Der CRISP-DM Prozess für Data Mining

Wozu einen standardisierten Prozess?

Der Prozess der Wissensentdeckung muss verlässlich und reproduzierbar sein – auch für Menschen mit geringem Data Mining Hintergrundwissen.

- Rahmen für “Speicher-” und “Wiedereinsetzpunkte”
- Hilft bei der Planung und der Verwaltung der Analyseaufgaben
- Leichterer Einstieg

CRISP-DM Standard

- CRoss Industry Standard Process for Data Mining
- Initiative begann September 1996
- Gefördert durch die EU
- Mehr als 200 Mitglieder der CRISP-DM SIG weltweit
  - DM Anbieter - SPSS, NCR, IBM, SAS, SGI, Data Distilleries, Syllogic, Magnify…
  - Berater - Cap Gemini, ICL Retail, Deloitte & Touche…
  - Anwender - BT, ABB, Lloyds Bank, AirTouch, Experian...

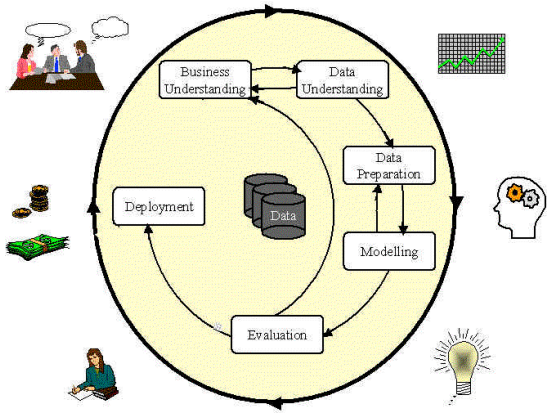
CRISP-DM

- Nicht-proprietär
- Neutral bezüglich
  - Anwendungsfeldern
  - Applikationen
- Als Leitlinie zu verstehen
- Erfahrungssammlung:
  - Analysetemplates

Phasen und Aufgaben

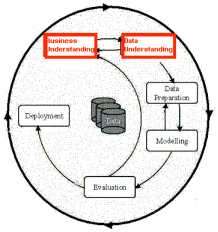
Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria* **Situation Assessment** *Inventory of Resources* *Requirements, Assumptions, and Constraints* *Risks and Contingencies* *Terminology* *Costs and Benefits* **Determine Data Mining Goal** *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*	*Data Set* *Data Set Description* **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*	**Select Modeling Technique** *Modeling Technique* *Modeling Assumptions* **Generate Test Design** *Test Design* **Build Model** *Parameter Settings* *Models* *Model Description* **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*

Phasen im DM – Prozess



Phasen im DM – Prozess (1 & 2)

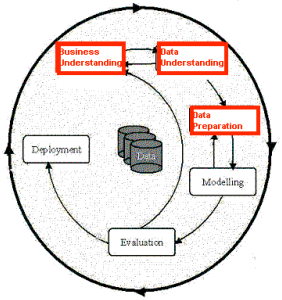
- Business Understanding:
  - Aufstellung des (Geschäfts-)ziels
  - Aufstellung des Data Mining Ziels
  - Aufstellung von Erfolgskriterien



- Data Understanding
  - Untersuchung der Daten und Prüfung der Qualität
  - Outlier entdecken

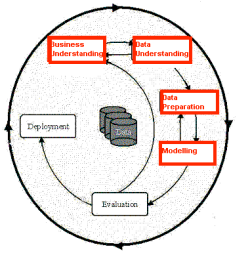
Phasen im DM – Prozess (3)

- Data Preparation:
  - Benötigt häufig bis zu 90% der Gesamtzeit
    - Datensammlung
    - Beurteilung
    - Konsolidierung und Säuberung
      - Tabellenlinks, Aggregationstiefe, fehlende Werte, …
    - Datenauswahl
      - Ausreißer?
      - Sampling?
      - Welche Variablen?
    - Transformationen – Erzeugung neuer Variablen



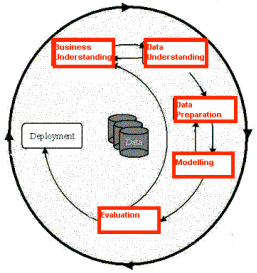
Phasen im DM – Prozess (4)

- Model Building
  (Data Mining Schritt)
  - Auswahl der Modellierungstechnik(en) wird auf Basis des erstellten Data Mining Ziels und der Daten getroffen
  - Modellbildung ist ein iterativer Prozess – welcher sich für überwachtes und unüberwachtes Lernen unterscheidet
    - Modelle können beschreibend oder vorhersagend sein
- Vielzahl von Verfahren während der Vorlesung



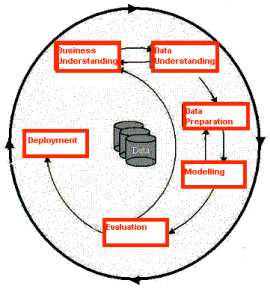
Phasen im DM – Prozess (5)

- Modell Evaluation
  - Evaluation des Modells: wie gut arbeitet es auf Testdaten?
  - Evaluationsmethoden und –kriterien hängen vom Modelltypen ab:
    - beispielsweise eine Fehlermatrix für Klassifikationsmodelle oder der mittlere quadratische Fehler für Regressionsmodelle
  - Interpretation des Modells:
    ob wichtig oder nicht, ob einfach oder schwer hängt erneut von den Zielen und den Modelltypen ab



Phasen im DM – Prozess (6)

- Deployment
  - Bestimmung, wie die Ergebnisse genutzt werden können
  - Wer wird sie nutzen?
  - Wie oft werden sie genutzt?
- Anwendung der Ergebnisse durch
  - Vorhersagen auf Datenbanken
  - Anwendung als (Geschäfts-)regeln
  - interaktive on-line Vorhersagen

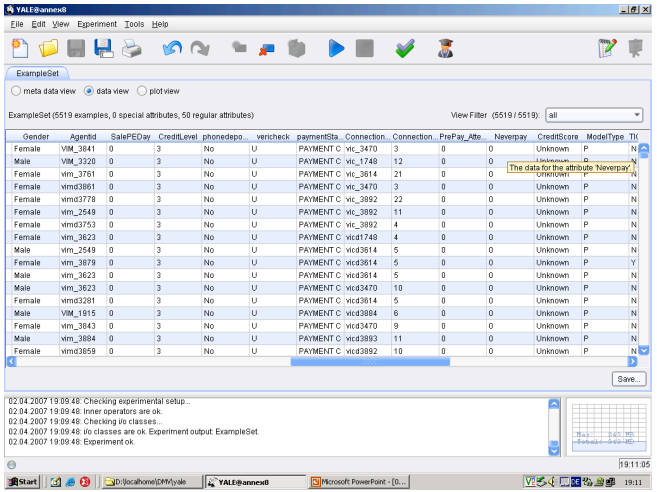


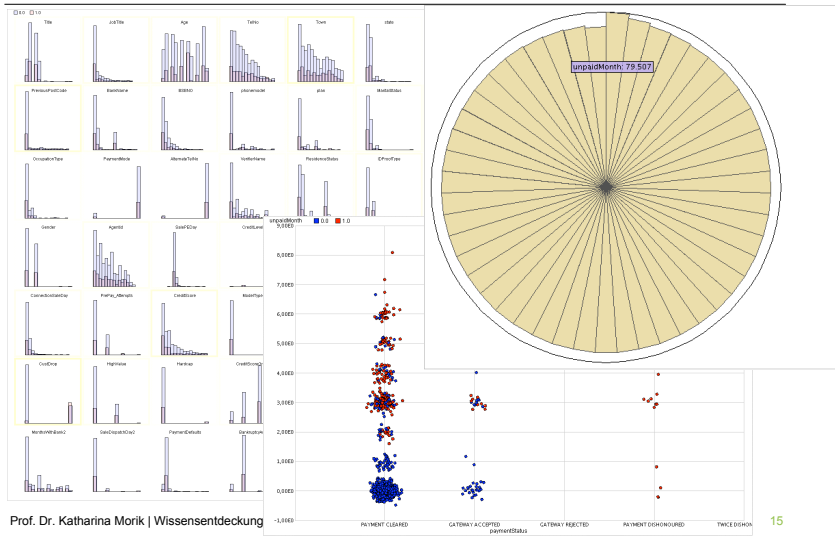
Beispiel: Analyse von Kundendaten

- Gegeben: Kundendaten eines Telefonanbieters
- Aufgabe: Bestimmung der Faktoren, welche Kunden zu “schlechten” Kunden machen, d.h. solchen Kunden mit schlechter Zahlungsmoral
- Screenshots etc. mit Hilfe von RapidMiner – Experimenten erstellt

Business & Data Understanding (1 & 2)

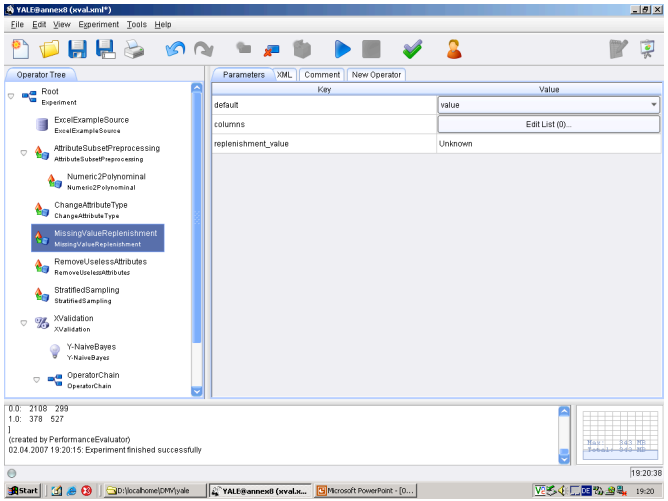
- Geschäftsziel: Reduktion der Zahl der Kunden, welche niemals zahlen werden und damit Erhöhung des Gewinns
- Erfolgskriterium: Reduktion dieser Anzahl um mind. 30%
- Sammlung der Daten bereits erfolgt:
  - Exceltabelle: jede Zeile ein Kunde, jede Spalte eine Eigenschaft (Merkmal, Attribut oder Variable)
  - Merkmale beschreiben Kunden und deren Verhalten
  - Die Spalte “Neverpay” soll vorhergesagt werden





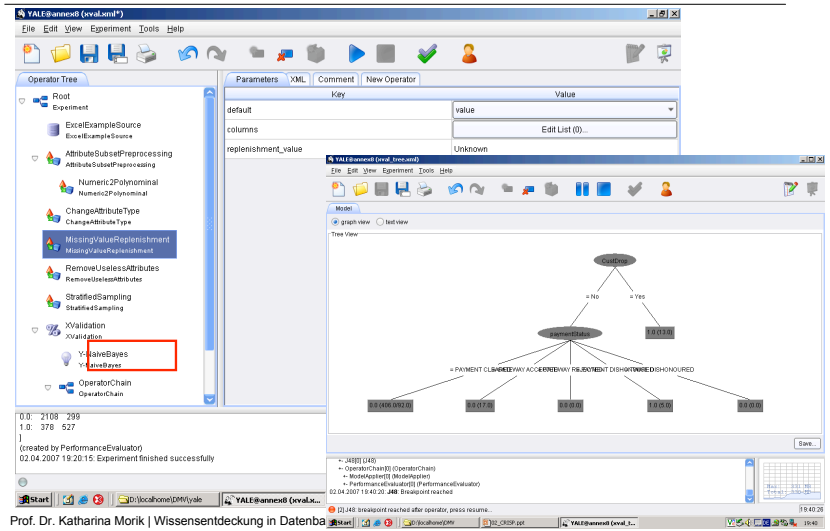
Data Preprocessing (3)

- Glück: Die Daten liegen bereits in 1 (!) Tabelle vor. Hauptteil der Arbeit also schon erledigt!
- Ansonsten: ETL (z.B. MiningMart)
- Noch nötig:
  - Umwandlung des numerischen Zielattributs Neverpay in ein nominales Klassifikationsattribut
  - Definition des Zielattributs Neverpay
  - Ersetzung fehlender Werte durch “Unknown”
  - Löschen sinnloser Merkmale
  - Ziehen eines stratifizierten Samples



Model Building (4)

- Lernen unterschiedlicher Modelle, hier
  - Naïve Bayes
  - Entscheidungsbaumlerner
  - SVM
  - …
- Hierzu ist nur der Austausch eines einzigen Operators nötig.



Model Evaluation (5)

- Evaluation der verschiedenen Modelle (Vorverarbeitungen) soll die Zahl der falschen Vorhersagen abschätzen…
- … in einem “realen” Anwendungsfeld
- Trick: nur Teil der Daten zum Lernen verwenden, den Rest für die Fehlermessung

Fehlerreduktion um mehr als 50%: reicht!

classification_error: 10.76% +/- 1.50% (mikro: 10.76%)			
	true 0.0	true 1.0	class precision
pred. 0.0	3756	207	94.78%
pred. 1.0	387	1169	75.13%
class recall	90.66%	84.96%	

Deployment (6)

- Gelernte Modelle und Vorverarbeitungsketten können in Systeme / Anwendungen integriert werden
- Analyse der Modelle gibt einen Einblick in das Verhalten von Kunden
- Modelle können nachtrainiert werden bei signifikanter Erweiterung der Datenlage

Was wissen Sie jetzt?

- Der CRISP-Prozess ist ein standardisierter Prozess zur Durchführung von Data Mining
- Zunächst sollten Erfolgskriterien definiert werden
- Data Preprocessing verursacht üblicherweise den größten Aufwand
- Evaluation liefert Abschätzung für den Realfall
- Schritte sollten mehrfach besucht werden bei neuen Erkenntnissen