IBM SPSS Modeler
CRISP-DM Handbuch

Hinweis

Vor Verwendung dieser Informationen und des darin beschriebenen Produkts sollten die Informationen unter „Bemerkungen" auf Seite 41 gelesen werden.

Produktinformation

Diese Ausgabe bezieht sich auf Version 17, Release 1, Modifikation 0 von IBM(r) SPSS(r) Modeler und alle nachfolgenden Releases und Modifikationen, bis dieser Hinweis in einer Neuausgabe geändert wird.

Diese Veröffentlichung ist eine Übersetzung des Handbuchs
IBM SPSS Modeler CRISP-DM Guide,
herausgegeben von International Business Machines Corporation, USA

Informationen, die nur für bestimmte Länder Gültigkeit haben und für Deutschland, Österreich und die Schweiz nicht zutreffen, wurden in dieser Veröffentlichung im Originaltext übernommen.

Möglicherweise sind nicht alle in dieser Übersetzung aufgeführten Produkte in Deutschland angekündigt und verfügbar; vor Entscheidungen empfiehlt sich der Kontakt mit der zuständigen IBM Geschäftsstelle.

Änderung des Textes bleibt vorbehalten.

Herausgegeben von:
TSC Germany
Kst. 2877
August 2015

Inhaltsverzeichnis

Vorwort

IBM® SPSS Modeler ist die auf Unternehmensebene einsetzbare Data-Mining-Workbench von IBM. Mit SPSS Modeler können Unternehmen und Organisationen die Beziehungen zu ihren Kunden bzw. zu den Bürgern durch ein tief greifendes Verständnis der Daten verbessern. Organisationen verwenden die mithilfe von SPSS Modeler gewonnenen Erkenntnisse zur Bindung profitabler Kunden, zur Ermittlung von Cross-Selling-Möglichkeiten, zur Gewinnung neuer Kunden, zur Ermittlung von Betrugsfällen, zur Reduzierung von Risiken und zur Verbesserung der Verfügbarkeit öffentlicher Dienstleistungen.

Die grafische Schnittstelle von SPSS Modeler erleichtert Benutzern die Anwendung ihres spezifischen Fachwissens, was zu leistungsfähigeren Vorhersagemodellen führt und die Zeit bis zur Lösungserstellung verkürzt. SPSS Modeler bietet zahlreiche Modellierungsverfahren, beispielsweise Algorithmen für Vorhersage, Klassifizierung, Segmentierung und Assoziationserkennung. Nach der Modellerstellung ermöglicht IBM SPSS Modeler Solution Publisher die unternehmensweite Bereitstellung des Modells für Entscheidungsträger oder in einer Datenbank.

Informationen zu IBM Business Analytics

Die Software IBM Business Analytics liefert umfassende, einheitliche und korrekte Informationen, mit denen Entscheidungsträger die Unternehmensleistung verbessern können. Ein umfassendes Portfolio aus Anwendungen fürBusiness Intelligence, Vorhersageanalyse, Finanz- und Strategiemanagement sowie Analysen bietet Ihnen sofort klare und umsetzbare Einblicke in die aktuelle Leistung und gibt Ihnen die Möglichkeit, zukünftige Ergebnisse vorherzusagen. Durch umfassende Branchenlösungen, bewährte Vorgehensweisen und professionellen Service können Unternehmen jeder Größe die Produktivität maximieren, Entscheidungen automatisieren und bessere Ergebnisse erzielen.

Als Teil dieses Portfolios unterstützt IBM SPSS Predictive Analytics-Software Unternehmen dabei, zukünftige Ereignisse vorherzusagen und proaktiv Maßnahmen zu ergreifen, um bessere Geschäftsergebnisse zu erzielen. Kunden aus Wirtschaft, öffentlichem Dienst und staatlichen Lehr- und Forschungseinrichtungen weltweit nutzen IBM SPSS-Technologie als Wettbewerbsvorteil für die Kundengewinnung, Kundenbindung und Erhöhung der Kundenumsätze bei gleichzeitiger Eindämmung der Betrugsmöglichkeiten und Minderung von Risiken. Durch die Einbindung von IBM SPSS-Software in ihre täglichen Operationen wandeln sich Organisationen zu "Predictive Enterprises", die Entscheidungen auf Geschäftsziele ausrichten und automatisieren und einen messbaren Wettbewerbsvorteil erzielen können. Wenn Sie weitere Informationen wünschen oder Kontakt zu einem Mitarbeiter aufnehmen möchten, besuchen Sie die Seite http://www.ibm.com/spss.

Technical Support

Kunden mit Wartungsvertrag können den Technical Support in Anspruch nehmen. Kunden können sich an den Technical Support wenden, wenn sie Hilfe bei der Arbeit mit IBM Produkten oder bei der Installation in einer der unterstützten Hardwareumgebungen benötigen. Zur Kontaktaufnahme mit dem Technical Support besuchen Sie die IBM Website unter http://www.ibm.com/support. Sie müssen bei der Kontaktaufnahme Ihren Namen, Ihre Organisation und Ihre Supportvereinbarung angeben.

Kapitel 1. Einführung in CRISP-DM

CRISP-DM-Hilfe - Übersicht

CRISP-DM steht für Cross-Industry Standard Process for Data-Mining und ist eine in der Branche bewährte Methode zur Anleitung Ihrer Data-Mining-Arbeit.

- Als **Methodologie** umfasst CRISP-DM Beschreibungen der typischen Phasen eines Projekts, die in jeder Phase auszuführenden Arbeiten und eine Erläuterung der Beziehungen zwischen diesen Aufgaben.
- Als **Prozessmodell** bietet CRISP-DM einen Überblick über den Data-Mining-Lebenszyklus.

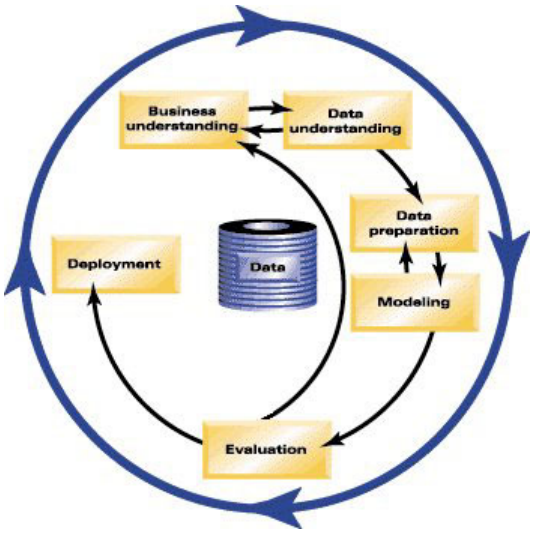

Abbildung 1. Data-Mining-Lebenszyklus

Das Lebenszyklusmodell setzt sich aus sechs Phasen zusammen. Mit Pfeilen werden die wichtigsten und häufigsten Abhängigkeiten zwischen den Phasen dargestellt. Die Reihenfolge der Phasen ist variabel. Es ist sogar so, dass in den meisten Projekten bei Bedarf zwischen den einzelnen Phasen vor- und zurückgewechselt wird.

Das CRISP-DM-Modell ist flexibel und kann einfach an die jeweiligen Bedürfnisse angepasst werden. Wenn Ihr Unternehmen beispielsweise Geldwäsche aufdecken möchte, ist es sehr wahrscheinlich, dass Sie große Datenmengen ohne spezifisches Modellierungsziel durchsuchen. Statt auf die Modellierung konzentriert sich Ihre Arbeit auf die Exploration und Visualisierung von Daten, um verdächtige Muster bei Finanzdaten aufzudecken. Mit CRISP-DM können Sie ein Data-Mining-Modell erstellen, das Ihre spezifischen Anforderungen erfüllt.

In solchen Fällen sind die Modellierungs-, Evaluierungs- und Bereitstellungsphasen möglicherweise weniger relevant als die Phasen zur Aufbereitung und Interpretation von Daten. Es ist jedoch nach wie vor wichtig, einige der in späteren Phasen aufgeworfenen Fragen für die langfristige Planung und zukünftige Data-Mining-Ziele zu berücksichtigen.

CRISP-DM in IBM SPSS Modeler

IBM SPSS Modeler berücksichtigt die CRISP-DM-Methodologie auf zwei Arten, um einmalige Unterstützung für ein effektives Data-Mining zu bieten.

- Das CRISP-DM-Projekttool hilft Ihnen bei der Organisation von Projektstreams, Ausgaben und Anmerkungen entsprechend den Phasen eines typischen Data-Mining-Projekts. Sie können zu jedem Zeitpunkt des Projekts anhand der Knoten für Streams und CRISP-DM-Phasen Berichte erstellen.
- Die Hilfe zu CRISP-DM führt Sie durch die Prozesse eines Data-Mining-Projekts. So umfasst sie Aufgabenlisten für jeden Schritt sowie Beispiele für den Einsatz von CRISP-DM in der Praxis. Sie können durch Klicken auf **CRISP-DM-Hilfe** im Hauptfenster des Hilfemenüs auf die CRISP-DM-Hilfe zugreifen.

CRISP-DM-Projekttool

Das CRISP-DM-Projekttool gewährt einen strukturierten Ansatz beim Data-Mining, der den Erfolg Ihres Projekts sicherstellen kann. Es ist im Grunde genommen eine Erweiterung des standardmäßigen IBM SPSS Modeler-Projekttools. Sie können sogar zwischen der CRISP-DM-Ansicht und der Standardklassenansicht umschalten, um Ihre Streams und Ausgaben nach Typ oder nach CRISP-DM-Phasen organisiert anzuzeigen.

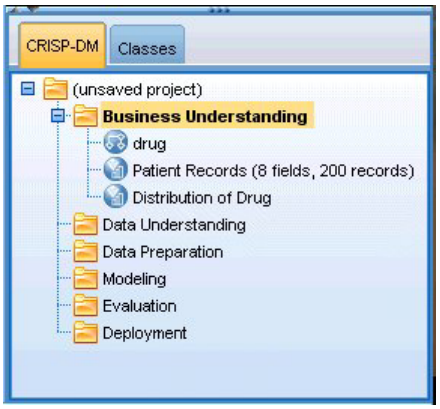

Abbildung 2. CRISP-DM-Projekttool

Mit der CRISP-DM-Ansicht des Projekttools können Sie folgende Aufgaben ausführen:

- Organisieren Sie Projektstreams und Ausgaben entsprechend den Data-Mining-Phasen.
- Zeichnen Sie die Ziele auf, die Ihr Unternehmen für jede Phase hat.
- Erstellen Sie benutzerdefinierte QuickInfos für jede Phase.
- Zeichnen Sie die Schlussfolgerungen auf, die aus einem bestimmten Diagramm oder Modell gezogen werden.
- Erzeugen Sie einen HTML-Bericht oder eine Aktualisierung zur Verteilung an das Projektteam.

Hilfe für CRISP-DM

IBM SPSS Modeler bietet ein Onlinehandbuch für das nicht gesetzlich geschützte CRISP-DM-Prozessmodell. Dieses Handbuch ist nach Projektphasen organisiert und bietet folgende Unterstützung:

- Überblick und Aufgabenliste für jede Phase von CRISP-DM
- Hilfe beim Erstellen von Berichten für verschiedene Etappen
- Praxisbeispiele, die zeigen, wie ein Projektteam mit CRISP-DM den Weg für das Data-Mining ebnen kann
- Verknüpfungen zu weiteren Ressourcen unter CRISP-DM

Sie können durch Klicken auf **CRISP-DM-Hilfe** im Hauptfenster des Hilfemenüs auf die CRISP-DM-Hilfe zugreifen.

Zusätzliche Ressourcen

Zusätzlich zum IBM SPSS Modeler-Support für CRISP-DM können Sie Ihr Wissen über Data-Mining-Prozesse auf verschiedensten Wegen erweitern.

- Besuchen Sie die Website des CRISP-DM Consortiums unter www.crisp-dm.org
- Lesen Sie das vom CRISP-DM Consortium verfasste CRISP-DM-Handbuch, das in dieser Version enthalten ist.
- Lesen Sie *Data Mining with Confidence*, Copyright 2002 von SPSS Inc., ISBN 1-56827-287-1.

Kapitel 2. Untersuchung der Geschäftsziele

Untersuchung der Geschäftsziele - Übersicht

Bereits vor der Arbeit mit IBM SPSS Modeler sollten Sie sich die Zeit nehmen und herausfinden, welchen Nutzen sich Ihr Unternehmen vom Data-Mining verspricht. Versuchen Sie, so viele Personen wie möglich in diese Diskussionen einzubeziehen und halten Sie die Ergebnisse fest. Im letzten Schritte dieser CRISP-DM-Phase wird erläutert, wie mithilfe der hier gesammelten Informationen ein Projektplan erstellt wird.

Auch wenn Ihnen diese Untersuchungen überflüssig erscheinen sollten, sie sind es nicht. Wenn die geschäftlichen Gründe für Ihre Data-Mining-Arbeit bekannt sind, haben alle Personen denselben Kenntnisstand, bevor wertvolle Ressourcen aufgewendet werden.

Bestimmen von Geschäftszielen

Ihre erste Aufgabe besteht darin, so viel Einblick wie möglich in die Geschäftsziele für das Data-Mining zu erhalten. Das sieht möglicherweise einfacher aus als es ist. Sie können jedoch später auftretende Risiken minimieren, wenn Sie Probleme, Ziele und Ressourcen klar definieren.

Die CRISP-DM-Methodologie bietet eine strukturierte Methode, wie Sie dies erreichen können.

Aufgabenliste

- Beginnen Sie, Hintergrundinformationen über die aktuelle Geschäftssituation zu sammeln.
- Dokumentieren Sie spezifische Geschäftsziele, die von Hauptentscheidungsträgern festgelegt werden.
- Vereinbaren Sie Kriterien, mit denen der Data-Mining-Erfolg aus Sicht des Unternehmens bestimmt wird.

Beispiel aus dem elektronischen Einzelhandel - Ermitteln von Geschäftszielen

Ein Web-Mining-Szenario mit CRISP-DM

Da immer mehr Unternehmen zum Verkauf im Internet übergehen, müssen etablierte Computer-/Elektronikhändler, die ihre Waren auf dem elektronischen Weg verkaufen (e-Einzelhändel), mit zunehmendem Wettbewerb von neueren Sites rechnen. Wenn das Unternehmen dann mit der Wirklichkeit konfrontiert wird, dass Web-Stores so schnell aus dem Boden sprießen, wie Kunden sich dem Internet zuwenden (oder schneller!), muss es Wege finden, um trotz der steigenden Kosten bei der Kundenakquisition gewinnbringend zu bleiben. Einer der Lösungsvorschläge ist es, die bestehenden Kundenbeziehungen auszubauen, um den Wert jedes aktuellen Kunden des Unternehmens zu maximieren.

Deshalb wird eine Studie mit den folgenden Zielen durchgeführt:

- Steigern der Cross-Sales durch bessere Empfehlungen.
- Verbessern der Kundenloyalität durch persönlichere Dienstleistungen.

Die Studie wird versuchsweise als erfolgreich angesehen, wenn folgende Punkte erfüllt sind:

- Cross-Sales steigen um 10 %.
- Kunden verbringen pro Internetbesuch mehr Zeit auf der Site oder rufen mehr Seiten auf.
- Studie wird pünktlich und innerhalb des Budgetrahmens abgeschlossen.

Erarbeiten des Geschäftshintergrunds

Wenn Sie die geschäftliche Situation Ihres Unternehmens kennen, wissen Sie auch, womit Sie arbeiten können, bezogen auf:

- Verfügbare Ressourcen (Personal und Material)
- Probleme
- Ziele

Dazu müssen Sie die aktuelle Geschäftssituation untersuchen, um echte Antworten auf Fragen zu erhalten, die das Ergebnis des Data-Mining-Projekts beeinflussen können.

Aufgabe 1 - Bestimmen der Organisationsstruktur

- Entwickeln Sie Organisationsdiagramme, um Unternehmensbereiche, Abteilungen und Projektgruppen darzustellen. Vergessen Sie nicht, die Namen und Verantwortlichkeiten der Manager mit aufzunehmen.
- Ermitteln Sie zentrale Personen im Unternehmen.
- Ermitteln Sie einen internen Sponsor, der finanzielle Unterstützung gewährt und/oder über Domänenwissen verfügt.
- Bestimmen Sie, ob ein Lenkungsausschuss existiert, und legen Sie eine Liste der Mitglieder bei.
- Ermitteln Sie Geschäftseinheiten, die vom Data-Mining-Projekt betroffen sein werden.

Aufgabe 2 - Beschreiben von Problembereichen

- Ermitteln Sie die Problembereiche wie Marketing, Kundenbetreuung oder wirtschaftliche Entwicklung.
- Beschreiben Sie das Problem ganz allgemein.
- Definieren Sie die Voraussetzungen des Projekts. Welche Absichten liegen dem Projekt zugrunde? Verwendet das Unternehmen bereits Data-Mining?
- Überprüfen Sie den Status des Data-Mining-Projekts innerhalb der Unternehmensgruppe. Wurde die Arbeit bereits zugelassen oder muss Data-Mining als wichtige Technologie für die Unternehmensgruppe "angepriesen" werden?
- Bereiten Sie bei Bedarf informative Präsentationen über Data-Mining für Ihr Unternehmen vor.

Aufgabe 3 - Beschreiben der aktuellen Lösung

- Beschreiben Sie alle Lösungen, die derzeit zur Behebung des Geschäftsproblems eingesetzt werden.
- Beschreiben Sie die Vor- und Nachteile der aktuellen Lösung. Beziehen Sie sich dabei auch auf den Akzeptanzgrad, den diese Lösung innerhalb des Unternehmens erreicht hat.

Definieren von Geschäftszielen

Jetzt wird es konkret. Als Ergebnis Ihrer Untersuchungen und Besprechungen sollten Sie ein konkretes primäres Ziel ausarbeiten, dem die Projektsponsoren und anderen Geschäftseinheiten, die von den Ergebnissen beeinflusst werden, zustimmen. Dieses Ziel wird sich im Laufe der Zeit von solch einem unpräzisen Ziel wie "Reduzierung der Kundenabwanderungsrate" zu spezifischen Data-Mining-Zielen entwickeln, die Ihre Analysen steuern werden.

Aufgabenliste

Vergessen Sie nicht, sich Notizen zu folgenden Punkten zu machen, um sie später in den Projektplan einbauen zu können. Formulieren Sie nur realistische Ziele.

- Beschreiben Sie das Problem, das Sie mithilfe von Data-Mining lösen möchten.
- Geben Sie alle Unternehmensfragen so genau wie möglich an.
- Bestimmen Sie alle anderen Unternehmensanforderungen (z. B. keine aktuellen Kunden verlieren, während die Cross-Selling-Möglichkeiten gesteigert werden).

- Legen Sie voraussichtliche Vorteile für das Unternehmen fest (z. B. Reduzierung der Abwanderung bei hochwertigen Kunden um 10 %).

Kriterien für den Unternehmenserfolg

Das große Ziel ist zwar möglicherweise klar, aber wissen Sie auch, wann Sie es erreicht haben? Es ist wichtig, die Art des Unternehmenserfolgs für Ihr Data-Mining-Projekt zu definieren, bevor Sie fortfahren. Erfolgskriterien unterteilen sich in zwei Kategorien:

- **Objektive Kriterien.** Diese Kriterien können so einfache Punkte wie eine bestimmte Erhöhung der Genauigkeit von Audits oder eine vereinbarte Reduzierung der Abwanderung sein.
- **Subjektive Kriterien.** Subjektive Kriterien wie die "Ermittlung von Clustern für effektive Behandlungen" sind schwieriger festzulegen. Sie können aber vereinbaren, wer die endgültige Entscheidung trifft.

Aufgabenliste

- Dokumentieren Sie die Erfolgskriterien für dieses Projekt so genau wie möglich.
- Achten Sie darauf, dass jedem Geschäftsziel ein entsprechendes Erfolgskriterium zugeordnet ist.
- Stellen Sie sicher, dass sich die Schiedspersonen der subjektiven Erfolgsmessung untereinander abstimmen. Notieren Sie sich möglichst ihre Erwartungen.

Bewerten der Situation

Nachdem Sie nun die Ziele klar definiert haben, ist es an der Zeit, die aktuelle Situation zu bewerten. In diesen Schritt gehören Fragen wie:

- Welche Art von Daten sind für die Analyse verfügbar?
- Verfügen Sie über das erforderliche Personal, um das Projekt abzuschließen?
- Was sind die größten Risikofaktoren bei diesem Projekt?
- Besitzen Sie einen Alternativplan für jedes Risiko?

Beispiel aus dem elektronischen Einzelhandel - Bewerten der Situation

Ein Web-Mining-Szenario mit CRISP-DM

Der e-Einzelhändler nutzt zum ersten Mal Web-Mining und das Unternehmen hat beschlossen, für die ersten Schritte einen Data-Mining-Spezialisten hinzuzuziehen. Eine der ersten Aufgaben des Beraters besteht darin, die Ressourcen des Unternehmens für das Data-Mining zu bewerten.

Personal. Es wird deutlich, dass interne Fachkenntnis in Bezug auf die Verwaltung von Serverprotokollen und Produkt- sowie Einkaufsdatenbanken vorhanden ist. Die Erfahrung beim Data Warehousing und der Datenbereinigung für Analysen ist jedoch gering. Deshalb sollte möglicherweise auch ein Datenbankspezialist konsultiert werden. Da das Unternehmen hofft, dass die Ergebnisse der Studie Teil eines fortlaufenden Web-Mining-Prozesses werden, muss das Management außerdem überdenken, ob Positionen, die während der aktuellen Arbeiten geschaffen werden, als feste Stellen eingerichtet werden.

Daten. Da es sich bei diesem Beispiel um ein etabliertes Unternehmen handelt, steht eine Fülle von Webprotokoll- und Einkaufsdaten zur Verfügung. Für diese Anfangsstudie wird das Unternehmen jedoch die Analysen auf Kunden beschränken, die sich auf der Site registriert haben. Wenn die Studie erfolgreich ist, kann das Programm erweitert werden.

Risiken. Neben den finanziellen Ausgaben für die Berater und die Zeit, die Mitarbeiter an der Studie gearbeitet haben, birgt dieses Projekt keine großen direkten Risiken. Die Zeit ist jedoch immer ein wichtiger Aspekt, weshalb das Anfangsprojekt für ein Finanzquartal geplant ist.

Momentan ist auch kaum zusätzlicher Cash-Flow zu verzeichnen, es ist also zwingend erforderlich, dass die Studie das Budget nicht überschreitet. Falls eines dieser Ziele gefährdet ist, haben die Geschäftsführer vorgeschlagen, den Umfang des Projekts zu verkleinern.

Ressourcenbestand

Eine genaue Bestandsaufnahme Ihrer Ressourcen ist das A und O. Sie können sich eine Menge Zeit und Kopfzerbrechen sparen, wenn Sie sich die Hardware-, Datenquellen- und Personalressourcen einmal genau ansehen.

Aufgabe 1 - Untersuchen von Hardwareressourcen

- Welche Hardware benötigen Sie?

Aufgabe 2 - Ermitteln von Datenquellen und Wissensschätzen

- Welche Datenquellen stehen für das Data-Mining zur Verfügung? Notieren Sie sich die Datentypen und -formate.
- Wie sind die Daten gespeichert? Haben Sie direkten Zugriff auf Data Warehouses oder operative Datenbanken?
- Planen Sie den Kauf externer Daten, z. B. demografische Informationen?
- Gibt es Sicherheitsprobleme, die den Zugriff auf die gewünschten Daten verhindern?

Aufgabe 3 - Ermitteln von Personalressourcen

- Haben Sie Kontakt zu Wirtschafts- und Datenexperten?
- Haben Sie Datenbankadministratoren und andere Mitarbeiter des Support-Teams bestimmt, die möglicherweise benötigt werden?

Wenn Sie diese Fragen gestellt haben, erfassen Sie eine Liste der Kontakte und Ressourcen für den Phasenbericht.

Anforderungen, Annahmen und Beschränkungen

Ihre Bemühungen zahlen sich eher aus, wenn Sie eine ehrliche Bewertung der Verbindlichkeiten für das Projekt erstellen. Zukünftige Probleme können vermieden werden, wenn Sie diese Bedenken so explizit wie möglich formulieren.

Aufgabe 1 - Bestimmen von Anforderungen

Die wesentlichste Anforderung liegt im bereits besprochenen Geschäftsziel. Bedenken Sie jedoch auch folgende Punkte:

- Unterliegen die Daten oder Projektergebnisse Sicherheitsbeschränkungen oder rechtlichen Beschränkungen?
- Kennt jeder die Projektplanungsanforderungen?
- Bestehen Anforderungen an die Bereitstellung der Ergebnisse (z. B. Veröffentlichung im Internet oder Einlesen von Bewertungen in eine Datenbank)?

Aufgabe 2 - Klären von Annahmen

- Gibt es wirtschaftliche Faktoren, die das Projekt beeinflussen könnten (z. B. Beratungsgebühren oder Konkurrenzprodukte)?
- Gibt es Annahmen zur Datenqualität?
- Wie möchte der Projektsponsor/das Managementteam die Ergebnisse anzeigen? Anders ausgedrückt, möchten sie das Modell an sich verstehen oder einfach nur die Ergebnisse anzeigen?

Aufgabe 3 - Überprüfen von Beschränkungen

- Besitzen Sie alle für den Datenzugriff nötigen Kennwörter?
- Haben Sie alle rechtlichen Beschränkungen hinsichtlich der Datennutzung überprüft?
- Sind alle finanziellen Einschränkungen im Projektbudget abgedeckt?

Risiken und Notfälle

Es ist ratsam, mögliche im Verlauf des Projekts auftretende Risiken zu bedenken. Risikotypen:

- Planungsrisiko (Was geschieht, wenn das Projekt länger als erwartet dauert?)
- Finanzrisiko (Was geschieht, wenn der Projektsponsor auf Budgetprobleme stößt?)
- Datenrisiko (Was geschieht, wenn die Daten von schlechter Qualität sind oder in zu geringen Mengen vorliegen?)
- Ergebnisrisiko (Was geschieht, wenn die Anfangsergebnisse weniger tief greifend als erwartet sind?)

Nachdem Sie die verschiedenen Risiken überdacht haben, erarbeiten Sie einen Notfallplan, um Katastrophen zu verhindern.

Aufgabenliste

- Dokumentieren Sie jedes mögliche Risiko.
- Dokumentieren Sie einen Notfallplan für jedes Risiko.

Terminologie

Um sicherzustellen, dass die Unternehmens- und Data-Mining-Teams "dieselbe Sprache sprechen", sollten Sie ein Glossar mit technischen Termini und Schlagwörtern erstellen, die geklärt werden müssen. Wenn beispielsweise das Wort "Abwanderung" in Ihrem Unternehmen eine bestimmte und einmalige Bedeutung hat, wäre eine explizite Erklärung dieser Bedeutung hilfreich und zum Vorteil des gesamten Teams. Gleichermaßen könnte das Team von der Erläuterung der Verwendung eines Gewinndiagramms profitieren.

Aufgabenliste

- Führen Sie eine Liste mit Termini oder Jargon-Ausdrücken, die für Teammitglieder verwirrend sind. Nehmen Sie sowohl Unternehmens- als auch Data-Mining-Terminologie auf.
- Überlegen Sie, diese Liste im Intranet oder in anderen Projektdokumentationen zu veröffentlichen.

Kosten-/Nutzen-Analyse

In diesem Schritt erhalten Sie eine Antwort auf die Frage **Was ist mein Endergebnis?** Als Teil der Endbeurteilung ist es wichtig, die Kosten des Projekts mit dem potenziellen Erfolgsnutzen zu vergleichen.

Aufgabenliste

Die Analyse sollte geschätzte Kosten für folgende Bereiche enthalten:

- Datenerfassung und sämtliche verwendete externe Daten
- Ergebnisbereitstellung
- Betriebskosten

Berücksichtigen Sie dann den Nutzen durch:

- Erreichen des primären Ziels
- Zusätzliche Einblicke durch Datenexploration
- Mögliche Vorteile durch bessere Datenuntersuchung

Bestimmen von Data-Mining-Zielen

Nun, da das Geschäftsziel klar definiert ist, wird es Zeit, es in eine Data-Mining-Realität umzuwandeln. Beispielsweise kann das Geschäftsziel "Abwanderung reduzieren" in ein Data-Mining-Ziel umgewandelt werden, das Folgendes umfasst.

- Identifizieren von hochwertigen Kunden anhand neuerer Kaufdaten

- Erstellen eines Modells mithilfe verfügbarer Kundendaten, um die Abwanderungswahrscheinlichkeit für jeden Kunden vorherzusagen
- Zuweisen eines Rangs anhand der Abwanderungsneigung und des Kundenwerts

Diese Data-Mining-Ziele können - sofern sie erreicht werden - vom Unternehmen genutzt werden, um die Abwanderung bei den wertvollsten Kunden zu reduzieren.

Sie sehen also, dass Unternehmen und Technologie für ein effektives Data-Mining Hand in Hand arbeiten müssen. Lesen Sie weiter, um spezifische Tipps zum Bestimmen von Data-Mining-Zielen zu erhalten.

Data-Mining-Ziele

Achten Sie bei der Zusammenarbeit mit Unternehmens- und Datenanalysten zum Definieren einer technischen Lösung für ein Geschäftsproblem immer darauf, konkret zu bleiben.

Aufgabenliste

- Beschreiben Sie den **Typ** des Data-Mining-Problems, wie z. B. Clustering, Vorhersage oder Klassifizierung.
- Dokumentieren Sie technische Ziele mithilfe spezifischer Zeiteinheiten, z. B. Vorhersagen mit einer dreimonatigen Gültigkeit.
- Geben Sie, wenn möglich, tatsächliche Zahlen für gewünschte Ergebnisse an, z. B. Abwanderungswerte für 80 % der aktuellen Kunden.

Beispiel aus dem elektronischen Einzelhandel - Data-Mining-Ziele

Ein Web-Mining-Szenario mit CRISP-DM

Mit der Unterstützung seines Data-Mining-Beraters konnte der e-Einzelhändler die Geschäftsziele des Unternehmens in Data-Mining-Begriffe umwandeln. Die Ziele für die Anfangsstudie, die dieses Quartal abgeschlossen werden soll, lauten:

- Verwenden Sie historische Informationen zu früheren Käufen, um ein Modell zu erstellen, das "verwandte" Elemente verknüpft. Wenn Benutzer eine Elementbeschreibung aufrufen, bieten Sie Links zu anderen Elementen in der verwandten Gruppe (**Warenkorbanalyse**).
- Bestimmen Sie mithilfe von Webprotokollen, was die verschiedenen Kunden suchen, und gestalten Sie die Site dann entsprechend neu, um diese Elemente hervorzuheben. Jedem "Kundentyp" wird eine andere Hauptseite der Site angezeigt (**Profilerstellung**).
- Versuchen Sie, mithilfe von Webprotokollen, also unter Berücksichtigung der Seite Ihrer Site, von der die Person kommt oder die sie besucht hat, zu bestimmen, welche Seite eine Person als nächstes aufrufen wird (**Sequenzanalyse**).

Erfolgskriterien für das Data-Mining

Erfolg muss auch in technischer Hinsicht definiert werden, um Ihre Data-Mining-Arbeiten auf dem richtigen Weg zu halten. Nutzen Sie das im oberen Abschnitt bestimmte Data-Mining-Ziel zum Formulieren von Benchmarks für den Erfolg. IBM SPSS Modeler bietet Tools wie den Evaluierungsknoten und den Analyseknoten, um Sie bei der Analyse der Genauigkeit und Gültigkeit Ihrer Ergebnisse zu unterstützen.

Aufgabenliste

- Beschreiben Sie die Methoden für die Modellbewertung (z. B. Genauigkeit, Leistung usw.).
- Definieren Sie Benchmarks zum Evaluieren des Erfolgs. Geben Sie spezifische Zahlen an.
- Definieren Sie subjektive Messungen so gut Sie können und bestimmen Sie die Person, die über Erfolg oder Misserfolg entscheidet.
- Überlegen Sie, ob die erfolgreiche Bereitstellung von Modellergebnissen Teil des Data-Mining-Erfolgs ist. Beginnen Sie bereits jetzt, die Bereitstellung zu planen.

Erstellen eines Projektplans

Nun sind Sie bereit, einen Plan für das Data-Mining-Projekt zu erstellen. Die Fragen, die Sie bisher gestellt, und die Geschäfts- und Data-Mining-Ziele, die Sie formuliert haben, bilden die Grundlage dieses Plans.

Schreiben des Projektplans

Der Projektplan ist das Masterdokument für Ihre gesamten Data-Mining-Arbeiten. Wenn er richtig erstellt wurde, kann der Projektplan alle mit diesem Projekt betrauten Personen über die Ziele, Ressourcen, Risiken und die Planung für alle Phasen des Data-Mining informieren. Möglicherweise möchten Sie den Plan sowie die während dieser Phase gesammelten Unterlagen im Intranet Ihres Unternehmens veröffentlichen.

Aufgabenliste

Stellen Sie beim Erstellen des Plans sicher, dass Sie die folgenden Fragen beantwortet haben:

- Haben Sie die Projektaufgaben und den vorgeschlagenen Plan mit allen beteiligten Personen diskutiert?
- Enthält der Plan Zeitschätzungen für alle Phasen oder Aufgaben?
- Haben Sie den Arbeitsaufwand und die Ressourcen aufgenommen, die für die Bereitstellung der Ergebnisse oder Geschäftslösungen erforderlich sind?
- Sind Entscheidungspunkte und Prüfanfragen im Plan hervorgehoben?
- Haben Sie Phasen gekennzeichnet, in denen üblicherweise mehrere Iterationen auftreten, z. B. die Modellierungsphase?

Beispiel eines Projektplans

In der nachfolgenden Tabelle ist der Überblicksplan für die Studie dargestellt.

Tabelle 1. Beispiel einer Planübersicht für ein Projekt

Phase	Zeit	Ressourcen	Risiken
Untersuchung der Geschäftsziele	1 Woche	Alle Analysten	Wirtschaftlicher Wandel
Datenuntersuchung	3 Wochen	Alle Analysten	Datenprobleme, Technologieprobleme
Datenaufbereitung	5 Wochen	Data-Mining-Berater, stundenweise Datenbankanalyst	Datenprobleme, Technologieprobleme
Modellierung	2 Wochen	Data-Mining-Berater, stundenweise Datenbankanalyst	Technologieprobleme, Unvermögen, ein geeignetes Modell zu finden
Evaluierung	1 Woche	Alle Analysten	Wirtschaftlicher Wandel, Unvermögen, Ergebnisse umzusetzen
Bereitstellung	1 Woche	Data-Mining-Berater, stundenweise Datenbankanalyst	Wirtschaftlicher Wandel, Unvermögen, Ergebnisse umzusetzen

Bewerten von Tools und Verfahren

Da Sie sich bereits für IBM SPSS Modeler als Ihr Tool zum Data-Mining-Erfolg entschieden haben, können Sie in diesem Schritt untersuchen, welche Data-Mining-Verfahren für Ihre Geschäftsanforderungen am besten geeignet sind. IBM SPSS Modeler bietet eine vollständige Palette an Tools für jede Phase des Data-Mining. Um zu entscheiden, wann Sie die verschiedenen Verfahren einsetzen sollten, lesen Sie den Abschnitt zur Modellierung in der Online-Hilfe.

Sind Sie bereit für den nächsten Schritt?

Bevor Sie Daten untersuchen und mit der Arbeit in IBM SPSS Modeler beginnen, sollten Sie die folgende Fragen beantwortet haben.

Aus Unternehmenssicht:

- Welchen Nutzen erhofft sich Ihr Unternehmen von diesem Projekt?
- Wie definieren Sie den erfolgreichen Abschluss Ihrer Arbeiten?
- Verfügen Sie über das erforderliche Budget und die notwendigen Ressourcen, um Ihre Ziele zu erreichen?
- Haben Sie Zugriff auf alle Daten, die für dieses Projekt benötigt werden?
- Haben Sie und Ihr Team die mit diesem Projekt verbundenen Risiken und Notfälle diskutiert?
- Ist das Projekt laut den Ergebnissen Ihrer Kosten-/Nutzen-Analyse rentabel?

Nachdem Sie die oben genannten Fragen beantwortet haben, haben Sie diese Antworten in Data-Mining-Ziele umgewandelt?

Aus Sicht des Data-Mining:

- Wie genau kann Data-Mining Sie beim Erreichen Ihrer Geschäftsziele unterstützen?
- Haben Sie eine Vorstellung, welche Data-Mining-Verfahren die besten Ergebnisse erzielen könnten?
- Wie erkennen Sie, wann Ihre Ergebnisse genau und effektiv genug sind? (*Haben wir eine Messung des Data-Mining-Erfolgs festgelegt?*)
- Wie werden die Ergebnisse der Modellierung bereitgestellt? Haben Sie die Bereitstellung in Ihrem Projektplan berücksichtigt?
- Umfasst der Projektplan alle Phasen von CRISP-DM?
- Sind Risiken und Abhängigkeiten im Plan aufgeführt?

Wenn Sie die oben genannten Fragen mit "ja" beantworten können, sind Sie bereit, sich die Daten genauer anzusehen.

Kapitel 3. Datenuntersuchung

Datenuntersuchung - Übersicht

In der Datenuntersuchungsphase von CRISP-DM werfen Sie einen genaueren Blick auf die für das Data-Mining verfügbaren Daten. Dieser Schritt ist überaus wichtig, um unerwartete Probleme während der nächsten Phase - der Datenaufbereitung - zu vermeiden, die normalerweise den längsten Teil des Projekts ausmacht.

Zur Datenuntersuchungsphase gehört der Zugriff auf Daten und deren Exploration mithilfe von Tabellen und Grafiken, die mit dem CRISP-DM-Projekttool in IBM SPSS Modeler organisiert werden können. Somit können Sie die Qualität der Daten bestimmen und die Ergebnisse dieser Schritte in der Projektdokumentation beschreiben.

Sammeln ursprünglicher Daten

An dieser Stelle in CRISP-DM sind Sie bereit, auf Daten zuzugreifen und diese in IBM SPSS Modeler zu importieren. Daten stammen aus einer Vielzahl von Quellen, z. B. aus:

- **Vorhandenen Daten.** Dazu gehört eine große Palette von Daten, wie z. B. Transaktionsdaten, Umfragedaten, Webprotokolle usw. Prüfen Sie, ob die vorhandenen Daten ausreichen, um Ihre Anforderungen zu erfüllen.
- **Erworbenen Daten.** Verwendet Ihr Unternehmen zusätzliche Daten, z. B. demografische Daten? Falls nicht, dann überlegen Sie, ob solche Daten eventuell benötigt werden.
- **Zusätzlichen Daten.** Wenn die oben genannten Quellen Ihren Anforderungen nicht genügen, sollten Sie möglicherweise Umfragen durchführen oder zusätzliche Nachverfolgungen starten, um die vorhandenen Datenspeicher zu ergänzen.

Aufgabenliste

Sehen Sie sich die Daten in IBM SPSS Modeler an und bedenken Sie die folgenden Fragen. Vergessen Sie nicht, sich Notizen zu Ihren Ergebnissen zu machen. Weitere Informationen finden Sie im Thema „Schreiben eines Berichts zur Datensammlung” auf Seite 14.

- Welche Attribute (Spalten) der Datenbank scheinen den meisten Erfolg zu versprechen?
- Welche Attribute scheinen irrelevant und können daher ausgeschlossen werden?
- Sind ausreichend Daten vorhanden, um verallgemeinerbare Schlussfolgerungen zu ziehen oder genaue Vorhersagen zu erstellen?
- Sind zu viele Attribute für Ihre ausgewählte Modellierungsmethode vorhanden?
- Verbinden Sie mehrere Datenquellen? Wenn ja, gibt es Bereiche, die beim Verbinden ein Problem darstellen könnten?
- Haben Sie bedacht, wie fehlende Werte in jeder Ihrer Datenquellen behandelt werden?

Beispiel aus dem elektronischen Einzelhandel - Sammeln ursprünglicher Daten

Ein Web-Mining-Szenario mit CRISP-DM

Der e-Einzelhändler in diesem Beispiel nutzt verschiedene wichtige Datenquellen, z. B.:

Webprotokolle. Die unbearbeiteten Zugriffsprotokolle enthalten alle Informationen darüber, wie Kunden auf der Website navigieren. Verweise auf Bilddateien oder andere nicht informative Einträge in den Webprotokollen müssen als Teil des Datenaufbereitungsprozesses entfernt werden.

Kaufdaten. Wenn ein Kunde eine Bestellung sendet, werden alle zu dieser Bestellung gehörenden Informationen gespeichert. Die Bestellungen in der Einkaufsdatenbank müssen den entsprechenden Sitzungen in den Webprotokollen zugeordnet werden.

Produktdatenbank. Die Produktattribute sind bei der Bestimmung "verwandter" Produkte nützlich. Die Produktinformationen müssen den entsprechenden Bestellungen zugeordnet werden.

Kundendatenbank. Diese Datenbank enthält zusätzliche Informationen, die von registrierten Kunden erfasst werden. Die Datensätze sind keineswegs vollständig, da viele Kunden die Fragebogen nicht ausfüllen. Die Kundeninformationen müssen den entsprechenden Käufen und Sitzungen in den Webprotokollen zugeordnet werden.

Momentan plant das Unternehmen nicht, externe Datenbanken zu erwerben oder Geld für Umfragen auszugeben, da die Analysten damit beschäftigt sind, die aktuell vorhandenen Daten zu verwalten. Zu einem späteren Zeitpunkt könnte das Unternehmen jedoch eine erweiterte Bereitstellung von Data-Mining-Ergebnissen in Erwägung ziehen, wofür der Erwerb zusätzlicher demografischer Daten zu nicht registrierten Kunden sehr nützlich wäre. Demografische Informationen wären auch nützlich, um aufzuzeigen, wie sich der Kundenstamm des e-Einzelhändlers vom durchschnittlichen Käufer im Internet unterscheidet.

Schreiben eines Berichts zur Datensammlung

Auf Grundlage der im vorherigen Schritt gesammelten Materialien können Sie nun beginnen, einen Bericht zur Datensammlung zu schreiben. Sobald der Bericht fertig gestellt ist, kann er der Projektwebsite hinzugefügt oder an das Team verteilt werden. Sie können ihn auch mit den Berichten kombinieren, die in den nächsten Schritten - Datenbeschreibung, Datenexploration und Qualitätsprüfung - vorbereitet werden. Diese Berichte führen Sie durch die Phase der Datenaufbereitung.

Beschreiben von Daten

Es gibt viele verschiedene Wege, Daten zu beschreiben. Die meisten Beschreibungen konzentrieren sich auf die Quantität und die Qualität der Daten, also auf die Menge an verfügbaren Daten und die Beschaffenheit der Daten. Nachfolgend sind einige wichtige Merkmale aufgeführt, auf die beim Beschreiben von Daten eingegangen werden sollte.

- **Menge an Daten.** Für die meisten Modellierungsverfahren gibt es im Zusammenhang mit der Datengröße stehende Kosten-Nutzen-Abwägungen (Trade-Offs). Mithilfe großer Datasets können genauere Modelle erstellt werden, sie können aber auch die Verarbeitungszeit verlängern. Sie sollten sich überlegen, ob auch ein Subset der Daten verwendet werden kann. Stellen Sie sicher, dass Ihre Aufzeichnungen für den abschließenden Bericht Größenstatistiken für alle Datasets enthalten und vergessen Sie nicht, beim Beschreiben der Daten sowohl die Anzahl der Datensätze als auch der Felder (Attribute) zu berücksichtigen.
- **Wertetypen.** Daten können in einer Vielzahl von Formaten vorliegen, z. B. **numerisch**, **kategorial** (Zeichenfolge) oder **boolesch** (wahr/falsch). Wenn Sie den Wertetyp beachten, können Sie Probleme während der späteren Modellierung vermeiden.
- **Codierungsschemata.** Häufig sind Werte in der Datenbank Darstellungen von Merkmalen wie Geschlecht oder Produkttyp. So verwendet z. B. ein Dataset die Werte *M* und *W*, um *männlich* und *weiblich* darzustellen, während andere Datasets dafür die numerischen Werte *1* und *2* verwenden. Dokumentieren Sie alle widersprüchlichen Schemata im Datenbericht.

Mit diesem erlangten Wissen sind Sie nun in der Lage, den Bericht zur Datenbeschreibung zu verfassen und Ihre Ergebnisse mit einem größeren Publikum zu teilen.

Beispiel aus dem elektronischen Einzelhandel - Beschreiben von Daten

Ein Web-Mining-Szenario mit CRISP-DM

In einer Web-Mining-Anwendung müssen viele Datensätze und Attribute verarbeitet werden. Denn obwohl der e-Einzelhändler, der dieses Data-Mining-Projekt durchführt, die Anfangsstudie auf die ca. 30.000 Kunden beschränkt hat, die sich auf der Site registriert haben, enthalten die Webprotokolle dennoch Millionen von Datensätzen.

Bei den meisten Wertetypen in diesen Datenquellen handelt es sich um symbolische Werte. Dabei ist es egal, ob es sich um Datums- und Zeitangaben, aufgerufene Webseiten oder Antworten auf Multiple-Choice-Fragen des Registrierungsfragebogens handelt. Einige dieser Variablen werden dazu verwendet, neue numerische Variablen zu erstellen, wie z. B. zur Anzahl der besuchten Webseiten und zur auf einer Webseite verbrachten Zeit. Zu den wenigen bereits vorhandenen numerischen Variablen in den Datenquellen gehören die aus der Produktdatenbank erhältlichen Angaben zur Menge jedes bestellten Produkts, zum beim Kauf ausgegebenen Betrag und die Spezifikationen zu Produktgewicht und -abmessungen.

In den Codierungsschemata für die verschiedenen Datenquellen gibt es wenige Überschneidungen, da die Datenquellen sehr unterschiedliche Attribute enthalten. Die einzigen Variablen, bei denen Überschneidungen auftreten, sind "Schlüssel" wie die Kunden-IDs und Produktcodes. Diese Variablen müssen in den Datenquellen identische Codierungsschemata besitzen, da es andernfalls nicht möglich wäre, die Datenquellen zu verbinden. Um diese Schlüsselfelder für das Verbinden neu zu codieren, ist eine zusätzliche Datenaufbereitung notwendig.

Schreiben eines Berichts zur Datenbeschreibung

Um Ihr Data-Mining-Projekt effektiv fortsetzen zu können, sollten Sie den Wert eines genauen Berichts zur Datenbeschreibung mithilfe der folgenden Metriken betrachten:

Datenquantität

- In welchem Format liegen die Daten vor?
- Ermitteln Sie die Methode, mit der die Daten erfasst wurden, z. B. ODBC.
- Wie groß ist die Datenbank (Anzahl der Zeilen und Spalten)?

Datenqualität

- Umfassen die Daten Merkmale, die für die Geschäftsfrage relevant sind?
- Welche Datentypen sind vorhanden (symbolisch, numerisch usw.)?
- Haben Sie Basisstatistiken für die Schlüsselattribute berechnet? Welchen Einblick in die Geschäftsfrage haben Sie dadurch erhalten?
- Können Sie relevanten Attributen Prioritäten zuweisen? Wenn nicht, sind Unternehmensanalysten verfügbar, die mehr Einblick gewähren können?

Untersuchen von Daten

Nutzen Sie diese Phase von CRISP-DM, um die Daten mit den in IBM SPSS Modeler verfügbaren Tabellen, Diagrammen und anderen Tools zur visuellen Darstellung zu untersuchen. Diese Analysen können dabei helfen, das Data-Mining-Ziel in Angriff zu nehmen, das Sie während der Phase Untersuchung der Geschäftsziele erarbeitet haben. Sie unterstützen Sie auch bei der Formulierung von Hypothesen und bei der Gestaltung der Datentransformationsaufgaben, die während der Datenaufbereitung stattfinden.

Beispiel aus dem elektronischen Einzelhandel - Untersuchen von Daten

Ein Web-Mining-Szenario mit CRISP-DM

Obwohl CRISP-DM an dieser Stelle eine erste Exploration vorschlägt, hat unser e-Einzelhändler herausgefunden, dass die Datenexploration bei Roh-Webprotokollen schwierig, wenn nicht gar unmöglich ist. Normalerweise müssen Webprotokolldaten zuerst in der Datenaufbereitungsphase verarbeitet werden,

um Daten hervorzubringen, die dann aussagekräftig untersucht werden können. Diese Abweichung von CRISP-DM macht noch einmal deutlich, dass der Prozess an Ihre bestimmen Data-Mining-Anforderungen angepasst werden kann und sollte. CRISP-DM arbeitet zyklisch und Data Miner wechseln in der Regel zwischen den Phasen vor und zurück.

Obgleich Webprotokolle vor der Exploration verarbeitet werden müssen, sind die anderen dem e-Einzelhändler zur Verfügung stehenden Datenquellen leichter für die Explorationen zugänglich. So bietet beispielsweise die Einkaufsdatenbank interessante Zusammenfassungen über Kunden, wie z. B. Informationen zu den ausgegebenen Beträgen, zu den pro Einkauf erworbenen Artikeln und zur Herkunft des Kunden. Zusammenfassungen der Kundendatenbank zeigen die Verteilung der Antworten auf die Punkte im Registrierungsfragebogen.

Die Exploration ist auch bei der Suche nach Fehlern in den Daten nützlich. Während die meisten Datenquellen automatisch generiert werden, werden die Informationen in der Produktdatenbank per Hand eingegeben. Durch kurze Zusammenfassungen der aufgeführten Produktabmessungen können Tippfehler wie "119-Zoll-Monitor" (anstatt "19-Zoll-Monitor") erkannt werden.

Schreiben eines Berichts zur Datenexploration

Beginnen Sie während der Erstellung von Diagrammen und der Anwendung von Statistiken auf die verfügbaren Daten damit, Hypothesen darüber zu bilden, wie die Daten die technischen und Geschäftsziele erfüllen können.

Aufgabenliste

Notieren Sie sich Ihre Ergebnisse, um sie später in den Bericht zur Datenexploration aufzunehmen. Stellen Sie sicher, dass Sie die folgenden Fragen beantworten:

- Welche Hypothesen haben Sie über die Daten gebildet?
- Welche Attribute scheinen für spätere Analysen erfolgversprechend?
- Haben Ihre Explorationen neue Merkmale über die Daten aufgezeigt?
- Inwiefern haben diese Explorationen Ihre ursprüngliche Hypothese geändert?
- Können Sie bestimmte Subsets an Daten für eine spätere Verwendung bestimmen?
- Sehen Sie sich Ihre Data-Mining-Ziele noch einmal an. Hat diese Exploration die Ziele geändert?

Überprüfen der Datenqualität

In den seltensten Fällen sind Daten perfekt. Tatsächlich weisen die meisten Daten Codierungsfehler, fehlende Werte oder andere Arten von Inkonsistenzen auf, die eine Analyse teilweise kompliziert machen. Um potenzielle Probleme zu vermeiden, können Sie vor der Modellierung eine gründliche Qualitätsanalyse durchführen.

Mithilfe der Berichterstellungstools in IBM SPSS Modeler (wie Data Audit-, Tabelllen- und andere Ausgabeknoten) können Sie nach den folgenden Typen von Problemen suchen:

- **Fehlende Daten** umfassen Werte, die leer sind oder als Nichtantwort (wie z. B. *$null$*, *?* oder *999*) codiert sind.
- **Datenfehler** sind in der Regel typografische Fehler, die bei der Dateneingabe verursacht wurden.
- Zu **Messfehlern** gehören Daten, die zwar korrekt eingegeben wurden, aber auf einem falschen Messschema basieren.
- **Codierungsinkonsistenzen** umfassen in der Regel nicht standardmäßige Maßeinheiten oder Wertinkonsistenzen wie die gleichzeitige Verwendung von *M* und *männlich* für das Geschlecht.
- Zu **ungültigen Metadaten** gehören mangelnde Übereinstimmungen zwischen der offensichtlichen Bedeutung eines Felds und der in einem Feldnamen oder einer Felddefinition angegebenen Bedeutung.

Vergessen Sie nicht, sich solche Bedenken hinsichtlich der Qualität zu notieren. Weitere Informationen finden Sie im Thema „Schreiben eines Berichts zur Datenqualität".

Beispiel aus dem elektronischen Einzelhandel - Überprüfen der Datenqualität

Ein Web-Mining-Szenario mit CRISP-DM

Die Überprüfung der Datenqualität findet oftmals während der Beschreibung und Exploration der Daten statt. Zu den Problemen, auf die der e-Einzelhändler gestoßen ist, gehören z. B.:

Fehlende Daten. Zu den bekannten fehlenden Daten zählen die nicht beantworteten Fragebogen einiger registrierter Benutzer. Ohne die durch den Fragebogen gelieferten zusätzlichen Informationen müssen diese Kunden möglicherweise von einigen der nachfolgenden Modelle ausgeschlossen werden.

Datenfehler. Die meisten Datenquellen werden automatisch generiert, sie stellen also kein großes Problem dar. Typografische Fehler in der Produktdatenbank können während des Explorationsprozesses gefunden werden.

Messfehler. Die größte potenzielle Quelle für Messfehler stellt der Fragebogen dar. Wenn Fragen unklug oder schlecht formuliert sind, führen sie möglicherweise nicht zu den Informationen, die sich der e-Einzelhändler erwünscht. Während des Explorationsprozesses ist es deshalb wiederum wichtig, Fragen mit einer ungewöhnlichen Antwortvielfalt besonders zu beachten.

Schreiben eines Berichts zur Datenqualität

Aufbauend auf der Exploration und Überprüfung der Datenqualität sind Sie jetzt in der Lage, einen Bericht vorzubereiten, der Sie durch die nächste Phase von CRISP-DM führen wird. Weitere Informationen finden Sie im Thema „Überprüfen der Datenqualität" auf Seite 16.

Aufgabenliste

Wie bereits besprochen, gibt es verschiedene Typen von Datenqualitätsproblemen. Bevor Sie mit dem nächsten Schritt fortfahren, überdenken Sie die folgenden Qualitätsprobleme und planen Sie eine Lösung. Dokumentieren Sie alle Antworten im Bericht zur Datenqualität.

- Haben Sie fehlende Attribute und leere Felder erkannt? Wenn ja, gibt es einen Grund für diese fehlenden Werte?
- Gibt es Inkonsistenzen in der Schreibweise, die bei späterem Verbinden oder Umwandeln Probleme verursachen könnten?
- Haben Sie Abweichungen untersucht, um zu bestimmen, ob sie "Rauschen" oder andere Phänomene darstellen, die weiter analysiert werden sollen?
- Haben Sie eine Plausibilitätsprüfung für Werte durchgeführt? Notieren Sie sich offensichtliche Konflikte (z. B. Jugendliche mit hohem Einkommensniveau).
- Haben Sie überlegt, Daten auszuschließen, die keinen Einfluss auf Ihre Hypothesen haben?
- Sind die Daten in Flatfiles gespeichert? Wenn ja, wurden konsistente Trennzeichen in der Datei verwendet? Enthält jeder Datensatz dieselbe Anzahl an Feldern?

Sind Sie bereit für den nächsten Schritt?

Bevor Sie die Daten für die Modellierung in IBM SPSS Modeler aufbereiten, bedenken Sie folgende Punkte:

Wie gut ist Ihr Datenverständnis?

- Sind alle Datenquellen klar identifiziert und zugänglich? Sind Sie sich eventueller Probleme und Beschränkungen bewusst?

- Haben Sie Schlüsselattribute aus den verfügbaren Daten ermittelt?
- Haben Ihnen diese Attribute beim Formulieren der Hypothesen geholfen?
- Haben Sie sich die Größe aller Datenquellen notiert?
- Können Sie gegebenenfalls ein Subset der Daten nutzen?
- Haben Sie Basisstatistiken für jedes gewünschte Attribut berechnet? Haben Sie aussagekräftige Informationen erhalten?
- Haben Sie explorative Diagramme verwendet, um weitere Einblicke in Schlüsselattribute zu erhalten? Haben diese Einblicke zur Umformulierung Ihrer Hypothesen geführt?
- Welche Datenqualitätsprobleme liegen für dieses Projekt vor? Haben Sie einen Plan, wie Sie diese Probleme angehen können?
- Sind die Schritte zur Datenaufbereitung klar und eindeutig? Wissen Sie beispielsweise, welche Datenquellen Sie verbinden und welche Attribute Sie filtern oder auswählen werden?

Da Sie zwischenzeitlich Geschäfts- und Datenverständnis erlangt haben, wird jetzt es Zeit, Ihre Daten mit IBM SPSS Modeler für die Modellierung vorzubereiten.

Kapitel 4. Datenaufbereitung

Datenaufbereitung - Übersicht

Die Datenaufbereitung ist einer der wichtigsten und sehr oft zeitaufwendigsten Aspekte des Data-Mining. Tatsächlich wird sogar geschätzt, dass die Datenaufbereitung in der Regel 50-70 % der Zeit und Arbeiten eines Projekts in Anspruch nimmt. Wenn Sie angemessene Zeit und Energie in die vorherigen Phasen Untersuchung der Geschäftsziele und Datenuntersuchung investieren, kann dieser Aufwand minimiert werden. Sie müssen jedoch nach wie vor einen großen Teil der Arbeitskraft einsetzen, um die Daten für das Data-Mining vorzubereiten und zu packen.

Abhängig von Ihrem Unternehmen und Ihren Zielen umfasst die Datenaufbereitung in der Regel die folgenden Aufgaben:

- Verbinden von Datasets und/oder Datensätzen
- Auswählen eines beispielhaften Subsets an Daten
- Aggregieren von Datensätzen
- Ableiten neuer Attribute
- Sortieren der Daten für die Modellierung
- Entfernen oder Ersetzen leerer oder fehlender Werte
- Aufteilen in Training- und Testdatasets

Auswählen von Daten

Auf Grundlage der in der vorherigen CRISP-DM-Phase durchgeführten ersten Datensammlung können Sie nun die für Ihre Data-Mining-Ziele relevanten Daten auswählen. Im Allgemeinen gibt es zwei Methoden zur Datenauswahl:

- Die **Auswahl von Elementen (Zeilen)** umfasst Entscheidungen, wie z. B. welche Konten, Produkte oder Kunden einbezogen werden sollen.
- Bei der **Auswahl von Attributen oder Merkmalen (Spalten)** müssen Entscheidungen über den Nutzen der Merkmale wie Transaktionsmenge oder Haushaltseinkommen getroffen werden.

Beispiel aus dem elektronischen Einzelhandel - Auswählen von Daten

Ein Web-Mining-Szenario mit CRISP-DM

Viele Entscheidungen des e-Einzelhändlers, welche Daten ausgewählt werden sollen, wurden bereits in früheren Phasen des Data-Mining-Prozesses getroffen.

Auswählen von Elementen. Die Anfangsstudie wird auf die (ca.) 30.000 Kunden beschränkt, die sich auf der Site registriert haben. Deshalb müssen Filter eingerichtet werden, mit denen Einkäufe und Webprotokolle von nicht registrierten Benutzern ausgeschlossen werden können. Außerdem sollten weitere Filter erstellt werden, um Aufrufe von Bilddateien oder anderen nicht informativen Einträgen in den Webprotokollen zu entfernen.

Auswählen von Attributen. Die Einkaufsdatenbank enthält vertrauliche Informationen zu den Kunden des e-Einzelhändlers. Es ist deshalb wichtig, Attribute wie Kundenname, Adresse, Telefonnummer und Kreditkartennummern herauszufiltern.

Einbeziehen oder Ausschließen von Daten

Wenn Sie eine Entscheidung hinsichtlich der einzubeziehenden oder auszuschließenden Subsets an Daten treffen, vergessen Sie nicht, die Gründe für Ihre Entscheidung zu dokumentieren.

Fragen, die Sie berücksichtigen sollten:

- Ist ein bestimmtes Attribut für Ihre Data-Mining-Ziele relevant?
- Schließt die Qualität eines bestimmten Datasets oder Attributs die Gültigkeit Ihrer Ergebnisse aus?
- Können Sie solche Daten verwerten?
- Gibt es Beschränkungen jeglicher Art bei der Verwendung bestimmter Felder wie *Geschlecht* oder *Rasse*?

Unterscheiden sich Ihre hier getroffenen Entscheidungen von den in der Datenuntersuchungsphase formulierten Hypothesen? Wenn ja, dokumentieren Sie unbedingt Ihre Argumentation im Projektbericht.

Bereinigen von Daten

Beim Bereinigen Ihrer Daten sehen Sie sich die Probleme in den Daten, die Sie für Ihre Analyse verwenden möchten, genauer an. Es gibt mehrere Möglichkeiten, die Daten mit den Datensatz- und Feldoperationsknoten in IBM SPSS Modeler zu bereinigen.

Tabelle 2. Bereinigen von Daten

Datenproblem	**Mögliche Lösung**
Fehlende Daten	Schließen Sie Zeilen oder Merkmale aus. Oder füllen Sie Leerstellen mit einem geschätzten Wert auf.
Datenfehler	Nutzen Sie die Logik, um die Fehler manuell zu erkennen und die Daten zu ersetzen. Oder schließen Sie Merkmale aus.
Codierungsinkonsistenzen	Entscheiden Sie sich für ein Codierungsschema, wandeln Sie die Werte um und ersetzen sie.
Fehlende oder ungültige Metadaten	Untersuchen Sie manuell verdächtige Felder und machen Sie die richtige Bedeutung ausfindig.

Der Bericht zur Datenqualität, den Sie während der Datenuntersuchungsphase vorbereitet haben, enthält nähere Informationen zur den Problemtypen, die speziell auf Ihre Daten zutreffen. Sie können ihn als Ausgangspunkt für die Datenbearbeitung in IBM SPSS Modeler verwenden.

Beispiel aus dem elektronischen Einzelhandel - Bereinigen von Daten

Ein Web-Mining-Szenario mit CRISP-DM

Der e-Einzelhändler nutzt den Datenbereinigungsprozess, um die im Bericht zur Datenqualität vermerkten Probleme zu behandeln.

Fehlende Daten. Kunden, die den Onlinefragebogen nicht vollständig ausgefüllt haben, müssen später eventuell von einigen Modellen ausgeschlossen werden. Diese Kunden könnten erneut gebeten werden, den Fragebogen auszufüllen. Das kostet jedoch Zeit und Geld, die der e-Einzelhändler nicht aufbringen bzw. das er nicht ausgeben kann. Er kann jedoch die Kaufunterschiede zwischen den Kunden, die die Fragebogen ausfüllen und denen, die sie nicht ausfüllen, in einem Modell darstellen. Wenn diese beiden Kundengruppen ähnliche Kaufgewohnheiten aufweisen, sind die fehlenden Fragebogen weniger Besorgnis erregend.

Datenfehler. Hier können die während des Explorationsprozesses gefundenen Fehler korrigiert werden. Zum größten Teil wird jedoch die richtige Dateneingabe auf der Website erzwungen, bevor ein Kunde eine Seite an die Backend-Datenbank sendet.

Messfehler. Schlecht formulierte Fragen in einem Fragebogen können die Qualität der Daten sehr stark beeinflussen. Wie schon bei fehlenden Fragebogen ist dieses Problem schwierig, da möglicherweise weder Zeit noch Geld zur Verfügung stehen, Antworten auf eine Ersatzfrage zu erfassen. Bei problematischen Elementen ist es möglicherweise die beste Lösung, zum Auswahlprozess zurückzukehren und diese Elemente aus den weiteren Analysen herauszufiltern.

Schreiben eines Berichts zur Datenbereinigung

Es ist wichtig, Ihre Arbeiten zur Datenbereinigung in einem Bericht zu dokumentieren, um Änderungen an den Daten nachverfolgen zu können. Zukünftige Data-Mining-Projekte werden davon profitieren, wenn ausführliche Informationen zu Ihrer Arbeit schnell verfügbar sind.

Aufgabenliste

Es ist sinnvoll, beim Schreiben des Berichts die folgenden Fragen zu berücksichtigen:

- Welche Typen von Rauschen traten in den Daten auf?
- Welche Ansätze haben Sie verfolgt, um das Rauschen zu entfernen? Welche Verfahren waren erfolgreich?
- Gab es Fälle oder Attribute, die nicht verwertet werden konnten? Vergessen Sie nicht, Daten zu vermerken, die aufgrund von Rauschen ausgeschlossen wurden.

Erstellen neuer Daten

Es kommt häufig vor, dass Sie neue Daten erstellen müssen. Es könnte zum Beispiel hilfreich sein, eine neue Spalte zu erstellen, in der der Erwerb einer verlängerten Garantie für jede Transaktion gekennzeichnet wird. Dieses neue Feld, *Garantie_erworben*, kann einfach mit dem Dichotomknoten in IBM SPSS Modeler erstellt werden.

Es gibt zwei Methoden, neue Daten zu erstellen:

- Ableiten von Attributen (Spalten oder Merkmale)
- Generieren von Datensätzen (Zeilen)

IBM SPSS Modeler bietet eine Vielzahl an Methoden, Daten mit seinen Datensatz- und Feldoperationsknoten zu erstellen.

Beispiel aus dem elektronischen Einzelhandel - Erstellen von Daten

Ein Web-Mining-Szenario mit CRISP-DM

Beim Verarbeiten von Webprotokollen können viele neue Attribute erstellt werden. Für die in den Protokollen aufgezeichneten Ereignisse möchte der e-Einzelhändler möglicherweise Zeitmarken erstellen, die Besucher und Sitzungen identifizieren sowie die zugegriffene Seite und die Art der Aktivität notieren, die das Ereignis darstellt. Einige dieser Variablen, wie z. B. die Zeit zwischen Ereignissen einer Sitzung, werden zum Erstellen weiterer Attribute verwendet.

Zusätzliche Attribute können als Ergebnis einer Verbindung oder Neustrukturierung von Daten erzeugt werden. Wenn beispielsweise die Webprotokolle mit einem Ereignis pro Zeile zusammengefasst werden, sodass jede Zeile eine Sitzung darstellt, werden neue Attribute erstellt, die die Gesamtanzahl an Aktionen, die insgesamt aufgewendete Zeit und die Gesamtanzahl an Einkäufen während der Sitzung erfassen. Werden die Webprotokolle mit der Kundendatenbank verbunden, sodass jede Zeile einen Kunden darstellt, werden gleichzeitig neue Attribute erstellt, die die Anzahl der Sitzungen, die Gesamtanzahl an Aktionen, die insgesamt aufgewendete Zeit und die Gesamtanzahl an Einkäufen jedes Kunden festhalten.

Nach dem Erstellen neuer Daten führt der e-Einzelhändler einen Explorationsprozess durch, um sicherzustellen, dass die Datenerstellung richtig ausgeführt wurde.

Ableiten von Attributen

In IBM SPSS Modeler können Sie die folgenden Feldoperationsknoten zum Herleiten neuer Attribute verwenden:

- Erstellen Sie mit einem **Ableitungsknoten** neue Felder, die von vorhandenen Feldern abgeleitet sind.
- Erstellen Sie mithilfe eines **Dichotomknotens** ein Flagfeld.

Aufgabenliste

- Berücksichtigen Sie beim Ableiten von Attributen die Datenanforderungen der Modellierung. Erwartet der Modellierungsalgorithmus einen bestimmten Typ von Daten, z. B. numerische Daten? Wenn ja, führen Sie die erforderlichen Transformationen durch.
- Müssen die Daten vor der Modellierung normalisiert werden?
- Können fehlende Attribute mithilfe von Aggregation, Durchschnittsbildung oder Induktion erstellt werden?
- Basierend auf Ihrem Hintergrundwissen, gibt es wichtige Fakten (wie die Zeit, die ein Benutzer auf der Website verbracht hat), die von vorhandenen Feldern abgeleitet werden können?

Integrieren von Daten

Es ist durchaus üblich, dass Sie mehrere Datenquellen für dasselbe Set an Geschäftsfragen haben, z. B. wenn Sie möglicherweise Zugriff auf die Hypothekendarlehensdaten sowie käuflich erworbene demografische Daten für dasselbe Set an Kunden haben. Falls diese Datasets dieselbe eindeutige ID (z. B. Personalausweisnummer) aufweisen, können Sie sie mithilfe dieses Felds in IBM SPSS Modeler zusammenführen.

Beim Integrieren von Daten werden zwei grundlegende Verfahren unterschieden:

- **Verbinden** von Daten: Hierbei werden zwei Datasets mit ähnlichen Datensätzen, aber unterschiedlichen Attributen zusammengeführt. Das Verbinden der Daten geschieht mithilfe derselben Schlüsselkennung für jeden Datensatz (wie z. B. Kunden-ID). Die sich daraus ergebenden Daten nehmen in Spalten oder Merkmalen zu.
- **Anhängen** von Daten: Hierbei werden mindestens zwei Datasets mit ähnlichen Attributen, aber unterschiedlichen Datensätzen, integriert. Die Integration der Daten erfolgt anhand eines ähnlichen Felds (wie z. B. Produktname oder Vertragslaufzeit).

Beispiel aus dem elektronischen Einzelhandel - Integrieren von Daten

Ein Web-Mining-Szenario mit CRISP-DM

Bei mehreren Datenquellen hat der e-Einzelhändler viele verschiedene Möglichkeiten, die Daten zu integrieren:

- **Hinzufügen von Kunden- und Produktattributen zu Ereignisdaten.** Um Webprotokollereignisse mithilfe von Attributen aus anderen Datenbanken zu modellieren, muss jede Kunden-ID, Produktnummer und Bestellnummer, die mit jedem Ereignis verknüpft ist, richtig identifiziert werden und die entsprechenden Attribute zu den verarbeiteten Webprotokollen müssen zusammengeführt werden. Beachten Sie, dass die zusammengeführte Datei die Kunden- und Produktinformationen immer dann repliziert, wenn ein Kunde oder Produkt mit einem Ereignis verknüpft wird.
- **Hinzufügen von Kauf- und Webprotokollinformationen zu Kundendaten.** Um den Wert eines Kunden in einem Modell darzustellen, müssen seine Kauf- und Sitzungsinformationen aus den entsprechenden Datenbanken herausgesucht, zusammengezählt und mit der Kundendatenbank verbunden werden. Dieser Vorgang umfasst das Erstellen neuer Attribute, wie im Abschnitt zum Erstellen von Daten erläutert.

Nach dem Integrieren von Datenbanken führt der e-Einzelhändler einen Explorationsprozess durch, um sicherzustellen, dass die Datenintegration richtig ausgeführt wurde.

Integrationsaufgaben

Wenn Sie nicht ausreichend Zeit investiert haben, um ein Datenverständnis zu entwickeln, kann die Integration von Daten zu einem komplexen Vorgang werden. Denken Sie über Elemente und Attribute nach, die am relevantesten für die Data-Mining-Ziele sind, und beginnen Sie anschließend mit der Integration Ihrer Daten.

Aufgabenliste

- Integrieren Sie mithilfe von Zusammenführungs- oder Anhangknoten in IBM SPSS Modeler die Datasets, die für die Modellierung als nützlich erachtet werden.
- Überlegen Sie, ob Sie die resultierende Ausgabe speichern, bevor Sie mit dem Modellierungsprozess fortfahren.
- Nach dem Zusammenführen können Daten durch **Aggregieren** von Werten vereinfacht werden. Bei der Aggregation werden neue Werte durch Zusammenfassen mehrerer Datensätze und/oder Tabellen berechnet.
- Möglicherweise müssen Sie auch neue Datensätze generieren (z. B. den durchschnittlichen Abzug aus kombinierten Steuererklärungen mehrerer Jahre).

Formatieren von Daten

Als letzten Schritt vor der Modellierung sollten Sie überprüfen, ob bestimmte Verfahren ein besonderes Datenformat oder eine spezielle Datenreihenfolge erfordern. Es ist beispielsweise nicht unüblich, dass die Daten für einen Sequenzalgorithmus vor dem Ausführen des Modells vorsortiert sein müssen. Selbst wenn das Modell diese Sortierung für Sie übernehmen kann, sparen Sie möglicherweise Verarbeitungszeit, wenn Sie vor der Modellierung einen Sortierknoten nutzen.

Aufgabenliste

Berücksichtigen Sie beim Formatieren von Daten die folgenden Fragen:

- Welche Modelle möchten Sie verwenden?
- Benötigen diese Modelle ein besonderes Datenformat oder eine spezielle Datenreihenfolge?

Wenn Änderungen empfohlen werden, können die Verarbeitungstools in IBM SPSS Modeler Sie bei der notwendigen Datenbearbeitung unterstützen.

Sind Sie bereit für die Modellierung?

Beantworten Sie unbedingt die folgenden Fragen, bevor Sie Modelle in IBM SPSS Modeler erstellen.

- Kann auf alle Daten in IBM SPSS Modeler zugegriffen werden?
- Konnten Sie auf Grundlage Ihrer ersten Exploration und Ihrer Datenuntersuchung relevante Subsets an Daten auswählen?
- Haben Sie die Daten effektiv bereinigt und nicht verwertbare Elemente entfernt? Dokumentieren Sie alle Entscheidungen im Abschlussbericht.
- Wurden mehrere Datasets richtig integriert? Sind Probleme beim Zusammenführen aufgetreten, die dokumentiert werden sollten?
- Haben Sie die Anforderungen der Modellierungstools, die Sie verwenden möchten, recherchiert?
- Gibt es Formatierungsprobleme, die Sie vor der Modellierung lösen können? Dazu gehören sowohl Überlegungen zur erforderlichen Formatierung als auch Aufgaben, die die Modellierungszeit verkürzen könnten.

Wenn Sie die oben genannten Fragen beantworten können, sind Sie für das Kernstück des Data-Mining bereit - die Modellierung.

Kapitel 5. Modellierung

Übersicht über die Modellierung

An diesem Punkt nun beginnt sich Ihre mühevolle Arbeit auszuzahlen. Die von Ihnen aufbereiteten Daten wurden in das Analysetool in IBM SPSS Modeler importiert und die Ergebnisse geben ersten Aufschluss über das Geschäftsproblem, das sich während der Phase Untersuchung der Geschäftsziele dargestellt hat.

Die Modellierung wird in der Regel in mehreren Schritten durchgeführt. Normalerweise führen Data-Mining-Experten verschiedene Modelle mit den Standardparametern aus und stimmen die Parameter dann ab oder kehren zur Datenaufbereitungsphase zurück, um die Änderungen durchzuführen, die vom Modell ihrer Wahl gefordert werden. In den seltensten Fällen kann die Data-Mining-Frage eines Unternehmens mit nur einem Modell und einer einzigen Ausführung zufriedenstellend beantwortet werden. Und genau das macht das Data-Mining so interessant. Es gibt viele Wege, ein vorhandenes Problem zu betrachten, und IBM SPSS Modeler bietet eine breite Palette an Tools, die Sie dabei unterstützen.

Auswählen der Modellierungsverfahren

Obgleich Sie bereits eine Vorstellung davon haben, welche Typen der Modellierung am besten für die Anforderungen Ihres Unternehmens geeignet sind, ist es jetzt an der Zeit, klare Entscheidungen hinsichtlich der zu verwendenden Modelle zu treffen. Die Bestimmung des geeignetsten Modells basiert in der Regel auf den folgenden Überlegungen:

- **Den für das Mining verfügbaren Datentypen.** Beispielsweise, ob die gewünschten Felder kategorial (symbolisch) sind?
- **Ihren Data-Mining-Zielen.** Möchten Sie einfach nur Einblick in Transaktionsdatenbestände erhalten und interessante Kaufmuster aufdecken? Oder müssen Sie einen Score erzielen, der z. B. die Neigung, mit der Rückzahlung eines Studentendarlehens in Verzug zu geraten, verdeutlicht?
- **Bestimmten Modellierungsanforderungen.** Erfordert das Modell eine bestimmte Datengröße oder einen bestimmten Datentyp? Benötigen Sie ein Modell mit einfach darzustellenden Ergebnissen?

Weitere Informationen zu den Modelltypen in IBM SPSS Modeler und deren Anforderungen finden Sie in der Dokumentation oder Onlinehilfe zu IBM SPSS Modeler.

Beispiel aus dem elektronischen Einzelhandel - Modellierungsverfahren

Die vom e-Einzelhändler verwendeten Modellierungsverfahren werden von den Data-Mining-Zielen des Unternehmens bestimmt:

Verbesserte Empfehlungen. In seiner einfachsten Form umfasst dies das Clustern von Bestellungen, um zu bestimmen, welche Produkte am häufigsten zusammen erworben werden. Kundendaten und selbst Aufzeichnungen über Seitenbesuche können für verbesserte Ergebnisse hinzugefügt werden. Für diesen Modellierungstyp eignen sich das TwoStep- oder das Kohonen-Netzclustering-Verfahren. Im Anschluss können die Cluster mit einem C5.0-Regelset erstellt werden, um zu bestimmen, welche Empfehlungen zu einem beliebigen Zeitpunkt während des Webseitenbesuchs des Kunden am geeignetsten waren.

Verbesserte Sitenavigation. Fürs Erste konzentriert sich der e-Einzelhändler auf das Identifizieren von Seiten, die oft verwendet werden, bei denen der Benutzer aber mehrmals klicken muss, um sie zu finden. Dazu gehört die Anwendung eines Sequenzierungsalgorithmus auf die Webprotokolle, um den "eindeutigen Pfad" zu generieren, den Kunden auf der Website verfolgen können, und dann die spezielle Suche nach Sitzungen, die eine hohe Anzahl an Seitenaufrufen aufweisen, ohne dass eine Aktion durchgeführt

wird (oder bevor eine Aktion durchgeführt wird). Später, in einer tief greifenderen Analyse, können mithilfe von Clustering-Verfahren verschiedene "Typen" von Aufrufen und Besuchern identifiziert werden und der Inhalt der Site kann entsprechend dem Typ organisiert und dargestellt werden.

Auswählen des richtigen Modellierungsverfahrens

In IBM SPSS Modeler steht eine Vielzahl von Verfahren zur Verfügung. Häufig verwenden Data-Mining-Experten mehr als ein Verfahren, um das Problem aus mehreren Richtungen in Angriff zu nehmen.

Aufgabenliste

Berücksichtigen Sie bei der Entscheidung, welche(s) Modell(e) verwendet werden soll(en), ob die folgenden Probleme Ihre Wahl beeinflussen:

- Setzt das Modell voraus, dass die Daten in Test- und Trainingssets aufgeteilt werden?
- Verfügen Sie über ausreichend Daten, um zuverlässige Ergebnisse für das festgelegte Modell zu erzielen?
- Erfordert das Modell ein bestimmtes Datenqualitätsniveau? Können Sie dieses Niveau mit den aktuellen Daten erfüllen?
- Verfügen Sie über den korrekten Datentyp für ein bestimmtes Modell? Wenn nicht, können Sie die notwendigen Umwandlungen mithilfe von Datenbearbeitungsknoten vornehmen?

Weitere Informationen zu den Modelltypen in IBM SPSS Modeler und deren Anforderungen finden Sie in der Dokumentation oder Online-Hilfe zu IBM SPSS Modeler.

Annahmen der Modellierung

Wenn Sie beginnen, die Modellierungstools Ihrer Wahl einzugrenzen, machen Sie sich unbedingt Notizen zum Entscheidungsprozess. Dokumentieren Sie alle Datenannahmen sowie alle Datenbearbeitungen, die Sie vorgenommen haben, um die Anforderungen des Modells zu erfüllen.

Beispielsweise müssen bei den Knoten der logischen Regression und des neuronalen Netzes die Datentypen vor der Ausführung vollständig **instanziiert** (Datentypen sind bekannt) sein. Das bedeutet, dass Sie einen Typknoten zum Stream hinzufügen und ihn ausführen müssen, sodass er die Daten durchleitet, bevor Sie ein Modell erstellen und ausführen. Auf ähnliche Weise profitieren Vorhersagemodelle wie C5.0 von der Umschichtung der Daten bei der Vorhersage von Regeln für seltene Ereignisse. Wenn Sie diese Art der Vorhersage treffen, erhalten Sie oft bessere Ergebnisse, indem Sie einen Balancierungsknoten in den Stream einfügen und die ausgewogeneren Subsets in das Modell einlesen.

Vergessen Sie nicht, diese Entscheidungen zu dokumentieren.

Generieren eines Testdesigns

Als letzten Schritt vor der eigentlichen Modellierung sollten Sie sich einen Moment Zeit nehmen und erneut überdenken, wie die Ergebnisse des Modells getestet werden sollen. Das Generieren eines umfassenden Testdesigns besteht aus zwei Teilen:

- Beschreiben der Kriterien für die "Güte" eines Modells
- Definieren der Daten, an denen diese Kriterien getestet werden

Die **Güte** eines Modells kann auf verschiedene Weisen gemessen werden. Bei überwachten Modellen, wie z. B. C5.0 und C&R-Baum, schätzt die Gütemessung in der Regel die Fehlerrate eines bestimmten Modells. Bei nicht überwachten Modellen, wie z. B. den Kohonen-Clusternetzen, umfassen die Messungen möglicherweise Kriterien wie die einfache Interpretation, Bereitstellung oder die erforderliche Bearbeitungszeit.

Bedenken Sie immer, dass die Modellierung ein schrittweiser Prozess ist. Das bedeutet, dass Sie normalerweise die Ergebnisse verschiedener Modelle testen, bevor Sie sich für das Modell entscheiden, das Sie verwenden und bereitstellen möchten.

Schreiben eines Testdesigns

Beim Testdesign handelt es sich um eine Beschreibung der Schritte, die Sie zum Testen der erstellten Modelle ausführen. Da die Modellierung einen schrittweisen Prozess darstellt, ist es wichtig zu wissen, wann die Parameter nicht weiter angepasst, sondern eine andere Methode oder ein anderes Modell ausprobiert werden sollten.

Aufgabenliste

Berücksichtigen Sie beim Erstellen des Testdesigns die folgenden Fragen:

- Welche Daten werden zum Testen der Modelle verwendet? Haben Sie die Daten in Trainings-/Testsets partitioniert? (Das ist ein weit verbreiteter Ansatz in der Modellierung.)
- Wie messen Sie den Erfolg von überwachten Modellen (wie C5.0)?
- Wie messen Sie den Erfolg von nicht überwachten Modellen (wie den Kohonen-Clusternetzen)?
- Wie oft würden Sie ein Modell nochmals mit angepassten Einstellungen ausführen, bevor Sie einen anderen Modelltyp versuchen?

Beispiel aus dem elektronischen Einzelhandel - Testdesign

Ein Web-Mining-Szenario mit CRISP-DM

Die Kriterien, anhand derer die Modelle beurteilt werden, hängen von den in Betracht gezogenen Modellen und den Data-Mining-Zielen ab:

Verbesserte Empfehlungen. Bis die verbesserten Empfehlungen den Live-Kunden vorgestellt werden, gibt es keine völlig objektive Methode, sie zu bewerten. Der e-Einzelhändler kann jedoch verlangen, dass die Regeln, die die Empfehlungen generieren, einfach genug sind, um aus Unternehmenssicht sinnvoll zu sein. Andererseits sollten die Regeln komplex genug sein, um unterschiedliche Empfehlungen für verschiedene Kunden und Sitzungen zu generieren.

Verbesserte Sitenavigation. Angesichts der Beweise dafür, auf welche Seiten die Kunden auf der Website zugreifen, kann der e-Einzelhändler das aktualisierte Site-Design hinsichtlich eines einfachen Zugriffs auf wichtige Seiten objektiv bewerten. Wie bei den Empfehlungen ist es aber auch hier schwer im Voraus zu bewerten, wie gut sich Kunden an die neu gestaltete Site anpassen. Wenn es die Zeit und die finanziellen Mittel erlauben, sind einige Tests zur Benutzerfreundlichkeit angebracht.

Erstellen der Modelle

Nun sollten Sie gut darauf vorbereitet sein, die Modelle zu erstellen, die Sie so lange geprüft haben. Lassen Sie sich Zeit und den Spielraum, mit einer Reihe verschiedener Modelle zu experimentieren, bevor Sie endgültige Schlussfolgerungen ziehen. Die meisten Data-Mining-Experten erstellen in der Regel mehrere verschiedene Modelle und vergleichen die Ergebnisse, bevor Sie sie bereitstellen oder integrieren.

Um Ihre Fortschritte mit einer Reihe von Modellen nachzuverfolgen, sollten Sie sich unbedingt die für jedes Modell verwendeten Einstellungen und Daten notieren. Somit können Sie später die Ergebnisse mit anderen diskutieren und Ihre Schritte bei Bedarf zurückverfolgen. Am Ende des Modellierungsprozesses verfügen Sie über drei Informationen, die Sie in den Data-Mining-Entscheidungen verwenden können:

- **Parametereinstellungen** umfassen die Notizen, die Sie zu Parametern mit den besten Ergebnissen gemacht haben.
- Die tatsächlich erstellten **Modelle**.

- **Beschreibungen der Modellergebnisse**, einschließlich Leistungs- und Datenproblemen, die während der Ausführung des Modells und der Exploration seiner Ergebnisse aufgetreten sind.

Beispiel aus dem elektronischen Einzelhandel - Modellierung

Ein Web-Mining-Szenario mit CRISP-DM

Verbesserte Empfehlungen. Cluster werden für verschiedene Ebenen der Datenintegration erstellt. Dabei wird zuerst nur mit der Einkaufsdatenbank begonnen und anschließend werden Kunden- und Sitzungsinformationen mit aufgenommen. Für jede Integrationsebene werden Cluster unter wechselnden Parametereinstellungen für die TwoStep- und Kohonen-Netzalgorithmen erstellt. Für jeden dieser Cluster werden einige C5.0-Regelsets mit unterschiedlichen Parametereinstellungen generiert.

Verbesserte Sitenavigation. Der Sequenzmodellierungsknoten wird verwendet, um Kundenpfade zu erzeugen. Der Algorithmus erlauben die Spezifikation eines minimalen Supportkriteriums, was bei der Fokussierung auf die gebräuchlichsten Kundenpfade nützlich ist. Es wurden verschiedene Einstellungen für die Parameter getestet.

Parametereinstellungen

Die meisten Modellierungsverfahren verfügen über eine Vielzahl an Parametern oder Einstellungen, die zur Steuerung des Modellierungsprozesses angepasst werden können. Entscheidungsbäume können beispielsweise durch Anpassen der Baumtiefe, Aufteilungen und eine Reihe anderer Einstellungen gesteuert werden. In der Regel erstellen die meisten Personen zuerst ein Modell mit den Standardoptionen und verfeinern dann in nachfolgenden Sitzungen die Parameter.

Sobald Sie die Parameter ermittelt haben, die die genauesten Ergebnisse liefern, sollten Sie nicht vergessen, den Stream und die erzeugten Modellknoten zu speichern. Aufzeichnungen zu den optimalen Einstellungen können außerdem hilfreich sein, wenn Sie das Modell automatisieren oder mit neuen Daten neu erstellen möchten.

Ausführen der Modelle

In IBM SPSS Modeler ist das Ausführen von Modellen eine einfache Aufgabe. Sobald Sie den Modellknoten in den Stream eingefügt und beliebige Parameter bearbeitet haben, führen Sie das Modell einfach aus, um darstellbare Ergebnisse zu erzielen. Ergebnisse werden im Navigationsbereich "Generierte Modelle" auf der rechten Seite des Arbeitsbereichs angezeigt. Klicken Sie mit der rechten Maustaste auf ein Modell, um die Ergebnisse zu durchsuchen. Bei den meisten Modellen können Sie das generierte Modell in den Stream einfügen, um die Ergebnisse weiter zu evaluieren und bereitzustellen. Außerdem können die Modelle in IBM SPSS Modeler gespeichert werden, um sie später einfach wiederverwenden zu können.

Modellbeschreibung

Vergessen Sie beim Untersuchen der Ergebnisse eines Modells nicht, sich Notizen zu Ihren Erfahrungen mit dem Modell zu machen. Sie können Notizen über das Dialogfeld zu Anmerkungen auf Knotenebene oder das Projekttool direkt im Modell speichern.

Aufgabenliste

Dokumentieren Sie für jedes Modell Informationen wie:

- Können Sie sinnvolle Schlussfolgerungen aus diesem Modell ziehen?
- Gewährt das Modell neue Einblicke oder deckt es ungewöhnliche Muster auf?
- Gab es beim Ausführen des Modells Probleme? Wie angemessen war die Verarbeitungszeit?
- Bestanden bei dem Modell Schwierigkeiten mit Datenqualitätsproblemen, wie z. B. eine große Anzahl an fehlenden Werten?
- Gab es Inkonsistenzen in der Berechnung, die vermerkt werden sollten?

Bewerten des Modells

Da Sie nun ein Set erster Modelle zur Verfügung haben, betrachten Sie diese genauer und bestimmen Sie, welche dieser Modelle präzise und effektiv genug sind, um zum endgültigen Modell zu werden. "Endgültig" kann hier mehrere Bedeutungen haben, wie z. B. "bereit für die Bereitstellung" oder "Darstellen interessanter Muster". Wenn Sie den in einem früheren Schritt erstellten Testplan berücksichtigen, kann Ihnen das bei der Bewertung aus der Sicht Ihres Unternehmens helfen.

Umfassende Modellbewertung

Für jedes zur Diskussion stehende Modell ist es sinnvoll, eine methodische Bewertung anhand der in Ihrem Testplan generierten Kriterien durchzuführen. An dieser Stelle können Sie das generierte Modell dem Stream hinzufügen und mithilfe von Evaluierungsdiagrammen oder Analyseknoten die Effektivität der Ergebnisse analysieren. Sie sollten außerdem überdenken, ob die Ergebnisse logischen Sinn ergeben oder ob sie allzu einfach für Ihre Geschäftsziele sind (wie das z. B. bei einer Sequenz der Fall ist, die Einkäufe wie Wein > Wein > Wein aufzeigt).

Nachdem Sie Ihre Auswertung abgeschlossen haben, sortieren Sie die Modelle nach Rangordnung anhand von objektiven (Modellgenauigkeit) und subjektiven (Benutzerfreundlichkeit oder Interpretation der Ergebnisse) Kriterien.

Aufgabenliste

- Evaluieren Sie die Ergebnisse Ihres Modells mithilfe der Data-Mining-Tools in IBM SPSS Modeler, wie Evaluierungsdiagramme, Analyseknoten oder Kreuzvalidierungsdiagramme.
- Überprüfen Sie die Ergebnisse anhand Ihres Verständnisses des Geschäftsproblems. Beraten Sie sich mit Datenanalysten oder anderen Experten, die möglicherweise Einblick in die Relevanz bestimmter Ergebnisse haben.
- Überlegen Sie, ob die Ergebnisse eines Modells einfach bereitzustellen sind. Erfordert Ihr Unternehmen, dass Ergebnisse über das Internet bereitgestellt oder zurück an das Data Warehouse gesendet werden?
- Analysieren Sie den Einfluss Ihrer Ergebnisse auf Ihre Erfolgskriterien. Erfüllen Sie die während der Phase "Untersuchung der Geschäftsziele" festgelegten Ziele?

Wenn Sie die oben genannten Probleme erfolgreich behandeln konnten und der Meinung sind, dass die aktuellen Modelle Ihre Ziele erfüllen, sollten Sie sich einer gründlicheren Evaluierung der Modelle und einer endgültigen Bereitstellung zuwenden. Berücksichtigen Sie andernfalls die gelernten Dinge und führen Sie die Modelle mit angepassten Parametereinstellungen erneut aus.

Beispiel aus dem elektronischen Einzelhandel - Modellbewertung

Ein Web-Mining-Szenario mit CRISP-DM

Verbesserte Empfehlungen. Eines der Kohonen-Netze und ein TwoStep-Cluster erzielen jeweils akzeptable Ergebnisse und dem e-Einzelhändler fällt es schwer, eine Wahl zwischen beiden zu treffen. Das Unternehmen hofft, mit der Zeit beide Modelle verwenden zu können, und akzeptiert die Vorschläge, in denen die beiden Verfahren übereinstimmen, und studiert eingehender die Situationen, in denen sie sich unterscheiden. Mit einem geringen Arbeitsaufwand und angewendetem Fachwissen kann der e-Einzelhändler weitere Regeln entwickeln, um Unterschiede zwischen den beiden Verfahren zu beseitigen.

Der e-Einzelhändler ist auch der Meinung, dass die Ergebnisse, die die Sitzungsinformationen enthalten, erstaunlich gut sind. Es weist vieles darauf hin, dass die Vorschläge mit der Sitenavigation verbunden werden könnten. Ein Regelset, das definiert, welche Seite der Kunden wahrscheinlich als Nächstes aufrufen wird, kann in Echtzeit verwendet werden, um den Site-Inhalt direkt zu beeinflussen, während der Kunde im Internet surft.

Verbesserte Sitenavigation. Das Sequenzmodell bietet dem e-Einzelhändler eine hohe Stufe des Vertrauens, dass bestimmte Kundenpfade vorhergesagt werden können und Ergebnisse liefern, die eine überschaubare Anzahl an Änderungen des Sitedesigns vorschlagen.

Behalten des Überblicks über geänderte Parameter

Es ist an der Zeit, sich die Modelle auf Grundlage des während der Modellbewertung Gelernten noch einmal anzusehen. Sie haben zwei Möglichkeiten:

- Sie passen die Parameter der vorhandenen Modelle an.
- Sie wählen ein anderes Modell aus, um Ihre Data-Mining-Probleme zu lösen.

In beiden Fällen müssen Sie zur Modellerstellung zurückkehren und diese Aufgabe so lange wiederholen, bis die Ergebnisse erfolgreich sind. Machen Sie sich keine Sorgen, falls Sie diesen Schritt mehrmals ausführen. Es ist durchaus üblich, dass Data-Mining-Experten Modelle mehrere Male evaluieren und ausführen, bevor Sie das Modell finden, das Ihren Anforderungen entspricht. Das spricht auch dafür, mehrere Modelle auf einmal zu erstellen und die Ergebnisse zu vergleichen, bevor die Parameter für jedes Modell angepasst werden.

Sind Sie bereit für den nächsten Schritt?

Bevor Sie mit der abschließenden Evaluierung der Modelle beginnen, überlegen Sie noch einmal, ob Ihre erste Bewertung gründlich genug ausgefallen ist.

Aufgabenliste

- Sind die Ergebnisse der Modelle für Sie verständlich?
- Ergeben die Modellergebnisse aus rein logischer Sicht für Sie Sinn? Gibt es offensichtliche Inkonsistenzen, die weiter untersucht werden sollten?
- Beantworten die Ergebnisse auf den ersten Blick die Geschäftsfrage Ihres Unternehmens?
- Haben Sie Analyseknoten und Lift- oder Gewinndiagramme verwendet, um die Modellgenauigkeit zu vergleichen und zu evaluieren?
- Haben Sie mehrere Modelltypen untersucht und die Ergebnisse verglichen?
- Sind die Ergebnisse Ihres Modells bereitstellbar?

Wenn die Ergebnisse Ihrer Modellierung genau und relevant erscheinen, sollten Sie nun vor der endgültigen Bereitstellung eine genauere Evaluierung durchführen.

Kapitel 6. Evaluierung

Evaluierung - Übersicht

Zum jetzigen Zeitpunkt haben Sie den Großteil Ihres Data-Mining-Projekts bereits abgeschlossen. Außerdem haben Sie in der Modellierungsphase ermittelt, dass die erstellten Modelle entsprechend den in einer früheren Phase definierten **Kriterien für den Data-Mining-Erfolg** technisch richtig und effektiv sind.

Bevor Sie fortfahren, sollten Sie jedoch die Ergebnisse Ihrer Arbeiten mithilfe der **Kriterien für den Unternehmenserfolg** evaluieren, die Sie zu Beginn des Projekts festgelegt haben. Dieser Schritt ist wichtig, um sicherzustellen, dass Ihr Unternehmen auch Nutzen aus den erzielten Ergebnisse ziehen kann. Beim Data-Mining werden zwei Typen von Ergebnissen erzielt:

- Die endgültigen **Modelle**, die Sie in der vorherigen Phase von CRISP-DM ausgewählt haben.
- Sämtliche Schlussfolgerungen oder Rückschlüsse, die aus den Modellen selbst und aus dem Data-Mining-Prozess gezogen werden. Diese werden als **Ergebnisse** bezeichnet.

Evaluieren der Ergebnisse

In dieser Phase formalisieren Sie Ihre Bewertung, ob die Projektergebnisse die Kriterien für den Unternehmenserfolg erfüllen oder nicht. Dafür ist ein klares Verständnis der festgelegten Geschäftsziele erforderlich. Stellen Sie also unbedingt sicher, Hauptentscheidungsträger in die Projektbewertung mit einzubeziehen.

Aufgabenliste

Zuerst müssen Sie Ihre Bewertung, ob die Data-Mining-Ergebnisse die Kriterien für den Unternehmenserfolg erfüllen, dokumentieren. Berücksichtigen Sie in Ihrem Bericht die folgenden Fragen:

- Sind Ihre Ergebnisse klar und in einer Form dargelegt, die einfach präsentiert werden kann?
- Gibt es besonders ungewöhnliche oder einzigartige Ergebnisse, die hervorgehoben werden sollten?
- Können Sie den Modellen und Ergebnissen in der Reihenfolge ihrer Anwendbarkeit auf die Geschäftsziele einen Rang zuordnen?
- Ganz allgemein gesehen, wie gut entsprechen diese Ergebnisse den Geschäftszielen Ihres Unternehmens?
- Welche zusätzlichen Fragen haben diese Ergebnisse aufgeworfen? Wie könnten Sie diese Fragen aus Unternehmenssicht formulieren?

Nachdem Sie die Ergebnisse evaluiert haben, erstellen Sie eine Liste der zugelassenen Modelle, die in den Abschlussbericht aufgenommen werden sollen. Diese Liste sollte Modelle enthalten, die sowohl die Data-Mining-Ziele als auch die Geschäftsziele Ihres Unternehmens erfüllen.

Beispiel aus dem elektronischen Einzelhandel - Evaluieren der Ergebnisse

Ein Web-Mining-Szenario mit CRISP-DM

Das Gesamtergebnis der ersten Erfahrung des e-Einzelhändlers mit Data-Mining lässt sich aus Unternehmenssicht ziemlich leicht kommunizieren: Die Studie hat Produktempfehlungen, von denen erhofft wird, dass sie Verbesserungen darstellen, und ein besseres Site-Design hervorgebracht. Das verbesserte Site-Design basiert auf den Suchsequenzen der Kunden, die die Sitefunktionen anzeigen, die Kunden wünschen, die sie aber nur über mehrere Schritte erreichen. Der Nachweis darüber, dass die Produktempfehlungen wirklich besser sind, ist wesentlich schwerer zu erbringen, da die Entscheidungsregeln kompliziert wer-

den können. Um den Abschlussbericht erstellen zu können, versuchen die Analysten einige allgemeine Trends in den Regelsets zu ermitteln, die einfacher erläutert werden können.

Rangbewertung der Modelle. Da mehrere der ersten Modelle aus Unternehmenssicht sinnvoll erschienen, wurde die Rangbewertung in dieser Gruppe anhand von statistischen Kriterien, einfacher Auswertbarkeit und Vielfältigkeit vorgenommen. So gab das Modell verschiedene Empfehlungen für unterschiedliche Situationen.

Neue Fragen. Die wichtigste Frage, die sich aus der Studie ergab, lautet: Wie kann der e-Einzelhändler mehr über seine Kunden erfahren? Die Informationen in der Kundendatenbank spielen bei der Bildung von Clustern für Empfehlungen eine wichtige Rolle. Während besondere Regeln für Empfehlungen an Kunden, deren Informationen fehlen, verfügbar sind, sind diese Empfehlungen allgemeiner gehalten als die für registrierte Kunden.

Überprüfungsprozess

Effektive Methodologien umfassen in der Regel Zeit für eine Reflektion der Erfolge und Schwächen des gerade abgeschlossenen Prozesses. Data-Mining bildet dabei keine Ausnahme. Zur Philosophie von CRISP-DM gehört es, dass Sie aus Ihren Erfahrungen lernen, sodass zukünftige Data-Mining-Projekte effektiver werden.

Aufgabenliste

Zuerst sollten Sie die Aktivitäten und Entscheidungen für jede Phase zusammenfassen, einschließlich der Schritte zur Datenaufbereitung, der Modellierung usw. Überdenken Sie anschließend für jede Phase die folgenden Fragen und unterbreiten Sie Verbesserungsvorschläge.

- Hat diese Phase zum Wert der endgültigen Ergebnisse beigetragen?
- Gibt es Wege, diese bestimmte Phase oder diesen bestimmten Vorgang zu rationalisieren oder zu verbessern?
- Welche Misserfolge oder Fehler gab es in dieser Phase? Wie können sie zukünftig vermieden werden?
- Gab es Sackgassen, wie z. B. bestimmte Modelle, die sich als ergebnislos erwiesen haben? Gibt es Möglichkeiten, solche Sackgassen vorherzusehen, sodass Arbeiten effektiver geleitet werden können?
- Gab es Überraschungen (gute oder schlechte) jeglicher Art während dieser Phase? Gibt es im Nachhinein nahe liegende Möglichkeiten, solche Vorkommen vorherzusagen?
- Gibt es alternative Entscheidungen oder Strategien, die möglicherweise in einer bestimmten Phase eingesetzt wurden? Notieren Sie sich solche Alternativen für zukünftige Data-Mining-Projekte.

Beispiel aus dem elektronischen Einzelhandel - Überprüfungsbericht

Ein Web-Mining-Szenario mit CRISP-DM

Als Folge der Überprüfung des Prozesses des ersten Data-Mining-Projekts hat der e-Einzelhändler ein besseres Verständnis der Zusammenhänge zwischen den einzelnen Schritten im Prozess erlangt. Zuerst ist der e-Einzelhändler nur zögerlich im CRISP-DM-Prozess "zurückgegangen", nun erkennt er aber, dass das zyklische Wesen des Prozesses seine Leistung steigert. Durch die Prozessüberprüfung hat der e-Einzelhändler auch Verständnis hinsichtlich folgender Punkte erlangt:

- Das Zurückgehen zum Explorationsprozess ist immer dann gerechtfertigt, wenn etwas Ungewöhnliches in einer anderen Phase des CRISP-DM-Prozesses auftritt.
- Die Datenaufbereitung, insbesondere von Webprotokollen, erfordert Geduld, da sie sehr viel Zeit in Anspruch nehmen kann.
- Es ist überaus wichtig, sich auf das vorliegende Geschäftsproblem zu konzentrieren, denn wenn die Daten für die Analyse bereit sind, ist es allzu einfach, mit der Bildung eines Modells zu beginnen, ohne auf das große Ganze zu achten.

- Sobald die Modellierungsphase abgeschlossen ist, ist die Untersuchung der Geschäftsziele bei der Entscheidung, wie die Ergebnisse umgesetzt werden können und welche weiteren Studien berechtigt sind, sogar noch wichtiger.

Bestimmen der nächsten Schritte

Bisher haben Sie Ergebnisse erstellt, Ihre Data-Mining-Erfahrungen evaluiert und fragen sich möglicherweise nun: **Was kommt als Nächstes?** Diese Phase unterstützt Sie dabei, diese Frage im Hinblick auf Ihre Geschäftsziele für das Data-Mining zu beantworten. Im Wesentlichen haben Sie jetzt zwei Möglichkeiten:

- **Sie fahren mit der Bereitstellungsphase fort.** In der nächsten Phase setzen Sie die Modellergebnisse in Ihrem Geschäftsprozess um und erstellen einen Abschlussbericht. Selbst wenn Ihre Data-Mining-Arbeiten nicht erfolgreich waren, sollten Sie die Bereitstellungsphase von CRISP-DM nutzen, um einen Abschlussbericht für den Projektsponsor zu erstellen.
- **Sie gehen einen Schritt zurück und verfeinern Ihre Modelle oder tauschen sie aus.** Wenn Sie der Meinung sind, dass Ihre Ergebnisse zwar fast, aber nicht ganz optimal sind, überlegen Sie sich, ob Sie einen weiteren Modellierungsprozess durchlaufen möchten. Sie können die in dieser Phase gelernten Dinge verwenden, um die Modelle zu verfeinern und bessere Ergebnisse zu erzielen.

Die Entscheidung, die Sie hier treffen, hängt mit der Genauigkeit und Relevanz der Modellierungsergebnisse zusammen. Wenn die Ergebnisse Ihre Data-Mining- und Geschäftsziele erfüllen, sind Sie bereit für die Bereitstellungsphase. Egal, wie Ihre Entscheidung ausfällt, vergessen Sie nicht, den Evaluierungsprozess genau zu dokumentieren.

Beispiel aus dem elektronischen Einzelhandel - Nächste Schritte

Ein Web-Mining-Szenario mit CRISP-DM

Der e-Einzelhändler ist sowohl von der Genauigkeit als auch von der Relevanz der Projektergebnisse ziemlich überzeugt und fährt deshalb mit der Bereitstellungsphase fort.

Gleichzeitig ist das Projektteam bereit, einen Schritt zurückzugehen und einige der Modelle zu erweitern, um Vorhersageverfahren einzubinden. Momentan warten Sie auf die Abgabe der Abschlussberichte und grünes Licht von den Entscheidungsträgern.

Kapitel 7. Bereitstellung

Bereitstellung - Übersicht

Bei der Bereitstellung nutzen Sie Ihre neuen Einblicke, um Verbesserungen in Ihrem Unternehmen vorzunehmen. Das kann eine formale Integration wie eine Implementierung eines IBM SPSS Modeler-Modells bedeuten, das Abwanderungsraten ausgibt, die dann in ein Data Warehouse eingelesen werden können. Bereitstellung kann aber auch heißen, dass Sie die Einblicke, die Sie durch das Data-Mining erhalten haben, nutzen, um Änderungen in Ihrem Unternehmen auszulösen. Beispielsweise haben Sie vielleicht alarmierende Muster in Ihren Daten erkannt, die eine Änderung im Verhalten von Kunden im Alter von über 30 Jahren anzeigen. Diese Ergebnisse sind möglicherweise nicht ausdrücklich in Ihren Informationssystemen integriert, aber sie sind zweifelsohne für das Planen und Treffen von Marketingentscheidungen nützlich.

Im Allgemeinen umfasst die Phase "Bereitstellung" von CRISP-DM zwei Typen von Aktivitäten:

- Planen und Überwachen der Bereitstellung von Ergebnissen
- Abschließen von Nachbereitungsaufgaben wie das Erstellen eines Abschlussberichts und Durchführen einer Projektbewertung

Je nach den Anforderungen Ihres Unternehmens müssen Sie möglicherweise einen oder beide dieser Schritte ausführen.

Planen der Bereitstellung

Obwohl Sie möglicherweise bestrebt sind, die Früchte Ihrer Data-Mining-Arbeiten schnell mit anderen zu teilen, sollten Sie sich die Zeit nehmen, einen Plan für eine reibungslose und umfassende Bereitstellung der Ergebnisse zu erstellen.

Aufgabenliste

- In einem ersten Schritt sollten Sie Ihre Resultate - Modelle und Ergebnisse - zusammenfassen. Somit können Sie bestimmen, welche Modelle in Ihr Datenbanksystem integriert werden können und welche Ergebnisse Ihren Kollegen vorgestellt werden sollten.
- Erstellen Sie für jedes bereitstellbare Modell einen Stufenplan für die Bereitstellung und Integration in Ihre Systeme. Notieren Sie alle technischen Details wie Datenbankanforderungen für die Modellausgabe, z. B. wenn Ihr System die Bereitstellung der Modellausgabe in einem durch Tabulator getrennten Format erforderlich macht.
- Erstellen Sie für jedes endgültige Ergebnis einen Plan, um diese Informationen an die für die Strategieerstellung verantwortlichen Personen weiterzugeben.
- Gibt es alternative Bereitstellungspläne für beide Typen von Ergebnissen, die erwähnt werden sollten?
- Überlegen Sie sich, wie die Bereitstellung überwacht wird. Wie wird z. B. ein Modell, das mit IBM SPSS Modeler Solution Publisher bereitgestellt wurde, aktualisiert? Wie entscheiden Sie, wann das Modell nicht mehr anwendbar ist?
- Bestimmen Sie alle Bereitstellungsprobleme und erstellen Sie einen Plan für Notfälle. Die Entscheidungsträger möchten z. B. möglicherweise weitere Informationen zu den Modellierungsergebnissen und fordern weitere technische Details an.

Beispiel aus dem elektronischen Einzelhandel - Planen der Bereitstellung

Ein Web-Mining-Szenario mit CRISP-DM

Eine erfolgreiche Bereitstellung der Data-Mining-Ergebnisse des e-Einzelhändlers setzt voraus, dass die richtigen Personen die richtigen Informationen erhalten.

Entscheidungsträger. Die Entscheidungsträger müssen über die Empfehlungen und vorgeschlagenen Änderungen an der Site informiert werden und kurze Erläuterungen dazu erhalten, wie diese Änderungen helfen können. Vorausgesetzt, die Ergebnisse der Studie werden akzeptiert, müssen die Personen informiert werden, die diese Änderungen implementieren.

Webentwickler. Die Personen, die die Website pflegen, müssen die neuen Empfehlungen und die Organisation der Site-Inhalte berücksichtigen. Informieren Sie sie darüber, welche Änderungen aufgrund zukünftiger Studien erfolgen *könnten*, sodass sie bereits jetzt die Vorbereitungen dafür treffen können. Wenn Sie das Team für sofortige Site-Erstellungen anhand von Sequenzanalysen in Echtzeit vorbereiten, könnte dies später hilfreich sein.

Datenbankexperten. Die Mitarbeiter, die die Kunden-, Einkaufs- und Produktdatenbanken pflegen, sollten immer darüber Bescheid wissen, wie die Informationen aus den Datenbanken verwendet werden und welche Attribute den Datenbanken in zukünftigen Projekten hinzugefügt werden könnten.

Darüber hinaus muss das Projektteam mit jeder dieser Gruppen in Kontakt stehen, um die Bereitstellung von Ergebnissen zu koordinieren und zukünftige Projekte zu planen.

Planen von Überwachung und Anpassung

Bei einer vollständigen Bereitstellung und Integration von Modellierungsergebnissen könnten Ihre Data-Mining-Arbeiten dauerhaft sein. Wenn z. B. ein Modell bereitgestellt wird, um mehrere aufeinander folgende Einkäufen mit einem elektronischen Warenkorb vorherzusagen, muss dieses Modell aller Wahrscheinlichkeit nach regelmäßig evaluiert werden, um seine Effektivität zu gewährleisten und ständige Verbesserungen vorzunehmen. Ähnlich muss ein Modell, das zur Verbesserung der Kundenbindung bei hochwertigen Kunden bereitgestellt wird, voraussichtlich optimiert werden, sobald ein bestimmter Bindungsgrad erreicht ist. Das Modell wird dann möglicherweise geändert und erneut verwendet, um Kunden auf einer niedrigeren, aber immer noch profitablen Ebene der Wertepyramide beizubehalten.

Aufgabenliste

Machen Sie sich Notizen zu den folgenden Problemen und stellen Sie sicher, dass Sie diese in den Abschlussbericht aufnehmen.

- Welche Faktoren oder Einflüsse (wie Marktwert oder saisonale Schwankungen) müssen für jedes Modell oder Ergebnis aufgezeichnet werden?
- Wie kann die Gültigkeit und Genauigkeit jedes Modells gemessen und überwacht werden?
- Wie bestimmen Sie, wann ein Modell "abgelaufen" ist? Geben Sie Einzelheiten zu Schwellenwerten für die Genauigkeit oder erwarteten Änderungen in Daten usw. an.
- Was passiert, wenn das Modell abgelaufen ist? Können Sie das Modell einfach mit aktuelleren Daten erneut erstellen oder geringfügige Anpassungen vornehmen? Oder sind die Änderungen so tief greifend, dass ein neues Data-Mining-Projekt erforderlich ist?
- Kann dieses Modell nach Ablauf für ähnliche Geschäftsprobleme verwendet werden? An dieser Stelle wird eine gute Dokumentation wichtig, um den Geschäftszweck für jedes Data-Mining-Projekt zu bewerten.

Beispiel aus dem elektronischen Einzelhandel - Überwachung und Anpassung

Ein Web-Mining-Szenario mit CRISP-DM

Die unmittelbare Aufgabe beim Überwachen besteht darin, zu bestimmen, ob die neue Organisation der Site und die verbesserten Empfehlungen wirklich funktionieren. Anders ausgedrückt, können Benutzer

auf direkteren Wegen zu den gesuchten Seiten gelangen? Ist eine Steigerung bei den Cross-Sales der empfohlenen Elemente zu verzeichnen? Nachdem der e-Einzelhändler das Modell einige Wochen überwacht hat, kann er den Erfolg der Studie bestimmen.

Neu registrierte Benutzer können automatisch aufgenommen werden. Wenn sich Kunden bei der Site registrieren, können die aktuellen Regelsets auf ihre Informationen angewendet werden, um die Vorschläge zu bestimmen, die ihnen unterbreitet werden sollten.

Eine weitaus schwierigere Aufgabe ist die Entscheidung, wann die Regelsets für die Bestimmung der Empfehlungen aktualisiert werden sollen. Bei der Aktualisierung der Regelsets handelt es sich nicht um einen automatischen Prozess, da die Cluster-Erstellung ein Eingreifen des Benutzers hinsichtlich der Eignung einer vorgegebenen Clusterlösung erforderlich macht.

Da zukünftige Projekte komplexere Modelle erstellen, werden die Notwendigkeit und der Umfang der Überwachung fast sicher steigen. Wenn möglich, sollte der Großteil der Überwachung automatisch erfolgen, mit regelmäßig geplanten Berichten, die zur Bewertung zur Verfügung stehen. Alternativ dazu könnte die Erstellung von Modellen, die sofortige Vorhersagen bieten, ein Weg sein, den das Unternehmen einschlagen möchte. Das erfordert jedoch mehr Perfektion vom Team als das erste Data-Mining-Projekt.

Erstellen eines Abschlussberichts

Im Abschlussbericht werden nicht nur offene Punkte aus vorherigen Dokumentationen aufgenommen und abgeschlossen, sondern er kann auch zur Übermittlung Ihrer Ergebnisse verwendet werden. Obwohl diese Aufgabe einfach erscheint, ist es wichtig, Ihre Ergebnisse den verschiedenen Personen zu präsentieren, die ein wirtschaftliches Interesse an den Ergebnissen haben. Dazu gehören sowohl technische Administratoren, die für die Implementierung der Modellierungsergebnisse verantwortlich sind, als auch Marketing- und Managementsponsoren, die anhand Ihrer Ergebnisse Entscheidungen treffen.

Aufgabenliste

Betrachten Sie zuerst die Zielgruppe, der Sie Ihren Bericht vorstellen. Handelt es sich um technische Entwickler oder marktorientierte Manager? Möglicherweise müssen Sie unterschiedliche Berichte für jede Zielgruppe erstellen, wenn deren Anforderungen gänzlich verschieden sind. In jedem Fall sollte Ihr Bericht die Mehrzahl der folgenden Punkte umfassen:

- Eine umfassende Beschreibung des ursprünglichen Geschäftsproblems
- Das zum Durchführen des Data-Mining verwendete Verfahren
- Die Kosten des Projekts
- Notizen zu allen Abweichungen vom ursprünglichen Projektplan
- Eine Zusammenfassung der Data-Mining-Resultate - Modelle und Ergebnisse
- Einen Überblick über den vorgeschlagenen Plan für die Bereitstellung
- Empfehlungen für weitere Data-Mining-Arbeiten, einschließlich interessanter Aspekte, sich sich der Exploration und Modellierung ergeben haben

Vorbereiten der Abschlusspräsentation

Zusätzlich zum Projektbericht müssen Sie möglicherweise die Projektergebnisse einem Team von Sponsoren oder verwandten Abteilungen vorstellen. In diesem Fall können Sie in Ihrem Bericht weitgehend dieselben Informationen verwenden, diese aber unter einem breiteren Blickwinkel präsentieren. Die Diagramme und Grafiken in IBM SPSS Modeler können leicht für diese Art der Präsentation exportiert werden.

Beispiel aus dem elektronischen Einzelhandel - Abschlussbericht

Ein Web-Mining-Szenario mit CRISP-DM

Die größte Abweichung vom ursprünglichen Projektplan ist auch ein interessanter Hinweis für zukünftige Data-Mining-Arbeiten. Ziel des ursprünglichen Plans war es herauszufinden, wie Kunden pro Besuch mehr Zeit auf der Site verbringen und mehr Seiten ansehen können.

Wie sich herausstellt, erhält man glückliche Kunden nicht einfach nur dadurch, indem man sie im Internet hält. Häufigkeitsverteilungen der pro Sitzung aufgewendeten Zeit, aufgeteilt danach, ob eine Sitzung zu einem Kauf führte, haben herausgefunden, dass die Sitzungszeiten für die meisten Sitzungen, die zu Käufen führten, zwischen die Sitzungszeiten von zwei Clustern mit Sitzungen ohne Kauf fallen.

Nachdem dies bekannt ist, liegt die Aufgabe darin herauszufinden, ob die Kunden, die viel Zeit auf der Site verbracht haben, ohne etwas zu kaufen, nur im Internet surfen oder einfach nicht die gewünschten Artikel finden können. Als nächsten Schritt müssen Sie herausfinden, wie die gesuchten Artikel angeboten werden können, um Einkäufe zu fördern.

Durchführen einer abschließenden Projektbewertung

Dies ist der letzte Schritt der CRISP-DM-Methodologie. Hier haben Sie die Möglichkeit, Ihre abschließenden Eindrücke zu formulieren und die während des Data-Mining-Prozesses gelernten Lektionen zusammenzutragen.

Aufgabenliste

Sie sollten eine kurze Befragung der wesentlich am Data-Mining-Prozess beteiligten Personen durchführen. Folgende Fragen sollten während dieser Befragungen berücksichtigt werden:

- Wie ist Ihr Gesamteindruck vom Projekt?
- Was haben Sie während des Prozesses gelernt? Sowohl über das Data-Mining im Allgemeinen als auch über die verfügbaren Daten?
- Welche Teile des Projekts verliefen gut? Wo traten Schwierigkeiten auf? Gab es Informationen, die bei der Beseitigung der Unklarheiten hätten behilflich sein können?

Nachdem die Data-Mining-Ergebnisse bereitgestellt wurden, können Sie auch die Personen befragen, die von den Ergebnissen beeinflusst werden, wie z. B. Kunden oder Geschäftspartner. Ihr Ziel sollte es hier sein, zu bestimmen, ob sich das Projekt gelohnt und die Vorteile geboten hat, die es schaffen wollte.

Die Ergebnisse dieser Befragungen können mit Ihren eigenen Eindrücken über das Projekt in einem Abschlussbericht zusammengefasst werden. Dieser Abschlussbericht sollte sich auf die Lektionen konzentrieren, die Sie aus der Erfahrung des Data-Mining Ihrer Datenspeicher gelernt haben.

Beispiel aus dem elektronischen Einzelhandel - Abschließende Bewertung

Ein Web-Mining-Szenario mit CRISP-DM

Befragungen der Projektmitglieder. Der e-Einzelhändler erfährt, dass Projektmitglieder, die von Anfang bis Ende am engsten mit der Studie verbunden waren, größtenteils begeistert von den Ergebnissen sind und sich auf zukünftige Projekte freuen. Die Datenbankgruppe scheint vorsichtig optimistisch. Sie schätzt zwar den Nutzen der Studie, weist aber auf die zusätzliche Belastung der Datenbankressourcen hin. Während der Studie stand ein Berater zur Verfügung, bei weiteren Studien wird ein zusätzlicher Mitarbeiter für die Datenbankwartung notwendig sein, da der Umfang des Projekts wächst.

Kundenbefragungen. Das Feedback der Kunden war bisher weitgehend positiv. Ein nicht gut durchdachtes Problem waren die Auswirkungen, die die Änderung des Site-Designs auf etablierte Kunden haben. Nach einigen Jahren haben die registrierten Kunden eine bestimmte Erwartung hinsichtlich des Site-Aufbaus entwickelt. Das Feedback von registrierten Benutzern ist nicht so positiv wie das von nicht registrierten Kunden. Und einige lehnten die Änderungen gänzlich ab. Dem e-Einzelhändler muss dieses Prob-

lem bewusst sein und er muss sorgfältig abwägen, ob die Änderungen genug neue Kunden bringen, sodass er das Risiko eingehen möchte, vorhandene Kunden zu verlieren.

Bemerkungen

Diese Informationen wurden für weltweit angebotene Produkte und Dienstleistungen erarbeitet.

Möglicherweise bietet IBM die in dieser Dokumentation beschriebenen Produkte, Services oder Funktionen in anderen Ländern nicht an. Informationen über die gegenwärtig im jeweiligen Land verfügbaren Produkte und Services sind beim zuständigen IBM Ansprechpartner erhältlich. Hinweise auf IBM Lizenzprogramme oder andere IBM Produkte bedeuten nicht, dass nur Programme, Produkte oder Services von IBM verwendet werden können. Anstelle der IBM Produkte, Programme oder Services können auch andere, ihnen äquivalente Produkte, Programme oder Services verwendet werden, solange diese keine gewerblichen oder anderen Schutzrechte von IBM verletzen. Die Verantwortung für den Betrieb von Produkten, Programmen und Services anderer Anbieter liegt beim Kunden.

Für in diesem Handbuch beschriebene Erzeugnisse und Verfahren kann es IBM Patente oder Patentanmeldungen geben. Mit der Auslieferung dieses Handbuchs ist keine Lizenzierung dieser Patente verbunden. Lizenzanforderungen sind schriftlich an folgende Adresse zu richten (Anfragen an diese Adresse müssen auf Englisch formuliert werden):

IBM Director of Licensing
IBM Europe, Middle East & Africa
Tour Descartes
2, avenue Gambetta
92066 Paris La Defense
France

Trotz sorgfältiger Bearbeitung können technische Ungenauigkeiten oder Druckfehler in dieser Veröffentlichung nicht ausgeschlossen werden. Die hier enthaltenen Informationen werden in regelmäßigen Zeitabständen aktualisiert und als Neuausgabe veröffentlicht. IBM kann ohne weitere Mitteilung jederzeit Verbesserungen und/oder Änderungen an den in dieser Veröffentlichung beschriebenen Produkten und/oder Programmen vornehmen.

Verweise in diesen Informationen auf Websites anderer Anbieter werden lediglich als Service für den Kunden bereitgestellt und stellen keinerlei Billigung des Inhalts dieser Websites dar. Das über diese Websites verfügbare Material ist nicht Bestandteil des Materials für dieses IBM Produkt. Die Verwendung dieser Websites geschieht auf eigene Verantwortung.

Werden an IBM Informationen eingesandt, können diese beliebig verwendet werden, ohne dass eine Verpflichtung gegenüber dem Einsender entsteht.

Lizenznehmer des Programms, die Informationen zu diesem Produkt wünschen mit der Zielsetzung: (i) den Austausch von Informationen zwischen unabhängig voneinander erstellten Programmen und anderen Programmen (einschließlich des vorliegenden Programms) sowie (ii) die gemeinsame Nutzung der ausgetauschten Informationen zu ermöglichen, wenden sich an folgende Adresse:

IBM Software Group
ATTN: Licensing
200 W. Madison St.
Chicago, IL; 60606
USA

Die Bereitstellung dieser Informationen kann unter Umständen von bestimmten Bedingungen - in einigen Fällen auch von der Zahlung einer Gebühr - abhängig sein.

Die Lieferung des in diesem Dokument beschriebenen Lizenzprogramms sowie des zugehörigen Lizenzmaterials erfolgt auf der Basis der IBM Rahmenvereinbarung bzw. der Allgemeinen Geschäftsbedingungen von IBM, der IBM Internationalen Nutzungsbedingungen für Programmpakete oder einer äquivalenten Vereinbarung.

Alle in diesem Dokument enthaltenen Leistungsdaten stammen aus einer kontrollierten Umgebung. Die Ergebnisse, die in anderen Betriebsumgebungen erzielt werden, können daher erheblich von den hier erzielten Ergebnissen abweichen. Einige Daten stammen möglicherweise von Systemen, deren Entwicklung noch nicht abgeschlossen ist. Eine Gewährleistung, dass diese Daten auch in allgemein verfügbaren Systemen erzielt werden, kann nicht gegeben werden. Darüber hinaus wurden einige Daten unter Umständen durch Extrapolation berechnet. Die tatsächlichen Ergebnisse können davon abweichen. Benutzer dieses Dokuments sollten die entsprechenden Daten in ihrer spezifischen Umgebung prüfen.

Alle Informationen zu Produkten anderer Anbieter stammen von den Anbietern der aufgeführten Produkte, deren veröffentlichten Ankündigungen oder anderen allgemein verfügbaren Quellen. IBM hat diese Produkte nicht getestet und kann daher keine Aussagen zu Leistung, Kompatibilität oder anderen Merkmalen machen. Fragen zu den Leistungsmerkmalen von Produkten anderer Anbieter sind an den jeweiligen Anbieter zu richten.

Aussagen über Pläne und Absichten von IBM unterliegen Änderungen oder können zurückgenommen werden und repräsentieren nur die Ziele von IBM.

Diese Veröffentlichung enthält Beispiele für Daten und Berichte des alltäglichen Geschäftsablaufs. Sie sollen nur die Funktionen des Lizenzprogramms illustrieren und können Namen von Personen, Firmen, Marken oder Produkten enthalten. Alle diese Namen sind frei erfunden; Ähnlichkeiten mit tatsächlichen Namen und Adressen sind rein zufällig.

Marken

IBM, das IBM Logo und ibm.com sind Marken oder eingetragene Marken der IBM Corporation in den USA und/oder anderen Ländern. Weitere Produkt- und Servicenamen können Marken von IBM oder anderen Unternehmen sein. Eine aktuelle Liste der IBM Marken finden Sie auf der Webseite "Copyright and trademark information" unter www.ibm.com/legal/copytrade.shtml.

Intel, das Intel-Logo, Intel Inside, das Intel Inside-Logo, Intel Centrino, das Intel Centrino-Logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium und Pentium sind Marken oder eingetragene Marken der Intel Corporation oder der Tochtergesellschaften des Unternehmens in den USA und anderen Ländern.

Linux ist eine eingetragene Marke von Linus Torvalds in den USA, anderen Ländern oder beidem.

Microsoft, Windows, Windows NT und das Windows-Logo sind Marken der Microsoft Corporation in den USA und/oder anderen Ländern.

UNIX ist eine eingetragene Marke von The Open Group in den USA und anderen Ländern.

Java und alle auf Java basierenden Marken und Logos sind Marken oder eingetragene Marken der Oracle Corporation und/oder ihrer verbundenen Unternehmen.

Weitere Produkt- und Servicenamen können Marken von IBM oder anderen Unternehmen sein.

Index