011) sowie Dbouk und Zaarour (2017) verwendet haben, noch nicht vorlag.

Vielmehr lag als unternehmensinterne Datengrundlage ein detaillierter Auszug aus der Buchhaltung, inklusive der Vorjahreszahlen, vor. Diese sogenannte Saldenbilanz bot eine Übersicht über sämtliche Bilanz- und Erfolgsrechnungspositionen und wurde vom Prüfteam vorab angefordert.

Dechow et al. (2011) gehen mit ihrem Modell über den Einbezug reiner Finanzgrössen hinaus und verlangen, die Branchenzugehörigkeit und Mitarbeiterfluktuation als weitere Grössen zu erheben. Nach Gesprächen mit der Personalabteilung des Mandanten wurden diese personalbezogenen Daten angefordert und auch die Unternehmung in einer Branchentaxonomie verortet.

Losgelöst von unternehmensinternen Daten mussten zudem die *Features* für eine möglichst grosse Zahl an Unternehmen, die einerseits Manipulationen ihrer Abschlüsse zu verzeichnen hatten, wie auch für solche ohne derartige Vorkommnisse gesammelt werden. Dies war im Rahmen der Zwischenrevision, trotz ausführlicher Recherche, in der vorgesehenen Zeit nicht möglich.

Erlangung des Geschäfts- und Datenverständnisses Im Vergleich zum Datenanalysenansatz nach dem *CRISP-DM* war eine aktive Erlangung des Geschäftsverständnisses nicht mehr notwendig, da dieses Thema bereits mit dem Prüfteam im Rahmen der Besprechung zur Risikobeurteilung aufgenommen worden war bzw. das Wissen in Form von Arbeitsblättern aus den Vorjahren vorlag.

Demgegenüber war die Erlangung eines Datenverständnisses von erheblicher Bedeutung, mussten doch für die korrekte Berechnung der Aggregate Lücken in Form von Abschlussposten geschlossen werden.

Datenvorbereitung Zur Umwandlung der Daten der Saldenbilanz in eine für das Modell verwertbare Form waren folgende Schritte nötig:

1. *Digitalisierung der Daten:* Aufgrund einer zeitweiligen Absenz eines Buchhaltungsmitarbeiters, der eine elektronische Version der Saldenbilanz per Zwischenrevision hätte zur Verfügung stellen können, mussten die rund 300 Hauptbuchkonten manuell in ein Hilfsprogramm *(Caseware)* übertragen werden.

Bestimmte Bilanzbestände wurden jedoch nicht angepasst – wie beispielsweise das Delkredere –, während andere Bestände noch gar nicht gebucht waren, wie die Steuerrückstellungen für das laufende Jahr.

2. *Berechnung der Aggregate:* Während für die Risikoeinschätzung des Mandatsmanagements eine physische Version der Saldenbilanz ausreichend war, musste für die Zwecke der Fallstudie ein Übertrag in ein spezielles Auditprogramm *(Caseware)* durchgeführt werden, das neben zahlreichen weiteren Funktionen erlaubte, einzelne Hauptbuchkonten einem Abschlussposten wie Debitoren oder Sachanlagen zuzuordnen und somit Aggregate zu berechnen, wie sie die Rechnungslegungsstandards als Abschlussposten vorsehen. Darüber hinaus erlaubt die Software, die so erfassten Aggregationen über mehrere Abschlussperioden zu speichern, zu reproduzieren und gegebenenfalls anzupassen.

3. *Approximierung von Abschlussposten:* Für die Schaffung einer Datenbasis, die die Features von Dikmen und Küçükkocaoğlu (2010) sowie Dechow et al. (2011) vorhält, mussten die entsprechenden fehlenden Werte – wie ein Delkredere – approximiert sowie nachgebucht werden, wobei nach den Regeln einer doppelten Buchhaltung eine Anpassung von Delkredere auch eine Anpassung des Unternehmenserfolgs erfordert.

   Die Approximation stützte sich auf Heuristiken und Annahmen, die aus dem Geschäfts- und Umfeldverständnis der Einheit hervorgingen. So wurde für das Delkredere angenommen, dass der stabile Kundenstamm ohne wesentliche Marktveränderungen ein gleichbleibendes Risikoprofil aufweist, weshalb eine lineare Beziehung zwischen Delkredere und Höhe des Debitorenbestands vermutet werden konnte. Gleichzeitig hatte sich die Steuersituation der Unternehmung nicht geändert, d.h., es waren keine neuen steuerrelevanten Arbeitsstätten eingerichtet oder andere spezielle Vorkommnisse registriert worden. Ausserdem blieb die Steuergesetzgebung für die Gesellschaft unverändert. Daher konnte mithilfe des kantonalen Steuerrechners die Rückstellung bzw. der Steueraufwand berechnet und als temporäre Nachbuchungen im *Caseware* vorübergehend erfasst werden.

4. *Export in Analysesoftware:* In einem nächsten Schritt mussten die Daten ins Analyse-Softwarepaket *Jupyter Notebooks* mit den Zusatzkomponenten Pandas, Scikit-Learn und Matplotlib überführt werden. Zu diesem Zweck wurden die Möglichkeiten für einen Export evaluiert. Sehr schnell stand fest, dass das Caseware einen Export in den Formaten CSV oder PDF zulässt. Die Daten wurden entsprechend als CSV exportiert, weil Daten im PDF nur schwierig wieder in ein strukturiertes Datenformat zu konvertieren gewesen wären.

   In einem speziell dafür vorgesehenen Jupyter Notebook für die Risikoanalyse konnten die Daten wieder importiert werden und standen für Analysen zur Verfügung.

5. *Berechnung der unternehmensinternen Features:* Die *Features*, die in die Modelle einfliessen, verwendeten unabhängige Variablen basierend auf abgeleiteten Werten, welche sich aus Bilanz- und Erfolgsrechnungsgrössen berechnen lassen. Daher wurde zur Ermittlung der Werte der verwendeten Features als nächstes die Berechnung der Variablen programmatisch umgesetzt.

Modellierung der Daten Bedingt durch den Umstand, dass die benötigten unternehmensexternen Daten, insbesondere die *Labels* für manipulierte und nicht manipulierte Finanzabschlüsse, nicht verfügbar waren, war es in der Folge unmöglich, während der Zwischenrevision im Rahmen der Risikobeurteilung ein Modell zu schätzen.

Die Tatsache jedoch, dass sowohl Dikmen und Küçükkocaoğlu (2010) als auch Dechow et al. (2011) in ihren Artikeln die Parameter bzw. die Koeffizienten ihres Modells publiziert haben, erlaubte, eine Schätzung für das untersuchte Unternehmen per Zwischenrevision und die Wahrscheinlichkeit einer Manipulation zu berechnen. Dieser Schritt war bei Dbouk und Zaarour (2017) nicht möglich, weil die Parameter des *naiven Bayes-Modells* fehlten. Für die Zwischenrevision galt, dass das Resultat des Modells von Dikmen und Küçükkocaoğlu (2010) eine Nichtmanipulation für die vorliegenden Daten vorhersagt, während das Modell von Dechow et al. (2011) gleichfalls keine materiellen Fehler im Abschluss prognostizierte.

Evaluation der Modelle Wie oben beschrieben, konnte kein Modell für eine grössere Anzahl von Unternehmen trainiert werden, da sich die nötigen Daten nicht innerhalb nützlicher Frist akquirieren liessen. Zur Einschätzung der Leistungsfähigkeit bzw. ihrer Prognosegüte musste auf die in den Artikeln beschriebenen Leistungskennzahlen abgestellt werden.

Für die betrachteten Modelle von Dikmen und Küçükkocaoğlu (2010) und Dechow et al. (2011) wird die Leistungsfähigkeit als vielversprechend eingeschätzt. Die Wahrscheinlichkeit, dass mit den Modellen Unternehmen mit Manipulationen oder Fehlern im Abschluss aufgedeckt werden, liegt bei 82 % für Dikmen und Küçükkocaoğlu (2010) respektive bei 69 % für Dechow et al. (2011).

Dieser Einschätzung muss jedoch entgegengestellt werden, dass die Modelle auf Daten trainiert wurden, die geografisch und zeitlich stark eingegrenzt sind:

- So trainierten Dikmen und Küçükkocaoğlu (2010) ihre Modelle auf Daten, die sich auf türkische Unternehmen im Zeitraum von 1990 bis 1999 bezogen.
- Dechow et al. (2011) stellten für ihre Trainingsdaten auf US-amerikanische Unternehmen zwischen 1997 und 2005 ab.

Diese Kritik bedeutet keinesfalls, dass die Modelle wertlos sind. Vielmehr ist bei Anwendung der Parameter Vorsicht geboten, da die Prognosegüte potenziell von den publizierten Werten abweichen kann. Es bedarf daher der vorherigen Beschaffung der Daten aus Sekundärquellen oder einer Erhebung aufgrund der Daten aus anderen Abschlussprüfungen der Prüfgesellschaften.

Bereitstellung der Modelle In einem letzten Schritt sieht das *CRISP-DM-Modell* die Bereitstellung des Modells im Produktionsbetrieb vor, sodass es einem grösseren Publikum einfach zugänglich wird und ohne grosse Hindernisse genutzt werden kann. Im vorliegenden Fall soll die Bereitstellung ermöglichen, auf einfache Weise eine Risikoeinschätzung in Bezug auf das Earnings Management und die Wahrscheinlichkeit für wesentliche Fehldarstellungen zu erhalten.

Unter der Annahme, dass die Modelle eine Prognosegüte aufweisen, wie sie in der Literatur bei Dikmen und Küçükkocaoğlu (2010) sowie bei Dechow et al. (2011) zu finden ist, wurde die Risiko- bzw. Fehleranfälligkeitsschätzung auf zwei separaten *Jupyter-Notebooks* bereitgestellt.

Beide Notebooks erlauben es, Prognosen zu generieren:

- In einem ersten Schritt wird der Katalog für die benötigten Features bzw. Metriken – beispielsweise Umsatzwachstum oder Lagerindex – berechnet.
- Anschliessend werden diese Metriken als Ausgang für die Prognose im Modell herangezogen.

Damit konnte das Modell dem Prüfteam in pragmatischer Weise zur Verfügung gestellt und einfache Prognosen ermöglicht werden.

Implikationen für Prozesse sowie Herausforderungen

- *Akquisition externer Daten*

  Dem *CRISP-DM-Modell* liegt die Prämisse zugrunde, dass die Daten für die Analyse bereits vorliegen. Diese Bedingung ist während einer Abschlussprüfung jedoch nicht immer gegeben – insbesondere in Bezug auf unternehmensexterne Beobachtungen.

  Im vorliegenden Anwendungsfall werden für das Training der Modelle Datensätze mit manipulierten und nicht manipulierten Abschlüssen (Dikmen & Küçükkocaoğlu, 2010) bzw. mit wesentlichen Fehldarstellungen (Dechow et al., 2011) benötigt.

  Die Beschaffung eines solchen Datensatzes erweist sich während eines einzigen Mandats als schwierig, zeitintensiv und bisweilen sogar unmöglich. Es gilt jedoch zu bedenken, dass ein solcher Datensatz, wenn er erst einmal zusammengetragen wurde, über das eigentliche Mandat hinaus genutzt werden kann. Zudem haben darauf aufbauende Modelle ebenfalls über einen Kunden hinaus Gültigkeit.

  Folglich wäre es unter Effizienzgesichtspunkten sinnvoll, die Sammlung, Aufbereitung und Modellierung von Risikoeinschätzungen zentral durchzuführen und die trainierten Modelle den Wirtschaftsprüfern und ihren Mandanten zur Verfügung zu stellen, für den Fall, dass sich diese Modelle als wertvoll herausstellten.

- *Herausfordernde Aufbereitung interner Daten*

  Die Prüfungsstandards sehen vor, dass der Wirtschaftsprüfer fortwährend seine Risikoeinschätzung anpasst, wenn er neue relevante Sachtatbeständen wahrnimmt. Aus praktischer Sicht ist die Risikoeinschätzung jedoch in der Planung des Mandats von herausragender Bedeutung: Aufgrund des vom Prüfer eingeschätzten Risikos werden die Ressourcen, insbesondere die Grösse und Zusammensetzung des Prüfteams sowie einzubeziehende Spezialisten, geplant und die zeitliche Ausdehnung des Audits festgelegt. Einen wichtigen Zeitpunkt stellt die Zwischenrevision dar, mit der die Abschlussprüfung detailliert geplant wird.

  In der Fallstudie ergab sich für das Prüfteam eine Situation, in der zum Zeitpunkt der Zwischenrevision kein für das Modell verwertbarer Abschluss vorlag, der in operationalisierbarer Form verwendbar gewesen wäre. Stattdessen mussten Abschlussposten – wie das Delkredere – approximiert werden.

  Damit war das Prüfteam bei dieser durch maschinelles Lernen unterstützten Risikoeinschätzung in der Planungsphase einem zusätzlichen und mit Herausforderungen belegten Aufwand ausgesetzt, welcher am Jahresende, mit dem Vorliegen einer Jahresrechnung, nicht entstanden wäre.

4.3.3 ML-unterstützte Aufdeckung doloser Handlungen

Die Mehrzahl der wissenschaftlichen Literatur über die Anwendung des maschinellen Lernens bzw. alternativer Datenanalyse-Techniken im Audit befasst sich mit der Aufdeckung doloser Handlungen (Gierbl, 2021).

Dieser Befund ist erstaunlich, stellt die Identifikation von dolosen Handlungen üblicherweise spezifische Herausforderungen an die Verfügbarkeit von Daten, die die Anwendung von Methoden des maschinellen Lernens erheblich erschweren: Denn die Fälle, die bei der Abschlusserstellung auf dolose Handlungen hinweisen, sind seltener als die Fälle, in denen keine dolosen Handlungen vorliegen. Es wird in diesem Zusammenhang von einem *Haystack-Problem* gesprochen, weil eine Nadel im Heuhaufen gesucht wird (J. L. Perols et al., 2017). Damit ergeben sich Datensätze, die unausgewogen sind und bei denen die Anzahl der Abschlüsse ohne dolose Handlungen im Vergleich zu denen mit dolosen Handlungen weiterhin

überwiegt. Das Vorliegen solcher unausgewogenen Datensätze erfordert die Anwendung spezieller Techniken (J. L. Perols et al., 2017).

Der grossen Anzahl von einschlägigen Artikeln steht ein breites Spektrum an verschiedenen Ansätzen, die zur Anwendung gebracht werden, gegenüber. So finden Techniken des überwachten Lernens wie *Support-Vektor-Maschinen*, *künstliche neuronale Netze*, *logistische Regressionsmodelle* und *naive Bayes-Modelle* ihren Einsatz sowie Ansätze aus dem unüberwachten Lernen wie *Clustering*, die *Warenkorb Analyse* oder *selbstregulierende Karten*..

Ansätze in der Literatur Nach Thiprungsri und Vasarhelyi (2011) eignet sich das *Clustering* für die Identifikation von dolosen Handlungen. Es hat den Vorteil, dass zur Erkennung vorab kein entsprechendes *Label* – eine Klassifizierung von Transaktionen in dolose und bzw. nicht dolose – für jede Beobachtung gegeben sein muss.

Im Artikel definierten die Autoren auf Grundlage der vorhandenen Daten neue unabhängige Variablen als Teil des *Feature Engineerings*. Dieser erste Schritt sollte Datenqualitätsprobleme lösen und die wichtigsten Dimensionen einfangen, in denen sich dolose Handlungen manifestieren. Die auf diese Weise berechneten und abgeleiteten Grössen wurden normalisiert, um einem irreführenden Einfluss von Skalenunterschieden vorzubeugen.

In einem weiteren Schritt wurde das eigentliche Clustering durchgeführt, indem auf das übliche *k-Means Clustering* zurückgegriffen wurde. Zudem wurden unterschiedliche Cluster getestet, bis eine Erhöhung der Clusterzahl das Modell nicht mehr massgeblich verbessert hat.

Für die Identifikation von Anomalien bzw. auffälligen Cluster wurde in zwei Schritten vorgegangen.

- *kleine Wahrscheinlichkeit der Clusterzugehörigkeit*

  Beobachtungen, die nur eine geringe Chance haben, einem Cluster zugerechnet zu werden – im Beispiel der Autoren bei Unterschreitung einer Wahrscheinlichkeit für Clusterzugehörigkeit der Grenze von 0,6 %.

- *kleiner Cluster*

Cluster mit einer sehr geringen Anzahl an Beobachtungen – bei Unterschreitung der Grenze von 1 % der Beobachtungen – werden als Anomalien deklariert.

Unter Anwendung dieses Verfahrens und seiner Grenzwerte konnten im Artikel acht *Cluster* identifiziert werden, von denen drei *Cluster* die Kriterien für eine Anomalie erfüllten. Von den ursprünglich 40'000 Geschäftsvorfällen, die ins *Clustering* miteinbezogen worden waren, wurden 169 einem auffälligen *Cluster* zugerechnet, wobei zusätzlich noch 568 der Beobachtungen nur schwer einem *Cluster* zugeordnet werden konnten (Kriterium 1). Thiprungsri und Vasarhelyi (2011) qualifizierten damit rund 2 % aller Transaktionen als auffällig.

Auch Jans et al. (2010) verwenden einen *Clusteransatz* für die Erkennung von dolosen Handlungen und greifen auf das *Internal Fraud Risk Framework* (IFR; Jans et al., 2009) zurück.

Nach diesem Modell wird im ersten Schritt ein potenziell gefährdeter Unternehmensprozess ausgewählt.

Für den identifizierten Prozess müssen Daten gesammelt, manipuliert und angereichert werden. Die Manipulation umfasst das Organisieren der Daten in eine Struktur und ein Format, die für die Verarbeitung notwendig sind. Die Anreicherung sieht die Generierung von neuen Attributen vor, die aus einer Kombination oder Transformation von bestehenden Attributen resultiert – beispielsweise die Berechnung von Verhältnissen wie Transaktionen pro Mitarbeiter. Es folgt ein dritter Schritt, bei dem die Beobachtungen in Verhaltensdaten umgewandelt werden. Im Vergleich zum vorhergehenden Schritt fliesst bei der Bildung dieser Variablen Domänenwissen mit ein und geht damit über die rein technischen bzw. mechanischen Transformationen hinaus. Ein Beispiel ist die Generierung des neuen Attributs *Anzahl der preisbedingten Änderungen* für einen Einkaufsauftrag (Purchase Order).

Es folgt eine univariate Analyse im vierten Schritt, um einen Einblick in die Verhaltensdaten zu erhalten, gefolgt von einem multivariaten *Clustering*, um damit typische Muster in den Abläufen innerhalb des Prozesses offenzulegen. Es wird davon ausgegangen, dass Transaktionen sich den meisten Mustern zuordnen lassen und abweichende Muster mit relativ wenigen Beobachtungen das Potenzial haben,

dolose Transaktionen darzustellen.

Abschliessend werden Domänenexperten damit betraut, die aufgezeichneten Transaktionen, die in die Gruppe mit abweichenden Mustern fallen, einer genaueren Prüfung zu unterziehen und einer der vier Gruppen zuzuordnen:

- dolose Handlung
- Umgehung von Abläufen
- Fehler oder Irrtümer
- Extremwerte

Für ihren Artikel extrahierten Jans et al. (2010) *Purchase Orders* als Abzug im txt-Format aus SAP und behielten diejenigen Purchase Orders, die in einer Rechnung mündeten und gleichfalls ein gewisses Alter noch nicht überschritten hatten. Sie erhielten damit eine Grundgesamtheit von etwa 33'000 Einträgen, die nach ihrer Beschaffung transformiert wurden, indem beispielsweise neue Attribute wie *Anzahl der Mutationen* der Purchase Order gebildet wurden. Auf der Grundlage von Domänenwissen leiteten die Autoren anschliessend derivative Attribute ab, die den Lebenszyklus einer Bestellung widerspiegeln. Diese Attribute bestanden aus dem Namen des Buchhalters, dem Namen des Lieferanten, der Bestellgruppe, dem Bestelltyp, der Anzahl der Mutationen, der Anzahl der Mutationen nach der letzten Freigabe und der Anzahl der Preisanpassungen nach der letzten Freigabe.

Mit Ende der Aufbereitung wurden die Daten mithilfe eines *latenten Klassenclustering (LCC)* in mehrere Gruppen aufgespalten. Beim LCC handelt sich um einen alternativen Ansatz, der den Autoren zufolge direkt Wahrscheinlichkeiten für Transaktionen und ihre Clusterzugehörigkeiten mitgibt, mit unterschiedlichen Skalen in den Variablen (kategoriale und stetige Skalen) umgehen kann und zudem ein Informationskriterium zur Bestimmung der Clusterzahl nutzt. Damit erhielten die Autoren drei verschiedene Cluster:

- *Cluster 1* umfasste 76.6 % der Purchase Orders.
- *Cluster 2* umfasste 22.1 % der Purchase Orders.
- *Cluster 3* hatte einen Anteil von 1.25 % der Purchase Orders.

Aus quantitativen Erwägungen wurde *Cluster 3* analysiert und die in ihm enthaltenen Transaktionen von Domänenexperten geprüft. Es stellte sich heraus, dass es sich bei neun Bestellungen um Umgehungen der Prozesse handelte, während eine fehlerhafte Bestellung aufgedeckt werden konnte.

Ein weiterer Ansatz des unüberwachten Lernens wird von Tackett (2013) angewandt. Der Autor macht sich die *Assozierungsanalyse* zunutze, um dolose Handlungen aufzudecken.

Mit der Assoziationsanalyse wird üblicherweise ermittelt, was Konsumenten zusammen kaufen. Diese Methode wird daher auch als *Warenkorbanalyse* bezeichnet. Im Vergleich zu traditionell statistischen Analysen bietet sie den Vorteil, dass sie simultan mehrere Variablen und individuelle Kombinationen miteinbeziehen kann, um eine Prognose durchzuführen.

Das Resultat der Assoziationsanalyse sind eine Menge an Regeln, die als Wenn-Dann-Aussagen mit den Wahrscheinlichkeiten ihres Auftretens strukturiert sind. Das Beispiel für eine Warenkorbanalyse in einem Supermarkt verdeutlicht diesen Zusammenhang: Wenn ein Kunde Brot und Butter kauft, besteht eine Wahrscheinlichkeit von 90 %, dass er auch Milch kauft.

Tackett (2013) geht es darum, mithilfe dieser Analyse dolose Handlungen aufzudecken, indem er in einem ersten Schritt eine Schlüsselvariable identifiziert. Dolose Handlungen ereignen sich jedoch im Verborgenen, daher steht nicht direkt die dolose Transaktion selbst im Mittelpunkt des Interesses, sondern vielmehr Transaktionen oder Ereignisse, die typischerweise mit dolosen Handlungen einhergehen. Im Beispiel von Tackett (2013) werden Rückerstattungen als eine mögliche Variable mit Potenzial für dolose Handlungen erkannt. Diesbezüglich stellt sich die Frage, ob es Ereignisse – oder eine Kombination von Ereignissen – gibt, die eine übliche Rückerstattung auszeichnen und genutzt werden können, um damit Transaktionen zu identifizieren, die eingehend analysiert werden müssen.

Die *Wenn-Komponente* wird in der Assoziationsanalyse auch als Prämisse bezeichnet und die *Dann-Komponente* als Folgerung – im Butter-Brot-Beispiel sind die Prämissen das Brot und die Butter, während die Milch die Folgerung aus dem Kauf der beiden Produkte ist. Als weitere Komponente zeigt der *Support* das Vorkommnis eines bestimmten Musters im Datensatz an. Häufig wird dieser dazu verwendet, um sehr seltene Muster auszuschliessen. Die *Konfidenz* eines Musters

gibt an, mit welcher Wahrscheinlichkeit die Folgerung tatsächlich auf die Prämisse folgt. Der zuletzt genannten Variable, dem *Lift*, wird schliesslich die grösste Bedeutung beigemessen. Er ähnelt der Korrelation und misst, wie viel wahrscheinlicher es ist, das Muster bei gegebener Prämisse und der Folgerung im Datensatz zu finden, wenn zufällig ein Eintrag mit der Prämisse gewählt würde. Damit ist diese Variable für die Auswertungszwecke besonders von Interesse.

Die beschriebenen Analyseschritte hat der Autor für Rückerstattungen bei einem US-amerikanischen Detailhandelswarengeschäft praktisch durchgeführt. Dazu wurde ein Datensatz aufbereitet, der Beobachtungen zu folgenden Variablen enthielt:

- Identifikation der Filiale bzw. des Geschäfts
- Abteilung
- Identifikation des Produkts
- Typ der Transaktion (Verkauf/Rückerstattung)
- Bezahlungsart(bar/Kreditkarte)
- Betrag
- Mitarbeiteridentifikation

Vor der Verarbeitung der Daten im Analysesoftwarepaket mussten die Beträge in kategoriale Variablen überführt werden, weil die Assoziationsanalyse nur mit derartigen Daten umgehen kann. Als interessante Variable wurde der Transaktionstyp ausgewählt. Ausserdem war ein minimaler Support von 1 und 5 % als Minimumkonfidenz vorgesehen.

Mit einem entsprechenden Analysesoftwarepaket ausgestattet, wurden durch Wirtschaftsprüfer die Transaktionen der Top-Zehn-Mitarbeiter – gemessen am Lift – untersucht. Bei acht Mitarbeitern konnten dolose Handlungen nachgewiesen werden.

Im Bereich der überwachten maschinellen Lernansätze schlagen J. L. Perols et al. (2017) die Verwendung von Support-Vektor-Maschinen vor, die auf speziell vorverarbeitete Daten angewandt werden, um die Prognosegüte zu maximieren.

Dem Vorgehen der Autoren folgend werden die Daten in zwei Schritten aufbereitet:

- *Subsampling nicht doloser Transaktionen*

  Die Autoren verwenden ein Multi-Subset-Beobachtungs-Untersampling nach Chan und Stolfo (1998), um dem Problem der Rarität zu begegnen, d.h. dem Umstand Rechnung zu tragen, dass die Transaktionen mit dolosen Handlungen selten auftreten bzw. aufgedeckt werden. Das Verfahren sieht vor, dass die Daten in mehrere Untergruppen aufgeteilt werden, wobei jede Untergruppe sämtlichen Transaktionen mit dolosen Handlungen entspricht. Gleichzeitig werden aus dem Pool der nicht dolosen Handlungen für jede Untergruppe Fälle ausgewählt, sodass eine ausgewogene Gruppe entsteht.

  Für jede Gruppe wird ein Modell berechnet, um die Wahrscheinlichkeit von dolosen Handlungen prognostizieren zu können. Für die tatsächliche Vorhersage werden die Prognosen sämtlicher Modelle zusammengeführt.

- *Multi-Subset-Variablen-Untersampling*

  Die Unterrepräsentation von Fällen mit dolosen Handlungen ist zusätzlich mit dem *Fluch der Dimensionalität* behaftet, der nach Bellman (1961) dazu führt, dass die Grösse des Datensatzes exponentiell mit der Anzahl der unabhängigen Variablen steigen muss. Dabei ist die Anzahl der in der Forschung zu dolosen Handlungen vorgeschlagenen *Features* sehr gross, während gleichfalls die Anzahl an Beobachtungen mit dolosen Handlungen relativ gering ist.

  J. L. Perols et al. (2017) schlagen ein neues Verfahren vor, um diesem Problem zu begegnen. Das Verfahren sieht vor, dass die unabhängigen Variablen zufällig und ohne Zurücklegen in Untergruppen eingeteilt werden. Jede dieser Untergruppen enthält die gleiche Anzahl an Beobachtungen, aber dafür mehrere nicht überlappende Gruppen von unabhängigen Variablen. Für jede Untergruppe wird ein Modell erstellt, das die Wahrscheinlichkeit für dolose Handlungen vorhersagen kann, indem sämtliche Modelle für die Wahrscheinlichkeit einer dolosen Handlung kombiniert werden.

Bei der Messung der Performance des erweiterten Ansatzes stützen sich J. L. Perols et al. (2017) auf einen Datensatz, wie ihn Beasley (1996) oder Brazel et al. (2009) verwenden und der neben 51 Unternehmen mit Abschlüssen, die auf dolose Handlungen hinweisen, 15'934 ohne derartige Sachverhalte kennt.

Für die Features greifen J. L. Perols et al. (2017) bei ihrem Modell für dolose Handlungen auf solche zurück, wie sie von J. Perols (2011) bzw. Dechow et al. (2011) zuvor verwendet wurden. Zusätzlich nehmen sie weitere 48 Variablen von Cecchini et al. (2010) hinzu, die Veränderungen in Niveaus, potenzielle Niveauveränderungen sowie abnormale prozentuale Veränderungen von üblicherweise manipulierten Finanzabschlusseinträgen oder Verhältnissen aufnehmen. Sämtliche Variablen lassen sich aus den Compustat-Daten berechnen.

Als Optimierungsziel legen die Autoren eine übliche Grösse *Erwartete Kosten der Fehlklassifizierung* fest und empfehlen nach Versuchen mit Untergruppen die Verwendung eines zwölffachen Beobachtungs-Untersamplings.

In Anlehnung an Beneish (1997) publizieren die Autoren die Metrik *Erwartete Kosten einer Fehlklassifikation* als Leistungskennzahl und erhalten im optimierten Fall einen Wert von 0.142.

Erhebung der benötigten Daten Für die Identifikation von dolosen Handlungen kommen Ansätze des überwachten und unüberwachten maschinellen Lernens zum Einsatz. Im Folgenden sollen die drei Ansätze geprüft werden, die nach Lantz (2015) derzeit hauptsächlich eine Verwendung finden:

- (Unüberwachtes) Clustering
- (Unüberwachte) Assoziationsanalyse
- Support-Vektor-Maschinen als Vertreter des überwachten Lernens

Ein Bereich, der stets mit einer erhöhten Gefahr für dolose Handlungen verbunden ist, insbesondere für dolose Handlungen durch das Management, bezieht sich nach ISA 240.26 (IAASB, 2009c) auf Umsätze bzw. Erlöse. Betroffen sind speziell Geschäftsbeziehungen mit Drittparteien, weil bei einem Konzernabschluss Transaktionen unter Teilgesellschaften in der Konsolidierung verschwinden und damit keinen Einfluss auf die Ergebnisse haben.

Die untersuchte Organisation stellte anlässlich der Zwischenrevision eine Liste der Verkaufstransaktionen für die Berichtsperiode zur Verfügung. In Tabelle 9 sind die *Features* aufgeführt:

Nr	Attribut	Beschreibung
1	Datum	Datum der Verbuchung im System
2	Buchungs-Nr.	Identifikationsnummer der Transaktion
3	Buchungstext	Freitext für Beschreibung der Transaktion
4	Debitor-Name	Name der Gegenpartei
5	Debitor-Nr.	Identifikationsnummer des Käufers
6	Produktgruppe	Name der Produktegruppe
7	Betrag in CHF	Betrag in Schweizer Franken

Tabelle 9: Features, die im Auszug aus den Umsatztransaktionen enthalten waren

Nach einer eingehenden Durchsicht der Transaktionen mit rund 2'000 Einträgen wurden die Attribute mit dem Management besprochen. Dabei liess sich klären, dass die konzerninternen Transaktionen anhand der Debitorennummern zu erkennen sind.

Im Hinblick auf die Vorgehensweise von Tackett (2013) sowie Jans et al. (2009) waren zusätzliche Attribute erforderlich: So wurde zu jeder Transaktion noch ein Attribut für die verantwortliche Person in der Buchhaltung und das Valutadatum – also das Datum, an dem die Transaktion rechtsverbindlich wurde – eingefordert.

Anhand der Debitorenkennzeichnung wurden konzerninterne und Transaktionen mit Dritten vorab in einem Notebook gefiltert und die resultierenden Saldos mit der Saldobilanz abgeglichen, um einen Überblick über die Vollständigkeit und Korrektheit des Auszugs zu erhalten. Es wurden jedoch keine Abweichungen entdeckt. Zudem wurden bei einer Zufallsstichprobe 25 Einträge direkt mit dem ERP-System abgeglichen. Zur Benutzung eines Modells des überwachten Lernens war es erforderlich, neben den unternehmensinternen Daten auch externe Daten zu sammeln. Dazu wurde ein Sample von 51 Firmen erstellt, welche dolose Handlungen zu verzeichnen hatten – was vergleichbar mit anderen Studien wie von Beasley (1996) oder Brazel et al. (2009) bzw. J. L. Perols et al. (2017) ist.

Es wurden nur Geschäftsjahre in die Berechnungen einbezogen, in denen eine dolose Handlung zum ersten Mal in einem Accounting and Audit Enforcement Release (AAER) eine Erwähnung findet. Es wird davon abgesehen, ein Unternehmen

mehrfach im Datensatz zu behalten, um sicherzustellen, dass das Unternehmen mit einer dolosen Handlung nicht zugleich im Trainingsset und im Validierungsset erscheint.

J. Perols (2011) verwendet AAERs, für die Section 10(b) Rule 10b-5 explizit eine SEC-Untersuchung vorsahen oder bei denen dolose Handlungen erwähnt wurden, um Unternehmen mit diesen Merkmalen zu identifizieren. Folgende Unternehmen mit dolosen Handlungen wurden aus der Studie ausgeschlossen: (1) Finanzinstitute (da für diese die Features nicht greifen), (2) Unternehmen, bei denen für das erste Jahr keine dolosen Handlungen vorlagen, (3) Erlasse im Zusammenhang mit den Buchprüfungsgesellschaften, nicht profitorientierte Unternehmen, dolose Handlungen mit Bezug auf die Registrierungsstatements sowie Unternehmen mit fehlenden *Compustat*-Finanzdaten.

Dazu wurden weitere Unternehmen hinzugenommen, für welche keine dolosen Handlungen bekannt waren und deren Finanzdaten in *Compustat* zu finden waren. Für die ausgewählten Unternehmen sind die erklärende und unabhängige Variablen erhoben worden, die in der einschlägigen Literatur eine Verwendung finden (z.B. bei Dechow et al., 2011). Die relevanten Features konnten vollständig aus den *Compustat*-Daten abgeleitet werden.

Für die Schlussrevision wurden vom Management erneut die gleichen internen Daten angefragt und schliesslich als Excel-File übernommen, während für die unternehmensexternen Daten auf den oben beschriebenen Datensatz zurückgegriffen werden konnte.

Erlangung des Geschäfts- und Datenverständnisses Auch bei der Aufdeckung von dolosen Handlungen konnte auf die Risikoanalyse für das Geschäftsverständnis abgestellt werden bzw. war ein solches bereits für die Erhebung der Daten notwendig.

Ferner wurden für die unternehmensinternen Daten bereits, wie oben ausgeführt, anlässlich der Datenbeschaffung, die Attribute des Datensatzes mit dem Management detailliert diskutiert. In weiteren ausführlichen Gesprächen mit dem Management ging es um die Klärung folgender Aspekte aus operativer und buchhalterischer Perspektive:

- Üblicher Anteil von Drittparteien

- Typen von Verkaufsgegenständen an Drittparteien
- Vereinbarung über die Übergabe von Gefahren und Risiken (Incoterms)
- Ausgestaltung der Handelsfinanzierung
- Zahlungsmodalitäten
- Rabattierungen und Umsatzboni

Das Management gab eine Auskunft zu sämtlichen Punkten und zeigte dem Team bei einer Werksbesichtigung die dazugehörigen Produktgruppen in der Produktion.

Zudem wurden vorab univariate Analysen für die einzelnen Attribute durchgeführt sowie die Datenreihen zur Analyse visuell aufbereitet.

Datenvorbereitung Zur Vorbereitung der Daten auf die Modellierung für das *Clustering*, die *Assoziationsanalyse* und die *Support-Vektor-Maschinen* waren folgende Schritte notwendig:

1. *Separation der Sollbuchungen*

   Die Durchsicht der Transaktionen ergab sehr schnell, dass insbesondere für das *Clustering* eine Ausscheidung von Sollbuchungen in den Umsatzkonten vorzunehmen ist. Diese Sollbuchungen stammten hauptsächlich aus stornierten Lieferungen und deren buchhalterischer Entsprechung. Mithilfe von Stornierungen wurden Teile der Buchungen wieder rückgängig gemacht (folglich wurde auch die zugehörige Habenbuchung berücksichtigt und ausgeschieden).

   Daneben wurden jeweils zum Monatsende durch das Controlling Abgrenzungen gebucht, die in den Folgemonaten wieder aufgelöst wurden und damit Sollbuchungen verursachten.

2. *Integration von externen Daten*

   Ein Teil der abgesetzten Produkte an Drittparteien stand in Verbindung mit Wintersport. Die Abnehmer der Produktionsanlagen waren zu Teilen Kunden aus dem Bereich des alpinen Sports.

Es lag daher nahe, eine Variable *Schneefall* für jeden einzelnen Tag für die fünf grössten Wintersportorte einzuführen, die jeder Beobachtung mitgegeben wurde.

3. *Einführung neuer Features*

   Aufbauend auf den eingeforderten Daten wurden neue Features bzw. unabhängige Variablen gebildet. (1) So wurde eine Variable *Saison* eingeführt, die abgeleitet aus dem Valuta-Datum eine Ganzzahl für die Saison generierte (Frühling 1, Sommer 2, Herbst 3 und Winter 4). (2) Zusätzlich wurde ein neues Feature für den *Tag im Jahr* aus dem Datum abgeleitet, beispielsweise 51 für den 51. Tag im Jahr. (3) Als weiteres Feature war ein Indikator (0 oder 1) vorgesehen, der angibt, ob eine Buchung während der ordentlichen Geschäftszeit getätigt wurde. Im Rahmen der Datenvorbereitung war bereits eine Liste aller Führungspersonen eingeholt sowie ein *Dummy-Variable* eingefügt worden, die *Personen in leitender Stellung* kennzeichnet. Die vom Controlling vorgenommenen und nicht aufgelösten Umsatzabgrenzungen liegen ihrer Natur nach ausserhalb des ordentlichen Geschäftsgangs und wurden daher speziell als *Abschlussbuchungen* gekennzeichnet.

4. *Standardisierung und Diskretisierung der Wertebereiche*

   Der *Clustering-Algorithmus* wie auch die *Support-Vektor-Maschinen* sind nicht skaleninvariant. Daher mussten die Werte für die unabhängigen Variablen vorab in einen einheitlich Wertebereich transformiert werden. Für die Zwecke der Analyse wurden die numerischen Werte standardisiert.

   Die Assoziationsanalyse kann einzig mit kategorialen bzw. *Dummy-Variablen* umgehen. Entsprechend mussten für die numerischen Werte Kategorien definiert werden, indem fünf Quantile für die gebuchten Beträge gebildet wurden. Anschliessend wurden die kategorialen Variablen wie Debitoren, Produktgruppen und Jahreszeiten mithilfe des *One-Hot-Encodings* in Dummy-Variablen überführt.

Zur Vorbereitung des Erkennens der dolosen Handlungen nach der Methode von J. Perols (2011) mithilfe der Support-Vektor-Maschine stand die Kodierung

der Berechnung der unabhängigen Variablen im Vordergrund, zumal mit den bezogenen *Compustat-Daten* die Grundlage für die Modellierung an sich geschaffen war.

Alle Schritte der Datenaufbereitung, d.h. das Feature-Engineering, die Integration externer Daten, die Berechnung von Features, wurden in *Jupyter Notebooks* durchgeführt und zum einfachen Nachvollzug von Dritten ausführlich dokumentiert. Es war nicht notwendig, über die üblichen Python-Pakete wie *Pandas* hinauszugehen, die für die Datenmanipulation bzw. Datenintegration genutzt werden können.

Modellierung der Daten Nach der Erarbeitung einer soliden Datengrundlage für die Modellberechnung zur Erkennung von dolosen Handlungen wurden die folgenden vier Modelle gebildet:

1. *k-Means Clustering*

   Für die Modellierung der Cluster wurde dem Vorgehen von Thiprungsri und Vasarhelyi (2011) gefolgt und ein k-Means Clustering durchgeführt, wobei die Identifikation der optimalen Zahl der Cluster gleich den Autoren über die Praktiker- bzw. Schultermethode mithilfe visueller Überprüfungen erfolgte (vgl. Abbildung 23).

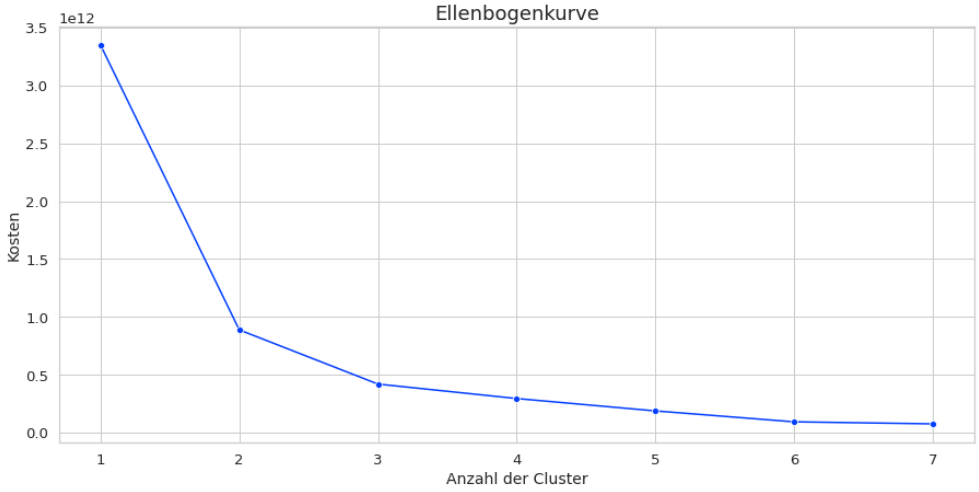


Abbildung 23: Ellenbogenkurve zur Bestimmung der optimalen Clusterzahl

Mit Festlegung der optimalen Clusterzahl wurden die einzelnen *Cluster* genauer untersucht. Es ergaben sich die folgenden Cluster und *Clustergrössen* (siehe Tabelle 10):

Cluster	Anzahl Transaktionen
A	10'032 Transaktionen
B	113 Transaktionen
C	15 Transaktionen

Tabelle 10: k-Means Cluster und die Anzahl der enthaltenen Transaktionen

2. *Assoziationsanalyse*

   Für die Assoziationsanalyse galt, dass das Datenanalysepaket die Modellierung mit dieser Technik nicht direkt implementierte. Es konnte jedoch das externe Paket *mlxtend* bezogen werden, das im *Jupyter Notebook* einen derartigen Ansatz zuliess.

   Auf Grundlage der kategorialen Variablen, die in Dummy-Variablen überführt wurden, konnten die Regeln modelliert werden. Dabei wurden für die Parameter für den benötigten Support und für die Konfidenz 1 % respektive 5% nach den Empfehlungen der Literatur gewählt (Tackett, 2013).

   Die ermittelten Daten, bestehend aus den Elementgruppen bzw. Regeln, Lift, Support, Konfidenz sowie weiteren Metriken, wurden abschliessend für die Debitorenbelastungen als *Konsequenz* – im Sinne der Assoziationsanalyse – gefiltert und absteigend nach *Lift* sortiert.

3. *Support-Vektor-Maschine*

   Dem Ansatz von J. Perols (2011) folgend, wurde auf Grundlage der externen Daten ein Modell mithilfe der *Support-Vektoren-Maschine* geschätzt.

   Nach der Empfehlung des Autors wurde für das Training ein Subsampling in den Abschlüssen ohne dolose Handlungen vorgenommen.

Evaluation der Modelle Für die Support-Vektor-Maschine wurde eine Prognose für das Auftreten von dolosen Handlungen getätigt und mit den unüberwachten

Ansätzen Gruppen von Transaktionen identifiziert, die potenziell für dolose Handlungen anfällig sind.

Das Clustering orientierte sich am Vorgehen von Jans et al. (2010). In den Clustern B und C wurde eine Stichprobe von Transaktionen gezogen. Die ermittelten Geschäftsvorfälle wurden anschliessend von einem Sachverständigen (Wirtschaftsprüfer) analysiert.

Für Cluster B waren beispielsweise hohe Beträge typisch. Die Unterstichprobe zeigte, dass sie Transaktionen enthielt, die jeweils am Monatsende bzw. als Auflösung zu Monatsbeginn als Umsatzabgrenzungen gebucht wurden. Im Fall des Clusters C erkannte der Wirtschaftsprüfer, dass eine Abstraktion aus den Daten sehr schwierig war und sich hauptsächlich durch die uniformen Buchungen auszeichnete, die nach der Prüfung keine wesentlichen Sachtatbestände hervortreten liessen.

Ein ähnliches Bild ergab die Analyse der Transaktionen, die in der Assoziationsanalyse einen hohen *Lift* aufwiesen. Auch bei ihnen wurde eine Unterstichprobe für die zehn auffälligsten Regeln gezogen und diese genauer analysiert. Insgesamt konnten mit den Prüfungen des Sachverständigen keine auffälligen Transaktionen aufgedeckt werden (dieses Ergebnis deckte sich mit der unabhängig durchgeführten Prüfung des Kontos durch das Prüfteams).

Das Modell von J. Perols, 2011 prognostiziert für den Abschluss der Organisation keine Vorkommnisse von dolosen Handlungen (für das *Traingsset* betragen die erwarteten Kosten einer Fehlklassifikation 0.152, was den offengelegten Werten nahekommt). Dies deckt sich auch mit der Einschätzung des ordentlichen Prüfteams.

Bereitstellung der Modelle Für die einzelnen Schritte – vom Einlesen der Daten über ihre Transformation bis hin zur Modellierung und Evaluation – wurde versucht, eine einfach zu operationalisierende Implementierung zu schaffen, die es dem Prüfer ermöglicht, aufgrund neuer Daten – beispielsweise im Folgejahr – im besten Fall auf Knopfdruck eine Gruppe von Transaktionen mit erhöhtem Risiko für dolose Handlungen zu identifizieren.

Das Einlesen, Prüfen, Transformieren und Integrieren konnten vollständig in einem *Jupyter Notebook* untergebracht und der Code darin sichergestellt werden. Zudem konnten zum Zweck der Nachvollziehbarkeit die einzelnen Programmschrit-

te vollumfänglich dokumentiert werden.

Die Analysen zur Einschätzung des Risikos für dolose Handlungen hat zwei Notebooks hervorgebracht: (1) ein Notebook, mit dem hauptsächlich unternehmensinterne Daten modelliert wurden – betroffen waren folglich das *Clustering* und die *Assoziationsanalyse*, (2) ein weiteres Notebook, mit dem ein Modell aufgrund von unternehmensexternen Daten berechnet wurde (*Support-Vektor-Maschinen-Ansatz*).

Ausgestattet mit diesen drei Notebooks bei gegebenen Softwarepaketen ist es dem Prüfer nun einfach möglich, die Analysen selbstständig nachzuvollziehen bzw. in der Folgeperiode anzuwenden.

Implikationen für Prozesse sowie Herausforderungen Die Anwendung der Ansätze führte zu Ergebnissen, die für den Prozess der Wirtschaftsprüfung und die damit verbundenen Herausforderungen hilfreich waren. Eine grundlegende Erkenntnis bestand darin, dass zwei sehr unterschiedliche Anwendungsfälle des maschinellen Lernens vorliegen: Im ersten Fall beruht das Training der Modelle nicht auf den Daten der geprüften Unternehmung, sondern es werden Daten anderer Unternehmen herangezogen. In einem zweiten Fall stützen sich Modelle für das Training hauptsächlich auf unternehmensinterne Daten und bilden auf diese Weise einen zweiten Anwendungsfall. Dabei kann die Integration von externen Daten, d.h. unabhängige Variablen, auch im zweiten Anwendungsfall helfen, die Prognosegüte zu verbessern. Das abgrenzende Merkmal besteht darin, dass die abhängige Variable stets aus dem geprüften Unternehmen hervorgeht.

Für den ersten Anwendungsfall mit unternehmensexternen Daten wurden neben den bereits beschriebenen, auch neuen Implikationen offengelegt:

1. *Klare Definition der unabhängigen Variablen*

   Im Hinblick auf die berechneten Werte der Abschlussposten wurde deutlich, dass es eine klare Definition der unabhängigen Variablen braucht, die im Modell Verwendung finden. Dies ist notwendig, um die Grössen über die verschiedenen Abschlüsse hinweg konsistent berechnen zu können, zumal eine inkonsistente Berechnung die Validität der Ergebnisse in Zweifel zöge.

In Bezug auf die Grössen, wie sie J. Perols (2011) verwenden, mussten Annahmen über die Komposition getroffen werden.

Für die trainierten Modelle des maschinellen Lernens aufgrund von unternehmensinternen Beobachtungen konnten die folgenden Implikationen offengelegt werden.

1. *Einfachheit des Clustering*

   Das *Clustering* hat sich als besonders einfach in seiner Anwendung herausgestellt, weil diese Methode nur über sehr wenig zu optimierende Parameter verfügt.

   So musste im Falle des *k-means Clusterings* lediglich zur Bestimmung der optimalen Clusterzahl die Schultermethode angewandt werden, während bei der Assoziationsanalyse und der Support-Vektor-Maschine wesentlich mehr Hyperparameter optimiert werden mussten.

   Auch die vorausgehende Datenaufbereitung, insbesondere im Falle der Modelle mit unternehmensinternen Daten und bei dem Ansatz von J. L. Perols et al. (2017), kann als wenig umständlich oder sogar einfach bezeichnet werden.

Die Aufdeckung von dolosen Handlungen und ihre besondere Behandlung stellen den Prüfer vor folgende Herausforderungen:

- *Unausgewogene Datensätze*

  Unausgewogene Datensätze spielen keine Rolle, wenn Cluster- oder Assoziationsanalyse durchgeführt werden.

  Die Anwendung von Methoden des überwachten maschinellen Lernens hingegen setzt voraus, dass die Eigenheit bzw. Unausgewogenheit der Datensätze berücksichtigt wird. Es wird in diesem Kontext vorgeschlagen, Stichproben bei Unternehmen ohne dolose Handlungen durchzuführen, um ausgewogene Datensätze zu generieren. Des Weiteren wird als Gegenmittel gegen den Fluch der Dimensionalität empfohlen, Unterstichproben in den unabhängigen Variablen zu bilden (J. Perols, 2011). Wird aufgrund derartig generierter

Datensätze eine Vielzahl von Modellen berechnet, können diese im *Ensemble* mithilfe einer Mehrheitsabstimmung über Klassen eine Vorhersage treffen.

- *Datenknappheit bei unternehmensexternen Beobachtungen*

  Bei einer Aufbereitung der Datensätze, wie im vorhergehenden Schritt, fällt die Anzahl der Beobachtungen dennoch öfters klein aus, zumal diese durch die Zahl an Fällen mit dolosen Handlungen determiniert wird. Mit einem reduzierten Trainingsset muss mithilfe der Kreuzvalidierung ein Validierungsset (oder gar eine Vielzahl davon) generiert werden, um dennoch Hyperparameter trainieren zu können.

Wird ein Verfahren gewählt, welches sich auf unternehmensinterne Beobachtungen stützt, so zeichnen sich folgende Herausforderungen ab:

- *Grosse Zahl an Anomalien*

  Der einfachen Anwendung des Clustering-Ansatzes steht der Umstand entgegen, dass damit eine hohe Zahl an potenziellen Anomalien aufgedeckt wurde.

  Für eine Bewältigung dieser möglichen Anomalien wurde auf Unterstichproben in den jeweils ausgewählten Clustern abgestellt, um die Merkmale zu identifizieren, die die Beobachtungen miteinander verbinden. Insgesamt handelt es sich um 213 Beobachtungen, die detailliert analysiert werden mussten.

4.3.4 ML-unterstützte analytische Prüfungshandlungen

Ansätze in der Literatur Auch für den Bereich der analytischen Prüfungshandlugen sind in der einschlägigen Literatur geeignete Ansätze des maschinellen Lernens aufgeführt. So verwendet Koskivaara (2000b) künstliche neuronale Netze, um Kontensalden für Finanzabschlüsse zu prognostizieren.

Für diesen Zweck bediente sich Koskivaara (2000b) der *One-Step Ahead*-Prognose (Einen-Schritt-Voraus-Prognose), die vorsieht, dass Werte aus einer Reihe von vorhergegangenen Perioden, beispielsweise x_{t-2}, x_{t-1} und x_t, als Inputvektoren für das neuronale Netzwerk herangezogen werden und den nächsten Wert x_{t+1} prognostizieren (siehe Abbildung 24). Entlang der Zeitachse kann so ein Trainingsdatensatz aufgebaut werden. Gleichzeitig lassen sich die *gelabelten* Daten aus dem Datensatz ableiten.

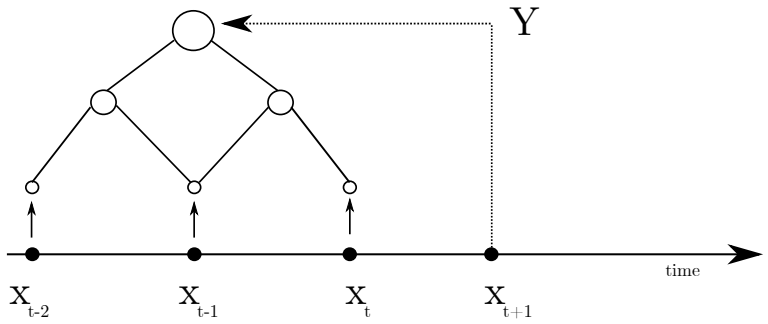


Abbildung 24: Prinzip der One-Step-Ahead-Prognose nach Bishop et al. (1995)

Die Modellierung der Daten erfolgt mittels eines mehrschichtigen, vorwärts gerichteten neuronalen Netzes und einer *Sigmoide* als Aktivierungsfunktion. Weiter besteht die Datengrundlage des Autors aus Finanzabschlüssen von 31 Firmen der produzierenden Wirtschaft über eine Periode von vier Jahren hinweg – allesamt aus der gleichen Branche (Koskivaara, 2000b). Die Variablen, welche ins Modell fliessen, sind in Tabelle wiedergegeben 11.

Nettoumsatz	sonstige Einkünfte
Materialaufwand	Veränderung Lager
externer Dienstleistungsaufwand	Personalkosten
Nettomargen	administrative Kosten
Nettoerfolg	Anzahl der Mitarbeiter
Debitorenbestand	Finanzvermögen
Kurzfr. Verbindlichkeiten	sonstiger Aufwand

Tabelle 11: Übersicht über verwendete unabhängige Variablen im Modell von Koskivaara (2000a)

Analog zu anderen Ansätzen wie dem Clustering müssen die Daten mithilfe einer Skalierung standardisiert bzw. normalisiert werden.

Für die Vorverarbeitung der Daten hat Koskivaara (2000a) vier verschiedene Ansätze eingesetzt:

1. Im *Modell A* wurden alle Variablen linear skaliert, womit die Dynamik und die Beziehungen zwischen den Variablen erhalten bleiben.

2. Im *Modell Y* wurden die Daten lediglich für jedes Jahr einzeln linear skaliert.

3. Das *Modell C* skaliert die Daten linear für eine gesamte Organisationseinheit.

4. Das *Modell C & Y* skaliert die Daten linear, sowohl für die Geschäftseinheit als auch für das Jahr.

Einbezogen ins Training wurden die 16 Variablen für die beiden vorangegangenen Jahre – entsprechend also insgesamt 32 – als Inputvariablen in das neuronale Netz, die für eine Prognose des dritten Jahres benutzt wurden. Dem Modell wurden noch vier Neuronen hinzugefügt, die die Jahre der Inputvariablen kennzeichnen und jeweils einen Wert 0 oder 1 annehmen, wenn das Jahr der Daten mit dem Neuron korrespondierte, sowie ein Neuron zur Kennzeichnung der Unternehmung. Als Ausgangsneuronen wurden wiederum die 16 Variablen gewählt, die prognostiziert werden.

Optimiert wurde das Netz mit der Zielfunktion, die die Minimierung des Root Mean Square Error (RMSE; Wurzel des mittleren Abweichungsfehlers) vorsah.

Das verwendete Trainingsset bestand aus 25 Unternehmen mit den Daten der Finanzabschlüsse von vier Jahren. Hinzu kamen sechs Organisationen, für die die ersten drei Jahre gleichwohl als Trainingsdaten verwendet wurden und das letzte Jahr als Testset für die Evaluation der Prognosegüte diente.

Die in Tabelle 12 aufgeführten Resultate sind das Ergebnis der Optimierung des neuronalen Netzes, das den RMSE der Zielabschlussposten minimiert. Werden damit für ein Testset von sechs Unternehmensjahren, die zuvor nicht für das Training benutzt wurden, Prognosen erstellt, so ergibt sich innerhalb einer Bandbreite von 10 % die Anzahl korrekter oder falscher Werte, wie sie in der Tabelle gezeigt sind.

Insgesamt hat das Modell A am besten abgeschnitten und 91 der 96 Werte so vorgesagt, dass sie innerhalb einer Bandbreite von plus oder minus 10 % zu liegen kamen (siehe Tabelle 12).

In einer Folgestudie modellierten Koskivaara und Back (2007) einzelne Kontensalden für eine Reihe von öffentlichen Betrieben wie Gesundheitsdienstleister, Hafenverwaltung und soziale Wohlfahrt. Als Grundlage dienten die Saldenbilanzen für eine Zeitspanne von fünf bis sieben Jahren, die teilweise Hunderte Konten umfassten. Das letzte Jahr in der Zeitreihe wurde jeweils als Testjahr herangezogen und entsprechend prognostiziert. Für eine vereinfachte Modellierung wurde die

	Modell A	Modell Y	Modell C	Model C&Y
Ebenen	5	3 C	6	4
Neuronen pro Ebene	37-32-26-20-16	37-27-16	37-33-29-25-21-16	37-27-16
Trainingsepochen	3002	4091	3627	2593
RSME	0.0048932	0.069367	0.132907	0.187114
Anzahl Elemente	96	96	96	96
Korrekte Elemente	91	89	61	61
Falsche Elemente	5	7	35	35
Prozentsatz Korrekt	95 %	93 %	64 %	64 %
Prozentsatz Falsch	5 %	7 %	36 %	36 %

Tabelle 12: Ergebnisse des künstlichen neuronalen Netzes im Artikel von Koskivaara (2000a)

Zahl der Konten reduziert, indem im ersten Schritt Konten entfernt wurden, für die über die Jahre nicht durchgehend monatliche Endbestände vorlagen. Anschliessend wurden zwei Gruppen von Konten gebildet: Die erste Gruppe wies gemessen an der monatlichen Abweichung die zehn stabilsten Endbestände auf (Modell A) und die zweite zehn von einem internen Wirtschaftsprüfer ausgewählte besonders interessante Konten (Modelle B).

Die Zeitreihe wurde aus Monatsendwerten der jeweiligen Konten gebildet und diente bis auf die letzten zwölf Monate als Trainingsgrundlage für das Modell. Für die zu beurteilende Prognosegüte wurden die Resultate des Modells jeweils mit den tatsächlichen Endbeständen verglichen sowie den budgetierten Werten gegenübergestellt.

Wie bereits bei Koskivaara (2000a) wurden in einem Vorverarbeitungsschritt die numerischen Daten der *Features* linear in einen Wertebereich zwischen 0 und 1 transformiert. Diese Transformation erfolgte auf zwei Arten: Einerseits wurden die Werte jeweils für ein einzelnes Konto über die Zeit hinweg in den Wertebereich $[0, 1]$ transformiert und andererseits wurde die Transformation über die gesamte Periode und gleichzeitig über alle Konten hinweg angewandt.

Für jedes Modell wurden verschiedene Tests durchgeführt, um die optimalen Parameter und Netzwerktopologien zu finden. Als Eingangsneuronen wurden schlussendlich (bis auf einen Fall) stets alle Konten aus Modell A und B gleichzeitig

verwendet, um damit alle Konten entweder aus Modell A oder Modell B zu prognostizieren. Eine Konfiguration mit 22 Eingabeneuronen, zehn Neuronen in der verdeckten Schicht und zehn Neuronen in der Ausgabeschicht hat sich zusammen mit einer Sigmoiden-Aktivierungsfunktion als die erfolgreichste herausgestellt.

Die Autoren berechneten zudem die Prognosen unter Zuhilfenahme alternativer Methoden, um deren Güte relativ einschätzen zu können. Sie bildeten mögliche Prognosen, indem sie (1) auf den letztjährigen Wert zurückgriffen, (2) den Durchschnitt der letzten Jahre ermittelten und (3) die durchschnittliche Abweichung berechneten. Dabei erkannten sie, dass das neuronale Netz den alternativen Methoden überwiegend überlegen ist und in den Fällen, in denen die alternativen Ansätze bessere Resultate lieferten, die Abweichungen äusserst gering waren.

Koskivaara und Back (2007) führten zur Evaluation der Prognosegüte gleichfalls einen Vergleich mit den budgetierten Kontenendständen der verschiedenen Organisationen durch. Durchschnittlich wiesen die neuronalen Netze eine Abweichung von 12 % vom tatsächlichen Wert auf, während die budgetierten Werte um 19 % abwichen. In elf Fällen lagen die Prognosen des Modells näher beim tatsächlichen Wert. Bei den klassischen Budgetierungen stellte sich dieser Zustand in acht Fällen ein.

Erhebung der benötigten Daten Die Erhebung der Daten stützte sich auf unternehmensinterne und -externen Quellen.

Die unternehmensexterne Datensätze stammten wiederum von *Compustat* und enthielten die 16 hauptsächlich finanziellen Daten, wie sie auch Koskivaara (2000a) verwendete. Wie bei der Vorgehensweise der Autoren wurden diese Daten von Unternehmen herangezogen, die zu derselben Branche gehörten. Die Daten für die Modellierung wurden für den Zeitraum von 2016 bis 2019 bezogen. Ein Rückgriff auf Daten, die bei der Erkennung von dolosen Handlungen bereits Verwendung fanden, war nicht möglich, da sie über verschiedene Branchen hinweg akquiriert wurden.

Zur Reproduktion des Ansatzes nach Koskivaara und Back (2007) konnte allein auf unternehmensinterne Daten abgestellt werden. Analog zu diesem Vorgehen wurden Daten für Saldenbilanzkonten aus zwei Gruppen herangezogen:

1. *Daten für Konten mit wenig Variabilität*

Für die letzten vier Jahre wurden monatliche Abschlussbestände für diejenigen Konten angefordert, die die geringste Varianz aufwiesen. Die Jahresendbestände der letzten vier Jahre wurden herangezogen – und nicht auf Monatsendsalden abgestellt, um die geringste Varianz zu ermitteln. Dies war eine pragmatische Entscheidung, anderenfalls hätten 48 Konten und deren Monatsendbestände analysiert werden müssen. Die auf diese Weise gewonnenen Konten wurden erneut einer Durchsicht unterzogen und diejenigen entfernt, die von minimalem Interesse waren, wie beispielsweise IC-Bestände bzw. IC-Flussgrössen. Es blieben Konten übrig, die unter anderem den Nettoumsatz und Personalaufwand beeinhalteten.

2. *Vom Wirtschaftsprüfer als interessant gewertete Konten*

   Die langjährigen Wirtschaftsprüfer des Kunden wurden gebeten, zehn Konten auszuwählen, die sie als besonders relevant einschätzten. Aufgrund ihrer analytischen Prüfungshandlungen anlässlich der Mandatsplanung und aufgrund ihres Wissens über nicht routinemässige Transaktionen wurden die monatlichen Daten für diese Konten vom Controlling eingefordert.

Als Erweiterung des Modells nach Koskivaara und Back (2007) wurden zusätzlich externe Datenquellen einbezogen. So wurden aufgrund des Geschäftsverständnisses die folgenden Zeitreihen aufbereitet:

- Witterungsdaten
- Wechselkursdaten
- Makroökonomische Daten (Allgemeines Zinsniveau, Veränderung des Bruttoinlandproduktes, Arbeitslosenquote u.a.)

Erlangung des Geschäfts- und Datenverständnisses Die Erlangung des Geschäftsverständnisses wurde anlässlich der Risikoanalyse durchgeführt, noch bevor der Prozess der Datenanalyse seinen Anfang nahm. Sie ist eine notwendige Voraussetzung für die Identifikation relevanter Drittquellen, die dem Modell zur Erhöhung der Prognosegüte hinzugefügt wurden.

Ferner konnte für das Datenverständnis überwiegend auf die Erkenntnisse abgestellt werden, die bereits im Rahmen der Aufdeckung von dolosen Handlungen vorlagen und mithilfe von Diskussionen mit dem Management, univariaten Analysen und visueller Aufbereitung erarbeitet wurden.

Datenvorbereitung Um die Daten im Analysesoftwarepaket in einer operationalisierbaren Form zu überführen, mussten die folgenden vorbereitenden Schritte ergriffen werden:

1. *Integration von externen Daten*

   Für eine Verwertbarkeit im Modell von Koskivaara und Back (2007) war es nötig, sämtliche unternehmensexternen und -internen Daten in Monatsaggregate umzuwandeln. Im Fall der Temperaturdaten lagen stündliche Beobachtungen vor, sodass sich unter Rückgriff auf alle verfügbaren Stunden ein monatlicher Durchschnittwert errechnen liess. Für die Wechselkursdaten, die als Tagesmittelkurs vorlagen, wurde eine ähnliche Vorgehensweise gewählt: Es wurden Monatsaggregate gebildet und anschliessend die Mittelwerte berechnet. Die makroökonomischen Daten lagen bereits auf Monats- bzw. Quartalsbasis vor und bedurften daher keiner weiteren vorbereitenden Massnahmen.

   So transformiert konnten die Daten anschliessend über eine *Jahr-Monat* Kombination (bspw. 2019-04) zusammengeführt bzw. integriert werden.

2. *Einführung neuer Features*

   Der Ansatz von Koskivaara (2000a) sieht vor, dass in der Eingabeschicht im neuronalen Netz zusätzliche Neuronen vorhanden sind, die das Jahr als Indikatorvariablen festhalten. Entsprechend war es notwendig, die Daten in einer Art und Weise aufzubereiten, dass die entsprechenden *Dummy-Variablen* je Trainingsvektor vorhanden waren.

   Für das Modell von Koskivaara und Back (2007) war es zudem nötig, Variablen für das Jahr und den Monat abzuleiten.

3. *Standardisierung der Wertebereiche*

Neuronale Netze sind bezüglich ihrer unabhängigen Variablen nicht skaleninvariant und mussten daher noch in standardisierte Wertebereiche überführt werden.

Modellierung der Daten Zwei sehr unterschiedliche Modelle wurden aufgrund der erhobenen Daten berechnet: ein Modell, das sich massgeblich auf Jahresabschlusswerte von Unternehmen in der gleichen Branche abstützte (Modell 1), und ein zweites Modell, das vorwiegend auf Daten der Unternehmung selbst abstellte (Modell 2), bzw. um externe Daten erweitert wurde (Modell 2b).

Für beide Modelle galt, dass mit dem Einsatz von neuronalen Netzen eine zusätzliche Komplexität einherging: Bei der Verwendung dieser Methoden waren eine Reihe von *Hyperparameter* wie die Netzwerktopologie, die *Lernrate* oder die *Regularisierung* zu wählen und zu optimieren, die die Prognosegüte beeinflussten. Während für die Netzwerktopologie auf die Konfiguration aus der Literatur zurückgegriffen werden kann, waren die *Lernrate* und die *Regularisierung* als Teil des Trainings zu optimieren.

Um die Parameter innerhalb der relativ limitierten Datenbestände zu ermitteln, die die Prognosegüte maximieren, wurden dazu die Beobachtungen in drei Gruppen (eine Trainings-, Test- und Validierungsgruppe) eingeteilt. Die Validierungsgruppe wurde mithilfe einer *10-Fold-Kreuzvalidierungsmethode* aus den Trainingsdaten gewonnen. Für die Optimierung wurden jeweils 10 % der Beobachtungen in zufälliger und rotierender Weise herangezogen.

Das Datenanalyse-Tool bot als Teil der zur Verfügung gestellten Methoden auch die Berechnung der Gewichte des künstlichen neuronalen Netzes an, jedoch ohne die Unterstützung von GPUs, also speziellen Prozessoren, die die äusserst effiziente Berechnung von Matrizen ermöglichen, die wiederum für künstliche neuronale Netze besonders wichtig sind. Das Modell konnte gleichwohl auf dem Hauptprozessor (CPU) berechnet werden, weil die Datenmengen relativ überschaubar waren.

Im Vergleich zu dem *Clustering*, den *Assoziationsanalysen* oder dem linearen Regressionsmodell hat sich der Zeitbedarf für das Modell mit Monatsdaten als eher hoch herausgestellt: Die Berechnung des neuronalen Netzes mit der Optimierung der Hyperparameter dauerte fast einen ganzen Tag. Im Gegensatz dazu erfolgte die Berechnung der Gewichte des Modells mit Jahresdaten innerhalb weniger Stunden.

Die Netzwerktopologien wurden den in der Literatur genannten Topologien grundsätzlich nachempfunden: Für das Modell mit Jahresaggregaten wurden 37 Eingabeneuronen, je 32, 26 und 20 Neuronen in der verdeckten Schicht sowie zehn Ausgabeneuronen gewählt.

Für die Prognose von Monatsdaten wurden 22 Eingabeneuronen, zehn Neuronen in der verdeckten Schicht und ein Ausgabeneuron gewählt (womit nur ein Endbestand prognostiziert wurde).

Evaluation Wird eine Evaluation der Leistungsfähigkeit des Modells aufgrund des *Testsets*, bestehend aus sechs Unternehmensjahren, durchgeführt, so ergeben sich für Modell 1 die Leistungskennzahlen in Tabelle 13.

Ebenen	5
Neuronen pro Ebene	37-32-26-20-16
Trainingsepochen	3426
RSME	0.0037542
Anzahl Elemente	96
Korrekte Elemente	88
Falsche Elemente	8
% Korrekt	92 %
% Falsch	8 %

Tabelle 13: Ergebnisse des künstlichen neuronalen Netzes im Artikel nach Koskivaara (2000a)

Wurde das Modell auf die tatsächlichen Werte des Unternehmens angewandt, so prognostizierte es in 15 von 16 Fällen (94 %) den tatsächlichen Jahresabschlussbestand korrekt.

Für Modell 2 wurde, wie der Konfiguration der Netzwerktopologie zu entnehmen ist, aufgrund der 22 Eingabewerte lediglich die Prognose für ein Konto (Verkäufe an Drittkunden) vorgenommen, was der klassischen analytischen Prüfungshandlung entspricht.

Der Tabelle 14 ist zu entnehmen, dass die Prognosequalität des erweiterten, mit externen Daten angereicherten Modells wesentlich besser abschneidet.

Der mit dem Modell prognostizierte Wert für Modell 2b wich dabei um +7,64 % oder rund 32'000 Schweizer Franken vom tatsächlich ausgewiesenen Wert ab.

	Modell 2	Modell 2b
Ebenen	5	5
Neuronen pro Ebene	22-10-10	30-10-10
Trainingsepochen	3426	4538
RSME	0.0037542	0.0015343
Anzahl Elemente	10	10
Korrekte Elemente	6	8
Falsche Elemente	4	2
% Korrekt	60 %	80 %
% Falsch	40 %	20 %

Tabelle 14: Ergebnisse des künstlichen neuronalen Netzes im Artikel nach Koskivaara (2000a)

Bereitstellung des Modells Die Anwendung der logistischen Regressionsmodelle, des Clusterings oder der Assoziationsanalyse ergab stets die Möglichkeit, dass sich ein Notebook mit Code und Bemerkungen aufstellen liess, das aus einem Satz neuer Daten bzw. Beobachtungen direkt die Berechnung eine Prognose ermöglichte.

Im Falle des Einsatzes von neuronalen Netzwerken konnte gleichfalls ein Jupyter Notebook implementiert werden, das die Daten lädt, die notwendigen Transformationen vornimmt und ein Modell schlussendlich berechnet.

Diese Bereitstellung spielt im Kontext der analytischen Prüfungshandlungen aber nur bei Modell 1, das auf unternehmensexternen Daten beruht, eine Rolle. Bei Modell 2 ist dies von untergeordneter Bedeutung, da sich die Daten, auf denen das Modell trainiert wird, mit jeder im ERP-System erfassten Transaktion verändern. Dies impliziert wiederum, dass damit stets die aktuelle Entwicklung in der Organisation und im Umfeld berücksichtigt werden kann und für eine optimale Prognosegüte stets wieder neu trainiert werden muss.

Implikationen für Prozesse sowie Herausforderungen Bei vorliegendem Mandaten ergeben sich aus der Anwendung der neuronalen Netze bei analytischen Prüfungshandlungen die folgenden Prozessimplikationen:

1. *Erlangung des Geschäfts- und Datenverständnisses*

   Die Erlangung des Geschäftsverständnisses wurde bereits im Rahmen der

Risikoanalyse – also noch bevor Modelle aufgestellt wurden – durchgeführt. Im Gespräch mit Sachverständigen und aufgrund von Recherche konnten damit Drittquellen für externe Daten identifiziert werden, die später Eingang ins Modell fanden und zu einer erhöhten Prognosegüte führten.

2. *Akquisition der Daten*

   Wie in den vorangegangenen Prüfprozessen müssen auch bei analytischen Prüfungshandlungen stets deren Daten erhoben werden.

   Die ERP-Systeme leisten im vorliegenden Fall gute Dienste und erlauben es, unternehmensinterne Transaktionen in einem gängigen Dateiformat zu extrahieren.

   Daneben gilt es, auch unternehmensexterne Datenbestände anzubinden bzw. zu erschliessen, um sie als unabhängige Variablen in den Modellen zu verwerten.

3. *Zentrale Bereitstellung von externen Daten*

   Die Modelle profitieren von externen (Dritt-)Daten. Es ist davon auszugehen, dass externe Daten wie Wechselkurse oder Zinssätze nicht nur im vorliegenden Fall geeignet sind. Gleichzeitig stand die Beschaffung der Daten wie Wetter- oder Konjunkturdaten mit einem grösseren zeitlichen Aufwand in Verbindung. Folglich ist es aus Effizienzgesichtspunkten sinnvoll, derartige Daten zentral zu verwalten, zu bewirtschaften und zur Verfügung zu stellen.

4. *Rechen- bzw. zeitintensive Modellberechnungen*

   Für die Unterstützung der analytischen Prüfungshandlungen durch maschinelles Lernen kamen erstmals neuronale Netze zum Einsatz. Dieser Ansatz hat wie auch die Support-Vektor-Maschinen den Vorteil, dass die Zusammenhänge der einzelnen Variablen und ihr Beitrag für die Prognose *a priori* nicht bekannt sein müssen und sich dadurch sehr gut für die Berechnung eines Erwartungswerts in der Wirtschaftsprüfung eignen.

   Der Einsatz der Support-Vektor-Maschine hat sich unter Zuhilfenahme des Analysepakets *Scikit-Learn* jedoch als sehr rechenintensiv erwiesen. Selbst

bei Verwendung einer moderaten Zahl von Beobachtungen dauerte die Optimierung des Hyperparameters mehrere Stunden. Folglich ist vorab diesbezüglich ein hoher zeitlicher Berechnungsaufwand zu berücksichtigen.

5. *Geringe Bedeutung der Bereitstellung der Modelle*

   Die Bereitstellung der Modelle hat nur eine geringe Bedeutung, wenn massgeblich auf interne Daten abgestellt wird.

   ERP-Systeme erfassen fortwährend die Transaktionen der Unternehmung, deren Daten in die Berechnung der Modelle einfliessen. Aufgrund dieses Stroms an neuen Daten, die die aktuelle Entwicklung innerhalb und ausserhalb der Unternehmung einfangen, bietet es sich an, die Modelle stets neu zu trainieren.

Als Herausforderung bei der Anwendung von Methoden des maschinellen Lernens im Kontext der analytischen Prüfungshandlungen hat sich wiederum die Erschliessung von externen Datenquellen herausgestellt. Darüber hinaus wurden die folgenden Prüfsteine erkannt:

- *Konfiguration des künstlichen neuronalen Netzes*

  Bei der Anwendung von neuronalen Netzen wurden Parameter erkannt, die konfiguriert bzw. variiert werden können, um die Prognosegüte zu optimieren. Zu diesen gehören:

  - Netzwerktopologie
  - Lernrate
  - Regularisierung
  - Aktivierungsfunktion

  Entsprechend erfordert der erfolgreiche Umgang mit derartigen Methoden ein Wissen über Heuristiken und praktische Erfahrung.

- *Bedarf an Rechenleistung*

Die Berechnung der *künstlichen neuronalen Netze* wie auch der Support-Vektor-Maschinen hat aufgrund der gewählten Methoden im Vergleich zu anderen Prüfprozessen deutlich mehr Zeit in Anspruch genommen.

Neben der Berücksichtigung im Prozess kann diesem Umstand auch mit neuerer Hardware entgegnet werden, die es zulässt, dass sich diese Berechnungen deutlich effizienter durchführen lassen.

- *Datenablage*

  Die Datensammlung, die mit dem Prüfprozess einherging, kann einen Nutzen über das analysierte Mandat hinaus stiften, wenn die Daten zugänglich gemacht werden.

  Für die dauerhafte Nutzung dieser Daten ist jedoch eine strukturierte Ablage der Rohdaten und gegebenenfalls der berechneten *Features* erforderlich, indem die Metadaten wie die Datenherkunft, Felder und Datentypen mit einbezogen werden.

4.3.5 ML-unterstützte Analyse der Fortführungsfähigkeit

In der Phase des Mandatsabschlusses wird maschinelles Lernen eingesetzt, um die für Rechnungslegungszwecke zentrale buchhalterische Annahme der Unternehmensfortführungsfähigkeit zu analyisieren bzw. zu prüfen.

Ansätze in der Literatur So verwendet Jones (2017) eine modifizierte Version von Entscheidungsbäumen – das sogenannte *Boosting* –, um die Fortführungsfähigkeit zu modellieren. Beim *Boosting* wird als Erweiterung zu den klassischen Entscheidungsbäumen nicht nur ein Baum berechnet, sondern es wird simultan eine grosse Anzahl von Bäumen aufbereitet, wobei nur eine Stichprobe an Variablen für einen einzelnen Baum herangezogen wird. Auf diese Weise entsteht eine grössere Anzahl von Entscheidungsbäumen, die aufgrund dieser Untergruppe von Variablen eine Entscheidung treffen, ob die Unternehmensfortführung gefährdet ist oder nicht. Mithilfe eines Mehrheitsabstimmungsverfahren wird über alle Bäume hinweg die Prognose abgegeben, um zu einer abschliessenden Einschätzung im Modell zu gelangen.

Für die Modellierung zieht der Autor eine grosse Anzahl von *Features* hinzu, die aus den folgenden Bereichen stammen:

- *buchhaltungsorientiert*

  In Anlehnung an Altman (2002) sowie Beaver et al. (2005) verwendet Jones (2017) Verhältnisse und Indikatoren für Liquidität und Solvenz, für Kapitalstruktur und Verschuldungsgrad, für Profitabilität und Eigenkapitalrendite, für Umsatz und Anlageeffizienz, für die Qualität des operativen und des freien Cashflows und Wachstumsraten für den operativen Cashflow. Gleichfalls finden ausserplanmässige Abschreibungen und Wertberichtigungen Eingang in das Modell, während auch signifikante Reduktionen bei Investitionen in immaterielle Unternehmenswerte, Investitionsausgaben und Dividendenzahlungen aufgenommen werden.

- *marktwertorientiert*

  Unter den marktorientierten Grössen für die Prognose der Unternehmensfortführung hat der Autor in Anlehnung an Shumway (2001) die Marktkapitalisierung, das Markt-Buchwert-Verhältnis, abnormale Aktienrenditen und die Preisvolatilität in das Modell mit einbezogen.

- *eigentümerstrukturorientiert*

  Einschlägige Studien (z.B. Shleifer und Vishny, 1997) zeigen, dass eine Konzentration bei der Eigentümerstruktur mit dem Risiko von finanziellen Schieflagen in Verbindung steht. Entsprechend finden Konzentrations- bzw. Eigentümerstrukturvariablen Eingang in das Modell, wie Prozentsatz an institutionellen Anlegern und Prozentsatz der Aktien, die von den Top-5- oder Top-10-Anlegern gehalten werden.

- *analystenorientiert*

  Analysten haben mit ihren Empfehlungen einen erheblichen Einfluss auf das Risiko von finanziellen Problemen. Aus diesem Grund hat Jones (2017) Variablen eingeführt, die derartige Sachverhalte berücksichtigen. Er verwendet dazu die *Features* Herabstufungen in Empfehlungen durch Analysten und die Anzahl beobachtender Analysten.

- *kreditratingorientiert*

  Die Herabstufung von Anleihen im Rating wird oftmals als Stellvertreter für finanzielle Engpässe einer Unternehmung herangezogen und enthält häufig neben öffentlichen auch private Informationen (Hillegeist et al., 2004). Entsprechend fliessen Einschätzungen von Ratingagenturen als unabhängige Variable ins Modell ein.

- *konjunkturorientiert*

  Die einschlägige Literatur zu finanziellen Problemen legt nahe, dass konjunkturelle Einbrüche die Wahrscheinlichkeit für eine Unternehmensfortführung vermindern (z.B. Koopman et al., 2009). Jones (2017) verwendet in seinem Modell die folgenden makroökonomischen *Features*: Niveau und Wachstum des Bruttoinlandprodukts, Inflationsindex, Zinsniveau, Verhältnis der öffentlichen Verschuldung zum BIP, Arbeitslosenquote, einen Rezessionsindikator, Konsumentenstimmung, Anleiheverzinsung, Einkaufspreisindizes und weitere Grössen.

- *entschädigungsorientiert*

  Abschliessend vermutet der Autor einen Zusammenhang zwischen Entlöhnung des Top-Managements und dem Risiko von Solvenzproblemen, weil die Höhe der Entschädigung mit einer höheren Managementqualität korrelieren könnte. Entsprechend führt er ein Feature für die Entschädigung des Top-Managements ein.

Neben den oben genannten Variablen werden Kontrollvariablen für die Grösse von Unternehmen verwendet, zumal der Autor darlegt, dass jüngere und kleinere Unternehmen tendenziell schneller in finanzielle Schieflage geraten. Zudem bindet er eine Reihe von *Dummy-Variablen* für verschiedene Branchen mit ins Modell ein.

Für das Training verwendet der Autor die Datenbank Capital IQ Service von *Standard and Poor's* und extrahiert damit 1'115 Insolvenzen über eine Zeitspanne von 23 Jahren in den Jahren von 1987 bis 2003. Für diese Firmen werden die Attribute gemäss der Aufstellung oben entnommen, und zwar jeweils bis drei Jahre vor der Insolvenz. Gleichzeitig entnimmt der Autor auch einen Auszug an US-Firmen, welche nicht in finanzielle Schieflage geraten sind und öffentlich gehandelt

werden. Eine wenige Unternehmen mussten im Rahmen der Datenaufbereitung aus der Stichprobe der bankrotten Firmen ausgenommen werden, da für sie keine Finanzabschlüsse in der Datenbank vorhanden waren.

Mit seinem Modell erreicht der Autor die in Tabelle 15 gezeigte Prognosegüte (B steht hierbei für *bankrott* und NB für *nicht bankrott*).

	Erwart. B	Erwart. NB	%	Total %
Tatsächlicher B	1'453	34	97.73	95.78
Tatsächlicher NB	532	10'051	94.98	

Tabelle 15: Prognosegüte des Modells von Jones (2017)

Anandarajan et al. (2001) verwenden für die Prognose von Insolvenzen bzw. einer eingeschränkten Fortführungsfähigkeit ein künstliches neuronales Netz, indem sie für die Optimierung der Parameter einen *backpropagation-of-error* und einen *genetischen Algorithmus* einsetzen.

Als Features werden für das Modell eine Reihe von qualitativen Grössen verwendet, die als Dummy-Variablen in das Modell eingehen:

- *Negativer operativer Cashflow*

  Auf der Grundlage der Einsichten verschiedener Autoren wie Bell und Tabor (1991) antizipiert ein negativer operativer Cashflow potenziell zukünftige Zahlungsschwierigkeiten oder Probleme mit dem *Working Capital*.

- *Reduktion oder Ausfall von Dividendenzahlungen*

  In der Reduktion oder im Ausfall von Dividendenzahlungen sehen die Autoren nach Dielman und Oppenheimer (1984) eine erhöhte Wahrscheinlichkeit für eine finanzielle Notlage.

- *Zahlungsausfall auf Anleihen*

  Als weitere Variable wird der Zahlungsausfall auf Anleihen nach K. C. Chen und Church (1992) verwendet, die darin ein klares Signal für künftige Insolvenzen sehen.

- *Restrukturierung von Anleihen*

Vorgängige Studien verorten die Restrukturierung von Anleihen als eine Stufe auf dem Insolvenzspektrum (Turetsky & McEwen, 2001), die potenziell eine Vorstufe für eine tatsächliche Insolvenz bildet. Dementsprechend nehmen die Autoren auch diese qualitative Variable als Dummy in das Modell auf.

Für die Auswahl an quantitativen Variablen bedienen sich Anandarajan et al. (2001) bei Zmijewski (1984). Das Modell von Zmijewski (1984) schätzt die Wahrscheinlichkeit eines Bankrotts mithilfe folgender Verhältnisse:

- $Unternehmenserfolg/Bilanzsumme$
- $Fremdkapital/Bilanzsumme$
- $Umlaufsvermögen/KurzfristigeVerbindlichkeiten$

Für das Training ihres Modells haben die Autoren ihre Daten der *Compact-Disclosure-Datenbank* für den Zeitraum von Dezember 1989 bis März 1996 entnommen. Neben den Unternehmen, die in dieser Periode Bankrott gegangen sind, wurden als Kontrollgruppe Unternehmen ausgewählt, die eines der qualitativen Kriterien für finanzielle Notlagen nach Turetsky und McEwen (2001) erfüllten und anschliessend wieder eine finanzielle Stabilität erreichten.

Die Unternehmen wurden in drei Schritten selektiert: Im ersten Schritt erfolgte die Auswahl von Unternehmen, die finanzielle Schwierigkeiten im Feld *all text* der *Compact-Disclusure-Datenbank* vermerkt hatten. Im zweiten Schritt wurden Unternehmen ausgeschlossen, die nicht öffentlich gehandelt und nicht in den Vereinigten Staaten inkorporiert waren. Im dritten Schritt wurden Unternehmen aus der Stichprobe entfernt, für die keine Daten in der *Compustat-Datenbank* verfügbar waren. Dieser Prozess führte dazu, dass 265 Unternehmen übrig blieben, die in finanziellen Schwierigkeiten waren.

Für das Training wurden die Daten in zwei Gruppe unterteilt: ein Trainingsset und ein Testset im Verhältnis von 80:20. Das Training der neuronalen Netze erfolgte auf dem Trainingsset.

Die Prognosegüte für das klassische neuronale Netzwerk ist der Tabelle 16 zu entnehmen.

	Erwart. B	Erwart. NB	%	Total %
Tatsächlicher B	76	5	93.83	89.10
Tatsächliche NB	6	14	70.00	

Tabelle 16: Prognosegüte des klassischen neuronalen Netzes mit *backpropagation-of-error* nach Anandarajan et al. (2001)

	Erwart. B	Erwart. NB	%	Total %
Tatsächlicher B	84	4	95.45	95.19
Tatsächliche NB	1	15	93.75	

Tabelle 17: Prognosegüte des neuronalen Netzes mit genetischem Lernalgorithmus nach Anandarajan et al. (2001)

Tabelle 17 zeigt die Ergebnisse, wenn der genetische Lernalgorithmus für das Training des neuronalen Netzes herangezogen wird.

Die Ergebnisse belegen, dass der genetische Algorithmus die höchste Prognosegüte aufweist (mit 95 % für die bankrotten und 94 % Prognosegüte für die gesundeten Unternehmen), während die Güte beim Ansatz mit einem klassischen Gradientenabstiegsverfahren für insolvente Unternehmen bei 94 % und die gesundeten Unternehmen bei 70 % lag.

Einen anderen Ansatz wählten Martens et al. (2008), indem sie die Leistungsfähigkeit einer Support-Vektor-Maschine, eines logistisches Regressionsmodells und eines *AntFarmers* für die Prognose von Verletzungen der Fortführungsfähigkeitannahme evaluierten.

Als binäre abhängige Variable verwendeten die Autoren *Report*, die erfasst, ob für ein Unternehmen ein Vermerk mit Einschränkung zur Fortführungsfähigkeit vorliegt oder nicht. Die unabhängigen Variablen bestehen aus den vierzehn finanziellen und nicht finanziellen qualitativen Grössen, wie sie in Tabelle 18 ersichtlich sind.

Die Autoren nutzen als Quelle für die Daten die *Worldscope-Datenbank*, welche den Zugriff auf Unternehmensdaten von AMEX, NASDAQ und NYSE für den Zeitraum von 2002 bis 2004 zulässt. Die Daten der dort verfügbaren Unternehmen wurden danach gefiltert, ob sie sämtliche Variablen enthalten, wie sie im Forschungsaufbau vorgesehen waren. Damit wurde ein abgeleiteter Datensatz geschaffen, der insgesamt 10'318 Unternehmensjahre umfasste, von denen lediglich 271 Jahre auf

Variable	Definition
Abhängige Variable	
REPORT	Einschränkung bezüglich der Fortführung im Testat
Unabhängige Variable	
CR	Current Ratio
LNTA	Logarithmus Naturalis des Gesamtvermögens
CFO/TL	Operativer Cashflow geteilt durch die Summe des Fremdkapitals
TL/TA	Verschuldungsquote
Δ TL/TA	jährliche Veränderung der Verschuldungsquote
NITA	Jahresgewinn im Verhältnis zum Gesamtkapital
Δ Ni/TA	Veränderung des Verhältnisses von Jahresgewinn zu Gesamtkapital
NEG NI	1 bei Jahresverlust, sonst 0
NEG OI	1 bei negativem operativem Cashflow, sonst 0
Big-4	1 wenn eine Big-4-Prüfgesellschaft das Unternehmen prüft, sonst 0
RE/TA	thesaurierte Gewinne geteilt durch das Gesamtkapital
Δ RE/TA	jährliche Veränderung der thesaurierten Gewinne im Verhältnis zum Gesamtkapital
Δ WC/TA	jährliche Veränderung des Working Capital im Verhältnis zum Gesamtkapital
NW/TL	Eigenfinanzierungsgrad

Tabelle 18: Features und ihre Definitionen, wie sie von Martens et al. (2008) verwendet wurden

Prüfurteile entfielen, die eine Einschränkung bei der Fortführungsfähigkeit enthielten, während die übrigen Unternehmen Testate ohne Einschränkung ausgestellt bekamen.

Die Forscher wählten für die Bildung des Trainings- und Testsets zwei Ansätze, um dem Umstand Rechnung zu tragen, dass es weit mehr Unternehmen gibt, die ein uneingeschränktes Prüfurteil ausgestellt bekamen als solche, die eine Einschränkung enthielten:

- *abgeglichene Unternehmenswahl*

Bei der abgeglichenen Unternehmenswahl wurden Unternehmen ohne Einschränkung im Prüfurteil derart ausgewählt, dass sie jeweils annähernd einem Unternehmen mit Einschränkung entsprachen, indem für ein Unternehmen mit Einschränkung ein anderes mit annähernd identischer Bilanzsumme und Übereinstimmung in den ersten beiden SIC-Klassifikationsziffern gesucht wurde.

- *ausbalancierte Unternehmenswahl*

  Als Alternative zum Abgleich der Unternehmen wurde bei der ausbalancierten Unternehmenswahl eine Zufallsstichprobe in den Unternehmen ohne Einschränkung vorgenommen, sodass eine identische Anzahl Unternehmen in beiden Klassen erreicht wurde.

Martens et al. (2008) verwendeten beide Datensätze, um ein Logit-Modell, eine Support-Vektor-Maschine und ein AntMiner-Modell zu schätzen. Die daraus resultierenden Genauigkeiten zeigt Tabelle 19.

	Abgeglichen	Ausbalanciert
Genauigkeit		
Logit	72.5	81.2
SVM	73.7	80.7
AntMiner	68.2	76.2

Tabelle 19: Genauigkeit der Prognose in Prozent nach Martens et al. (2008)

Erhebung der Daten Die Datenbeschaffung in der Fallstudie erfolgte in zwei Schritten: So wurden im ersten Schritt die unternehmensexternen – zur Berechnung des Modells – und im zweiten Schritt die unternehmensinternen Daten beschafft.

Bei den unternehmensexternen Daten ziehen Anandarajan et al. (2001) die *Compact Disclosure Database* heran, um Organisationen mit Fortführungsschwierigkeiten zu identifizieren, wobei die Autoren als Ergänzung für die finanziellen Daten auf die *Compustat-Datenbank* zurückgreifen, während Jones (2017) die Daten aus der *Capital-IQ-Datenbank* von Standard and Poor's verwendet und dort rund 100 Variablen für sein Modell extrahiert. Martens et al. (2008) nutzen die *Worldscope-Datenbank*, um sowohl Prüfurteile mit Einschränkungen in der

Fortführungsfähigkeit zu ermitteln und Kenngrössen für die entsprechenden Finanzabschlüsse zu ermitteln.

Für die Zwecke dieser Studie wurden gleichfalls mehrfach unternehmensexterne Daten aus der *Compustat-Datenbank* entnommen: Sie erlaubte einen schnellen Zugriff auf finanzielle, aber auch auf nicht finanzielle Daten von Unternehmen. Zudem konnte die *Audit-Analytics*-Datenbank abgefragt werden, die sich mit der Compustat-Datenbank verbinden liess und insbesondere ein Feld bereithielt, das indizierte, ob für die Fortführungsannahme der Unternehmung im Prüfurteil eine Einschränkung bestand.

Es konnten jedoch nicht sämtliche 100 Variablen einbezogen werden, wie sie Jones (2017) verwendet. Vielmehr wurde auf *Features* abgestellt, die von Anandarajan et al. (2001) und Martens et al. (2008) vorgeschlagen wurden.

Für die in den Modellen verwendeten unternehmensinternen Kenngrössen galt, dass sie bereits überwiegend für die Erhebungen der vorangegangenen Phasen und die darin verwendeten Ansätzen vorlagen. Wo diese Kenngrössen noch nicht vorhanden waren, wurde abermals das Controlling bemüht, das die fehlenden Grössen beibringen konnte.

Erlangung des Geschäfts- und Datenverständnisses Über das Wissen hinaus, das zur geprüften Organisation anlässlich der Risikobeurteilung erlangt wurde, konnten während der vorhergehenden Prüfphase (Prüfungsdurchführung) ausreichend weitere Kenntnisse über das Unternehmen gewonnen werden, sodass keine weiteren Schritte notwendig waren.

Es wurden daher lediglich punktuell weitere Informationen vom Management bezüglich der Daten angefordert.

Datenvorbereitung Die Vorbereitung der Daten erforderte ein mehrstufiges Vorgehen:

1. *Korrektur von Lücken*

   Bei einigen akquirierten Daten fehlten Werte. In diesen Fällen wurden – sofern nur wenige Werte nicht vorhanden waren – die Datenlücken mit dem Durchschnittswert aller Daten für dieses *Feature* ersetzt.

Bei zu grossen Lücken (mehr als fünf Werte) wurden die Daten vollständig aus dem Datensatz gelöscht.

2. *Integration der Daten*

   Die Akquisition der Daten führte zu drei Tabellen: (1) Die erste Tabelle enthielt neben der Unternehmensidentifikation der geprüften Organisation auch den Berichtstext. (2) Die zweite Tabelle umfasste die finanziellen Unternehmensdaten sowie die SIC-Einordnung. (3) Die dritte Tabelle bestand aus dem Kreditrating der Unternehmung und einem Unternehmensschlüssel.

   Als Teil der Datenvorbereitung mussten die Tabellen miteinander verbunden werden. Hierzu wurde zunächst die Kreditratingtabelle über die Unternehmenskennung in die Unternehmensdaten integriert und anschliessend erfolgte die Verbindung der Unternehmensdaten über die Unternehmensidentifikation mit der Berichtstabelle.

3. *Unternehmensabgleich*

   Gestützt auf die Einsichten in der einschlägigen Literatur vergleichen Martens et al. (2008) die Unternehmen, denen eine eingeschränkte Unternehmensfortführung im Bericht des Prüfers attestiert wird, mit vergleichbaren Unternehmen – in Bezug auf die Bilanzsumme und die Ziffern in der SIC-Kategorisierung – für den Datensatz. Dies wurde für die Fallstudie ebenfalls mithilfe des entsprechenden Python-Codes automatisiert nachvollzogen.

4. *Bildung neuer Features*

   Für sämtliche von den Autoren verwendeten Variablen war eine Verrechnung erforderlich, um neue Beziehungen zu bilden. Diese Transformationen konnten für alle verwendeten Variablen durchgeführt und als neue Features abgespeichert werden.

   Eine Reihe von Daten liegen als kategoriale Werte vor, wie beispielsweise ein Feld für den Namen der Prüfgesellschaft eines Unternehmens. Um diese für die Modelle in operationalisierbarer Form vorzubereiten, mussten sie in Dummy-Variablen überführt werden.

5. *Standardisierung der Wertebereiche*

   Für die Anwendung spezifischer Ansätze wie die Support-Vektor-Maschine wird empfohlen, den Wertebereich der unabhängigen Variablen vorab zu standardisieren. Entsprechend wurden sämtliche Werte für den späteren Gebrauch in der Modellierungsphase in den Wertebereich $[0, 1]$ transformiert.

Modellierung der Daten Für die Modellierung der Klassifikation, ob eine Unternehmung in ihrer Fortführungsfähigkeit eingeschränkt ist, wurde auf die Ansätze der Autoren und ihre Konfigurationen zurückgegriffen.

- *Modellierung über geboostete Entscheidungsbäume*

  Aufgrund der guten Prognosegüte wurde in einem ersten Schritt die Fortführungsfähigkeit nach Jones (2017) mithilfe geboosteter Entscheidungsbäume modelliert. Dabei wurden die Hyperparameter *n_estimators* (Anzahl der zu schätzenden Entscheidungsbäume) und *max_depth* (Grösse der Entscheidungsbäume) unter Zuhilfenahme eines dritten Datensatzes (Validierungsdatensatz) optimiert, der mithilfe des *10-Fold-Kreuzvalidierungsansatz* aus dem Trainingsdatensatz gewonnen wurde.

- *Modellierung mithilfe einer Support-Vektor-Maschine*

  Nach Martens et al. (2008) wurde auch eine Support-Vektor-Maschine auf den Daten berechnet. Wiederum war für die zwei Parameter C und *gamma* eine Hyperparameteroptimierung notwendig. Dabei definiert C die Komplexität der Oberfläche, die zunehmend genauer die Trainingsdaten abbildet, und *gamma*, welchen Einfluss eine einzelne Trainingsbeobachtung ausübt. Entsprechend wurden in einem Gridsearch-Ansatz – der die Permutationen in den Parameterwerten systematisch abarbeitet – die Support-Vektoren errechnet. Die Hyperparameter wiederum wurden mit einem Validierungsset optimiert, das mit einem 10-Fold-Kreuzvalidierungsansatz abgeleitet und auf das Trainingsset angewendet wurde.

- *Modellierung mittels eines künstlichen neuronalen Netzes*

Als weitere nicht parametrische Methode wurde ein künstliches neuronales Netz berechnet. Als Hyperparameter lagen in diesem Modell die *Regularisierung* vor, also das Mass, mit dem die Gewichte mit grossem Einfluss abgestraft werden, und die Lernrate, die festlegt, in welchen Schrittgrössen sich der stochastische Gradient bewegt. Diese wurden abermals mit einem Gridsearch systematisch mithilfe eines Validierungssets (gebildet aus einem 10-Fold-Auswahlverfahren) optimiert. Zur Berechnung der Gewichte kamen ein Adam-Gradientenabstiegsalgorithmus nach Kingma und Ba (2014) und verschiedene Netzwerktopologien zum Einsatz.

- *Modellierung eines logistischen Regressionsmodells*

  Um zu einer Einschätzung über die Leistungsfähigkeit der *modernen* Ansätze des maschinellen Lernens im Vergleich zu einem parametrischen, und daher einfach zu verstehenden Modell, zu gelangen, wurde abschliessend ein logistisches Regressionsmodell geschätzt.

Evaluation Die Leistungsfähigkeit der verschiedenen Modelle wurde mithilfe des Testsets im Anschluss an deren Schätzung getestet. Im Folgenden werden nur die Konfusionsmatrizen des erweiterten Modells dargestellt, weil für den reduzierten Satz an Variablen von Anandarajan et al. (2001) die Resultate durchweg 20 - 30% hinter dem erweiterten Modell von Martens et al. (2008) zurückblieben.

In Tabelle 20 sind die Leistungskennzahlen nach dem *Random Forest*-Modell aufgeführt.

	Erwart. B	Erwart. NB	%	Total %
Tatsächlicher B	51	3	94.44	93.33
Tatsächlicher NB	5	61	92.42	

Tabelle 20: Going-Concern Konfusionsmatrix für das Random-Forest Modell

Wird hingegen ein *künstliches neuronales Netz* zur Prognose herangezogen, ergeben sich die in Tabelle 21 wiedergegebenen Werte.

Tabelle 22 zeigt die Leistungsfähigkeit bei Verwendung einer *Support-Vektor-Maschine*.

Abschliessend stellt Tabelle 23 die Resultate der *logistischen Regression* dar.

	Erwart. B	Erwart. NB	%	Total %
Tatsächlicher B	50	4	92.53	90.84
Tatsächlicher NB	7	59	89.40	

Tabelle 21: Going-Concern Konfusionsmatrix eines künstlichen neuronalen Netzes

	Erwart. B	Erwart. NB	%	Total %
Tatsächlicher B	49	5	90.74	91.67
Tatsächlicher NB	5	61	92.42	

Tabelle 22: Going-Concern Konfusionsmatrix einer Support-Vektor-Maschine

Bereitstellung der Modelle Die vier Ansätze wurden anlässlich der Zwischenrevision in vier separaten *Jupyter Notebooks* über sämtliche Schritte der Data-Science-Pipeline hinweg aufgesetzt. Dies umfasste:

- Laden der Daten in den Arbeitsspeicher
- Durchführung der notwendigen Transformationen – wie Umwandlung und Standardisierung
- Aufteilung der Daten in Trainings- und Testdaten
- Berechnung der Modellparameter inklusive einer Optimierung der Hyperparameter mithilfe eines Gridsearches
- Erstellung von Prognosen für Testdaten und Berechnung der Prognosegüte

Für die Anwender des Modells sind viele Schritte unnötig: Für sie ist es lediglich wichtig, dass aufgrund eines Eingabevektors für die Daten der geprüften Organisation eine Prognose über die Fortführungsfähigkeit ermöglicht wird. Folglich kann für die Anwender selbst ein eigenes *Jupyter Notebook Light* erstellt werden, das die Berechnung der Features und eine Prognose aufgrund abgegebener Modellparameter bereitstellt.

Implikationen für Prozesse und Herausforderungen Die bereits vorher festgestellten Implikationen und Herausforderungen gelten nach wie vor, auch wenn sich Probleme im Zusammenhang mit der Zwischenrevision und dem Prüfzeitpunkt zur Schlussrevision lösen liessen.

	Erwart. B	Erwart. NB	%	Total %
Tatsächlicher B	44	10	81.48	78.33
Tatsächlicher NB	16	50	75.76	

Tabelle 23: Going-Concern Konfusionsmatrix einer logistischen Regression

Zudem konnten folgende Implikationen für den Prüfprozess erkannt werden.

- *Aufteilung in Berechnung und Anwendung*

  Im Gegensatz zu Modellen, deren Datengrundlage hauptsächlich auf Beobachtungen der Unternehmung selbst beruht, kann bei Trainings mit unternehmensexternen Daten folgende Art der Arbeitsteilung praktiziert werden:

  Im ersten Schritt wird wie bislang ein *Notebook* entworfen, das Daten lädt, transformiert, das Modell berechnet und abschliessend Modellparameter für Anwender bereitstellt.

  Danach wird in einem zweiten Schritt ein *Anwender-Notebook* erstellt, das die Daten der Unternehmung in ein passendes Format transformiert und dem Modell zuführt, das nicht mehr trainiert wird, sondern lediglich aufgrund eines Eingangsvektors eine Prognose ausgibt.

  Das so entwickelte *Jupyter Notebook-Light* stellt eine deutlich reduzierte Codebasis zur Verfügung und ist gleichfalls weniger fehleranfällig, weil kein Training mehr stattfindet.

- *Dokumentation sämtlicher Prozessschritte*

  Während der Zwischenrevision konnten die Modellparameter bereits aufgrund der Verfügbarkeit der externen Daten berechnet werden. Es hat sich nach unterschiedlichen Gesichtspunkten als äusserst hilfreich herausgestellt, sämtliche Programmierschritte vollständig und direkt zu dokumentieren:

  - Die Wiederholung der Berechnung in einer Folgeperiode konnte wesentlich vereinfacht werden.
  - Etwaige Anpassungen und Verbesserungen liessen sich sehr einfach einführen.

– Der Nachvollzug der Arbeit war gewährleistet, was wiederum eine Qualitätskontrolle des Mandatsmanagements ermöglichte und die Kommunikation erleichterte.

4.3.6 ML-unterstützte Berichterstattung

Den eigentlichen Abschluss der Prüfarbeiten bzw. des Prüfungsprozesses bildet die Berichterstattung, in der ein Urteil darüber gefällt und dokumentiert wird, ob ein geprüfter Abschluss in wesentlichen Teilen den vorgegebenen Standards entspricht – möglicherweise sind Ergänzungen aufgrund etwaiger zusätzlicher Anforderungen aus dem jeweiligen Rechtsraum notwendig. So ist unter Umständen ein zusätzliches Urteil über die Implementierung bzw. Funktionsfähigkeit eines internen Kontrollsystems erforderlich.

Ein Prüfurteil kann in seinem Standardwortlaut ohne Einschränkungen ausgestellt werden, falls keine wesentlichen Fehldarstellungen identifiziert wurden. Sind hingegen wesentliche Fehler festgestellt worden, wird eine Qualifizierung des Prüfurteils vorgenommen und damit vom Standardwortlaut abgewichen.

Ansätze in der Literatur In der einschlägigen Literatur finden sich Meinungen, die die automatisierte Modellierung der Prüfurteil für notwendig erachten. Gaganis et al. (2007) schlagen für diesen Zweck drei Methoden vor: Die Autoren führen eine Nächste-Nachbarn-Klassifikation durch, wählen eine Diskrimantenanalyse und verwenden abschiessend ein logistisches Regressionsmodell.

Die verwendete abhängige Variable und die unabhängigen Finanz- und Nicht-Finanz-Features sind in Tabelle 24 aufgeführt.

Mit LIQ bzw. mit der Quick-Ratio wollen die Autoren die Liquiditätssituation des Unternehmens einfangen. Gestützt auf den Ergebnissen von Spathis (2003) nimmt die Qualifikation der Prüfurteile mit zunehmender Verschlechterung der Liquiditätssituation zu. Gleichwohl stellen die Autoren fest, dass auch der umgekehrte Fall dokumentiert ist: Ireland (2003) stellt fest, dass eine hohe Liquidität die Wahrscheinlichkeit einer Qualifikation erhöht, weil damit eine überhöhte Bewertung im Anlagevermögen einhergehen kann.

Weiter führen die Autoren die unabhängige Variabel PROF (Gesamtkapitalrendite) ein. Sie stützen sich auf Studien, die darlegen, dass eine verminderte Gesamt-

Variable	Definition
Abhängige Variable	
Audit Opinion	Dummy-Variable mit Wert 1, wenn eingeschränkt, sonst 0
Unabhängige Variable	
PROF	Gesamtkapitalrendite
LIQ	Quick-Ratio
GROWTH	Jährliches Wachstum des Gesamtkapitals
LEV	Eigenkapitalquote
RATING	Rating von 0-5 (5 steht für ein Hochrisiko)
INDUSTRY	Dummy-Variable mit 1, wenn verarbeitende Industrie sonst 0

Tabelle 24: Miteinbezogene Features im Modell von Gaganis et al. (2007)

kapitalrendite mit einer erhöhten Wahrscheinlichkeit zur Qualifizierung einhergeht (Loebbecke et al., 1989; Summers und Sweeney, 1998; Laitinen und Laitinen, 1998; Beasley et al., 1999; Spathis, 2002). Als dritte finanzielle Variable verwenden Gaganis et al. (2007) *GROWTH*, das jährliche Wachstum des Gesamtkapitals. Laitinen und Laitinen (1998) begründen die Anwendung dieser Variablen mit der Erkenntnis, dass ein geringeres Wachstum mit einer höheren Wahrscheinlichkeit einer Qualifikation verbunden ist.

Die letzte unabhängige finanzielle Variable *LEV* bestimmt den finanziellen Leverage und wird im Modell als Verhältnis vom Eigenkapital zum Gesamtkapital definiert. Ein höherer Wert für die *LEV* bedeutet eine höhere Kapitaldeckung für die Aktionäre zur Finanzierung des Unternehmensvermögens. Der Anteil der Verbindlichkeiten am Gesamtkapitel ist geringer womit auch das das Risiko sinkt, in eine finanzielle Schieflage zu geraten. Diese Variable wird verwendet, da von mehreren Autoren belegt werden konnte, dass mit steigender Wahrscheinlichkeit eines Untergangs der Unternehmung auch eine steigende Wahrscheinlichkeit für eine Einschränkung besteht (Bell und Tabor, 1991; McKeown et al., 1991; Reynolds und Jere, 2000).

Als unabhängige nicht finanzielle Variable verwenden die Autoren *RATING*, die dem Kreditrating entspricht, weil eine erhöhte Wahrscheinlichkeit für finanzielle Schräglagen mit einer erhöhten Wahrscheinlichkeit für Qualifikationen korre-

liert. Konkret setzen die Autoren den *QuiScore* ein, der die Klassifikation eines Unternehmens für die Wahrscheinlichkeit eines Bankrotts in den nächsten zwölf Monaten abbildet. Basierend auf dem Score, der in einem Wertebereich von 0 bis 100 liegt, werden fünf Gruppen gebildet, die für einen Wert von 0 bis 20 als *High Risk* und im gegenteiligen Fall bei einem *Score* von 81 bis 100 als *Secure* eingestuft werden.

Zuletzt nutzen die Autoren die Dummy-Variable *INDUSTRY* zur Kennzeichnung der Unternehmenszugehörigkeit zur verarbeitenden Industrie oder zum Handel, die dem Sample zugrunde liegt.

Die Datengrundlage der Autoren besteht aus 485 Abschlüssen von Unternehmen aus Grossbritannien und Irland, die in den Jahren 1998-2003 eine Einschränkung im Prüfurteil erfahren haben. Zudem wählten sie 970 Firmen aus, die keine Einschränkung im Testat aufwiesen.

Gaganis et al. (2007) berücksichtigten dabei nur Abschlüsse von Unternehmen mit Einschränkungen im Testat, die

1. einen Finanzabschluss in der genannten Zeitperiode publizierten,
2. eine Einschränkung im Finanzabschluss verzeichneten,
3. sich als Grossunternehmen qualifizierten (mit einem Gesamtkapital grösser als 27 Mio. Euro, einem Umsatz von mehr als 40 Mio. Euro und mehr als 250 Mitarbeitern) und
4. entweder in der verarbeitenden Industrie oder im Handel tätig waren.

Demgegenüber wurden zufällig 1'000 Unternehmen aus den Branchen Industrie und Handel ausgewählt. 30 Unternehmen mussten jedoch aufgrund fehlender Informationen im Abschlussbericht des Prüfers ausgeschlossen werden. Das von den Autoren aufgestellte logistische Regressionsmodell hat für die Einschränkungen im Prüfurteil die Genauigkeiten ergeben, wie sie in Tabelle 25 gezeigt werden.

Als Alternative schlagen Kirkos, Spathis, Nanopoulos et al. (2007) die Verwendung eines Entscheidungsbaums, eines neuronalen Netzes sowie eines *Bayesischen Believe Networks* vor – die zuletzt genannte Methode wurde jedoch aufgrund der geringen gegenwärtigen Anwendung nicht weiterverfolgt.

	Unqualifiziert	Qualifiziert
Industrie	68.40	44.90
Handel	44.90	52.70
Generell	70.20	44.90

Tabelle 25: Genauigkeiten des logistischen Regressionsmodells von Gaganis et al. (2007)

Bei der Auswahl der unabhängigen Variablen stützen sich die Autoren auf die einschlägige Literatur und orientieren sich dabei hauptsächlich an Gaganis et al. (2007). Ausgehend von einer grossen Palette von Variablen führen sie im ersten Schritt eine Einweg-Varianz-Analyse durch und evaluieren, ob für die einzelnen Variablen die Differenz in den zwei Gruppen signifikant ist. Falls die Unterschiede gemessen am *p-Wert* nicht signifikant sind, werden die Variablen ausgeschlossen, sodass die folgenden unabhängigen Variablen übrig bleiben: Umsatz, Jahresgewinn vor Steuern, Working Capital, Verschuldungsgrad, Eigenkapitalrendite, Gesamtkapitalrendite, Return-on-Assets, QuiScore, Vorliegen eines Grossunternehmens (*Dummy-Variable*), Bilanzsumme, kurzfristige Verbindlichkeiten, langfristige Verbindlichkeiten, Eigenkapital, Revisionshonorar, *Altman's Z-Score* sowie das Prüfhonorar im Verhältnis zum Gesamtkapital.

Für das Training des Modells stützen sich Kirkos, Spathis, Nanopoulos et al. (2007) auf die Daten aus Grossbritannien und Irland. In diesem Datensatz identifizierten sie für den Zeitraum von 1995 bis 2004 insgesamt 225 Unternehmen, die eine Einschränkung im Prüftestat sahen. Für ein ausgeglichenes Trainingsset zogen sie weitere 225 Unternehmen, die den Unternehmen mit Einschränkung glichen, d.h., bei denen zwei SIC-Ziffern übereinstimmten.

Die Modellierung der Entscheidungsbäume und neuronalen Netze bedurfte einer spezifischen Parametrierung:

- *Entscheidungsbäume*

  Es existieren verschiedene Algorithmen, die aufgrund der Trainingsdaten Entscheidungsbäume erstellen. Die Autoren wählen in ihrer Forschungsarbeit einen *C4.5-Entscheidungsbaumalgorithmus*, der u.a. mit fehlenden Werten und kontinuierlichen Wertebereichen umgehen kann.

- *Neuronale Netze*

  Für das neuronale Netz wählen Kirkos, Spathis, Nanopoulos et al. (2007) eine Topologie, die eine verdeckte Schicht an Neuronen enthält. Die optimale Anzahl der Neuronen der verdeckten Schicht ermittelten die Autoren mit einer 10-Fold-Kreuzvalidierungsmethode. Dabei hat sich gezeigt, dass zehn Neuronen in der verdeckten Schicht im Trainingsset die beste Prognosegüte erreichen. Daneben wurde für die Lernrate ein Wert von 0.15 gewählt.

	unqualifiziert	qualifiziert
C4.5	78.76	76.62
Neural Network	83.78	78.44

Tabelle 26: Genauigkeit in der Klassifizierung der einzuschränkenden Abschlüsse nach Kirkos, Spathis, Nanopoulos et al. (2007)

Die Ergebnisse in Tabelle 26 legen nahe, dass das neuronale Netz mit einer Genauigkeit von 78,44 % die beste Prognosegüte bietet, wenn es darum geht, eingeschränkte bzw. einzuschränkende (Jahres-)Abschlüsse zu identifizieren. Gleichwohl liegt der Ansatz mit einem Entscheidungsbaum, der eine einfache Interpretation der Ergebnisse erlaubt, in dieser Kategorie nur rund 2 % hinter dem Klassifizierungssieger.

Erhebung der Daten Wiederum analog zur Risikobeurteilung und Beurteilung der Unternehmensfortführungsfähigkeit entstammen die Daten, die in die Modelle einfliessen, hauptsächlich aus unternehmensexternen Quellen und nutzen lediglich als Eingabevektor die Daten der Unternehmung selbst.

Für die Gewinnung der externen Datensätze wurde folgende Vorgehensweise gewählt:

- *Unternehmensexterne Daten*

  Für die unternehmensexternen Daten wurde abermals auf die *Compustat-Datenbank* zugegriffen. Dies ermöglichte die Berechnung der Kennzahlen für beispielsweise Liquidität, Profitabilität und Verschuldung für verschiedenste Unternehmen.

Gleichzeitig wurden aus der *Audit-Analytics-Datenbank* die Berichte für alle Unternehmen im Zeitraum von 2015 bis 2020 bezogen. Die entsprechenden Testate mit Einschränkung liessen sich relativ einfach ermitteln, weil die Wirtschaftsprüfer ihren Bericht nach ISA 700 (IAASB, 2009j) abfassen und Einschränkungen stets mit dem gleichen Wortlaut einleiten (*„Basis for Qualified Opinion [...]“*). Dem Datensatz konnten zudem die Kreditratings und die SIC-Unternehmenskategorien zugeordnet werden.

- *Unternehmensinterne Daten*

  Für die unternehmensinternen Daten stellte das Controlling die benötigten Informationen bereit, falls sie noch nicht vorlagen.

  Für das Kreditrating wurde die Einschätzung zum Konzern herangezogen, da auf Stufe der Einheit kein derartiges Rating vorlag.

Erlangung des Geschäfts- und Datenverständnisses Im abschliessenden Prüfprozess, der Berichterstattung, waren keine zusätzlichen Informationen über die Unternehmung mehr notwendig. Zu einzelnen Punkten wurden lediglich Daten geprüft und ihre Qualität mit dem verantwortlichen Mitarbeiter der Organisation diskutiert.

Datenvorbereitung Die Schritte, welche im Rahmen der Datenvorbereitung durchgeführt werden mussten, waren mit denen der Unternehmensfortführung deckungsgleich und umfassten:

1. *Korrektur von Lücken*
2. *Integration der Daten*
3. *Abgleich der Unternehmen*
4. *Bildung neuer Features*
5. *Standardisierung der Daten*

Modellierung der Daten Der Vorgehensweise von Gaganis et al. (2007) und Kirkos, Spathis, Nanopoulos et al. (2007) folgend wurden anhand der aufbereiteten Daten drei Modelle gebildet, die eine Prognose für die Qualifizierung oder Nichtqualifizierung eines Prüfurteils zuliessen.

Die Daten wurden in Trainingsdaten (80 %) und in Testdaten (20 %) zufällig eingeteilt und anschliessend die folgenden Modelle generiert:

1. *Logistisches Regressionsmodell*

   Es wurde ein logistisches Regressionsmodell auf Grundlage der gesamten Trainingsdaten sowie der von Gaganis et al. (2007) vorgeschlagenen unabhängigen Variablen berechnet.

2. *Entscheidungsbaum*

   Als weiteres Modell wurde in Anlehnung an Kirkos, Spathis, Nanopoulos et al. (2007) ein Entscheidungsbaum modelliert. In Abweichung zum Vorgehen der Autoren wurde aus pragmatischen Gründen – ein C4.5-Entscheidungsbaum-Algorithmus sieht die Scikit-Learn-Analysesoftware im Standard nicht vor – auf den CART-Algorithmus zurückgegriffen, der einfach verfügbar ist und vergleichbare Ergebnisse liefert.

3. *Künstliches neuronales Netz*

   Zum Schluss stand der Versuch, ein neuronales Netz für die Identifikation von zu qualifizierenden Prüfurteilen zu trainieren. Ausgehend von den Konfigurationen der Autoren wurden mehrere Grössen und eine bestimmte Anzahl von versteckten Schichten auf einem Validierungsset getestet. So wurden initial eine verdeckte Schicht mit 14, 10, 8 und 24 Neuronen sowie eine Topologie bestehend aus zwei Schichten mit jeweils 14, 10 und 8 Neuronen getestet.

   Die *Regularisierung* und die *Lernrate* wurden mit einer *10-fold-Kreuzvarlidieren*, gebildet auf dem Trainingsset, und einem Gridsearch in einer Art und Weise optimiert, dass sich die besten Prognosen für das jeweilige Validierungsset ergaben.

Evaluation Die Darstellung der Konfusionsmatrix soll Aufschluss über die Leistungsfähigkeit der jeweiligen Modelle geben (Q steht hierbei für qualifiziert und NQ für nicht qualifiziert). Die Berechnung ergab für ein logistisches Regressionsmodell die in Tabelle 27 angegebenen Leistungskennzahlen.

	Erwart. Q	Erwart. NQ	%	Total %
Tatsächlicher Q	33	3	91.66	65.82
Tatsächlicher NQ	24	19	44.18	

Tabelle 27: Konfusionsmatrix für die Qualifizierung des Prüfurteils mithilfe des logistischen Regressionsmodells

Bei Verwendung eines Entscheidungsbaums ergeben sich die in Tabelle 28 ermittelten Werte.

	Erwart. Q	Erwart. NQ	%	Total %
Tatsächlicher Q	28	8	77.78	56.96
Tatsächlicher NQ	26	17	39.53	

Tabelle 28: Konfusionsmatrix für die Qualifizierung von Prüfurteilen mithilfe des Entscheidungsbaum

Für die Anwendung eines künstlichen neuronalen Netzes im Rahmen der Modellierung von möglichen Einschränkungen im Prüfurteil stellten sich die in Tabelle 29 genannten Werte ein.

	Erwart. Q	Erwart. NQ	%	Total %
Tatsächlicher Q	27	9	75.00	68.35
Tatsächlicher NQ	16	27	62.79	

Tabelle 29: Konfusionsmatrix für die Qualifizierung des Prüfurteils mithilfe eines künstlichen neuronalen Netzes

Die Ergebnisse machen deutlich, dass das logistische Regressionsmodell die Leistungsfähigkeit, gemessen an der Genauigkeit, von Gaganis et al. (2007) zu erreichen vermochte und diese gar übertraf.

Die Werte von Kirkos, Spathis, Nanopoulos et al. (2007) wurden hingegen nicht erreicht und lagen rund 10 % hinter den publizierten Resultaten. Dabei gilt es zu bemerken, dass die für die Prüfung wichtigen *Typ-1 Fehler* jedoch sehr vergleichbar waren.

Bereitstellung der Modelle Die Bereitstellung der Modelle erfolgte analog zur Analyse der Unternehmensfortführung, da sich dieser Ansatz bewährt hat. Insbesondere auch die Bereitstellung eines zusätzlichen *Jupyter Notebook Light*, das den Anwendern eine Generierung einfacher Prognosen für eine Unternehmung erlaubt.

Implikationen für Prozesse und Herausforderungen Die bereits in den vorangegangenen Prüfprozessen diskutierten Implikationen gelten auch für die mit dem maschinellen Lernen unterstützte Berichterstattung. Darüber hinaus ergaben sich weitere Implikationen:

- *Strukturierte Prüfdokumentation*

  In der letzten Prüfphase, der Prüfbeendigung, spielt der Review des Mandatsmanagements eine bedeutende Rolle. Ziel dieser Phase ist zu gewährleisten, dass die Prüfarbeiten mit der angemessenen Prüfevidenz abgeschlossen wurden, um aus den gewonnenen Erkenntnissen ein belastbares Prüfurteil abzuleiten. Ferner muss sichergestellt sein, dass die Prüfung nach den einschlägigen Vorgaben durchgeführt wurde.

  Während der Prüfarbeiten muss stets festgehalten werden, welche Prüfziele erreicht werden sollen, welche Arbeitsschritte unternommen wurden und welche Resultate bzw. Konklusionen aus den Erhebungen abzuleiten sind.

  Als Paradigma wurde diese Aufteilung für alle Arbeitsblätter bzw. *Jupyter-Notebooks* übernommen und die Dokumentation dreigeteilt:

  1. Definition des Prüfziels
  2. Durchgeführte Prüfungshandlungen mit einem vollständig kommentierten Code
  3. Resultate und Konklusionen aus der Prüfarbeit

  Mit der Anwendung dieser Dokumentationsschritte konnten die Kommunikation mit dem Mandatsmanagement wesentlich erleichtert und die Einhaltung der Prüfvorgaben sichergestellt werden.

- *Spezifikation der verwendeten Grössen*

  In Bezug auf die Verwendung von Variablen im Training anlässlich der Berichterstattung hat sich erneut gezeigt, dass diese Grössen – wie ein Betriebsaufwand – sehr genau spezifiziert sein müssen, damit die Funktionsweise der Modelle und deren hohe Prognosegüte sichergestellt wird und für die Anwender bei der Berechnung nur wenig Unsicherheit besteht.

4.4 Fallstudie 2

4.4.1 Rahmenbedingungen

Prüfkunde In einer zweiten Fallstudie wurde maschinelles Lernen in einer mittelgrossen Organisation angewendet, die als öffentlich-rechtliche Unternehmung bzw. als technische Betriebe Leistungen in einer Gemeinde bereitstellt. Neben diesen technischen Betrieben, wurde die Prüfgesellschaft zudem beauftragt, die Gemeinderechnung sowie weitere kleinere Einheiten zu prüfen, die nicht Gegenstand dieser Fallstudien sind.

Für die Gemeinde mit 10'000 Einwohnern stellen die technischen Betriebe folgende Dienstleistungen zur Verfügung:

- *Energieversorgung*

  Die Organisation versorgt Privat- und Geschäftskunden mit Strom. Dabei wir ein Teil der Energie selbst hergestellt und ein anderer Teil von einem kantonalen Stromversorger bezogen.

- *Wasserversorgung*

  Die technischen Werke stellen Wasser für Haushalte und Unternehmen der Gemeinde und für die umliegenden Kommunen zur Verfügung. Dieses Wasser wird überwiegend aus den eigenen Quellen bezogen.

- *Kommunikation*

  Ausserdem bietet die Organisation Telekommunikationsdienstleistungen an. Das breite Kommunikationsangebot schliesst digitales Fernsehen und Radio, Festnetztelefonie, Mobilfunk und Internet ein.

Die Unternehmung ist in dieser Form seit zehn Jahren aktiv und erbringt jährlich eine Jahresabschlussrechnung, bestehend aus Bilanz, Erfolgsrechnung, Investitionsrechnung, Geldflussrechnung und Anhang, zuhanden ihrer Bürger. Die Rechnungslegung folgt den einschlägigen kommunalen und kantonalen Vorgaben. Daneben haben die Betriebe auch ein internes Kontrollsystem aufgebaut, das die Prüfgesellschaft jährlich prüft.

Über alle Dienstleistungen hinweg erwirtschaften die Betriebe jährlich einen Umsatz von rund 10 Mio. Schweizer Franken. Diesen Einnahmen stehen Ausgaben in etwa gleicher Höhe gegenüber. Die Bilanzsumme beträgt per Revisionszeitpunkt rund 20 Mio. Schweizer Franken und ist massgeblich als Sachanlagen im Bereich des Tiefbaus gebunden. Der Eigenkapitalanteil ist mit über 95 % sehr hoch, sodass keine Zweifel an der Bonität der Betriebe bestehen.

Prüfteam Wie bereits für die Fallstudie 1 wurde auch für diese Fallstudie ein grösseres Team von der Prüfgesellschaft abgestellt.

Das Team umfasste folgende Personen:

1. *Mandatsleiter bzw. -partner:* Der Mandatsleiter war für die Beurteilung der Risiken und die daraus folgende Prüfstrategie verantwortlich. Aufgrund der Komplexität des Mandats mit den zu beachtenden rechtlichen Auflagen, die über den Rahmen eines eines üblichen Audits hinausgeht – insbesondere hinsichtlich der Gemeinderechnung und der weiteren Betriebe –, war der Mandatspartner in die Prüfung operativ eingebunden und übte damit eine Rolle aus, die nicht nur die Pflege der Kundenbeziehung und die Sicherstellung der Prüfqualität beinhaltete.

2. *Mandatsmanager:* Dem Mandatsmanager kam im Rahmen der Abschlussprüfung die Aufgabe zu, das Mandat mit dem Kunden aufzusetzen, sodass während der Abschlussprüfung sämtliche Unterlagen zur Verfügung standen und gleichwohl für das Prüfteam auch die benötigte Infrastruktur mit entsprechend konfigurierter Software vorhanden war. Während der Abschlussprüfung vor Ort jedoch übernahm der Mandatspartner die operative Leitung und stand für Rückfragen zur Rechnungslegung oder dem Prüfansatz zur Verfügung.

3. *Mandatsassistenten:* Die Prüfgesellschaft setzte für die Abwicklung der Prüfung insgesamt drei Assistenten ein, die mit der operativen Prüfungsdurchführung betraut waren.

Der Studienleiter durfte dieses Team während der Abschlussprüfung begleiten.

Infrastruktur Dem Studienleiter wurde ein aktuelles Notebook zur Verfügung gestellt, das in Bezug auf die Installation neuer Software und Komponenten keinerlei Beschränkungen unterlag.

Prüfungsdurchführung Der Studienleiter konnte die Prüfung der oben genannten Jahresrechnung im Februar 2020 während der gesamten Phase der Abschlussprüfung mit dem Prüfteam begleiten. Aufgrund der beschränkten Grösse des Mandats liess sich die Prüfung in dieser Zeit durchführen.

Fertiggestellt wurde die Abschlussprüfung am 28. Februar 2020 mit der Berichterstattung zuhanden der Gemeindeversammlung.

4.4.2 Vorgehen in der Fallstudie

Das Vorgehen in Fallstudie 2 orientiert sich grundsätzlich an jenem der ersten Fallstudie, stellt die Ansätze jedoch in konsolidierter Weise dar. Dies hat folgende Hintergründe:

1. *Theoretische Grundlagen vorhanden:* Die Grundlage für die Vorgehensweise in den einzelnen Prüfphasen wurde bereits anlässlich der Zwischenrevision in der Fallstudie 1 erarbeitet und ist damit bereits vorhanden.

2. *Kleiner Umfang der Prüfung:* Es handelt sich beim Prüfkunden in Fallstudie 2 gemessen am Umsatz um ein vergleichsweise kleines Mandat. Entsprechend ist auch der Prüfung durch die Prüfgesellschaft weniger Zeit eingeräumt worden.

3. *Evaluation mit gleichen Resultaten:* Auch für die Evaluation gilt, dass diese aufgrund der Verwendung übereinstimmender (unternehmensexterner) Daten, die gleichen Ergebnisse zutage förderte, weshalb für diesen Schritt auf Fallstudie 1 verwiesen sei.

Fallstudie 1 zeigt, dass eine grundsätzliche Unterscheidung der Ansätze möglich ist und sich diese in zwei Gruppen einteilen lassen.

1. *ML-Ansätze aufbauend auf externen Daten*

Die Mehrheit der Ansätze in der einschlägigen Literatur, die auf externen Daten aufbauen, sind für die Prüfphasen der Risikobeurteilung und Berichterstattung vorgesehen. Diesen Ansätzen ist gemein, dass sich das einmal berechnete Modell auf mehrere Unternehmen anwenden lässt und für das Training hauptsächlich auf unternehmensexterne Daten zurückgegriffen wird.

2. *ML-Ansätze aufbauend auf internen Daten*

   Eine Minderzahl der in der Forschung beschriebenen Ansätze lässt die Modellierung allein aufgrund von internen Daten zu. So werden für die analytischen Prüfungshandlungen und teilweise auch die Aufdeckung doloser Handlungen lediglich interne Daten verwendet. Die Modellierung muss für jede Unternehmung einzeln durchgeführt werden.

4.4.3 ML-Ansätze aufgrund externer Beobachtungen

Betrachtete Ansätze Bei den betrachteten Ansätzen, die eine Prognose aufgrund von externen abhängigen Variablen ermöglichen, wurde auf die in Fallstudie 1 untersuchten und angewandten Methoden (siehe Tabelle 30) zurückgegriffen.

Erhebung der Daten Die Akquisition der Daten für die Modellierung der Prognose der verschiedenen Modelle erfolgte in zwei Schritten:

1. *Akquisition unternehmensexterner Daten*

   Die Modelle erfordern in den Bereichen Risikoanalyse, deliktische Handlungen, Unternehmensfortführung und Berichterstattung unabhängige und abhängige Merkmale für zahlreiche Beobachtungen, die in der Regel Grössen aus den Unternehmensabschlüssen bzw. deren Testaten sind. Diese Daten wurden fast vollständig aus der Compustat-Datenbank entnommen und mussten nur an wenigen Stellen mit Daten aus anderen Datenquellen ergänzt werden (siehe Fallstudie 1).

   Entsprechend liess sich die Prognose fast vollständig mit den bereits in Fallstudie 1 zur Verfügung stehenden Daten durchführen.

2. *Akquisition unternehmensinterner Daten*

Autoren	Methode
Risikobeurteilung	
Dikmen und Küçükkocaoğlu (2010)	logistische Regression
Dbouk und Zaarour (2017)	naive Bayes
Dechow et al. (2011)	logistische Regression
Evaluation Deliktische Handlungen	
J. L. Perols et al. (2017)	Support-Vektor-Maschine
Fortführungsfähigkeit	
Jones (2017)	RandomForests
Anandarajan et al. (2001)	Neuronales Netz
Martens et al. (2008)	Support-Vektor-Maschine
	Logistische Regression
Berichterstattung	
Gaganis et al. (2007)	logistische Regression
Kirkos, Spathis, Nanopoulos et al. (2007)	Entscheidungsbäume
	Neuronales Netz

Tabelle 30: Übersicht über Autoren und zugehörige Methoden

Zur Generierung einer Prognose, wie beispielsweise, ob ein erhöhtes Risiko für den Prüfkunden erwartet oder gegebenenfalls eine eingeschränkte Unternehmensfortführungsfähigkeit vorliegt, besteht der nötige Inputvektor für das Modell stets aus unternehmensspezifischen Werten, wie beispielsweise aus dem Umsatzwachstum.

Wie bereits in Fallstudie 1 sind diese Daten in den allermeisten Fällen das Resultat verschiedener Aggregationsstufen bzw. Abschlussposten in Bilanz und Erfolgsrechnung. Als Prüfobjekt der eigentlichen Abschlussprüfung müssen diese Daten damit vorliegen, was in Fallstudie 2 mit der Schlussrevision der Fall war.

Der Abruf der Abschlusssalden auf der Stufe der Jahresrechnung erfolgte per Abschlussprüfung vom internen Portal der Prüfgesellschaft, weil der Prüfkunde diese Informationen bereits vorab zur Verfügung gestellt hatte. Daneben war es möglich, auf die verschiedenen Aggregationsstufen der Pos-

ten mithilfe zusätzlicher Dokumentation zuzugreifen.

Erlangung des Daten- und Geschäftsverständnisses Das Verständnis des Geschäfts und der Daten musste wiederum separat für die unternehmensexternen und die unternehmensinternen Daten erarbeitet werden:

- *Unternehmensexterne Daten*

  Ein Verständnis für die unternehmensexternen Daten wurde aus der ersten Fallstudie übernommen.

  Viele Autoren extrahierten aufgrund der SIC-Klassifizierungen der Unternehmen die Daten für die Modelle. Obwohl es sich um eine öffentlich-rechtliche Einrichtung handelt, wurde entschieden, das Unternehmen als Elektrizitätsproduzent zu klassifizieren (die ersten beiden SIC-Ziffern 49). Das Unternehmen generiert 90 % seiner Umsätze im Elektrizitätsgeschäft, das damit die absolut wichtigste Sparte ist.

- *Unternehmensinterne Daten*

  Im Rahmen eines zu Beginn des Prüfungsprozesses geführten Gesprächs wurden die einzelnen Posten mit dem Prüfkunden und der Mandatsleitung diskutiert, zumal die Rechnungslegungsgrundsätze einige Ermessensspielräume und Wahlrechte kennen, die einen potenziellen Einfluss auf den Eingabevektor haben können.

Vorbereitung der Daten Zur Verwendung der Daten in den Modellen waren wiederum zahlreiche Schritte nötig.

1. *Bereinigung der Daten bzw. Korrektur von Lücken*

   Der erste Schritt sah die Bereinigung der Daten von Datenlücken vor. So liegen in der *Compustat-Datenbank* Lücken für Finanzabschlüsse vor, die ausgeschlossen bzw. mithilfe von Mittelwerten korrigiert wurden.

2. *Bildung neuer Features*

   Nahezu alle Variablen, die in die einzelnen Modelle eingehen, sind das Resultat einer Verrechnung mit Vorperioden, beispielsweise für die Berechnung

der Wachstumsrate oder einer anderen Kennzahl, um z.B. den Lagerumschlag zu bestimmen. Entsprechend mussten diese Werte berechnet und als neue unabhängige Variablen zur Verfügung gestellt werden.

3. *Integration der Datenquellen*

   Die Beobachtungen, die in die Modelle einfliessen, nehmen teilweise Bezug auf verschiedene Datenbestände. So werden für die Prognose einer Einschränkung im Testat folgende Daten hinzugezogen:

   - Unternehmen mit ihrem Prüfurteil für verschiedene Jahre
   - Unternehmensfinanzdaten
   - Bonitätsausweise

   Folglich müssen sie zusammengeführt bzw. integriert werden.

4. *Unternehmensabgleich*

   Mehrere Autoren greifen für das Training ihrer Modelle auf einen Vergleich mit anderen Unternehmen zurück. Für jedes Unternehmen, das beispielsweise eine Einschränkung im Testat verzeichnet, wird ein vergleichbares Unternehmen (häufig aufgrund einer Übereinstimmung mit den ersten zwei Ziffern in der SIC-Klassifizierung) ohne Einschränkung hinzugezogen.

5. *Standardisierung bzw. Normalisierung der Wertebereiche*

   Verschiedene Modelle, die zum Einsatz kommen, sind nicht skaleninvariant – wie beispielsweise die Support-Vektor-Maschine oder künstliche neuronale Netze. Die Wertebereiche der unabhängigen Variablen müssen daher standardisiert bzw. normalisiert werden.

Modellierung der Daten Die Berechnungen der Modelle, wie sie für die erste Fallstudie vorbereitet wurden, eignen sich auch für die vorliegende Fallstudie und mussten daher nicht noch einmal neu erstellt werden.

Die Modelle für vergleichbare Unternehmen und Branchen der Fallstudie 1 wurden agnostisch trainiert. Folglich sind die daraus resultierenden Prognosen auch für die Fallstudie 2 gültig.

Evaluation Von einer Evaluation bzw. Validierung der Modelle wird gleichfalls abgesehen. Es gilt, dass die Modelle unverändert bleiben konnten und damit die bereits erhobenen Leistungskennzahlen weiterhin Geltung haben.

Bereitstellung der Modelle Die Bereitstellung der Modelle in einer Form, dass sie einem Prüfer eine einfache Prognose ermöglichen, lehnt sich nahtlos an die Erkenntnisse der ersten Fallstudie an.

Es wurden wiederum für jedes Modell zwei Notebooks aufgesetzt:

1. *Primäres Juypter Notebook*

   Analog zur ersten Fallstudie wurde ein primäres Notebook zum Modelltraining zusammengestellt. Dazu war es gleichfalls notwendig,

   - die Daten zu laden,
   - die Daten zu sichten und zu analysieren,
   - Transformations- und Integrationsschritte durchzuführen,
   - Features zu bilden,
   - die Modelle zu berechnen.

2. *Sekundäres Jupyter Notebook*

   Die Light-Version des primären *Jupyter Noteboks* gewährleistet mithilfe von Datenverarbeitungsschritten und mitgegebenen Parametern die Generierung von Prognosen.

Das primäre *Jupyter-Notebook* konnte von der ersten Fallstudie übernommen werden, während die sekundären Notebooks für die Aufbereitung der unternehmensspezifischen Grössen neugestaltet werden mussten.

4.4.4 ML-Ansätze aufgrund unternehmensinterner Beobachtungen

Betrachtete Ansätze Es wurde keine spezifischen Abhängigkeit von Ansätzen des maschinellen Lernens und den Fallstudien erkannt, sodass im Folgenden wiederum auf die Methoden, wie sie für die Fallstudie 1 bereits verwendet wurden, zurückgegriffen werden konnte.

Konkret fanden die in Tabelle 31 aufgeführten Methoden Verwendung.

Autoren	Methode
Analytische Prüfungshandlungen	
Koskivaara (2000a)	künstl. neuronales Netz
Koskivaara und Back (2007)	künstl. neuronales Netz
Dolose Handlungen	
Thiprungsri und Vasarhelyi (2011)	Clustering
Jans et al. (2010)	Clustering
Tackett (2013)	Assoziationsanalyse

Tabelle 31: Übersicht über Autoren und zugehörige Methoden

Erhebung der Daten Als Eingangsdaten für die Modelle mussten wiederum unternehmensexterne sowie -interne Daten beschafft werden.

1. *Unternehmensinterne Daten*

   Der Risikoanalyse des Mandatsmanagements folgend wurden Umsatzdaten für einzelne Produkte aus dem ERP-System bezogen.

   Als Schwierigkeit für das Controlling, das die Daten zur Verfügung stellte, hat sich der Abzug über die verschiedenen Jahre erwiesen, zumal der Kontenplan mehrmals angepasst wurde. Es waren somit verschiedene Umstrukturierungen erforderlich, um die Werte aus vorher geprüften Abschlüssen zu reproduzieren.

   Damit sichergestellt war, dass keine Fehler in diesem Prozess unterlaufen sind, wurden die resultierenden Kontensalden mit den Jahresrechnungen der vergangenen Jahre abgeglichen. Gleichzeitig erfolgte für eine Stichprobe von je 25 Einträgen ein Abgleich zwischen erhaltenem Buchungsjournal und dem ERP-System.

   Schlussendlich konnte das Controlling die Buchungen von fünf vergangenen Jahren aus dem ERP-System in das *CSV-Format* exportieren.

2. *Unternehmensexterne Daten*

   Für die zweite Fallstudie wurden unternehmensexterne Daten hinzugezogen, um die Prognosegüte positiv zu beeinflussen. Für die Ermittlung erklärender Features wurde im Rahmen der Mandatsplanung eine Liste von Spezialisten

in der Organisation – wie der technische Leiter – erstellt und sodann systematisch ein regelmässiger Austausch vereinbart. Anhand der gewonnenen Informationen ergab sich eine Reihe von potenziellen erklärenden Variablen, die für die Prognosen relevant waren (z. B. Bevölkerungsentwicklung, Schulferien, Witterungsbedingungen usw.).

Diese Daten wurden vollständig aus unterschiedlichen externen Quellen in einem für *Jupyter Notebooks* lesbares Format bereitgestellt.

Erlangung des Datenverständnisses Für die öffentlich verfügbaren unternehmensexternen Daten liegen Dokumentationen an der Datenquelle vor. Diese Möglichkeit ist bei unternehmensinternen Daten für diese Organisationseinheit nicht gegeben.

Obwohl die Daten während der Datenakquisition bereits besprochen wurden, konnten mit dem Controlling die 45 Merkmale diskutiert werden, die eine Transaktion kennzeichnen. Die Merkmale reichten von der Kundennummer über die verkaufte Leistung und das Gegenkonto bis hin zum verantwortlichen Sachbearbeiter und Verbuchungszeitpunkt. Eine Limitierung der Daten lag nicht vor, jedoch mussten weiteren Tabellen angefragt werden, die beispielsweise für jeden Kunden erweiterte Merkmale enthielten.

Vorbereitung der Daten Für die Verwendung in den Modellen waren abermals folgende Schritte der Datenaufbereitung notwendig:

1. *Bereinigung der Daten bzw. Korrektur von Lücken*

   Bei den Daten, die nicht aus Fallstudie 1 wiederverwendet werden konnten, mussten zu Beginn Datenlücken bereinigt werden. Diese Daten wurden – abgestützt auf *Best Practices* – entweder ganz aus der Beobachtung gestrichen oder mithilfe von Mittelwerten über die Features ergänzt.

2. *Integration der Datenquellen*

   Für sämtliche Modelle wurden zehn Datenquellen erschlossen, die in der Vorverarbeitung über das Datum, die Kundenkennung oder Fremdwährungsnamen miteinander verbunden wurden.

3. *Bildung neuer Features*

   Neue Features wurden aus den Daten abgeleitet, um anschliessend die Modelle zu trainieren.

   Für die analytischen Prüfungshandlungen wurden neue Features für die *Jahreszeit* und den *Monat im Jahr* aus den Valuta-Daten berechnet. Zudem wurden für die Produktgruppen *Dummy-Variablen* angelegt.

   Für die Erkennung von dolosen Handlungen wurden zusätzlich verhaltensorientierte *Features* gebildet (z.B. *Buchungen am Wochenende* und *Beträgen im oberen Perzentil*).

4. *Standardisierung der Wertebereiche*

   Für die Verwendung der Ansätze gilt, dass nicht alle skaleninvariant sind, wie die *Support-Vektor-Maschine* oder künstlichen neuronalen Netze. Die Wertebereiche der unabhängigen Variablen müssen daher in Abhängigkeit des Modells standardisiert werden.

Modellierung der Daten Für die Berechnung der Modelle nach den oben angeführten Methoden wurde zudem ein lineares Regressionsmodell als *Benchmark*-Ankerpunkt für die Einschätzung der relativen Prognosegüte der restlichen Modelle verwendet.

Die Modellierung stützte sich auf Methoden des überwachten Lernens mit einem reduzierten Datensatz, bei dem vorab 20 % der Beobachtungen zur späteren Evaluation der Leistungsfähigkeit aufgehoben wurden. Damit entstanden sowohl ein Trainings- als auch ein Testdatensatz.

Die gewählten Methoden machten es ferner notwendig, eine Reihe von Hyperparametern zu optimieren:

1. *Random Forest*

   Für den Random Forest wurden aus einer grossen Menge an möglichen Parametern die *Anzahl verwendeter Entscheidungsbäume* und die *maximale Baumtiefe* optimiert.

2. *Support-Vektor-Maschine*

Die Support-Vektor-Maschine wurde hinsichtlich der Parameter C (Regularisierung) und *gamma* optimiert.

3. *Künstliche Neuronale Netze*

   Schliesslich wurden die neuronalen Netze in Bezug auf die *Lernrate* und ihrer *Regularisierung* optimiert.

Für alle diese Ansätze gilt, dass die Optimierung mittels einer randomisierten (Grid-)Suche durchgeführt wurde. Die dafür notwendigen Daten wurden aus dem Trainingsset mit einem *10-Fold-Kreuzvalidierungsset* abgeleitet.

Die Geschwindigkeit für die Berechnung der Modelle deckte sich mit der aus Fallstudie 1. So benötigte die Berechnung der Support-Vektor-Maschine aufgrund der grossen Anzahl von Beobachtungen längere Zeit, während der Arbeitsspeicher erheblich beansprucht wurde, auch wenn sich die Anwendung des Ansatzes als relativ einfach herausstellte. Weniger ressourcenintensiv bei gegebenen Beobachtungen hat sich die Berechnung neuronaler Netze herausgestellt, die in angemessener Zeit berechnet werden konnten, auch wenn dazu eine Vielzahl von Topologien ausprobiert werden musste. Problemlos liessen sich die Modelle mit dem Random Forest berechnen: Die Resultate waren bereits in wenigen Minuten verfügbar, während die Hardware nur einer mittleren Belastung ausgesetzt war.

Für die Aufdeckung von dolosen Handlungen anhand von internen Unternehmensdaten wurde mit der Clustering-Methode und der Assoziationsanalyse gearbeitet. Im Falle des Clustering musste als Teil der Modellierung die Anzahl der Cluster gewählt werden. Für die zu bestimmende optimale Anzahl der Gruppen wurde die Ellenbogenmethode eingesetzt. Die Parametrisierung der Assoziationsanalyse erfolgte nach den Vorgaben aus Fallstudie 1, wobei für beide nicht überwachten Methoden galt, dass sie vergleichsweise schnell berechnet werden konnten.

Bereitstellung der Modelle Eine Bereitstellung der Modelle wurde durch die Persistierung der Parameter in einem spezifischen Datenformat (*Pickles-Files*) und durch die Programmierung in Jupyter Notebooks mit entsprechender ausführlicher Kommentierung ermöglicht.

Trotzdem kommt der Bereitstellung wenig Bedeutung zu. So produziert der Prüfkunde konstant neue Daten im ERP-System, indem er beispielsweise seine Leistungen an den Endkunden vertreibt und dafür Rechnungen ausstellt. Dieses Vorgehen impliziert, dass das Modell in regelmässigen Abständen neu berechnet werden kann, wenn es wesentliche Veränderungen in der Unternehmung oder dem Umfeld gibt.

Insofern eignet sich die Verwendung der *Jupyter Notebooks*, um eine Nachvollziehbarkeit und eine einfache Reproduzierbarkeit zu gewährleisten, damit sich bei Folgeprüfungen sehr effizient neue Modelle berechnen und Prognosen erstellen lassen.

5 Analyse

5.1 Angewandte Konstrukte

5.1.1 Zwei Konstrukte

Es haben sich zwei Konstrukte im Laufe der Forschungsarbeit mit dem Prüfteam und den Prüfmandanten herausgebildet. Diese unterscheiden sich massgeblich in Bezug auf die herangezogenen Daten:

- *Unternehmensexterne Labels*

  Ein erstes Konstrukt verwendet für das Training der Modelle ausschliesslich unternehmensexterne Beobachtungen.

  Ist ein Modell schliesslich trainiert, werden bei einem Prüfkunden unternehmensinterne Daten als Eingabevektor für eine Prognose herangezogen (siehe Abbildung 25).

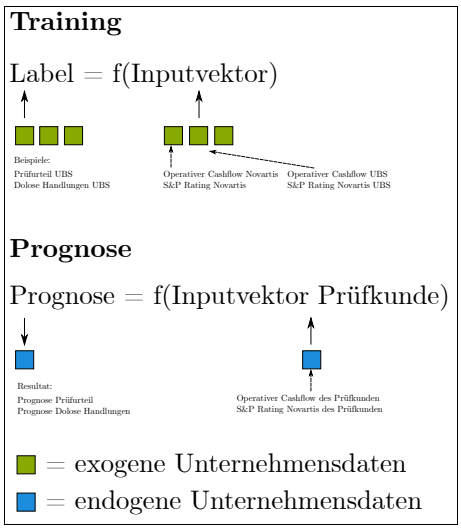


Abbildung 25: Training aufgrund externer Daten mit Prognose eines Vektors (eigene Darstellung)

  Dieser *exogene Anwendungsfall* unterscheidet sich von einem zweiten Fall, weil bei seinem Training ausschliesslich *Labels* (abhängige Variablen) aus dem *Unternehmensumfeld* Verwendung finden.

Damit kommt Datenbeständen, die ausserhalb einer Mandatsbeziehung liegen – seien es Daten, die die Prüfgesellschaft selbst vorhält oder die von Drittdatenbanken abgerufen werden müssen –, eine wichtige Rolle zu.

Dieses Konstrukt wurde bei folgenden Prüfgebieten der Abschlussprüfung eingesetzt:

- Risikobeurteilung
- Beurteilung der Anfälligkeit für dolose Handlungen
- Analytischen Prüfungshandlungen
- Beurteilung der Unternehmensfortführungsfähigkeit
- Aufdeckung wesentlicher Fehldarstellungen bzw. Evaluation, ob Qualifizierung des Prüfurteil vorzunehmen ist

Die aufgelisteten Prüfgebiete zeigen, dass Ereignisse, die viele der Modelle versuchen vorherzusagen, wie beispielsweise das Vorliegen einer dolosen Handlung oder einer eingeschränkten Unternehmensfortführung, nur selten bei einer Prüfung des Jahresabschlusses aufgedeckt werden und damit in den Datensätzen stark untervertreten sind. Dies hat Auswirkungen auf die Datengewinnung, die Datenmodellierung und die Interpretation der Ergebnisse.

- *Unternehmensinterne Beobachtung*

  Ein zweites Konstrukt setzt beim Training der Modelle massgeblich auf *unternehmensinterne Beobachtungen*. So werden als *Features* bzw. unabhängige Variablen beispielsweise Kontenbestände, Debitorenname und Buchungstexte verwendet, während als *Labels* bzw. abhängige Variablen z.B. Kontensalden herangezogen werden. Die Vorgehensweise wird in Abbildung 26 illustriert.

  Der Rückgriff auf unternehmensinterne Daten schliesst die Nutzung externer Daten jedoch nicht aus. Die Integration von externen Daten verbessert im Gegenteil die Prognosegüte der damit gebildeten Modelle und erlaubt mithilfe des Rückgriffs auf unabhängige Daten die Einbeziehung von Daten mit hoher Verlässlichkeit.

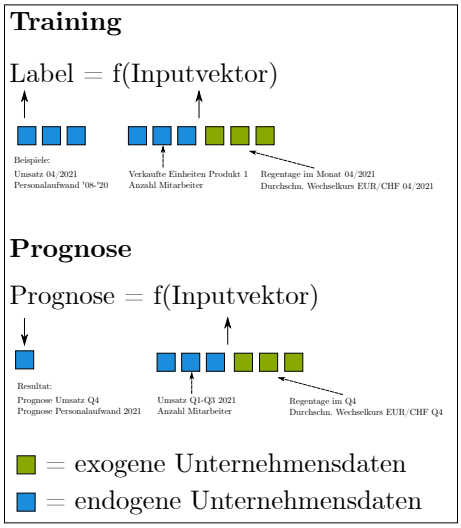


Abbildung 26: Training aufgrund interner Daten mit der Prognose eines Vektors (eigene Darstellung)

Dabei muss jedoch jeder aufgenommene externe Datenpunkt in Beziehung zur Beobachtung gesetzt werden. Konkret ist es sehr wohl möglich und auch wünschenswert, zur Verbesserung der Prognose für die Absatzmenge den Fremdwährungskurs einzubeziehen, wenn diese Ergänzung nach dem Geschäftsverständnis sinnvoll erscheint. Dieser Kurs muss jedoch, falls monatliche Werte für den Umsatz vorliegen, auch auf monatlicher Basis, d.h. als Monatsmittelkurs oder Monatsendkurs, der monatlichen Umsatzbeobachtung zugewiesen werden.

Das Konstrukt ist für folgende Prüfgebiete einsetzbar:

- Analytische Prüfungshandlungen
- Beurteilung für die Anfälligkeit von dolosen Handlungen

Wie der Liste zu entnehmen ist, trifft die Beurteilung der Anfälligkeit für dolose Handlungen für beide Konstrukte zu, weil eine derartige Beurteilung auf der Ebene des Abschlusses als Ganzes und auf der Ebene von einzelnen Transaktionen oder Saldenbeständen möglich ist.

Da sich dieser Anwendungsfall dadurch kennzeichnet, dass beim Training der Modelle lediglich unternehmensinterne *Labels* zum Einsatz kommen, soll er

im Folgenden als *endogener Anwendungsfall* bezeichnet werden.

5.1.2 Notwendige Infrastruktur

Während sich die beiden Anwendungsfälle bezüglich ihrer Datenherkunft erheblich voneinander unterscheiden, kann bei der Infrastruktur auf denselben *Softwarestack* und auf die gleiche Hardware zurückgegriffen werden:

- *Eingesetzte Software*

  Im Laufe der Fallstudie haben sich folgende zusätzliche Ansprüche offenbart, die die Software zur effektiven und effizienten Umsetzung maschinellen Lernens im Prüfumfeld erfüllen muss:

  - Laden und Transformation der Daten
  - Visualisierung der Daten
  - Bereinigung der Daten
  - Integration verschiedener Datenquellen
  - Berechnung verschiedener Modelle
  - Speicherung und Archivierung von Trainingsdaten
  - Speicherung und Laden der Parameter für die Modelle
  - Dokumentation der Schritte
  - Lokale Verarbeitung

  Laden und Transformation der Daten: Üblicherweise werden im Vorfeld zur Datenanalyse Techniken und Software eingesetzt, die es erlauben, aus einem existierenden Datenbestand Daten zu extrahieren, aufzubereiten, zu transformieren und anschliessend in ein neues System zu überführen.

  Für die Fallstudien war es nicht notwendig, die Daten selbst aus den Buchhaltungs- bzw. ERP-Systemen zu extrahieren. Die geprüften Unternehmen konnten diese Daten stets selbst generieren und waren es aus früheren Anfragen der Prüfer bereits gewohnt, diese Daten für andere Zwecke sehr detailliert aufzubereiten.

Die Transformation der Daten hat sich hingegen als äusserst wichtig erwiesen. Es war notwendig, bei der Verwendung der einzelnen Ansätze Aggregate zu bilden, wenn beispielsweise tägliche Umsatzdaten vorlagen und das Modell kumulierte Monatswerte verlangte oder wenn Verhältnisse berechnet werden mussten, wie beispielsweise ein Debitorenumschlag, der eine Verrechnung von Debitorenbestand und Umsatz vorsieht.

Visualisierung der Daten: Als Teil der Datenvorbereitung war es während der Fallstudie wichtig, die Daten – beispielsweise die Entwicklung des Personalbestands für einen Zeitraum – darzustellen und damit Rückschlüsse auf die Vollständigkeit und Korrektheit zu erhalten. So konnten Lücken in den Zeitreihen identifiziert, aber auch abrupte Niveauverschiebungen erkannt werden, die auf Fehler hinwiesen.

Bereinigung der Daten: Die mit der Visualisierung oder in späteren Schritten identifizierten Datenqualitätsprobleme mussten behoben werden. Dazu wurden fehlende Daten und Datenreihen ausgeschlossen, Fehler mit Mittelwerten der Datenreihen ersetzt oder gar auf neue Datenquellen zurückgegriffen.

Integration verschiedener Datenquellen: Verschiedene Modelle haben von der Integration von unterschiedlichen Datenquellen profitiert. Dies erforderte, dass Datenquellen über gemeinsame Merkmale kombiniert werden konnten.

Berechnung verschiedener Modelle: Die Software sollte es zulassen, unterschiedlichste Ansätze des *maschinellen Lernens* wie Regressionsmodelle oder künstliche neuronale Netze zu verwenden.

Speicherung und Archivierung der Daten: Es bot sich während der Abschlussprüfung an, die Daten einerseits in ihrer noch rohen Form und andererseits als *Features* historisiert abzulegen.

Damit war es möglich, entweder schnell auf Datenbestände mandatsübergreifend zuzugreifen oder längere, historische Datenreihen vorzuhalten.

Speicherung und Laden der Parameter der Modelle: Das Speichern der trainierten Modelle ermöglicht deren effiziente Wiederverwendung.

Wäre eine Bereitstellung der Modelle in dieser Form nicht möglich, müssten

sie vor jeder Anwendung erneut trainiert werden, was bei komplexen Modellen zeitintensiv ist und zudem eine potenzielle Quelle für Fehler darstellt.

Dokumentation: Die Dokumentation aller vorangegangenen Schritte erlaubt es nachzuvollziehen, wie die Berechnungen und Prognosen generiert wurden. Zudem ermöglicht sie Prüfern, eine Qualitätskontrolle sicherzustellen und die einschlägigen Anforderungen aus den Prüfungsstandards zu erfüllen.

Lokale Verarbeitung: Obschon die *Cloudprovider* wie *Microsoft Azure Cloud*, *Google-Cloud-Plattform* oder *Amazons Webservices* unterschiedliche Tools und zusätzlich auch spezialisierte Ressourcen zur Verfügung stellen, die die Datenhaltung und das Training von Modellen ermöglichen und beschleunigen, wurde explizit von dieser Option abgesehen.

Für diesen Verzicht spielen verschiedene Erwägungen eine Rolle: Die Übertragung von Daten in die Cloud stellt ein Sicherheitsrisiko dar. Die Daten könnten beim Transfer oder in der Cloud selbst exponiert werden. Die Folgen wären nicht nur Reputationsrisiken, sondern der Datentransfer wäre auch nicht mit den Compliance-Vorschriften der Prüfgesellschaften vereinbar. Darüber hinaus wurde in der Zusammenarbeitsvereinbarung zwischen dem Studienleiter und der Prüfgesellschaft explizit ein Datentransfer ausgeschlossen.

Obwohl die anderen Autoren unterschiedliche Softwarepakete für die Modellierung ihrer Daten einsetzten, hat sich die Entscheidung für eine Kombination aus *Python 3.7* mit *Jupyter Notebooks*, *Scikit-Learn*, *Pandas* und *Matplotlib*, ergänzt um einzelne spezifische Pakete, die beispielsweise eine Assoziationsanalyse erlauben, bewährt.

- *Eingesetzte Hardware*

  Die eingesetzte Hardware sollte es ermöglichen, die Modelle mit dem gewählten *Software-Stack* effizient zu berechnen.

  Mithilfe eines handelsüblichen Laptops und eines Prozessors der aktuellen Generation war es möglich, die ausgesuchten Modelle zu berechnen, weil die verwendeten Datenmengen auch bei Anwendung der Kreuzvalidierungen nicht in den *Big-Data*-Bereich reichten.

5.2 Exogener Anwendungsfall

5.2.1 Übersicht

Für die Anwendung des maschinellen Lernens bei Verwendung von ausschliesslich unternehmensexterner *Labels* wurde ein spezifisches Konstrukt entwickelt und in beiden Abschlussprüfungen verwendet.

Das Konstrukt wird sowohl in der Phase der Prüfungsplanung eingesetzt, in der es die Risikobeurteilung und die Beurteilung von dolosen Handlungen unterstützt, als auch in der Mandatsabschlussphase, in der sich der Fokus auf die Einschätzung der Unternehmensfortführung und der Notwendigkeit der Qualifizierung des Prüfurteils verschiebt. Ferner ist es auch für die Prüfungsdurchführung geeignet, wo es im Sinne einer analytischen Prüfungshandlung aufgrund von Branchendaten die Prognose für einzelne Abschlussposten erlaubt.

Darstellen lässt sich das Konstrukt als eine Reihe von Unterprozessschritten, die in Abbildung 27 wiedergegeben werden. Während des gesamten Prozesses werden sämtliche Schritte in der Datenakquisition, -verarbeitung und -modellierung dokumentiert, um damit eine Qualitätskontrolle sowie einen Rückgriff auf diese Grundlage bei künftigen Prüfungen zu ermöglichen, aber gleichzeitig auch die einschlägigen Vorgaben an die Prüfungsdokumentation zu erfüllen.

Darüber hinaus hält die Dokumentation fest, wie die einzelnen Variablen, welche in den Modellen Verwendung finden, definiert sind. Ansonsten bestünde die Gefahr, dass das Modell inkonsistent angewandt und die vormals getestete Prognosegüte kompromittiert würden.

Das Konstrukt gliedert sich in zwei Teile: Ziel des ersten Teils ist Berechnung des Modells, das später als Grundlage für die Generierung der Prognose dient. Im zweiten Teil folgt die Anwendung des Modells, das zur schlussendlichen Prognose führt. Dabei kann die Modellierung mandats- und zeitpunktunabhängig durchgeführt werden, während die Generierung der Prognose im Rahmen der Abschlussprüfung erfolgt.

5.2.2 Unterprozessschritte der Modellbildung

1. *Definition des Prüfziels und des Datenbedarfs*

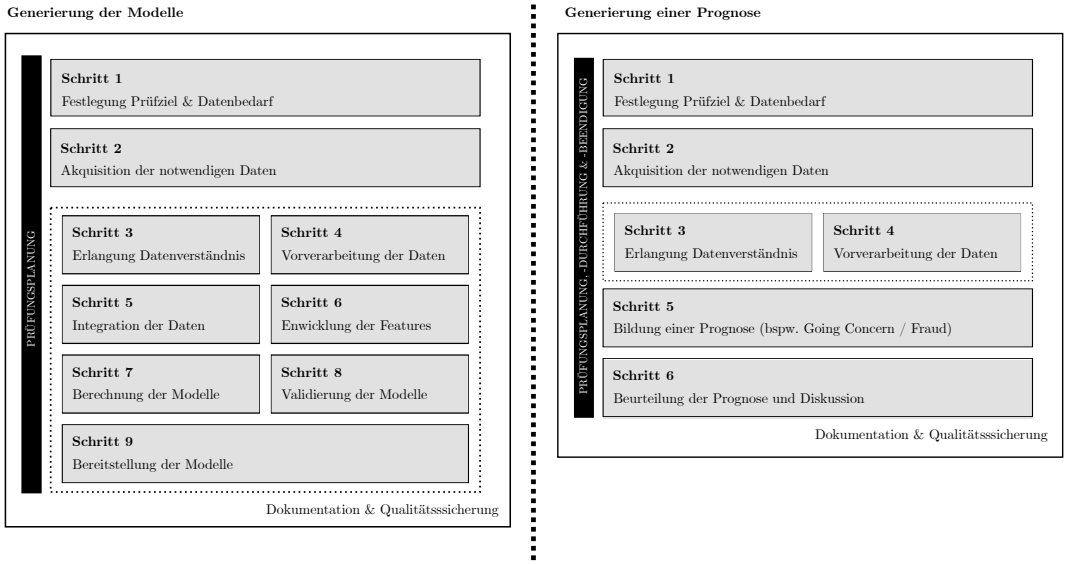


Abbildung 27: Unterprozessschritte im exogenen Anwendungsfall (eigene Darstellung)

Den Ausgangspunkt für das Konstrukt bildet die Definition des Prüfziels und des Datenbedarfs, wobei das Prüfziel den Datenbedarf determiniert.

Es bot sich in der Fallstudie an, Prüfziel und Datenbedarf für jedes Prüfgebiet in einem einzelnen *Jupyter-Notebook* zu erfassen, um damit das Vorgehen vollumfänglich zu dokumentieren und die Qualitätskontrolle zu erleichtern.

Mit der Dokumentation direkt in *Jupyter Notebooks* konnten zudem von Referenzierung auf andere Arbeitsblätter abgesehen werden, was die Kommunikation und den Review erleichtert hat.

2. *Akquisition der notwendigen Daten*

   Während die Datenakquisition im *CRISP-DM* eine untergeordnete Rolle spielt, kommt diesem Schritt in der Wirtschaftsprüfung grosse Bedeutung zu.

   In den Fallstudien wurden die benötigten Datensätze auf verschiedenen Ebenen, wie in Abbildung 28 ersichtlich, ausgehend vom einzelnen Wirtschaftsprüfer, gesucht.

   Es hat sich gezeigt, dass ein Wirtschaftsprüfer relativ selten mit qualifizierten

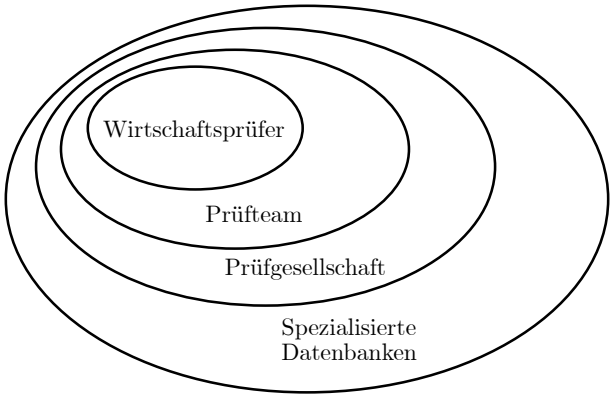


Abbildung 28: Sphären der Datensammlung (eigene Darstellung)

Prüfurteilen oder dolosen Handlungen in Kontakt kommt. Auch die Ausdehnung des Mandantenkreises mit dem Einbezug der Erfahrung des gesamten Prüfteams führte zu keiner befriedigenden Anzahl derartiger Abschlüsse. Vielmehr hat es sich erwiesen, dass auch im Prüfteam nur die wenigsten und erfahrensten Prüfer mit Abschlüssen in Kontakt gekommen sind, die diese Risikofaktoren enthielten.

Als *Ultimo Ratio* mussten die Daten daher aus externen Quellen wie der *Audit-Analytics-* oder aus der *Compustat-Capital-IQ-Datenbank* extrahiert werden.

Neben der reinen Beschaffung der Daten muss gleichfalls eine Einschätzung der Quellen vorgenommen und dokumentiert werden, um damit den Erfordernissen der Prüfungsstandards Rechnung zu tragen.

3. *Erlangung eines Datenverständnisses*

   Die beschafften Daten müssen einer kritischen Analyse unterzogen werden, um damit Datenqualitätsprobleme zu identifizieren und diese in einem späteren Schritt zu korrigieren.

   Diesem Zweck dient im ersten Schritt eine univariate Analyse der Datenreihen. Lage- und Streuungsmasse verschaffen einen schnellen Eindruck zur Plausibilität der Werte.

   Eine visuelle Aufbereitung der Datenreihen und deren Analyse in Form von

Histogrammen helfen, Lücken in den Datenreihen offenzulegen, die in den nächsten Schritten zu korrigieren sind.

4. *Vorverarbeitung der Daten*

   Im weiteren Schritt folgt die Beseitigung von Datenqualitätsproblemen. Dazu müssen in Abhängigkeit der verfügbaren Datenmenge Beobachtungen komplett ausgeschlossen werden, bei denen Werte fehlen, aber gleichzeitig ein grosser Datenbestand verfügbar ist. Steht jedoch nur eine geringe Anzahl von Beobachtungen zur Verfügung, werden die Datenlücken mithilfe von Mittelwerten über das entsprechende *Feature* aufgefüllt.

   Den Abschluss der Vorverarbeitung bildet sodann die Aggregation der Daten. Dazu mussten in den Fallstudien Wechselkursdaten, die als Tagesendkurse vorlagen, in einen für den Monat repräsentativen Wert überführt werden. In diesem Zusammenhang war entweder die Berechnung von Monatsmittelkursen oder die Auswahl des letzten Wertes zum Monatsende als Stellvertreter erforderlich.

5. *Integration der Daten*

   Mit Abschluss der Datenaufbereitung können die Daten miteinander verbunden werden, um aus einer Reihe von fragmentierten Datensätzen jeweils einzelne Beobachtungen zu generieren.

   Für diesen Zweck war es notwendig, eine Bezugsgrösse zu finden, die die einzelnen Quellen miteinander verband. In Bezug auf die Kreditratings und Unternehmen erfüllte die Unternehmensidentifikation diese Aufgabe. Bei den Temperaturdaten kamen dagegen das Datum der Beobachtung und der Monat des Abschlusses zum Einsatz.

6. *Entwicklung der Features*

   Für die Eingabe in die Modelle ist es zudem notwendig, *neue Features* aus den Beobachtungen abzuleiten und in einer analytischen Basistabelle zusammenzufassen. So wurde beispielsweise aus dem beobachteten Jahresabschlussdatum das Jahr extrahiert.

Der Umstand, dass gewisse Methoden des maschinellen Lernens nicht skaleninvariant sind, erfordert zudem eine Standardisierung bzw. Normalisierung der Daten. Gleichfalls müssen zuvor in seltenen Fällen eine Diskretisierung vorgenommen und kategoriale Variablen in *Dummyvariablen* überführt werden.

7. *Berechnung der Modelle*

   Liegen die Daten in einem Format vor, mit dem die verschiedenen Ansätze des maschinellen Lernens umgehen können, muss dem Umstand Rechnung getragen werden, dass viele der Datensätze unausgeglichen sind, sodass ein *Rebalancing* unumgänglich ist, indem mithilfe von Zufallsstichproben sichergestellt wird, dass für die verschiedenen Klassen – wie ein manipulierter bzw. nicht manipulierter Abschluss – jeweils gleich viele Beobachtungen vorliegen. Der Steigerung der Prognosegüte dient zudem ein *Mehrfach-Subsampling*.

   Der nächste Schritt erfordert einen Ausschluss von Beobachtungen, um zu einem späteren Zeitpunkt mit dem *Testset* die Güte der Prognose zu prüfen. Für die vorliegenden Fälle wurden jeweils etwa 20 % der Beobachtungen einer Zufallsstichprobe für Schritt 8 weggelegt.

   Es folgt das Training der Modelle auf dem reduzierten Datenbestand. Für einige Ansätze wie die Support-Vektor-Maschine oder das künstliche neuronale Netz ist eine Optimierung der *Hyperparameter* notwendig, weshalb in den Fallstudien jeweils mithilfe der *10-fold-Kreuzvalidierung* bei jedem Training ein Validierungsset ausgeschieden wurde.

8. *Validierung der Prognosegüte*

   Nach dem Training muss im Rahmen der Modellvalidierung bestimmt werden, inwieweit das Modell die Prognosen mit einer vorherbestimmten Qualität ausführt.

   Ermöglicht wurde diese Qualitätskontrolle, indem Prognosen mit dem im vorhergehenden Schritt beiseitegelegten *Testset* generiert und dieses Ergebnis mit den tatsächlichen Realisierungen abgeglichen wurde. Unter Zuhilfenahme von Leistungskennzahlen wie der MSE oder Konfusionsmatrizen für

die Klassenzugehörigkeiten – ob beispielsweise ein Abschluss potenziell manipuliert war oder nicht – liessen sich die Leistungsfähigkeiten der Modelle differenziert analysieren.

9. *Bereitstellung des Modells*

   Die bereitgestellten Modelle mit den Parametern, die eine Prognose für weitere Unternehmen zulassen, müssen abschliessend lokal persistiert werden. In den Fallstudien wurden sie auf einem zentralen Ordner abgelegt, was eine mandatsübergreifende Anwendung ermöglichte.

Im Gegensatz zum Vorgehen im *CRISP-DM Modell* entfällt bei der Generierung der Modelle die Erlangung eines Geschäftsverständnisses. Dies ist darauf zurückzuführen, dass die Aufstellung der Modelle weitestgehend *unternehmensagnostisch* erfolgte.

5.2.3 Unterprozesschritte der Prognosegenerierung

Im exogenen Anwendungsfall gestaltete sich die Prognose mithilfe der vormals generierten Modelle sehr leicht. Dazu ist Folgendes notwendig:

1. *Festlegung des Prüfziels und des Datenbedarfs*

   Die Festlegung des Prüfziels wie auch des Datenbedarfs gestaltet sich wenig umständlich: Während das Prüfziel sich aus der Risikobeurteilung ableitet, lässt sich der Datenbedarf aus der Spezifikation des einzusetzenden Modells entnehmen.

   Wird beispielsweise mit dem Modell das Ziel verfolgt, eine Einschätzung zur Unternehmensfortführung zu generieren, so spezifiziert die dem Modell zugehörige Dokumentation den unternehmensinternen Datenbedarf.

2. *Akquisition der notwendigen Daten*

   Die Daten, die für die Berechnung der Prognose notwendig sind, müssen sodann erhoben werden. Typischerweise handelt es sich hierbei um unternehmensinterne Daten, die als Abschlussposten in die Modelle einfliessen.

3. *Erlangung eines Datenverständnisses*

   Obwohl eine grosse Vertrautheit mit den Daten, insbesondere wenn sie als Abschlussposten ins Modell eingehen, vorliegt, müssen sie vor einer Prognose kritisch hinterfragt werden, falls dies nicht bereits anlässlich der Risikobeurteilung der Fall war.

4. *Vorverarbeitung der Daten*

   Bei Bedarf muss in einem Vorverarbeitungsschritt die Korrektur der Daten erfolgen. So war es anlässlich der Fallstudien nötig, Korrekturbuchungen vorzunehmen, um verlässliche Prognosen zu ermöglichen.

   Als Teil der Vorverarbeitung kann es zudem erforderlich sein, Verrechnungen vorzunehmen – falls beispielsweise ein Debitorenumschlag benötigt wird – und die Werte gegebenenfalls zu standardisieren.

5. *Bildung der Prognose*

   Der erarbeitete Eingabevektor, der die Merkmale der Unternehmung aufnimmt, dient anschliessend im Modell für eine Prognose.

6. *Beurteilung der Prognose und Diskussion*

   Sollten die Prognose ergeben, dass Risikofaktoren bestehen, so müssen diese abschliessend im Team oder mit dem Prüfkunden besprochen werden.

5.2.4 Herausforderungen

Bei der Anwendung des Konstrukts während der praktischen Prüfarbeit ergaben sich an verschiedenen Stellen Herausforderungen.

- *Spezifikation der verwendeten Features*

  Im exogenen Anwendungsfall wird ein Modell auf Daten – insbesondere als *Features* – von unterschiedlichen Unternehmen trainiert. Anschliessend folgt eine Prognose aufgrund eines Eingabevektors für die zu prüfende Unternehmung.

  Features können in diesem Zusammenhang beispielsweise der Jahresgewinn, Debitorenumschläge oder Lagerbestände sein.

Damit für das Training wie auch für die Generierung einer Prognose eine konsistente Berechnung dieser Aggregate möglich ist, müssen diese Features vorab genau spezifiziert werden – beispielsweise ob es sich beim Lagerwert um den Brutto- oder den Nettolagerwert handelt.

Bei einem Verzicht können Fehlprognosen generiert oder die Modellgüte überschätzt werden.

- *Bezug der Daten*

  Prüfer, Prüfteam sowie die einzelnen Prüfgesellschaften verfügen über ungenügend Daten, um damit aussagekräftige Modelle zu trainieren, da nur selten dolose Handlungen aufgedeckt oder Qualifikationen in den Prüfurteilen vorgenommen werden.

  Folglich muss für die Berechnung der Modelle auf externe Datenquellen zurückgegriffen werden. Eine dieser Datenquellen stellt die Unternehmung *Audit Analytics* zur Verfügung, die eine Datenbank u.a. für Prüfurteile, Prüfungshonorare und Prüfgesellschaften vorhält und diese gleichfalls mit Daten der *S&P Compustat (Capital-IQ-Datenbank* für Unternehmensdaten wie Bilanzsumme, Unternehmenserfolg etc.) kombiniert.

- *Unausgewogene Datensätze*

  Im exogenen Anwendungsfall werden fast ausschliesslich Klassen prognostiziert. So wird eine Vorhersage getroffen, ob für ein Unternehmen die Annahme der Unternehmensfortführung gilt (Ja/Nein) oder ob ein Prüfurteil eingeschränkt werden soll (Ja/Nein).

  Gewisse Fälle, wie ein eingeschränktes Prüfurteil, sind eher selten, sodass diese auch in den Datensätzen unterrepräsentiert sind. Ein Training auf diesen *unausgewogenen* Datensätzen würde die Leistungsfähigkeit erheblich verzerren.

  Es sind deshalb Techniken bei der Modellbildung anzuwenden, die diesem Umstand Rechnung tragen. Die Literatur verweist diesbezüglich auf *Unterstichproben*, bei denen aus der Klasse der überrepräsentierten Beobachtungen eine Stichprobe gezogen wird, die in Bezug auf die Grösse den unter-

repräsentierten Beobachtungen entspricht, sodass eine ausgeglichene Grundgesamtheit entsteht. Zudem können Mehrfach-Unterstichproben die Prognosegüte weiter erhöhen.

Gelegentlich wird empfohlen, eine Abstimmung der Beobachtungen aus den verschiedenen Klassen vorzunehmen und so zum Beispiel für jeweils eine Unternehmung aus der unterrepräsentierten Klasse ein vergleichbares Unternehmen – bezüglich der Branche und Unternehmensgrösse – aus der überrepräsentierten Klasse auszuwählen.

- *Kleine Datenmengen*

  Die im exogenen Anwendungsfall berechneten Modelle stützen sich üblicherweise auf relativ kleine Datenmengen.

  Für die Schätzung komplexerer Modelle, die keine grosse Spezifikation erfordern, sind *Hyperparameter* zu optimieren. Dieser Schritt bedarf neben der Ausscheidung eines Testsets auch eine Reserve in Form eines Validierungssets, sodass die Datenbasis zusätzlich eingeschränkt wird. Um diesem Umstand zu begegnen, können *Kreuzvalidierungsansätze* angewandt werden, die aus dem Trainingsset entweder aufgrund von Zufallsstichproben oder aber systematisch Daten für ein Validierungsset auswählen.

Ferner bestehen folgende potenzielle Herausforderungen:

- *Notwendigkeit zur Approximation von Abschlussposten*

  In der Phase der Prüfungsplanung nutzen Modelle zur Risikoeinschätzung bereits finanzielle Daten auf der Stufe der Abschlussposten.

  In Abhängigkeit von den Abschlussprozessen der Prüfkunden besteht die Möglichkeit, dass diese Abschlussposten zum Planungszeitpunkt noch nicht vollständig bzw. nicht korrekt vorliegen, weil Abschlussbuchungen noch nicht durchgeführt wurden.

  Folglich müssen in einem solchen Fall – möglicherweise unter Rückgriff auf Heuristiken und das Wissen um die Einheit und ihr Umfeld – Abschlussbestände approximiert werden.

5.3 Endogener Anwendungsfall

5.3.1 Übersicht

Für das maschinelle Lernen, das für Aufgaben eingesetzt wird, die unternehmensinterne Beobachtungen voraussetzen, wurde ein eigenes Konstrukt für die Wirtschaftsprüfung erarbeitet, das wiederum bei beiden Abschlussprüfungen zum Einsatz kam.

Der Anwendungsbereich des Konstrukts richtet sich hierbei auf die Phase der Prüfungsplanung und -durchführung: Es eignet sich sowohl zur Unterstützung bei analytischen Prüfungshandlungen als auch für die Beurteilung bzw. Identifizierung von dolosen Handlungen.

Das Konstrukt lässt sich als eine Folge aus Unterprozessen darstellen (siehe Abbildung 29). Wie bereits im Konstrukt zuvor wurden während des gesamten Prozesses das detaillierte Vorgehen und sämtliche Ergebnisse dokumentiert, um die Prüfziele zu erreichen, Auflagen zu erfüllen und für weitere Prüfungen wie auch die Qualitätskontrolle zur Verfügung zu stehen.

Während ausgewählte unternehmensexterne Daten lediglich für eine Fallstudie relevant waren – beispielsweise die örtlichen Bevölkerungsdaten für die zweite Fallstudie oder die Wechselkursdaten für die erste Fallstudie –, waren andere Daten für beide Fallstudien unverzichtbar. So halfen die Witterungsdaten in beiden Fällen bei der Verbesserung der Prognosegüte für die Umsätze.

Um die Daten in beiden Fällen einsetzen zu können, stand die Möglichkeit offen, diese strukturiert in einem Datawarehouse zu hinterlegen. Aufgrund der vielen Attribute, die in den Witterungsdaten enthalten sind, und des relativ aufwendigen Setups eines solchen Datenspeichers wurde pragmatisch ein *Datalake* mithilfe einer lokalen *MongoDB*-Instanz aufgesetzt.

5.3.2 Unterprozessschritte

1. *Erlangung des Geschäftsverständnisses*

   In der Planungsphase werden zur Einschätzung des Risikos von wesentlichen Fehlaussagen im Abschluss vom Prüfteam verschiedene Prüfungshandlungen durchgeführt. Aufbauend auf der Risikoanalyse wird ein Prüfplan entworfen,

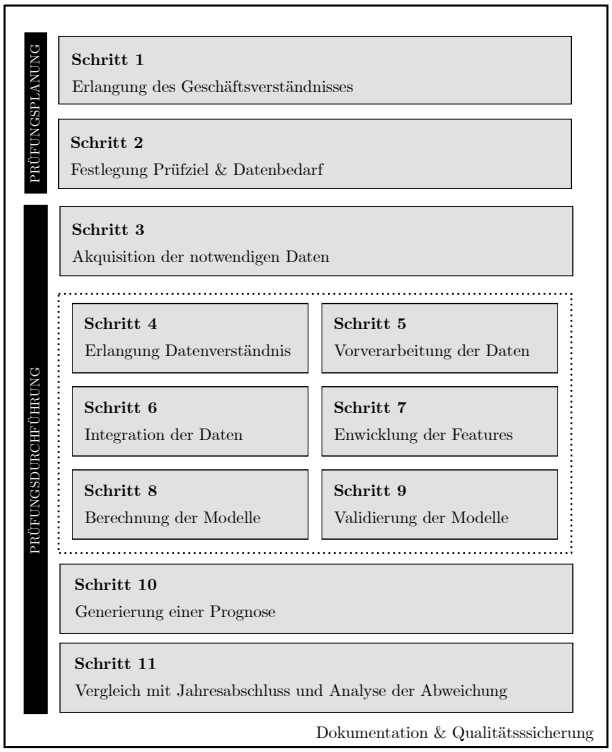


Abbildung 29: Training aufgrund externer Daten mit Prognose eines Vektors (eigene Darstellung)

der die Prüfung von Aussagen im Abschluss für einzelne Konten und Transaktionen vorsieht wie auch deren Art und zeitlichen Umfang.

Sind für einzelne Prüfgebiete analytische Prüfungshandlungen vorgesehen, die maschinelles Lernen einschliessen, so ist in diesen Bereichen der Aufbau eines detaillierten Geschäftsverständnisses erforderlich, da im endogenen Anwendungsfall insbesondere die Integration verschiedener Datenquellen – interne wie auch externe – verlässliche Prognosen erlaubt. Zur Identifikation dieser Datenquellen ist es notwendig, bereits anfangs ein gutes Verständnis über die Unternehmung und ihr Umfeld zu erarbeiten.

So hat es sich in den Fallstudien bewährt, mit internen und externen Sachverständigen das Gespräch zu suchen und in Interviews erklärende Variablen aufzudecken – diese wurden um Recherchen ergänzt.

2. *Definition des Prüfziels und Datenbedarf*

Danach ist es erforderlich, die Prüfziele auf der Ebene Aussagen im Abschluss in den einzelnen Prüfgebieten zu definieren, die sich mit den Prüfungshandlungen erreichen lassen.

Zudem gilt es, den Datenbedarf, gestützt auf das Geschäftsverständnis, abzustecken. Die frühe Evaluation des Datenbedarfs ermöglichte es, Datenquellen – beispielsweise Witterungsdaten – vorab zu evaluieren und nötige Transformationen zu planen.

3. *Akquisition der Daten*

   Das erste Konstrukt war mit dem Problem der *Unausgeglichenheit* behaftet, weil die Anzahl der Beobachtungen relativ gering ausfiel.

   Im endogenen Anwendungsfall ist das Problem weit weniger ausgeprägt. Vielmehr lassen sich aus den ERP- bzw. Buchhaltungssystemen die Daten der Prüfmandanten sehr direkt, schnell und in grösseren Mengen in einem maschinenlesbaren Format (*.txt- bzw. .csv-Format*) extrahieren.

   Die Akquisition der erweiterten Daten aus dem Unternehmensumfeld, die als potenziell hilfreich identifiziert werden, ist gleichwohl für Marktdaten wie Wechselkurse, Börsenkurse und Ratings relativ einfach: Derartige Daten finden sich an vielen Stellen im Internet in einem maschinenlesbaren Format. Gleiches gilt auch für makroökonomische Zeitreihen wie BIP, Arbeitslosenquoten oder Zinssätze, die in der Regel direkt von den Zentralbanken zur Verfügung gestellt werden.

   Schwieriger hingegen hat sich die Akquisition sehr spezifischer Daten, wie Temperatur- oder Niederschlagsdaten, erwiesen. Derartige Daten werden teilweise kommerziell angeboten oder können wie im Fall der Witterungsdaten direkt von Organisationen wie der *National Ocean and Atmospheric Administration (NOAA)* bezogen werden.

4. *Erlangung eines Datenverständnisses*

   Aufgrund der grossen Datenmengen und der unternehmensinternen, potenziell wenig dokumentierten Daten kommt der Erlangung des Datenverständnisses eine grosse Bedeutung zu.

Anders als in Fallstudie 1 wiesen die ERP-Daten in Fallstudie 2 eine sehr hohe Heterogenität auf und waren ebenfalls weniger gut dokumentiert. Entsprechend mussten für die Erlangung eines Datenverständnisses folgende zusätzlichen Informationen eingeholt werden:

- Inwieweit sind die Daten vollständig bzw. weisen sie Lücken auf, beispielsweise aufgrund von fehlerhaften Systemen oder Migrationen?
- Welche Bedeutung haben die einzelnen Felder und wie sind diese Felder kodiert?
- Welche weiteren Felder können aus dem ERP- bzw. Buchhaltungssystem entnommen werden?
- Gibt es weitere assoziierte Daten in anderen Systemen oder Subsystemen, beispielsweise in der Personalzeiterfassung?

Der Beantwortung dieser Fragen dienten folgende Massnahmen: Zunächst wurden das Management, die IT-Abteilung und die Buchhaltung befragt. Anschliessend folgte wie im ersten Anwendungsfall eine eingehende Analyse der Daten:

- Durch *univariate Analysen* der einzelnen Features und abhängigen Variablen wurde festgestellt, dass die Temperaturdaten für einzelne Datenreihen eine Spannweite von -1'000 bis 255 Grad Celsius aufweisen. Dieses Ergebnis sprengte die physikalischen Grenzen und deutete auf Datenprobleme hin.
- Ausserdem wurden die Datenreihen *visualisiert*, um deren zeitliche Vollständigkeit oder andere nicht plausible Daten zu ermitteln.

Zudem wurden wiederum die Datenquellen analysiert und ihre Eignung für analytische Prüfungshandlungen dokumentiert.

5. *Vorverarbeitung der Daten*

   Analog zum exogenen Anwendungsfall ist es nötig, die Daten vorab aufzubereiten, um sie in eine Form zu bringen, die eine Verwertung in den Modellen erlaubt.

Folgende Schritte sind hierfür notwendig:

- Analyse und Behebung der Datenqualitätsprobleme
- Bildung von Aggregaten (z.B. Monatsmittelwerte)

6. *Integration der Daten*

   Die gewonnenen Daten stammten in grösserem Umfang als im ersten Anwendungsfall aus unterschiedlichen Quellensystemen.

   Sämtliche Daten müssen daher über je ein gemeinsames Merkmal verbunden und zu einer einzigen Datenbasis zusammengeführt werden.

   Konkret mussten in den Fallstudien beispielsweise monatliche Witterungsdaten und Umsatzdaten als Monatsaggregate miteinander über eine *Monat-Jahr*-Kombination verbunden werden.

7. *Entwicklung der Features*

   Genauso wie im exogenen Anwendungsfall gilt, dass für die Berechnung der Modelle verschiedenste *Features* aufbereitet werden müssen.

   So ist es nötig:

   - abgeleitete Daten zu ermitteln – wie Tag des Jahres aus dem Datum
   - diskrete Klassen und *Dummy-Variablen* zu bilden
   - sowie die Daten zu standardisieren bzw. zu normalisieren

8. *Berechnung der Modelle*

   Im Unterschied zum ersten Konstrukt ist es nicht notwendig, die Daten vorab in Bezug auf die Klassen auszugleichen, zumal die Datensätze relativ umfangreich ausfallen.

   Die nun in geeigneter Weise zur Verfügung stehenden Daten lassen sich wiederum einem *Trainingsset* und einem *Testset* zuordnen. Hierfür wurden in der Fallstudie mithilfe einer Zufallsstichprobe 20 % der Beobachtungen als *Testset* für eine spätere Analyse der Prognosegüte zurückgelegt.

   In den Fallstudien wurden numerische Werte wie monatliche Umsätze prognostiziert, sodass die Auswahl der anwendbaren Modelle eingeschränkt war.

Für die Modelle wurden die in Tabelle 32 dargestellten Methoden auf die Daten angewandt.

Methode	Trainingsalgorithmus
Regressionsmodell	Kleinste-Quadrate
Ensemble Methoden	Random Forest
Support-Vektor-Maschine	Kernel SVM
künst. neuronale Netze	Stochastischer Gradientenabstieg

Tabelle 32: Angewandte Methoden bei analytischen Prüfungshandlungen

Es folgte die Schätzung eines Regressionsmodells, um eine Vergleichsgrundlage für die Leistungsfähigkeit der Modelle zu erhalten. Diesem Modell, das eine einfache Interpretation zuliess, wurden drei *Blackbox-Modelle* gegenüberstellt. Bei allen diesen Modellen war eine Spezifizierung *a priori* nicht erforderlich, einige *Hyperparametern* mussten jedoch während des Trainings optimiert werden. Die dafür benötigten Daten wurden für alle Modelle aus dem Trainingsset mithilfe eines *10-Fold-Kreuzvalidierungsansatzes* entnommen.

Der *Random Forest* war in seiner Anwendung sehr einfach und konnte gleichfalls das Modell innerhalb relativ kurzer Zeit berechnen. Ähnlich einfach gestaltete sich die Berechnung der Support-Vektoren. Im Gegensatz zum *Random Forest*-Ansatz hat sich diese Methode jedoch als sehr langsam herausgestellt, wenn die Daten in einer grösseren Menge an Beobachtungen vorlagen. Schliesslich gestaltete sich die Modellierung von neuronalen Netzen als komplex. Verschiedene Aspekte mussten vorab konfiguriert werden, ohne dass dafür Heuristiken zur Verfügung standen – beispielsweise für die Wahl der Netzwerktopologie.

Für den Bereich der Aufdeckung von dolosen Handlungen kamen Methoden des unüberwachten Lernens zum Einsatz (siehe Tabelle 33).

Methode	Trainingsalgorithmus
Clustering	k-Means
Assoziationsanalyse	AIS-Algorithmus

Tabelle 33: Angewandte Methoden bei unüberwachten Lernansätzen

Der gewählte Algorithmus (k-Means) für das Clustering machte die Identifizierung der optimale Clusterzahl aufgrund von Visualisierungen mithilfe des Ellenbogenansatzes als pragmatische Vorgehensweise notwendig. Die Berechnung sowohl der Cluster als auch der Regeln in der Assoziationsanalyse gestaltete sich im Vergleich zu anderen Ansätzen des überwachten Lernens als relativ schnell.

9. *Validierung der Modelle*

   Die Berechnung der Prognosegüte erfolgt wiederum auf Basis der Daten aus dem *Testset*. Dazu werden die Abweichungen der Prognosen von den tatsächlich realisierten Werten unter Zuhilfenahme des *Mean Square Error (MSE)* berechnet und gegebenenfalls mit dem *Benchmark-Modell* und der *Toleranzwesentlichkeit* verglichen.

   Dieses Vorgehen ist bei den Ansätzen zur Aufdeckung von dolosen Handlungen nicht möglich. Dort müssen anhand von Unterstichproben die Cluster auf Anomalien untersucht werden.

10. *Generierung der Prognose*

    Für die analytischen Prüfungshandlungen gilt, dass nun ein Eingabevektor erarbeitet werden kann, auf dessen Grundlage das Modell eine Prognose für die Abschlussbestände generiert.

11. *Diskussion der Ergebnisse*

    In einem finalen Schritt lassen sich die Prognosen den tatsächlichen Werten in der Buchhaltung gegenüberstellen und mögliche wesentliche Abweichungen im Team und mit dem Kunden evaluieren.

Auf die Bereitstellung, wie sie im *CRISP-DM*-Ablauf vorgesehen ist, konnte im endogenen Anwendungsfall gänzlich verzichtet werden. Denn anders als beim ersten Konstrukt veränderte sich die Datenbasis sehr schnell, weil täglich neue Transaktionen hinzukamen. Dies bedeutete wiederum, dass die Modelle in regelmässigen Abständen neu berechnet werden können und müssen.

5.3.3 Herausforderungen

Wie bereits der exogene war auch der endogene Anwendungsfall mit spezifischen Herausforderungen verbunden.

- *Identifikation von erklärenden Variablen*

  Im endogenen Anwendungsfall leisten unternehmensexterne Daten eminent wichtige Beiträge, indem ihr Einbezug als erklärende Variablen die Prognosegüte erhöht. Dabei ist es für aktuelle Ansätze des maschinellen Lernens wie den *Random Forest*, die *Support-Vektor-Maschinen* oder die *künstlichen neuronalen Netzen* nicht notwendig zu spezifizieren, wie diese Variablen berücksichtigt werden müssen – die Algorithmen erkennen dies selbstständig.

  Die Herausforderung besteht in der Identifikation von Datenreihen, die fähig sind, einen Beitrag zur Prognose zu leisten. In den Fallstudien hat es sich bewährt, potenzielle Variablen mit einer grösseren Anzahl von Fachspezialisten der Unternehmung selbst und mit Spezialisten ausserhalb der Organisation im Rahmen der Prüfungsplanung zu erörtern. Gleichfalls können im Rahmen einer Recherche derartige Variablen ermittelt werden.

  Neben dem Beitrag zur Prognosegüte liegt der Wert des Einbezugs von externen Variablen auch darin begründet, dass sie sich als besonders verlässlich qualifizieren, werden sie doch nicht vom geprüften Unternehmen bereitgestellt und sind damit unabhängig.

- *Beschaffung unternehmensexterner Daten*

  Sind mögliche unternehmensexterne Variablen einmal identifiziert, müssen sie in einem weiteren Schritt beschafft werden.

  In den Fallstudien wurde deutlich, dass verschiedene Datenquellen für Marktdaten vorlagen, die mit einfachen Mitteln abrufbar waren. So war es leicht, Wechselkursdaten oder Aktienkurse zu beziehen, die in einem sehr einfach zu integrierenden Format auf verschiedenen Plattformen beispielsweise als *.csv-Dateien* tagesaktuell zur Verfügung stehen.

  Die Erkenntnis gilt auch für makroökonomische Variablen wie Konjunkturdaten, Arbeitslosenquote oder Aussenhandelsbilanz: Diese Daten wurden von

der Schweizer Nationalbank bezogen, während Statistiken zur Bevölkerungsentwicklung bis auf die Ebene der Gemeinden den Datenbeständen des Statistischen Bundesamtes entnommen wurden.

Die grösste Herausforderung bestand in der Erschliessung spezialisierter Datensätze wie den stündlichen Witterungsdaten der Wetterstationen. Ein hohes Mass an Onlinerecherchearbeit war erforderlich, um diese Daten abzurufen und anschliessend – aufgrund ihrer Eigenheiten und ihres Umfangs – in ein operationalisierbares Format zu überführen.

- *Strukturierte Datenablage*

  Über die Mandate hinweg hat sich gezeigt, dass in der Datenbeschaffung ein Synergiepotenzial besteht. So können spezialisierte Datenquellen wie die Witterungsdaten in unterschiedlicher Weise anwendbar sein.

  Die Daten müssen jedoch strukturiert abgelegt werden, um diese Synergiepotenziale zu ermöglichen. Daher war es erforderlich, neben den Datensätzen auch *Metadaten* mitzugeben, die es zu einem späteren Zeitpunkt ermöglichen, die Felder schnell und einfach zu interpretieren.

  Als pragmatischer Datenspeicher hat sich ein *Datalake* auf mit *MongoDB* erwiesen, der die strukturierte Speicherung von Meta- und Rohdaten ermöglichte.

5.4 Generalisierbarkeit

Nach Kasanen et al. (1993) ist die Generalisierbarkeit der entwickelten Konstrukte stets kritisch zu hinterfragen. Im vorliegenden Fall kann eine Evaluation, inwieweit die oben beschriebenen Prozesse und Herausforderungen verallgemeinerbar sind, hinsichtlich der Prüfunternehmen und der Prüfkunden durchgeführt werden.

Generalisierbarkeit im Hinblick auf die Prüfgesellschaften Die als Destillat der Fallstudie hervorgegangenen Prozesse wurden bei einem Prüfunternehmen mittlerer Grösse entwickelt.

Nichtsdestotrotz steht einer Anwendung bei noch grösseren Prüfgesellschaften nichts im Weg. Im Gegenteil, grössere Prüfgesellschaften bewirtschaften die wichtigsten börsenkotierten Unternehmen der Schweiz, was mit einer intensiveren Prüfung der Kunden und deren Prozessen, Konten und Transaktionen einhergeht. Damit lassen sich auch die dargestellten Unterprozessschritte noch ausgedehnter durchführen.

- Es besteht die Möglichkeit einer detaillierten Auseinandersetzung anhand von Kundengesprächen, von Recherchen oder dem Rückgriff auf Spezialisten, welche externe erklärende Faktoren identifizieren, die sich ins Modell miteinbeziehen lassen.

- Es steht mehr Zeit für die Optimierung von Modellen und ihrer Hyperparameter zur Verfügung. Es können umfangreichere Datenmengen bewältigt werden, was die Anwendung komplexerer Modelle (wie beispielsweise von Schreyer et al., 2020 vorgeschlagen) ermöglicht.

Auf der anderen Seite des Spektrums dürfte der Nutzen für kleinere Prüfgesellschaften, die traditionell auch kleinere Prüfkunden bedienen, geringer ausfallen da zumindest der Prozess, der auf unternehmensinternen Beobachtungen beruht, für jede geprüfte Unternehmung spezifisch angewandt wird. Für die Prozesse, die sich auf unternehmensexterne Beobachtungen stützen, ist aufgrund der zentralisierten Modellierbarkeit von einer Anwendungsmöglichkeit auszugehen.

Generalisierbarkeit hinsichtlich Prüfkunden Neben den unterschiedlichen Prüfgesellschaften, die als Anwender die Prozesse durchlaufen, sind auch die Prüfkunden sehr unterschiedlich, auf die die Prozesse angewendet werden.

Für die Durchführung analytischer Prüfungshandlungen aufgrund von unternehmensinternen Beobachtungen oder die Aufdeckung doloser Handlungen sind umfangreichere Datenmengen erforderlich. Es existieren eine Reihe von Organisationstypen, die diese Vorbedingung nicht erfüllen. Damit dürfte das entsprechende Konstrukt nicht für Gesellschaften geeignet sein, die nur wenige Transaktionen zu verzeichnen haben, wie Holdinggesellschaften und Gesellschaften, die lediglich Immobilien oder Fonds verwalten.

Die Anwendbarkeit bzw. Nützlichkeit hinsichtlich des exogenen Anwendungsfalls wird durch die Prognosegüte der Modelle, bei entsprechend verwendeten Trainingsdaten, determiniert. In der Fallstudie wurden die Daten aus der Compustat-Datenbank speziell für Unternehmen der verarbeitenden Industrie bzw. der Telekommunikation entnommen. Es sind weitere Tests erforderlich, inwieweit die Anwendung eines solchen Datensatzes oder anderer aus der Datenbank entnommener Beobachtungen im Hinblick auf Fehler ersten und zweiten Typs in der Konfusionsmatrix sinnvoll ist.

5.5 Einfluss von SARS-CoV-2 auf Ergebnisse der Fallstudien

Die Forschung zu diesen Fallstudien fiel teilweise in die Zeit der SARS-CoV-2 Pandemie, die in der Schweiz am 16. März 2020 zu weitreichenden Einschränkungen im öffentlichen Leben und einem *Lock-Down* führte (swissinfo.ch, 2021). In den folgenden Wochen wurden die Massnahmen schrittweise wieder gelockert und die Unternehmen nahmen ihre Tätigkeiten wieder vor Ort in ihren Büros auf.

Dies traf auch auf das Prüfteam für Fallstudie 1 zu, das sich nach einer Zwischenprüfung direkt beim Unternehmen für die Schlussrevision in den Büroräumlichkeiten der Prüfgesellschaft zusammenfand. Der persönliche Austausch mit Ansprechpersonen der geprüften Organisation fand fortan *ad-hoc* über Onlinegespräche bzw. in regelmässig wiederkehrenden Onlinekonferenzen statt. Für den Austausch von Daten hatte sich bereits vor der Pandemie mit der geprüften Or-

ganisation eine Onlineplattform etabliert, was sich als sehr effektiv herausstellte.

Der Studienleiter kommunizierte einerseits über Onlinekonferenzen wie auch im direkten Kontakt mit Mitarbeitern und stützte sich für den Datenaustausch auf das interne Dateifreigabesystem. Insofern glich sein Vorgehen dem der Fallstudie 2, ohne den persönlichen Kontakt.

Darüber hinaus wurde im Vergleich zur Zwischenrevision in Fallstudie 1 bzw. der Abschlussprüfung in Fallstudie 2 keine wesentlich verminderte Effizienz oder Effektivität in der ortsungebundenen Prüfarbeit festgestellt: Auskünfte wie auch Daten wurden zeitgerecht und vollständig bereitgestellt, sodass von keinem wesentlichen Einfluss der Pandemie auf die Studie ausgegangen wird.

6 Konklusion

6.1 Gewonnene Erkenntnisse

Die vorliegende Forschungsarbeit befasst sich mit zwei Fragestellungen, die die Einführung von maschinellem Lernen in der praktischen Wirtschaftsprüfung untersuchen:

- *RQ1: Wie lässt sich in der Praxis der Wirtschaftsprüfungsprozess mit dem archetypischen Vorgehen des maschinellen Lernens verbinden?*
- *RQ2: Welchen operativen Herausforderungen begegnet der Prüfer in der praktischen Anwendung maschinellen Lernens im Prozess der Wirtschaftsprüfung und wie kann er diese überwinden?*

Beide Forschungsfragen wurden mithilfe von Fallstudien bei zwei Jahresabschlussprüfungen beantwortet, bei denen die Prüfteams begleitet wurden.

Methodologisch stütze sich diese Studie auf das von Kasanen et al. (1993) vorgeschlagene Vorgehen des *konstruktiven Forschungsansatzes*.

6.1.1 Integration des maschinellen Lernens in den Prüfprozess

Bei der Implementierung von Methoden des maschinellen Lernens, unabhängig der verwendeten Ansätze, hat sich in der praktischen Arbeit unmittelbar die Erkenntnis eingestellt, dass hinsichtlich der verwendeten Datenbasis grundsätzlich zwei verschiedene Anwendungsfälle zu unterscheiden sind.

1. *Unternehmensexterne Daten:* Die Daten, die zum Training der Modelle herangezogen werden, insbesondere die *Labels*, stammen nicht aus der Unternehmung selbst, sondern werden aus Drittquellen bezogen (exogener Anwendungsfall). Die Unternehmensdaten des Prüfkunden fliessen ausschliesslich als Eingabevektor für die Prognose in das Modell ein.
2. *Unternehmensinterne Daten:* Das Training der Modelle beruht im zweiten Anwendungsfall (endogener Anwendungsfall) auf unternehmensinternen Daten, insbesondere die *Labels*. Diese können durch externe Daten ergänzt

werden, beispielsweise durch Markt- oder Umweltdaten wie Wechselkurse, Konjunktur- oder Witterungsdaten.

Obwohl sich die beiden Anwendungsfälle lediglich in der Datenbasis unterscheiden, die für das Training der Modelle herangezogen werden, müssen daraus zwei Konstrukte gebildet werden, da die Datenbasis verschiedene Implikationen für die Integration mit sich bringt, zumal sich in beiden Fallstudien die Datenmengen, Datenprobleme (z.B. unausgeglichene Datensätze) und Anforderungen an den Prozess und die Hardware fundamental unterschieden.

Abstützung auf unternehmensexterne Beobachtungen Der exogene Anwendungsfall findet Anwendung in folgenden Prüfgebieten:

1. Risikobeurteilung mit Bezug auf den zu prüfenden Abschluss
2. Beurteilung des Risikos von dolosen Handlungen
3. Analytische Prüfungshandlungen
4. Einschätzung der Unternehmensfortführungsfähigkeit (engl. *going concern*)
5. Beurteilung des Qualifikationsbedarfs des Prüfurteils

Damit ist der Anwendungsfall auch für alle Phasen von der Prüfungsplanung über die Prüfungsdurchführung bis zum Prüfungsabschluss relevant.

Die unternehmensexternen Daten dienen zum Training der Modelle, die eine Aussage über Prüfgebiet machen und entsprechend eine Kategorie *„Ja, aufgrund der Datenlage ist die Annahme der Unternehmensfortführungfähigkeit eingeschränkt“* bzw. *„Nein, die Unternehmensfortführungsfähigkeit ist gegeben“* prognostizieren.

Die Fallstudien haben ergeben, dass für die Datenquelle weder die Erfahrung der Prüfer noch die des Prüfteams ausreichen, um genügend Daten bereitstellen zu können, da das Problem der *Datenrarität* vorliegt. So gibt es nur wenig Datensätze, die beispielsweise eine grosse Zahl von dolosen Handlungen enthalten. Ferner ist davon auszugehen, dass nicht einmal im internationalen Verbund der Prüfgesellschaft zu erwarten ist, dass genügend derartige Fälle vorliegen, um valide

Prognosen treffen zu können. Entsprechend müssen als Datenquellen internationaler Datenbanken herangezogen werden, die diese Merkmale vorhalten. Das Training der Modelle kann, wie aus den Fallstudien hervorgeht, komplett unabhängig von der eigentlichen Prüfung stattfinden.

Als Prozess zur Generierung der Modelle des maschinellen Lernens hat sich das *CRISP-DM-Modell* als paradigmatischer Prozess bewährt, der ausgehend von den Daten aus fünf einheitlichen Prozessschritten besteht:

1. Erlangung eines Daten- und Geschäftsverständnisses
2. Vorbereitung der Daten
3. Berechnung der Modelle
4. Validierung der Modelle
5. Bereitstellung der Modelle

Der Prozess wurde unabhängig von der Abschlussprüfung entwickelt, womit die Verfügbarkeit von Daten durch das *CRISP-DM–Modell* nahezu implizit vorausgesetzt wird, die Datenakquise entsprechend nicht berücksichtigt ist, während die Erlangung eines Geschäftsverständnisses verzichtbar ist, zumal die resultierenden Modelle *unternehmensagnostisch* angewendet werden können.

Der Prozess muss darum um zwei vorausgehende Schritte, (1) die Festlegung von Prüfzielen sowie Datenbedarf und (2) die Datenakquise, ergänzt werden, wobei sich von einer Erlangung des Geschäftsverständnisses absehen lässt.

Während der Abschlussprüfung selbst kann der Prozess wesentlich verkürzt werden: So bedarf es, wenn die vormals trainierten Modelle in der Praxis angewendet werden sollen, lediglich Abschlussposten der Unternehmung als Eingabevektoren für die Generierung einer Prognose. Dazu müssen die Daten zuvor gesichtet und analysiert sowie anschliessend mithilfe von Transformationen in eine operationalisierbare Form überführt werden. Die Berechnung, Validierung und Bereitstellung der Modelle entfallen: An ihre Stelle treten die Berechnung der Prognose und die Analyse des Ergebnisses im Team und mit den Kunden.

Abstützung auf unternehmensinterne Beobachtungen Für den endogenen Anwendungsfall des maschinellen Lernens aufgrund unternehmensinterner Daten wurde erkannt, dass dieser in zwei Prüfgebieten auftritt:

1. Beurteilung des Risikos von dolosen Handlungen
2. Analytische Prüfungshandlungen

Damit war der Anwendungsfall für die Prüfungsplanung als auch für die Prüfungsdurchführung relevant.

Die für das Training der Modelle herangezogenen Daten entstammten dem *ERP-System* und weiteren externen Quellen, die die von der Organisation produzierten Daten anreicherten. Die auf diese Weise in einem überwachten Lernverfahren trainierten Modelle wurden mit der Fähigkeit ausgestattet, numerische Prognose zu generieren, um beispielsweise Umsätze für die nächsten drei Monate vorherzusagen. Lernverfahren der unüberwachten Art ermöglichten hingegen die Zuordnung von Transaktionen in Cluster und damit eine Evaluation von Anomalien.

Das *CRISP-DM-Modell* musste erneut erweitert werden, um den Ansprüchen der Wirtschaftsprüfung zu genügen. So bedarf es eines vertieften Verständnisses für die Grössen, die mithilfe analytischer Prüfungshandlungen im späteren Verlauf der Prüfung prognostiziert werden sollen – wie beispielsweise erwartete Umsätze für eine Produktgruppe. Für diese Grössen müssen erklärende Variablen gefunden werden, die Recherchearbeiten oder der Diskussion mit unternehmensinternen oder -externen Sachverständigen entstammen. Damit verschiebt sich die Erlangung des Geschäftsverständnisses in die Phase der Prüfungsplanung und initiiert damit den Datenanalyseprozess. Anschliessend gilt es, die Erhebung und Transformation der identifizierten erklärenden Variablen zu planen.

Während im ersten Anwendungsfall der *Bereitstellung des Modells* eine wesentliche Bedeutung zukommt, ist dieser Schritt im zweiten Anwendungsfall obsolet. Die Wiederverwendung eines Modells ist aufgrund der Dynamik der Datenquelle – dem ERP-System – nicht angezeigt, die in kurzer Zeit viele Daten produziert. Vielmehr trägt das Training auf den jüngst produzierten Daten positiv zur Prognosegüte bei, vergrössert sich doch die Datenmenge über die Zeit, die für ein Training zur Verfügung steht.

6.1.2 Herausforderungen bei der Anwendung

Herausforderungen bei Abstützung auf unternehmensexterne Beobachtungen Folgende Herausforderungen zeigten sich anlässlich der Implementierung von maschinellem Lernen in der Prüfpraxis, wenn Ansätze verwendet wurden, die auf unternehmensexternen Beobachtungen beruhten:

1. *Bezug der Daten:* Dolose Handlungen werden selten aufgedeckt, während Prüfurteile nur in ausserordentlichen Fällen qualifiziert werden. Umfangreichere Datensätze zu solchen Risikofaktoren liegen darum häufig ausserhalb des Prüfteams vor, wenn nicht sogar ausserhalb der Prüfgesellschaft.

   Folglich bestand eine der wesentlichen Herausforderungen für die Modellierung derartiger Risikofaktoren in der Beschaffung der entsprechenden Daten, die in Anlehnung an Empfehlungen in der einschlägigen Literatur aus spezialisierten Datenbanken generiert wurden.

2. *Unausgewogenen Datensätze:* Die Seltenheit des Auftritts von Risikofaktoren führt zu *nicht balancierten* Datensätzen. Werden sie mit Methoden des maschinellen Lernens bearbeitet, so müssen Vorkehrungen getroffen werden, damit die Leistungsfähigkeit gesichert und die Prognosegüte erhalten bleibt. Einige Autoren berücksichtigen diesen Umstand, indem sie *Unterstichproben* der Daten durchführen sowie Beobachtungen mit Risikofaktoren und der Unterstichprobe abgleichen – beispielsweise über die Bilanzsumme und SIC-Ziffern der Unternehmenskategorien.

3. *Limitierte Datenbasis:* Aufgrund der insgesamt limitierten Datenbasis reduziert sich auch die Datenbasis, die anlässlich des Trainings der Modelle herangezogen wird.

   Bedeutsam wird die Herausforderung, wenn die Modelle eine Optimierung von Hyperparametern vorsehen. In diesen Fällen müssen Kreuzvalidierungsansätze – wie eine *10-fold-Kreuzvalidierung* – einbezogen werden, um diese Parameter dennoch schätzen zu können, ohne dabei die Prognosegüte stark zu beeinträchtigen.

Herausforderungen bei Abstützung auf unternehmensinterne Beobachtungen Für den endogenen Anwendungsfall des maschinellen Lernens ergaben sich im Rahmen der beiden Fallstudien folgende Hauptherausforderungen:

1. *Identifizierung von erklärenden Variablen:* Viele Modelle des überwachten Lernens führten insbesondere dann zu einer akzeptablen Prognosegüte, wenn sie neben den unternehmensinternen Daten zusätzlich mit Daten aus unternehmensexternen Quellen ergänzt wurden.

   Zu diesem Zweck mussten für die einzelnen Modelle, nachdem die Risikoanalyse hilfreiche Konten und Transaktionen ergeben hatte, Recherchen durchgeführt werden. Ausserdem hat es sich als sinnvoll herausgestellt, mit Sachverständigen aus finanzfernen Bereichen der Organisationen – wie dem Engineering oder dem Verkauf – das Gespräch zu suchen und mit den daraus gewonnenen Erkenntnissen Variablen zu identifizieren, die die zu modellierenden Grössen beeinflussten.

2. *Beschaffung unternehmensexterner Daten:* Für identifizierte erklärende Variablen mussten Datenquellen erschlossen werden.

   Es hat sich gezeigt, dass viele Daten, insbesondere Marktdaten (Kreditratings, Wechselkurs- oder Konjunkturdaten), relativ einfach in einem für die Software geeigneten Format verfügbar und verarbeitbar sind. Andere Daten, wie Witterungsdaten, finden sich hingegen erste nach einiger Recherche, und können dann gar als *Big Data* vorliegen, sodass eine ausgedehnte Vorverarbeitung vorgesehen und geleistet werden muss.

3. *Strukturierte Datenablage:* Nach der Identifikation und Akquirierung konnten die unternehmensexternen Daten über die Mandate hinweg für die Modellierung verschiedener Grössen verwendet werden.

   Dafür war es nötig, die Daten zusammen mit den Metadaten abzulegen, um die übergreifende Datenhaltung sicherzustellen. Im Rahmen der Fallstudie wurde dieser Schritt mithilfe eines *Datalake* umgesetzt.

6.2 Limitierungen

Hinsichtlich der Unternehmen, die als Prüfkunden in die Fallstudien eingingen, galt, dass sie über eine hochqualitative Buchführung und Rechnungslegung verfügten, sich keinen finanziellen Schwierigkeiten gegenübersahen, über sehr qualifiziertes Personal verfügten und weder Einschränkungen im Testat verzeichnen mussten noch dolose Handlungen bekannt wurden. In dieser Arbeit konnte der Fall einer Unternehmung, bei welcher die Finanzabschlüsse fragwürdig und der Prüfprozesse erschwert sind, entsprechend nicht betrachtet werden.

Von einer Diskussion von Implikationen für den Prozess wie auch von Herausforderungen, welche sich als Konsequenz der über mehrere Perioden erstreckenden Anwendung von maschinellem Lernen in der Abschlussprüfung ergeben, musste im Rahmen dieser Arbeit abgesehen werden, da die durchgeführten Fallstudien mit neun Monaten lediglich einen Auditzyklus – wenn auch bei zwei verschiedenen Mandanten – betrachteten.

Weiter konnten die Auswirkungen der Verlagerung der Datenanalyse in die *Cloud* nicht studiert werden. Als Teil des *Non-Disclosure Agreements* zur Forschungszusammenarbeit wurde vereinbart, dass die prüfungsrelevanten Kundendaten die Prüfgesellschaft nicht verlassen durften, womit von einer Untersuchung des Transfers von Daten, ihrer Verarbeitung und Modellierung sowie Bereitstellung auf entfernter Infrastruktur abgesehen werden musste.

Als implizite Prämisse dieser Arbeit wurde angenommen, dass das nötige Fachwissen für die Anwendung der Konstrukte vorhanden ist. So muss es dem Prüfer in der Praxis alleine oder in Arbeitsteilung möglich sein, Daten zu akquirieren, zu verarbeiten und Modelle zu entwickeln, die verlässliche Prognosen ermöglichen. Von einer Untersuchung, welche Implikationen die Absenz eines solchen Wissens im Prozess gehabt hätte, wurde abgesehen.

Generell wurde in dieser Arbeit der Fokus auf die Buchprüfungspraxis gelegt, ohne organisatorische und rechtliche-regulatorische Fragestellungen zu berücksichtigen. Damit liegt der Fokus primär auf dem Prozess und den technologischen Herausforderungen, welche mit maschinellem Lernen in der Abschlussprüfung einhergehen.

Des Weiteren wurde in dieser Forschungsarbeit keine Analyse vollständiger

Datenbestände durchgeführt, wie beispielsweise bei Schreyer et al. (2017), die sämtliche *Journalbuchungen* in das Training ihres Modells – beispielsweise zur Aufdeckung von dolosen Buchungen – miteinbeziehen. Dies hatte praktische Gründe: Einerseits hätte es sich als sehr viel aufwendiger erwiesen, sicherzustellen, dass der extrahierte Datenbestand tatsächlich den Daten aus dem ERP entspricht, und andererseits war absehbar, dass die Datenmengen die Verarbeitungskapazitäten der zur Verfügung gestellten Infrastruktur überschreiten würden.

Die Fallstudien wurden mit ausschliesslich in der Schweiz beheimateten Unternehmen durchgeführt. Ein anderer Rechtsrahmen als der schweizerische fand damit keine Beachtung.

6.3 Ausblick für weiterführende Forschung

Im Rahmen der Untersuchung der Fallstudien haben sich neue Fragestellungen ergeben, die Ansatzpunkte für die weiterführende Forschung bilden.

So wurden die Prüfteams über einen längeren Zeitraum von etwa neun Monaten begleitet, sodass in diese Zeit sowohl Zwischenprüfungen wie auch Jahresendprüfungen fielen. Damit konnte limitiert longitudinale Erfahrung mit den Konstrukten gesammelt werden. Es liessen sich in dieser Zeitspanne bereits Effizienzgewinne durch eine wiederholte Anwendung der Berechnungen und der Berechnungslogik erkennen. Von grossem Interesse wäre jedoch zu evaluieren, wie hoch die Effizienzgewinne in Folgeperioden ausfallen.

Für Prüfgesellschaften besteht jedoch das Problem, dass das initiale Aufsetzen derartiger mit maschinellem Lernen unterstützter Prüfprozesse mit hohen Investitionen verbunden ist. Folglich böte es sich sehr an, in künftigen Studien zu untersuchen, in welchem Masse Effizienzgewinne die Investitionen zu rechtfertigen wissen bzw. wie ein Kosten-Nutzen-Kalkül für die Prüfgesellschaften für die längerfristige Nutzung aussieht. Den beiden vorgestellten Konstrukten ist gemein, dass ihnen ein Synergiepotenzial über die Mandate hinweg innewohnt. Für Modelle, die sich auf das Training unternehmensexterner Beobachtungen stützen, kann die Verteilung der Modelle oder gar ihre zentralisierte Anwendung ein zusätzliches Synergiepotenzial entfalten. In der Anwendung von Methoden, die für das Training der Modelle unternehmensinterne Daten verwenden, kann die zentrale Bereitstellung von Da-

ten als *Features* Effizienzgewinne herbeiführen. Gegebenenfalls sind dies aber nicht die einzigen Ansatzpunkte für derartige Verbesserungen, und es stellt sich die Frage, an welchen Stellen sich zusätzliches Synergiepotenzial ausschöpfen liesse, um damit das Kosten-Nutzen-Kalkül weiter zu verbessern.

7 Literaturverzeichnis

Abbasi, A., Albrecht, C., Vance, A. & Hansen, J. (2012). Metafraud: a meta-learning framework for detecting financial fraud. *Mis Quarterly*, 1293–1327.

Agrawal, M. (1994). Rakesh, and Ramakrishnan Srikant." Fast algorithms for mining association rules." *Proc. 20th int. conf. very large data bases, VLDB*, *1215*.

Ahn, B., Cho, S. & Kim, C. (2000). The integrated methodology of rough set theory and artificial neural network for business failure prediction. *Expert systems with applications*, *18*(2), 65–74.

Altman, E. I. (2002). *Bankruptcy, credit risk, and high yield junk bonds*. Wiley-Blackwell.

Amani, F. A. & Fadlalla, A. M. (2017). Data mining applications in accounting: A review of the literature and organizing framework. *International Journal of Accounting Information Systems*, *24*, 32–58.

Anandarajan, M., Lee, P. & Anandarajan, A. (2001). Bankruptcy prediction of financially stressed firms: An examination of the predictive accuracy of artificial neural networks. *Intelligent Systems in Accounting, Finance & Management*, *10*(2), 69–81.

Arens, A. A., Elder, R. J., Beasley, M. S. & Hogan, C. E. (2016). Auditing and assurance services. *Auditing and Assurance Services*.

Argyrou, A. & Andreev, A. (2011). A semi-supervised tool for clustering accounting databases with applications to internal controls. *Expert Systems with Applications*, *38*(9), 11176–11181.

Ata, H. A. & Seyrek, I. H. (2009). The use of data mining techniques in detecting fraudulent financial statements: An application on manufacturing firms. *Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, *14*(2).

Bay, S., Kumaraswamy, K., Anderle, M. G., Kumar, R. & Steier, D. M. (2006). Large scale detection of irregularities in accounting data. *Data Mining, 2006. ICDM'06. Sixth International Conference on*, 75–86.

Beasley, M. S. (1996). An empirical analysis of the relation between the board of director composition and financial statement fraud. *Accounting review*, 443–465.

Beasley, M. S., Carcello, J. V., Hermanson, D. R., of Sponsoring Organizations of the Treadway Commission, C. et al. (1999). Fraudulent financial reporting: 1987-1997: an analysis of US public companies.

Beaver, W. H., McNichols, M. F. & Rhie, J.-W. (2005). Have financial statements become less informative? Evidence from the ability of financial ratios to predict bankruptcy. *Review of Accounting studies*, *10*(1), 93–122.

Bell, T. B. & Tabor, R. H. (1991). Empirical analysis of audit uncertainty qualifications. *Journal of Accounting Research*, *29*(2), 350–370.

Bella, M. B., Eloff, J. H. & Olivier, M. S. (2009). A fraud management system architecture for next-generation networks. *Forensic science international*, *185*(1-3), 51–58.

Bellman, R. (1961). Adaptive control processes: a guided tour princeton university press. *Princeton, New Jersey, USA*, 96.

Beneish, M. D. (1997). Detecting GAAP violation: Implications for assessing earnings management among firms with extreme financial performance. *Journal of accounting and public policy*, *16*(3), 271–309.

Benhayoun, N., Chairi, I., El Gonnouni, A. & Lyhyaoui, A. (2013). Financial intelligence in prediction of firm's creditworthiness risk: evidence from support vector machine approach. *Procedia Economics and Finance*, *5*, 103–112.

Bishop, C. M. et al. (1995). *Neural networks for pattern recognition*. Oxford university press.

Brachman, R. J., Khabaza, T., Kloesgen, W., Piatetsky-Shapiro, G. & Simoudis, E. (1996). Mining business databases. *Communications of the ACM*, *39*(11), 42–48.

Brazel, J. F., Jones, K. L. & Zimbelman, M. F. (2009). Using nonfinancial measures to assess fraud risk. *Journal of Accounting Research*, *47*(5), 1135–1166.

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984). Cart. *Classification and Regression Trees; Wadsworth and Brooks/Cole: Monterey, CA, USA*.

Calderon, T. G. (1999). Neural networks and preliminary information risk assessment in an auditing environment. *Accounting Enquiries*, *8*(2), 245.

Cecchini, M., Aytug, H., Koehler, G. J. & Pathak, P. (2010). Detecting management fraud in public companies. *Management Science*, *56*(7), 1146–1160.

Cerullo, M. J. & Cerullo, M. V. (2006). Using neural network software as a forensic accounting tool. *Information Systems Control Journal*, *2*, 33.

Chakraborty, S. & Sharma, S. K. (2007). Prediction of corporate financial health by artificial neural network. *International Journal of Electronic Finance*, *1*(4), 442–459.

Chan, P. K. & Stolfo, S. J. (1998). Learning with non-uniform class and cost distributions: Effects and a distributed multi-classifier approach. *In Workshop Notes KDD-98 Workshop on Distributed Data Mining*.

Charalambous, C., Charitou, A. & Kaourou, F. (2000). Comparative analysis of artificial neural network models: Application in bankruptcy prediction. *Annals of operations research*, *99*(1-4), 403–425.

Chen, H.-J., Huang, S. Y. & Lin, C.-S. (2009). Alternative diagnosis of corporate bankruptcy: A neuro fuzzy approach. *Expert Systems with Applications*, *36*(4), 7710–7720.

Chen, H.-J., Huang, S.-Y. & Kuo, C.-L. (2009). Using the artificial neural network to predict fraud litigation: Some empirical evidence from emerging markets. *Expert Systems with Applications*, *36*(2), 1478–1484.

Chen, K. C. & Church, B. K. (1992). Default on debt obligations and the issuance of going-concern opinions. *Auditing*, *11*(2), 30.

Chen, M.-Y. (2013). A hybrid ANFIS model for business failure prediction utilizing particle swarm optimization and subtractive clustering. *Information Sciences*, *220*, 180–195.

Chen, Y., Zhang, L. & Zhang, L. (2013). Financial distress prediction for Chinese listed manufacturing companies. *Procedia Computer Science*, *17*, 678–686.

Chye Koh, H. & Kee Low, C. (2004). Going concern prediction using data mining techniques. *Managerial Auditing Journal*, *19*(3), 462–476.

Coakley, J. R. (1995). Using pattern analysis methods to supplement attention directing analytical procedures. *Expert Systems with Applications*, *9*(4), 513–528.

Coakley, J. R. & Brown, C. E. (1993). Artificial neural networks applied to ratio analysis in the analytical review process. *Intelligent Systems in Accounting, Finance and Management*, *2*(1), 19–39.

Comunale, C. L., Rosner, R. L. & Sexton, T. R. (2010). The auditor's assessment of fraud risk: A fuzzy logic approach. *Journal of Forensic & Investigative Accounting*, *3*(1), 149–194.

Dbouk, B. & Zaarour, I. (2017). Towards a machine learning approach for earnings manipulation detection. *Asian Journal of Business and Accounting*, *10*(2), 215–251.

Debreceny, R. S. & Gray, G. L. (2010). Data mining journal entries for fraud detection: An exploratory study. *International Journal of Accounting Information Systems*, *11*(3), 157–181.

Dechow, P. M., Ge, W., Larson, C. R. & Sloan, R. G. (2011). Predicting material accounting misstatements. *Contemporary accounting research*, *28*(1), 17–82.

Deng, Q. (2009). Application of support vector machine in the detection of fraudulent financial statements. *Computer Science & Education, 2009. ICCSE'09. 4th International Conference on*, 1056–1059.

Dickens, L. & Watkins, K. (1999). Action research: rethinking Lewin. *Management learning*, *30*(2), 127–140.

Dielman, T. E. & Oppenheimer, H. R. (1984). An examination of investor behavior during periods of large dividend changes. *Journal of Financial and Quantitative Analysis*, *19*(2), 197–216.

Dikmen, B. & Küçükkocaoğlu, G. (2010). The detection of earnings manipulation: the three-phase cutting plane algorithm using mathematical programming. *Journal of Forecasting*, *29*(5), 442–466.

Eimers, P. (2016). *Exploring the Growing Use of Technology in the Audit, with a Focus on Data Analytics* (Techn. Ber.). International Audit & Assurance Standards Board (IAASB) Data Analytics Working Group. https://research.vu.nl/en/publications/e8777ba2-a42b-4683-8bec-2344b920225f

Ernst&Young. (2016). *Insights-Driven Digital Innovation*. http://www.ey.com/Publication/vwLUAssets/EY-insights-%20driven-digital-innovation/$FILE/EY-insights-driven-digital-innovation.pdf

Etheridge, H. L., Sriram, R. S. & Hsu, H. K. (2000). A comparison of selected artificial neural networks that help auditors evaluate client financial viability. *Decision Sciences*, *31*(2), 531–550.

Fanning, K. M. & Cogger, K. O. (1998). Neural network detection of management fraud using published financial data. *Intelligent Systems in Accounting, Finance & Management*, *7*(1), 21–41.

Fedorova, E., Gilenko, E. & Dovzhenko, S. (2013). Bankruptcy prediction for Russian companies: Application of combined classifiers. *Expert Systems with Applications*, *40*(18), 7285–7293.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, *7*(2), 179–188.

Friedman, J., Hastie, T. & Tibshirani, R. (2001). *The elements of statistical learning* (Bd. 1). Springer series in statistics Springer, Berlin.

Gaganis, C., Pasiouras, F., Spathis, C. & Zopounidis, C. (2007). A comparison of nearest neighbours, discriminant and logit models for auditing decisions. *Intelligent Systems in Accounting, Finance & Management: International Journal*, *15*(1-2), 23–40.

Gauss, C. F. (1809). *Theoria motus corporum coelestium in sectionibus conicis solem ambientium auctore Carolo Friderico Gauss*. sumtibus Frid. Perthes et IH Besser.

Gierbl, A. S. (2021). *Data analytics in external auditing* (Diss.). Universität St. Gallen.

Gierbl, A. S., Schreyer, M., Leibfried, P. & Borth, D. (2020). Künstliche Intelligenz in der Prüfungspraxis-Eine Bestandsaufnahme aktueller Einsatzmöglichkeiten und Herausforderungen. *Expert Focus*, *2020*(09), 612–617.

Goodfellow, I., Bengio, Y. & Courville, A. (2016). *Deep learning*. MIT Press.

Green, B. P. & Choi, J. H. (1997). Assessing the risk of management fraud through neural network technology. *Auditing*, *16*(1), 14.

Gupta, R. & Gill, N. S. (2012). Financial statement fraud detection using text mining. *Editorial Preface*, *3*(12), 189–191.

Hammersley, M. (2004). Action research: a contradiction in terms? *Oxford Review of Education*, *30*(2), 165–181.

Hastie, T. & Tibshirani, R. (1987). Generalized additive models: some applications. *Journal of the American Statistical Association*, *82*(398), 371–386.

Hastie, T., Tibshirani, R. & Friedman, J. (2009). The elements of statistical learnin. *Cited on*, 33.

Hillegeist, S. A., Keating, E. K., Cram, D. P. & Lundstedt, K. G. (2004). Assessing the probability of bankruptcy. *Review of accounting studies*, *9*(1), 5–34.

Hinton, G. E., Osindero, S. & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, *18*(7), 1527–1554.

Hoogs, B., Kiehl, T., Lacomb, C. & Senturk, D. (2007). A genetic algorithm approach to detecting temporal patterns indicative of financial statement fraud. *Intelligent Systems in Accounting, Finance & Management: International Journal*, *15*(1-2), 41–56.

Huang, S.-M., Tsai, C.-F., Yen, D. C. & Cheng, Y.-L. (2008). A hybrid financial analysis model for business failure prediction. *Expert Systems with Applications*, *35*(3), 1034–1040.

Huang, S.-Y., Tsaih, R.-H. & Lin, W.-Y. (2012). Unsupervised neural networks approach for understanding fraudulent financial reporting. *Industrial Management & Data Systems*, *112*(2), 224–244.

IAASB. (2009a). *ISA 200: Overall Objectives of the Independent Auditor and the Conduct of an Audit in Accordance with International Standards on Auditing*. New York: International Auditing, Assurance Standards Board.

IAASB. (2009b). *ISA 220, Quality Management for an Audit of Financial Statements*. New York: International Auditing, Assurance Standards Board.

IAASB. (2009c). *ISA 240: The Auditors Responsibilities. Relating to Fraud in an Audit of Financial Statements*. New York: International Auditing, Assurance Standards Board.

IAASB. (2009d). *ISA 300: Planning an Audit of Financial Statements*. New York: International Auditing, Assurance Standards Board.

IAASB. (2009e). *ISA 315: Identifying and assessing the risk of material misstatement*. New York: International Auditing, Assurance Standards Board.

IAASB. (2009f). *ISA 320: Audit Materiality*. New York: International Auditing, Assurance Standards Board.

IAASB. (2009g). *ISA 500: Audit Evidence*. New York: International Auditing, Assurance Standards Board.

IAASB. (2009h). *ISA 520: Analytical Procedures*. New York: International Auditing, Assurance Standards Board.

IAASB. (2009i). *ISA 570: Going Concern*. New York: International Auditing, Assurance Standards Board.

IAASB. (2009j). *ISA 700: Forming an opinion and reporting on fiancial statements*. New York: International Auditing, Assurance Standards Board.

Ireland, J. C. (2003). An empirical investigation of determinants of audit reports in the UK. *Journal of Business Finance & Accounting*, *30*(7-8), 975–1016.

Issa, H. & Kogan, A. (2014). A predictive ordered logistic regression model as a tool for quality review of control risk assessments. *Journal of Information Systems*, *28*(2), 209–229.

Issa, H., Sun, T. & Vasarhelyi, M. A. (2016). Research Ideas for Artificial Intelligence in Auditing: The Formalization of Audit and Workforce Supplementation. *Journal of Emerging Technologies in Accounting*, *13*(2), 1–20.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). *An introduction to statistical learning* (Bd. 112). Springer.

Jans, M., Alles, M. G. & Vasarhelyi, M. A. (2014). A field study on the use of process mining of event logs as an analytical procedure in auditing. *The Accounting Review*, *89*(5), 1751–1773.

Jans, M., Lybaert, N. & Vanhoof, K. (2009). A Framework for Internal Fraud Risk Reduction at IT Integrating Business Processes: The IFR 2 Framework. *International Journal of Digital Accounting Research*, *9*(1).

Jans, M., Lybaert, N. & Vanhoof, K. (2010). Internal fraud risk reduction: Results of a data mining case study. *International Journal of Accounting Information Systems*, *11*(1), 17–41.

Jans, M., Van Der Werf, J. M., Lybaert, N. & Vanhoof, K. (2011). A business process mining application for internal transaction fraud mitigation. *Expert Systems with Applications*, *38*(10), 13351–13359.

Jo, H., Han, I. & Lee, H. (1997). Bankruptcy prediction using case-based reasoning, neural networks, and discriminant analysis. *Expert Systems with Applications*, *13*(2), 97–108.

Jones, S. (2017). Corporate bankruptcy prediction: a high dimensional analysis. *Review of Accounting Studies*, *22*(3), 1366–1422.

Kaplan, R. S. (1998). Innovation action research: creating new management theory and practice. *Journal of management accounting research*, *10*, 89.

Kasanen, E., Lukka, K. & Siitonen, A. (1993). The constructive approach in management accounting research. *Journal of management accounting research*, *5*, 243.

Kasgari, A. A., Divsalar, M., Javid, M. R. & Ebrahimian, S. J. (2013). Prediction of bankruptcy Iranian corporations through artificial neural network and Probit-based analyses. *Neural Computing and Applications*, *23*(3-4), 927–936.

Kasurinen, T. (2002). Exploring management accounting change: the case of balanced scorecard implementation. *Management accounting research*, *13*(3), 323–343.

Keating, P. J. (1995). A framework for classifying and evaluating the theoretical contributions of case research in management accounting. *Journal of management accounting research*, *7*, 66.

Keenoy, C. (1958). The impact of automation on the field of accounting. *Accounting Review*, 230–236.

Kelleher, J. D. & Tierney, B. (2018). *Data science*. MIT Press.

Kingma, D. P. & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kirkos, E., Spathis, C. & Manolopoulos, Y. (2007). Data mining techniques for the detection of fraudulent financial statements. *Expert systems with applications*, *32*(4), 995–1003.

Kirkos, E., Spathis, C., Nanopoulos, A. & Manolopoulos, Y. (2007). Identifying qualified auditors' opinions: a data mining approach. *Journal of Emerging Technologies in Accounting*, *4*(1), 183–197.

Koh, H. C. & Tan, S. S. (1999). A neural network approach to the prediction of going concern status. *Accounting and Business Research*, *29*(3), 211–216.

Kokina, J. & Davenport, T. H. (2017). The emergence of artificial intelligence: How automation is changing auditing. *Journal of Emerging Technologies in Accounting*, *14*(1), 115–122.

Koopman, S. J., Kräussl, R., Lucas, A. & Monteiro, A. B. (2009). Credit cycles and macro fundamentals. *Journal of Empirical Finance*, *16*(1), 42–54.

Korol, T. (2013). Early warning models against bankruptcy risk for Central European and Latin American enterprises. *Economic Modelling*, *31*, 22–30.

Koskivaara, E. (2000a). Artificial neural network models for predicting patterns in auditing monthly balances. *Journal of the operational research society*, *51*(9), 1060–1069.

Koskivaara, E. (2000b). Different pre-processing models for financial accounts when using neural networks for auditing. *ECIS 2000 Proceedings*, 3.

Koskivaara, E. & Back, B. (2007). Artificial neural network assistant (ANNA) for continuous auditing and monitoring of financial data. *Journal of Emerging Technologies in Accounting*, *4*(1), 29–45.

Kotsiantis, S., Koumanakos, E., Tzelepis, D. & Tampakas, V. (2006). Forecasting fraudulent financial statements using data mining. *International Journal of Computational Intelligence*, *3*(2), 104–110.

Krambia-Kapardis, M., Christodoulou, C. & Agathocleous, M. (2010). Neural networks: the panacea in fraud detection? *Managerial Auditing Journal*, *25*(7), 659–678.

Labro, E. & Tuomela, T.-S. (2003). On bringing more action into management accounting research: process considerations based on two constructive case studies. *European accounting review*, *12*(3), 409–442.

Laitinen, E. K. & Laitinen, T. (1998). Qualified audit reports in Finland: evidence from large companies. *European Accounting Review*, *7*(4), 639–653.

Lantz, B. (2015). *Machine learning with R*. Packt Publishing Ltd.

Lee, K., Booth, D. & Alam, P. (2005). A comparison of supervised and unsupervised neural networks in predicting bankruptcy of Korean firms. *Expert Systems with Applications*, *29*(1), 1–16.

Legendre, A. M. (1806). *Nouvelles méthodes pour la détermination des orbites des comètes; par AM Legendre*. chez Firmin Didot, libraire pour lew mathematiques, la marine.

Lehtiranta, L., Junnonen, J.-M., Kärnä, S. & Pekuri, L. (2015). The constructive research approach: Problem solving for complex projects. *Designs, methods and practices for research of project management*, 95–106.

Lenard, M. J., Alam, P. & Madey, G. R. (1995). The application of neural networks and a qualitative response model to the auditor's going concern uncertainty decision. *Decision Sciences*, *26*(2), 209–227.

Leo, B., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984). Classification and regression trees. *Wadsworth International Group*.

Li, H., Sun, J., Li, J.-C. & Yan, X.-Y. (2013). Forecasting business failure using two-stage ensemble of multivariate discriminant analysis and logistic regression. *Expert systems*, *30*(5), 385–397.

Liang, J. & Lv, W. (2009). Research on detecting technique of financial statement fraud based on fuzzy genetic algorithms BPN. *Management Science and Engineering, 2009. ICMSE 2009. International Conference on*, 1462–1468.

Loebbecke, J. K., Eining, M. M. & Willingham, J. J. (1989). Auditors experience with material irregularities-frequency, nature, and detectability. *Auditing-A Journal of Practice & Theory*, *9*(1), 1–28.

Louwers, T. J., Ramsay, R. J., Sinason, D. H., Strawser, J. R. & Thibodeau, J. C. (2013). *Auditing and assurance services*. McGraw-Hill/Irwin New York, NY.

Lukka, K. (2000). The key issues of applying the constructive approach to field research. *Reponen, T.(ed.)*, 113–28.

Lukka, K. & Tuomela, T.-S. (1998). Testattuja ratkaisuja liikkeenjohdollisiin ongelmiin: konstruktiivinen tutkimusote. *Yritystalous*, *4*(98), 23–29.

Martens, D., Bruynseels, L., Baesens, B., Willekens, M. & Vanthienen, J. (2008). Predicting going concern opinion with data mining. *Decision Support Systems*, *45*(4), 765–777.

Mattessich, R. (1995). Conditional-normative accounting methodology: incorporating value judgments and means-end relations of an applied science. *Accounting, Organizations and Society*, *20*(4), 259–284.

McKee, T. E. (2007). Altman's 1968 bankruptcy prediction model revisited via genetic programming: new wine from an old bottle or a better fermentation process? *Journal of Emerging Technologies in Accounting*, *4*(1), 87–101.

McKeown, J. C., Mutchler, J. F. & Hopwood, W. (1991). Towards an explanation of auditor failure to modify the audit opinions of bankrupt companies. *Auditing-a Journal Of Practice & Theory*, *10*, 1–13.

McKinnon, J. (1988). Reliability and validity in field research: some strategies and tactics. *Accounting, Auditing & Accountability Journal*, *1*(1), 34–54.

Melendez, C. (2016). *Artificial intelligence gets into auditing, what's next?* Verfügbar 10. Oktober 2013 unter http://www.infoworld.com/article/3044468/application-development/artificial-intelligence-gets-into%5C--auditing-whats-next.html

Min, J. H. & Lee, Y.-C. (2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert systems with applications*, *28*(4), 603–614.

Nyce, C. & Cpcu, A. (2007). Predictive analytics white paper. *American Institute for CPCU. Insurance Institute of America*, 9–10.

Öğüt, H., Aktaş, R., Alp, A. & Doğanay, M. M. (2009). Prediction of financial information manipulation by using support vector machine and probabilistic neural network. *Expert Systems with Applications*, *36*(3), 5419–5423.

O'leary, D. E. (1998). Using neural networks to predict corporate failure. *Intelligent Systems in Accounting, Finance & Management*, *7*(3), 187–197.

Olson, D. L., Delen, D. & Meng, Y. (2012). Comparative analysis of data mining methods for bankruptcy prediction. *Decision Support Systems*, *52*(2), 464–473.

Owusu-Ansah, S., Moyes, G. D., Babangida Oyelere, P. & Hay, D. (2002). An empirical analysis of the likelihood of detecting fraud in New Zealand. *Managerial Auditing Journal*, *17*(4), 192–204.

Pai, P.-F., Hsu, M.-F. & Wang, M.-C. (2011). A support vector machine-based model for detecting top management fraud. *Knowledge-Based Systems*, *24*(2), 314–321.

PCAOB. (2010). Audit Evidence. PCAOB Auditing Standard No. 15.

Peel, M. (1989). The going-concern qualification debate: some UK evidence. *The British Accounting Review*, *21*(4), 329–350.

Perols, J. (2011). Financial statement fraud detection: An analysis of statistical and machine learning algorithms. *Auditing: A Journal of Practice & Theory*, *30*(2), 19–50.

Perols, J. L., Bowen, R. M., Zimmermann, C. & Samba, B. (2017). Finding needles in a haystack: Using data analytics to improve fraud prediction. *The Accounting Review*, *92*(2), 221–245.

PricewaterhouseCoopers. (2016). *Blockchain and Smart Contract Automation: How Smart Contracts Automate Digital Business*. https://www.pwc.com/us/en/technology-forecast/2016/blockchain/pwc-smart-contract-automation-digital-business.pdf

Puolamäki, E. (2000). Beyond the normative empiricism: how to use emerging construc- tive approaches in management accounting research to scientifically exploit the pragmatic quest for control? *Second Conference on New Directions in Management Accounting: Innovations in Practice and Research*, (5), 71–90.

Rafiei, F. M., Manzari, S. & Bostanian, S. (2011). Financial health prediction models using artificial neural networks, genetic algorithm and multivariate discriminant analysis: Iranian evidence. *Expert Systems with Applications*, *38*(8), 10210–10217.

Ramamoorti, S., Bailey Jr, A. D. & Traver, R. O. (1999). Risk assessment in internal auditing: a neural network approach. *Intelligent Systems in Accounting, Finance & Management*, *8*(3), 159–180.

Raphael, J. (2015). *How Artificial Intelligence Can Boost Audit Quality*. http://%20ww2.cfo.com/auditing/2015/06/artificial-%20intelligence-can-boost-audit-quality/

Ravisankar, P., Ravi, V., Rao, G. R. & Bose, I. (2011). Detection of financial statement fraud and feature selection using data mining techniques. *Decision Support Systems*, *50*(2), 491–500.

Reinartz, T. & Wirth, R. (1995). The need for a task model for knowledge discovery in databases. *Workshop notes Statistics, Machine Learning, and Knowledge Discovery in Databases. MLNet Familiarisation Workshop, Heraklion, Crete*, 19–24.

Revisionsaufsichtsbehörde, E. (2016). *Geschäftsbericht*.

Revisionsaufsichtsbehörde, E. (2017). *Geschäftsbericht*.

Reynolds, J. & Jere, R. (2000). Does Size Matter? The Influence of Large Client on Office-Level Auditor Reporting Decisions. *Journal of Accounting and Economics*, *30*, 375–400. https://doi.org/10.1016/S0165-4101(01)00010-6

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386.

Samuel, A. L. (1969). Some studies in machine learning using the game of checkers. II—Recent progress. *Annual Review in Automatic Programming*, *6*, 1–36.

Schäfer, A., Steiner, A. & Keller, B. (2017). Data Analytics in der Wirtschaftsprüfung. *expertFOCUS*, *11*, 798–802.

Schreyer, M., Sattarov, T., Borth, D., Dengel, A. & Reimer, B. (2017). Detection of anomalies in large scale accounting data using deep autoencoder networks. *arXiv preprint arXiv:1709.05254*.

Schreyer, M., Sattarov, T., Gierbl, A. S., Reimer, B. & Borth, D. (2020). Learning Sampling in Financial Statement Audits using Vector Quantised Variational Autoencoder Neural Networks. *Proceedings of the International Conference on Artificial Intelligence (ICAIF)'20*.

Serrano-Cinca, C. & Gutiérez-Nieto, B. (2013). Partial least square discriminant analysis for bankruptcy prediction. *Decision Support Systems*, *54*(3), 1245–1255.

Shirata, C. Y., Takeuchi, H., Ogino, S. & Watanabe, H. (2011). Extracting key phrases as predictors of corporate bankruptcy: Empirical analysis of annual reports by text mining. *Journal of emerging technologies in accounting*, *8*(1), 31–44.

Shleifer, A. & Vishny, R. W. (1997). A survey of corporate governance. *The journal of finance*, *52*(2), 737–783.

Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *The journal of business*, *74*(1), 101–124.

Spathis, C. T. (2002). Detecting false financial statements using published data: some evidence from Greece. *Managerial Auditing Journal*.

Spathis, C. T. (2003). Audit qualification, firm litigation, and financial information: an empirical analysis in Greece. *International Journal of Auditing*, *7*(1), 71–85.

Summers, S. L. & Sweeney, J. T. (1998). Fraudulently misstated financial statements and insider trading: An empirical analysis. *Accounting Review*, 131–146.

Sun, T. (2019). Applying deep learning to audit procedures: An illustrative framework. *Accounting Horizons*, *33*(3), 89–109.

Susskind, R. & Susskind, D. (2015). *The future of the professions: How technology will transform the work of human experts*. Oxford University Press, USA.

swissinfo.ch. (2021). *Lockdown: Vor einem Jahr hat der Bundesrat die Schweiz abgeriegelt*. https://www.swissinfo.ch/ger/lockdown--vor-einem-jahr-hat-der-bundesrat-die-schweiz-abgeriegelt/46451438 (accessed: 21.12.2021)

Tackett, J. A. (2013). Association rules for fraud detection. *Journal of Corporate Accounting & Finance*, *24*(4), 15–22.

Tang, T.-C. & Chi, L.-C. (2005). Neural networks analysis in business failure prediction of Chinese importers: A between-countries approach. *Expert Systems with Applications*, *29*(2), 244–255.

Thiprungsri, S. & Vasarhelyi, M. A. (2011). Cluster Analysis for Anomaly Detection in Accounting Data: An Audit Approach. *International Journal of Digital Accounting Research*, *11*.

Tinoco, M. H. & Wilson, N. (2013). Financial distress and bankruptcy prediction among listed companies using accounting, market and macroeconomic variables. *International Review of Financial Analysis*, *30*, 394–419.

Treuhand-Kammer. (2009). *Schweizer Handbuch der Wirtschaftsprüfung - Band 2 - Abschlussprüfung*.

EXPERTsuisse. (2013). *Schweizer Prüfungsstandards*.

EXPERTsuisse. (2016). *Schweizer Handbuch der Wirtschaftsprüfung - Band Ordentliche Revision*.

Tsai, C.-F. & Wu, J.-W. (2008). Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert systems with applications*, *34*(4), 2639–2649.

Tsaih, R.-H., Lin, W.-Y. & Huang, S.-Y. (2009). Exploring fraudulent financial reporting with GHSOM. *Pacific-Asia Workshop on Intelligence and Security Informatics*, 31–41.

Turetsky, H. F. & McEwen, R. A. (2001). An empirical investigation of firm longevity: A model of the ex ante predictors of financial distress. *Review of Quantitative Finance and Accounting*, *16*(4), 323–343.

Virdhagriswaran, S. & Dakin, G. (2006). Camouflaged fraud detection in domains with complex relationships. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 941–947.

Westbrook, R. (1995). Action research: a new paradigm for research in production and operations management. *International Journal of Operations & Production Management*.

Westermann, C. & Spindler, C. (2017). Anomalie-Erkennung in der Wirtschaftsprüfung mithilfe von Maschine Learning. *expertFOCUS*, *11*, 791–797.

Wirth, R. & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, 29–39.

Yang, Z., Platt, M. B. & Platt, H. D. (1999). Probabilistic neural networks in bankruptcy prediction. *Journal of Business Research*, *44*(2), 67–74.

Youn, H. & Gu, Z. (2010). Predicting Korean lodging firm failures: An artificial neural network model along with a logistic regression model. *International Journal of Hospitality Management*, *29*(1), 120–127.

Yu, Z., Guang, Y. & Zi-qi, J. (2013). Violations detection of listed companies based on decision tree and K-nearest neighbor. *Management Science and Engineering (ICMSE), 2013 International Conference on*, 1671–1676.

Yue, D., Wu, X., Shen, N. & Chu, C.-H. (2009). Logistic regression for detecting fraudulent financial statement of listed companies in China. *Artificial Intelligence and Computational Intelligence, 2009. AICI'09. International Conference on*, *2*, 104–108.

Zhang, G., Hu, M. Y., Patuwo, B. E. & Indro, D. C. (1999). Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis. *European journal of operational research*, *116*(1), 16–32.

Zhou, W. & Kapoor, G. (2011). Detecting evolutionary financial statement fraud. *Decision Support Systems*, *50*(3), 570–575.

Zmijewski, M. E. (1984). Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting research*, 59–82.

Zouboulidis, E. & Kotsiantis, S. (2012). Forecasting fraudulent financial statements with committee of cost-sensitive decision tree classifiers. *Hellenic Conference on Artificial Intelligence*, 57–64.

Curriculum Vitae – Sven Bösiger

Personalien

Geburtsdatum	24. Juni 1982
Geburtsort	Basel, Schweiz
Nationalität	Schweiz
E-Mail	sven.boesiger@unisg.ch

Ausbildung

2016 – 2023	**Universität St. Gallen (HSG), Schweiz** Doktoratsstudium in Betriebswirtschaftslehre KPMG-Lehrstuhl für Accounting & Auditing
2012 – 2015	**Universität St. Gallen (HSG), Schweiz** Masterstudium in Rechnungswesen und Finanzen Master of Arts HSG
2006 – 2008	**Universität Basel, Schweiz** Masterstudium in Economics and Business Master of Science
2002 – 2006	**Universität Basel, Schweiz** Bachelorstudium in Economics and Business Bachelor of Arts

Berufserfahrung

Seit 2018	**gateB, Schweiz** Senior Data Scientist Leiter Software Engineering
2016 – 2018	**Universität St. Gallen (HSG), Schweiz** Wissenschaftlicher Mitarbeiter
2012 – 2016	**KPMG Schweiz, Zürich** Mitarbeiter im Consulting, Tax und Audit