We make your business more intelligent

Data Mining

Janine Oswald, plus-IT GmbH

07. April 2009
.Net User Group, Regensburg

Agenda

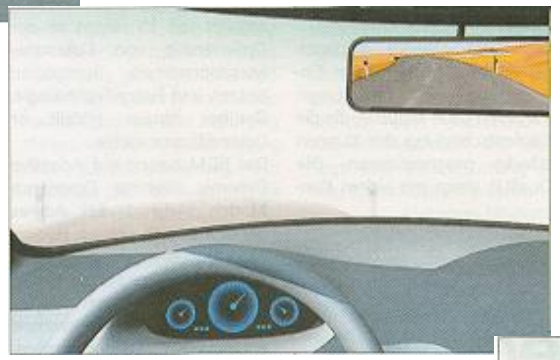
Data Mining - Einführung

Data Mining Grundlagen, Begriffsdefinition
Data Mining Werkzeuge
Branchenlösungen
Data Mining Prozess mit SQL Server 2008
Data Mining Algorithmen mit SQL Server 2008
Data Mining Neuerungen mit SQL Server 2008
Demo: Data Mining mit BIDS
Data Mining mit Office 2007
Zusammenfassung

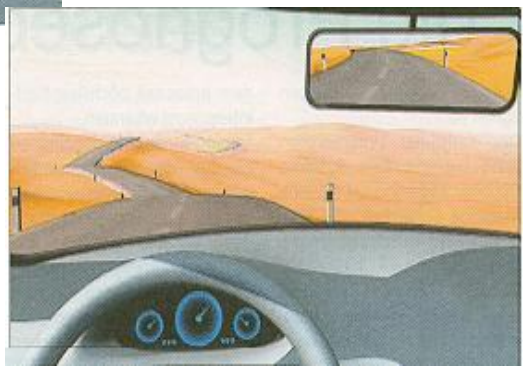
Entscheidungsunterstützung

für das Management

Ohne entscheidungsrelevante Informationen befinden sich Unternehmen oft auf „Blindfahrt".



Data Warehouse, OLAP und Reporting erlaubt den Blick nach hinten in die Vergangenheit.



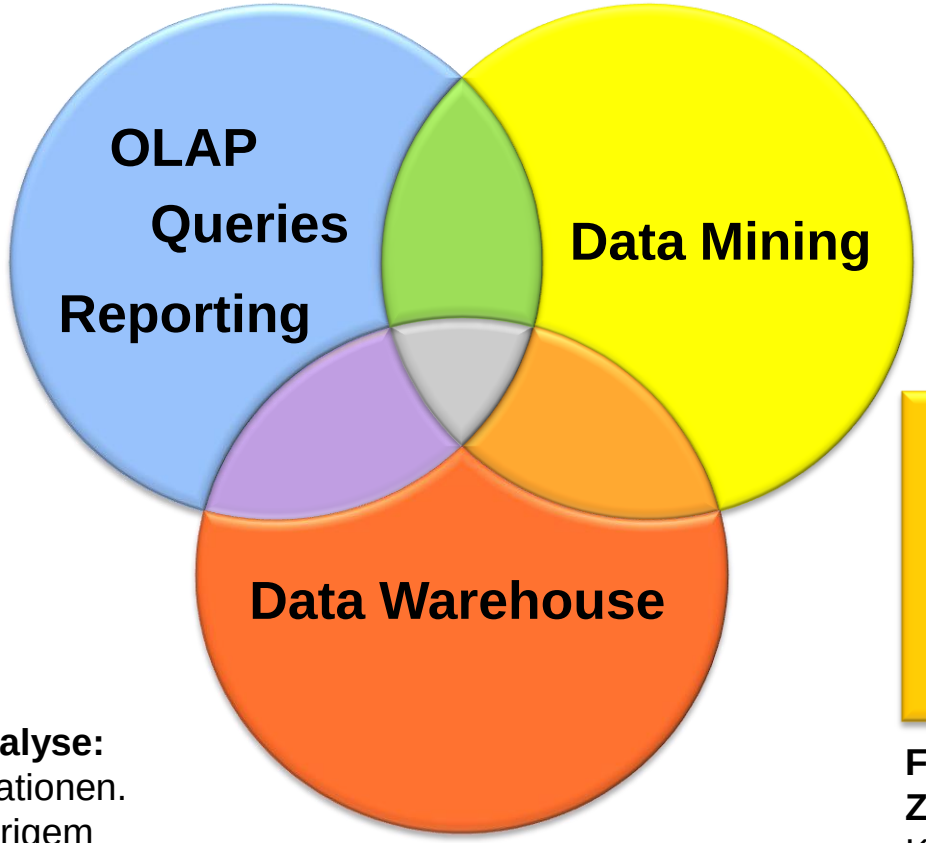
Beim Einsatz von Data Mining-Verfahren können zusätzlich fundierte Vorhersagen über den künftigen Weg getroffen werden.

Quelle:
Computer Zeitung
Nr. 9/2005, S. 17.

Data Mining – Was ist das?

Data Mining - Positionierung

•Was geschah?
•Was ist gerade?
•Historische Sicht
•Manuell/Interaktiv/Reaktiv



•Warum und wie geschieht etwas?
•Historie, Gegenwart & Zukunft
•Automatisierte Verfahren

Wie hoch war das Umsatzvolumen auf allen Girokonten im PLZ-Bereich 18 im 4. Quartal 2006?

Welche Kunden könnten demnächst abwandern und welche typischen Merkmale haben diese Kunden?

Verifikationsgetriebene Analyse:
Gezielte Abfrage von Informationen.
Quantitative Fragen mit niedrigem Freiheitsgrad.

Finden bisher unbekannter Zusammenhänge:
Korrelationen, Muster, Trends.
Qualitative Fragen mit hohem Freiheitsgrad.

Quelle: Rösch, Vortragsfolien, KDD-Tagung 2002

Data Mining

Was ist das?

Was ist nicht Data Mining?

- SQL / Ad Hoc Query / Reports
- Online Analytical Processing (OLAP)
- Statistik

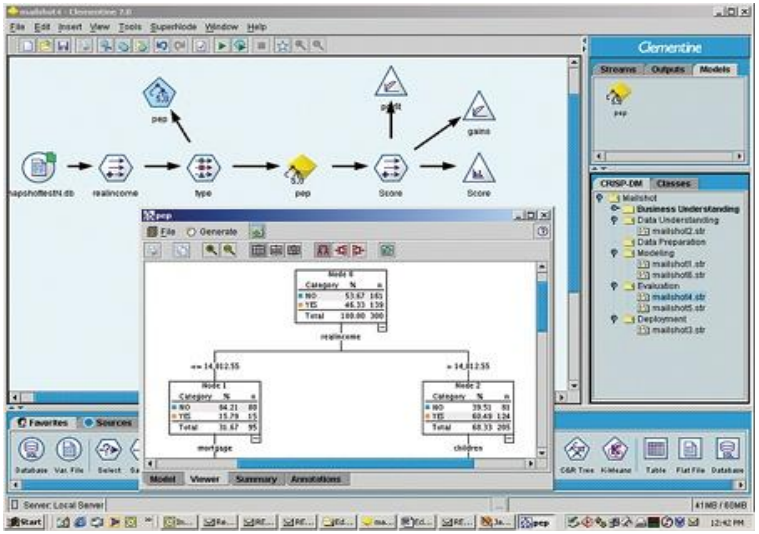
- *to mine for* heißt „schürfen nach“
- Data Mining =
  - + nichttriviales
  - + automatisches Schürfen
  - + nach bedeutsamen Mustern/Zusammenhängen
  - + in Datenbanken
  - + um sie dem Anwender als interessantes Wissen zu präsentieren.
- „Data Mining is the semi-automatic discovery of pattern, associations, anomalies, structures, and changes in large data sets.“ (Grossmann, 1998)
- Basis sind Methoden und Verfahren aus der Statistik und der künstlichen Intelligenz (KI)
- Data Dredging, Fishing, Data Grubbing, Database Exploration, Knowledge Extraction, Information Discovery/Harvesting, Data Archaeology
- KDD (Knowledge Discovery in Databases) und Data Mining

Was macht Data Mining?

Data Mining Werkzeuge

- Auswahl wesentlicher Werkzeuge
  - + Enterprise Minier SAS
  - + Clementine von SPSS
  - + MS SQL Server 2005/2008 (Analysis Services)
  - + Intelligent Miner von IBM
  - + DISCOVERER von Prudential Systems (prudsys)
  - + BASKET ANALYZER von prudsys
  - + KXEN
- Open Source Werkzeuge
  - + RapidMiner (Rapid I)
  - + WEKA
  - + Orange
  - + Knime
  - + CLUTO

Quelle: www.sas.com



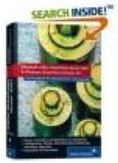
Quelle: Clementine Dokumentation

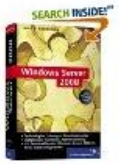
Data Mining bei Amazon

Kunden, die diesen Artikel gekauft haben, kauften auch:

Seite 1 von 13


Microsoft Windows Server 2008 - Die Neuerungen im Überblick von Thomas Joos
(3)


Office SharePoint Server 2007 und Windows SharePoint Serv... von Ulrich B. Boddenberg
(14) EUR 49,90


Windows Server 2008: Technologie, Lösungen, Einsatzszenar... von Ulrich B. Boddenberg
(7) EUR 49,90


SQL Server 2005 - Der schnelle Einstieg. Abfragen, Transa... von Klemens Konopasek
(9) EUR 29,95


Microsoft Office SharePoint Server 2007 - Das Handbuch von Bill English
(2) EUR 59,00


Microsoft SQL Server 2005 - Das Handbuch von Marci Frohock Garcia
(2) EUR 69,00

- Ergebnis: Recommendation List
- Collaborative Filtering
- Einsatz sinnvoll, wenn Informationsmenge und Nutzerzahl sehr hoch
- Kunden bei der Produktauswahl unterstützen
- Personalisierung steigert Bindung zum Shop
- Kundenwünsche erfassen

Warenkorbanalyse:

Kunden, die Bücher von Hajo Hippner gekauft haben, haben auch Bücher dieser Autoren gekauft:

- Manfred Bruhn
- Gregor Stokburger
- Michael Brendel
- Matthias F. Uebel
- Martin Stadelmann

Recommender:

Anhand des Sachgebietes nach ähnlichen Produkten suchen:

- ☐ Customer Relationship Management (CRM)
- ☐ Data Mining (EDV)
- ☐ Internet
- ☐ Wirtschaft, Recht
- ☐ Wirtschaft
- ☐ Werbung, Marketing

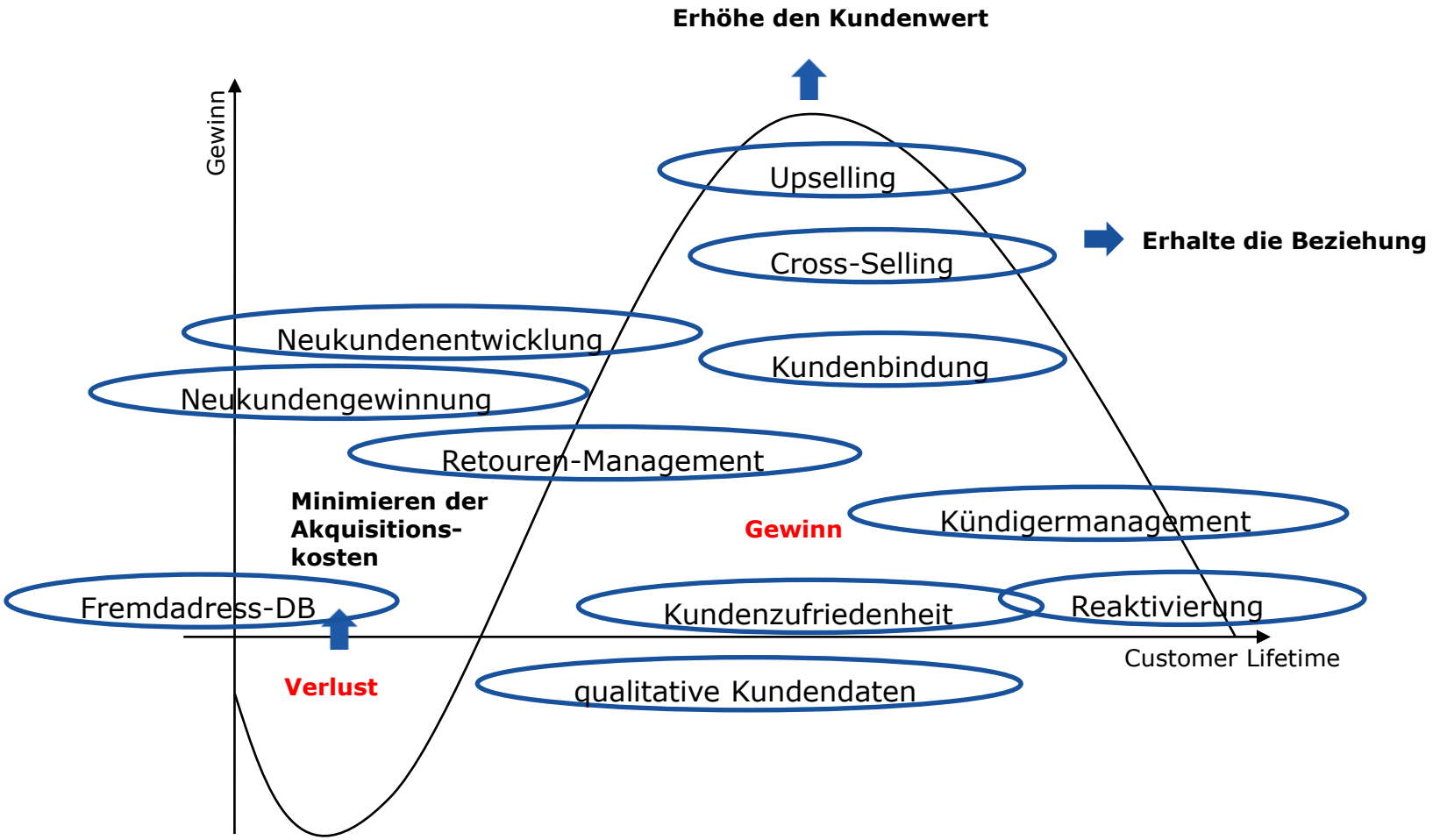
Branchenlösungen

	Handel	Banken/Vers.	Energie
Produktempfehlungen/Warenkorbanalyse	☺	☺	
Customer Targeting	☺		
Revisionsanalyse	☺		
Checkout-Couponing	☺		
Up- und Cross-Selling	☺	☺	
Mailingoptimierung/Kampagnenopt.	☺	☺	
Betrugserkennung/Fraud Detection	☺	☺	
Storno-/Kündigerprävention		☺	☺
Forderungsmanagment			☺

Telekommunikation, Gesundheitswesen, Touristik, Media etc.

Mit Data Mining ...

Kunden nachhaltig gewinnen und binden



verstehen – vorhersagen - agieren

Data Mining - Sucess Storys

- **BankFinancial:** 7x höhere Response Rate, 80 % Kostenreduktion bei Neukundenwerbung (bessere Vorauswahl affiner Kunden in Mailingaktion)
- **Softmap.com:** dreifache Gewinnsteigerung aus Online-Verkäufen durch Recommendation-Engine und Verdoppelung der Verweildauer auf der Webseite.
- **Verizon:** Deutliche Einsparungen durch geeignete Maßnahmen zur Kündigerprävention und Kundenbindung, 60 % reduzierte Marketingkosten.
- **Union Investment:** „…die Kundengruppen sind heute genau definiert, so dass die Mailings nicht an alle Kunden, sondern gezielt nach Lebensumständen und Anforderungen versendet werden. So konnte die Abschlussquote neuer Verträge bereits um 30 bis 100 Prozent im Vergleich zu Mailings ohne Data Mining Einfluss erhöht werden."
- ***Viseca Card Services SA verbessert ihr Risikomanagement:*** „Seit der Einführung des neuen Systems konnten wir die Anzahl der Neukunden, welche innerhalb der ersten 9 Monate in die Inkassoabteilung transferiert werden mussten, um 50% reduzieren".

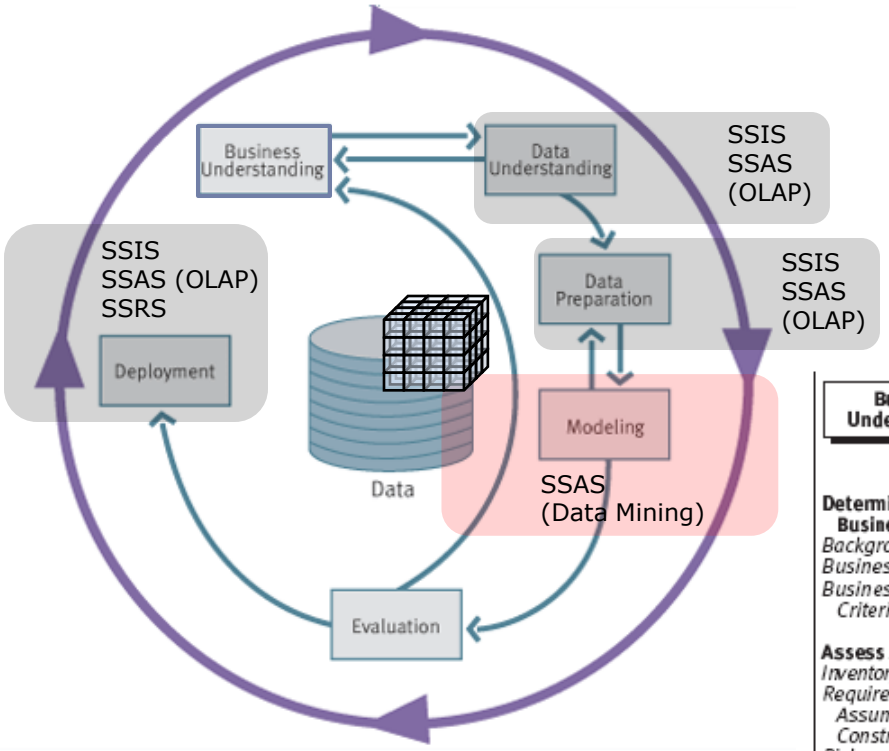
Data Mining mit MS SQL Server 2008

Schlüsselbegriffe der Datenanalyse

- **Variable** (Merkmal, Feld, Spalte)
  - + Diskrete Variable z. B. Alter, Geschlecht
  - + Stetige (kontinuierliche) Variable, z. B. Einkommen, Temperatur
- **Attribut** (*State, Value*)
  - + z. B. Beziehungsstatus: Attribute: verheiratet, ledig
- **Fall** (*Case*)
  - + z. B. Zeile in einer Tabelle mit Spalten, die Merkmale repräsentieren
  - + alle Zeilen, die zu der gleichen Transaktion gehören bilden einen Fall
- **Schlüssel** (*Keys*)
  - + case key (z. B. Primary Key)
  - + nested key
- **Eingabespalten und Ausgabespalten**

CRISP-DM – Prozessmodel

CRoss Industrie Standard Process for Data Mining



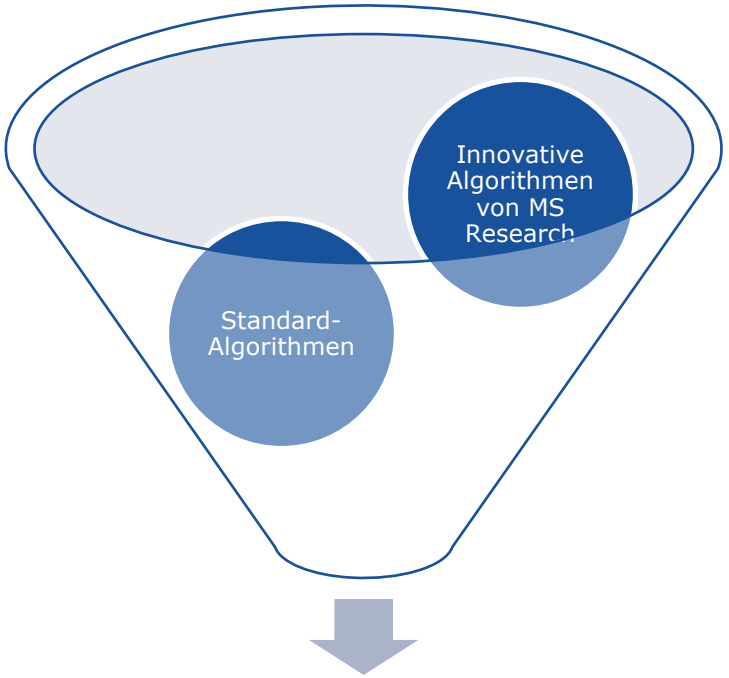
1. Business Understanding (10 %)
2. Data Understanding(20-30 %)
3. Data Preparation(50-80 %)
4. Modeling (10 %)
5. Evaluation (5 %)
6. Deployment (5-10 %)

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria* **Assess Situation** *Inventory of Resources* *Requirements, Assumptions, and Constraints* *Risks and Contingencies* *Terminology* *Costs and Benefits* **Determine Data Mining Goals** *Data Mining Goals* *Data Mining Success Criteria* **Produce Project Plan** *Project Plan* *Initial Assessment of Tools and Techniques*	**Collect Initial Data** *Initial Data Collection Report* **Describe Data** *Data Description Report* **Explore Data** *Data Exploration Report* **Verify Data Quality** *Data Quality Report*	**Select Data** *Rationale for Inclusion/ Exclusion* **Clean Data** *Data Cleaning Report* **Construct Data** *Derived Attributes* *Generated Records* **Integrate Data** *Merged Data* **Format Data** *Reformatted Data* *Dataset* *Dataset Description*	**Select Modeling Techniques** *Modeling Technique* *Modeling Assumptions* **Generate Test Design** *Test Design* **Build Model** *Parameter Settings* *Models* *Model Descriptions* **Assess Model** *Model Assessment* *Revised Parameter Settings*	**Evaluate Results** *Assessment of Data Mining Results w.r.t. Business Success Criteria* *Approved Models* **Review Process** *Review of Process* **Determine Next Steps** *List of Possible Actions* *Decision*	**Plan Deployment** *Deployment Plan* **Plan Monitoring and Maintenance** *Monitoring and Maintenance Plan* **Produce Final Report** *Final Report* *Final Presentation* **Review Project** *Experience Documentation*

Quelle: CRISP-DM CONSORTIUM (2000), www.crisp-dm.org

Algorithmen-Typen

zur Lösung typischer Geschäftsprobleme



Breites Spektrum von Möglichkeiten zur Erstellung optimaler Modelle

- Klassifikationsalgorithmen
- Regressionsalgorithmen
- Segmentierungsalgorithmen (Clustering)
- Assoziationsalgorithmen
- Sequenzanalysealgorithmen

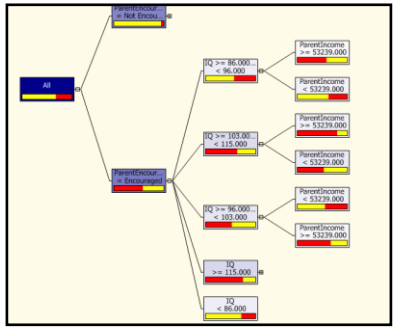
	Decision Trees	Naive Bayes	Clustering	Sequence Clustering	Time Series	Association	Neural Network
Klassifikation	👍	✌	✌	✌		✌	👍
Regression	👍	✌	✌	✌			👍
Segmentierung			👍	👍			✌
Assoziationsanalyse	👍	👍	✌	✌		👍	👍
Zeitreihenanalyse					👍		

Methodenauswahl abhängig von der Data Mining-Aufgabe und dem Datenmaterial

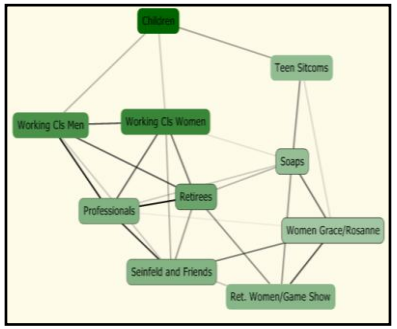
👍 = 1. Wahl

✌ = 2. Wahl

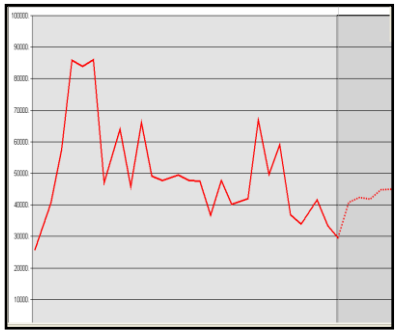
Data Mining Algorithmen (SSAS)



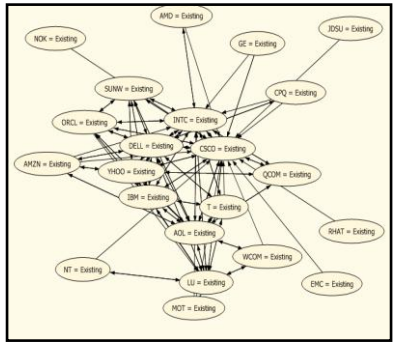
Decision Trees



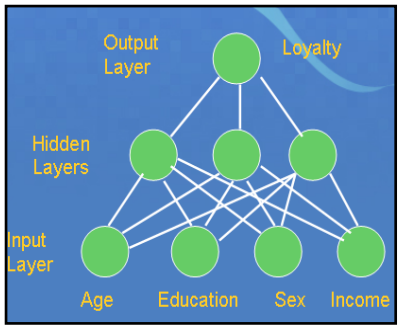
Clustering



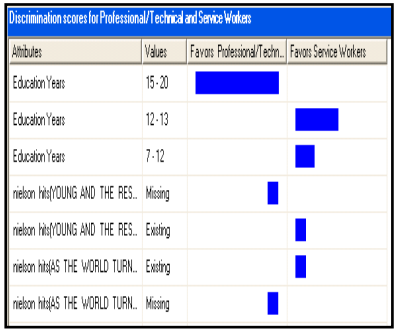
Time Series



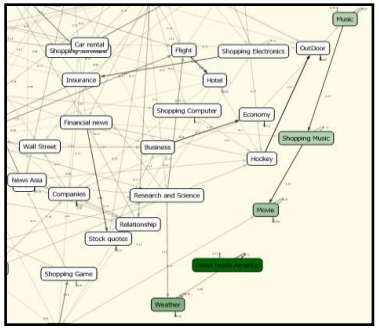
Association



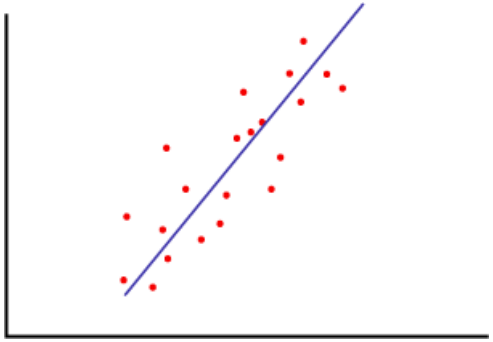
Neural Nets



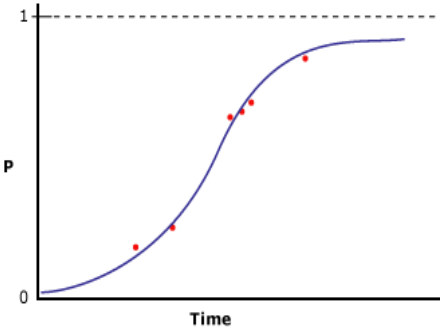
Naïve Bayes



Sequence Clustering



Linear Regression

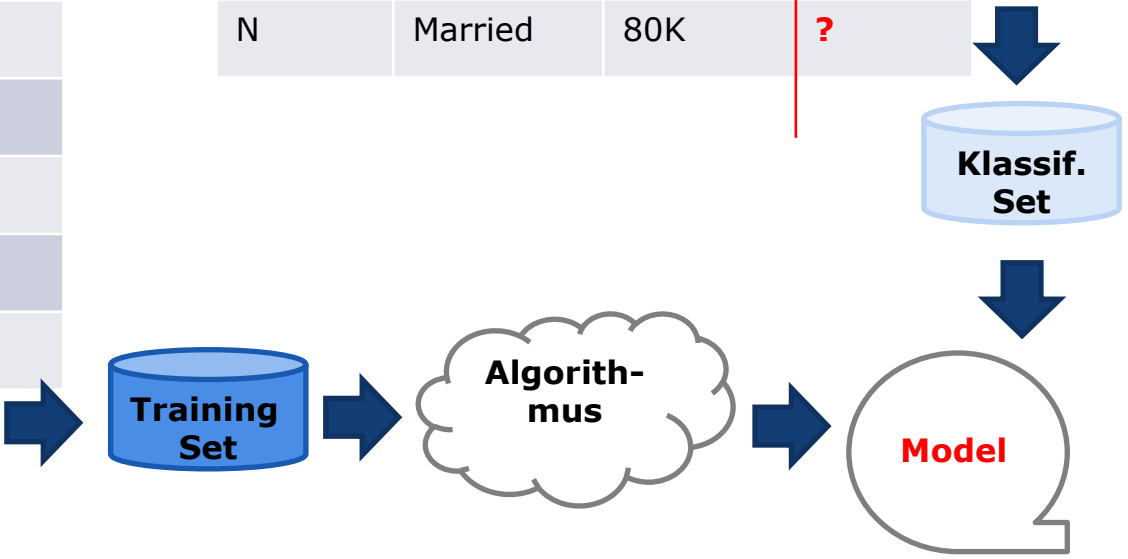


Logistic Regression

Klassifikation

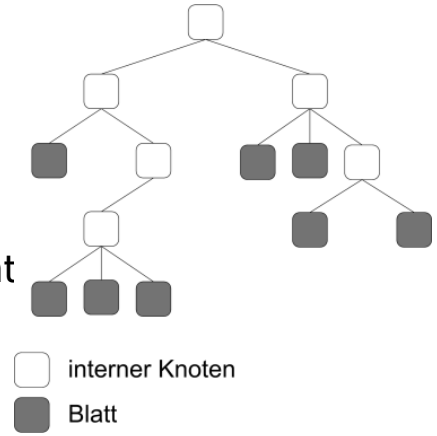
	discrete	discrete	continuous	class
Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Y	Single	125K	**N**
2	N	Married	100K	**N**
3	N	Single	70K	**N**
4	Y	Married	120K	**N**
5	N	Divorced	95K	**Y**
6	N	Married	60K	**N**
7	Y	Divorced	220K	**N**
8	N	Single	85K	**Y**
9	N	Married	75K	**N**
10	N	Single	90K	**Y**

Refund	Marital Status	Taxable Income	Cheat
Y	Single	75K	?
N	Married	50K	?
Y	Married	150K	?
Y	Dicorced	90K	?
N	Single	40K	?
N	Married	80K	?

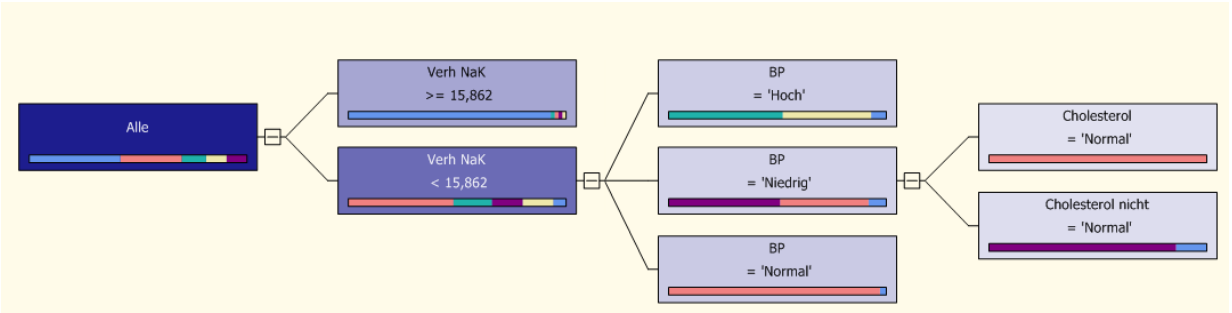


Microsoft Decision Trees

Microsoft Entscheidungsstrukturen



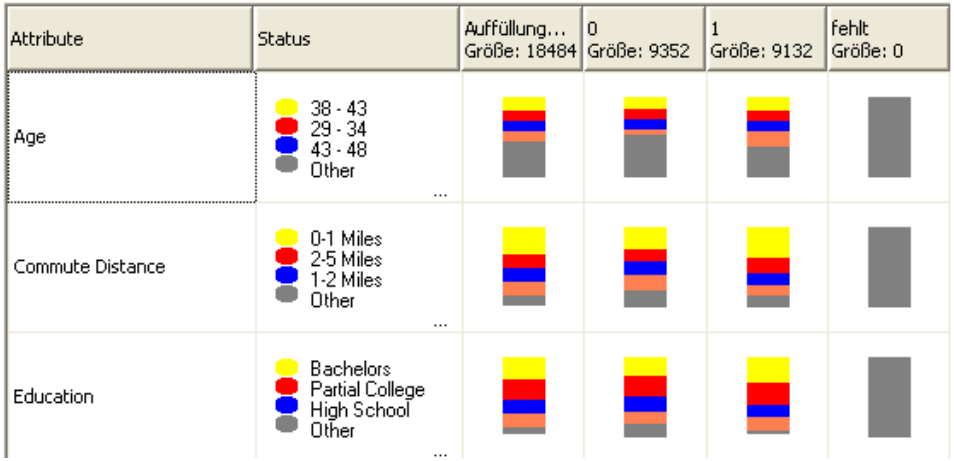
- Entscheidungsbaum: Ablaufdiagramm ähnlich einer Baumstruktur
- Aufgabe: Individuen *bekannten* Klassen zuordnen
  - + Bonitätsbeurteilung: Individuen: Bankkunden, Klassen: kreditwürdig, nicht kreditwürdig
- Top-Down-Ansatz: Attribut suchen, mit dem die Daten am besten klassifiziert werden können (höchsten Informationsgehalt)
- Hybrider Entscheidungsbaumalgorithmus: Klassifikations- und Regressionsalgorithmus
- beliebtes, häufig eingesetztes Verfahren, da Baumstruktur leicht zu interpretieren
- System zur Regelinduktion: *WENN* Bedingung *DANN* Folgerung



„Mit dem Entscheidungsbaumverfahren werden Gruppen nach dem Hierarchieprinzip gebildet und die jeweils für die Gruppenbildung relevanten Variablen identifiziert."

Microsoft Naive Bayes

- Klassifikationsalgorithmus für Vorhersagemodellierung
- stützt sich auf von Reverend Thomas Bayes im Jahre 1763 begründete Bayes Theorem für bedingte Wahrscheinlichkeiten
- Naive Annahme der Unabhängigkeit der Attribute
- Für jedes Attribut der Eingabespalten werden die Wahrscheinlichkeiten anhand jeder möglicher Ausprägungen der vorhersagbaren Spalte berechnet. Als Ergebnis wird die Klasse mit der höchsten Wahrscheinlichkeit zurückgegeben.
- Algorithmus weniger rechenintensiv als andere –> gut geeignet für erste Datenerkundung, um Beziehungen zwischen den Eingabe-Spalten und Vorhersage-Spalten aufzudecken



Microsoft Naive Bayes-Viewer

DMX

- DMX - Data Mining eXtensions
  - + SQL ähnliche Sprache für Erstellung von Abfragen für DM-Modelle
  - + DDL-Anweisungen (Data Definition Language)
    - Erstellen, Modifizieren, Löschen, Export/Import
  - + DML-Anweisungen (Data Manipulation Language)
    - Trainieren, Abfragen
  - + DM-Funktionen
    - Predict(), PredictProbability, CaseLikelihood, Cluster(), etc.
  - + User-defined functions, Parametrisierte Abfragen

1. Modell erstellen

```
CREATE MINING MODEL <model>
(
        <column definition>
) USING <algorithmus>
```

2. Modell trainieren

```
INSERT INTO <model>
(
   <model column list>
)  <data source query>
```

3. Modell abfragen

```
SELECT <select list>
FROM <model>
PREDICTION JOIN
<data source query> AS <alias>
ON <cloumn mapping>
```

```
<column definition>:= <name><data type> <content type>
```

SQL Server 2008 – Neuerungen

- **Verbesserung der Engine und der Algorithmen**
  - + Anforderungen von SQL-Server 2005 DM-Anwendern (offene Wünsche)
- **Verbesserung im Bereich Mining-Strukturen**
  - + Aufteilung in Trainings- und Test-Partitionen
  - + Abfragen gegen Struktur-Cases und Struktur-Spalten
    - Ermöglicht Drillthrough aus Cluster-Modell um zusätzlich Daten anzuzeigen, die nicht im Modell benutzt werden sollen (z.B. eine Mail-Adresse)
  - + Filterung von Daten beim Aufbau von Mining-Modellen
    - Bsp.: Erstelle getrennte Modelle für männliche und weibliche Kunden
  - + Kreuz-Validierung (k-Fold Cross-Validation)
    - Erleichtert Verstehen der Modell-Genauigkeit
    - Automatischer Test des Modells gegen mehrere Subsets von Trainingsdaten und Vergleich der Ergebnisse
- **Verbesserungen im Bereich Zeitreihen (Time Series)**
  - + Zusätzlich zum ARTxp-Algorithmus verfügbar: ARIMA
  - + Der bekannteste und verbreiteste Zeitreihen-Algorithmus
  - + Akzeptable Vorhersagen bei Projektion auf mehr als 10 Schritte

Quelle: Vach, M.: Data Mining leicht gemacht, Microsoft 2008.

DEMO

- Die Marketingabteilung von Adventure Works verfügt über eine Liste potenzieller Neukunden, die sie mit einer gezielten Marketing-Kampagne anschreiben und zum Kauf eines Fahrrades anregen möchte.
- Zur Reduzierung von Kosten sollen die Flyer/Kataloge nur an jene Kunden gesendet werden, die mit höherer Wahrscheinlichkeit auf diese reagieren
- Das Unternehmen speichert die Informationen in einer Datenbank mit demographischen Daten und Reaktionen auf vorherige Mailingaktionen
- Muster erkennen, die auf einen potenziellen Fahrradkäufer schließen lassen durch den Vergleich mit Kunden die ähnliche Merkmale aufweisen und bereits in der Vergangenheit Fahrräder des Unternehmens gekauft haben.

Data Mining mit Office 2007

Data Mining an jedem Arbeitsplatz

- Kostenloses add-in Package für Office Excel 2007, Office Visio 2007
- 3 add-ins

Tabellenanalysetool für Excel 2007

Data Mining Client für Excel 2007

Data Mining Vorlagen für Visio 2007

Quelle: www.sqlserverdatamining.com

- Voraussetzungen
  - + Windows Server 2003 SP2; Windows Server 2008; Windows Vista SP1; Windows XP SP3
  - + .NET Framework 2.0
  - + Microsoft Office 2007 mit .NET-Programmierunterstützug
    - Professional, Professional Plus, Ultimate, Enterprise
  - + Verbindung mit SQL Server Analysis Services 2005/2008
    - Enterprise oder Standard

Zusammenfassung

- Data Mining ist das halbautomatische Extrahieren von Mustern aus großen Datenbeständen, um sie dem Anwender als interessantes Wissen zu präsentieren
- Zu den Aufgaben von DM zählen die Segmentierung (Clustering), Klassifikation, Abhängigkeitsanalyse, Prognose
- Data Mining adressiert viele Geschäftsprobleme in nahezu alle Branchen und hilft bei der Neukundengewinnen, Reduzierung von Marketingkosten, Kundenbindung, Gewinnsteigerung durch Recommendation Engines, uvm.
- Mit MS SQL Server ist Data Mining für jeden zugänglich, vollständig, integriert und erweiterbar
- 9 Algorithmen
- Data Mining mit MS Office 2007 möglich: Anwender und Entwickler nutzen gewohnte Umgebung

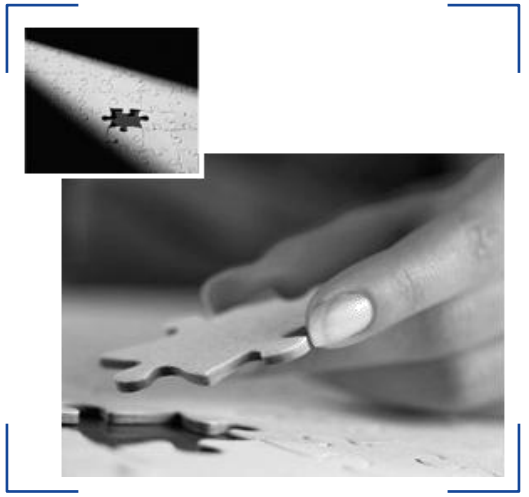
Literatur und Infos zu Data Mining

- **Literatur:**
- HIPPNER (2001): Hippner, Hajo (Hrsg.) et al.: Handbuch Data Mining im Marketing. Braunschweig/Wiesbaden.
- AZEVEDO/BROSIUS/DEHNERT/NEUMANN/SCHEERER (2006): Azevedo, Pedro; Brosius, Gerhard; Dehnert, Stefan; Neumann, Berthold; Scheerer, Benjamin: Business Intelligence und Reporting mit Microsoft SQL Server 2005. Unterschleißheim.
- BROSIUS/SCHEERER/WOLF (2008): Brosius, Gerhard; Scheerer, Benjamin; Wolff, Ulrich: Business Intelligence mit Office 2007 und SQL Server. Unterschleißheim.
- TANG/MACLENNAN (2008): Tang, ZhaoHui; MacLennan, Jamie: Data Mining with SQL Server 2008. Indianapolis, Indiana.
- **Online Quellen:**
- SQL Server data mining: http://www.sqlserverdatamining.com/
- Jamie‘ Junk: http://blogs.msdn.com/jamiemac/
- Kdnuggets Portal: http://www.kdnuggets.com/
- **Videos:**
- DM mit Excel 2007: http://msdn.microsoft.com/en-us/library/dd299410.aspx
- DM mit Excel 2007: http://www.sqlservercentral.com/articles/Video/65057/
- http://www.microsoft.com/emea/spotlight/sessionh.aspx?videoid=867

Ihr Zugang zum Plus!

VIELEN DANK FÜR DIE AUFMERKSAMKEIT!

FRAGEN ?



plus-IT GmbH
Lina-Ammon-Str. 3
D-90471 Nürnberg

Fon +49 911 8176678-0
Fax +49 911 8176678-7
E-mail nuernberg@plus-it.de
www.plus-it.de

Ihr Ansprechpartner

Michael Deinhard
- Niederlassungsleiter Nürnberg –

Lina-Ammon-Str. 3
D-90471 Nürnberg
Tel.: 0911 / 81 766 780
0151 57 116 116

michael.deinhard@plus-it.de
www.plus-it.de