Masterlehrgang der FHWien der WKW
MSc Designing Digital Business (Berufsakademie)

Einsatz von Predictive Analytics zur Ermittlung von Cross-Selling-Potenzialen im Sanitär-Heizung-Klima-Großhandel

Angestrebter akademischer Grad:
Master of Sciences MSc

Verfasst von: Lukas Elmar Amann
Matrikelnummer: 51848626
Abschlussjahr: 2020
Betreut von: Dr. Thomas Biruhs
Lehrgangsort: Dornbirn
Lehrgangsstart: WS 2018

Ich versichere hiermit,

- diese Arbeit selbständig verfasst, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und mich auch sonst keiner unerlaubten Hilfe bedient zu haben,
- diese Arbeit bisher weder im In- noch Ausland in irgendeiner Form als Prüfungsarbeit vorgelegt zu haben,
- die Übereinstimmung dieser Arbeit mit jener Version, die der Betreuung vorgelegt und zur Plagiatsprüfung hochgeladen wurde,
- mit der Veröffentlichung dieser Arbeit durch die Bibliothek der FHWien der WKW einverstanden zu sein, die auch im Fall einer Sperre nach Ablauf der genehmigten Frist erfolgt.

_____

Ort, Datum Unterschrift VerfasserIn

Ich stimme der Veröffentlichung samt Upload der elektronischen Version meiner Masterarbeit durch die Bibliothek der FHWien der WKW in deren Online-Katalog zu. Im Fall einer Sperre der Masterarbeit erfolgt die Veröffentlichung samt Upload erst nach Ablauf der genehmigten Sperrfrist. Diese Zustimmungserklärung kann ich jederzeit schriftlich widerrufen.

_____

Ort, Datum Unterschrift VerfasserIn

Abstract

Der Sanitär-Heizung-Klima-Großhandel befindet sich im digitalen Wandel. Prognosen zeigen, dass der Anteil von in Webshops gekauftem Profimaterial in den nächsten Jahren deutlich zunehmen wird. Dies bringt neben neuen Chancen auch einige Herausforderungen mit sich. Die Branche muss deshalb ihre Geschäftsmodelle anpassen, um gegenüber den großen Online-Versandhändlern weiterhin konkurrenzfähig zu bleiben. Die beste Erfolgsstrategie liegt hierbei in der Digitalisierung von Serviceleistungen, welche den Zweck verfolgen, einen Mehrwert zu bieten und gleichzeitig die Kundenbindung zu erhöhen. Eine innovative Lösung, welche sich im B2C Bereich bereits etabliert hat, im B2B Bereich allerdings noch weitgehend unerforscht ist, bildet dabei Cross-Selling in Kombination mit Predictive Analytics.

Ziel dieser Masterarbeit ist es, mit Hilfe von Predictive Analytics einen Cross-Selling-Prototypen für den SHK-Großhandel zu entwerfen, der dem Kunden/der Kundin auf Basis seiner/ihrer historischen Daten weiterführende Artikel im Webshop vorschlägt.

Der theoretische Teil befasst sich mit der Untersuchung von geeigneten Algorithmen und Methoden, die im Prototypen zur Anwendung kommen. Beleuchtet wurden auch die Probleme, welche derartige Systeme mit sich bringen können. Für die Evaluierung und Beantwortung der theoretischen Forschungsfragen, wurden mittels qualitativer Inhaltsanalyse nach Mayring insgesamt fünf Experten-/Expertinneninterviews durchgeführt, um die Anforderungen und Einflussfaktoren an ein solches System zu ermitteln. Im Praxisteil wurde ein Prototyp realisiert, der das Wissen aus der theoretischen und der empirischen Forschung kombiniert.

Die Ergebnisse zeigen, dass es keine vordefinierte Regel für die Erstellung von Prognosen gibt und die Methodenwahl immer individuell anhand der Aufgabe erfolgen sollte. Probleme können dabei durch fehlerhafte und inkonsistente Daten entstehen, aber auch die gesellschaftlichen Folgen algorithmischer Entscheidungsfindung können umstritten sein. Die Auswertung der qualitativen Inhaltsanalyse zeigt, dass der Vorteil von solchen Systemen hauptsächlich in der beschleunigten Bestellmöglichkeit, sowie als Orientierungshilfe bei auslaufendem Material gesehen wird. Eine erste Auswertung des erstellten Prototyps liefert eine vielversprechende Trefferquote von fast 50%. Das Resultat offenbart großes Potential, welches dazu beitragen kann, die Wettbewerbsfähigkeit in der SHK-Branche zu stärken. Zukünftige Forschungen könnten sich dabei mit weiteren Einflussfaktoren oder Methoden zur Prognose von Warenbestellungen im SHK-Großhandel beschäftigen.

Inhaltsverzeichnis

Abkürzungsverzeichnis

B2B	Business to Business
B2C	Business to Consumer
SHK	Sanitär-Heizung-Klima
KDD	Knowledge Discovery in Databases
CRISP-DM	Cross Industry Standard Process for Data-Mining
UI	User Interface
UX	User Experience
GWS	Gas-Wasser-Sanitär
SQL	Structured Query Language
SVM	Support Vector Machine
ERP	Enterprise Resource Planning
DWH	Data Warehouse
CSV	Comma-separated values
KNN	Künstliche neuronale Netze
PIM	Produktinformationsmanagement

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Der Einsatz von Querverkäufen (Cross-Selling) verspricht Unternehmen günstige und attraktive Wachstumsmöglichkeiten. Der Grundgedanke lautet: Durch den Verkauf von zusätzlichen ergänzenden Produkten, der Kundschaft einen Mehrwert zu bieten und gleichzeitig höhere Umsatzziele zu erreichen. (Malms & Schmitz, 2008, S. 30)

1.1 Problemstellung

Im Rahmen einer repräsentativen Marktumfrage mit 450 Teilnehmern/Teilnehmerinnen aus der deutschen Sanitär-Heizung-Klima-(SHK) Branche, stellte sich heraus, dass neun Prozent aller Einkäufe über Online-Shops abgewickelt werden. Prognosen zeigen, dass der Anteil von in Webshops gekauftem Profimaterial bis 2021 deutlich steigen wird. Führende Akteure/Akteurinnen in der Bauwirtschaft erwarten, dass große Online-Versandhändler wie Amazon in den nächsten zehn Jahren auch im professionellen Baustoffhandel zu Marktführern werden. (BauInfoConsult GmbH, 2018) Der Großhandel befindet sich im Umbruch, die Zukunft ist von einem kompetitiven Konkurrenzfeld geprägt. Digitalisierung und E-Commerce verstärken diese Entwicklung. Aus diesem Grund müssen Unternehmen ihre Geschäftsmodelle anpassen, um weiterhin konkurrenzfähig zu bleiben. Die besten Erfolgsstrategien liegen hierfür in der Digitalisierung von Serviceleistungen, die dem Kunden/der Kundin einen Mehrwert bieten. (KMU Forschung Austria, 2018)

Die bereits genannten Studien zeigen, dass zukünftig der Fokus vermehrt auf den digitalen Vertriebsweg gelegt werden muss. Es wird deshalb nicht mehr nur ausreichend sein einen Online-Shop anzubieten, sondern diesen mit einem echten technologischen Mehrwert auszustatten. (Inman & Nikolova, 2017, S. 1) Innovative Methoden, wie beispielsweise Cross-Selling mittels Predictive Analytics, wäre ein solcher Ansatz, um sich hier von der Konkurrenz abzuheben. (Burow, Gerards, & Demmer, 2017, S. 50) Die prädiktive Analytik verfolgt das Ziel, Muster aufzudecken und Beziehungen in Daten zu erfassen, um beispielsweise eine Vorhersage der nächsten Schritte des Kunden/der Kundin, basierend auf dem was bisher gekauft wurde, zu treffen. Dabei können aber auch diverse Probleme auftreten, beispielsweise, wenn Daten nur zufällig miteinander korrelieren, Datenpunkte fehlen oder der Einfluss von Faktoren nicht messbar ist und es dadurch zu einer Verzerrung der Ergebnisse kommt. (Gandomi & Haider, 2015, S. 143)

1.2 Forschungsstand

Aus Forschungssicht lässt sich zum Themengebiet des Cross-Selling bis rückwirkend zu den 90er-Jahren, ein starker Fokus auf die Branche der Finanzdienstleister beobachten, mit weitgehend sehr praxisorientierter Sichtweise. (Schäfer, 2002, S. 5) In der aktuelleren wissenschaftlichen Literatur geht es vorwiegend um die Effekte auf die Kundenbeziehung und wie sich die Kundenzufriedenheit und das Vertrauen auf die Cross-Selling-Bereitschaft auswirkt. (Wagner, 2008, S. 43-45; Malms & Schmitz, 2008, S. 31-32; Aurier & N'Goala, 2009, S. 304-307)

Was in den letzten Jahren aufgrund des zunehmenden Fortschritts und der immer günstiger werdenden Hardware, im stark umkämpften Einzel- und Versandhandel zugenommen hat, sind die Möglichkeiten der Unternehmen Verbraucherverhalten mittels Predictive Analytics zu prognostizieren, um damit Cross-Selling-Modelle abzuleiten. (McCarthy, Halawi, & Ceccucci, 2019, S. 2) Ein Großteil der Arbeiten im Handelsbereich zu Predictive Analytics und Data Science konzentrieren sich dabei auf den B2C Bereich. Auch wenn schon Lösungen für den B2B Bereich entworfen wurden, verblassen diese im Vergleich zu dem, was im B2C gelernt und angewendet wurde. Mitunter hängt dies damit zusammen, dass sich die beiden genannten Bereiche im Wesentlichen stark unterscheiden und Erkenntnisse sich dadurch nicht einfach umlegen lassen. Boire stellt deshalb fest, dass Predictive Analytics im B2B Bereich noch ein weitgehend unerschlossenes Gebiet darstellt. (Boire, 2017, S. 222) Zu einem ähnlichen Ergebnis kommt auch Lilien, welcher kritisiert, dass die Aufmerksamkeit akademischer Arbeiten fast ausschließlich auf den B2C Bereich konzentriert ist. (Lilien, 2016, S. 543)

Untermauert wird diese These von Reid und Plank (Reid & Plank, 2000), welche Anfang der 2000er anhand einer Auswertung der Top Vier Marketingzeitschriften (Journal of Marketing, Journal of Marketing Research, Journal of Marketing Science, Journal of Consumer Research) festgestellt haben, dass pro Jahr nur zwischen einer und fünf Publikationen im Bereich B2B veröffentlicht werden. Ein weiterer wissenschaftlicher Artikel aus dem Jahre 2009 mit dem Titel „Relative Presence of Business-to-Business Research in the Marketing Literature“ von Laplaca und Katrichis zeigt, dass sich auch nach fast zehn Jahren wenig geändert hat und Publikationen in den besagten Zeitschriften im B2B Bereich immer noch Seltenheitswert haben. (LaPlaca & Katrichis, 2009, S. 1-22) Ein anderer Forschungstrend fokussiert sich auf die verschiedenen Kundengruppen und auf die Herleitung der Cross-Selling-Angebote. Die Analyse basiert hierbei hauptsächlich auf Datenerhebungen, die im Zeitablauf immer wiederkehrend durchgeführt und dann mit Hilfe von Prognosemodellen ausgewertet werden. (Knott, Hayes, & Scott, 2002) Die enormen Datenmengen verhindern in vielen Fällen zeitnahe und flexible Aus-

wertungen, beim modernen Cross-Selling kommen deshalb Analysetools zum Einsatz, um daraus Potenziale abzuleiten. (Wagner, 2008, S. 42) Nicht immer wird dabei strategisch vorgegangen. Prinzie und Van den Poel zeigen auf, dass Cross-Selling-Aktivitäten vorwiegend auf subjektiven Erfahrungen und Intuition der Manager/Managerinnen beruhen. (Prinzie & Van den Poel, 2006, S. 714) Basierend auf dieser Erkenntnis, ist es nachvollziehbar, dass durch mangelhafte Planung gewisse Erwartungen und Wünsche von Cross-Selling bei Kunden/Kundinnen nicht zufriedenstellend erfüllt werden können. Dabei ist gerade der Kunde/die Kundin derjenige/diejenige welcher/welche über den Erfolg von Cross-Selling-Methoden entscheidet, nämlich ob der Kunde/die Kundin die Leistung in Anspruch nimmt oder nicht. (Shah, Kumar, Qu, & Chen, 2012, S. 3)

1.3 Forschungsfragen

Hauptforschungsfrage

- Wie müsste ein Cross-Selling Algorithmus für den SHK-Großhandel gestaltet sein, um den Anforderungen der Zielgruppe zu entsprechen?

Theoretische Subforschungsfragen

1. Welche Methoden und Formeln eignen sich für die Prognosenerstellung?
2. Welche Probleme können beim Einsatz von Predictive Analytics entstehen?

Empirische Subforschungsfragen

1. Wie schätzen Experten/Expertinnen die Praktikabilität eines solchen Cross-Selling Algorithmus ein?
2. Welche Daten und Faktoren empfehlen SHK-Experten/Expertinnen für eine zielführende Vorhersage?

Werkleitende Subforschungsfrage

- Welche Ergebnisse sind von einem Cross-Selling Prototypen zu erwarten, der anhand von Informationen der Experten/Expertinnen und unter Einsatz von Predictive Analytics Methoden für den SHK-Großhandel entwickelt wurde?

1.4 Zielsetzung

Predictive Analytics ist derzeit eines der wichtigsten Anwendungsgebiete von Big Data. (Iffert, 2016, S. 17) Ziel dieser Thesis ist es anhand einer empirischen Untersuchung festzustellen, welche Daten und Faktoren im SHK-Großhandel beim Bestellen von Produkten über den Online Shop eine entscheidende Rolle spielen. (Maitzen, 2016, S. 56) Anhand dessen soll ein Prototyp programmiert werden, welcher dem Kunden/der Kundin weiterführende Produkte (Cross-Selling) vorschlägt.

1.5 Umsetzung

Der Theorieteil bildet die Basis der Forschung und beschäftigt sich mit der Funktionsweise und dem aktuellen Stand von Predictive Analytics, dabei sollen auch bestehende Konzepte analysiert werden, um die nötigen Informationen zur Beantwortung der theoretischen Fragen auszuarbeiten. Das Ganze erfolgt auf Basis einschlägiger Forschungs- und Fachliteratur. Im empirischen Teil wird die qualitative Forschung mit Experten-/Expertinneninterviews angewandt. Dabei sollen Experten/Expertinnen der SHK-Branche zu den Subforschungsfragen interviewt werden. Die Datenauswertung erfolgt nach den Regeln der qualitativen Inhaltsanalyse nach Mayring. Anhand der gewonnenen Daten soll in weiterer Folge auch ein Prototyp des Cross-Selling Algorithmus in einem SHK-Großhandelsbetrieb programmiert und dokumentiert werden.

2 Begriffsabgrenzungen und Definition

2.1 Cross-Selling

Cross-Selling ermöglicht Unternehmen die Kundenbeziehung weiter auszubauen, um dem Kunden/der Kundin durch den Querverkauf von zusätzlichen Produkten oder Leistungen einen Mehrwert zu bieten, von dem beide Seiten profitieren. (Malms & Schmitz, 2008, S. 30) Beim modernen Cross-Selling werden Analysetools verwendet, um die vergangenen Einkäufe der Kunden/Kundinnen zu analysieren und daraus Cross-Selling Potenziale abzuleiten. (Wagner, 2008, S. 42)

2.2 Data-Mining

Data-Mining zielt darauf ab, nützliche Muster aus großen Datenmengen zu finden und zu extrahieren. Es wird häufig bei Geschäftsprozessen angewendet, um kritische Entscheidungen zu treffen, kann aber bei allen Arten von Daten eingesetzt werden. Data-Mining umfasst verschiedene Techniken wie Assoziations-, Klassifizierungs-, Clustering-, Vorhersageanalysen usw. und gilt als eine der wichtigsten Entwicklungen in der Informationstechnologie. (Lokanatha & Venkatadri, 2011, S. 19)

2.3 Predictive Analytics

Predictive Analytics beschreibt eine fortgeschrittene Analysemethode, welche unter Verwendung von empirischen Methoden, Vorhersagen von Ereignissen oder Trends basierend auf zuvor beobachteten historischen oder neuen Daten berechnet. Die Grundlage bildet oftmals Data-Mining in Kombination mit weiteren komplexen Algorithmen. Dabei werden nicht nur Modelle generiert, sondern auch der Aufbau und das Testen von Theorien unterstützt. (Shmueli & Koppius, 2011, S. 553-555)

2.4 Datenmodell

Datenmodelle beschreiben alle in der Datenbank verfügbaren Objekte und ihre Beziehungen zueinander. Bei der Entwicklung von Informationssystemen dienen Datenmodelle dazu, Strukturen in den Daten zu finden und festzulegen. (Ester & Sander, 2000, S. 16)

2.5 Big Data

Big Data bezeichnet große Mengen an komplexen Daten, die mittels fortschrittlicher Technologie erfasst, gespeichert, analysiert und verteilt werden können. Als Merkmal für die Beschreibung haben sich die fünf V's etabliert:

- Volumen (Volume)
- Vielfalt (Variety)
- Geschwindigkeit (Velocity)
- Wert (Value)
- Glaubwürdigkeit (Veracity)

(Gandomi & Haider, 2015, S. 138)

2.6 Data Science

Data Science ist eine Wissenschaft, die sich mit Hilfe von Techniken und Methoden aus der Mathematik, Statistik, Stochastik und Informatik mit der Extraktion von Wissen aus großen Datenmengen beschäftigt. (Provost & Fawcett, 2013, S. 4-5)

2.7 Big Data Analytics

Big Data Analytics beschreibt Datenbanken und Data-Mining Technologien, die ein Unternehmen einsetzen kann, um umfangreiche komplexe Daten auszuwerten, mit dem Ziel die Unternehmensleistung zu erhöhen. (Kwon, Lee, & Shin, 2014, S. 387) Big Data Analytics beschreibt dabei alle Schritte von der Datenspeicherung, über die Verwaltung, bis hin zur Analyse der Daten. (Chen, Chiang, & Storey, 2012, S. 1166)

2.8 Statistik

Statistik wird im Data-Mining oft verwendet, um wichtige Kennzahlen aus Daten zu erheben, häufig auch, wenn sich Kennzahlen nicht verallgemeinern lassen, wie beispielsweise das durchschnittliche Einkommen der Bürger/Bürgerinnen eines Landes. Generell sollte Statistik immer im Einklang mit dem Businessproblem stehen. Die Statistik hilft auch, Hypothesen zu erstellen oder festzustellen, ob ein beobachtetes Muster wahrscheinlich gültig ist. Viele Techniken zum Extrahieren von Modellen oder Mustern haben ihren Ursprung in der Statistik. (Provost & Fawcett, 2013, S. 35-36)

2.9 In-Memory-Technologie

Im Gegensatz zu Datenbanken, welche sich auf einem Festspeicher befinden, ermöglicht die In-Memory-Technologie die Daten in den Arbeitsspeicher des Computers auszulagern. Dies hat den Vorteil, dass die Daten im Schnitt 10- bis 100-mal schneller verarbeitet werden können als auf einem „klassischen“ Datenträger. (Loos, et al., 2011, S. 383-384) Entwickelt wurde diese Technologie bereits in den 90er Jahren und bekam durch die fallenden Arbeitsspeicherpreise der letzten Jahre noch einmal einen ordentlichen Schub. Ein Nachteil besteht darin, dass es sich bei Arbeitsspeichern um einen „nicht-persistenten“ Speicher handelt, bei einem Absturz kann es somit zu einem Datenverlust kommen. (Matt, 2012, S. 229)

2.10 Prototyp

Prototypen werden im Bereich Software Engineering als eine Vorgehensweise zur Erstellung eines noch nicht endgültigen Softwaresystems angesehen, dessen Ziel es ist, frühestmöglich eine lauffähige Version zu haben, an der Veränderungen und Optimierungen vorgenommen werden können. (Hallmann, 1990, S. 11)

3 Theoretische Grundlagen

3.1 Predictive Analytics

In den vergangenen Jahren haben die Unternehmen erheblich in die Infrastruktur investiert, um die Möglichkeiten der Datensammlung zu verbessern. Aufgrund der Verfügbarkeit dieser Datenmengen, sind die Unternehmen bestrebt, den größtmöglichen Nutzen daraus zu ziehen. Konnten Unternehmen früher Statistiker/Statistikerinnen, Entwickler/Entwicklerinnen und Analysten/Analystinnen einsetzen, um die Daten auszuwerten, so ist es heutzutage aufgrund der Größe und des Umfangs der Daten fast unmöglich, diese manuell auszuwerten. Dieser Umstand hat dazu geführt, dass immer mehr Unternehmen auf Techniken von Data Science und Data-Mining zurückgreifen. (Provost & Fawcett, 2013, S. 1-2)

3.1.1 Was ist Predictive Analytics?

„It is human nature to want to know and predict what the future holds." Eine Methode, die uns dabei unterstützen soll, ist Predictive Analytics. Predictive Analytics befasst sich mit der Vorhersage von Ereignissen basierend auf zuvor beobachteten historischen oder aktuellen Daten. Die Daten werden gesammelt, aufbereitet und mit Hilfe von modernen Algorithmen transformiert. Daraus lassen sich Prognosemodelle generieren, welche uns einen Ausblick auf die Zukunft werfen lassen. (Mishra & Silakari, 2012, S. 4434)

Predictive Analytics entwickelte sich aus mehreren Disziplinen, wovon einige bereits vor mehreren hundert Jahren zum Auffinden von Mustern verwendet wurden. Diese umfassen Mustererkennung, Statistik, maschinelles Lernen, künstliche Intelligenz und Data-Mining.

Die prädiktive Analytik ist datengesteuert, somit leiten Algorithmen die Schlüsseleigenschaften der Modelle ab und nicht die Analysten/Analystinnen aufgrund von Annahmen. (Abbott, 2014, S. 3-4) Aufgrund der zunehmenden Digitalisierung der Wirtschaft und Gesellschaft wächst der Bedarf von Big Data Analytics. Oftmals werden die Begriffe Big Data, Data-Mining, maschinelles Lernen und Predictive Analytics in unternehmensstrategischen Diskussionen verwendet, ohne dass zwischen den einzelnen Methoden und Disziplinen differenziert wird. Somit bleibt oft unklar, was sich hinter den Methoden genau verbirgt. Der Begriff Data-Mining wird schon seit vielen Jahren verwendet, während Predictive Analytics erst seit einigen Jahren zur Anwendung kommt. Data-Mining ist häufig der erste Prozessschritt in der Datenanalyse, der einen mehrstufigen Prozess zur Wissensgenerierung umfasst. Somit dient Data-Mining als Überbegriff für alle datengestützten Analyse- und Prognoseverfahren. Dieser leitet sich wiederum vom Begriff Big Data Analytics ab, der die Bereiche Data-Mining, Machine Learning und Predictive

Analytics umfasst. Diese Betrachtungsweise soll dabei helfen, den relativ neuen Terminus „Predictive Analytics“ besser einzuordnen. Der Fokus des Data-Minings liegt hauptsächlich in der explorativen Analyse bestehender Datenstrukturen, wie beispielsweise Texte oder Web-Inhalte, während sich Predictive Analytics auf zukünftige Ereignisse oder Trends konzentriert. (Brühl, 2019, S. 1-4)

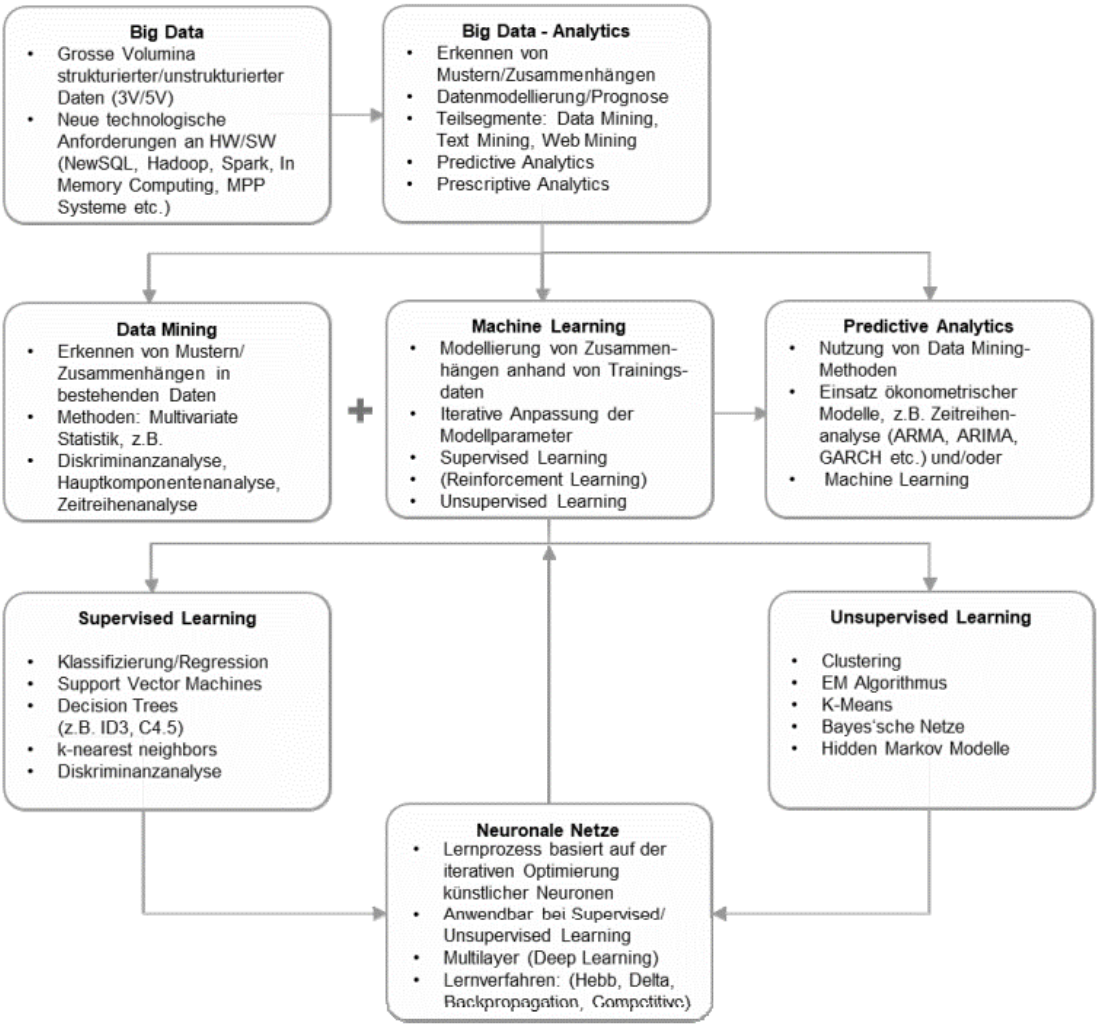


Abbildung 1. Big Data, maschinelles Lernen und Predictive Analytics.

Quelle: (Brühl, 2019, S.3)

Predictive Analytics stellt auch ein Teilgebiet von Business Intelligence dar. Dabei werden gezielt Daten verarbeitet, um Entwicklungen und Leistungen zu verstehen, aber auch um zu prognostizieren und bewerten. Hierfür kann je nach Fragestellung unter drei Arten von Analytics unterschieden werden. (McCarthy, Halawi, & Ceccucci, 2019, S. 10)

- **Descriptive Analytics**: Damit soll die Frage geklärt werden, was passiert ist. Darunter fallen Ad hoc Abfragen oder auch periodische Reports. Das Ziel ist es Problembereiche zu finden oder über Geschäftsbereiche zu informieren.
- **Predictive Analytics**: Hierbei wird versucht anhand von Daten oder mittels mathematischen Verfahren bestimmte Muster zu finden, welche dann in einer Prognose verwendet werden können. Die Frage lautet deshalb „Was wird passieren und weshalb?“. Häufig verwendete Verfahren sind Data-Mining oder Prognoserechnungen. Ziel ist es Prognosen und deren Erklärungen zu erstellen.
- **Prescriptive Analytics**: Hierbei werden mathematische Methoden angewandt, um Handlungsempfehlungen zu generieren. Als Basis dienen Daten und/oder Expertenwissen, welche für die Erstellung der Empfehlungen benötigt werden. Typische Anwendungsgebiete sind Simulationen oder Expertensysteme. Das Ergebnis ist entweder eine unterstützende Information für den Handelnden/die Handelnde, oder eine Handlungsalternative. (Christ & Ebert, 2015, S. 300-301)

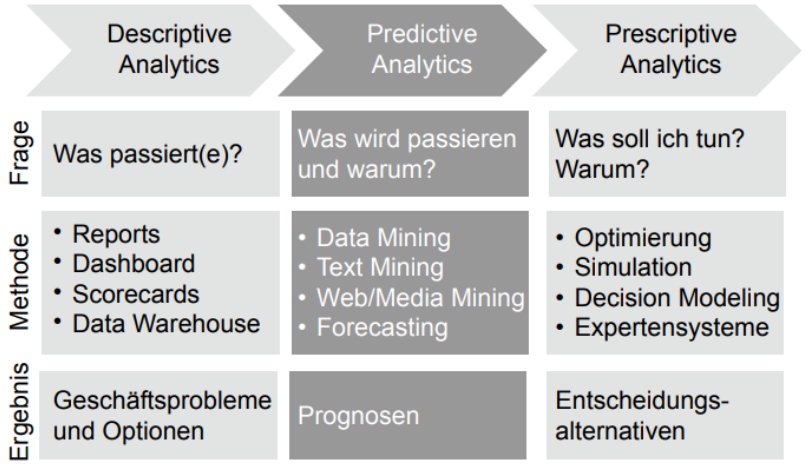


Abbildung 2. Taxonomie von Analytics.

Quelle: (Christ & Ebert, 2015, S.300)

3.1.2 Technische Voraussetzungen

Abhängig von der Größe und Komplexität der Daten ist bereits eine Excel-Funktion ausreichend, um beispielsweise mit Hilfe der exponentiellen Glättung eine Bedarfsplanung zu errechnen. Sinnvoller ist der Einsatz von spezieller Software, die den Nutzer/die Nutzerin bei der Wahl des Prognosemodells unterstützen kann und diese dann an Folgesysteme weiterleitet. (Iffert, 2016, S. 17-18)

Es gibt eine große Anzahl an Data-Mining Techniken, die Wichtigsten sind:

- **Assoziation**: Ermöglicht das Auffinden von Häufigkeiten und Mustern in den Datensätzen, um daraus Schlussfolgerungen abzuleiten.
- **Klassifikation**: Die Klassifizierung ermöglicht es Daten in Kategorien zusammenzufassen, um sie vergleichbar zu machen.
- **Regression**: Regressionsanalysen helfen dabei, Beziehungen in Datenstrukturen zu identifizieren.
- **Clustering**: Clustering wird verwendet, um ähnliche Objekte basierend auf ihren Charaktereigenschaften und Ähnlichkeiten zu gruppieren. (Alharan, Al-Haboobi, & Alsagheer, 2017, S. 134)

Die aktuell erhältlichen Predictive Analytics Softwarepakete bieten nicht nur die mathematischen Basisfunktionen, sondern auch bereits vordefinierte Modelle, welche die zuweilen recht komplizierten Berechnungen mittels bereits vorgedachter Logik dem Anwender/der Anwenderin möglichst benutzerfreundlich präsentieren sollen. So sind verschiedene Standardauswertungen wie Warenkorbanalysen, oder Kundenklassifizierungen meistens schon als Vorlagen abrufbar. Komplexer wird es allerdings, wenn mehrere Faktoren in das Data-Mining Modell einfließen sollen. Derartige Weiterentwicklungen sind nicht trivial und erfordern Personen, welche über das nötige Analyse- und Statistik-Know-how, sowie IT-Wissen verfügen. (Iffert, 2016, S. 17-18)

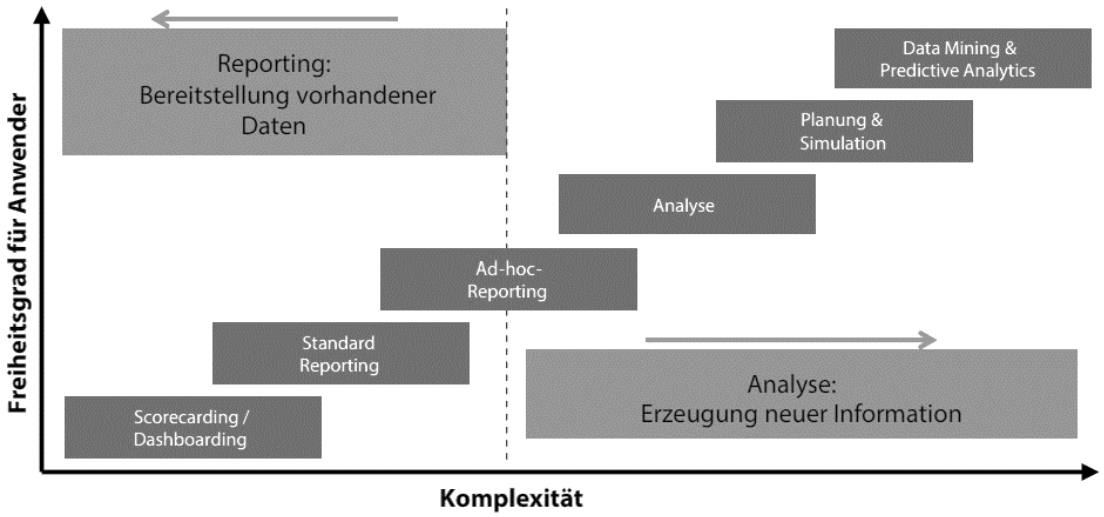


Abbildung 3. Komplexität von Analysepraktiken

Quelle: http://barc.de/predictive abgerufen am 04.12.2019

3.1.3 Predictive Analytics Projekte umsetzen

Für den Erfolg von Predictive Analytics Vorhaben ist es unumgänglich, dass der Projektablauf effektiv gestaltet wird. Hierbei haben sich zwei Methoden als praktikabel erwiesen. Zum einen die KDD-Methodik und zum anderen das CRISP-DM als Standard-Prozessmodell für Data-Mining. (Iffert, 2016, S. 18-19)

3.1.3.1 KDD-Methodik

Fayyad definiert den Begriff KDD wie folgt: "Knowledge Discovery in Databases describes the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data." Damit wird beschrieben, dass es sich hierbei um einen nicht trivialen Prozess handelt, dessen Aufgabe es ist, Muster aus Datensätzen zu extrahieren und diesen Eigenschaften zuzuordnen. Diese Eigenschaften sollen für einen Großteil des Datensatzes gültig sein und leicht verständliche Zusammenhänge innerhalb des Datensatzes beschreiben.
Fayyad unterteilt in seiner Abbildung fünf Phasen für sein Prozessmodell. (Fayyad, Piatetsky-Shapiro, & Padhraic , 1996, S. 39)

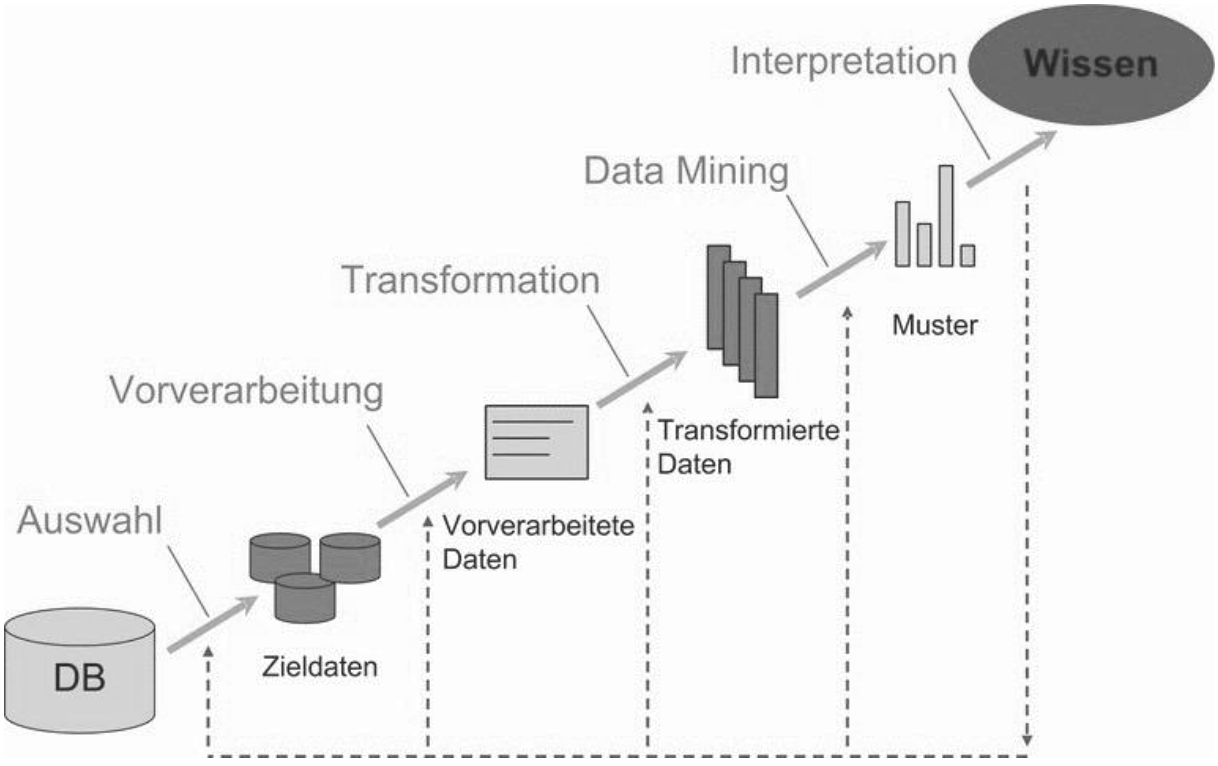


Abbildung 4. KDD-Prozess nach Fayyad, Piatetsky-Shapiro, & Padhraic

Quelle: (Fayyad, Piatetsky-Shapiro, & Padhraic., 1996, S.41)

In der ersten Phase erfolgt die Auswahl der Zieldaten, die für das Projekt relevant sind. Diese Zieldaten werden in weiterer Folge in einer Vorverarbeitung bereinigt und etwaige Qualitätsmängel oder fehlerhafte Werte behoben. In dieser Phase können die Daten bereits mit weiteren Attributen angereichert werden, um die spätere Auswertung zu erleichtern. In der Transformationsphase werden die Daten in die gewählte Analyseform gebracht, beispielsweise das Umwandeln von Datentypen. Anschließend werden mathematische Funktionen angewendet, um bestimmte Muster im Datenmodell sichtbar zu machen. Diese Muster müssen dann von Experten/Expertinnen beurteilt und interpretiert werden, um das Wissen zu generieren. Sollte das Ergebnis nicht zufriedenstellend sein, kann ein Rücksprung in eine vorherige Phase erfolgen. (Fayyad, Piatetsky-Shapiro, & Padhraic , 1996, S. 39-41)

3.1.3.2 CRISP-DM

Im Gegensatz zum KDD-Modell nach Fayyad (Fayyad, Piatetsky-Shapiro, & Padhraic , 1996), ist der CRISP-DM (Cross-industry standard process for data-mining) ein Branchen- und Industriestandard. Ausgangspunkt ist hier eine betriebswirtschaftliche Problemstellung. Hierbei wird der zyklische Charakter von Projekten in den Vordergrund gestellt. (Göpfert & Breiter, 2015, S. 1220)

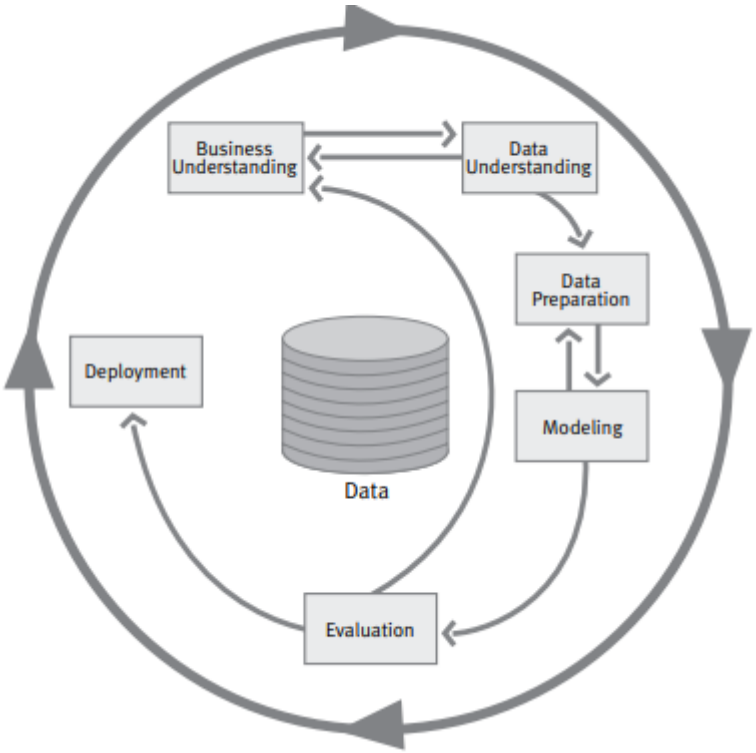


Abbildung 5. Phasen des CRISP-DM

Quelle: (Göpfert & Breiter, 2015, S. 1221)

Am Anfang steht hierbei der Konsens zwischen den Fachabteilungen und der mit der Durchführung beauftragten Mitarbeiter/Mitarbeiterinnen an. Hierbei soll das grundlegende Problem und die Zielsetzung des Projekts besprochen werden (Business Understanding). Anschließend wird die Datenbasis analysiert, inwieweit sie zur Lösung des Problems beiträgt (Data Understanding). Dabei sind auch Rücksprünge zur vorherigen Phase möglich. In der Datenaufbereitung (Data Preparation), soll die Datenbasis von Qualitätsmängeln befreit werden. Hier kann auch schon eine Transformation der Daten in Richtung Algorithmus erfolgen. Im Modelling (Modeling) werden dann durch den Einsatz von Algorithmen Muster erkannt, welche dann in der Evaluation (Evaluation) mit dem Projektziel abgeglichen werden. Nach der erfolgreichen Evaluation kann die Anwendung oder die Daten dem Unternehmen zur Verfügung (Deployment) gestellt werden. (Göpfert & Breiter, 2015, S. 1221)

3.1.4 Anwendungsbeispiele

Der größte europäische Softwarekonzern SAP arbeitet schon seit 2012 mit Predictive Analytics und Cross-Selling Methoden. Daraus entstanden ist ein Programm mit dem Namen „SAP Precision Retailing“. Ziel der Softwarelösung für Einzelhandelsunternehmen ist es, den registrierten Kunden/Kundinnen personalisierte Angebote und Vorschläge in Echtzeit zu unterbreiten. Dabei werden nach jedem Einkaufsvorgang Daten über den Kunden/die Kundin und dessen Vorlieben gesammelt. Nach einer Auswertung der Daten, können dem Kunden/der Kundin personalisierte und kontextabhängige Angebote in Echtzeit angezeigt werden. (Binckebanck & Elste, 2016, S. 101) SAP gibt dabei an, dass die Konversionsraten bei Verkaufsaktionen um bis zu 20 Prozent höher sind, die durchschnittliche Größe des Warenkorbs auf 15 Prozent wächst und der Erfolg von Up- und Cross-Selling Angeboten um 10 Prozent steigt. Dabei wird gezielt versucht, die Entscheidung eines Kunden/einer Kundin zu beeinflussen und das nicht nur vor dem Einkauf, sondern auch während des gesamten Einkaufs. Das Ganze wird im Wesentlichen damit realisiert, dass kleinere alltägliche Dinge gezielt an der Kasse platziert werden, während der Kunde/die Kundin in der Warteschlange steht. (King, 2012) Auf der einen Seite bieten solche Systeme kostengünstige Möglichkeiten für Einzelhändler personalisierte Werbung anzubieten, auf der anderen Seite steigen dabei auch die Bedenken der Kunden/Kundinnen hinsichtlich Privatsphäre und missbräuchlicher Verwendung der Daten. (Inman & Nikolova, 2017, S. 17) Teilweise werden derartige Einzelhandelstechnologien, welche zu sehr in die Privatsphäre der Kunden/Kundinnen eingreifen, eher gemieden, womit der ganze Vorteil wieder wegfällt. (White, 2004, S. 41)

Solche Datenschutzbedenken sind längst keine Seltenheit mehr, wie ein aktueller Fall des Österreichischen Vereins für Konsumenteninformation (VKI) zeigt. So kritisieren sie in ihrem Artikel den Zusammenschluss großer österreichischer Handelsunternehmen zum Zwecke der Einführung einer Bonuskarte, durch den der Konsument/die Konsumentin Vorteilspunkte sammeln kann. Dabei wird vor allem der Umgang mit den persönlichen Daten und die mühsame Datenlöschung, sowie Kündigung bemängelt. (Ecker, 2019) Ein anderes Beispiel aus dem Einzelhandel ist Walmart. Die größte Einzelhandelskette im B2C Bereich, errechnete 2004 vor Eintreffen des Hurrikan Frances, mit Hilfe von Predictive Analytics, dass ihre Geschäfte nicht nur Taschenlampen brauchen würden, sondern auch Erdbeer-Pop-Tarts. Die Software erkannte, dass vor großen Umweltkatastrophen die Verkäufe von Pop-Tarts um das Siebenfache steigen würden. Der Grund war, dass aufgrund des Hurrikans in weiten Teilen der USA der Strom sowie die Gasleitungen ausfielen. Das Teiggebäck hatte den Vorteil, dass es keine Zubereitung brauchte. In Folge dieser Erkenntnis, gelang es Walmart frühzeitig die Geschäfte mit den nötigen Mengen zu beliefern, um den benötigten Bedarf zu decken. (Lee & Kang, 2015, S. 76)

Auch außerhalb des Retail Bereiches gibt es erfolgreiche Beispiele bei der Umsetzung von Predictive Analytics Projekten. UPS, einer der weltweit führenden Logistikanbieter, verwendet seit 2013 die gesammelten Daten zur intelligenten Steuerung ihres Logistiknetzwerkes. Dafür entwickelten sie 2017 auch ein eigenes System mit dem Namen ORION (On Road Integrated Optimization and Navigation). Das Flottenmanagementsystem verwendet Fahrzeugtelematik in Kombination mit Predictive Analytics, um die Fahrzeuglieferrouten dynamisch für jeden/jede der insgesamt 55.000 Fahrer/Fahrerinnen weltweit zu optimieren. Mit Hilfe dieser Methode ist es jedes Jahr möglich, die Liefermeilen um 100 Millionen und die Kohlenstoffemissionen um 100.000 Tonnen zu reduzieren. Die Verwendung von Big Data Analysen führte dabei zu einer höheren Effizienz im Betrieb, Kraftstoff- und Fahrzeugeinsparungen und zu einer Reduzierung der Treibhausgase. (Samuels, 2017)

Versicherungsbetrug ist in den Vereinigten Staaten von Amerika laut dem CAIF (Coalition Against Insurance Fraud) ein Multimilliardenproblem und kostet jedes Jahr ca. 80 Milliarden Dollar. Versicherungsunternehmen investieren deshalb viel Zeit und Ressourcen in das Aufdecken und Verhindern von Versicherungsbetrug. Verschiedene Studien belegen, dass die Anzahl der Fälle jedes Jahr zunehmend ist. Aber nicht nur die Betrugsfälle haben zugenommen, sondern auch die Möglichkeiten diese zu entdecken. Die zunehmende Verfügbarkeit von Datenquellen bietet Versicherungsgesellschaften neue Möglichkeiten diese mit Hilfe von Data-Mining auszuwerten, zu erkennen und auch proaktiv zu unterbinden. (Power, 2015, pp. 1-2) Auch

bei der Festlegung von Versicherungsprämien kommt Predictive Analytics zum Einsatz. Versicherungsgesellschaften berechnen damit, wie hoch die Wahrscheinlichkeit ist, dass ein Fahrer/eine Fahrerin einen Unfall haben wird. Die Fahrverhaltensdaten stammen dabei aus speziellen Boxen, welche im Auto des Versicherungsnehmers installiert werden, oder werden anhand einer App mittels GPS und Mobilfunkdaten ausgewertet. (McCarthy, Halawi, & Ceccucci, 2019, S. 4) Auch die Polizei hat längst die Vorteile von Predictive Analytics bei der Bekämpfung und Prävention von Kriminalität entdeckt. Wurden früher Straftaten auf analogen Karten eingetragen, erfolgt heutzutage alles digital. Anhand dieser Daten lassen sich für die Kriminalfahnder Muster ableiten, die die Polizeipräsenz in den betroffenen Gebieten erhöht, um damit präventiv gegen Kriminalität vorzugehen. Die Daten werden dabei täglich aktualisiert, um die Polizeistreifen dynamisch zu verteilen. Es werden längst nicht nur mehr vergangene Straftaten als Datenbasis herangezogen, auch andere Variablen wie beispielsweise Bevölkerungsdichte, Verkehrsmuster, Geodaten und soziale Netzwerke fließen in die Algorithmen mit ein. (Lersch & Chakraborty, 2020, S. 87-99)

3.1.5 Risiken und Herausforderungen

Die Verwendung von Big Data Analysen zur Entscheidungsfindung für Regierungen und Finanzunternehmen kann diesen Prozess erheblich beschleunigen und verbessern. Weltweit existieren zahlreiche automatische Entscheidungsalgorithmen, die in einer Vielzahl von Bereichen eingesetzt werden, insbesondere diejenigen, mit denen die Eignung einer Person für Versicherungen oder Kredite beurteilt werden kann oder auch zur Segmentierung von Risikogruppen. (Pérez-Martin, Pérez-Torregrosa, & Vaca, 2018, S. 448-449) Robinson, Harlan und Rieke kritisieren dabei, dass der menschliche Faktor komplett wegfällt und Algorithmen anhand vordefinierter Variablen entscheiden. Solche Systeme bergen allerdings das Risiko, unfaire oder unüberlegte Ergebnisse zu liefern, die auf verzerrte oder diskriminierende Programmierung zurückzuführen sind. Diese Besorgnis über die potenziell diskriminierenden Aspekte von Big Data Analysen ist weit verbreitet, und Forscher/Forscherinnen aus einer Vielzahl wichtiger Branchen, in denen Big Data derzeit verwendet wird, haben Beispiele aus der Praxis identifiziert. (Robinson, Harlan, & Rieke, 2014)

Abbott sieht die Hauptursachen warum Predictive Analytics Modelle scheitern können, hauptsächlich in prozessbezogenen Problemen und fasst diese in vier Gruppen zusammen. (Abbott, 2014, S. 12)

- **Probleme beim Modellieren**

  Eines der größeren Probleme bei Vorhersagemodellen aus Sicht des Analysten/der Analystin ist die Überanpassbarkeit. Dies geschieht, wenn das Modell zu komplex wird und die Interpretation der Daten nicht mehr zuverlässig funktioniert. Speziell neue Daten werden dann nicht immer korrekt interpretiert. Ein weiteres Hindernis können übereifrige Analysten/Analystinnen sein, die versuchen das komplette Modell in einem Entwurf zu erstellen. Dabei gestaltet sich die Fehlersuche oder überhaupt der Abschluss des Projektes als zu zeitintensiv. Wie in Abbildung 6 dargestellt, ist es oftmals ratsamer mit einfacheren Modellen zu starten, diese zu validieren und dann sukzessive zu erweitern. (Abbott, 2014, S. 13-14)

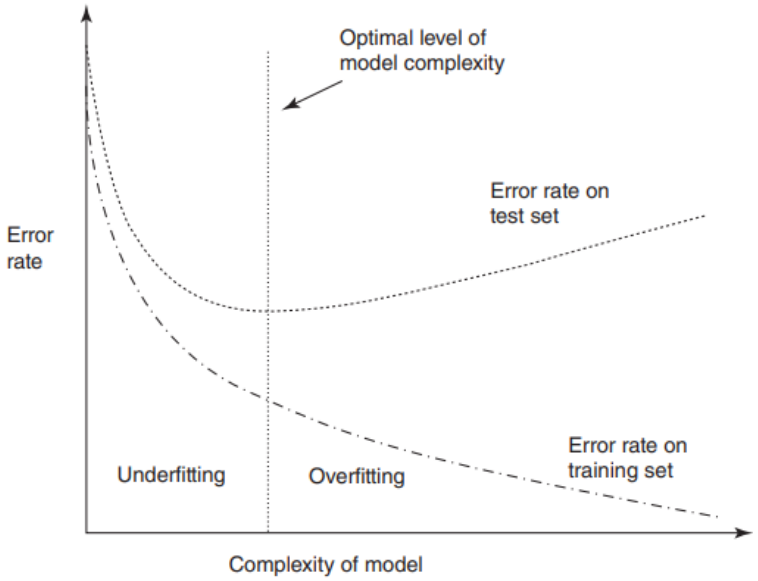


Abbildung 6. Darstellung der optimalen Modellkomplexität

Quelle: (Larose, 2015, S.163)

- **Fehlende Unterstützung des Managements**

  Predictive Analytics Projekte bündeln oft eine Menge an Ressourcen im Unternehmen, umso wichtiger ist dabei die Unterstützung des Managements. Sollte also das Management das Projekt nicht vollends unterstützen, kann es mitunter schwierig werden dieses umzusetzen. (Abbott, 2014, S. 12)

- **Inkonsistente Daten**

  Predictive Analytics Modelle brauchen oftmals Daten aus verschiedenen Systemen. Diese müssen mittels Schlüssel im Datenmodell miteinander verbunden werden. Sollten diese Schlüsselfelder in den Tabellen nicht existieren, können Projekte scheitern, bevor sie überhaupt begonnen haben. Eine andere Problematik ergibt sich bei Datensätzen die

laufend aktualisiert werden, so kann es beispielsweise bei einseitig aktualisierten Adressen oder Postleitzahlen passieren, das Geo-Analysen nicht mehr nachvollziehbar sind. (Larose, 2015, S. 20-23)

- **Probleme bei der Bereitstellung**

  Eine weitere Hürde bei der erfolgreichen Umsetzung von Predictive Analytics Projekten kann die Bereitstellung des Modells sein. In der Regel sind diese Modelle rechnerisch nicht sehr komplex, moderne Prozessoren können diese Operationen binnen weniger Sekunden problemlos verarbeiten. Was allerdings viel Zeit in Anspruch nehmen kann, ist die Bereitstellung der Daten. Hierbei müssen teilweise mehrere Gigabyte an Datensätzen extrahiert und transformiert werden, um eine homogene Datenbasis zu schaffen. Problematisch kann es auch bei Analysen werden, die Entscheidungen fordern, die binnen weniger Sekunden benötigt werden, wie beispielsweise Wetter- oder Aktienmarkt-Modellen. (Abbott, 2014, S. 14)

3.2 Data-Mining und Wissensfindung in Daten

Die Zahl wissensbasierter Systeme ist in den letzten Jahren rasant gestiegen. Immer mehr Prozesse werden von Computersystemen übernommen. Entscheidungsunterstützende Software wird in der Wirtschaft oder bei Banken eingesetzt, um komplexe Zusammenhänge zu verstehen. Die rasante Entwicklung im Soft- und Hardware-Bereich macht es schier unmöglich hier noch einen Überblick zu bewahren. Dennoch haben alle Systeme eines gemeinsam, intelligentes Denken und Handeln zu simulieren. Deshalb muss Wissen gezielt verarbeitet und dargestellt werden. Aus diesem Grund gibt es hier nicht einfach die beste Methode, vielmehr muss aus unterschiedlichen Ansätzen diejenige Methode gewählt werden, welche optimal zur Lösung des Problems beiträgt. (Beierle & Kern-Isberner, 2019, S. 1) Um einen effizienten Cross-Selling Algorithmus zu finden, spielt Data-Mining bei Auswertungen zu historischen Transaktionsdaten eine wichtige Rolle. Während Warenkorbanalysen hauptsächlich den statistischen Zusammenhang herstellen, welcher Kunde/welche Kundin welche Produkte zusammen kauft, bietet uns Data-Mining die Möglichkeit, weitere Beziehungen in den Daten zu finden und diese als Attribut abzuspeichern. Chen und Tung beschreiben Data-Mining als einen nicht trivialen Prozess dessen Ziel es ist, unbekannte aber potenziell nützliche Informationen in Daten zu entdecken. Das neu entdeckte Wissen lässt sich in einer Vielzahl an Anwendungen einsetzen, wie Marktanalysen, Betrugserkennung, Kundenbeziehungsmanagement und anderen geschäftlichen Entscheidungsprozessen. (Chen & Tung, 2006, S. 1504-1505)

3.2.1 Datenverständnis

Datenverständnis spielt beim Modellieren eine große Rolle. Nachdem alle relevanten Daten für ein Projekt gesammelt wurden, ist es die Aufgabe des Datenanalysten/der Datenanalystin diese zu untersuchen. Er/sie ist auch die erste Person, welcher/welche die Daten in der Modellierungsphase als erstes überhaupt betrachtet und somit auch alle Unvollkommenheiten und Probleme identifizieren muss, die bisher unbekannt waren oder ignoriert wurden. Abbott (Abbott, 2014, S. 35) empfiehlt deshalb, die Daten zu visualisieren, um zum einen Probleme oder ungültige Werte zu identifizieren und zum anderen einen Überblick über das große Ganze zu bekommen. Nicht immer sind Personen oder Fehleingaben die Ursache für fehlerhafte Daten, durch den Einsatz verschiedener Software und Schnittstellen kann es häufig zu Transformationsfehlern kommen. Dies tritt auf, wenn die Software einen Datentyp falsch interpretiert. Beispielsweise werden führende Nullen oftmals entfernt oder es kommt zu Fehlinterpretationen aufgrund von zahlreichen unterschiedlichen Datumsformaten. Nicht immer sind die Probleme dabei offensichtlich, in einer Transaktionstabelle beispielsweise, in der unterschiedliche Währungen vorhanden sind, können Umsätze erst nach einer Umrechnung auf die Landeswährung miteinander summiert werden, dasselbe gilt bei Mengen und Einheiten. Es wäre nicht zielführend Gewichts- und Längenmaße zu addieren. Deshalb müssen speziell bei Summierungen, die Daten immer im Kontext zueinander betrachtet werden. (Abbott, 2014, S. 35-36)

Ein weiterer wichtiger Aspekt, der beim Modellieren von Daten berücksichtigt werden muss, ist deren Lebenszyklus. Es gibt durchaus Bereiche in denen historische Daten bis weit in die Vergangenheit nützlich sind. In der Regel jedoch, haben Daten in einem Unternehmen ein bestimmtes Ablaufdatum. Dieses beginnt mit der Speicherung der Daten und nimmt mit zunehmendem Fortschreiten der Zeit immer mehr ab. Diese abnehmende Nützlichkeit, wie in Abbildung 7 dargestellt, hat mehrere Gründe. (Inmon, Linstedt, & Levins, 2019, S. 33-35)

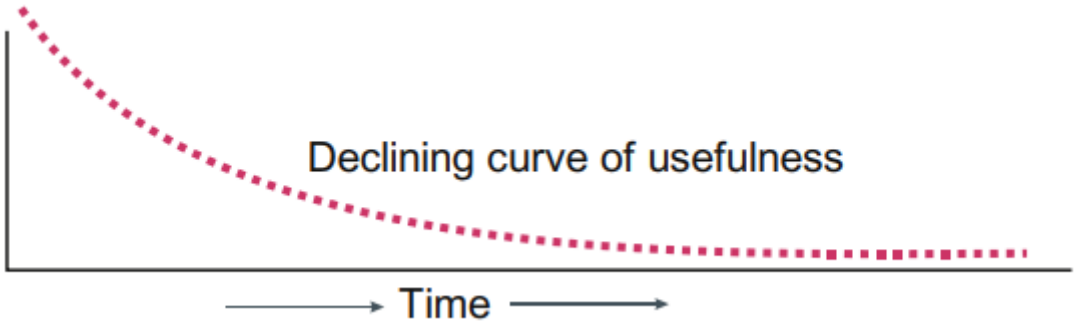


Abbildung 7. Abnehmende Nützlichkeit von Daten

Quelle: (Inmon, Linstedt, & Levins, 2019, S. 35)

Zum einen werden Entscheidungen von Unternehmen stark von aktuellen Trends und wirtschaftlichen Faktoren beeinflusst und zum anderen entwickeln sich Unternehmen oft weiter und erschließen neue Geschäftsfelder. Diese beiden Faktoren können sich innerhalb von Jahren stark verändern, was die Ableitung von historischen Daten auf die Zukunft äußerst schwierig gestaltet und somit den Wert der Daten über die Zeit gesehen, drastisch sinken lässt. (Inmon, Linstedt, & Levins, 2019, S. 35-37)

3.2.2 Aufgaben des Data-Mining

Das Ziel von Data-Mining ist es, unter Anwendung von effizienten Algorithmen gültige Muster in Daten zu finden. (Fayyad, Piatetsky-Shapiro, & Padhraic , 1996, S. 37)

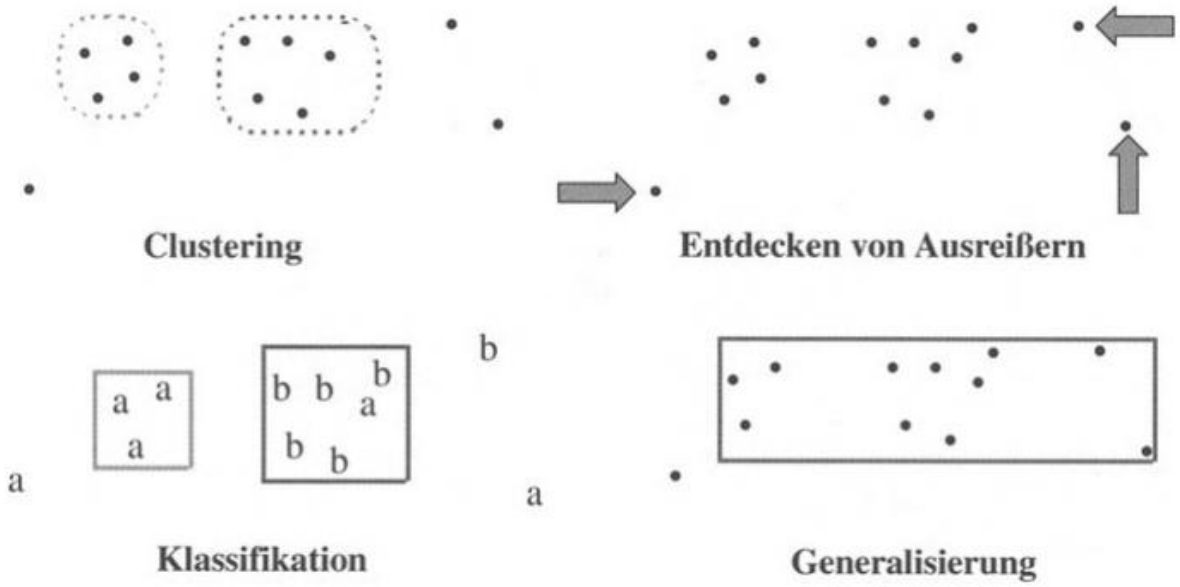


Abbildung 8. Data-Mining Aufgaben

Quelle: (Ester & Sander 2000, S.5)

Die wichtigsten Data-Mining Aufgaben lassen sich wie folgt zusammenfassen:

- **Clustering/Ausreißer:** Ziel ist es die Daten in Gruppen so aufzuteilen, dass Objekte innerhalb eines Clusters möglichst ähnlich sind, aber die Cluster zueinander möglichst unähnlich. Als Ausreißer werden Objekte definiert, die keinem Cluster zugeordnet werden können.
- **Klassifikation**: Aufgrund eines vorher definierten Samplings, werden neue Objekte automatisch klassifiziert und mit Attributwerten ausgestattet.

- **Assoziationsregeln**: Hierbei wird versucht, Zusammenhänge oder Logiken in Daten zu finden und herzuleiten, wie z.B. WENN A UND B DANN C.
- **Generalisierung**: Dabei sollen Daten und deren Attribute möglichst kompakt zusammengefasst und generalisiert werden. Ziel ist es die Anzahl der Datensätze zu reduzieren. (Ester & Sander, 2000, S. 4-5)

3.2.3 Entwicklung eines wissensbasierten Systems

Der zentrale Aspekt eines wissensbasierten Systems ist die Wissensbasis selbst. Entscheidend hierbei ist es, welche Schlussfolgerungen das System aus dem Wissen ziehen kann und wie es sich dabei verhält. Die Fähigkeit Schlussfolgerungen ziehen zu können, ist ein zentraler Aspekt des menschlichen Wesens und eröffnet die Möglichkeit, die Dinge des täglichen Lebens zu meistern. So schlussfolgert der Mensch zum Beispiel, dass wenn es heute regnet, die Straße nass ist oder das Kind, wenn es über starke Bauschmerzen klagt, vielleicht eine Blinddarmentzündung hat. (Brachman & Levesque, 2004, S. 7-10)

Dass dieses menschliche Schließen nicht immer einfach darzustellen ist, lässt sich mit Hilfe einer Formel veranschaulichen:

$$(W, B) \in R$$

Dabei steht *W* für Wissen und *B* für neues Wissen, welches man von *W* ableiten kann, um dadurch eine Schlussfolgerung *R* zu ziehen. Wenn es zum Beispiel draußen regnet (*W*), kann man daraus schließen, dass die Straße nass (*B*) ist. Also gilt die Schlussfolgerung (W_{Regen}, B_{nass}) ∈ *R*. Steht man allerdings an einer roten Ampel und es regnet (W_{Regen}, B_{Rot}) ∉ R, würde man sagen, dass es sich dabei nicht um eine logische Schlussfolgerung handelt, da die Ampel bei Regen nicht zwingend auf Rot geschaltet ist. In Abhängigkeit davon, ob *W* oder *B* gegeben ist, ergeben sich unterschiedliche Betrachtungsweisen:

1. Ist *W* gegeben, so kann mittels der Schlussfolgerung *R* das fehlende *B* prognostiziert werden.
2. Ist *B* gegeben, so wurde *B* aus *W* abgeleitet.
3. Ist sowohl *B* als auch *W* gegeben, liefert *R* einen Test, ob die Schlussfolgerung korrekt ist.

Um also menschliches Schließen nachbilden zu können, gilt es die Schlussfolgerung zu charakterisieren. (Bibel, Hölldobler, & Schaub, 1993, S. 112-114)

C. S. Pierce (Hartshorne & Weiss, 1931) unterscheidet bei der Charakterisierung von Schlussfolgerungen unter drei Arten, welche hier mit Beispielen angeführt werden:

1. **Deduktion**: Mit dem Wissen, dass zum Leuchten einer Taschenlampe, eine volle Batterie benötigt wird und, dass bei einer gegebenen Taschenlampe die Batterie leer ist, kann man daraus schließen, dass sich die Taschenlampe nicht einschalten lässt.
2. **Induktion**: Bei regelmäßiger Betrachtung von Taschenlampen mit leerer Batterie, die sich nicht einschalten lassen, kann man daraus schließen, dass Taschenlampen mit leerer Batterie nicht leuchten können.
3. **Abduktion**: Bei der dritten Art wird eine Erklärung für eine Beobachtung gesucht. Weiß man, dass eine Taschenlampe mit leerer Batterie nicht leuchtet und weiß man, dass sich eine gegebene Taschenlampe nicht einschalten lässt, schließt man daraus, dass die Batterie leer sein muss. (Hartshorne & Weiss, 1931)

Bei der Deduktion geht man immer davon aus, dass die Schlussfolgerungen korrekt sind. Genauso wie bei der Induktion. Anders verhält es sich bei der Abduktion, es wurde zwar eine Erklärung gefunden, hier geht man aber davon aus, dass das abgeleitete Wissen nicht unbedingt wahr ist, sondern vielleicht doch andere Ursachen haben könnte. Für wissensbasierte Systeme sind Schlussfolgerungen, die stets korrekt sein müssen, nicht praktikabel. (Beierle & Kern-Isberner, 2019, S. 24) Der Aufbau eines wissensbasierten Systems kann zuweilen recht komplex werden, deshalb empfehlen Beierle und Kern-Isberner (Beierle & Kern-Isberner, 2019, S. 19-20) die Entwicklung in acht Schritte aufzuteilen:

1. **Problembeschreibung**: Im ersten Schritt geht es darum, das Problem hinreichend zu beschreiben und auch die Funktionalität festzulegen.
2. **Wissensquellen**: Die Art und Quelle der Daten müssen festgelegt werden, dies können z.B. Datenbanken, Bücher oder menschliche Experten/Expertinnen sein.
3. **Design**: Wie sollen die Strukturen für die Wissensdarstellung aufgebaut sein und welche Benutzerschnittstellen werden benötigt.
4. **Entwicklungswerkzeug**: In Abhängigkeit zur bisherigen Erkenntnis muss entschieden werden, welche Entwicklungswerkzeuge sich am besten für die Aufgabe eignen.
5. **Entwicklung eines Prototyps:** Gerade für die Erstellung eines komplexen wissensbasierten Systems, ist die Erstellung eines Prototyps unumgänglich, hier wird auch entschieden, ob die ursprünglich geforderten Funktionalitäten hinreichend erfüllt werden.
6. **Testen des Prototyps:** In dieser Phase werden verschiedene Szenarien durchgespielt, um die Funktionen zu testen.

7. **Verfeinerung und Generalisierung**: Hier werden neue Aspekte miteingebunden oder neue Funktionen hinzugefügt, die sich erst im Laufe des Projektes ergeben haben.
8. **Wartung und Pflege**: Fehlerbehebung sowie Anpassungen und neue Funktionen. (Beierle & Kern-Isberner, 2019, S. 19-20)

3.2.4 Expertensystem

Expertensysteme definieren sich dadurch, dass das Wissen von Experten/Expertinnen stammt, darin liegt auch der wesentliche Unterschied zu einem wissensbasierten System. Ziel hierbei ist es mit Hilfe von Wissen, welches normalerweise nur erfahrenen oder gut ausgebildeten Personen zur Verfügung steht, dem Anwender/der Anwenderin zugänglich zu machen.
Als Experten/Expertinnen werden Personen bezeichnet, welche überdurchschnittliche Fähigkeiten haben, Probleme in einem speziellen Gebiet zu lösen. Sie verwenden oft Erfahrungen oder heuristisches Wissen, um spezielle Aufgaben zu lösen, haben gutes Allgemeinwissen und handeln oft intuitiv richtig. (Gottlob, Frühwirth, & Horn, 1990, S. 14)

Bei der Entwicklung eines Expertensystems sind drei Personengruppen beteiligt:

- **Benutzer/Benutzerinnen:** Arbeiten mit dem System, besitzen nur rudimentäres Wissen über das Fachgebiet.
- **Experten/Expertinnen:** Stellen das Fachwissen für die Software-Entwickler/Entwicklerinnen zur Verfügung.
- **Software-Entwickler/Entwicklerinnen:** Die Hauptaufgabe des Software-Entwicklers/der Software-Entwicklerin liegt darin, das Wissen in einem Programm darzustellen. Allerdings umfasst seine/ihre Tätigkeit noch viele andere Bereiche. Er/Sie muss die Anforderungen an das System mit dem Benutzer/der Benutzerin genau eruieren und abschätzen, ob diese mit einem Expertensystem überhaupt lösbar ist. Anschließend muss er/sie das Wissen vom Experten/der Expertin in eine Repräsentationsform übertragen, um aber die meist unstrukturierte Form des Wissens zu verstehen, muss sich der Entwickler/die Entwicklerin im Selbststudium ein Fachwissen über das Gebiet aneignen. Weiters muss er/sie sich Gedanken über das UI und UX Design machen. (Gottlob, Frühwirth, & Horn, 1990, S. 14-16)

3.2.5 Entscheidungsbaum

Entscheidungsbäume dienen als Klassifizierungsmethode für Datenobjekte und helfen bei Lösungen für Entscheidungsprobleme. (Larose, 2015, S. 318) Entscheidungsbäume bestehen aus einem Wurzelknoten, von dem aus beliebig viele Knoten zu mindestens jeweils zwei Entscheidungsblättern führen. Die Knoten am Ende werden Endknoten genannt und bezeichnen eine Entscheidungsklasse. Die Knoten generieren so lange untergeordnete Knoten bis entweder die Untergruppen sehr klein sind, oder durch weitere Aufteilung keine statistisch signifikanten Untergruppen mehr erzeugt werden können. Bei einem großen Baum kann es passieren, dass einige Ausreißer oder falsche Werte dabei sind, in diesem Fall sollte der Baum so beschnitten werden, dass die Genauigkeitsrate dadurch nicht beeinflusst wird. (Alharan, Al-Haboobi, & Alsagheer, 2017, S. 135)

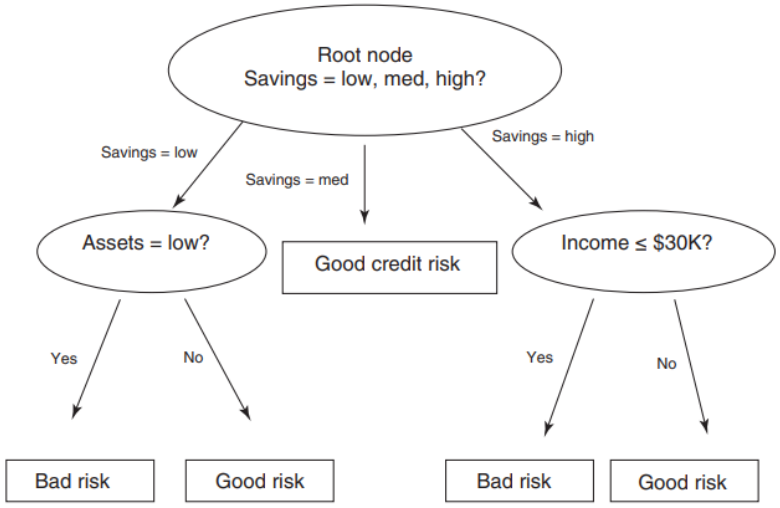


Abbildung 9. Einfacher Entscheidungsbaum

Quelle: (Larose, 2015, S.318)

Entscheidungsbäume werden oft als Entscheidungshilfe bei Banken für die Vergabe von Krediten verwendet. Hierbei wird der Baum vom Analysten/der Analystin anhand von festgelegten Variablen definiert. Beantragt ein Kunde/eine Kundin dann einen Kredit, wird er/sie gezielt nach den Entscheidungsparametern gefragt, beispielsweise das bereits gesparte Vermögen und das jährliche Einkommen. Anhand dieser Informationen wird mit Hilfe des Entscheidungsbaums über das Kreditrisiko entschieden. Um einen Entscheidungsbaum zu erstellen, sollte folgendes beachtet werden (Larose, 2015, S. 318):

1. Die Erstellung eines Entscheidungsbaumes ist ein manueller Prozess, deshalb müssen davor bereits Zielvariablen aus den Datensets berechnet werden.
2. Die Datensets sollten reichhaltig und vielfältig sein, um einen vollständigen Entscheidungsbaum abbilden zu können. Das Fehlen von Informationen kann für die Klassifizierung von Daten problematisch oder unmöglich sein.
3. Die Eingangsparameter müssen klar definiert sein. Das heißt, hat man einen Knoten abgeschlossen, kann nicht aufgrund von neuen Erkenntnissen zu einem vorherigen Punkt zurückgekehrt werden. (Larose, 2015, S. 318)

3.2.6 Naive Bayes

Die Bayes Klassifikation, benannt nach dem englischen Mathematiker Thomas Bayes, ist eine statistische Methode, die davon ausgeht, dass sämtliche Attribute voneinander unabhängig sind. Für die Berechnung wird die Bayes Formel verwendet, wobei *p* für die Wahrscheinlichkeit steht, *B* für das Zielattribut und *A* für ein Ereignis. (Provost & Fawcett, 2013, S. 237)

$$p(B \mid A) = \frac{p(A \mid B) \cdot p(B)}{p(A)}$$

Um herauszufinden, welches Zielattribut am wahrscheinlichsten ist, werden für ein Ereignis mehrere Berechnungen durchgeführt. (Provost & Fawcett, 2013, S. 238)
Das Ganze lässt sich an einem Beispiel abbilden. Bei der Beantwortung einer Frage in einem Multiple Choice Test, kennt ein Student/eine Studentin entweder die Antwort, oder der Student/die Studentin kann die Frage nur durch erraten beantworten. Angenommen, die Wahrscheinlichkeit, dass der Student/die Studentin die Antwort kennt, beträgt 75% und die Wahrscheinlichkeit, dass der Student/die Studentin die Antwort nicht kennt, beträgt 25%. Geht man nun davon aus, dass es für jede Multiple Choice Frage fünf Auswahlmöglichkeiten gibt und der Student/die Studentin, die Antwort nur erraten kann, ergibt sich daraus eine 1/5 = 20% Wahrscheinlichkeit das er/sie richtig liegt. Wie hoch ist die bedingte Wahrscheinlichkeit, dass der Student/die Studentin die Antwort auf eine Frage wusste, wenn der Student/die Studentin sie richtig beantwortet hat? (Sheldon, 2010, S. 67)

Abbildung 10. Baumdiagramm der prozentuellen Wahrscheinlichkeiten

Quelle: Eigene Darstellung in Anlehnung an (Sheldon, 2010, S.67)

Definiert man nun *A* als die Variable, dass der Student/die Studentin die Frage richtig beantwortet und *B* als die Wahrscheinlichkeit, dass der Student/die Studentin die Antwort kennt, resultiert daraus das Ergebnis *p*. (Sheldon, 2010, S. 67)

$$= \frac{0{,}75}{0{,}75 * 1 + 0{,}25 * 0{,}20}$$

$$p = 0{,}9375$$

Abbildung 11. Berechnung der Wahrscheinlichkeit nach Bayes

Quelle: Eigene Darstellung in Anlehnung an (Sheldon, 2010, S.67)

3.2.7 Betrachtungszeitraum

Der Betrachtungszeitraum von Daten kann für Vorhersagemodelle einen entscheidenden Faktor darstellen. Dies hängt im Wesentlichen vom Ziel des Modells ab. Um beispielsweise die Erderwärmung vorherzusagen, macht ein größerer Betrachtungszeitraum durchaus Sinn, während die Siegchancen einer Fußballmannschaft eher von der Leistung in der näheren Vergangenheit abhängen. (Mishra & Silakari, 2012, S. 24)

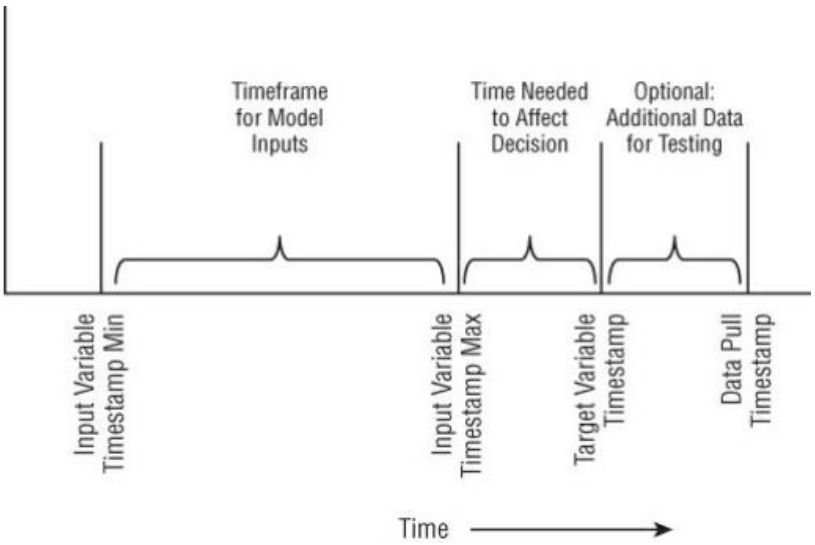


Abbildung 12. Festlegen der Betrachtungszeitraumvariablen

Quelle: (Abbot., 2014, S.26)

Um also den Datumsbereich festzulegen, müssen zwei Variablen definiert werden. Diese bestehen aus dem Start- und dem Enddatum. Anschließend wird definiert, in welchem Zeitraum die Prognose gültig ist. Optional kann zusätzlich noch ein Testbereich festgelegt werden, um die Ergebnisse zu validieren. (Abbott, 2014, S. 25-26)

3.2.8 Assoziationsanalyse

Die Assoziationsanalyse ist eines der wenigen Verfahren im Data-Mining, mit dem sich Zusammenhänge und Abhängigkeiten in Daten entdecken lassen. Assoziationsanalysen kommen häufig bei Handelsunternehmen oder bei Versicherungsgesellschaften zum Einsatz, um beispielsweise Beziehungen in Produkten darzustellen oder zur Wahrscheinlichkeitsrechnung von Schadensfällen. Um die Assoziationsanalyse besser zu verstehen, müssen zunächst ein paar Begriffe definiert werden:

Item Menge I = $\{i_1,\ldots,i_m\}$ (beispielsweise Artikel eines Lebensmittelmarktes)

Transaktion T = (Warenzusammenstellung eines Kunden/einer Kundin)

Datenbasis D = $(T_1,\ldots,T_n)$ (Zusammenfassung aller Kaufaktionen in einem Zeitraum)

(Bankhofer & Vogel, 2008, S. 261)

„Eine Assoziationsregel stellt dann eine Regel der Form „wenn Item(menge) X, dann Item(menge) Y“ ($X \rightarrow Y$) dar, wobei die Menge X im Regelrumpf und die Menge Y im Regelkopf disjunkt und echte Teilmengen der Item Menge I sind. Eine Transaktion T erfüllt eine Regel $X \rightarrow Y$ genau dann, wenn $(X \cup Y) \subset T$, das heißt, wenn alle Items, die durch diese Regel abgebildet werden, auch in der Transaktion vorkommen.“ Damit eine Regel bewertet werden

kann, benötigt man zwei weitere Faktoren in Form des Supports sowie der Confidence. Der Support gibt die Häufigkeit der Item Menge in der Datenbasis an. Der Confidence-Faktor die Wahrscheinlichkeit über die Stärke des Zusammenhangs. (Bankhofer & Vogel, 2008, S. 262-263)

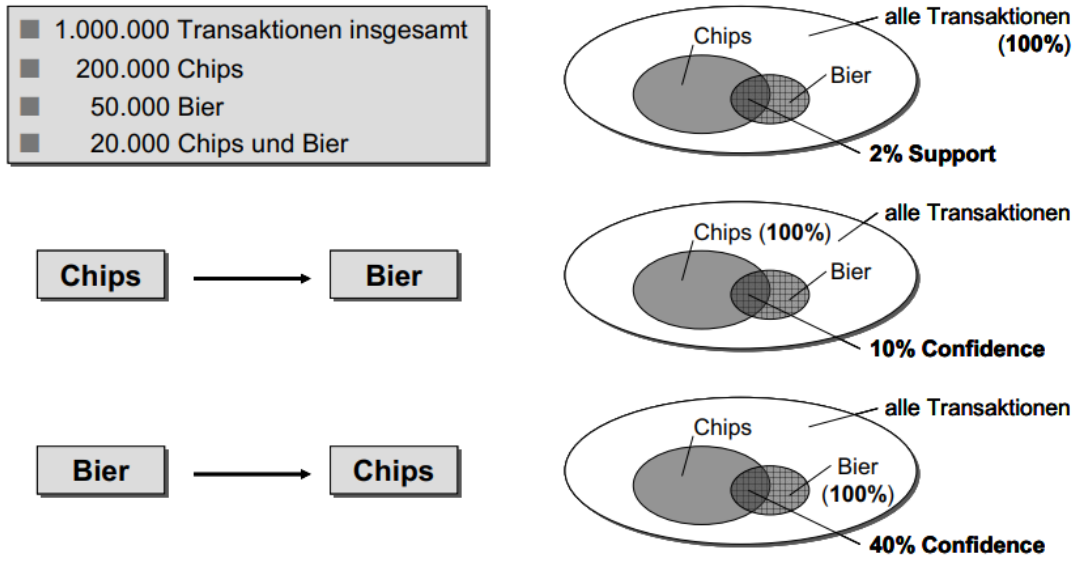


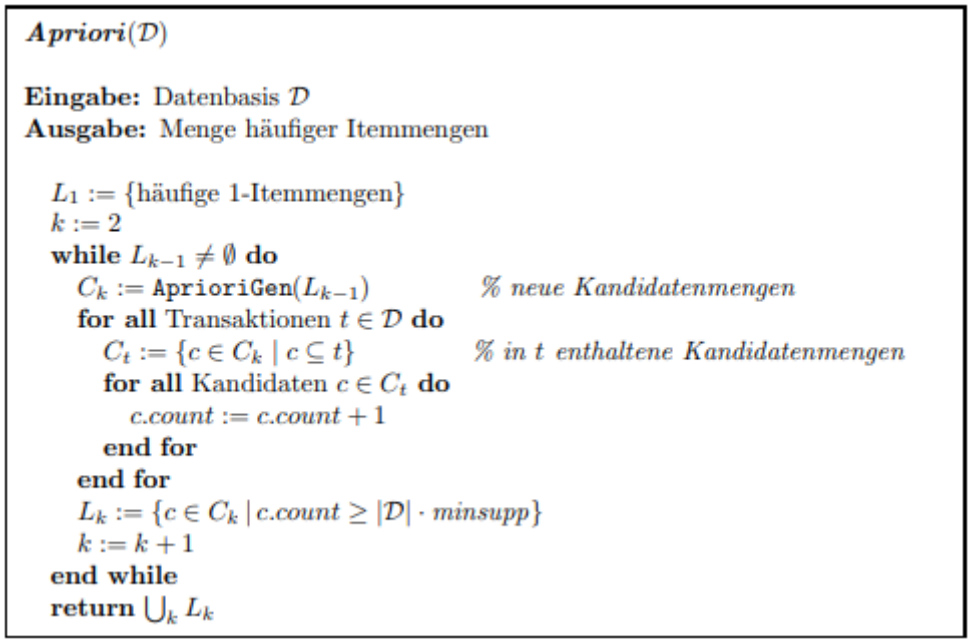
Abbildung 13. Support und Confidence einer Assoziationsregel

Quelle: (Bankhofer & Vogel, 2008, S.263)

Wie in Abbildung 13 dargestellt, könnte sich folgende Regel ableiten lassen: Eine gemeinsame Anordnung von Bier und Chips könnte möglicherweise zu einer Zunahme von Verbundverkäufen der beiden Produkte führen. Gleichzeitig sollten nicht beide Produkte ermäßigt sein, wenn dann überhaupt nur Bier, was zu einer Zunahme der Chips Verkäufe führen könnte.
(Bankhofer & Vogel, 2008, S. 263)

3.2.8.1 Apriori-Algorithmus

Eine weit verbreitete Methode Warenkörbe zu analysieren, ist der Apriori-Algorithmus. Dieser ermöglicht es Zusammenhänge innerhalb der Transaktionen zu assoziieren, um daraus Regeln abzuleiten. (Ester & Sander, 2000, S. 159)



Apriori$(\mathcal{D})$

Eingabe: Datenbasis $\mathcal{D}$
Ausgabe: Menge häufiger Itemmengen

$L_1 := \{$häufige 1-Itemmengen$\}$
$k := 2$
while $L_{k-1} \neq \emptyset$ **do**
  $C_k :=$ `AprioriGen`(L_{k-1}) *% neue Kandidatenmengen*
  for all Transaktionen $t \in \mathcal{D}$ **do**
    $C_t := \{c \in C_k \mid c \subseteq t\}$ *% in t enthaltene Kandidatenmengen*
    for all Kandidaten $c \in C_t$ **do**
      $c.count := c.count + 1$
    end for
  end for
  $L_k := \{c \in C_k \mid c.count \geq |\mathcal{D}| \cdot minsupp\}$
  $k := k + 1$
end while
return $\bigcup_k L_k$

Abbildung 14. Der Apriori-Algorithmus

Quelle: (Beierle & Kern-Isberner, 2019, S.152)

Die Grundidee zum Auffinden von Artikelmengen basiert auf folgender Eigenschaft:
„Jede Teilmenge einer häufig auftretenden Artikelmenge muss selbst auch häufig sein."
Um diese Eigenschaft zu nutzen, werden die Artikelmengen der Größe nach angeordnet. Bestimmt werden die frequenten Artikelmengen durch Zählen in der Datenbank. Angefangen wird mit einelementigen Artikelmengen, dann zweielementigen und so weiter. Durch das Zählen müssen im nächsten Schritt nicht mehr alle möglichen zweielementigen Artikelmengen durchgezählt werden, sondern nur noch solche, von denen man weiß, dass sie öfters vorkommen. In jedem weiteren *k* Durchgang werden dann Kandidaten für die *k*-elementigen häufigen Artikelmengen L_k generiert, mit den schon berechneten (*k - 1*) -elementigen häufigen Artikelmengen aus L_{k-1}. Der Algorithmus iteriert dann solange bis keine frequenten Artikelmengen mehr gefunden werden. Anschließend wird der Support der Kandidaten in einer kompletten Iteration durch $\mathcal{D}$ (Datenbasis) gezählt. Die Kandidaten, die den minimalen Support *(minsupp)* erreichen, werden in L_k aufgenommen. (Ester & Sander, 2000, S. 160-162)

3.2.8.1.1 Generierung der Kandidaten

AprioriGen beschreibt die Bildung der Kandidaten im *k*-Schritt. Die Prozedur aus L_{k-1} erzeugt dabei eine Obermenge von L_k. Um diesen Vorgang besser veranschaulichen zu können, kann man folgendes Beispiel nehmen (Ester & Sander, 2000, S. 162-163):

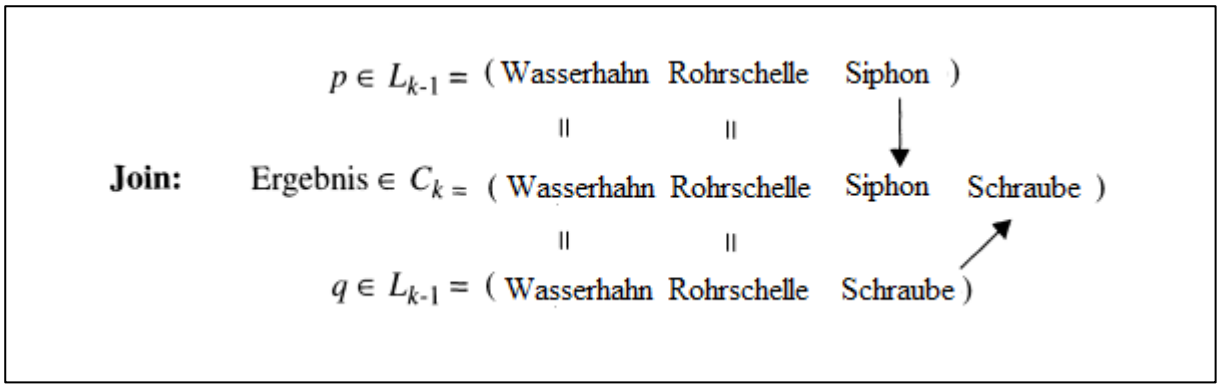


Abbildung 15. Kandidatengenerierung *AprioriGen*

Quelle: Eigene Darstellung in Anlehnung an (Ester & Sander, 2000, S.163)

Es sind die häufigsten Artikelmengen L_3 gegeben:
{(Wasserhahn Rohrschelle Siphon),
(Wasserhahn Rohrschelle Schraube),
(Wasserhahn Siphon Schraube),
(Wasserhahn Siphon Seifenspender),
(Rohrschelle Siphon Schraube)}
Nach dem Verbinden *(join)* enthält die Kandidatenmenge C_4 die beiden Elemente:
{(Wasserhahn Rohrschelle Siphon Schraube),
(Wasserhahn Siphon Schraube Seifenspender)}.
Da die Teilmenge
(Wasserhahn Schraube Seifenspender)
nicht in L_3 enthalten ist, wird *(Wasserhahn Siphon Schraube Seifenspender)* wieder entfernt.
Somit bleibt nur *(Wasserhahn Rohrschelle Siphon Schraube)* in der Kandidatenmenge erhalten.
(Ester & Sander, 2000, S. 162-163)

3.2.9 Clustering

Clustering ist ein weiteres wichtiges Data-Mining Verfahren und kommt häufig bei der Bildverarbeitung zur Szenenvervollständigung zum Einsatz. Das Ziel von Clustering ist es, Daten so zu kategorisieren, dass Objekte innerhalb des Clusters möglichst ähnlich zueinander sind und Objekte aus anderen Clustern möglichst unähnlich. Um das Clustering Verfahren verwenden zu können, müssen die Daten bestimmte ähnliche Merkmale aufweisen. Außerdem muss beachtet werden, dass die Daten innerhalb des Clusters unterschiedliche Formen, Größen und Dichten haben können. (Ester & Sander, 2000, S. 45)

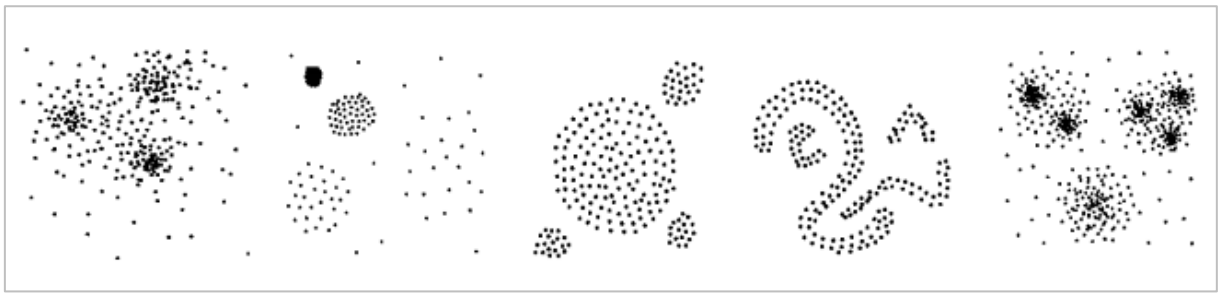


Abbildung 16. 2D Clusterstrukturen mit unterschiedlicher Charakteristik

Quelle: (Ester & Sander, 2000, S.45)

Wie in Abbildung 16 ersichtlich, werden die Ähnlichkeiten der Objekte durch die Abstände zwischen den Punkten dargestellt. Je kleiner die Distanz, desto ähnlicher sind sich die Objekte. (Ester & Sander, 2000, S. 46)

3.2.9.1 Hierarchische Cluster

Beim hierarchischen Cluster handelt es sich um ein hierarchisches System, bei dem durch das sukzessive Gruppieren von Objekten und im weiteren Verlauf durch Fusion von Gruppen, größere Gruppen konstruiert werden, bis schlussendlich nur noch ein Cluster übrig ist. Für einen hierarchischen Cluster benötigt man eine Distanzfunktion die auch „linkage“ genannt wird. (Provost & Fawcett, 2013, S. 166)

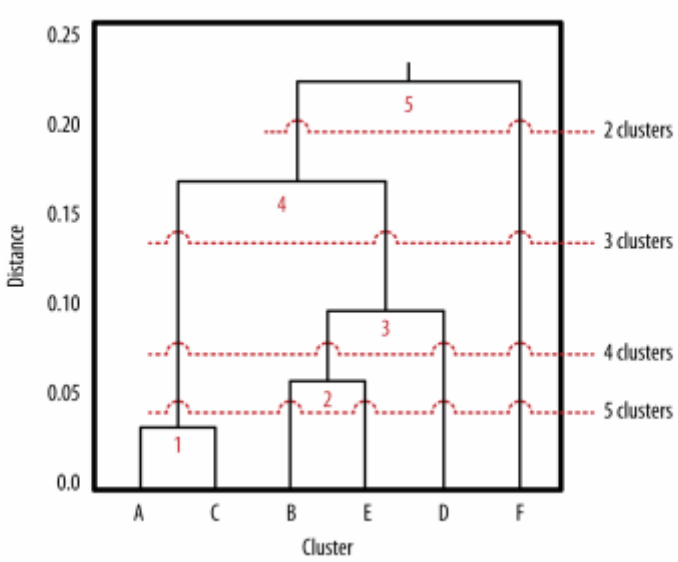


Abbildung 17. Aufbau eines hierarchischen Clusters

Quelle: (Provost & Fawcett, 2013, S.165)

Für die Berechnung der Distanz gibt es je nach Datentyp und Einsatzzweck unterschiedliche Formeln für die Berechnung. (Ester & Sander, 2000, S. 47)

Datensätze mit numerischen Werten:

- **Euklidische Distanz:** Ermittelt die räumliche Distanz zweier Punkte beziehungsweise Attribute. Dies kann entweder ein mathematischer oder geografischer Raum sein.

$$dist_E(v,w) = \sqrt{\sum_i (v_i - w_i)^2}$$

- **Manhattan-Distanz**: Dient ebenfalls der Ermittlung der räumlichen Distanz, allerdings wird hier, wie bei einem Schachbrett, jeder Schritt auf der Achse gezählt.

$$dist_{Man}(v,w) = \sum_i |v_i - w_i|$$

- **Maximum-Metrik**: Ermittelt den Abstand zwischen allen Attributen und wählt anschließend den größten aus. Somit werden Vergleiche immer nur zwischen zwei Attributen durchgeführt.

$$dist_{Max}(v,w) = \max_i(|v_i - w_i|)$$

Datensätze mit kategorischen Attributwerten:

$$dist(x, y) = \sum_{i=1}^{d} \delta(x_i, y_i),$$

Endliche Mengen:

$$dist(x, y) = \frac{|x \cup y| - |x \cap y|}{|x \cup y|}$$

Textdokumente:

$$dist(D_1, D_2) = 1 - \frac{\langle g(c(D_1)), g(c(D_2))\rangle}{\|g(c(D_1))\| \cdot \|g(c(D_2))\|}$$

Eine andere Methode ohne die Verwendung einer Funktionsgleichung, ist die Erstellung einer Distanzmatrix. Dabei werden die Distanzen zwischen den Objekten einfach paarweise abgespeichert. (Ester & Sander, 2000, S. 47)

3.2.9.2 k-Means Cluster

Der *k*-Means Algorithmus ist eine weitere Methode, die sich für die Gruppierung von Objekten in einem Cluster einsetzen lässt. Aufgrund des geringen Speicherbedarfs und der effizienten Berechnung der Clusterzentren, eignet sich der Algorithmus gut, für die Analyse großer Datenmengen. Der Algorithmus geht dabei in einer bestimmten Reihenfolge vor:

1. Als erstes muss der Benutzer/die Benutzerin definieren, wie viele Cluster *k* er/sie erstellen will.
2. Als nächstes werden dem *k* zufällige Datenpunkte als Abstand zu den Zentren zugeordnet.
3. Anschließend wird für jeden Datensatz das nächstgelegene Clusterzentrum gesucht.
4. Abschließend wird der Punkt Zwei solange wiederholt, bis sich die Lage der Zentren nicht mehr ändert. (Larose, 2015, S. 529)

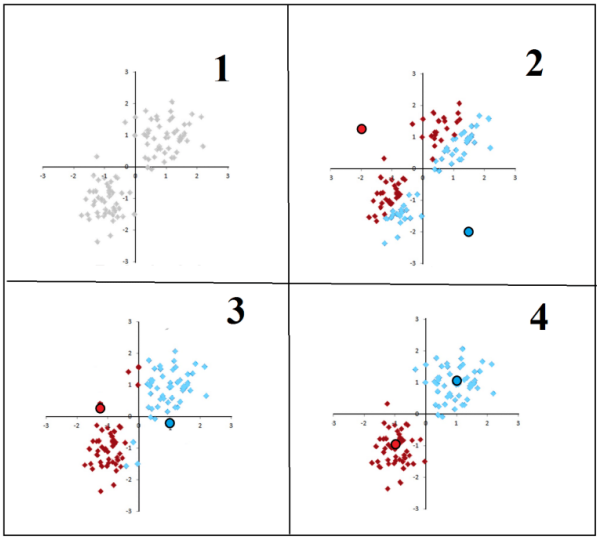


Abbildung 18. Schritte des *k-Means* Algorithmus

Quelle: https://www.researchgate.net/figure/Steps-of-the-K-mean-clustering-algorithm_fig5_321051036 abgerufen am 04.04.2020

3.2.10 Regressionsanalyse

Regression ist eine Data-Mining-Technik, mit der ein Bereich mit numerischen Werten (kontinuierliche Werte) für einen bestimmten Datensatz aus einer Reihe unabhängiger Eingabevariablen vorhergesagt werden kann. Beispielsweise kann die Regression verwendet werden, um die Kosten eines Produkts oder einer Dienstleistung unter Berücksichtigung anderer Variablen vorherzusagen. Die letzten Jahre haben sich Regressionsanalysen als mächtiges Werkzeug für Umwelt- und Trendanalysen erwiesen. (Yang, Liu, Tsoka, & Papageorgiou, 2016, S. 156-157)

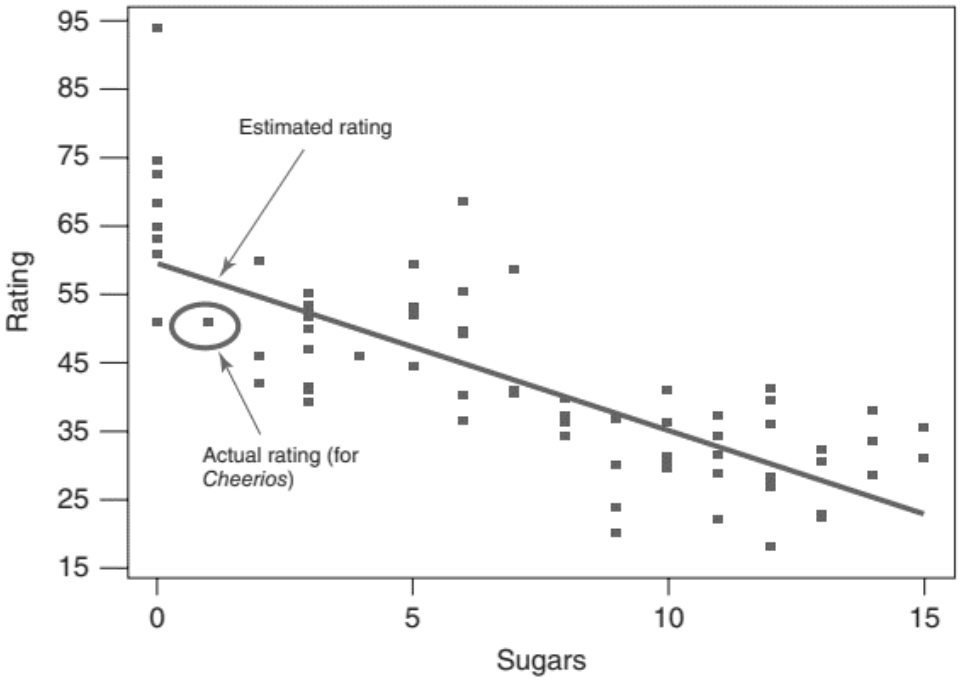


Abbildung 19. Nährwertbewertung im Vergleich zum Zuckergehalt von 77 Zerealien

Quelle: (Larose, 2015, S.173)

In der Literatur existieren eine große Anzahl an Regressionsanalysemethoden, die wichtigsten sind:

- **Lineare Regression**: Eines der am häufigsten verwendeten Verfahren bei der die Ausgabevariable als lineare Kombination der Eingabevariable vorhergesagt werden kann.
- **Support Vector Machine**: Ist ein sehr etablierter Lernalgorithmus, der auf mathematischen Verfahren aus der Mustererkennung basiert. Die grundlegende Idee der SVM besteht darin, die Menge von Objekten durch eine Hyperebene in zwei Klassen zu unterteilen und zu kategorisieren.
- **Kriging:** Ist ein statistisches Verfahren zur Interpolation. D.h. man kann Werte an Orten, für die keine Daten vorliegen, durch umliegende Messwerte interpolieren.
- **Random Forest:** Bestehen aus einer großen Anzahl einzelner Entscheidungsbäume, die als Gruppe fungieren. Jeder Entscheidungsbaum darf eine Klassenvorhersage treffen und die Klasse mit den meisten Stimmen entscheidet über die Vorhersage des Modells.
- **KNN:** Der KNN-Algorithmus geht davon aus, dass ähnliche Dinge in unmittelbarer Nähe existieren. Der Zweck besteht darin, in einer Datenbank die Klassifizierung eines neuen Stichprobenpunkts vorherzusagen.

- **MARS**: Ist eine andere Variante der Regressionsanalyse und kann als Erweiterung linearer Modelle angesehen werden, die automatisch Nichtlinearität und Wechselwirkung zwischen Variablen modelliert.
- **MLP**: Beschreibt ein künstliches neuronales Netzwerk. Dabei wird untersucht, wie Modelle des biologischen Gehirns verwendet werden können, um schwierige Rechenaufgaben, wie es beim maschinellen Lernen der Fall ist, zu lösen. (Yang, Liu, Tsoka, & Papageorgiou, 2016, S. 157-158)

3.2.11 Zeitreihenanalyse

Zeitreihenanalysen untersuchen die Entwicklung von Werten im zeitlichen Verlauf. Zu den Aufgaben der Zeitreihenanalyse gehört neben der Analyse von historischen Daten auch die Prognose über Ereignisse auf Grundlage der bisherigen Werte. Hierbei werden die Zeitreihen auf häufig wiederkehrende Verlaufsmuster untersucht, die auf bestimmte gleichwirkende Einflussfaktoren zurückzuführen sind. Diese Einflussfaktoren sind das Ergebnis von mehreren Komponenten, die wichtigsten sind:

- **Trendkomponente**: Beschreibt die konstante Entwicklungsrichtung der Reihe. Diese kann sowohl steigend als auch fallend sein. Beispielsweise die Entwicklung des Aktienmarktes während einer Wirtschaftskrise.
- **Saisonkomponente**: Berücksichtigt saisonale Schwankungen, die innerhalb eines Jahres auftreten. So steigt beispielsweise der Flugverkehr in den Urlaubsmonaten.
- **Konjunkturkomponente**: Berücksichtigt zyklische Schwankungen mit einer Periodenlänge von über einem Jahr. Ursache dafür kann eine schwache Wirtschaft mit der folgenden höheren Arbeitslosigkeit sein.
- **Restkomponente**: Umfasst alle Einflüsse, die nur einmalig auftreten und nicht durch die vorher genannten Komponenten erfasst werden. Dies können zum Beispiel Fehler bei der Datenerhebung oder Wetterumschwünge sein. (Holland & Scharnbacher, 2010, S. 79-81)

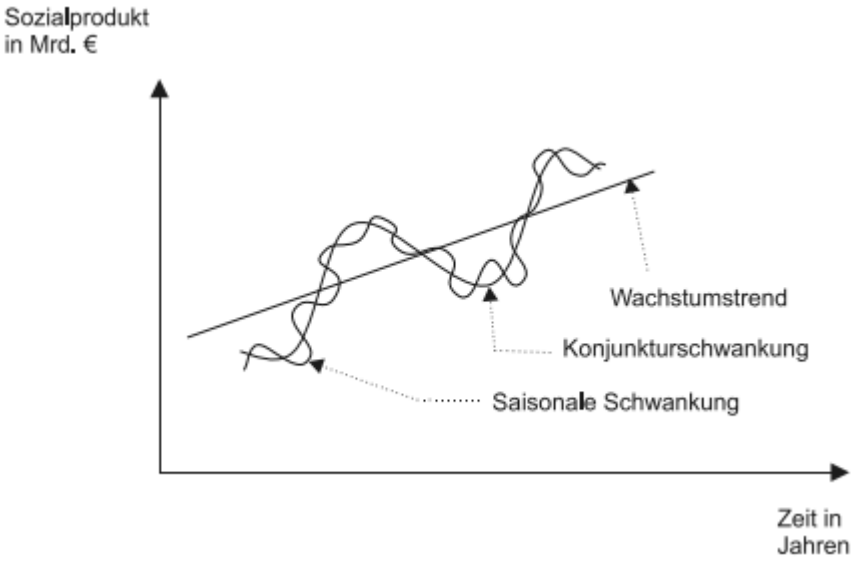


Abbildung 20. Beispiel für eine Zeitreihe

Quelle: (Holland & Scharnbacher, 2010, S.80)

3.2.12 Maschinelles Lernen

Maschinelles Lernen ist ein Teilgebiet der Künstlichen Intelligenz (KI). Der Begriff lässt sich nur schwer exakt definieren, es gibt viele unterschiedliche Ansätze zum maschinellen Lernen. In vereinfachter Form betrachtet handelt es sich um ein System, welches Dinge wahrnehmen, verstehen, handeln und lernen kann. (Brühl, 2019, S. 5)

Die Hauptaufgabe von maschinellem Lernen ist die korrekte Vorhersage von Daten basierend auf erlernten Regeln. Die Regeln lernt es aus Mustern, welche es aus Daten generalisiert. Wie in Abbildung 21 ersichtlich, existieren unterschiedliche Typen von maschinellem Lernen. (Vieira, Hugo, Pinaya, & Andrea, 2020, S. 9)

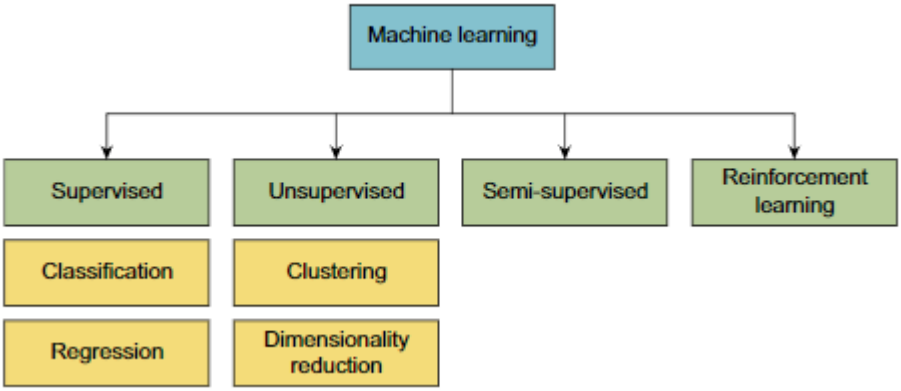


Abbildung 21. Typen des maschinellen Lernens

Quelle: (Vieira, Hugo, Pinaya, & Andrea, 2020, S. 9)

- **Supervised Learning**: Beim überwachten Lernen kennt der Algorithmus die Eingangsparameter sowie das zu erwartende Ergebnis. Ziel hierbei ist es, aus Trainingsdaten Modelle zu erstellen. Wenn diese erstellt sind, können dem Modell unbekannte Daten geliefert werden und das System berechnet das Ergebnis. Diese Methode wird häufig bei der Erkennung von Objekten in Bildern angewandt.
- **Unsupervised Learning:** Beim unüberwachten Lernen hingegen, weiß der Algorithmus nicht wonach er suchen soll. Er erkennt lediglich ähnliche Muster und teilt diese in einem Cluster oder in Kategorien auf. Werden beispielsweise Fotos analysiert, so kann der Algorithmus zwischen Menschen und Tieren unterscheiden, ohne diese jedoch zu benennen.
- **Semi-supervised Learning**: Beim halbüberwachten Lernen ist nur ein Teil der Bezeichnungen für die Zielvariablen vorhanden. Diese Methode behebt dieses Problem, indem es dem Modell ermöglicht, die verfügbaren unbeschrifteten Daten in das überwachte Lernen zu integrieren.
- **Reinforcement Learning:** Beim verstärkten Lernen geht es darum, dass das System anhand der Interaktion mit der Umgebung lernt, um die erhaltenen Belohnungen zu maximieren. Dabei weiß das System nicht welche Aktion für eine bestimmte Situation die beste ist, sondern es lernt diese anhand einer Belohnung, die positiv oder negativ sein kann. (Vieira, Hugo, Pinaya, & Andrea, 2020, S. 9-12)

Ein Beispiel für maschinelles Lernen ist das sogenannte „Black Box Vorhersagemodell“. Es entstand aufgrund der zunehmenden Komplexität von Geschäftsprozessen und dem Neigen der Entscheidungsträger/Entscheidungsträgerinnen, häufig Entscheidungen auf Basis subjektiver Erfahrungen zu treffen. Forschungen haben jedoch gezeigt, dass Unternehmen bessere Leistungen erzielen, wenn die Basis für ihre Entscheidungen datengesteuert ist. Dies soll nicht heißen, dass komplett auf menschliche Entscheidungen verzichtet werden soll, vielmehr sollen Entscheidungsträger/Entscheidungsträgerinnen durch solche Systeme unterstützt werden. Hierbei werden mit Hilfe von maschinellem Lernen einheitliche Erklärungen generiert, welche die Entscheidungsträger/Entscheidungsträgerinnen mit Was-Wäre-Wenn-Analysen unterstützen. (Bohanec, Borštnar, & Šikonja, 2017, S. 416-420)

3.2.12.1 Künstliche neuronale Netze

Künstliche neuronale Netze (engl. *artificial neural networks)* werden als Spezialbereich des maschinellen Lernens angesehen. Diese können sowohl im Bereich des überwachten als auch im unüberwachten Lernen eingesetzt werden. (Brühl, 2019, S. 6) KNN sind informationsverarbeitende Systeme, deren Struktur und Funktionsweise dem Gehirn von Menschen und Tieren nachempfunden ist. Sie bestehen im Wesentlichen aus einer großen Anzahl parallel arbeitender Einheiten die auch Neuronen genannt werden. Diese Neuronen, wie in Abbildung 22 dargestellt, senden sich Informationen als Aktivierungssignale über gerichtete Verbindungen zu. (Kruse, et al., 2015, S. 7)

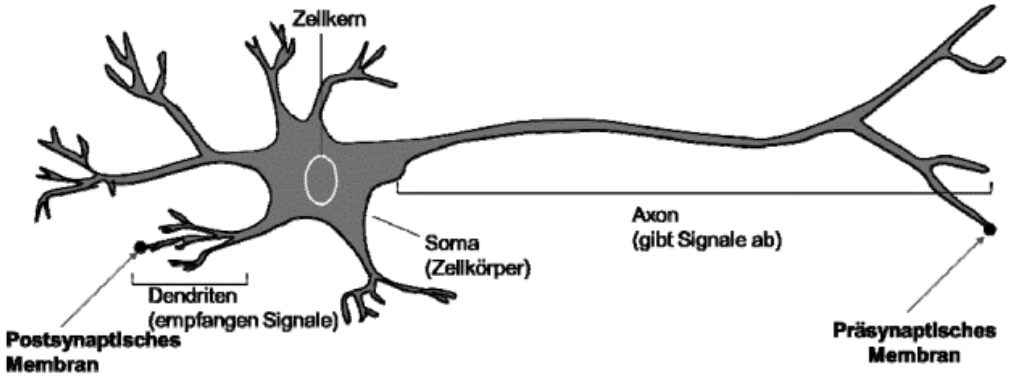


Abbildung 22. Aufbau von Nervenzellen (Neuronen)

Quelle: (Brühl, 2019, S. 7)

Das menschliche Gehirn besteht aus ca. 100 Milliarden Neuronen, welche zuständig sind für die Aufnahme, Speicherung, Verarbeitung und Weitergabe von Informationen. Die Dendriten empfangen dabei die Informationen als elektrische Signale und leiten diese an den Zellkörper (Soma) weiter. Sobald die Summe der über die Dendriten empfangenen Eingangssignale einen bestimmten Schwellenwert überschreitet, leitet das Axon die elektrischen Impulse an die Synapsen weiter. Die Synapsen dienen als Neurotransmitter und stellen den Kontakt und die Informationsübertragung zu anderen Neuronen her. Je nach Komplexität und Aufgabe werden so Milliarden von Neuronen untereinander verknüpft und simultan genutzt, um beispielsweise Muster zu erkennen. (Brühl, 2019, S. 6) Heutzutage wird KNN dazu verwendet, bestimmte kognitive Fähigkeiten des Menschen nachzubilden. Dies kann beispielsweise nützlich sein, um Anomalien auf medizinischen Bildern zu entdecken, Prognosen am Finanzmarkt zu erstellen oder bei der Gesichtserkennung. (Kruse, et al., 2015, S. 7-8)

3.3 Beantwortung der theoretischen Subforschungsfragen

Welche Methoden und Formeln eignen sich für die Prognosenerstellung?

Ereignisse vorherzusagen liegt in unserer menschlichen Natur und so haben sich im Laufe der Geschichte verschiedene Techniken und Methoden etabliert, um uns dabei zu unterstützen. (Mishra & Silakari, 2012, S. 4434) Der erste Schritt bei einer Prognosenerstellung ist dabei die effektive Gestaltung des Projektablaufs. Hierbei haben sich zwei Methoden als praktikabel erwiesen. Zum einen die KDD-Methodik und zum anderen das CRISP-DM als Standard-Prozessmodell für Data-Mining. Beide beschreiben einen Ablauf, um Wissen aus Rohdaten zu generieren, unterscheiden sich aber im Kern voneinander. Bei der KDD-Methodik handelt es sich um einen nicht trivialen Prozess, dessen Aufgabe es ist, Muster aus Datensätzen zu extrahieren und diesen Eigenschaften zuzuordnen. Diese sollen dann für einen Großteil des Datensatzes gültig sein, um leicht verständliche Zusammenhänge zu beschreiben. Beim CRISP-DM handelt es sich um einen Branchen- und Industriestandard. Ausgangspunkt ist hier eine betriebswirtschaftliche Problemstellung wobei der zyklische Charakter von Projekten in den Vordergrund gestellt wird. Welche der beiden Methoden schlussendlich gewählt wird, hängt im Wesentlichen von der Aufgabenstellung und dem Ziel der Prognose ab. (Fayyad, Piatetsky-Shapiro, & Padhraic , 1996, S. 39-41; Göpfert & Breiter, 2015, S. 1220)

Damit man etwas prognostizieren kann, bedarf es auch immer einer Wissensbasis. Die Art und Quelle der Daten müssen festgelegt werden, dies können z.B. Datenbanken, Bücher oder menschliche Experten/Expertinnen sein. Weiters muss in Abhängigkeit zum Projektziel entschieden werden, welche Entwicklungswerkzeuge sich am besten für die Aufgabe eignen. Dabei geht es nicht nur um die Software, sondern auch um die zu verwendenden Analysemethoden. (Beierle & Kern-Isberner, 2019, S. 19-20) Eine der aktuelleren Methoden ist dabei Predictive Analytics, welche sich mit der Vorhersage von Ereignissen basierend auf zuvor beobachteten historischen oder aktuellen Daten, unter Zuhilfenahme von modernen Algorithmen befasst. (Mishra & Silakari, 2012, S. 4434) Predictive Analytics entwickelte sich aus mehreren Disziplinen, diese umfassen Mustererkennung, Statistik, maschinelles Lernen, Künstliche Intelligenz und Data-Mining. (Brühl, 2019, S. 1-4) Der Begriff Data-Mining wird schon seit vielen Jahren verwendet, während Predictive Analytics erst seit einigen Jahren zur Anwendung kommt. Data-Mining ist häufig der erste Prozessschritt in der Datenanalyse, der einen mehrstufigen Prozess zur Wissensgenerierung umfasst. (Iffert, 2016, S. 17-18)

Ein großer Teilbereich von Data-Mining bildet dabei die Statistik. Diese wird bei Analysen häufig verwendet, um wichtige Kennzahlen aus Daten zu erheben, häufig auch, wenn sich

Kennzahlen nicht verallgemeinern lassen. Statistik sollte generell immer im Einklang mit dem Business Problem stehen, hilft aber auch beim Erstellen von Hypothesen oder zu Überprüfung, ob ein beobachtetes Muster gültig ist. (Provost & Fawcett, 2013, S. 35-36) Statistik unterstützt ebenfalls bei der Entscheidungs- und Wahrscheinlichkeitstheorie, wenn es darum geht, abzuschätzen, wie hoch die Chancen auf das Eintreffen eines Ereignisses sind. (Sheldon, 2010, S. 67) Einige Algorithmen wurden dabei speziell für den Handel entworfen, so auch der Apriori-Algorithmus, welcher zur Kategorie der Assoziationsanalysen gehört. Wendet man diesen mittels einer statistischen Software auf einen Warenkorb an, so lassen sich Beziehungen in Produkten errechnen, sowie deren Wahrscheinlichkeit mit anderen Produkten gekauft zu werden. Zusätzlich lässt sich mit dem Support Faktor die Häufigkeit der beiden Produkte in der Datenbasis anzeigen. (Bankhofer & Vogel, 2008, S. 261-263) Eine weitere Möglichkeit Muster in Daten zu entdecken, bildet dabei die Zeitreihenanalyse. Hierbei werden die Zeitreihen auf häufig wiederkehrende Verlaufsmuster untersucht, die auf bestimmte gleichwirkende Einflussfaktoren zurückzuführen sind. (Holland & Scharnbacher, 2010, S. 79-81) Nicht immer wird für eine Vorhersage menschliche Intelligenz benötigt, beim maschinellen Lernen reicht es aus die Regeln vorzugeben, damit das System eigenständig das Ergebnis berechnet. (Vieira, Hugo, Pinaya, & Andrea, 2020, S. 9) Zusammenfassend kann also gesagt werden, dass es kein Modell und keine gültige Regel zur Erstellung einer Prognose gibt, vielmehr geht es darum, aus verschiedenen statistischen Verfahren und Algorithmen diejenigen Methoden auszuwählen, welche sich am besten für die Erhebung der benötigten Informationen eignen. (Beierle & Kern-Isberner, 2019, S. 1)

Welche Probleme können beim Einsatz von Predictive Analytics entstehen?

Nicht immer verlaufen Predictive Analytics Projekte reibungslos oder fehlerfrei, es drängt sich deshalb die Frage auf, welche Probleme oder Hürden bei der Entwicklung oder im Einsatz entstehen können. Bei der Realisierung eines Projektes im Bereich Predictive Analytics, werden oftmals viele Ressourcen im Unternehmen beansprucht. Dies startet bei der Konzeption und Ausarbeitung des Projektes im Team bis hin zum Testen und Validieren der Anwendung durch Experten/Expertinnen. Viele ambitionierte Projekte sind schon wegen fehlender Unterstützung gescheitert oder haben schlussendlich nicht den gewünschten Vorstellungen entsprochen. Deshalb ist hier entscheidend, sich die Unterstützung durch das Management zu sichern, um die nötigen Ressourcen während der Projektphase verfügbar zu haben. Eine weitere Hürde können schlechte oder inkonsistente Daten darstellen. Dies ist häufig der Fall, wenn Daten nicht regelmäßig aktualisiert werden, wie beispielsweise Anschriften oder Stammdaten. Auch fehlende

Schlüsselfelder in den verschiedenen Datenbanken können dazu führen, dass sich Daten schlecht oder gar nicht verbinden lassen und es so zu keinem zufriedenstellenden Ergebnis kommt. (Abbott, 2014, S. 12-13) Nicht immer entstehen Probleme außerhalb des Wirkungsbereiches des Analysten/der Analystin, oftmals ist auch er/sie es selbst welcher/welche das Vorhersagemodell zu komplex gestaltet, die Fehlersuche somit zeitintensiver wird, was wiederum den Abschluss des Projektes verzögert. Weitere Probleme, die sich durch ein zu komplexes Datenmodell ergeben können, ist die Überanpassbarkeit. Dadurch funktioniert die Interpretation der Daten nicht mehr zuverlässig. Speziell neue Daten werden dann nicht immer zuverlässig interpretiert. (Larose, 2015, S. 163) Gerade wenn Entscheidungen kurzfristig getroffen werden müssen, wie beispielsweise am Aktienmarkt oder in Wettermodellen, können komplexe Algorithmen kontraproduktiv sein, wenn es um die schnelle Bereitstellung der Informationen geht. (Abbott, 2014, S. 14) Abseits von prozessrelevanten Problemen ergeben sich im Einsatz auch Fragen von Objektivität. Solche Systeme bergen oft das Risiko, unfaire oder unüberlegte Ergebnisse zu liefern, die auf verzerrte oder diskriminierende Programmierung zurückzuführen sind. (Robinson, Harlan, & Rieke, 2014) Weltweit existieren zahlreiche automatische Entscheidungsalgorithmen, die in einer Vielzahl von Bereichen eingesetzt werden, mit denen die Eignung einer Person für Versicherungen oder Kredite beurteilt werden kann. (Pérez-Martin, Pérez-Torregrosa, & Vaca, 2018, S. 448-449) Forscher/Forscherinnen zeigen sich besorgt über die diskriminierenden Aspekte von Big Data Analysen, wovon einige bereits identifiziert wurden. (Robinson, Harlan, & Rieke, 2014) Aber auch Datenschutz kann ein großes Thema sein. Die meisten Prognosesysteme für den Einzelhandel beziehen ihre Daten anhand des Einkaufsverhaltens des/der Kunden/Kundin, welche in Form von Apps oder Bonuskarten generiert werden. (Ecker, 2019) Häufig sind sich Kunden//Kundinnen nicht bewusst, wofür ihre Daten verwendet werden und somit steigen auch die Bedenken hinsichtlich der Privatsphäre und missbräuchlichen Verwendung der Daten. (Inman & Nikolova, 2017, S. 17)

4 Erhebung und Auswertung der empirischen Ergebnisse

Das folgende Kapitel beschäftigt sich mit der empirischen Untersuchung hinsichtlich der Praktikabilität und den Einflussfaktoren von Cross-Selling in Kombination mit Predictive Analytics im SHK-Großhandel. Dabei wird die Methodenwahl und deren Ergebnisse näher erläutert.

4.1 Erhebungsmethode

Um die Hauptforschungsfrage: „Wie müsste ein Cross-Selling Algorithmus für den SHK-Großhandel gestaltet sein, um den Anforderungen der Zielgruppe zu entsprechen?“ beantworten zu können, wurden fünf leitfadenorientierte Interviews mit Experten und Expertinnen aus dem Bereich Sanitär-Heizung-Klima durchgeführt.

4.2 Auswertungsmethode

Für die Auswertung der Experten- und Expertinnen-Interviews wurde die qualitative Inhaltsanalyse nach Mayring gewählt. Mayring beschreibt die qualitative Inhaltsanalyse als eine Methode zur Analyse von Kommunikation, die das Ziel verfolgt, systematisch, theorie- sowie regelgeleitet eine eindeutig überprüfbare Auswertung des Textmaterials durchzuführen, um daraus Rückschlüsse auf relevante Kommunikationsaspekte zu ermöglichen. (Mayring, 2015, S. 13)

4.3 Sampling

Als Experten/Expertinnen wurden Personen in Führungspositionen der Sanitär-Heizung-Klima-Branche in West-Österreich sowie der Ost-Schweiz ausgewählt. Es wurden nur solche Betriebe interviewt, welche bereits Bestellungen in Online Shops getätigt haben. Die Daten darüber wurden mir von einem SHK-Großhandel zur Verfügung gestellt. Die fachspezifischen Kenntnisse sollen Aufschluss darüber geben, nach welchen Kriterien solche Bestellungen getätigt werden und wie sich diese auf eine Kaufprognose auswirken können. Letztendlich sollen diese Ergebnisse helfen, eine Methode zu entwickeln, anhand derer weiterführende Artikel vorgeschlagen werden können.

Tabelle 1: Zusammenstellung der Experten-/Expertinnnen-Interviews

Proband	Position im Unternehmen	Datum des Interviews	Ort des Interviews
A	Geschäftsführer	29.10.2019	Betriebssitz
B	Geschäftsführer	30.10.2019	Betriebssitz
C	Geschäftsführer	31.10.2019	Betriebssitz
D	Assistenz der GL	26.11.2019	Betriebssitz
E	Geschäftsführer	10.12.2019	Betriebssitz

Quelle: Eigene Darstellung

4.4 Operationalisierung

Um den Kriterien qualitativer Forschung gerecht zu werden, wurde bei der Umsetzung darauf geachtet, den klassischen Gütekriterien von Reliabilität (Genauigkeit) und Validität (Gültigkeit) zu entsprechen. (Mayring, 2015, S. 124-125)

Die Reliabilität und Validität wurden anhand standardisierter, klarer und verständlich formulierter Fragen sichergestellt, welche sich aus der Theorie ergaben und somit dem neuesten Stand der Forschung entsprechen. Um die Glaubwürdigkeit und Nachvollziehbarkeit zu gewährleisten, wurde die Vorgehensweise umfangreich dokumentiert. Interviews wurden digital aufgezeichnet und vollständig transkribiert. (Mayring, 2015, S. 53)

4.5 Durchführung der Interviews

Die Interviews fanden im Zeitraum vom 29.10.2019 – 12.12.2019 statt. Die Probanden/Probandinnen wurden im Vorfeld telefonisch über das Forschungsthema informiert. Alle erklärten sich einverstanden, sich die Zeit für ein Interview zu nehmen. Ansonsten wurden keine weiteren Details besprochen. Für die Audio-Aufnahme wurde ein Smartphone samt spezieller Recorder Software verwendet, welche später transkribiert wurde.

Die Probanden/Probandinnen erklärten sich einverstanden, dass ihre Interviews aufgezeichnet und anonymisiert werden. Die Interviews fanden an den jeweiligen Betriebsstätten statt.

4.5.1 Kategorienbildung

Das Ziel eines theoriebezogenen Kategoriensystems ist es, Themen, Inhalte und Aspekte aus dem Material zu filtern. In Folge ist es möglich, den paraphrasierten Text mittels Kategoriensystem zusammenzufassen. (Mayring, 2015, S. 103)

Tabelle 2: Übersicht Kategoriensystem

Nr. Kategorie	**Aussage**	**Fragen**
1	Fragen zum Betrieb und der Organisation	1.1-1.3
2	Vorgehensweise beim Bestellen über den Webshop sowie Vor- und Nachteile	2.1-2.5
3	Erfahrungen und Usability von Cross-Selling-Systemen	3.1-3.5
4	Erfassung von Daten und Faktoren für den Algorithmus	4.1-4.4

Quelle: Eigene Darstellung

4.5.2 Zusammenfassende Inhaltsanalyse

Bei der qualitativen Inhaltsanalyse werden die Texte systematisch in das theoriegeleitete Kategoriensystem eingegliedert. Dabei kann unter drei Formen unterschieden werden:

1. Die Zusammenfassung
2. Die Explikation
3. Die Strukturierung

In dieser Thesis wurde die Zusammenfassung gewählt. Hierbei werden die aufgezeichneten Audio-Dateien transkribiert, paraphrasiert, generalisiert und schlussendlich reduziert. (Mayring, 2015, S. 114-115)

4.6 Darstellung der empirischen Ergebnisse

Kategorie 1: Fragen zum Betrieb und der Organisation

Alle interviewten Experten/Expertinnen arbeiten bei kleineren Handwerksbetrieben die vorwiegend im Sanitär-Heizung- oder GWS-Bereich tätig sind. Einen Großteil der Experten/Expertinnen bilden die Geschäftsführer/Geschäftsführerinnen selbst. Aufgrund der Größe der Betriebe, werden die meisten administrativen Aufgaben auch von diesen selbst erledigt. Der Einkauf wird hauptsächlich über auslaufende Lagerstände oder Bestelllisten angestoßen. Alle Betriebe haben ihr eigenes kleines Lager wo ein Großteil der Ware nach dem Einkauf angeliefert wird. Größere Produkte werden meistens direkt auf die Baustelle selbst bestellt. In den meisten Fällen organisiert den Einkauf der Geschäftsführer/die Geschäftsführerin oder in seltenen Fällen auch ein Einkäufer bzw. eine Einkäuferin.

Kategorie 2: Vorgehensweise beim Bestellen über den Webshop sowie Vor- und Nachteile

Als primärer Hauptgrund den Webshop anderen Bestellvarianten vorzuziehen, wird die Zeitersparnis genannt. Weitere Faktoren waren die geringere Fehlerquote und die Möglichkeit zu jeder Uhrzeit bestellen zu können. Alle Experten/Expertinnen waren sich einig, dass der Online-Einkauf in ihrer Branche in Zukunft zunehmen wird. Negativ wurde gewertet, dass der persönliche Kontakt verloren geht und Mitarbeiter/Mitarbeiterinnen umgeschult werden müssen. Als positiv wurde gewertet, dass die Informationen jederzeit abgerufen werden können und die Bestelldauer sich dadurch verkürzt. Der Umweg über klassische Bestellverfahren (Telefon, E-Mail, ...) erfolgt hauptsächlich wegen fehlender Informationen oder unbekannten Produkten, bei denen noch eine persönliche Beratung erforderlich ist. Je nach Geschäftsfall wird der Einkauf hauptsächlich manuell über auslaufende Lagerstände oder Bestelllisten angestoßen.

Kategorie 3: Erfahrungen und Usability von Cross-Selling Systemen

Ein überwiegender Teil der Experten/Expertinnen, die bereits Onlineshopping betrieben haben, kennen die Cross-Selling-Methode bereits. Die Experten/Expertinnen finden die Idee grundsätzlich gut, wenn auch die Gründe dafür andere sind als im B2C Bereich. Bei den Vorschlägen sollte dabei immer auf die präferierten Hersteller des jeweiligen Kunden/der jeweiligen Kundin geachtet werden. Hauptgründe dafür sind die Qualität, Haftung und garantierte Ersatzteile. Die Experten/Expertinnen sind sich einig, dass sich die Bestelldauer mit dieser Methode verkürzen würde, sehen somit den Vorteil hauptsächlich als Zeitersparnis beim Bestellvorgang. Weitere Vorteile wären Minimierung der Fehlerquellen und eine Gedankenstütze, damit kein Material vergessen wird. Der Nutzen wird im Wesentlichen von der Erfahrung des jeweiligen Einkäufers/der jeweiligen Einkäuferin abhängig sein. Ein Großteil der Experten/Expertinnen könnten sich durchaus vorstellen, dass ein derartiges Feature eine Auswirkung auf die Wahl des Online Shops haben könnte.

Kategorie 4: Erfassung von Daten und Faktoren für den Algorithmus

Es werden bestimmte Hersteller bevorzugt, die Hauptgründe dafür sind die Qualität, Haftung und garantierte Ersatzteile. Komponenten ohne Abhängigkeit bilden eher die Ausnahme. Über das Jahr gesehen, wird ein Großteil des Materials regelmäßig zyklisch bestellt. Es gibt gewisse Verschiebungen in den Monaten aufgrund der Jahreszeit und den Abhängigkeiten zu den Bauprojekten. Die Experten/Expertinnen schlagen vor, das jeweilige Kaufverhalten der Kunden/Kundinnen im Webshop zu analysieren, daraus lassen sich Rückschlüsse auf die priorisier-

ten Hersteller und Artikel ziehen. Des Weiteren besitzen die Großhändler bereits wichtige Informationen zu den jeweiligen Bauprojekten und können abschätzen, welche Materialien der Handwerksbetrieb in Zukunft benötigen wird.

4.7 Beantwortung der empirischen Subforschungsfragen

Wie schätzen Experten/Expertinnen die Praktikabilität eines solchen Cross-Selling Algorithmus ein?

Die Auswertung hat gezeigt, dass ein überwiegender Teil der Experten/Expertinnen, die bereits Onlineshopping betrieben haben, die Cross-Selling-Methode bereits kennen. Die Idee wird grundsätzlich positiv aufgenommen, wenn auch die Gründe dafür andere sind als beispielsweise im B2C Bereich. (Malms & Schmitz, 2008, S. 30) Die Experten/Expertinnen sind sich einig, dass sich die Bestelldauer mit dieser Methode verkürzen würde, sehen den Vorteil somit hauptsächlich als Zeitersparnis beim Bestellvorgang. Weitere Vorteile, die erwähnt wurden, waren die Minimierung der Fehlerquellen und eine Gedankenstütze, damit kein Material vergessen wird. Die Experten/Expertinnen stellen fest, dass der Nutzen im Wesentlichen von der Erfahrung des jeweiligen Einkäufers/der jeweiligen Einkäuferin abhängig sein wird. Ein Großteil könnte sich durchaus vorstellen, dass ein derartiges Feature eine Auswirkung auf die Wahl des Onlineshops haben könnte. Alle Experten/Expertinnen waren sich einig, dass der Online-Einkauf in ihrer Branche in Zukunft zunehmen wird.

Welche Daten und Faktoren empfehlen SHK-Experten/Expertinnen für eine zielführende Vorhersage?

Beim Interview gaben alle Experten und Expertinnen an, dass die Betriebe ihre eigenen kleinen Zwischenlager betreiben und somit nicht nur direkt, sondern überwiegend auch auf Lager bestellen. Auslaufende Lagerstände bilden dabei häufig auch den Impuls, neue Ware zu bestellen. Weiters empfehlen die Experten und Expertinnen bei den Vorschlägen immer auf die präferierten Hersteller des jeweiligen Kunden/der jeweiligen Kundin zu achten. Die Gründe hierfür sind die Qualität, Haftung und die garantierten Ersatzteile. Die Auswertung hat auch ergeben, dass über das Jahr gesehen, ein Großteil des Materials regelmäßig zyklisch bestellt wird. Es kann dabei zu gewissen Verschiebungen in den Monaten aufgrund der Jahreszeit und den Abhängigkeiten der Bauprojekte kommen, der Großteil wird aber regelmäßig benötigt. Die Experten und Expertinnen empfehlen auch eine Analyse des Einkaufsverhaltens. Daraus würden sich wichtige Informationen zu den Kunden/Kundinnen ableiten lassen.

5 Prototyp: Cross-Selling-Algorithmus

Für den Projektablauf wurde die in der Literatur beschriebene Methode zur Entwicklung von wissensbasierten Systemen von Beierle und Kern-Isberner (Beierle & Kern-Isberner, 2019, S. 19-20) verwendet. Diese sieht einen achtstufigen Entwicklungsprozess vor, wobei diese Thesis nur den Bereich bis zum Testen des Prototypen abdecken soll. Die Wissensbasis beruht zum großen Teil auf den Erkenntnissen der im Interview befragten Experten/Expertinnen, deshalb handelt es sich hierbei zum Teil auch um ein sogenanntes Expertensystem. (Gottlob, Frühwirth, & Horn, 1990, S. 14) Für die Entwicklung des Algorithmus wurde der von Fayyad (Fayyad, Piatetsky-Shapiro, & Padhraic , 1996, S. 14-16) empfohlene KDD-Prozess angewandt, dessen Ziel es ist, Muster aus Datensätzen zu extrahieren und diesen Eigenschaften zuzuordnen.
Die verwendeten Werkzeuge und die Programmlogiken wurden vom Autor selbst gewählt, basierend auf den in der Theorie erforschten Methoden zur Wissensfindung in Daten.

5.1 Ziele und Rahmenbedingungen

Für die Umsetzung des Prototyps wurden folgende Rahmenbedingungen definiert:

- Anhand von anonymisierten Webshop-Daten eines Kunden aus der SHK-Branche, sollen Prognosen errechnet werden, welche Artikel zusätzlich verkauft werden könnten.
- Die Berechnung der Prognosen, soll mittels Einsatz von Predictive Analytics Methoden erfolgen.
- Die Cross-Selling Optionen sollen in einem Prototypen dargestellt werden, der Artikelkäufe in einem Webshop zu einem bestimmten Datum simuliert.
- Das Ergebnis soll anhand mehrerer Warenkörbe ausgewertet werden, welche nach der Erstellung des Prototyps angelegt wurden.

5.2 Abgrenzungen

Das Ergebnis dieser Arbeit soll keine allgemeingültige oder vollständige Software für die SHK-Branche darstellen. Vielmehr sollen mit Hilfe des Prototyps die Möglichkeiten und die Chancen dieser Technologie aufgezeigt und die werkleitende Subforschungsfrage beantwortet werden.

5.3 Technologien

- **RStudio**: Für die Berechnung des Apriori-Algorithmus wurde die Software RStudio verwendet, welche in der Basisversion kostenlos ist und unter der Lizenz AGPL V3 (Open Source License) vertrieben wird. RStudio ist eine vom Unternehmen RStudio, Inc. angebotene, integrierte Entwicklungsumgebung und grafische Benutzeroberfläche für die statistische Programmiersprache R. R ist ursprünglich eine prozedurale Programmiersprache die Befehlszeilen hintereinander ausführt, welche man z.B. in einem einfachen Textfile speichern kann und in der sogenannten R-Konsole ausführt. (RStudio , 2020)
- **SQL**: Um Daten von relationalen Datenbanken abzufragen, wurde auf die Datenbanksprache SQL zurückgegriffen. SQL wurde 1970 von IBM entwickelt und wird seither stetig verbessert und aktualisiert. Die Sprache hat ihren Ursprung in der Algebra und ist semantisch an die englische Umgangssprache angelehnt. (Harrington, 2010, S. 65-67)
- **QlikView**: Für die visuelle Darstellung des Prototyps kommt die Business Intelligence Software QlikView in der Version 12 zum Einsatz. Die Software wird als Freemium Modell vertrieben und ist für den privaten Gebrauch kostenlos. QlikView arbeitet mit der In-Memory Technologie mit Hilfe derer große Datenmengen in Echtzeit analysiert werden können. (QlikTech, 2017)

5.4 Datenbasis

Die Datenbasis wurde dem Autor von einem SHK-Großhandel zur Verfügung gestellt, aufgrund von Datenschutzrichtlinien aber soweit anonymisiert, dass keine Rückschlüsse auf Kunden/Kundinnen oder Betriebsprozesse möglich sind. Als Datengrundlage wurde ein Kunde/eine Kundin ausgewählt, der/die eine Webshop-Quote von über 80% aufweist.

5.5 Design

Die fertige Lösung soll in einen bestehenden Webshop implementiert werden. Dieser bietet bereits ein eigenes Cross-Selling Anzeigemodul, welches allerdings keine eigene Programmlogik besitzt und auf Informationen von anderen Systemen angewiesen ist. Insgesamt können pro eingegebenem Artikel im Warenkorb, jeweils vier Cross-Selling Optionen angezeigt werden. Diese Implementierung wird allerdings nicht in dieser Forschungsarbeit behandelt, sondern dient lediglich der Designvorgabe. Es soll lediglich eine Anzeigemaske dargestellt werden, wo ein Artikelkauf simuliert wird und die Prognosen mit ihrer Wahrscheinlichkeit dargestellt werden.

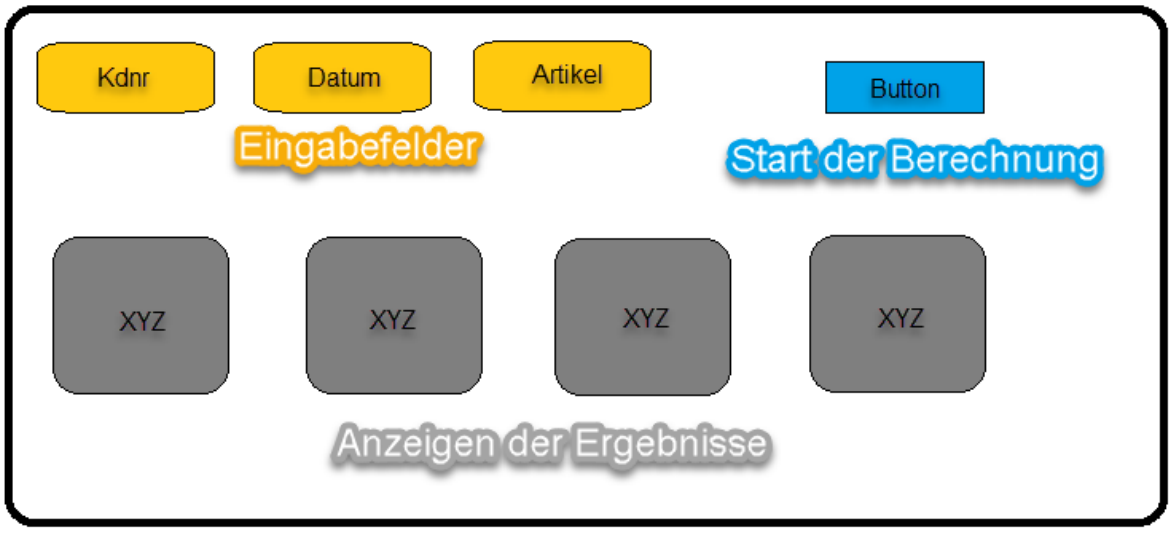


Abbildung 23. Design Entwurf für den Prototypen

Quelle: Eigene Darstellung

5.6 Projektdokumentation

5.6.1 Vorbereiten der Datenbasis

Im ersten Schritt werden die wichtigsten Webshop Daten (Kundennummer, Belegnummer, Artikelnummer, Menge, Einheit, Buchungsdatum) eines Kunden/einer Kundin mittels SQL und QlikView aus dem ERP-System extrahiert. Wie im empirischen Teil von den Experten/Expertinnen beschrieben, sind gewisse Artikel jahreszeitabhängig, deshalb wurde als Betrachtungszeitraum ein ganzes Jahr ausgewählt. Des Weiteren sollte jeder Kunde/jede Kundin einzeln ausgewertet werden, oder zumindest eine hohe Übereinstimmung beim Kaufverhalten aufweisen, da eine Durchmischung der Hersteller bei den Interviews als nicht zielführend beschrieben wurde. Der Datenload setzt sich aus Bewegungsdaten sowie Stammdaten zusammen. Um mit den Daten im weiteren Verlauf besser arbeiten zu können, wurden Transformationen vorgenommen und beispielsweise Datentypen umgewandelt und Kalenderdaten generiert. Abschließend wurden die Daten mittels eines *STORE* Befehles in einem QlikView Container Objekt zwischengespeichert.

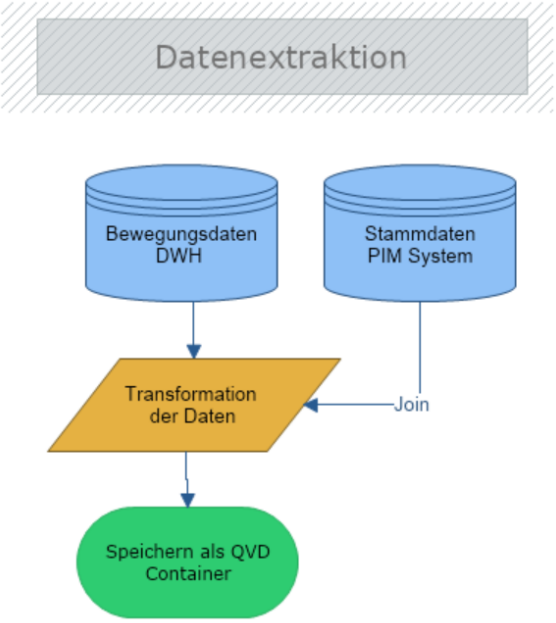


Abbildung 24. Datenextraktion und Transformation aus dem DWH

Quelle: Eigene Darstellung

5.6.2 Kennzahlen Dashboard

Dashboards helfen dabei die Datenbasis besser zu verstehen und analysieren zu können. Mit Hilfe von Dashboards kann man sich einen ersten Überblick verschaffen, um nutzbare Informationen aus riesigen Rohdatensätzen extrahieren zu können. Weiters lassen sich damit Probleme oder Unvollkommenheiten in der Datenbasis finden. Es soll auch dabei helfen, den Analyseprozess zu beschleunigen und Muster zu finden, die für das Projekt relevant sind. Zusätzlich können damit Ergebnisse aus dem Prototypen getestet und validiert werden. (Abbott, 2014, S. 35)

Abbildung 25. Zusammenfassung der Rohdaten

Quelle: Eigene Darstellung

In der Abbildung 25 kann man erkennen, dass insgesamt 224 Belege von dem Kunden im Jahr 2019 erstellt worden sind. In diesem Fall entspricht der Beleg einem bestellten Warenkorb im Webshop. In diesen Belegen finden sich 1350 unterschiedliche Artikel aus 3202 Zeilen wieder. Die Artikel wurden *Distinct* gezählt, das bedeutet, dass sie beim mehrfachen Auftreten nur einmal gezählt werden. In der rechten Spalte befindet sich eine Auflistung der am häufigsten bestellten Artikel und deren Mengen.

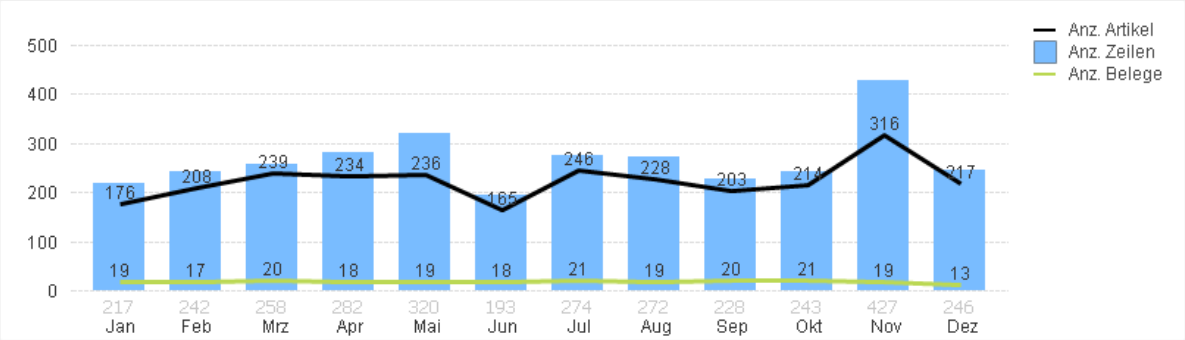


Abbildung 26. Diagramm mit dem zeitlichen Verlauf der Bestellungen

Quelle: Eigene Darstellung

Weitere Informationen, die sich aus der Abbildung 26 ablesen lassen, ist die Anzahl der Belege in den Monaten. Diese Balken verlaufen relativ konstant mit im Schnitt ungefähr 19 Bestellungen pro Monat, das könnte darauf hindeuten, dass wenig auf Lager bestellt wird, sondern Material eher kurzfristig eingelagert wird. Das würde sich auch mit den Informationen aus den Interviews decken, dass es sich hierbei nur um kleine Zwischenlager handelt. Ein ähnlicher Verlauf zeichnet sich bei den Artikeln ab, wobei hier die Kurven im Juni, sowie im Dezember kurzfristig nach einem Peak stark abfallen. Das kann auf eine kurzfristige Unterbrechung der Geschäftstätigkeit wie beispielsweise durch Urlaube zurückzuführen sein.

Artikelmengen pro Monat

Art.Nr. Eh. / Monat	Jan	Feb	Mrz	Apr	Mai	Jun	Jul	Aug	Sep	Okt	Nov	Dez
13449-STK	2	1	1	3	13	6	3	4	2	-	7	1
512444-STK	8	7	2	3	1	1	1	6	1	1	10	11
16100-STK	-	34	2	137	82	8	4	16	1	30	22	3
40211-STK	10	10	10	30	41	20	20	21	-	10	10	10
477557-STK	3	1	-	1	5	-	3	3	-	-	10	2
285734-STK	30	80	20	50	40	-	20	40	-	20	60	20
51089-STK	14	3	2	2	-	-	8	-	3	10	7	2
41230-STK	130	70	60	30	70	-	30	30	30	60	30	40
59617-GAR	6	1	1	11	-	-	1	4	5	6	1	7
478742-STK	-	11	-	26	-	-	1	13	-	-	5	12

Abbildung 27. Tabellenansicht der bestellten Artikelmengen in den Monaten

Quelle: Eigene Darstellung

Die Tabelle in der Abbildung 27 stellt die Artikel und die Mengen in den Monaten dar. Mit dieser Ansicht lässt sich leicht überprüfen, welche Artikel regelmäßig bestellt werden und welche Artikel eher saisonalen Charakter haben. Weiters kann überprüft werden welche Artikel hohe Schwankungen bei den einzelnen Bestellungen aufweisen. Um solche Variabilität zu überprüfen, eignen sich statistische Parameter wie das Streuungsmaß. (Holland & Scharnbacher, 2010, S. 51)

Art.Nr. Eh.	Summe Jahr	Mittelwert	Standard Abw.	Spannweite
512444-STK	52	2,08	2,16	9,00
40211-STK	192	9,14	2,71	9,00
285734-STK	380	21,11	4,71	20,00
477557-STK	28	1,56	0,86	3,00
41230-STK	580	34,12	6,18	20,00
51089-STK	51	3,00	2,60	9,00
512202-M	1600	106,67	49,52	150,00
41437-STK	250	17,86	4,26	10,00
41434-STK	270	22,50	4,52	10,00
59584-STK	70	5,83	4,43	9,00

Abbildung 28. Berechnen der Variabilität

Quelle: Eigene Darstellung in Anlehnung an (Holland & Scharnbacher, 2010, S. 54)

In der Abbildung 28 wird ein Ausschnitt der Daten und die Berechnung der verschiedenen Streumaße dargestellt. Der Mittelwert gibt dabei die durchschnittliche Bestellgröße pro Bestellung an. Der wichtigste Streuparameter in der Praxis ist die Standardabweichung, dieser beschreibt die durchschnittliche Entfernung der Werte zum Mittelwert. Umso kleiner dieser Parameter ist, desto gleichmäßiger ist die Größe der bestellten Menge bei diesem Artikel. Ein weiteres Streuungsmaß ist die Spannweite, dieses beschreibt die Differenz vom niedrigsten zum höchsten Wert. Diese Kennzahl dient nur dazu einen schnellen Überblick zu generieren, sie wird nämlich stark durch Extremwerte verfälscht. (Holland & Scharnbacher, 2010, S. 51-56)

5.6.3 Datenaufbau für R

Damit man den Apriori-Algorithmus in R verwenden kann, müssen die Daten, welche bereits als *TestDaten.qvd* vorgeladen wurden, als CSV exportiert werden. Dazu verwendet man die *STORE* Funktion von QlikView. Dabei werden nur die Belegnummer und die Artikelnummer mit *ORDER BY* sortiert und in einer vertikalen Form als CSV mit definiertem Trennzeichen exportiert. Werden weitere Parameter benötigt, können diese mit ausgegeben werden, entscheidend ist aber die Reihung nach Belegnummer und Artikel.

Tabelle 3: CSV Daten Beispiel

Belegnummer,Artikelnummer
R10001,21553
R10001,29001
R10002,59840
…,…

Quelle: Eigene Darstellung

5.6.4 Deskriptive Datenanalyse

Bevor die CSV Datei in R importiert werden kann, benötigt man noch ein zusätzliches Package. Dieses trägt den Namen „*arules*“ und erweitert die Software mit dem Apriori-Algorithmus. Nach der Installation kann die CSV Datei mittels der Funktion *read.transactions(“Pfad“ , format = "single", sep = ",", cols = 1:2)* in R eingelesen werden. Wichtig hierbei ist es, dass das Format und die Parameter korrekt angegeben werden. Bezogen auf das Beispiel in Tabelle 3 bedeutet das: Die Daten befinden sich in Spalte eins und zwei in vertikaler Anordnung, das Trennzeichen, welches die einzelnen Elemente voneinander trennt, wurde mit einem Beistrich definiert.

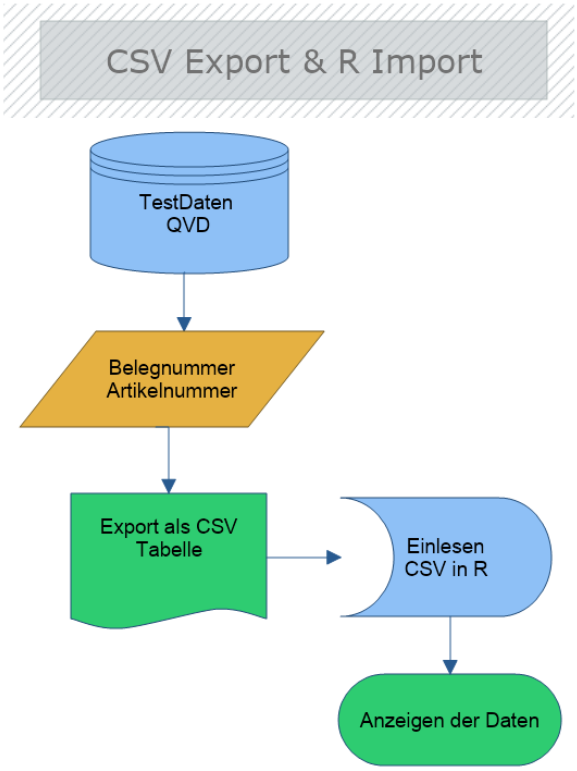


Abbildung 29. CSV Export in QlikView und Einlesen in R

Quelle: Eigene Darstellung

Nach dem Einlesen wird in der Konsole eine Zusammenfassung ausgegeben.

```
>  read.transactions(file = "C:\\Transaktionen.csv", format = "single", sep = ",", cols = 1:2)
transactions in sparse format with
 225 transactions (rows) and
 1351 items (columns)
```

Abbildung 30. Erfolgreicher Import der Daten

Quelle: Eigene Darstellung

Die Daten befinden sich jetzt in einem frei definierbaren Objekt mit dem Namen „*daten*“ und können nach Zuweisung eines Operators mittels der Funktion *summary(daten)* bereits deskriptiv analysiert werden.

```
most frequent items:
 512444   13449   16100   40211  285734 (Other)
     24      22      22      21      18    3068

element (itemset/transaction) length distribution:
sizes
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 30 31 32 33 34 35 36 37 41
 7 10 10 12 13  8  9  5 12 12 11 11  9 10  9  7  8  2  4  7  5  5  4  3  1  3  2  3  4  3  2  1  2  1  2  1  3

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    6.00   12.00   14.11   19.00   60.00
```

Abbildung 31. Darstellung der wichtigsten Kennzahlen in R

Quelle: Eigene Darstellung

In der Zusammenfassung in Abbildung 31, sieht man bereits die fünf meistverkauften Artikel sowie die Größe der Warenkörbe. Es existieren jeweils sieben Warenkörbe mit nur einem Artikel. Diese könnten theoretisch schon beim Laden der Daten weggefiltert werden, da sie für eine Assoziationsanalyse unbrauchbar sind. Für die verbesserte Darstellung lassen sich die Warenkorbgrößen in R auch mit einem Histogramm darstellen, dazu wird einfach die Funktion *hist(size(daten),col="lightgrey")* aufgerufen.

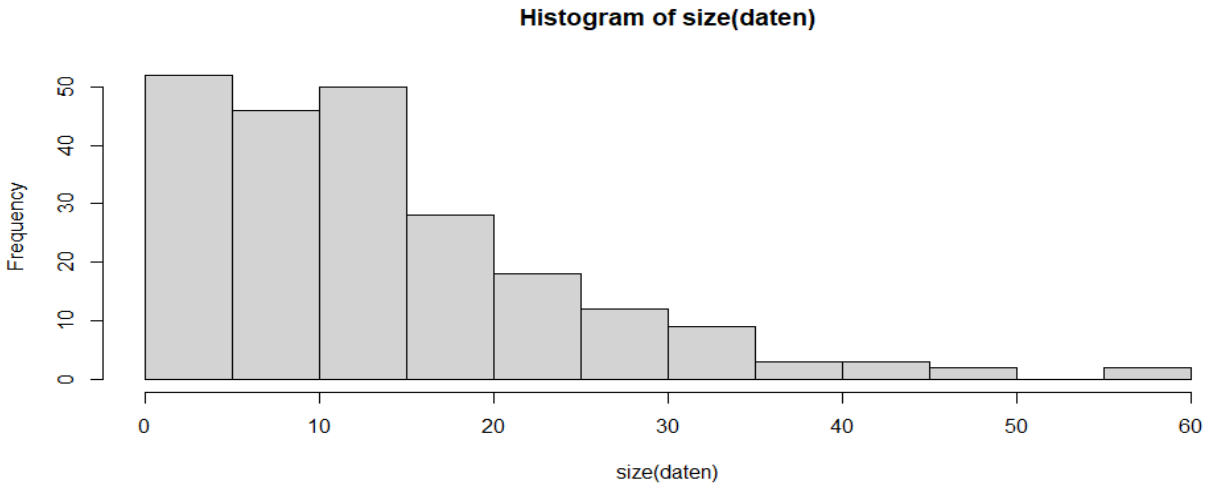


Abbildung 32. Histogramm der Warenkörbe und Anzahl Artikel

Quelle: Eigene Darstellung

Wie in Abbildung 32 ersichtlich, befinden sich die meisten Warenkörbe im linken Bereich. Somit liegt der durchschnittliche Warenkorb zwischen einem und zwanzig Artikeln. Nach zwanzig ist der Warenkorb stark abfallend und bildet somit eher die Ausnahme.

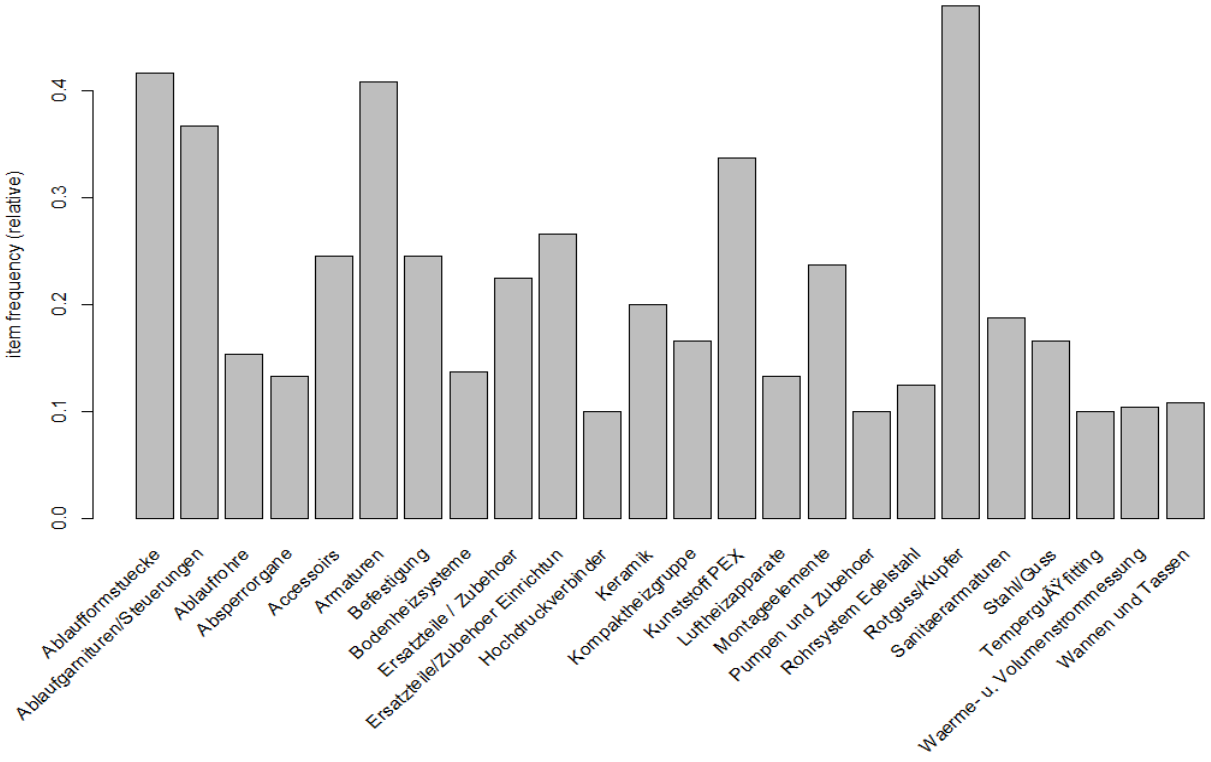


Abbildung 33. Relative Häufigkeiten der häufigsten Artikelgruppen

Quelle: Eigene Darstellung

Die am häufigsten bestellten Artikelgruppen werden in Abbildung 33 dargestellt. Diese Darstellung kann mit Hilfe der Funktion *itemFrequencyPlot (daten, support=0.1)* aufgerufen werden. Dabei wird zusätzlich noch nach relativer Häufigkeit größer 0,1 gefiltert. Diese Ansicht kann nützlich zur Identifikation von häufig bestellten Artikelgruppen sein, wenn es darum geht, bestimmte Gruppen zu priorisieren.

5.6.5 Sortimentsverbundanalyse

Verbundanalysen sind statistische Methoden, um nach Strukturen in den Daten zu suchen. Als Grundlage werden Ähnlichkeitsmuster erstellt, die für jedes Produktpaar die Häufigkeit des gemeinsamen Vorkommens in den Daten bestimmen. (Hahsler & Reutterer, 2006, S. 1) Dazu extrahiert man abermals die Rohdaten in eine CSV Datei, dieses Mal aber mit der Artikelgruppe anstatt den Artikeln. Mit Hilfe der Funktion *crosstable(daten)* lässt sich eine Ähnlichkeitsmatrix erstellen, die für jede Artikelgruppe die Häufigkeit des gemeinsamen Auftretens in Belegen bestimmt.

	Armaturen	Befestigung	Begleitheizband	Behaelter	Bodenablaeufe	Bodenheizsysteme
Ablaufformstuecke	47	29	0	1	4	19
Ablaufgarnituren/Steuerungen	52	19	2	0	4	17
Ablaufrohre	14	10	0	0	2	9
Absperrorgane	17	9	0	1	0	5
Accessoirs	42	16	2	1	1	12
Armaturen	98	25	2	1	6	20
Befestigung	25	59	0	0	3	9
Begleitheizband	2	0	2	0	0	0
Behaelter	1	0	0	1	0	0
Bodenablaeufe	6	3	0	0	7	0
Bodenheizsysteme	20	9	0	0	0	33
Dachentwaesserung	0	0	0	0	0	0

Abbildung 34. Ähnlichkeitsmatrix der Artikelgruppe

Quelle: Eigene Darstellung

Anhand dieser Basis lässt sich eine hierarchische Clusteranalyse erstellen. Dazu lädt man die Daten zuerst in eine Unähnlichkeitsmatrix und definiert mittels *itemFrequency* die relative Häufigkeit der einzelnen Artikelgruppen die in diesem Beispiel größer 5% vorkommen. *dissJacc <- dissimilarity(daten[, itemFrequency(daten) > 0.05], method = "Jaccard", which = "items")* Anschließend wird das Ergebnis mit der Funktion *hclust* als hierarchische Clusteranalyse in ein Objekt „*hcWard*“ geladen: *hcWard <- hclust(dissJacc, method = "ward.D")* Mit der *Plot* Funktion kann das Ergebnis als Dendrogramm visualisiert werden. (Hahsler & Reutterer, 2006, S. 8-9)

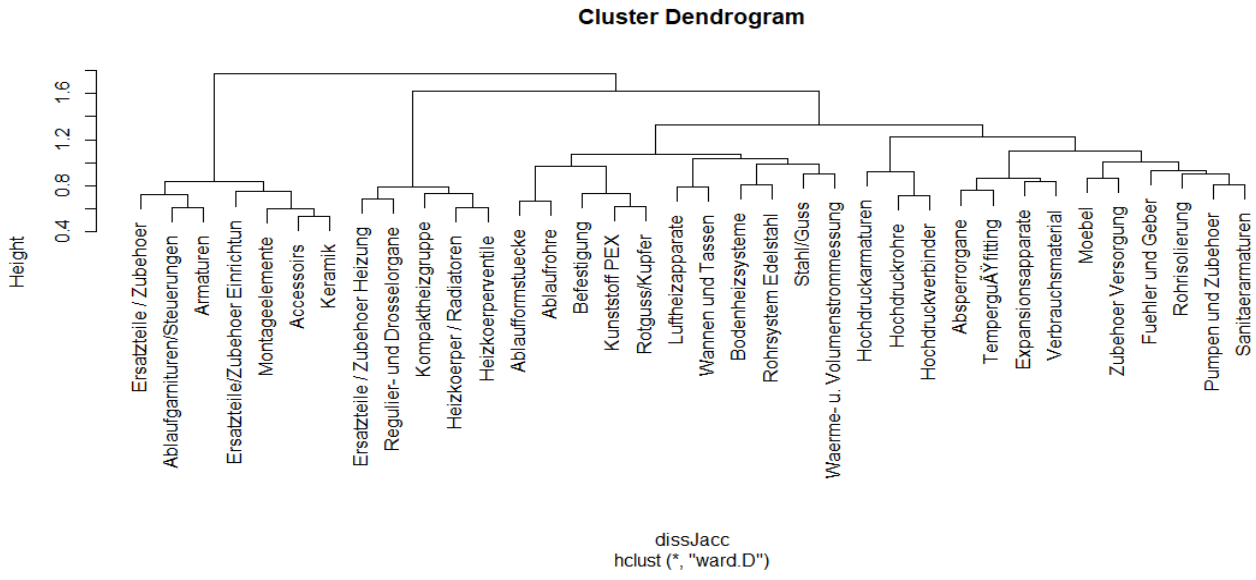


Abbildung 35. Dendrogramm der hierarchischen Clusteranalyse

Quelle: Eigene Darstellung

5.6.6 Apriori mit R berechnen

Bevor man mit dem Apriori-Algorithmus rechnen kann, muss zuerst eine Regel definiert werden. *regeln<-apriori(daten,parameter=list(support = 0.01, confidence = 0.10))*

Mit den Parametern Min Support = 1% und Min Konfidenz = 10% liefert die Funktion *apriori* eine entsprechend angepasste Trefferliste. (Hahsler & Reutterer, 2006, S. 14)

```
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target   ext
        0.1    0.1    1 none FALSE            TRUE       5    0.01      1     10  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 2
```

Abbildung 36. Definieren der Berechnungsparameter

Quelle: Eigene Darstellung

Die wichtigste Kennzahl in Abbildung 36 ist der „*Absolute minimum support count*“, dieser zählt die Transaktionen, die dem Minimum Support von 1% entsprechen. Je höher dieser Wert, desto besser die Datengrundlage. Mit zwei ist dieser Wert aber verhältnismäßig gering, was mit der hohen Anzahl unterschiedlicher Artikel bei diesem Kunden zusammenhängt. Um diesen Wert zu verbessern, könnte man entweder den Zeitraum der Betrachtung erhöhen, oder Kunden/Kundinnen, die ein ähnliches Einkaufsverhalten aufweisen, in einem Cluster zusammenfassen, um so mehr Datensätze zu erhalten. Was diese Regel für dieses Datenset bedeutet, kann man mit der *summary* Funktion abfragen. (Hahsler & Reutterer, 2006, S. 15-16)

```
> summary(regeln)
set of 12299 rules

rule length distribution (lhs + rhs):sizes
   1    2    3    4    5    6    7    8    9
   1 1080 2202 3216 3100 1884  679  128    9

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   3.000   4.000   4.414   5.000   9.000

summary of quality measures:
    support          confidence          lift            count
 Min.   :0.01333   Min.   :0.1067   Min.   : 1.00   Min.   : 3.000
 1st Qu.:0.01333   1st Qu.:0.7500   1st Qu.:10.23   1st Qu.: 3.000
 Median :0.01333   Median :1.0000   Median :16.88   Median : 3.000
 Mean   :0.01490   Mean   :0.8577   Mean   :19.96   Mean   : 3.353
 3rd Qu.:0.01333   3rd Qu.:1.0000   3rd Qu.:22.50   3rd Qu.: 3.000
 Max.   :0.10667   Max.   :1.0000   Max.   :75.00   Max.   :24.000
```

Abbildung 37. Zusammenfassung der Regel

Quelle: Eigene Darstellung

Insgesamt wurden 12299 Regeln erstellt. Die Kennzahl „*rule length distribution*“ gibt an, wie viele Artikel in einer Regel definiert sind. So existiert eine Regel, die nur einen Artikel enthält, dies kann vorkommen, wenn der Artikel nur einmal in einem einzelnen Beleg ohne andere Artikel bestellt wurde. Am häufigsten mit 3216 Regeln sind vier Artikel betroffen. Weitere Eigenschaften lassen sich aus der Zusammenfassung der Qualitätskennzahlen ableiten. So ergibt sich beispielsweise ein *Max Confidence* von 100%, was einer hundertprozentigen Übereinstimmung eines Artikels zu einem anderen entspricht. Das bedeutet, dass diese zwei oder mehrere Artikel immer gemeinsam bestellt wurden. Diese Erkenntnis deckt sich mit Aussage der Experten/Expertinnen, wonach Artikel häufig Abhängigkeiten zu anderen Artikeln aufweisen. Mit der Funktion *inspect* kann man einen Blick in die Regel werfen.

```
> inspect(sort(regeln, by = "support"))
      lhs                    rhs      support    confidence  coun
[1]   {531190}            => {531564} 0.06222222 0.9333333   14
[2]   {531564}            => {531190} 0.06222222 0.9333333   14
[3]   {16100}             => {13449}  0.06222222 0.6363636   14
[4]   {13449}             => {16100}  0.06222222 0.6363636   14
[5]   {478742}            => {16100}  0.05333333 0.8000000   12
[6]   {16100}             => {478742} 0.05333333 0.5454545   12
[7]   {477557}            => {13449}  0.05333333 0.7058824   12
[8]   {13449}             => {477557} 0.05333333 0.5454545   12
[9]   {123399}            => {204004} 0.04444444 1.0000000   10
[10]  {204004}            => {123399} 0.04444444 1.0000000   10
```

Abbildung 38. Top zehn Artikel mit Support und Wahrscheinlichkeit

Quelle: Eigene Darstellung

Wird also der erste Artikel 531190 bestellt, besteht eine 93,3% Wahrscheinlichkeit, dass der Artikel 531564 mitbestellt wird. Insgesamt konnte diese Regel vierzehnmal beobachtet werden. Vergleicht man den Artikel mit den Rohdaten, so kann man feststellen, dass der Artikel insgesamt 15-mal bestellt wurde und nur einmal ohne den Artikel 531564.

Artikel Bestellungen

Artikelnummer	Anzahl	Menge
531190	15	15

Abbildung 39. Vergleich der Rohdatenanalyse vom Dashboard

Quelle: Eigene Darstellung

Daraus ergibt sich die hohe Wahrscheinlichkeit im „*confidence*“. Um mit den Daten in QlikView weiter zu arbeiten, können diese mit dem Befehl *write.csv2(inspect(regeln), file = 'regeln.csv')* in eine CSV Datei exportiert werden.

5.6.7 Lagerreichweite für Trendartikel berechnen

Um Wissen aus Daten zu generieren, gilt es wie aus der Theorie abgeleitet, nach Mustern zu suchen, diese zu Klassifizieren und anschließend zu Generalisieren. (Ester & Sander, 2000, S. 4-5) Da wir aus den Interviews die Information erhalten haben, dass die Betriebe ihre eigenen Zwischenlager betreiben, kann man für Trendartikel als zusätzliches Attribut, die Lagerreichweite anhand der Vergangenheitsdaten berechnen. Ziel ist es die Prognosen noch weiter zu verbessern, da so Artikel, welche sich innerhalb der Lagerreichweite befinden, eine geringere Priorität zugewiesen bekommen. (Bentz, 1984, S. 221-223)

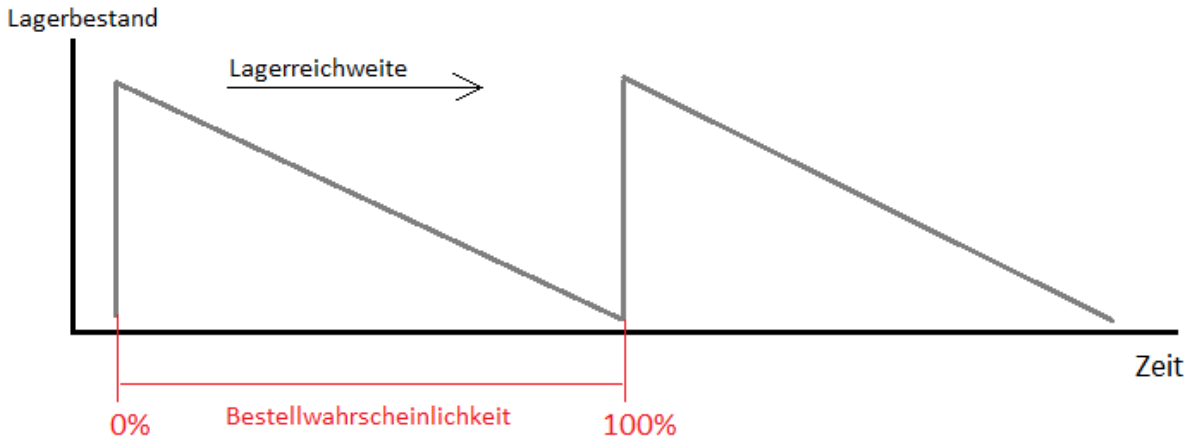


Abbildung 40. Bestellwahrscheinlichkeit im Verhältnis zur Lagerreichweite

Quelle: Eigene Darstellung

Die Formel zur Berechnung der Lagerreichweite lautet:

$$\text{Lagerreichweite in Tagen} = \frac{\varnothing\ \text{Lagerbestand}}{\varnothing\ \text{Bedarf pro Tag}}$$

Abbildung 41.Formel für die Lagerreichweiten Berechnung

Quelle: https://logistikknowhow.com/bestandsverwaltung/die-lagerreichweite abgerufen am 28.03.2020

Das Ganze lässt sich mit den Daten aus dem Dashboard nachbauen. Dazu verwendet man ein Diagrammobjekt aus QlikView und die Artikelnummer als Dimension. Die Formeln ergeben sich wie folgt:

Ø Tagesverbrauch: *sum(BasisMenge)/DayNumberOfYear(YearEnd(Year(Buchungsdatum)))*

Menge letzte Bestellung: *FirstSortedValue(BasisMenge, -Buchungsdatum)*

Datum letzte Bestellung: *Date(FirstSortedValue(Buchungsdatum, -Buchungsdatum))*

Lagerreichweite in Tagen: *[Menge letzte Bestellung]/[Ø Tagesverbrauch]*

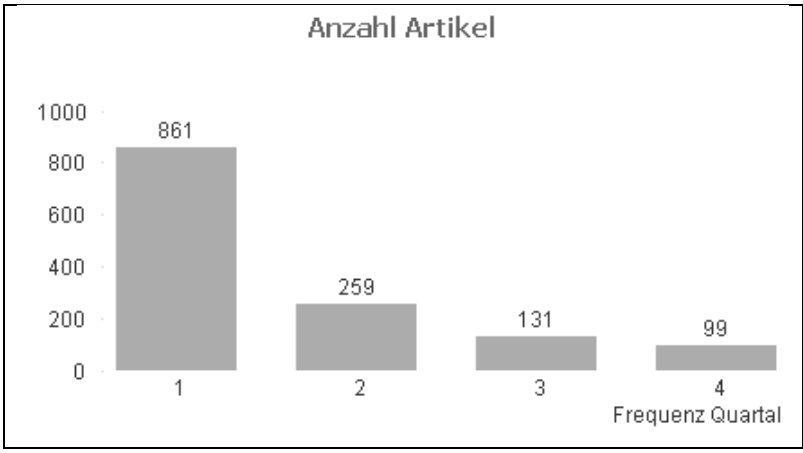
Voraussichtliche Bestellung*: Date(FirstSortedValue(Buchungsdatum, -Buchungsdatum)+[Lagereichweite Tage])*

Artikelnummer	Ø Tagesverbrauch	Menge letzte Bestellung	Datum letzte Bestellung	Lagereichweite Tage	voraus. Bestellung
16100	0,929	3	10.12.2019	3	13.12.2019
40315	0,173	1	27.11.2019	6	02.12.2019
512444	0,142	1	12.12.2019	7	19.12.2019
13449	0,118	1	10.12.2019	8	18.12.2019
40322	0,112	1	27.11.2019	9	05.12.2019
46298	0,107	1	19.11.2019	9	28.11.2019
478742	0,186	2	10.12.2019	11	20.12.2019
477557	0,077	1	12.12.2019	13	25.12.2019
43358	0,077	1	27.11.2019	13	10.12.2019

Abbildung 42. Berechnen der Lagerreichweite

Quelle: Eigene Darstellung

In Abbildung 42 wurden nun neue Daten und Informationen generiert, welche anschließend in das Datenmodell des Prototyps geladen werden. Eine weitere wichtige Information, die man aus dem Interview ableiten kann, ist der Unterschied zwischen saisonalen und regelmäßigen Artikelbestellungen. Da die Quartale im Datenmodell bereits erstellt wurden, kann man mit Hilfe eines Diagrammobjektes in QlikView die Anzahl der Artikel in den einzelnen Quartalen berechnen: *Count(Distinct Quartal)*.



Artikelnummer	Anzahl Quartal
10280	1
10281	1
10371	3
10373	2
10497	1
10512	1
10513	1
10514	1
10532	3
12376	2
12390	3
12432	1
12434	3

Abbildung 43. Vorkommen der Artikel in den Quartalen

Quelle: Eigene Darstellung

Betrachtet man die Grafik in Abbildung 43, so kann man daraus schließen, dass 861 und 259 Artikel nur in einem bzw. zwei Quartalen bestellt wurden. Alle anderen Artikel wurden innerhalb von drei oder vier Quartalen regelmäßiger bestellt. Man kann deshalb davon ausgehen, dass alle Artikel die in mehr als zwei Quartalen vorkommen, regelmäßiger bestellt werden. Somit berechnet man auch nur von diesen die Lagerreichweite. Die Artikelnummer sowie die Anzahl Quartale werden exportiert, um sie später im Datenmodell des Prototyps mittels Mapping als weiteres Attribut zuweisen zu können.

5.6.8 Datenmodell für Prototyp

Nachdem alle Daten aufbereitet und mit Attributen angereichert wurden, können diese nun im Datenmodell für den Prototypen zusammengefasst geladen werden.

- Bewegungsdaten + Stammdaten
- Regeln des Apriori-Algorithmus
- Lagerreichweite pro Artikel
- Artikelfrequenz Quartal

Das Generieren von Datenmodellen erfordert einige Schritte, um im weiteren Verlauf die Oberfläche von komplexen Berechnungen zu bewahren, die sich mitunter negativ auf die Performance auswirken können.

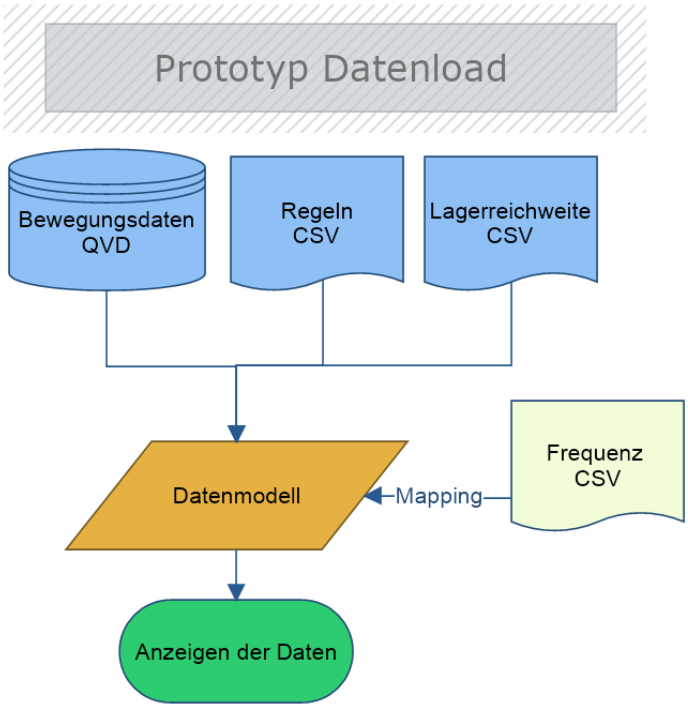


Abbildung 44. Datenload für den Prototyp

Quelle: Eigene Darstellung

Im ersten Schritt wird ein Attribut anhand der Quartale generiert. Ist die Frequenz höher als zwei und kommt der Artikel mindestens in drei Quartalen vor, so wird dieser als Trend- und alles was unter dem Wert liegt, als Saisonartikel klassifiziert:
if(Frequenz>2,'Trend','Saison') as Frequenz
Dieses Mapping wird anschließend mit dem Datenmodell verbunden, um dem Artikel ein neues Attribut „*Frequenz*“ zuzuweisen.
Als nächstes wird die prozentuelle Wahrscheinlichkeit der Trendartikel-Bestellungen anhand der zuvor berechneten Variablen des voraussichtlichen Bestelldatums errechnet. Dieses vorgehen wird im theoretischen Teil als Zeitreihenanalysen beschrieben. (Holland & Scharnbacher, 2010, S. 79-81) Diese ergibt sich, wie in Abbildung 42 ersichtlich, im Zeitverlauf vom Datum der letzten Bestellung bis zum Datum der voraussichtlichen Bestellung. Mit zunehmender Annäherung an das voraussichtliche Bestelldatum, erhöht sich auch die Wahrscheinlichkeit einer Kundenbestellung. Normalerweise ergibt sich die Basis zur Berechnung dieses Wertes anhand des aktuellen Tagesdatum zu dem sich der Kunde/die Kundin in den Webshop einloggt. Dieses Verhalten wird im Prototypen mit Hilfe einer Variabel simuliert, um die Unterschiede testen zu können. Steigt das simulierte Tagesdatum über das Voraussichtliche, so liegt die Wahrscheinlichkeit für diesen Teilwert bei 100%. Die gesamte Logik kann mit Hilfe von *If*-Anweisungen

und der in der Theorie beschriebenen Methode von Larose zur Erstellung von Entscheidungsbäumen, realisiert werden. (Larose, 2015, S. 318)

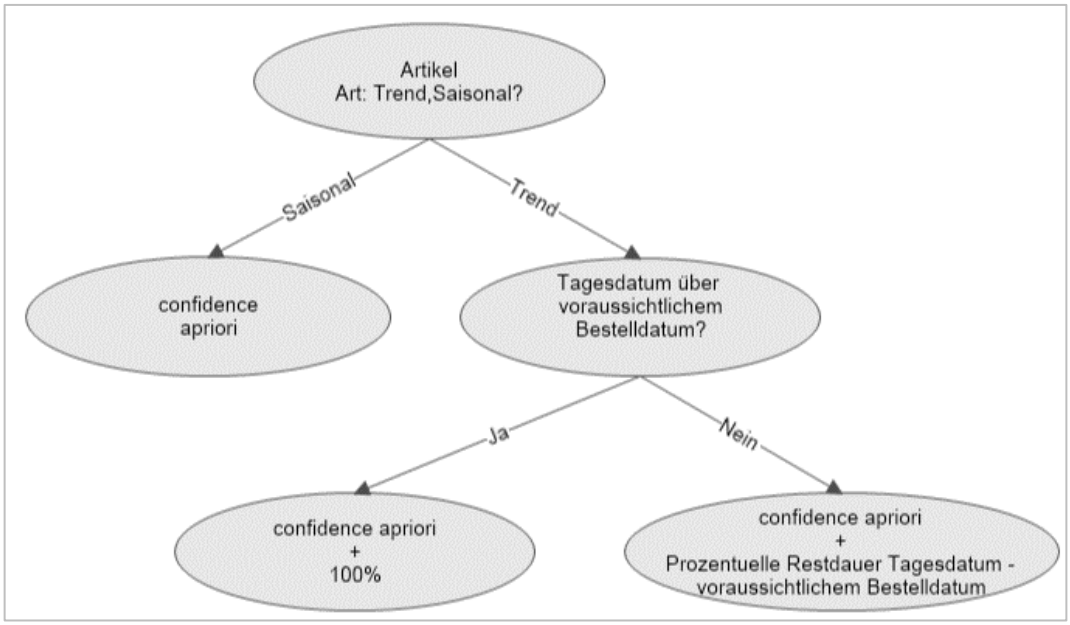


Abbildung 45. Entscheidungsbaum für die Bestellprognose

Quelle: Eigene Darstellung

Da es für die Prognose nur eine maximale Wahrscheinlichkeit von 100% geben kann, müssen bei Trendartikeln die beiden Werte gewichtet werden. Für die erste Iteration des Prototyps wurde eine 50/50 Gewichtung gewählt. In diesem Fall bedeutet das, sollte der *confidence* Wert bei 90% liegen und die Bestellwahrscheinlichkeit bei 80%, so würde sich eine Kaufwahrscheinlichkeit von 85% ergeben. (Provost & Fawcett, 2013, S. 237) Diese Gewichtung ist ein reiner Schätzwert und wird sich erst nach mehrmaligem Testen einpendeln. Ist das Datenmodell erstellt und einmal geladen, kann es bei neuen Daten entweder inkrementell oder komplett neu geladen werden.

5.6.9 Prototyp

Wie bereits unter Punkt 5.5 beschrieben, soll die fertige Lösung, pro eingegebenem Artikel vier Prognosen mit deren Wahrscheinlichkeiten anzeigen. Um den Prototypen mit unterschiedlichen Konstellationen zu testen, werden insgesamt drei Eingabefelder benötigt, die anhand der Abbildung 46 genauer erläutert werden.

Abbildung 46. Oberfläche des Prototyps

Quelle: Eigene Darstellung

1.) In diesem Feld wird die Kundennummer eingegeben. Im Webshop selbst wird das allerdings nicht mehr nötig sein, da eine direkte Zuordnung des/der Kunden/Kundin anhand vom Login möglich ist.
2.) Das Datum bildet die Basis für die Berechnung der Lagerreichweite. Auch dieses Feld ist ausschließlich für den Prototypen gedacht, da im Webshop einfach nur das aktuelle Tagesdatum als Basis herangezogen wird und die Prognosen täglich berechnet werden.
3.) Dieses Feld soll die Artikeleingabe des Benutzers/der Benutzerin im Warenkorb simulieren, nach Auswahl eines Artikels, werden darunter die Prognosen angezeigt.
4.) Mit diesem Button wird das Datenmodell neu generiert. Dies ist nur nötig, wenn sich die Kundennummer oder das Datum ändert.
5.) Der Prozentwert, gibt die Wahrscheinlichkeit an, wie sicher diese Prognose zutreffen wird. Je höher dieser Wert, desto wahrscheinlicher ihr eintreffen.
6.) Hier werden die vier prognostizierten Artikel in aufsteigender Reihenfolge angeordnet.
7.) Artikelgruppe der Artikelnummer

5.6.10 Auswertung der Ergebnisse

Zum Zeitpunkt der Erstellung dieser Arbeit, war es bereits möglich die Daten anhand von den im Jänner 2020 bestellten Warenkörben zu verifizieren. Als Beispiel wurden zwei Warenkörbe samt Artikeln herangezogen. Zur Überprüfung der Ergebnisse wurde jeweils ein Artikel vom Warenkorb im Prototyp eingegeben und verglichen, ob sich zumindest eine der vier Prognosen

im Warenkorb wiederfindet. Als Eingabeparameter wurde die Kundennummer, der Artikel sowie das Datum, an dem der Warenkorb bestellt wurde, verwendet. Für die Bewertung wurden drei Bewertungskriterien festgelegt, welche in Tabelle 4 in der Spalte Ergebnis eingetragen sind.

1. **Artikel nicht vorhanden (NV):** Der Artikel findet sich nicht im Prototypen wieder. Die Gründe hierfür können unterschiedlich sein, entweder wurde der Minimum *support* bzw. der Minimum *confidence* nicht erreicht, das heißt, es sind zu wenig Datenpunkte vorhanden, um eine Prognose abzuleiten, oder der Artikel wurde zum ersten Mal überhaupt von diesem Kunden/dieser Kundin bestellt.
2. **Artikel vorhanden (P„*Nummer*“):** Eine von vier Prognosen wurde gefunden. Treffen mehrere Prognosen zu, so wird diejenige mit der höchsten Wahrscheinlichkeit eingetragen.
3. **Prognose falsch (PF):** Eine oder mehrere Prognosen wurden gefunden, tauchen aber nicht im Warenkorb auf.

Tabelle 4: Warenkorb bestellt am 09.01.2020

Buchungsdatum	Artikel-nummer	ArtikelgruppenBez.	Menge	Einheit	Ergeb-nis
09.01.2020	14154	Rotguss Schraubfittings	20	STK	PF
09.01.2020	24937	Kupfer Rohre blank + isoliert	120	Meter	PF
09.01.2020	40952	Schallschutz Ablaufformstücke	30	Meter	P1
09.01.2020	41434	PE-Ablaufformstücke	20	STK	P1
09.01.2020	41437	PE-Ablaufformstücke	20	STK	P1
09.01.2020	41440	PE-Ablaufformstücke	10	STK	P2
09.01.2020	41455	PE-Ablaufformstücke	5	STK	P4
09.01.2020	41462	PE-Ablaufformstücke	10	STK	NV
09.01.2020	41463	PE-Ablaufformstücke	10	STK	NV
09.01.2020	41483	PE-Ablaufformstücke	24	STK	P1
09.01.2020	45350	Tempergussfittings schwarz	10	STK	NV
09.01.2020	45894	Tempergussfittings verzinkt	20	STK	NV
09.01.2020	75702	Schallschutz Ablaufformstücke	30	Meter	P3
09.01.2020	90929	Kupfer Rohre blank + isoliert	100	Meter	P1
09.01.2020	101643	Gelbarmaturen	1	STK	PF
09.01.2020	285734	PEX Formstücke	20	STK	P3
09.01.2020	285826	PEX-Rohre	40	Meter	PF
09.01.2020	512202	PEX-Rohre	100	Meter	P1

Quelle: Eigene Darstellung

Tabelle 5: Warenkorb bestellt am 22.01.2020

Buchungsdatum	Artikel-nummer	ArtikelgruppenBez.	Menge	Einheit	Ergebnis
22.01.2020	13513	Badezimmerarmaturen	4	STK	PF
22.01.2020	14159	Rotguss Schraubfittings	20	STK	P3
22.01.2020	14160	Rotguss Schraubfittings	20	STK	P1
22.01.2020	14164	Rotguss Schraubfittings	20	STK	P2
22.01.2020	28975	Kupfer Pressfittings	30	STK	NV
22.01.2020	29005	Kupfer Pressfittings	30	STK	PF
22.01.2020	40751	Ersatzteile Sanitär Install.	10	STK	PF
22.01.2020	46269	Rotguss Schraubfittings	20	STK	NV
22.01.2020	46274	Rotguss Schraubfittings	20	STK	NV
22.01.2020	46298	Rotguss Schraubfittings	10	STK	PF
22.01.2020	285734	PEX Formstücke	20	STK	P1
22.01.2020	41228	PE-Ablaufformstücke	20	STK	P2
22.01.2020	168372	Rotguss Schraubfittings	20	STK	NV
22.01.2020	264417	Regulier- und Mischventile	2	STK	NV

Quelle: Eigene Darstellung

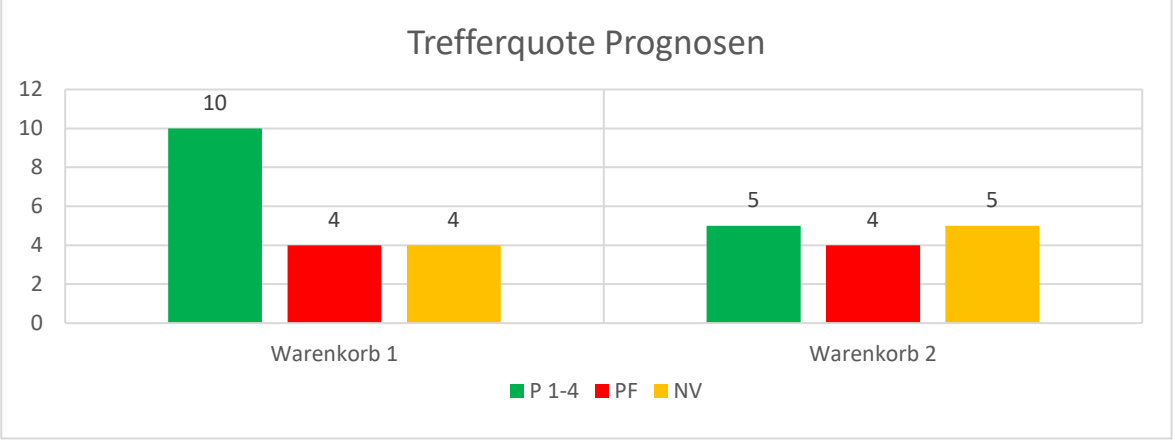


Abbildung 47. Trefferquote nach Auswertung der beiden Warenkörbe

Quelle: Eigene Darstellung

Nach Auswertung der beiden im Jänner angelegten Warenkörbe in der Abbildung 46 und unter der Berücksichtigung, dass es sich hierbei um die erste Iteration eines Prototyps handelt, zeichnet sich ein durchwegs positives Ergebnis ab. Von 32 getesteten Artikeln, stimmte die Prognose

insgesamt fünfzehn Mal mit zumindest einem anderen Artikel im Warenkorb überein. Bei jeweils acht Artikeln wurden zwar Prognosen gefunden, diese befanden sich aber nicht im Warenkorb, wurden somit falsch prognostiziert. Neun weitere Artikel wurden wegen mangelnder Datenbasis nicht im Prototypen angezeigt.

5.7 Beantwortung der Werkleitenden Subforschungsfrage

Die ursprüngliche Intention des Autors war es, die klassischen Methoden des Cross-Selling auf den B2B Bereich zu abstrahieren, um damit die gleichen Vorteile wie etwa Kundenbindung und den Verkauf von ergänzenden Produkten zur Umsatzsteigerung zu forcieren. (Malms & Schmitz, 2008, S. 30) Wie sich allerdings in den Experten und Expertinnen Interviews herausgestellt hat, sehen diese die Vorteile in ganz anderen Bereichen. Zum einen haben die Experten und Expertinnen großes Interesse daran, den Einkaufsprozess zu beschleunigen, damit sie sich auf den wertschöpfenden Teil ihrer Arbeit konzentrieren können und zum anderen kann es aufgrund der manuellen Lagerhaltung der Kunden/Kundinnen vorkommen, dass Artikel hin und wieder vergessen werden. Beide geschilderten Szenarien können vom Algorithmus zwar auch abgedeckt werden, entsprechen aber nicht mehr dem in der Literatur definierten Zweck des Cross-Selling. Bezogen auf die Datenbasis lässt sich festhalten, dass für den im Praxisteil angeführten Kunden die Basis als nicht ideal angesehen werden darf. Wie im Praxisteil angeführt, liegt der Wert, den das Programm zur Beurteilung der Datengrundlage ausgibt, unter dem Durchschnitt. (Hahsler & Reutterer, 2006, S. 15-16) Dieses Problem liegt primär an den vielen unterschiedlichen Artikeln und des zu geringen Transaktionsvolumens. Es kann deshalb davon ausgegangen werden, dass bei einer homogeneren Datenbasis, die Qualität der Ergebnisse noch verbessert wird. Aus Sicht eines Prototypen, dessen primäres Ziel es ist, die Ergebnisse so schnell wie möglich zu präsentieren (Hallmann, 1990, S. 11), spielt das noch keine große Rolle, in weiteren Iterationen sollte aber unbedingt auf diese Problematik eingegangen werden. Eine Verbesserung könnte mit Hilfe von einer Erweiterung der Datenbasis um weitere Jahre oder durch die Verknüpfung von Kunden/Kundinnen mit ähnlichen Einkaufsverhalten erzielt werden. Wie in der Theorie beschrieben, gibt es keine vordefinierte Regel oder Vorgehensweise zur Erstellung einer Prognose. (Beierle & Kern-Isberner, 2019, S. 1) Die Auswahl der Methoden für die Erstellung der Prognosen, basierten im Wesentlichen auf den Anforderungen der Experten und Expertinnen, sowie dem erlernten theoretischen Wissen des Autors. Die Ergebnisse zeigen, dass ungefähr die Hälfte der überprüften 32 Artikelpositionen richtig prognostiziert wurde. Der Rest wurde entweder nicht gefunden oder lieferte falsche Ergebnisse. Auch wenn die erste Bewertung des Prototyps aus Sicht des Autors als zufriedenstellend betrachtet

wird, so bedarf es hier noch einiger Optimierungen und weiterer Tests, um das Ergebnis seriös beurteilen zu können. Abschließend kann man festhalten, dass Prognosen aus einer Kombination von Experten/Expertinnen Wissen und Predictive Analytics Methoden durchaus funktionieren. In diesem Fall war es sogar ratsamer das Wissen der Experten/Expertinnen einfließen zu lassen, da neue Erkenntnisse geschaffen wurden, welche so in der Literatur nicht vorhanden waren.

6 Conclusio und Ausblick

Ziel dieser Arbeit war es, anhand von Leitfadeninterviews zu ermitteln, welchen Bedarf die Zielgruppe aus der SHK-Branche an einem Cross-Selling-Algorithmus hat und welche Parameter entscheidend für den Erfolg eines solchen Prognosesystems sind. Der theoretische Teil wurde auf Basis einschlägiger Forschungs- und Fachliteratur aus dem Bereich Data Science erstellt. Der daraus entstandene Prototyp, soll dabei einen ersten Einblick in die Möglichkeiten der Technologie geben und als Grundlage für weiterführende Forschungen dienen. In diesem abschließenden Kapitel werden die Ergebnisse dieser Arbeit mit der Beantwortung der Hauptforschungsfrage verdeutlicht.

6.1 Relevanz der Arbeit

Die Ergebnisse dieser Arbeit sind für SHK-Großhandelsbetriebe relevant, die unter Einsatz von Predictive Analytics Cross-Selling-Funktionen in ihren Webshop integrieren wollen. Wie in der Theorie erforscht wurde, bietet Predictive Analytics mehrere Möglichkeiten die Kundenwünsche noch besser abzudecken. Für die Umsetzung bedarf es allerdings Personen, welche über das nötige Analyse- und Statistik-Know-how sowie IT-Wissen verfügen.

6.2 Beantwortung der Hauptforschungsfrage

Die fortschreitende Digitalisierung der Wirtschaft und Gesellschaft macht sich auch in der SHK-Branche bemerkbar. Während Handwerksbetriebe noch relativ zurückhaltend mit der Verwendung von E-Commerce Services sind, wird sich dies in den nächsten Jahren drastisch ändern. Dessen sind sich auch die interviewten Experten und Expertinnen bewusst. Auch der SHK-Großhandel muss sich auf diese Entwicklung vorbereiten, um nicht zu riskieren von den großen Versandhändlern aus dem Markt gedrängt zu werden. Umso wichtiger ist es für Großhandelsunternehmen diesen Trend frühzeitig zu nutzen, um dem/der Kunden/Kundin zusätzliche Online Services anzubieten, welche die Kundenbindung erhöhen. Ein solches Werkzeug, welches genau diesen Zweck erfüllen kann und sich bereits im Einzelhandel bewährt hat, ist die Cross-Selling Methode. Wie sich im Zuge der Recherche herausgestellt hat, ist diese Methode im B2B Bereich noch weitgehend unerforscht, insbesondere mangelte es an Erkenntnissen über die Einflussfaktoren für erfolgreiches Cross-Selling. Deshalb musste für die Umsetzung dieser Methode, auf das Wissen von Experten und Expertinnen aus der SHK-Branche zurückgegriffen werden. Dabei konnten nicht nur Einblicke in den Einkaufsprozess der Betriebe erlangt werden, sondern es wurde auch Wissen generiert, welches nötig war, um den Prototypen zu realisieren. Die durchgeführten Interviews haben außerdem Aufschluss darüber

gegeben, dass anders als beim klassischen Cross-Selling, die Vorteile nicht in der zusätzlichen Information über ein ergänzendes oder ähnliches Produkt eine Rolle spielen, sondern die Möglichkeit den Einkaufsprozess zu beschleunigen. Auch ist es für die Experten und Expertinnen aufgrund ihrer teils manuellen Lagerhaltung interessant, über eventuell auslaufendes Material informiert zu werden, welches gegebenenfalls vergessen wurde. Beide Anforderungen können zwar auch von dieser Methode abgedeckt werden, entsprechen aber nicht mehr dem klassischen Zweck des Cross-Selling. Eine Gemeinsamkeit hat sich dann aber sowohl in der empirischen Forschung als auch in der theoretischen Auseinandersetzung ergeben. Wenn so eine Lösung gut umgesetzt ist und sich im Einsatz als praktikabel oder zeitsparend erweist, erhöht sich die Chance der Kundenbindung. Außerdem hat die Untersuchung gezeigt, dass sich die Herleitung solcher Prognosen für das Cross-Selling in den letzten Jahren dank der fortschreitenden Technologie stark geändert hat. Während solche Vorschläge früher noch oft auf menschlichen Erfahrungen basierten, wäre das heutzutage aufgrund der enormen Datenmengen und dem dynamischen Umfeld kaum mehr möglich. Die Anwendungsbeispiele aus der Theorie verdeutlichen dabei, dass große Unternehmen und Konzerne seit Jahren bereits erfolgreich an verschiedenen Methoden experimentieren, um mit Hilfe von Predictive Analytics ihre Geschäftsprozesse zu verbessern und zu optimieren. Nicht immer laufen dabei solche Projekte problemlos ab, oftmals scheitern diese an fehlendem Vertrauen durch das Management oder wegen schlechter und inkonsistenter Daten. So hat es sich auch im Laufe dieser Arbeit ergeben, dass die Datengrundlage für den Prototypen bei diesem ausgewählten Kunden aufgrund der Streuung nicht ideal ist. Als besondere Herausforderung in der Entwicklung von Algorithmen hat sich die effektive Gestaltung des Projektablaufs gezeigt. Als Grundlage für die Entwicklung des Prototyps in diesem Werk, wurde deshalb die KDD-Methodik gewählt, welche sich speziell für Erstellung eines nicht trivialen Prozesses eignet. Die KDD-Methodik liefert auch Antworten auf die Frage, welche Schritte notwendig sind, um aus Rohdaten Wissen zu generieren. Eine weitere Erkenntnis, die sich aus der Literatur ergeben hat, betrifft die Auswahl der Data-Mining-Methoden zur Erstellung von Prognosen. Es wurde dabei festgestellt, dass es keine allgemeingültigen Regeln für die Erstellung von Prognosen gibt, sondern immer individuell auf Basis des vorliegenden Problems die geeigneten Methoden ausgewählt werden sollten. Einige Algorithmen wie der Apriori wurden dabei speziell für den Handel entworfen. Für den Prototypen wurden verschiedene Methoden aus dem Data-Mining wie Assoziations-, Zeitreihen-, Klassifizierungsanalysen und Statistik kombiniert, um die Anforderungen der Zielgruppe umzusetzen. In diesem Zusammenhang bleibt auch die Frage offen, ob es sich bei der Auswahl der Data-Mining-Methoden

tatsächlich um die besten und effektivsten handelt, um die Aufgabe zu lösen. Bei der Umsetzung war es auch entscheidend, die definierten Parameter der Experten und Expertinnen aus der empirischen Forschung zu beachten, um keine ungewollten Ergebnisse zu erzeugen. Die Ergebnisse vom Prototypen zeigen, dass der Entwurf als durchaus vielversprechend angesehen werden darf. Nach einer Auswertung der Ergebnisse anhand von Warenkörben, die nach der Entwicklung des Prototypen erstellt wurden, lag die Treffergenauigkeit bei fast 50%. Um diesen Wert in weiterer Folge zu verbessern, müsste analysiert werden, worin die Gründe liegen, warum andere Artikel nicht korrekt oder gar nicht prognostiziert wurden. In Anbetracht der Ergebnisse und der Erkenntnisse in dieser Arbeit, lässt sich die Hauptforschungsfrage deshalb nur teilweise beantworten. Während bei der Programmierung die Anforderungen der Zielgruppe eingehalten wurden, ist es fraglich, ob die Gestaltung tatsächlich so für den SHK-Großhandel funktionieren würde. Es müssten weitere Tests durchgeführt und viel praktische Erfahrung gesammelt werden, um die Hauptforschungsfrage eindeutig beantworten zu können. Offen bleibt auch die Frage, wie sich der fertige Prototyp unter realen Bedingungen verhält und inwieweit der Service tatsächlich von den Kunden und Kundinnen in Anspruch genommen werden würde.

6.3 Limitationen

Die vorliegende Arbeit weist eine Anzahl von Limitierungen auf, die bei möglichen Schlussfolgerungen zu bedenken sind, sie soll vor allem erste richtungsweisende Erkenntnisse und Überlegungen liefern, für einen Bereich, der bisher wenig untersucht wurde. Da nur Experten und Expertinnen im Raum Westösterreich und Ostschweiz interviewt wurden, ist die Reichweite dieser Untersuchung begrenzt, und die in den Betrieben festgestellten Prozesse und Empfehlungen nicht zwangsläufig auf andere Regionen übertragbar oder verallgemeinerbar. Aufgrund der Komplexität und dem Umfang des Themas, wurde bei der Erstellung des Prototyps nur auf einen Bruchteil der Möglichkeiten aus dem Bereich des Predictive Analytics und des Data-Mining zurückgegriffen. Insbesondere die Auswahl der Methoden und Algorithmen, würde hier noch wesentlich mehr Spielraum für weitere Untersuchungen bieten. Weiterführende Forschungen würden diesbezüglich noch aussagekräftigere Ergebnisse liefern.

6.4 Ausblick

Die Digitalisierung ist eine Veränderungsdynamik, die unsere gesamte Wirtschaft betrifft, deshalb kann man davon ausgehen, dass das Thema Predictive Analytics in den nächsten Jahren noch mehr an Bedeutung gewinnen wird. Auch Cross-Selling hat sich in der Vergangenheit als

nützliches Instrument für die Kundenbindung erwiesen. Werden beide Methoden kombiniert, bietet sich die Möglichkeit, ein leistungsstarkes Werkzeug zu erschaffen, welches das Potenzial hat, die Position am Markt weiter zu stärken. Während die letzten Jahre hauptsächlich große Konzerne und Unternehmen die Vorteile von Predictive Analytics nutzten und es bereits einige Anwendungsbeispiele gibt, welche das Potenzial dieser Technologie aufzeigen, werden mit der Zeit wohl auch immer mehr kleinere und mittlere Unternehmen den Wert und die Chancen einer solchen Technologie erkennen. Der Prototyp zeigt dabei, wie schon mit einfachen Mitteln ein aussichtsreiches Ergebnis erzielt werden kann. Die Einbeziehung von Experten und Expertinnen ist dabei genauso wichtig, wie die richtige Wahl der Analysemethoden. Zukünftige Forschungen, könnten sich dabei mit weiteren Einflussfaktoren oder Methoden zur Prognose von Warenbestellungen im SHK-Großhandel beschäftigen. Auch der Einsatz in einem Webshop, sowie die daraus folgende Akzeptanz des Service könnte untersucht werden. Möglicherweise lässt sich das Werk auch als Basis für andere Branchen im B2B Bereich anwenden.

7 Literaturverzeichnis

Abbott, D. (2014). *Applied Predictive Analytics - Principles and Techniques for the Professional Data Analyst.* Indianapolis, USA: John Wiley Sons. Inc.

Alharan, A., Al-Haboobi, A., & Alsagheer, R. (Juni 2017). Popular Decision Tree Algorithms of Data Mining Techniques: A Review. *International Journal of Computer Science and Mobile Computing Vol. 6, Issue 6 ISSN: 2320–088X*, S. 133-142.

Aurier, P., & N'Goala, G. (September 2009). The Differing and Mediating Roles of Trust and Relationship Commitment in Service Relationship Maintenance and Development. *Journal of the Academy of Marketing Science Vol. 38, Issue 3*, S. 303-325. doi:10.1007/s11747-009-0163-z

Bankhofer, U., & Vogel, J. (2008). *Datenanalyse und Statistik - Eine Einführung für Ökonomen.* Wiesbaden, Deutschland: Gabler Verlag.

BauInfoConsult GmbH. (2018). *Studie: Zukunftsmarkt Online-Handel am Bau.* Düsseldorf.

Beierle, C., & Kern-Isberner, G. (2019). *Methoden wissensbasierter Systeme - Grundlagen, Algorithmen, Anwendungen* (Bde. 6., überarbeitete Auflage). Wiesbaden, Deutschland: Springer Fachmedien GmbH. doi:10.1007/978-3-658-27084-1

Bentz, S. (1984). Kennzahlensysteme zur Erfolgskontrolle des Verkaufs und der Marketing-Logistik: Entwicklung und Anwendung in der Konsumgüterindustrie. *Schriften zu Marketing und Management, No. 8*. doi:10.3726/b13616

Bibel, W., Hölldobler, S., & Schaub, T. (1993). Wissensreprasentation und Inferenz: Eine grundlegende Einführung. Braunschweig/Wiesbaden, Deutschland: Vieweg & Sohn Verlagsgesellschaft. doi:10.1007/978-3-322-86814-5

Binckebanck, L., & Elste, R. (2016). *Digitalisierung im Vertrieb - Strategien zum Einsatz neuer Technologien in Vertriebsorganisationen.* Wiesbaden: Springer Fachmedien. doi:10.1007/978-3-658-05054-2

Bohanec, M., Borštnar, M., & Šikonja, M. R. (April 2017). Explaining machine learning models in sales predictions. *Expert Systems With Applications Vol. 71*. doi:10.1016/j.eswa.2016.11.010

Boire, R. (September 2017). B2B predictive analytics: an untapped sector. *MOJ Proteomics & Bioinformatics Vol. 6, Issue 1*. doi:10.15406/mojpb.2017.06.00185

Brachman, R. J., & Levesque, H. J. (2004). Knowledge representation and reasoning. California, San Francisco, USA: Morgan Kaufmann Publishers.

Brühl, V. (2019). Big Data, Data Mining, Machine Learning und Predictive Analytics: Ein konzeptioneller Überblick. (G. U. Center for Financial Studies (CFS), Hrsg.) *CFS Working Paper Series, No. 617*.

Burow, L., Gerards, Y., & Demmer, M. (September 2017). Effektiv und effizient steuern mit Predictive Analytics. *Controlling & Management Review Ausgabe 9 ISSN: 2195-8262*, S. 48-56.

Chen, H.-c., Chiang, R., & Storey, V. C. (Dezember 2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly Vol. 36, Issue 4*. doi:10.2307/41703503

Chen, Y.-L., & Tung, C.-W. (Februar 2006). A data mining approach for retail knowledge discovery with consideration of the effect of shelf-space adjacency on sales. (D. o. Management, Hrsg.) *Decision Support Systems Vol. 42 Issue 3*. doi:10.1016/j.dss.2005.12.004

Christ, O., & Ebert, N. (23. Dezember 2015). Predictive Analytics im Human Capital. *HMD Praxis der Wirtschaftsinformatik*. doi:10.1365/s40702-015-0193-6

Ester, M., & Sander, J. (2000). *Knowledge Discovery in Databases - Techniken und Anwendungen.* Heidelberg, Berlin, Deutschland: Springer Verlag. doi:10.1007/978-3-642-58331-5

Fayyad, U., Piatetsky-Shapiro, G., & Padhraic , S. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine Vol 17 Nr 3*. doi:10.1609/aimag.v17i3.1230

Gandomi, A., & Haider, M. (2. April 2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, Volume 35, Issue 2, April 2015*. doi:10.1016/j.ijinfomgt.2014.10.007

Göpfert, T., & Breiter, A. (2015). *Informatik 2015 - Knowledge Discovery in Big Data: Herausforderungen durch Big Data im Prozess der Wissensgewinnung am Beispiel des CRISP-DM.* Bonn: Gesellschaft für Informatik e.V.

Gottlob, G., Frühwirth, T., & Horn, W. (1990). *Expertensysteme.* Wien: Springer-Verlag. doi:10.1007/978-3-7091-9094-4

Hahsler, M., & Reutterer, T. (2006). Warenkorbanalyse mit Hilfe der Statistik-Software R. In P. Schnedlitz, *Innovationen in Marketing und Handel* (S. 144-163). Linde-Verlag Ges.m.b.H.

Hallmann, M. (1990). *Prototyping komplexer Softwaresysteme - Ansätze zum Prototyping und Vorschlag einer Vorgehensweise.* Vieweg & Teubner Verlag. doi:10.1007/978-3-663-01528-4

Harrington, J. L. (2010). *SQL Clearly Explained (Third Edition).* Burlington: Morgan Kaufmann. doi:10.1016/B978-0-12-375697-8.50003-0

Hartshorne, C., & Weiss, P. (1931). The Collected Papers of Charles Sanders Peirce Vol. 2. Cambridge, Massachusetts , USA: Harvard University Press.

Holland, H., & Scharnbacher, K. (2010). *Grundlage der Statisik: Datenerfassung und -darstellung Maßzahlen, Indexzahlen, Zeitreihenanalyse 8. Aufl.* Wiesbaden, Deutschland: Gabler GWV Fachverlage GmbH.

Iffert, L. (2016). Predictive Analytics richtig einsetzen. In *Controlling & Management Review Sonderheft 1*. Würzburg, Deutschland.

Inman, J. J., & Nikolova, H. (März 2017). Shopper-Facing Retail Technology: A Retailer Adoption Decision Framework Incorporating Shopper Attitudes and Privacy Concerns. *Journal of Retailing Vol. 93, Issue1*. doi:10.1016/j.jretai.2016.12.006

Inmon, W., Linstedt, D., & Levins, M. (2019). *Data Architecture (Second Edition) - A Primer for Data Scientist.* doi:10.1016/B978-0-12-816916-2.00021-8

KMU Forschung Austria. (2018). *Studie: Zukunft des Großhandels.* Im Auftrag der Wirtschaftskammer Wien.

Knott, A., Hayes, A., & Scott, N. A. (Juni 2002). Next-product-to-buy models for cross-selling applications. *Journal of Interactive Marketing Vol. 16, Issue 3*, S. 59-75. doi:10.1002/dir.10038

Kruse, R., Borgelt, C., Braune, C., Klawonn, C., Moewes, C., & Steinbrecher, M. (2015). *Computational Intelligence - Eine methodische Einführung in Künstliche Neuronale Netze, Evolutionäre Algorithmen, Fuzzy-Systeme und Bayes-Netze 2., überarbeitete und erweiterte Auflage.* Wiesbaden: Springer Vieweg. doi:10.1007/978-3-658-10904-2_2

Kwon, O., Lee, N., & Shin, B. (Juni 2014). Data quality management, data usage experience and acquisition intention of big data analytics. *International Journal of Information Management 34, Issue 3*. doi:10.1016/j.ijinfomgt.2014.02.002

LaPlaca, P. J., & Katrichis, J. M. (März 2009). Relative Presence of Business-to-Business Research in the Marketing Literature. *Journal of Business-to-Business Marketing Vol. 16, Issue 1-2*. doi:10.1080/10517120802484213

Larose, C. D. (2015). *Data Mining and Predictive Analytics Second Edition.* Hoboken, New Jersey, USA: John Wiley & Sons, Inc.

Lee, J.-G., & Kang, M. (2015). Geospatial Big Data: Challenges and Opportunities. *Big Data Research Vol.2, Issue 2*. doi:10.1016/j.bdr.2015.01.003

Lersch, K. M., & Chakraborty, J. (2020). *Geographies of Behavioural Health, Crime, and Disorder*. Springer Nature Switzerland AG. doi:10.1007/978-3-030-33467-3_5

Lilien, G. L. (Februar 2016). The B2B Knowledge Gap. *International Journal of Research in Marketing Vol. 33*. doi:10.1016/j.ijresmar.2016.01.003

Lokanatha, R. C., & Venkatadri, M. (Februar 2011). A Review on Data mining from Past to the Future. *International Journal of Computer Applications Vol. 15 No. 7*. doi:10.5120/1961-2623

Loos, P., Lechtenböger, J., Vossen, G., Zeier, A., Krüger, J., Müller, J., . . . Winter, R. (2011). In-Memory Datenmanagement in betrieblichen Anwendungssystemen. *Wirtschaftsinformatik 53 Heft 6*. doi:10.1007/s12599-011-0188-y

Maitzen, P. (2016). *Attraktivität von CrossSelling-Angeboten aus Kundensicht.* Stuttgart, Deutschland: Springer Fachmedien. doi:10.1007/978-3-658-11647-7

Malms, O., & Schmitz, C. (18. Juni 2008). Cross-Selling-Potenziale - Nachhaltiges Wachstum realisieren. *Marketing Review St. Gallen 25*. doi:10.1365/s11621-008-0162-3

Matt, C. (August 2012). In-Memory-Technologien für Unternehmensanwendungen. *Controlling & Management 56, Issue 4*. doi:10.1365/s12176-012-0394-6

Mayring, P. (2015). *Qualitative Inhaltsanalyse: Grundlagen und Techniken, 12. Aufl.* Weinheim: Beltz.

McCarthy, M. V., Halawi, L., & Ceccucci, W. (2019). *Applying Predictive Analytics - Finding Value in Data.* Cham, Schweiz: Springer Nature Switzerland AG. doi:10.1007/978-3-030-14038-0

Mishra, N., & Silakari, D. (2012). Predictive Analytics: Trends, Applications, Oppurtunities. *International Journal of Computer Science and Information Technologies, Vol. 3 ISSN: 0975-9646*.

Pérez-Martin, A., Pérez-Torregrosa, A., & Vaca, M. (August 2018). Big Data techniques to measure credit banking risk in home equity loans. (D. o. Studies, Hrsg.) *Journal of Business Research Volume 89*. doi:10.1016/j.jbusres.2018.02.008

Prinzie, A., & Van den Poel, D. (Januar 2006). Investigating purchasing-sequence patterns for financial services using Markov, MTD and MTDg models. *European Journal of Operational Research Vol. 170 No. 3*. doi:10.1016/j.ejor.2004.05.004

Provost, F., & Fawcett, T. (2013). *Data Science for Business.* California, USA: O'Reilly Media, Inc.

Reid, D. A., & Plank, R. E. (Juni 2000). Business marketing comes of age: A Comprehensive review of the literature. *Journal of Business-to-Business Marketing Volume 7, Issue 2-3*, S. 9-186. doi:10.1300/J033v07n02_02

Schäfer, H. (2002). *Die Erschließung von Kundenpotentialen durch Cross-Selling.* Wiesbaden, Deutschland: Springer Fachmedien GmbH. doi:10.1007/978-3-663-09693-1

Shah, D., Kumar, V., Qu, Y., & Chen, S. (Mai 2012). Unprofitable Cross-Buying: Evidence from Consumer and Business Markets Vol. 76. (A. M. Association, Hrsg.) *Journal of Marketing*. doi:10.1509/jm.10.0445

Sheldon, R. M. (2010). *A first course in probability 8th ed.* New Jersey: Pearson Education Inc.

Shmueli, G., & Koppius, O. R. (Sept. 2011). Predictive Analytics in Information Systems Research. *MIS Quarterly Vol. 35 No. 3 pp. 553-572*. doi:10.2139/ssrn.1606674

Vieira, S., Hugo, W., Pinaya, L., & Andrea, M. (2020). *Machine Learning Methods and Applications to Brain Disorders.* London: Elsevier Inc. doi:10.1016/B978-0-12-815739-8.00001-8

Wagner, K. (Februar 2008). Cross-Selling: Offering the Right Product to the Right Customer at the Right Time. *Journal of Relationship Marketing*. doi:10.1300/J366v06n03_03

White, T. B. (2004). Consumer Disclosure and Disclosure Avoidance: A Motivational Framework. *Journal of Consumer Psychology Vol. 14, Issue 1-2*. doi:10.1207/s15327663jcp1401&2_6

Yang, L., Liu, S., Tsoka, S., & Papageorgiou, L. G. (Februar 2016). Mathematical programming for piecewise linear regression analysis. (E. Ltd., Hrsg.) *Expert Systems With Applications 44*. doi:10.1016/j.eswa.2015.08.034

7.1 Internetquellen

Ecker, M. (2019, September). *jö Bonus Club*. Retrieved April 10, 2020, from Verein für Konsumenteninformation: https://blog.vki.at/article/jö-bonus-club

King, R. (2012, Oktober). *SAP Precision Retailing could influence last-minute customer decisions*. Retrieved April 01, 2020, from zdnet: https://www.zdnet.com/article/sap-precision-retailing-could-influence-last-minute-customer-decisions/

Power, M. D. (2015, Mai). Sharing and Analyzing Data to Reduce Insurance. *Association for Information Systems*. Retrieved April 01, 2020, from http://aisel.aisnet.org/mwais2015/19

QlikTech. (2017). *What is QlikView*. Retrieved März 26, 2020, from https://help.qlik.com/de-DE/qlikview/November2017/Content/what-is.htm

Robinson, D., Harlan, Y., & Rieke, A. (2014, September). *Civil Rights, Big Data, and Our Algorithmic Future*. Retrieved März 12, 2020, from https://bigdata.fairness.io/introduction

RStudio . (2020). *About RStudio*. Retrieved März 26, 2020, from https://rstudio.com/about/

Samuels, M. (2017). *Big data case study: How UPS is using analytics to improve performance*. Retrieved April 01, 2020, from zdnet: https://www.zdnet.com/article/big-data-case-study-how-ups-is-using-analytics-to-improve-performance/

8 Anhang

Frage	Fragestellung	Kategorie
	Kategorie 1: Einleitung	
1.1	Beschreiben Sie bitte kurz ihren Betrieb, einfach das wichtigste was es ihrer Meinung nach zu erzählen gibt.	1
1.2	Erzählen Sie uns, wer Sie sind und welche Tätigkeit Sie in diesem Betrieb ausführen?	1
1.3	Wie ist in Ihrem Betrieb der Einkauf organisiert und wer ist dafür zuständig?	1
	Kategorie 2: Webshop Fragen	
2.1	Welche Motivation steckt hinter der Nutzung des Webshops?	2
2.2	Wie sehen Sie die Zukunft im Zusammenhang mit ihrer Branche und dem Online Einkauf?	2
2.3	Welche Vor- bzw. Nachteile entstehen ihrer Meinung nach durch die Nutzung von Webshops?	2
2.4	Was führt dazu, dass Sie keine Online Bestellungen tätigen und auf klassische Bestellverfahren zurückgreifen?	2
2.5	Wie ist ihre Vorgehensweise der Bestellung von Artikeln im Webshop?	2
	Kategorie 3: Usability	
3.1	Welche Erfahrungen hatten Sie mit Cross-Selling Systemen und kennen sie Beispiele wo diese Methode verwendet wurde?	3
3.2	Wie schätzen Sie die Praktikabilität eines solchen Cross-Selling Algorithmus ein?	3
3.3	Welche Auswirkungen würde so ein Feature Ihrer Meinung nach auf die Bestelldauer haben?	3
3.4	Inwieweit würde Ihnen ein vorgeschlagener weiterführender Artikel bei der Erfassung des Warenkorbs einen Vorteil bringen?	3
3.5	Was meinen Sie, wie wirkt sich der Einsatz einer solchen Technologie auf die Akzeptanz eines Webshops aus?	3
	Kategorie 4: Algorithmus	
4.1	Wie häufig bestellen Sie Artikel, die im Verarbeitungsprozess von einem anderen Artikel abhängig sind?	4
4.2	Gibt es Situationen, wo Sie bestimmte Hersteller bevorzugen obwohl es alternative oder günstigere Produkte geben würde und was könnten mögliche Gründe dafür sein?	4

4.3	Würde man Ihnen während dem Bestellvorgang eines Artikels einen weiterführenden Artikel anhand der Daten Ihrer bisher getätigten Einkäufe vorschlagen (Warenkorbanalyse), glauben Sie, dass diese Prognose zutreffen würde oder bestellen Sie öfter Artikel, die sie nicht regelmäßig benötigen?	4
4.4	Was wären Ihrer Meinung nach wichtige Daten und Faktoren, die man bei einem weiterführenden Artikel berücksichtigen muss, um die Qualität der Prognose zu erhöhen?	4

Anhang 1: Interviewleitfaden

Quelle: Eigene Darstellung

ID	Proband	Frage Nr.	Seite	Nr.	Paraphrase	Generalisierung	Reduktion
7	A	2.4	2	59-66	[...] noch nicht alle Produkte drinnen [...] Aber das ist eher selten der Fall, man telefoniert eher nicht mehr miteinander.	Wenn Informationen fehlen oder bei unbekannten Artikeln	K2: Der Umweg über klassische Bestellverfahren (Telefon, E-Mail, ...) erfolgt hauptsächlich wegen fehlender Informationen oder unbekannten Produkten, wo noch eine persönliche Beratung erforderlich ist.
7	B	2.4	2	88-92	Es gibt sicher immer Ausnahmen wo man anrufen muss, Sondergrößen etc. […] Viele Ersatzteile werden natürlich nicht online bestellt, sondern direkt ein Bild geschickt + Info und dann wird das passende Teil herausgesucht	~~Wenn Informationen fehlen oder bei unbekannten Artikeln~~	
7	C	2.4	1	48-49	Spezialfälle, wo man mit jemandem reden muss […] Wo man Beratung braucht […] man z.B. ein Foto schicken kann […]	~~Wenn Informationen fehlen oder bei unbekannten Artikeln~~	

7	D	2.4	1	38-41	Wenn man Informationen dazu braucht, die im Webshop nicht ersichtlich sind. Oder auch bei Preisverhandlungen, im Onlineshop sind die Preise ja fix.	~~Wenn Informationen fehlen~~ oder bei Preisverhandlungen.	
7	E	2.4	1	38-41	[…] neue Ware nicht sicher wie das verbaut wird, dann ist es schon gut, wenn man jemanden vom Fach anrufen kann.	Wenn zusätzliche Informationen zum Produkt benötigt werden.	

Anhang 2: Qualitative Inhaltsanalyse am Beispiel der Frage 2.4

Quelle: Eigene Darstellung