Datenqualität im Data Mining

Christian Gottermeier
Ruprecht-Karls-Universität Heidelberg
Fakultät für Wirtschafts- und Sozialwissenschaften
Lehrstuhl für Wirtschaftsinformatik
69117 Heidelberg
Christian.Gottermeier@AWI.Uni-Heidelberg.de

Zusammenfassung

Datenqualität stellt im Data Mining einen wesentlichen Erfolgsfaktor dar. Schätzungen zu folge müssen 70 bis 80 Prozent des allgemeinen Projektaufwands für die Aufbereitung der Daten veranschlagt werden. In dieser Arbeit werden die Anforderungen aufgelistet, die die Knowledge Discovery in Databases an die Daten stellt und welche Systemarchitekturen hierbei unterstützend wirken. Des Weiteren werden die potentiell auftretenden Probleme skizziert, sowie die Maßnahmen des Data Quality Managements und des Pre-Processings zur Steigerung der Datenqualität erläutert. Beide Ansätze werden in eine Methodologie integriert, die eine systematische Vorgehensweise bei der Datenbereinigung ermöglicht. Im Mittelpunkt dieses Beitrages steht das Pre-Processing. Die Handhabung fehlender Werte, der Umgang mit Ausreißern, falsche bzw. fehlerhafte Skalierungen und ungünstige Verteilungen sowie eine ausreichend große Datenbasis werden in diesem Artikel ebenso behandelt, wie die Auswirkungen mangelnder Datenqualität auf die Verfahren des Data Mining, insbesondere der Vorhersage- und Klassifikationsmodelle. Abschließend wird noch eine Einführung in das Data Quality Mining gegeben. Dieser Ansatz stellt gewissermaßen einen Paradigmenwechsel dar. Die für Data Mining-Analysen geforderte Datenqualität lässt sich auch durch typische Data Mining-Verfahren erzeugen. Prognosemodelle sowie Cluster- und Assoziationsanalysen können bei der Verbesserung einen wertvollen Beitrag leisten.

Keywords: Data Mining, Datenqualität, Pre-Processing, Data Quality Mining.

1 Einleitung

Der Begriff Data Mining steht für eine Forschungs- und Anwendungsrichtung deren Bestandteile schon existieren: [18] Verfahren der klassischen statistischen Datenanalyse, Anwendungen aus der künstlichen Intelligenz, der Mustererkennung und des maschinellen Lernens wurden in die sog. Knowledge Discovery in Databases (KDD) integriert. Mit dem Namen „KDD“ wird zudem eine starke Verbindung zur Datenbanktechnologie hergestellt.

Die Begriffe „Data Mining“ und „KDD“ werden in der Literatur häufig simultan verwendet. [20] Entgegen dieser Gleichstellung der Begriffe existiert allerdings auch die Sichtweise, KDD als den Gesamtprozess der Analyse zu betrachten. Die hierfür vorgesehenen Schritte, d.h. von der Planung über die eigentliche Analyse bis zur Implementierung der Ergebnisse, beinhalten dementsprechend auch das Data Mining mit der Modellierung der Verfahren. Mit der SEMMA-Methode[1] der SAS Institute Inc. {z.B.: [2]} existiert auch ein Modell, das den Ablauf einer Data Mining-Analyse vorgibt und abbildet, und so bei der Gestaltung des Prozessflusses ein systematisches Vorgehen ermöglicht.

Die Aufgabe des Data Mining ist die Entwicklung, Implementierung und Ausführung von Datenanalysemethoden. Ein Fokus wird dabei auf sehr große Datensätze mit komplexen Strukturen gelegt.

Die Hauptaufgaben sind:

I. Vorhersage- und Klassifikationsmodelle
II. Segmentierungen oder Clusterung
III. Assoziationsanalysen
IV. Dimensionsreduktion
V. Beschreibung von Strukturen

Vorhersage und Klassifikation unterscheiden sich hauptsächlich durch die Skalierungsmaße der Zielvariablen. Ist die abhängige Variable intervallskaliert, spricht man von Vorhersagemodellen. Klassifikationsmodelle dagegen besitzen einen binären, ordinalen oder nominalen Regressand. Man spricht von Klassifikation, weil die Klassenwahrscheinlichkeiten angegeben werden sollen. Vertreter dieser Modelle sind u.a. die Regressionsanalyse, das Entscheidungsbaumverfahren oder künstliche neuronale Netze.

[1] Die SEMMA-Methode wird in Kapitel 5 vorgestellt.

Die Besonderheit der Clusterung ist das Fehlen einer abhängigen Variablen. Als Modelle, die eine Klassenunterscheidung aufgrund der Homogenität bzw. Heterogenität der unabhängigen Variablen vornehmen, können beispielsweise das K-Means-Verfahren oder sog. selbstorganisierende Karten, die Kohonen-Netze, angeführt werden.

Eine Assoziationsanalyse soll Beziehungen und Abhängigkeitsverhältnisse zwischen Variablen aufdecken. Als Verfahren werden Assoziationsregeln zur Bestimmung von intratransaktionellen Mustern und Sequenzmuster zur Beschreibung intertransaktioneller Strukturen verwendet. Des Weiteren werden Link-Analysen angeboten, die Informationen in Form von Strukturen verbundener Objekte repräsentieren.

Die Dimensionsreduktion ist in komplexen Datenbeständen unerlässlich, um eine Visualisierung der Problemstellung zu ermöglichen. Auch wenn eine Reduktion auf zwei bis drei Dimensionen, die eine graphische Anschauung ermöglicht, nicht gelingt, ist eine Variablenselektion das Instrument, um Modelle zu vereinfachen und damit die Fähigkeit zur Modellentwicklung zu erhöhen.

Die Beschreibung von Strukturen in Daten ermöglicht Verständnis über den zu untersuchenden Datenbestand zu gewinnen. Die Ergebnisse aus den beschriebenen Aufgabengebieten, d.h. Prognosen, Segmentierungen, Assoziationsanalysen und Dimensionsreduktion, lassen sich zur Darstellung von Datenmustern und -strukturen heranziehen. Auf diese Weise lassen sich auch Aussagen über die Datenqualität treffen, welche Gegenstand der vorliegenden Arbeit ist.

Ein wesentliches Kennzeichen des Data Mining ist die Evaluierung von großen Datenmengen. Im ökonomischen Umfeld befindet sich die Größenordnung der Datenbanken im Giga- oder Terabyte Bereich. Data Mining-Analysen nutzen häufig hochdimensionale Data Warehouse-Architekturen, um ihre Daten zu beziehen. {[18] und [32]} Wie in allen Datenanalyseverfahren spielt die Datenqualität auch in der KDD eine große Rolle. Die hohe Dimensionalität der Daten erschwert die manuelle Behebung von Datenmängeln. Diese Handhabung ist zeitintensiv, fehleranfällig und daher ineffizient und sollte nach Möglichkeit vermieden werden. Bessere Ergebnisse werden meist mittels einer automatischen Behebung von Dateninkonsistenzen erzeugt, die in ein strukturelles Konzept eingebunden werden sollte. Vergegenwärtigt man sich die Statistik, dass 70 bis 80 Prozent des Aufwands einer Data Mining-Analyse auf die Datenvorbereitung fällt [21], verstärkt dies die Forderung nach einer ganzheitlichen Systematik.

Hypothesengesteuerte Konzepte, wie beispielsweise SQL-Abfragen, lassen teilweise noch ein Arbeiten mit Datenmängeln zu – wenngleich dies natürlich vermieden werden sollte. Durch genaue Kenntnis der Datenbasis und entsprechend gestalteten Aufrufen können auch bei unzureichender Datenqualität die erwünschten Ergebnisse erzielt werden. Datengesteuerte Ansätze, wie es Data Mining darstellt, laufen dagegen weitgehend automatisiert ab. Dies beinhaltet, dass die gefundenen Muster jedoch stets durch Datenfehler verfälscht sind.

Die im Verlauf der Arbeit vorgestellten Verfahren zur Gewährleistung von Datenqualität beziehen sich selbstverständlich nicht nur auf Bottom-Up-Analysen des Data Mining, die nach Schätzungen der Meta Group weniger als 5 Prozent aller Auswertungen ausmachen. [30] Die für eine Data Mining-Analyse erzeugte Datenqualität sollte daher auch für andere Zwecke genutzt werden können, d.h. für Top-Down-Analysen, Berichte sowie die Datenhaltung in Datenbanken und Data Warehouse-Architekturen. Die Modifizierung der Daten, die während einer Data Mining-Analyse vorgenommen werden, sollten – sofern gewollt – anschließend den Datenquellen oder dem Data Warehouse-System zurückgeführt werden können, um so auch an diesen Stellen die Datenqualität zu erhöhen. [13]

Aussagen über Datenqualität werden in Kapitel 2 ausführlich behandelt. Für eine Data Mining-Analyse kann jedoch eine allgemeine Forderung formuliert werden, die dann im Laufe dieser Arbeit genauer spezifiziert wird: Daten für eine Data Mining-Analyse sollten in denormalisierter Form vorliegen und regelmäßig aktualisiert werden, um damit eine gute Generalisierungsfähigkeit zu gewährleisten. Außerdem wird meistens eine niedrige Granularität gewünscht, wobei diese hoch detaillierten Daten vorbereitet und weitgehend vollständig seien sollten. [30] Der Begriff „vorbereitet“ und dessen genaue Bedeutung für die Datenqualität im Data Mining werden in den Kapiteln zu Data Quality Management und Pre-Processing detailliert erläutert.

Zuvor wird eine Methodologie für die KDD eingeführt, um damit eine Systematik in den Prozess der Datenqualität zu bringen. Sowohl Data Mining, als auch die Methoden zur Gewährleistung guter Datenqualität weisen eine hohe Komplexität des Analyseprozesses auf, welche durch die Vielzahl denkbarer Untersuchungsprobleme entsteht, die dazu in unterschiedlichster Weise miteinander kombinierbar sind. Die vorgestellte Methodologie erlaubt eine generelle Systematisierung bei allen Problemen, die entlang der Prozesskette auftreten können.

Die Vorteile des Data Mining, d.h. Automatisierbarkeit, effiziente Methoden für große Datenbestände und Transformation von Daten in Informationen, Muster oder Wissen, macht sich das Data Quality Mining [13] zu Eigen, das im letzten Kapitel eingeführt wird. Auf diese Weise soll mit den Verfahren des Data Mining unbekannte Zusammenhänge und Strukturen bei dem Auftreten von Datenmängeln entdeckt werden.

2 Datenqualität

Der Wandel von einer Industrie geprägten Gesellschaft hin zum sog. Informationszeitalter lässt sich seit der zweiten Hälfte des 20. Jahrhunderts beobachten, wobei diese Entwicklung durch den Siegeszug des Internets, der Globalisierung und der Öffnung der Märkte in den 1990er Jahren ihren zumindest zeitweiligen Höhepunkt fand. Informationen werden heute als Produktionsfaktor angesehen. Das „The Data Warehouse Institute“ [35] bezeichnet somit die Information treffend als die neue Währung dieser „New Economy“. Dementsprechend kommt den Daten die Rolle des Rohmaterials zu, die in einen strategischen Vermögensposten transformiert werden sollen. Die Qualität der Daten nimmt einen kritischen Teil in der Informationsgewinnung ein. Entsprechend des in der Datenverarbeitung geläufigen Sprichwortes „Garbage in, garbage out“ stellt schlechte Datenqualität einen Hauptgrund bei mangelhaften bzw. falschen Ergebnissen von Datenanalysenverfahren dar. In diesem Zusammenhang stellt sich folgende Frage: „Was ist (gute) Datenqualität?“

Aussagen über die Bedeutung von Datenqualität sind stets subjektiv geprägt, man könnte salopp formulieren: „Die Datenqualität muss dem genügen, was der Benutzer bzw. das System fordert“. Trotz dieser doch sehr allgemeinen Anforderung, kann man Begriffe und damit Konzepte finden, die fast immer vorausgesetzt werden: Integrität, Gültigkeit, Konsistenz und Aktualität.

In der Literatur {Vgl. im Folgenden: [12]} werden unterschiedliche Konzepte und Schwerpunkte propagiert, die verschiedene Schwerpunkte setzen. So schließen Wand und Wang [36], dass Datenqualitätsmängel bei Inkonsistenzen zwischen der Sicht auf das Informationssystem und der Sicht auf die reale Welt auftreten. Die so auftretenden Abweichungen können anhand von Datenqualitätsmerkmalen wie Vollständigkeit, Eindeutigkeit, Bedeutung und Korrektheit bestimmt werden. English [9] dagegen unterscheidet zwischen der Qualität der Datendefinition, der Architektur, der Datenwerte und der Datenpräsentation. Wang und Strong [37] bewerten in einer empirischen

Studie allgemeiner Datenqualitätsmerkmale anhand von vier Kategorien. Datenqualität wurde demnach kontextabhängig bestimmt oder aufgrund der Datenwerte (innere Datenqualität). Des Weiteren wurde die Zugangsqualität bewertet sowie eine vernünftige Darstellungsqualität der verschiedenen Zugriffe. Eine prozessorientierte Herangehensweise wurde von Jarke [17] gewählt, der Datenqualitätsmerkmale in die Vorgänge Entwicklung und Verwaltung, Softwareimplementierung sowie Datennutzung gliedert.

Aus diesen unterschiedlichen Herangehensweisen an die Thematik Datenqualität lässt sich für den Komplex Data Mining folgende Anforderungen formulieren: {Vgl. im Folgenden: [12]}

Interpretierbarkeit:

Aus dem Bereich der Semantik wird die Forderung erhoben, dass die Entitäten und Beziehungen sowie deren Attribute und Wertebereiche einheitlich und klar beschrieben sein sollen. Die einzelnen Informationsobjekte müssen eindeutig identifiziert werden können, wobei der zeitliche Bezug einzelner Informationsobjekte abgebildet und dokumentiert sein sollte. Letztendlich wird eine einheitliche Repräsentation fehlender Werte gefordert, d.h. es existiert eine Definition für fehlende Werte (Missing und Empty Value)[2], die dementsprechend abzubilden sind.

Nützlichkeit:

Die einzelnen Informationsobjekte müssen das Kriterium der Vollständigkeit erfüllen, d.h. alle Entitäten, Beziehungen und Attribute sind erfasst. Die Daten ermöglichen die Erfüllung der Aufgabe, wenn die Darstellung der Granularität den Ansprüchen angepasst wird. Alle Entitäten, Beziehungen und Attribute sind im notwendigen Detaillierungsgrad zu erfassen. Die letzten Forderungen des Kriteriums der Interpretierbarkeit beziehen sich auf die Relevanz. Die Datenwerte müssen auf einen relevanten Datenausschnitt beschränkt werden können, die sich dann auf den benötigten Zeitraum beziehen.

Glaubwürdigkeit:

Die Daten müssen inhaltlich mit der Datendefinition übereinstimmen und empirisch korrekt sein. Des Weiteren wird bezüglich der Integritätsbedingungen und der Wertebereichsdefinitionen Widerspruchsfreiheit gefordert. Eine konstante Glaubwürdigkeit der Daten ist Voraussetzung für das Kriterium der Zuverlässigkeit. Eine syntaktische Korrektheit gilt als erreicht,

[2] Siehe hierzu Abschnitt 5.1.

wenn die Daten mit der spezifizierten Syntax übereinstimmen. Abschließend muss auf die zeitliche Dimension eingegangen werden. Die Datenwerte sollten möglichst aktuell sein, d.h. eine einheitliche Erfassung bezogen auf den gegenwärtigen Zeitpunkt. Die Aktualität der Daten sollte mit einer zeitlichen Konsistenz verbunden sein.

Die verschiedenen Kriterien aus den Bereichen Interpretierbarkeit, Nützlichkeit und Glaubwürdigkeit geben einen Rahmen für gute Datenqualität bei einer Data Mining-Analyse vor. Ist diese nicht gewährleistet, sollte überlegt werden, wie sich diese Datenmängel begründen, damit anschließend korrigierend eingegriffen werden kann.

Mangelnde Datenqualität hat viele Ursachen: {Vgl. im Folgenden: [1] und [35]} Die Erfassung der Daten, insbesondere bei der manuellen Eingabe, stellt das erste Problemfeld dar. Eingabefehler können durch Tipp- und Schreibfehler entstehen oder sind auf mangelnde Konzentration bzw. ein nicht ausreichend ausgeprägtes Problembewusstsein zurückzuführen. Durch Definition und Implementierung von Standards sowie geeigneten Verifizierungsmaßnahmen kann hier Abhilfe geschaffen werden. Auf diese Weise wird auch die Problematik des falschen Contents in den Feldern behoben. Ein falscher Content entsteht, wenn ein Datenfeld missbraucht wird, um Daten aus einem anderen Themenbereich einzutragen.[3] Daten weisen zudem eine komplexe Struktur auf, so existieren unterschiedliche Datentypen (numerische und alpha-numerische Variablen sowie Datumsvariablen), sowohl im Einlese- als auch im Ausgabeformat. Eine unterschiedliche Verwendung von Formaten kann somit mögliche Fehlerquellen mit sich ziehen. Metadaten helfen mit Informationen über die Daten und ihrer Strukturen Fehler zu entdecken (siehe auch Kapitel 4). Das Hauptproblem bei der Entstehung der Datenqualität fällt der Migration von Daten zu. Daten stammen aus verschiedenen Quellen: relationale Datenbanken, Legacy Systeme, unstrukturierte Textdateien, unterschiedliche Fileformate aus dem Bereich Internet oder diverse Unternehmensapplikationen. Diese Daten sind heterogen und müssen – werden sie zusammengeführt – vereinheitlicht werden. In Kapitel 4 wird mit den ETL-Prozessen weiter auf diese Problematik eingegangen.

[3] Als Beispiel könnte die Datenkategorie „Telefon privat“ herangezogen werden. Da hier selten Einträge vorgenommen werden, gleichzeitig aber eine wichtige Kategorie fehlt, z.B. E-Mail, kann es passieren, dass diese Informationen dort „zwischengeparkt“ werden. Im operativen Geschäftsablauf bringt ein falscher Content in den Feldern meist keine Probleme mit sich, da sie von dem Benutzer leicht zu identifizieren sind. Automatisierte Analysen können dagegen diese Felder nicht von Feldern mit einem richtigen Content unterscheiden und liefern deshalb falsche Ergebnisse.

Datenqualität wird nicht positiv, sondern negativ in Form von Datenmängeln wahrgenommen. Die Wahrnehmung von schlechter Datenqualität ist häufig auch mit der Entstehung von Kosten verbunden. {[1] und [35]} Bei vielen Datenproblemfeldern lassen sich die Kosten direkt zuordnen und quantifizieren. Direkte Kosten entstehen durch fehlerhafte Rechnungen, verursacht durch falsche Datenwerte oder einen falschen Content. Ein weiteres Problem entsteht durch Dubletten, welche häufig Mehrfachversendungen zur Folge haben. Diese können allerdings auch durch eine uneinheitliche Schreibweise hervorgerufen werden. Abschließend, wenngleich sich noch viele weitere Beispiele finden lassen, kann hier noch der zusätzliche Aufwand genannt werden, der durch alle Formen von Dateninkonsistenzen entsteht.

Neben Kosten, die direkt auf Datenfehler zurückzuführen sind, können auch solche genannt werden, die sich indirekt, häufig erst später auswirken. An dieser Stelle lässt sich der Vertrauensverlust in die eigenen Entscheidungen oder das Entscheidungsinstrumentarium nennen, was zur Folge haben kann, dass beispielsweise Data Mining-Analysen abgesetzt werden. Probleme aus dem Bereich der direkten Kosten wie falsche Rechnungen, mehrfach verschickte Postsendungen oder falsche Berechnungen führen langfristig außerdem zu unzufriedenen Kunden oder gar zur Abwanderung. Die Umsetzung neuer Anforderungen, Richtlinien[4] oder Konzepte[5] wird durch mangelhafte Datenqualität drastisch erschwert. Zusammenfassend kann gesagt werden, dass eine schlechte Datenqualität zu konstant schlechten Entscheidungen und damit zu einer Minderung des Unternehmenserfolges führt. Schätzungen zufolge veranschlagen die Gesamtkosten schlechter Daten zwischen acht bis zwölf Prozent des Umsatzes. Datenbanken mit Kundeninformationen beinhalten ca. 15 bis 20 Prozent fehlerhafte Einträge. [33]

Der Nutzen aus guter Datenqualität lässt sich entsprechend formulieren: Geringere Kosten, schnellere Entscheidungen und Prozessabläufe, höhere Zufriedenheit und damit letztlich mehr Gewinn. Außerdem kann der „Single Version of the Truth“ [35] nicht genügend Nutzen zugesprochen wird, gewährleistet sie doch Entscheidungssicherheit. Somit sorgt eine gute Datenqualität auch dafür, dass bessere Data Mining-Lösungen erzeugt werden.

Das Ziel eine gute Datenqualität zu erreichen, beispielsweise für eine Data Mining-Analyse, lässt sich durch folgendes Schema bewerkstelligen: Entde-

[4] z.B. Basel II, Bilanzierungsvorschrift nach IAS oder Umsetzung von KonTraG.

[5] Ob Balanced Scorecard, Customer Relationship Management oder Supply Chain Management, Management-Konzepte, die Informationen verarbeiten, sind auf gute Datenqualität angewiesen.

cken – Bewerten – Erklären – Korrigieren. Methoden zur Gewährleistung guter Datenqualität sollten stets über die Schritte Entdecken und Korrigieren hinausgehen, da sonst nur auf Datenmängel reagiert werden kann. Ein proaktives Datenmanagement [33] entsteht, wenn die Schritte Bewertung und Erklärung hinzukommen. Auf diese Weise können stetig auftretende Probleme dauerhaft gelöst werden und zudem geklärt werden, wann die Korrektur stattfinden soll. Dies ist insofern wichtig, da ein kausaler Zusammenhang bestehen kann zwischen Datenfehler und der untersuchten Problemstellung[6].

3 Methodologie für systematisches Vorgehen

Mit der SEMMA-Methode existiert ein Modell für eine strukturierte Data Mining-Analyse. Die in diesem Kapitel vorgestellte Methodologie ermöglicht nun, die SEMMA-Methode integrierend, eine systematische Vorgehensweise für die gesamte KDD. Maßnahmen, die eine ausreichende Datenqualität für eine Data Mining-Analyse gewährleisten, lassen sich an verschiedenen Ansatzpunkten ansiedeln. Das Pre-Processing der Daten und ein geeignetes Data Quality Management sind Schlüsselwörter in diesem Prozess. Diese sind teils vor (Data Quality Management) und teils während des Prozesses (Pre-Processing) verankert. Gemäß der KDD wird nun eine Methodologie vorgestellt, anhand derer eine systematische Behandlung der Daten ermöglicht wird. Die Methodologie umfasst alle Prozessschritte der KDD – von der Quantifizierung der Ziele bis zur Implementierung der Ergebnisse. Durch die Zuordnung der Probleme in die einzelnen Ablaufschritte kann so der geforderten Systematisierung Folge geleistet werden.

Schritt 1 – Quantifizierung:

Im ersten Schritt der Methodologie erfolgt die Zielvereinbarung und Aufgabenzuweisung. Außerdem werden die Projekt- bzw. Implementierungsdauer festgelegt und etwaige Kompetenzen und Zuständigkeitsbereiche abgesprochen.

Schritt 2 – Anforderungsanalysen:

Im zweiten Schritt der Methodologie wird eine Anforderungsanalyse durchgeführt. Dies beinhaltet neben der Bestimmung der benötigten Variablen auch Aussagen und Forderungen bezüglich der Datenqualität.

---

[6] Datenfehler können Erklärungskraft für ein Prognoseverfahren besitzen. Vgl. hierzu die Indikator-Variablen aus Abschnitt 5.1.

Schritt 3 – Beschaffung der Daten:

Im dritten Schritt der Methodologie erfolgt die Beschaffung der Daten. Je nach Speicherort der Daten sind verschiedene Maßnahmen des Data Quality Managements anzusiedeln.

Schritt 4 – Data Mining mit SEMMA:

Im vierten Schritt der Methodologie wird die Data Mining-Analyse durchgeführt. Anhand der SEMMA-Methode, d.h. den einzelnen Schritten Sample, Explore, Modify, Model und Assess, wird der Benutzer geleitet. An dieser Stelle ist das Pre-Processing der Daten verankert.

Schritt 5 – Implementierung der Ergebnisse:

Im letzten Schritt der Methodologie werden die Ergebnisse der Data Mining-Analyse entsprechend der unter Quantifizierung genannten Ziele umgesetzt.

Dokumentation und Review:

Die Methodologie wird in den einzelnen Schritten durch einen Review begleitet. Zwischen den einzelnen Schritten wird geprüft, ob durch Änderungen Verbesserungen erzielt werden können. Dies schließt auch eine umfassende Dokumentation ein.

Das Thema Datenqualität, d.h. Probleme und entsprechende Maßnahmen, wird entlang der gesamten Methodologie behandelt. Sei es durch Vorgaben, Anforderungen, festgestellten Diskrepanzen oder Handlungen. Auf diese Weise erhält das Problem der Datenqualität die benötigte Aufmerksamkeit. In diesem Kapitel richtet sich der Fokus auf die Beschaffung der Daten. Dieser Schritt beinhaltet auch das Data Quality Management, welches Gegenstand des nächsten Kapitels ist. Das Pre-Processing der Daten erfolgt im vierten Schritt der Methodologie, dem Data Mining mit SEMMA.

Neben den formalen Qualitätsmerkmalen müssen auch die inhaltlichen Gesichtspunkte stimmen. Voraussetzung für eine Data Mining-Analyse ist die Verfügbarkeit der notwendigen Daten {Vgl. im Folgenden: [22]}. Diese lassen sich in folgende Kategorien unterteilen:

Grunddaten

Grunddaten enthalten alle Informationen, die zur Identifikation und allgemeinen Beschreibung des Kunden, des Produktes oder des Geschäftsablaufes herangezogen werden können. Die Grunddaten enthalten Stammdaten, soziodemographische, sozioökonomische und psychographische Daten sowie Finanzdaten. Stammdaten enthalten die verschiedenen Identifikationsmerkmale, d.h. Identifikationsnummern, Bezeichnungen und Namen. Die Entität

Kunde wird durch soziographische und sozioökonomische Daten beschrieben, hierzu zählen Variablen wie Alter, Geschlecht, Familienstand, Religionszugehörigkeit etc. Finanzdaten stammen bei Unternehmen häufig aus dem Controlling und geben u.a. Auskünfte über den Umsatz, Gewinn, Cash Flow, Eigenkapital-, Gesamtkapital- und Umsatzrendite, ROI, Selbstfinanzierungsgrad, durchschnittliche Debitoren- und Kreditorenlaufzeit sowie Schuldentilgungsdauer. Finanzdaten von Personen enthalten Informationen bezüglich des Einkommens, Verbindlichkeiten, Sicherheiten, Vermögenswerte, Anzahl an Krediten, Kundendeckungsbeiträge oder Kundenwerte.

Historische Daten

Stammt die Data Mining-Aufgabe aus dem Bereich des überwachten Lernens [18] muss für die Zielvariable eine Basis von bekannten Fällen existieren, deren betrachtete Objekte klassifiziert[7] sind. Das ausgewählte, aber noch nicht angepasste Modell muss nun so konfiguriert (iterativer Trainingsprozess) werden, dass die bekannten Fälle möglichst gut reproduziert (Generalisierungsfähigkeit) werden können. Historische Daten bilden also die Grundlage für Prognosemodelle aus dem Bereich des überwachten Lernens.

Historische Daten können aber auch – handelt es sich nicht um die zu erklärende Variable – in Form von in Anspruch genommenen Leistungen, vergangene Transaktionshäufigkeiten, gewählten Vertriebskanälen und Dauer der Kundenzugehörigkeit einen Kunden beschreiben. Historische Daten werden außerdem in Form von Aktions- und Reaktionsdaten verwendet, i.d.R. Dummy-Variablen, um ein Verhalten, das auf eine Handlung folgt, zu dokumentieren. Als Beispiel kann hier eine Marketingaktion angeführt werden, wobei in den Aktionsdaten Art, Umfang, Zeitpunkt, Kosten und Kommunikationskanal festgehalten, während in den Reaktionsdaten das Antwortverhalten sowie Reaktionsart und -weg gemessen wird.

Potentialdaten

Auskünfte über den künftigen Bedarf eines Kunden können den Potentialdaten entnommen werden, die den zeitlichen Nutzungsrahmen wie Vertragslaufzeiten, Produktlebenszyklus oder neue technische Entwicklungen darstellen.

Eine Data Mining-Analyse benötigt i.d.R. Variablen aus diesen Bereichen. Sollten Daten fehlen, von denen man sich erhofft, dass sie eine Erklärungs-

[7] Mögliche Kategorien sind binär (z.B.: 0, 1), mehrfach (z.B.: Aaa, Aa, …, Caa oder 1, …, 10), Funktions-Konstrukte oder Prognosen mit zeitlicher Komponente.

kraft für Prognosemodelle besitzen, sollten diese entweder über externe Daten oder Schätzungen in den Datensatz gelangen.

4 Data Quality Management

Data Quality Management ist ein umfangreiches Feld, das an dieser Stelle nur kurz angerissen werden kann. Es werden lediglich die Bereiche aufgegriffen zu denen ein Bezug zum Thema Datenqualität im Data Mining herzustellen ist. Dieses Kapitel beschäftigt sich daher hauptsächlich mit der Datenbereinigung während des Prozesses der Datenmigration. An dieser Stelle sei auf Kapitel 3 verwiesen, welches sich u.a. mit der Beschaffung der Daten für eine Data Mining-Analyse auseinandersetzt. Die Prozesse, die bei der Datenmigration beachtet werden müssen, werden nun genau so berücksichtigt, wie Überlegungen, die hinsichtlich einer Architektur (Data Warehouse und Metadaten-Management) zu treffen sind. Nachdem aufgezeigt wurde, inwieweit sich Data Mining in eine Data Warehouse-Umgebung integrieren lässt, werden die dort verankerten Maßnahmen zur Datenbereinigung, die sog. ETL-Prozesse beschrieben.

Data Quality Management sollte stets als unternehmensweiter Ansatz verstanden werden, und erfordert deshalb Änderungen in der Ablauf- und Aufbauorganisation. Damit Datenqualität auch wirklich unternehmensweit Einzug erhält, bedarf es der Schaffung eines diesbezüglichen Problembewusstseins. Operative Problemfelder bei denen die Maßnahmen des Data Quality Management angewandt werden, sind mannigfaltig: Fusionen, Abgleiche externer Daten, Systemwechsel und besonders die Migration von Daten.

Data Warehoue-Architektur

Das Data Warehouse (DWH) stellt ein unternehmensweites Konzept zur effizienten Bereitstellung und Verarbeitung entscheidungsorientierter Daten dar. Diese Daten weisen einen hohen Aggregationsgrad auf, haben einen Zeitraumbezug und sind den Bedürfnissen der Entscheidungsträger zur Durchführung ihrer Aufgaben angepasst. Inmon [16] fasst die Forderungen, die ein DWH erfüllen soll, folgendermaßen zusammen: „A data warehouse is a subject oriented, integrated, non-volatile, and time variant collection of data in support of management's decisions." Das DWH kann somit im Wesentlichen durch die Merkmale Themenorientierung, Vereinheitlichung, Dauerhaftigkeit und Zeitorientierung beschrieben werden. [30]

KDD und DWH können im Zuge einer geeigneten Strategie zur Verbesserung bei Entscheidungsfindungen in verschiedenen Geschäftsprozessen genutzt werden. Data Mining verlangt nicht zwingend nach einer DWH-Architektur, sie ist aber überaus hilfreich. Es werden nicht nur die benötigten Daten zur Verfügung gestellt, sondern dies bereits in integrierter und konsistenter Form. [30] Daten dieser Qualität sind sonst nur mit großem Aufwand aus den operativen Systemen zu beziehen. Dadurch, dass Data Mining-Werkzeuge auf Daten zugreifen können, die im DWH bereinigt und konsolidiert werden, entstehen außerdem keine Belastungen für die operativen Datenbanken durch den Data Mining-Prozess. [18] Anschließend können Ergebnisse des KDD-Prozesses direkt zurück auf das DWH übertragen werden.

Häufig wird im Data Mining auf die Variante der Data Marts[8] zurückgegriffen (sog. Data Mining Marts), die speziell auf die Data Mining-Anforderungen zugeschnittene Daten enthalten.

ETL-Prozesse

Die Migration der Daten in ein DWH wird über ETL-Prozesse gesteuert. {Vgl. im Folgenden: [1], [30] und [35]} Die Abkürzung ETL steht für „Extraction, Transformation, Loading“, wobei besonders in den Maßnahmen der Transformation die Datenqualität gesteuert wird. Im ersten Schritt der Prozesskette werden die Daten aus den Quellsystemen extrahiert. Innerhalb dieses Transfers werden schon erste geringfügige Umgestaltungen vorgenommen (Datenkodierung, Dekomprimierung, Umwandlung in Groß- oder Kleinschreibung, Datum- und Zeitkonvertierungen sowie Sortierungen). Im nächsten Schritt, der Transformation, können die Modifizierungen, sowohl auf Zeilen- oder Spaltenebene, als auch Tabellen übergreifend, durchgeführt werden. Die Transformationen umfassen die Gebiete Validierung und Verifizierung, Ableitung neuer Werte, Aggregation sowie Bereinigung. Zu Beginn der Transformations-Prozesse wird meistens ein Profiling durchgeführt, welches die Daten hinsichtlich ihrer Qualität analysiert. Die Ergebnisse dieser Analyse stellen häufig einen Leitfaden für die folgenden Prozessschritte dar. Anschließend wird im Prozess des Parsings, d.h. der Zerlegung von Strings, einzelne Datenelemente aus den Rohdaten lokalisiert, identifiziert und schließlich isoliert. Darauf bauen Standardisierungsverfahren auf, die eine einheitliche Schreibweise von Informationen ermöglichen, da nur so

[8] Data Marts stellen einen Teilausschnitt eines DWH dar, und halten lediglich auf den Anwender abgestimmte Informationen und Daten bereit. Neben Data Mining Marts existieren z.B. Marketing oder Finanz Marts.

Abfragen über Freitext oder Data Mining-Analysen das gewünschte Ergebnis liefern. Das Verifizieren, beispielsweise mit Hilfe von Referenztabellen[9], ermöglicht die Aufdeckung von Inkonsistenzen. Matchen und Clustern ermöglicht schließlich die Zusammenführung von Informationen. Im Match-Prozess werden Daten, die aus verschiedenen Quellen stammen, anhand eines Match-Codes verglichen und ermöglichen somit die Identifizierung von Dubletten. Diese können auch durch das Clustern von Daten entdeckt werden. Im letzten Schritt des ETL-Prozesses werden die Daten in das DWH geladen, wobei hier verschiedene Ladeverfahren angewandt werden können (Refresh bzw. Ersetzen, Einfügen der geänderten Daten oder Anhängen neuer Daten).

Metadaten

Metadaten sind beschreibende Informationen über Struktur und Inhalt der Daten und über Applikationen und Prozesse, die auf diese Daten zugreifen. [30] Fachliche Metadaten[10] fördern das Verständnis über Daten und Anwendungen. Indem über Metadaten Regeln bezüglich Transformationen oder Berechnungen sowie Darstellungsformen und fachliches Know-how in Fachsprache kommuniziert werden, wird damit auch ein nicht unwesentlicher Einfluss auf die Datenqualität ausgeübt. Die Informationen, welche die Metadaten enthalten, ermöglichen die Anpassung von semantischen und syntaktischen Unterschieden verschiedener Daten. So werden beispielsweise Synonyme erkannt oder Homonyme identifiziert. Auf diese Weise können Standardisierungen zielgerichtet ablaufen.

5 Pre-Processing

Eine Data Mining-Analyse, beispielsweise mit dem SAS® Enterprise Miner™ [2], hat immer Prozesscharakter. Die Modelle erfahren stets viele Anpassungsschritte bevor die Klassifikationsgüte als ausreichend für das jeweilige Problem bewertet wird. Die Reihenfolge einer Analyse erfolgt immer nach folgendem Schema: Pre-Processing, Modellierung und Bewer-

[9] Referenztabellen enthalten Konvertierungsinformationen, die für Soll-Ist-Wertevergleiche, Erstellung von Formaten und Auswertungen herangezogen werden. Für SAS-Einlese- und Ausgabeformate existieren spezielle SAS Translation-Tables. [30] Öffentliche Datenbanken mit Adress- und Telefonlisten ermöglichen den Vergleich mit der eigenen Datenbasis.

[10] Metadaten können in fachliche und technische Metadaten aufgeteilt werden. Technische Metadaten beschäftigen sich mit physischen Datenbeschreibungen: Daten-Speicherorte, Infrastruktur (Netzwerk, Server etc.), ETL-Prozesse oder Tabellenrelationen.

tung. Zur Orientierung bei der Auswahl der Werkzeuge und deren Anordnung, bietet sich die SEMMA-Methode der SAS Institute Inc. an, die das oben erwähnte Schema aufgreift und erweitert. Die einzelnen Schritte dieser Methode umfassen eine Reihe von Knoten und anhand der vorgegebenen Reihenfolge des Begriffes SEMMA werden das Pre-Processing, die Modellbildung und schließlich die Modellauswahl durchgeführt, wobei die Methode einen iterativen Prozess darstellt, indem Rückkopplungen ausdrücklich vorgesehen sind.

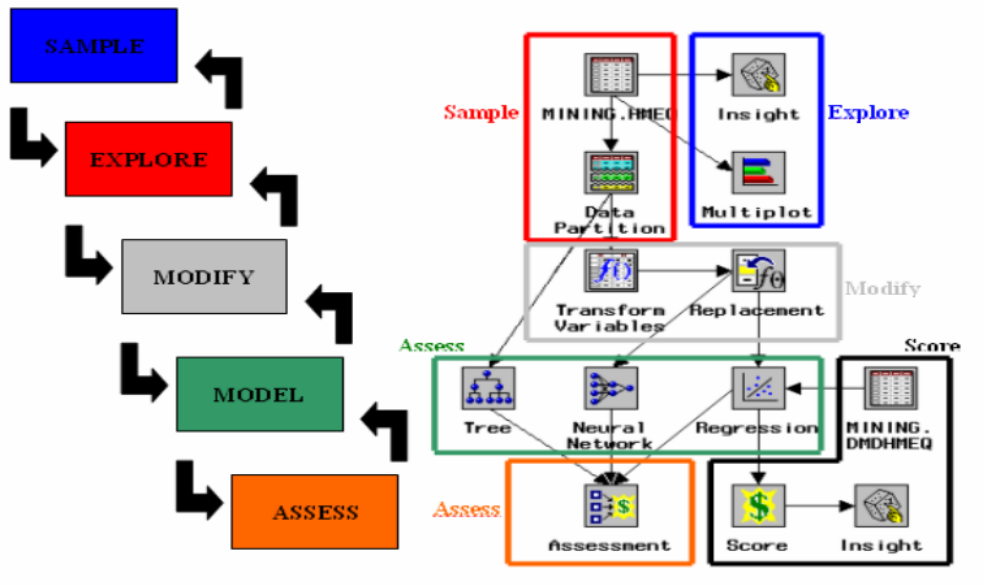

Abbildung 1: Die SEMMA-Methode und ein Prozessdiagramm des SAS® Enterprise Miner™

Als Abschluss einer Data Mining-Analyse kann mit dem Score-Knoten das entwickelte Modell (hier: Regression) auf neue Daten angewandt werden.

Die SEMMA-Methode steht im Einzelnen für: {Vgl. im Folgenden [2]}

Sample:

In den Knoten der Sample-Gruppe werden die Daten für die Analyse eingelesen, die Prädikatsmerkmale und der Einfluss der Variablen, die diese im

Modell annehmen sollen, festgelegt und Target-Profile definiert. Es erfolgt die Einteilung der Daten in die Kategorien Training, Validierung und Test. Außerdem besteht die Möglichkeit aus der Datenbasis Stichproben zu ziehen.

Explore:

Um eine Data Mining-Analyse durchführen zu können, sind Kenntnisse über den Datensatz, die Variablen und insbesondere der zugrunde liegenden Qualität erforderlich. Visualisierungen, das Aufdecken von Zusammenhängen inklusive Assoziationsanalyse und die Auswahl der Variablen werden in den Knoten der Explore-Gruppe durchgeführt.

Modify:

Die Erkenntnisse aus dem Explore-Knoten werden anschließend in den Knoten der Modify-Gruppe umgesetzt. Mangelnde Datenqualität verschlechtert die Klassifikationsgüte. Geeignete Modifizierungen stellen die größten Verbesserungspotentiale für Data Mining-Modelle dar. Das wichtigste Kriterium sind Einfügestrategien für fehlende Werte. Variablentransformationen beheben Probleme wie eine nicht vorhandene Normalverteilung. Neue Variablen können aus vorhandenen erzeugt werden. Die Ausreißerproblematik kann mit Filtern behoben werden. Außerdem besteht die Möglichkeit einer Clusteranalyse.

Model:

Klassifikations- und Vorhersagemodelle werden in Knoten der Model-Gruppe entwickelt. Zu den wichtigsten Modellen gehören die Regressionsanalyse, das Entscheidungsbaumverfahren und künstliche neuronale Netze. Dazu kommen Verfahren wie die Hauptkomponentenanalyse, das fallbasierte Schließen sowie kombinierte und benutzerdefinierte Modelle.

Assess:

Der Assessment-Knoten bietet die Möglichkeit konkurrierende Modelle vergleichend gegenüberzustellen. Dazu können entweder statistische Kennzahlen, die sog. Confusion Matrix oder Graphiken (Response, Captured Response, Lift Value und ROC-Charts) zur Visualisierung der Klassifikationsgüte verwendet werden.

Das Pre-Processing der Daten in der SEMMA-Methode wird in den Knoten der Sample-Gruppe und der Modify-Gruppe durchgeführt, wobei der Explore-Schritt Auskunft über die zu untersuchenden Problemstellungen gibt. Die folgenden Erläuterungen über das Pre-Processing der Daten kon-

zentrieren sich auf Klassifikations- und Prognosemodelle, da diese die größte Bandbreite aufweisen. Des Weiteren werden lediglich die Elemente des Pre-Processings aufgegriffen, die im Bezug zur Datenqualität stehen. Folgende Aufgaben sind im Pre-Processing gewöhnlich zu erledigen:

I. Einfügen von fehlenden Werten
II. Eliminierung von Ausreißern
III. Transformation der Variablen
IV. Generierung neuer Variablen
V. Partitionierung der Daten
VI. Selektion der Variablen

5.1 Fehlende Werte

Der Problemfall der fehlenden Werte stellt einen der wichtigsten Pre-Processing-Schritte dar. Fehlende Werte haben einerseits einen großen Einfluss auf die Modellgüte, andererseits ist die Handhabung sehr vielschichtig, so dass diesem Thema im Folgenden viel Platz eingeräumt wird. Fehlende Werte sind fehlende Einträge in den einzelnen Tupeln. Diese werden im SAS-System, sofern es sich um numerische Werte handelt, durch einen Punkt (.) dargestellt bzw. durch ein Leerzeichen (blank) gekennzeichnet, wenn es alphanumerische Werte sind.

Die Entstehung von fehlenden Werten liegt in zwei Ursachen begründet: Nicht-Existenz oder Nicht-Erfassung. Deshalb empfiehlt es sich fehlende Werte in Missing und Empty Values zu unterscheiden. [21] Missing Values sind demnach Werte, die nicht erfasst wurden, aber existieren. Dagegen sind Empty Values Werte, die aufgrund ihrer Nicht-Existenz nicht zu erfassen sind. Die Handhabung von Empty Values ist relativ einfach. Eine proaktive Möglichkeit im Rahmen des Data Quality Managements ist die Festlegung von Standards schon bei der Datenerfassung. Besteht die Möglichkeit, dass in einer Variablen auch leere Werte vorkommen können, sollte eine entsprechende Auswahlmöglichkeit integriert werden, beispielsweise durch die Einträge „other“ oder „none“. Diese Möglichkeit besteht allerdings nicht immer, vor allem wenn die Daten im operativen System für Zwecke fernab der Datenanalyse verwendet werden, was den Regelfall darstellt. Kann das Problem der Empty Values nicht durch die Implementierung von Standards gelöst werden, besteht die Möglichkeit Empty Values durch die Eingabe von Konstanten zu kennzeichnen, dies erfolgt dann ebenfalls durch die Bezeichnung „other“ oder „none“. Empty Values treten meistens bei nominalen

Werten auf, da neben den vorgegebenen Werten auch alternative Möglichkeiten existieren können, die mangels Übereinstimmung leer bleiben werden.[11] Wie in den folgenden Problemfeldern ist Hintergrund-, Experten- bzw. Fachwissen nötig um eine Unterscheidung treffen zu können: *Handelt es sich um einen Empty oder Missing Value?*

Interessanter, da für Analysen wichtiger und auch häufiger anzutreffen, sind die Missing Values. Im Folgenden werden die Begriffe Missing Values und fehlende Werte synonym verwendet.

Missing Values lassen sich anhand verschiedener Muster und aufgrund unterschiedlicher Mechanismen bezüglich ihres Auftretens charakterisieren. [15] Der erste Schritt bei der Handhabung von fehlenden Werten stellt die Überprüfung nach der Existenz von Mustern oder Strukturen innerhalb der fehlenden Werte dar. So kann beispielsweise bei Zeitreihen eine monotone Zunahme von Missing Values festgestellt werden, die auf Panelmüdigkeit oder Ausscheiden aus der Beobachtungsgruppe zurückzuführen sind. Als nächstes sollte geprüft werden, ob das Auftreten von Missing Values auf Mechanismen zurückzuführen sind. Folgende fehlende Werte-Mechanismen sind hier zu nennen: {Vgl. im Folgenden: [6] und [15]}

I.	MCAR	Missing Completely At Random
II.	MAR	Missing At Random
III.	NMAR	Non Missing At Random – Non Ignorable

Anhand eines kleinen – idealisierten – Beispiels werden die Mechanismen bei fehlenden Werten verdeutlicht. Dieses beinhaltet eine Identifikationsvariable, die als Primärschlüssel fungiert sowie die Variablen Alter und Einkommen. Alter nimmt in dem Beispiel die Rolle der erklärenden Variablen ein, während Einkommen die zu erklärende Variable darstellt. Der komplette Datensatz beinhaltet zwei verschiedene Altersgruppen (27 und 56 Jahre) und Einkommensklassen (Mittelwert: 1.424,25 € und 5.571,25 €).

[11] Werden beispielsweise in der Variable Beruf verschiedene Auswahlmöglichkeiten vorgegeben, ist es schwer mit einer geringen Anzahl an Überbegriffen alle Berufsfelder abzubilden. Die Folge kann ein fehlender Wert sein, der als Empty Value zu interpretieren ist.

Kompletter Datensatz

Customer ID	Age	Income
C-100001	27	1.428,00
C-100002	27	1.378,00
C-100003	27	5.650,00
C-100004	27	5.289,00
C-100005	56	1.489,00
C-100006	56	1.402,00
C-100007	56	5.890,00
C-100008	56	5.456,00

MCAR

Customer ID	Age	Income
C-100001	27	1.428,00
C-100002	27	.
C-100003	27	5.650,00
C-100004	27	.
C-100005	56	1.489,00
C-100006	56	.
C-100007	56	5.890,00
C-100008	56	.

MAR

Customer ID	Age	Income
C-100001	27	.
C-100002	27	1.378,00
C-100003	27	.
C-100004	27	5.289,00
C-100005	56	1.489,00
C-100006	56	1.402,00
C-100007	56	5.890,00
C-100008	56	5.456,00

NMAR

Customer ID	Age	Income
C-100001	27	1.428,00
C-100002	27	1.378,00
C-100003	27	.
C-100004	27	.
C-100005	56	1.489,00
C-100006	56	1.402,00
C-100007	56	.
C-100008	56	.

Abbildung 2: Mechanismen bei fehlenden Werten[12]

MCAR:

Fehlende Werte in der Variablen y sind MCAR, wenn die Wahrscheinlichkeit, dass Werte yi nicht vorhanden sind, weder von einer anderen Variablen x noch von y selbst abhängt. Die fehlenden Werte aus Abbildung 1 der Tabelle MCAR treten unabhängig vom Alter und der Einkommensklasse auf.

$$\Pr(MV^{13} \mid Age = 27) = \Pr(MV \mid Age = 56) \text{ [11]} \qquad (1)$$

[12] Eigene Darstellung.
[13] MV steht für Missing Value.

und

$$\Pr(MV|Age) = \Pr(MV) \text{ [11]} \quad (2)$$

MAR:

Fehlende Werte in der Variablen y sind MAR, wenn die Wahrscheinlichkeit, dass Werte yi nicht vorhanden sind, von einer anderen Variablen x, aber nicht von y selbst abhängt. Die fehlenden Werte aus Abbildung 1 der Tabelle MAR treten nur bei jungen Personen auf, unabhängig von ihrem Einkommen. Das Auftreten von fehlenden Werten hängt von der Variable AGE ab.

$$\Pr(MV|Age = 27) = 0{,}5 \quad (3)$$

und

$$\Pr(MV|Age = 56) = 0{,}0 \quad (4)$$

NMAR:

Fehlende Werte in der Variablen y sind NMAR, wenn die Wahrscheinlichkeit, dass Werte yi nicht vorhanden sind, von y selbst abhängt (und evt. von einer anderen Variable x). Die fehlenden Werte aus Abbildung 1 der Tabelle NMAR treten bei hohem Einkommen auf. Das Auftreten von fehlenden Werten hängt von der Variable INCOME ab. Ohne zusätzliche Kenntnisse ist es nicht möglich zwischen MCAR und NMAR zu unterscheiden, da bei beiden Mechanismen folgendes gilt:

$$\Pr(MV|Age) = \Pr(MV) \quad (5)$$

Zusammenfassend: Ohne Kenntnisse der zugrunde liegenden Abhängigkeiten, was gewöhnlich der Fall ist, ist es nicht möglich die Daten wirklichkeitsgerecht zu analysieren.

Nachdem mögliche Gründe für das Auftreten von fehlenden Werten dargestellt wurden, wird im nächsten Schritt die Handhabung von fehlenden Werten untersucht. Fehlende Werte besitzen zwei verschieden Problemfelder: Einerseits die Behandlungsweise der Missing Values bei dem Aufbau des Modells und andererseits während des Score-Vorgangs. Allerdings unterscheidet sich die Vorgehensweise nicht grundsätzlich.

Die einfachste Strategie bei dem Umgang mit fehlenden Werten ist die Complete Case Analysis, der sog. Listwise Deletion. Hierbei werden nur die Fälle für die Modellanpassung genutzt, die vollständig vorhanden sind. Datensätze, die einen fehlenden Wert aufweisen, werden eliminiert.

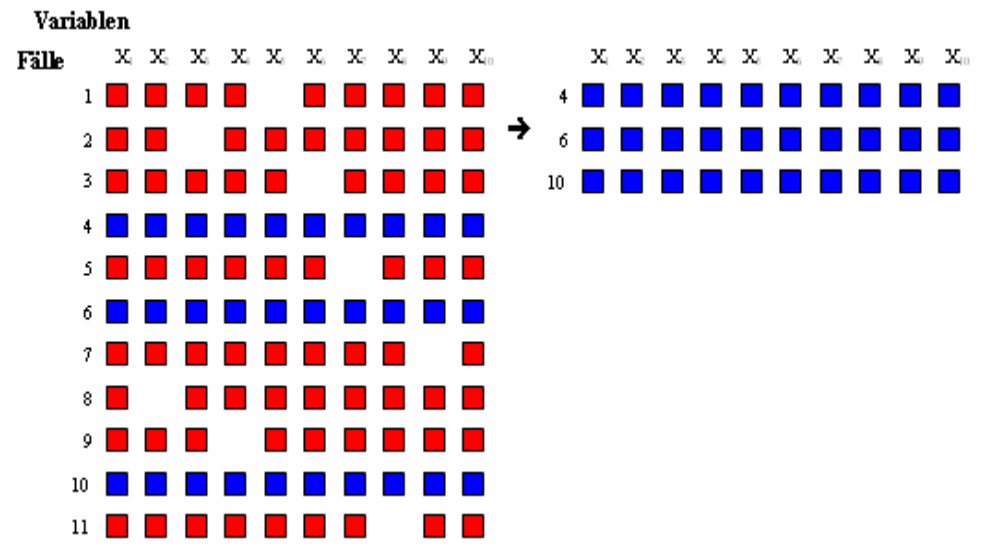

Abbildung 3: Complete Case Analysis – Listwise Deletion[14]

Abbildung 2 zeigt einen wesentlichen Nachteil der Complete Case Analysis: der hohe Datenverlust. Eine hohe Ausschussrate bei einer Vielzahl von Variablen kann schon bei relativ geringem prozentualem Anteil an Missing Values entstehen. Der Erwartungswert für komplette Fälle (CC für Complete Case) beträgt bei k Inputvariablen und einer Wahrscheinlichkeit von α für das Auftreten von fehlenden Werten:

$$E(CC) = (1 - \alpha)^k \text{ [28]} \qquad (6)$$

Ein weiterer Nachteil der Complete Case Analysis ist die Tatsache, dass die Annahme vorausgesetzt werden muss, dass fehlende Werte nach dem Mechanismus MCAR entstehen. Mögliche systematische Unterschiede zwischen kompletten und nicht-kompletten Fällen werden ignoriert. Es wird also vorausgesetzt, dass die Verteilung der kompletten Fälle und der fehlenden Werte übereinstimmt.

Zur Vermeidung des Verlustes von Daten und damit auch von Informationen ist es sinnvoll die fehlenden Werte zu ersetzen. Geeignete Einfügestrategien,

[14] Eigene Darstellung, in Anlehnung an [28].

abhängig von den Skalierungsmaßen der Variablen, sind in der Tat erfolgversprechender. Im Replacement-Knoten des SAS® Enterprise Miner™ [2] existiert eine große Bandbreite an Methoden, die zur Einfügung fehlender Werte geeignet sind. Eine Unterscheidung wird an dieser Stelle zwischen den Intervall- und Class-Variablen getroffen.[15].Die Einfügestrategien des Replacement-Knoten des SAS® Enterprise Miner™ reicht von einfachen Mittelwertstatistiken, wie arithmetisches Mittel, Median oder Modus, über Techniken des Entscheidungsbaumverfahrens bis hin zu komplexen Methoden mit M Schätzern[16] [14] wie Tukey's Biweight, Huber oder Andrew's Wave. Außerdem lassen sich Konstanten (Default Constant) z.B. "other" oder sonstige Werte (Set Value) einfügen. Letztendlich wird mit der Option „none" die Complete Case Analysis durchgeführt.

Anstatt eine Mean Imputation zu verwenden, wie arithmetisches Mittel, Modus oder Median, kann eine Einfügestrategie, welche die Durchschnittsberechnung in Abhängigkeit zu den anderen Variablen setzt, eine bessere Schätzung von Missing Values erwirken. Dieses Vorgehen wird Regression Imputation genannt. Hierbei werden k lineare Regressionen berechnet. Jede Input-Variable nutzt dabei die anderen Variablen als Regressoren. Damit besteht nicht nur die Möglichkeit bessere Schätzergebnisse zu erzielen, sondern man erreicht zusätzliche eine Überprüfung, ob eine Abhängigkeit zwischen Missing Value und den Regressor-Variablen besteht. Ein starker Kritikpunkt dieses Verfahrens liegt in der Möglichkeit, dass die Variablen, die als Regressoren fungieren, auch fehlende Werte besitzen. Somit müsste auch hier eine geeignete Einfügestrategie gefunden werden. Besser, da praktikabler ist das Verfahren der Cluster Imputation. Dort werden die Fälle in weitgehend homogene Untergruppen segmentiert. Anschließend wird für jede Gruppe eine Mean Imputation durchgeführt. Neue Fälle (Score-Vorgang) mit Missing Values bekommen den Cluster-Mittelwert, der den existierenden Fällen am nächsten kommt.

Eine zusätzliche Möglichkeit ist das Bilden von Indikator-Variablen, die ergänzend zur der verwendeten Einfügestrategie angeben, dass an dieser Stelle ein Missing Value vorgelegen hat. Dies kann sinnvoll sein, da die

[15] Mittelwertberechnungen können beispielsweise nur bei intervallskalierten Variablen durchgeführt werden.

[16] M Schätzer werden berechnet, um den Einfluss extremer Werte bei der Kennzeichnung der Lage durch ein Lagemaß zu verringern. Die einzelnen Werte werden bei der Berechnung des M Schätzers unterschiedlich gewichtet. Je stärker ein Wert von den übrigen Werten nach oben oder unten abweicht, desto geringer ist das Gewicht mit dem dieser Wert in den M Schätzer eingeht.

fehlenden Werte systematisch mit der Zielvariablen zusammenhängen können und so weitere Informationen beinhalten. [38] So kann in einer Kreditbewertungsanalyse ein Fehlen des Wertes in der Kategorie „Höhe der Verbindlichkeiten“, eng mit der Zielvariablen korreliert sein, da bei entsprechender Schuldenhöhe die Chance auf einen Kredit sinkt.

Der Umgang mit Missing Values hängt entscheidend von dem Modellierungsverfahren ab. Eine Besonderheit stellt in diesem Kontext das Entscheidungsbaumverfahren dar. Entscheidungsbäume als Verfahren der rekursiven Partitionierung ziehen u.a. einen Vorteil aus der Behandlung fehlender Werte. Hierzu sollte zuerst eine Betrachtung der parametrischen Verfahren wie der Regressionsanalyse stattfinden. Parametrische Verfahren verlangen komplette Fälle, ansonsten führt auch schon ein einziger fehlender Wert dazu, dass der gesamte Datensatz eliminiert wird (siehe: Complete Case Analysis). Mit den beschriebenen Einfügestrategien wird der Daten- und Informationsverlust umgangen.

Das Entscheidungsbaumverfahren dagegen sieht Missing Values als zusätzliche Ausprägung der Input Variablen an. Je nach Ausprägung, also ordinal oder nominal und damit ordnungserhaltend oder nicht, ergeben sich unterschiedliche Strategien für die Splitsuche. Eine nominale Inputvariable mit L Ausprägung und fehlenden Werten wird behandelt wie eine mit L+1 Ausprägungen. Dagegen verlangt eine Inputvariable mit ordinalen Ausprägungen eine Modifizierung der Splitsuche[17]. Das Entscheidungsbaumverfahren benötigt also keine Einfügestrategien, da es die fehlenden Werte als eigenständige Ausprägung ansieht.
Einfügestrategien sind auch ein probates Mittel zur Behebung von fehlenden Werten im Score-Vorgang. Hier nimmt das Entscheidungsbaumverfahren wiederum eine besondere Rolle ein: Durch Verwendung von sog. Surrogat-Splits kann das Einfügen von Missing Values gesteuert werden. Ein Surrogat-Split nimmt eine Einteilung aufgrund der Nachahmung von anderen Inputs basierend auf den dafür ausgewählten Splits vor.

5.2 Ausreißer

Werte, die als Ausreißer klassifiziert werden, weisen mindestens eines der folgenden Merkmale auf:

[17] Durch den oder die fehlenden Werte können die ordinalen Inputvariablen nicht mehr in eine ordnungserhaltende Reihenfolge gebracht werden, wie es bei ordinalskalierten Variablen üblicherweise der Fall ist.

I. Die Werte treten einzeln oder selten auf
II. Die Werte befinden sich am Rande des Wertebereichs
III. Die Werte liegen abseits der Mehrheit bzw. Mittelwerte der anderen Ausprägungen

Bei Ausreißern lässt sich eine Unterscheidung zwischen tatsächlich existierenden Werten und fehlerhaften Werten vornehmen. Die drei Merkmalsklassen können sowohl als Ausreißer als auch als Fehler interpretiert werden. Mithilfe des Filter Outliers-Knoten [2] wird der Benutzer in die Lage versetzt einen Datenfilter zu verwenden, der Beobachtungen ausschließt, die mittels diverser Einstellungen[18] in diesem Knoten als Ausreißer klassifiziert wurden. Auf diese Weise wird eine Modellverbesserung durch die erhöhte Stabilität der Parameterschätzungen erreicht. In den Filtern können sowohl statistische Regeln (z.B. außerhalb der x-fachen Standardabweichung) verwendet werden, als auch solche die auf Erfahrungen von Geschäftsbereichen (z.B. Personen, die älter als x Jahre sind) basieren.

Die Analyse von Ausreißern lässt sich auch fernab des SAS® Enterprise Miner™ regeln. Mit verschiedenen SAS-Prozeduren (PROC MEANS, PROC TABULATE und PROC UNIVARIATE) lassen sich schnell Informationen über die Variable gewinnen. {Vgl. im Folgenden: [19]} Auf diese Weise lässt sich eine explorative Datenanalyse durchführen, die mit Hilfe von Mittel- und Extremwerten mögliche Ausreißerkandidaten lokalisiert.[19] Diverse Optionen innerhalb von PROC UNIVARIATE ermöglichen zusätzliche Flexibilität. Mit der Option PCTLPTS können explizit bestimmte Perzentile ausgewählt werden. Durch Abfragen können nun Variablen[20] erzeugt werden, welche diejenigen Beobachtungen identifizieren, die außerhalb dieser Grenzen liegen.

Die Verwendung von Formaten kann dazu eingesetzt werden, bestimmten Wertebereichen der numerischen Variablen und den unterschiedlichen Ausprägungen der Character-Variablen gewisse Bezeichnungen (z.B. „Valid“, „Missing“ oder „Miscoded“) zuzuweisen, die in einem nachfolgenden Data

[18] Percentile, Standardabweichung vom Mittelwert, Abweichung vom Median oder Modus, etc.
[19] Es werden u.a. folgende Statistiken ausgegeben: Mittelwert, Standardabweichung, Varianz, Minimum, Maximum, Spannweite und Anzahl der Beobachtungen. PROC TABULATE ermöglicht zudem eine detailliertere Extremwertbetrachtung. Die fünf größten und kleinsten Werte einer Variablen werden standardmäßig ausgegeben.
[20] Die Option PCTLPRE definiert für die neu erstellten Variablen zur Identifizierung ein Präfix.

Step abgefragt werden können. In Abbildung 4 ist die Vorgehensweise mit PROC FORMAT nachzuvollziehen:

```
PROC FORMAT;
    VALUE $GENDER 'F','M' = 'Valid'
                  ' '     = 'Missing'
                  OTHER   = 'Miscoded';
    VALUE $DX '001' - '999' = 'Valid'
              ' '           = 'Missing'
                  OTHER     = 'Miscoded';
    VALUE $AE '0','1' = 'Valid'
              ' '     = 'Missing'
               OTHER  = 'Miscoded';
RUN;
```

Abbildung 4: PROC FORMAT zur Bestimmung von Ausreißern[21]

Die Behandlung von Ausreißern kann die Eliminierung des Wertes und anschließende Anpassung an die Verteilung der Variablen nach sich ziehen. Dies ist sinnvoll, wenn der Ausreißer keinen nennenswerten Einfluss auf die Zielvariable hat oder sogar falsch ist. So sorgt beispielsweise ab einem gewissen Niveau ein weiterer Anstieg des Einkommens für keine nennenswerten Steigerungen im Konsum. Hat der Ausreißer allerdings einen signifikanten Einfluss auf das Modell und stellt somit also einen interessanten Fall dar, sollte er unverändert im Datensatz bleiben.

5.3 Transformation von Variablen

Parametrische Verfahren wie die Regressionsanalyse erfordern gewisse Annahmen [4], beispielsweise eine Normalverteilung der intervallskalierten Variablen. Eine Normalverteilung ist symmetrisch – nicht-symmetrische Verteilungen weisen eine positive oder negative Schiefe auf (links- bzw. rechtsschief). Eine normal verteilte Variable weist eine Schiefe von Null aus, hier stimmt das arithmetische Mittel mit dem Median überein. In einem weiteren Kriterium der Normalverteilung wird von einer Kurtosis von ebenfalls Null ausgegangen. Die Normalverteilung einer Variablen kann mit dem Kolmogorov-Smirnov Z Test überprüft werden. [8]

[21] Screenshot des Editor-Fensters im SAS-Systems.

Tatsächlich sind viele numerische Variablen schief verteilt. Wenngleich auch eine Reihe von nicht-parametrischen Statistiken zur Verfügung stehen, die weniger starke Annahmen treffen, sollte anstatt auf parametrische Modelle zu verzichten, besser eine Anpassung der Variablen vorgenommen werden. Eine andere statistische Darstellungsform kann z.B. das Problem der Schiefe beheben und die Verteilung der Variablen normalisieren. In diesem Zusammenhang spricht man von einer Transformation der Variablen. Verschiedene statistische Transformationen können bei unterschiedlicher Schiefe eingesetzt werden: [8]

Verteilung ist:	**Transformation durch:**
- stark positiv schief	- negative reziproke Quadratwurzel
- moderat positiv schief	- Logarithmus Naturalis
- leicht positiv schief	- Quadratwurzel
- leicht negativ schief	- Quadrat
- schwer negativ schief	- 3. Potenz

Gemäß dem Central Limit Theorem[22] kann das Problem der fehlenden Normalverteilung vernachlässigt werden, da bei Ansteigen der Stichprobengröße sich die Verteilung an eine Normalverteilung annähert. [8]

Rekodierungen von Variablen können zur Steigerung der Aussagekraft der Variablen führen. {[24], [26] und [29]} So wird beispielsweise die nominale Variable Postleitzahl aufgrund ihrer Ausprägungen im SAS® Enterprise Miner™ als intervallskaliert angesehen. Da diese Variable weder in eine Reihenfolge zu bringen ist, noch ein Verhältnis ausdrückt, sollte an dieser Stelle eine rekodierende Transformation vorgenommen werden. Ordinalskalierte Variablen weisen häufig dieselbe Problematik auf, die durch den Einsatz von Dummy-Variablen oder der Bildung neuer Variablen gelöst werden kann. [28]

Numerische Variablen, die viele verschiedene Ausprägungen mit geringer Anzahl an Beobachtungen aufweisen, können an Aussagekraft gewinnen, wenn einzelne Ausprägungen in Intervalle zusammenfasst werden. {[24], [26] und [29]} Im Transform-Knoten des SAS® Enterprise Miner™ [2] kann hierfür die Transformation „Bucket“ gewählt werden, mit der man die Anzahl der Intervalle und die dazugehörigen Grenzen angeben kann. Für

[22] Vgl. hierzu das Gesetz der großen Zahlen und den zentraler Grenzwertsatz.

drei Intervalle kann man beispielsweise die Ausprägungen niedrig, mittel, hoch erzeugen. Weist die Majorität der Beobachtungen den Wert Null auf, während sich die restlichen Beobachtungen über die anderen Ausprägungen verteilen, macht es Sinn diese numerische Variable in eine binäre zu transformieren.

5.4 Generierung neuer Variablen

Transformationen erstellen neue Variablen auf Basis einer anderen Darstellungsform. Neue Variablen können allerdings auch aus mehreren Variablen generiert werden, wenn diese die Werte verschiedener Variablen der gleichen Thematik aggregieren, um auf diese Weise einen Gesamtüberblick zu ermöglichen. Existieren verschiedene Aktions- und Reaktionsdaten einer Thematik kann mit einer neuen Variablen geprüft werden, ob insgesamt überhaupt reagiert wurde. Zusammengesetzte Variablen haben häufig eine größere Aussagekraft als alle einzelnen Variablen für sich.

5.5 Partitionierung der Daten

Bei der Datenpartitionierung fällt neben der Einteilung der Daten in die Bereiche Training, Validierung und Test auch noch die Entscheidung über die richtige Datengröße[23] hinsichtlich der Fälle[24] an. Massiv große Datenbestände beeinträchtigen die Rechenleistung, so dass es – gerade bei ersten Analyseschritten – Sinn macht die Datenbasis zu verkleinern. Mit dem Sampling-Knoten [2] steht hierfür ein Werkzeug zur Verfügung.[25] Problematischer, da dies die Generalisierungsfähigkeit einschränkt, sind beschränkt kleine Datensätze. Zur Lösung dieses Problemfalles kann das Verfahren der Cross Validation herangezogen werden, um so die Datenbasis künstlich zu vergrößern.

Cross Validation [25] teilt den Datenbestand in n gleich große Datensätze auf. Für den Trainingsprozess stehen n-1 Datensätze zur Verfügung, der n-te wird für die Validierung genutzt. Dieser Prozess wird nun solange wieder-

[23] Die Frage nach der „richtigen“ Stichprobengröße ist nicht exakt zu beantworten und hängt von verschiedenen Faktoren ab: Erwartete Antwortrate der Zielgruppe oder Variablenanzahl. Als Minimum gibt Rud [23] den Faktor 25 pro Variable an. Eine Vergrößerung der Stichprobe verstärkt die Vorhersagekraft der Modelle.

[24] Die Datengröße hinsichtlich der Variablenanzahl wird in Anschnitt 5.6 Selektion der Daten diskutiert.

[25] Des Weiteren kann mit stratifizierten Stichproben das Verhältnis der Event- und Non-Event-Fälle angepasst werden.

holt bis jeder einzelne Datensatz genau einmal für die Validierung eingesetzt wurde. Die resultierenden Ergebnisse werden schließlich gemeinsam für die Modellanpassung genutzt.

5.6 Selektion der Daten

Die Dimension eines Datensatzes steht in Abhängigkeit zu der Anzahl der Inputvariablen. Die benötigte Rechenleistung für die Modellanpassung wird stärker von der Anzahl der Inputvariablen beeinflusst als von der Anzahl der Fälle. [27] Ein weiteres Problem wird von Breiman als der „Fluch der Dimensionalität“ [28] bezeichnet: Eine hohe Dimensionalität begrenzt die Fähigkeit, Beziehungen zwischen Variablen zu entdecken und zu modellieren. Eine höhere Dimensionalität führt zu einem sprunghaften Anstieg der Komplexität der Daten. In diesem Zusammenhang sei auf die Over- bzw. Underfitting-Problematik[26] hingewiesen. [27] Mit Hilfe des Bias[27] und der Varianz[28] kann diese Problemstellung untersucht werden (Bias-Variance-Trade-Off)[29]{[3], [11] und [27]}.

Die Lösung dieses Problems stellt die Dimensionsreduktion dar. Irrelevante oder redundante Variablen werden für die Modellentwicklung ignoriert. Die Dimensionsreduktion kann durchgeführt werden, ohne dass ein Zusammenhang zu dem verwendeten Analyseverfahren (sog. Filter-Verfahren [10]) besteht. Eine Auswahl der Variablen kann so beispielsweise durch das R2- bzw. Chi2-Kriterium des Variable Selection-Knoten [2] vorgenommen werden. Weitere Möglichkeiten in diesem Kontext sind die Hauptkomponentenanalyse[30], das Screening[31] oder das Clustern von Variablen. Außerdem kann das Entscheidungsbaumverfahren dazu verwendet werden, die entscheidungsrelevanten Variablen zu selektieren.

---

[26] Modelle, die „underfitten“ besitzt keine Aussagekraft, sind sog. Null-Modelle. Dagegen bezeichnet Overfitting ein zu komplex gestaltetes Modell, das eine Interpolation der Daten vornimmt.

[27] Der Bias eines Schätzers ist als die Abweichung des Erwartungswertes des Schätzers von dem tatsächlichen Wert definiert.

[28] Die Varianz des Schätzers ist die durchschnittliche quadrierte Abweichung des Schätzers von seinem Erwartungswert.

[29] Modelle mit Underfitting zeichnen sich durch einen hohen Bias und eine geringe Varianz aus, währenddessen Overfitting die Merkmale geringen Bias und hohe Varianz besitzt.

[30] Die Hauptkomponentenanalyse untersucht hauptsächlich die Varianzstruktur von Variablen und ermöglicht damit eine Variablenreduktion.

[31] Jede Variable wird einzeln auf ihren Wirkungszusammenhang zur Zielvariablen untersucht. Einflüsse von Variablen untereinander werden genauso wenig wie Interaktionen zwischen Variablen berücksichtigt.

Modellunterstützende Verfahren (sog. Wrapper-Ansätze [10]) stehen im engen Zusammenhang mit dem verwendeten Modellierungsverfahren: In der Regressionsanalyse unterstützen die Prozeduren Forward[32], Backward[33] und Stepwise[34]den Variablenauswahlprozess. {[34] und [28]}

6 Data Quality Mining

Der Begriff Data Quality Mining (DQM) wurde von Hipp, Güntzer und Grimmer [13] geprägt. Das Ziel des DQM ist mit Hilfe von Data Mining-Methoden Datenqualitätsmängel zu erkennen, zu bewerten, zu erklären und abschließend zu korrigieren. Die Muster, Strukturen und Zusammenhänge von Dateninkonsistenzen zu entdecken, ist eine Aufgabenstellung die gerade durch Data Mining-Verfahren sehr gut gelöst werden kann. In diesem Kapitel wird beschrieben, inwieweit die Cluster- und Assoziationsanalyse sowie die Vorhersage- und Klassifikationsmodellen in diesem neuen Ansatz Anwendung finden.

Datenqualitätsmaßnahmen prüfen i.d.R. den oder die Fehlerkandidaten einer Variablen anhand der übrigen Werte dieser Variablen. Datenfehler sind allerdings häufig nur dann zu identifizieren, wenn man die Wechselwirkung mit anderen Variablen berücksichtigt.

Die Herangehensweise des DQM, Data Mining-Verfahren für die Gewährleitung von Datenqualität heranzuziehen, wurde in verschiedenen Ansätzen unter alternativen Bezeichnungen weiterentwickelt bzw. um neue Aspekte ergänzt.

Exploratory Data Mining

Das Exploratory Data Mining (EDM) [7] stellt streng genommen keinen neuen Ansatz dar, da hierfür nur Elemente des Pre-Processings, der Visualisierung sowie der explorativen Datenanalyse verwendet werden. Diese Aufgaben können im SAS® Enterprise Miner™ mit den Knoten der Explore-

[32] Bei der „Forward Selection“ wird diejenige Variable als erste in die Regression aufgenommen, welche die größte Korrelation mit der abhängigen Variablen aufweist. Jede weitere Aufnahme einer Variablen wird dann wiederum aufgrund der Wechselbeziehung zwischen Input- und Ziel-Variable entschieden.

[33] Die „Backward Selection“ nimmt zuerst alle Variablen in das Modell auf und eliminiert danach Variablen auf der Basis von Signifikanz-Tests.

[34] Die „Stepwise Selection“ ist eine Variante der Forward-Prozedur, wobei einzelne Variablen in späteren Schritten wieder eliminiert werden können.

und Modify-Gruppe [2] gelöst werden und sind somit in der SEMMA-Methode verankert. Häufig werden mit EDM die Daten vorverarbeitet, um anschließend beispielsweise eine Klassifikation der Variablen durchzuführen. EDM muss allerdings nicht zwangsläufig Bestandteil einer Data Mining-Analyse mit anschließender Modellierung sein, sondern kann auch autonom eingesetzt werden, um vorhandene Datenmängel zu entdecken.

Clusteranalyse

Segmentierungen, mit den Forderungen nach Homogenität der Objekte innerhalb der Cluster und Heterogenität zwischen den Clustern [34], ermöglichen eine sehr Kontext bezogene Analyseform, verglichen mit Analyseverfahren, welche die Grundgesamtheit der Daten heranziehen. Das K-Means-Clusterverfahren konstruiert für eine Anzahl von K Clustern Clusterzentren. Der Modellansatz [5] besteht darin, dass die Clusterzentren der K Cluster so berechnet werden, dass die Streuungsquadratsumme in den Clustern minimiert wird. Dies entspricht der quadrierten euklidischen Distanz $d_{g,k}^2$ zwischen dem Objekt g und dem Clusterzentrum k. Diese Minimierungsaufgabe lässt sich auch als Fehlerstreuung interpretieren, also der Streuung in den Daten, die nicht durch die Cluster erklärt werden. Auf diese Weise wird ein Bezug zu Datenqualitätsmaßnahmen hergestellt.

Zuerst sollten die gebildeten Cluster hinsichtlich ihrer Größe betrachtet werden. Cluster mit einer relativ kleinen Größe sind potentielle Fehlerkandidaten. [33] Allerdings können kleine Cluster auch ein Nischensegment darstellen, wobei an dieser Stelle eine Unterscheidung zwischen besonders interessanten und uninteressanten Fällen zu treffen ist. Große Cluster können innerhalb der Attribute ebenfalls fehlerhaft sein. Dies kann überprüft werden, indem um jedes der Merkmale ein Konfidenzintervall (z.B. 95 oder 99 Prozent) gelegt wird, um auf diese Weise Randwerte zu finden, die als Ausreißer zu interpretieren sind.

Assoziationsanalyse

Durch die Assoziationsanalyse kann ein Fall bzw. eine Datenzeile als falsch erkannt werden, die lediglich aufgrund ihrer Zusammensetzung, d.h. aufgrund von Abhängigkeiten, als Fehlerkandidat zu identifizieren ist. Um das Assoziationsproblem formal zu beschreiben, {Vgl. im Folgenden [31]} betrachtet man eine Datenmenge D von Transaktionen t. Jede Transaktion besteht aus einer Menge einzelner Elemente mit jeweils unterschiedlicher

Häufigkeit des Auftretens. Eine Assoziationsregel wird mit X ➔ Y bezeichnet, d.h. das Auftreten von Element X führt zu dem Auftreten des Elements Y. Element X befindet sich im Regelrumpf, ist also die Wenn-Bedingung. Element Y wird in den Regelkopf gesetzt und entspricht folglich der Dann-Bedingung. Damit eine Transaktion die Assoziationsregel X ➔ Y erfüllt, muss $(X \cup Y) \subseteq t$ gelten, es müssen also alle Elemente in der Assoziationsregel enthalten sein. Für die Bewertung einer Regel existieren drei verschiedene Kriterien: Support, Konfidenz und Lift.

Die Konfidenz der Assoziationsregel beschreibt, für welchen Anteil der Transaktionen, die X enthalten, die Assoziationsregel X ➔ Y gilt.

$$\textit{Konfidenz (X} \rightarrow \textit{Y)} = \frac{\left|\{t \in D \mid (X \cup Y) \subseteq t\}\right|}{\left|t \in D \mid X \subseteq t\right|}. \text{ [31]} \qquad (7)$$

Das Konfidenz-Kriterium wird herangezogen, um eine Reihe von Transaktionen zu bewerten, die alle einer Entität zugehörig sind.

Hipp, Güntzer und Grimmer {Vgl. im Folgenden [13]} verwenden PKW-Typen und die dazugehörigen Ausstattungen, um den dargestellten Ansatz zu demonstrieren. Für einen bestimmten Autotyp, z.B. S-Klasse, werden pro mögliche Ausstattung Assoziationsregeln aufgestellt und die zugehörige Konfidenz berechnet. Desto höher dieser Wert ausfällt, umso glaubwürdiger wird die Kombination PKW-Typ ➔ Ausstattung. Eine niedrige Konfidenz einer Assoziationsregel bedeutet allerdings nicht, dass dies ein Datenfehler ist. Ein S-Klasse PKW mit einem Dieselmotor wird eine niedrigere Konfidenz besitzen als einer mit einem Benzinmotor. Ein Dieselmotor ist aber kein Datenfehler, sondern vielmehr ein seltener Fall. Da ein PKW aber mehrere Ausstattungen enthält, können für einen PKW mehrere Assoziationsregeln gebildet werden, deren Konfidenz zusammen in die Bewertung einfließen. Wie bei der Clusteranalyse ist auch bei dem Konglomerat von Assoziationsregeln die Unterscheidung zwischen Nischenprodukten und Fehlern zu treffen.

Predictive Modeling

Vorhersage- und Klassifikationsmodelle bilden mit Hilfe einer historischen Datenbasis, für welche die Klassifikation der Zielvariablen bekannt ist, ein Modell, dass für den Scoring-Prozess verwendet werden kann.

Im DQM kommen Verfahren dieser Data Mining-Klasse beispielsweise im Call Center-Bereich oder bei Online-Fragebögen zum Einsatz, um Lücken in den Datensätzen zu schließen. Fehlende Einträge sensitiver Daten, wie Alter oder Nationalität, können auf Basis eines Prognosemodells geschätzt werden. Als Regressor kann hier der Vorname herangezogen werden. Weitere soziodemographische Variablen, beispielsweise von Marktforschungsunternehmen, können des Weiteren zu Modellverbesserung beitragen.

Die Wahrscheinlichkeit einer Non-Response[35] bzw. einer falschen Angabe ist bei der Variable Einkommen besonders hoch. Aufgrund ihrer Vorhersagefähigkeit ist sie jedoch in vielen Analysen von entscheidender Bedeutung, so dass auch hier eine Schätzung sinnvoll ist, um anschließend wiederum bessere Modelle bilden zu können. Das Einkommen steht im engen Zusammenhang mit u.a. folgenden Variablen: Alter, Berufsjahre, Berufskategorie, Hausbesitzer, Wohnbezirk sowie gegebenenfalls registriertem Kaufverhalten.

Regelinduktion

Regelbasierte Data Mining-Werkzeuge {Vgl. im Folgenden [33]} ermöglichen die automatische Generierung von Business Rules. Die theoretische Grundlage für regelbasierte Verfahren bildet die Regelinduktion. Durch die Messung von statistischen Signifikanzen werden bestimmte Gegebenheiten in den Daten als „Wenn-Dann"-Regeln formuliert. Fehlerhafte Daten haben zur Folge, dass nicht alle Regeln vollständig zutreffen. Um die Regeln dennoch identifizieren zu können, berücksichtigt man bei der Regelinduktion Fehlertoleranzen – häufig Toleranzschwellen zwischen 95 und 99 Prozent. Abweichungen von den Regeln können so zur Identifikation von Fehlerkandidaten genutzt werden. Würden keine Fehlertoleranzen berücksichtigt, könnten über die Regelinduktion auch keine Fehlerkandidaten ermittelt werden.

7 Fazit

Die Einflüsse der Datenqualität lassen sich unternehmensweit feststellen, besonders deutlich treten sie allerdings bei Datenanalyseverfahren, wie dem Data Mining, hervor. Die notwendige Vorverarbeitung der Daten nimmt beim Aufbau einer Data Mining-Analyse einen nicht unbeträchtlichen Zeit-

[35] d.h. fehlender Wert.

anteil ein und hat entscheidenden Einfluss auf die Qualität der generierten Modelle.

Die Behebung von Datenqualitätsmängeln kann mit unterschiedlichen Konzepten an verschiedenen Ansatzpunkten verwirklicht werden. Häufig entsteht dabei ein uneinheitliches Vorgehen: unterschiedliche Maßnahmen von verschiedenen Personen, Korrekturen nur bei einem Teil der Daten, fehlende Dokumentation oder lediglich reaktives Handeln. Ein ganzheitliches Konzept ist daher von entscheidender Bedeutung, was mit der vorgestellten Methodologie dann auch realisiert wurde.

Ist eine Data Warehouse-Architektur in den Data Mining-Ablauf integriert, werden mit Hilfe von ETL-Prozessen und dem Verständnis der Daten, welches durch die Metadaten erzeugt wird, erste Korrekturmaßnahmen getroffen und schwerwiegende Datenfehler behoben. Eine Data Mining-Analyse benötigt allerdings darüber hinaus gehende Anpassungen, die dann im Pre-Processing unbedingt durchgeführt werden sollten. Die in dieser Arbeit erläuterten Aufgabenstellungen (fehlende Werte, Ausreißer, Transformation, Selektion und Generierung von Variablen sowie Partitionierung der Daten) sind bei einer Data Mining-Analyse nach ihrer Notwendigkeit zu prüfen und gegebenenfalls durchzuführen.

Der neue Ansatz des DQM bietet neue Möglichkeiten bei der Behebung von Datenmängeln. Da mit Hilfe des DQM die Datenqualität im Allgemeinen und für eine Data Mining-Analyse im Besonderen durch die Entdeckung von unbekannten Zusammenhängen und Strukturen gesteigert werden kann, sollte dieser Ansatz in die Lösungsexpertise zur Gewährleistung von Datenqualität integriert werden.

Literatur

[1] A SAS White Paper (o.J.): Exponentially Enhance the Quality of Your Data with SAS ETL^{Q}, Cary.

[2] A SAS White Paper (o.J.): Finding the Solution to Data Mining, Cary.

[3] Anders, U. (1995): Neuronale Netze in der Ökonometrie, Discussion Paper, Mannheim.

[4] Backhaus, K.; Erichson, B.; Plinke, W. und Weiber, W. (2000): Mutlivariate Analysemethoden: Eine anwendungsorientierte Einführung, Berlin, Heidelberg, New York.

[5] Badner, J. (1994): Clusteranalyse: Eine anwendungsorientierte Einführung, München.

[6] Chantala, K. and Suchindran, C. (o.J.): Multiple Imputation for Missing Data, www.cpc.unc.edu/services/computer/presentations/mi_presentation2.pdf (Stand: 23.10.2003). o.O.

[7] Dasu, T. und Johnson, T. (2003): Exploratory Data Mining and Data Cleansing, Hoboken, New Jersey.

[8] De Vaus, D.A (2000): Analysing Social Science Data, London.

[9] English, L.P. (1999): Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Improving Profits, New York.

[10] Freitas, A. (2002): Data Mining and Knowledge Discovery with Evolutionary Algorithms, Berlin, Heidelberg, New York.

[11] Hastie, T.; Tibshirani, R. and Friedman, J.H. (2001): The Elements of Statistical Learning: Data Mining, Inference and Prediction, Berlin, Heidelberg, New York.

[12] Helfert, M.; Winter, M. und Herrmann, C. (o.J): Datenqualitäts-management für Data Warehouse-Systeme – Technische und organisatorische Realisierung am Beispiel der Credit Suisse, St. Gallen.

[13] Hipp, J.; Güntzer, U. und Grimmer, U. (2001): Data Quality Mining – Making A Virtue of Necessity. In Proc.of the 6th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD 2001), pages 52-57, Santa Barbara, California.

[14] Hogg, R.V. and Craig, A.V. (1978): Introduction to Mathematical Statistics, New York.

[15] Hox, J. and de Leeuw, E. (1998): Ad Hoc Solutions for Missing Values, www.fss.uu.nl/ms/jh/papers/zad3p1.pdf (Stand: 23.10.2003). o.O.

[16] Inmon, W. (1993): Building the Data Warehouse, New York.

[17] Jarke, M.; Jeusfeld, M.; Quix, C. and Vassiliou, Y. (1999): Architecture and Quality in Data Warehouses: An Extended Repository Approach. In Information Systems, (24. Jg.), Nr. 11, o.O.

[18] Krahl, D. (1998): Data Mining, Bonn.

[19] Kroll, F. (2003): Analyse von Ausreißern mit der Base SAS Software, Vortrag auf disc03, Bonn.

[20] Küppers, B. (1999): Data Mining in der Praxis, Frankfurt am Main.

[21] Pyle, D. (2001): Data Preparation for Data Mining, San Francisco.

[22] Roßbach, P. und Moormann, P.(Hg.) (2001): Mikromarketing, Data Warehouse und Data Mining im CRM. In Customer Relationship Management in Banken, Frankfurt.

[23] Rud, O.P. (2001): Data Mining Cookbook, New York.

[24] SAS Institute Inc. (2002): Applying Data Mining Techniques Using Enterprise Miner™, Cary.

[25] SAS Institute Inc. (2000): Decision Tree Modeling, Cary.

[26] SAS Institute Inc. (2000): Getting Started with the Enterprise Miner™, Cary.

[27] SAS Institute Inc. (2000): Neural Network Modeling, Cary.

[28] SAS Institute Inc. (2000): Predictive Modeling Using Logistic Regression, Cary.

[29] SAS Institute Inc. (2000): Using Enterprise Miner™ Software: A Case Study Approach, Cary.

[30] SAS Institute Inc. (2001): Warehouse Architecture, Cary.

[31] Schinzer, H.; Bange, C. und Mertens, H. (1999): Data Warehouse und Data Mining: Marktführende Produkte im Vergleich, München.

[32] Schütte, R.; Rotthowe, T. und Holten R. (Hg.) (2001): Data Warehouse Management-Handbuch: Konzepte, Software, Erfahrungen, Berlin, Heidelberg, New York.

[33] Seidl, J. and de Vries, D. (2001): Proaktives Datenqualitäts-management als zentrale Aufgabe im Data Warehouse-Prozess, 6. Data Warehouse-Forum, St. Gallen.

[34] Stier, W. (1999): Empirische Forschungsmethoden, Berlin, Heidelberg, New York.

[35] The Data Warehouse Institute (2002.): Data Quality and the Bottom Line – Achieving Business Success through a Commitment to High Quality Data, Chatsworth.

[36] Wand, Y. and Wang, R.Y. (1996): Anchoring Data Quality Dimensions in Ontological Foundations. In Communications of the ACM, (39. Jg.), Nr. 11, o.O.

[37] Wang, R.Y. and Strong, D.M. (1996): Beyond Accuracy: What Data Quality Means to Data Consumers. In Journal of Management Information Systems, (12. Jg.), Nr. 4, o.O.

[38] Witten, I.H. und Frank, E. (2001): Data Mining – Praktische Werkzeuge und Techniken für das maschinelle Lernen, München.