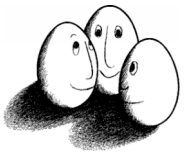
technische universität
dortmund

Fakultät für Informatik
LS 8

Der CRISP-DM Prozess für Data Mining

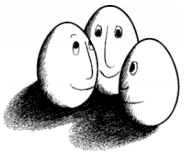
Prof. Dr. Katharina Morik



Wozu einen standardisierten Prozess?

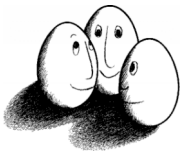
Der Prozess der Wissensentdeckung muss verlässlich und reproduzierbar sein – auch für Menschen mit geringem Data Mining Hintergrundwissen.

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



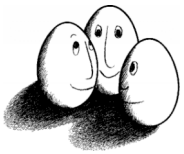
CRISP-DM Standard

- 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

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 Asessment of Tools and Techniques

Data Understanding

Collect Initial Data
Initial Data Collection Report

Describe Data
Data Description Report

Explore Data
Data Exploration Report

Verify Data Quality
Data Quality Report

Data Preparation

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

Modeling

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

Evaluation

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

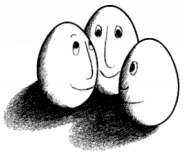
Deployment

Plan Deployment
Deployment Plan

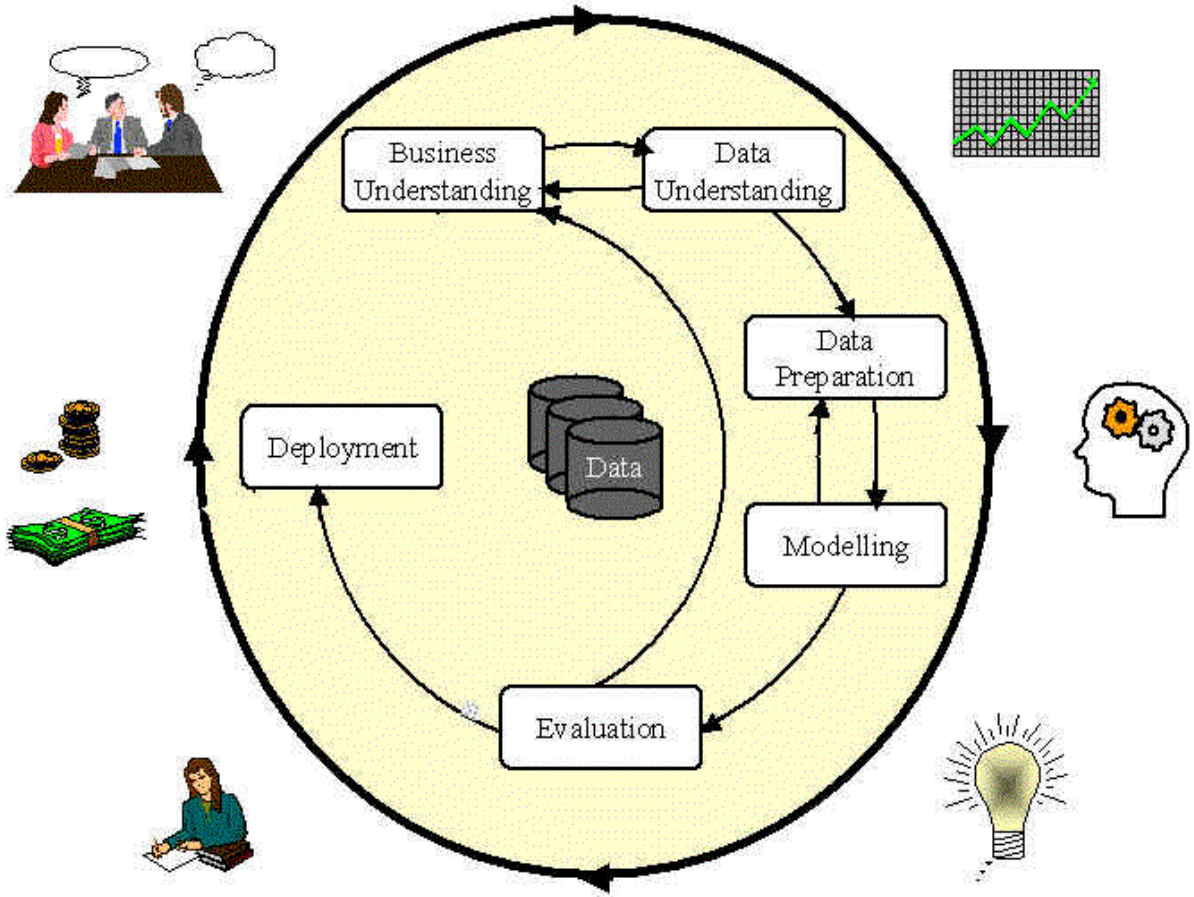
Plan Monitoring and Maintenance
Monitoring and Maintenance Plan

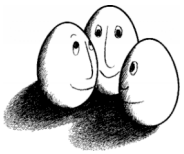
Produce Final Report
Final Report
Final Presentation

Review Project
Experience Documentation



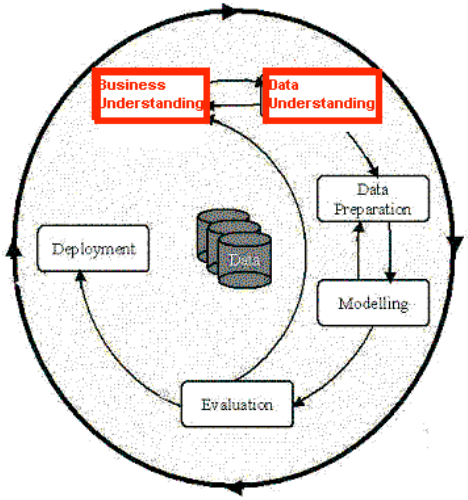
Phasen im DM – Prozess



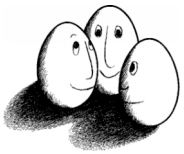


Phasen im DM – Prozess (1 & 2)

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

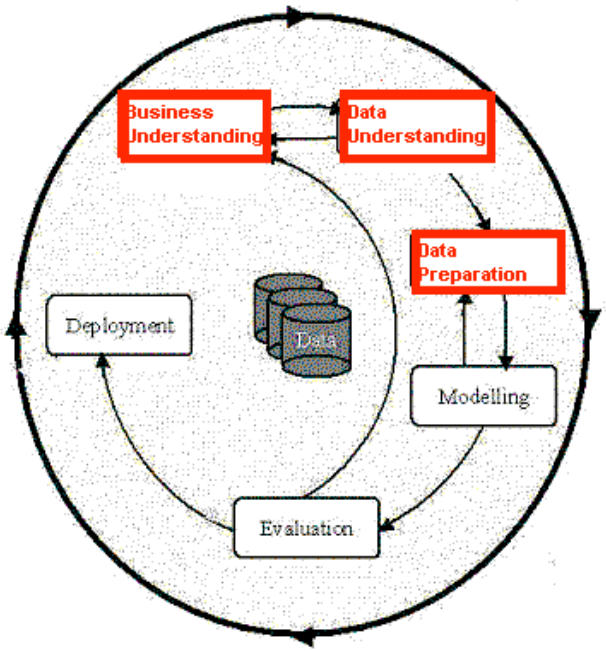


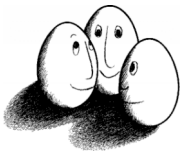
- Data Understanding
  - Untersuchung der Daten und Prüfung der Qualität
  - Outlier entdecken



Phasen im DM – Prozess (3)

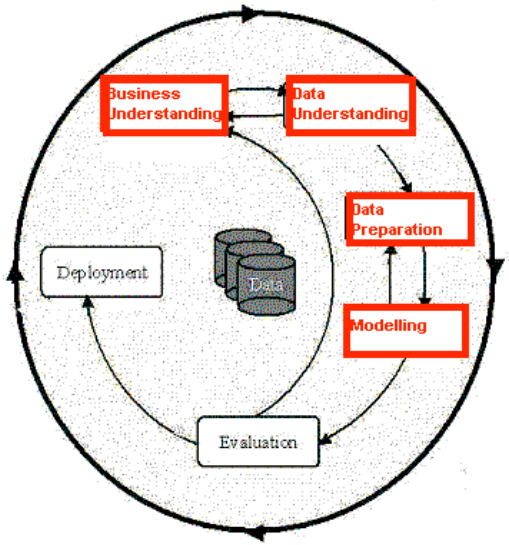
- Data Preparation:
  - Benötigt häufig bis zu 90% der Gesamtzeit
    - Datensammlung
    - Beurteilung
    - Konsolidierung und Säuberung
      - Tabellenlinks, Aggregationstiefe, fehlende Werte, …
    - Datenauswahl
      - Ausreißer?
      - Sampling?
      - Welche Variablen?
    - Transformationen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

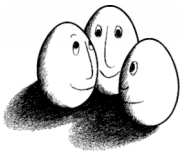




Phasen im DM – Prozess (4)

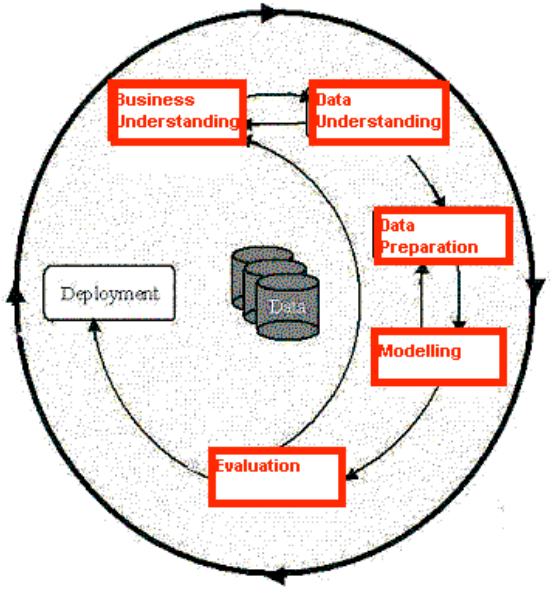
- Model Building
(Data Mining Schritt)
  - Auswahl der Modellierungstechnik(en) wird auf Basis des erstellten Data Mining Ziels und der Daten getroffen
  - Modellbildung ist ein iterativer Prozess – welcher sich für überwachtes und unüberwachtes Lernen unterscheidet
    - Modelle können beschreibend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

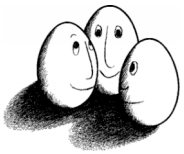




Phasen im DM – Prozess (5)

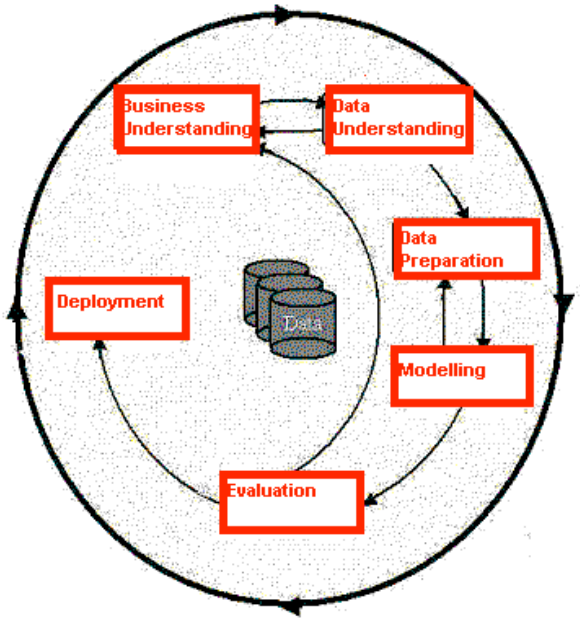
- Modell Evaluation
  - Evaluation des Modells: wie gut arbeitet es auf Testdaten?
  - Evaluationsmethoden und –kriterien hängen vom Modelltypen ab:
    - beispielsweise eine Fehlermatrix für Klassifikationsmodelle oder der mittlere quadratische Fehler für Regressionsmodelle
  - Interpretation des Modells: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

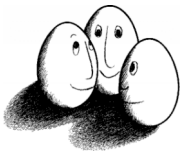




Phasen im DM – Prozess (6)

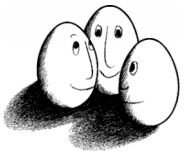
- Deployment
  - Bestimmung, wie die Ergebnisse genutzt werden können
  - Wer wird sie nutzen?
  - Wie oft werden sie genutzt?
- Anwendung der Ergebnisse durch
  - Vorhersagen auf Datenbanken
  - Anwendung als (Geschäfts-)regeln
  - interaktive on-line Vorhersagen





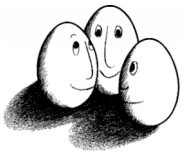
Beispiel: Analyse von Kundendaten

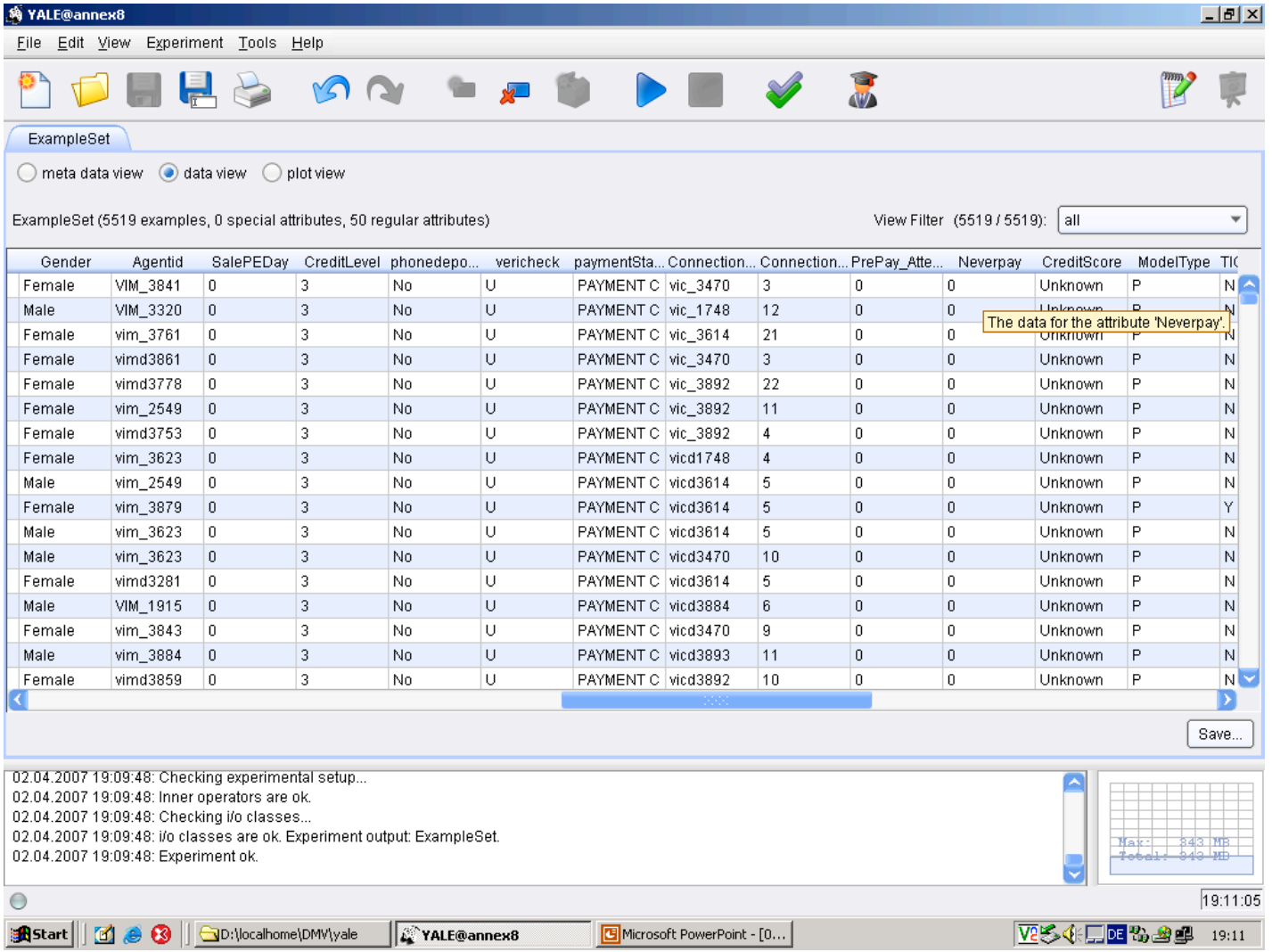
- Gegeben: Kundendaten eines Telefonanbieters
- Aufgabe: Bestimmung der Faktoren, welche Kunden zu “schlechten” Kunden machen, d.h. solchen Kunden mit schlechter Zahlungsmoral
- Screenshots etc. mit Hilfe von RapidMiner – Experimenten erstellt

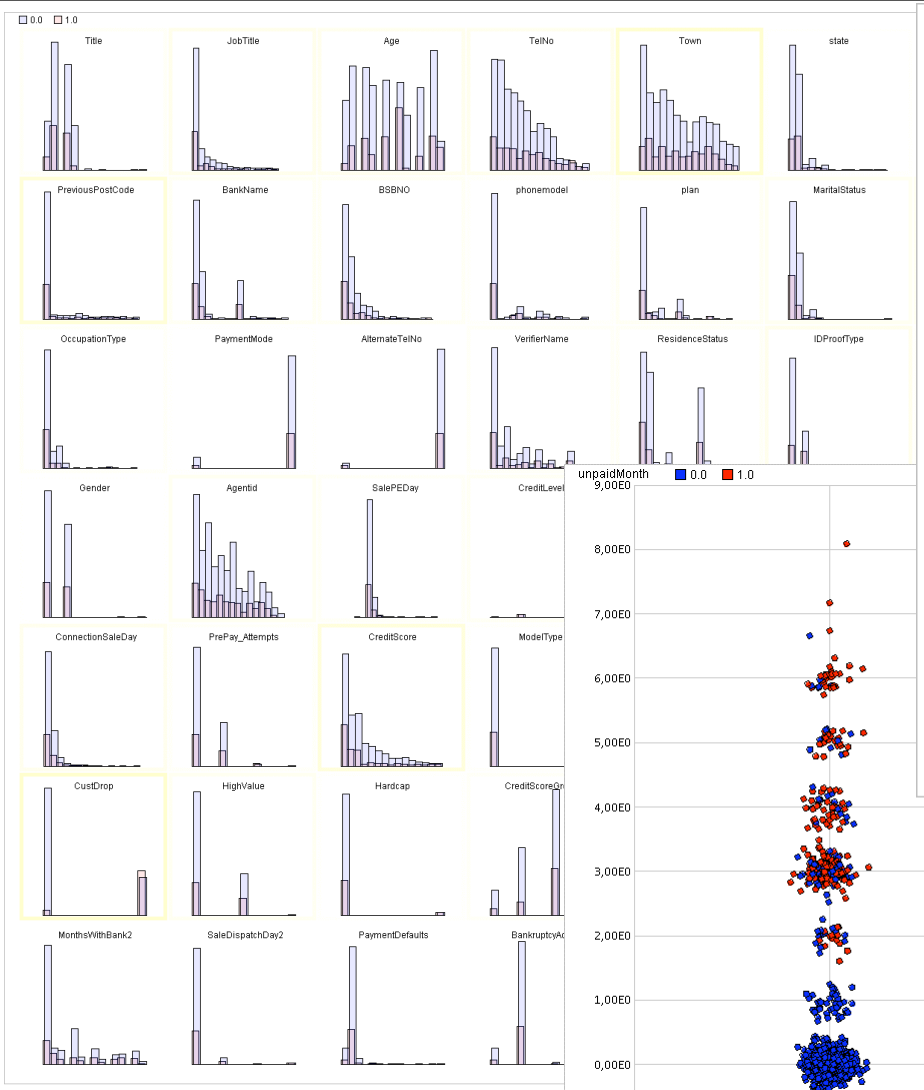


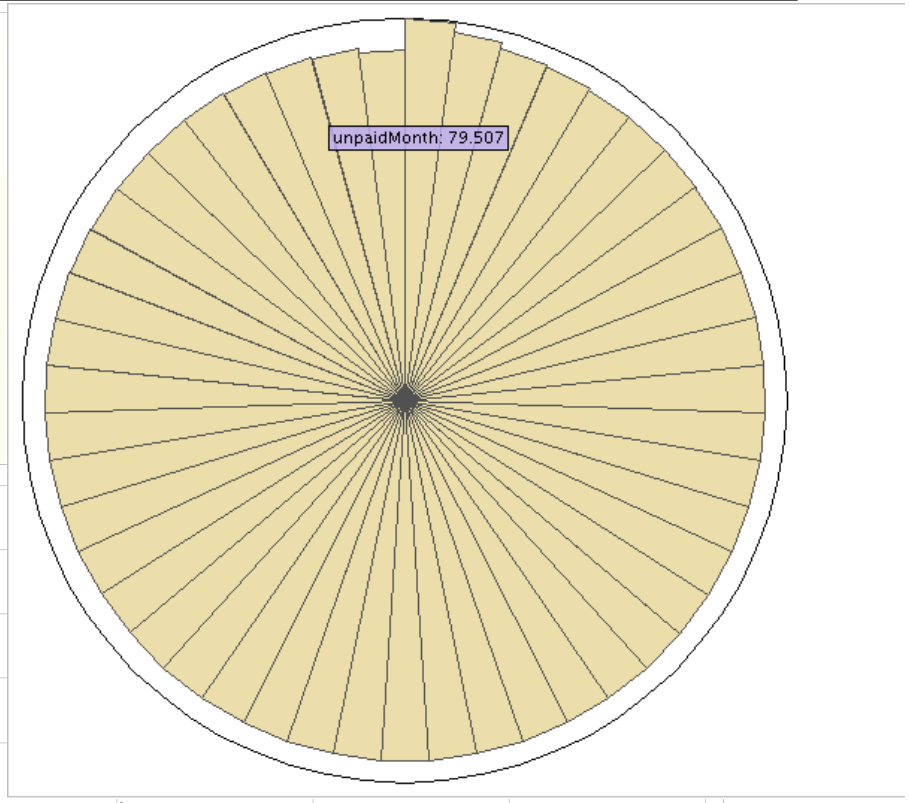
Business & Data Understanding (1 & 2)

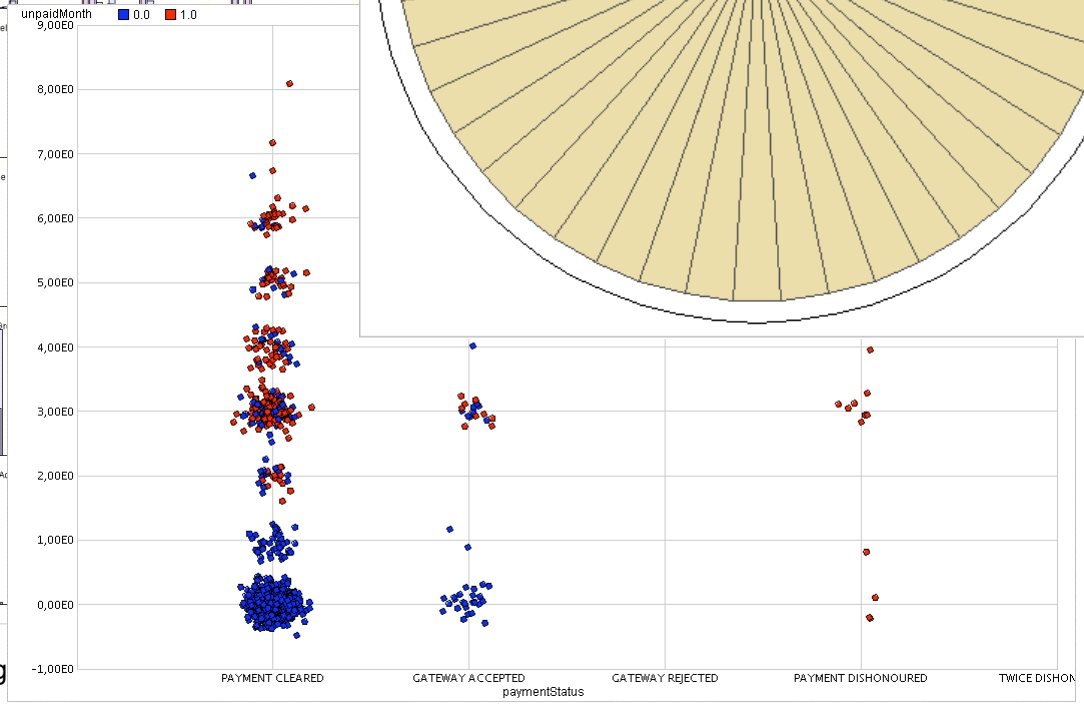
- Geschäftsziel: Reduktion der Zahl der Kunden, welche niemals zahlen werden und damit Erhöhung des Gewinns
- Erfolgskriterium: Reduktion dieser Anzahl um mind. 30%
- Sammlung der Daten bereits erfolgt:
  - Exceltabelle: jede Zeile ein Kunde, jede Spalte eine Eigenschaft (Merkmal, Attribut oder Variable)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

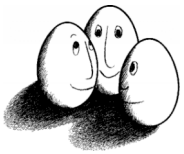



YALE@annex8
File Edit View Experiment Tools Help
ExampleSet
meta data view
data view
plot view
ExampleSet (5519 examples, 0 special attributes, 50 regular attributes)
View Filter (5519 / 5519): all
Gender Agentid SalePEDay CreditLevel phonedepo... vericheck paymentSta... Connection... Connection... PrePay_Atte... Neverpay CreditScore ModelType TI
Female VIM_3841 0 3 No U PAYMENT C vic_3470 3 0 0 Unknown P N
Male VIM_3320 0 3 No U PAYMENT C vic_1748 12 0 0 Unknown P N
Female vim_3761 0 3 No U PAYMENT C vic_3614 21 0 0 Unknown P N
Female vimd3861 0 3 No U PAYMENT C vic_3470 3 0 0 Unknown P N
Female vimd3778 0 3 No U PAYMENT C vic_3892 22 0 0 Unknown P N
Female vim_2549 0 3 No U PAYMENT C vic_3892 11 0 0 Unknown P N
Female vimd3753 0 3 No U PAYMENT C vic_3892 4 0 0 Unknown P N
Female vim_3623 0 3 No U PAYMENT C vicd1748 4 0 0 Unknown P N
Male vim_2549 0 3 No U PAYMENT C vicd3614 5 0 0 Unknown P N
Female vim_3879 0 3 No U PAYMENT C vicd3614 5 0 0 Unknown P Y
Male vim_3623 0 3 No U PAYMENT C vicd3614 5 0 0 Unknown P N
Male vim_3623 0 3 No U PAYMENT C vicd3470 10 0 0 Unknown P N
Female vimd3281 0 3 No U PAYMENT C vicd3614 5 0 0 Unknown P N
Male VIM_1915 0 3 No U PAYMENT C vicd3884 6 0 0 Unknown P N
Female vim_3843 0 3 No U PAYMENT C vicd3470 9 0 0 Unknown P N
Male vim_3884 0 3 No U PAYMENT C vicd3893 11 0 0 Unknown P N
Female vimd3859 0 3 No U PAYMENT C vicd3892 10 0 0 Unknown P N
The data for the attribute 'Neverpay'.
Save...
02.04.2007 19:09:48: Checking experimental setup...
02.04.2007 19:09:48: Inner operators are ok.
02.04.2007 19:09:48: Checking i/o classes...
02.04.2007 19:09:48: i/o classes are ok. Experiment output: ExampleSet.
02.04.2007 19:09:48: Experiment ok.
Max: 343 MB
Total: 343 MB
19:11:05
Start
D:\localhome\DMW\yale
YALE@annex8
Microsoft PowerPoint - [0...
19:11


0.0
1.0
Title
JobTitle
Age
TelNo
Town
state
PreviousPostCode
BankName
BSBNO
phonemodel
plan
MaritalStatus
OccupationType
PaymentMode
AlternateTelNo
VerifierName
ResidenceStatus
IDProofType
Gender
Agentid
SalePEDay
ConnectionSaleDay
PrePay_Attempts
CreditScore
ModelType
CustDrop
HighValue
Hardcap
MonthsWithBank2
SaleDispatchDay2
PaymentDefaults


unpaidMonth: 79.507

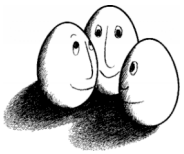

unpaidMonth
0.0
1.0
9,00E0
8,00E0
7,00E0
6,00E0
5,00E0
4,00E0
3,00E0
2,00E0
1,00E0
0,00E0
-1,00E0
PAYMENT CLEARED
GATEWAY ACCEPTED
GATEWAY REJECTED
PAYMENT DISHONOURED
paymentStatus

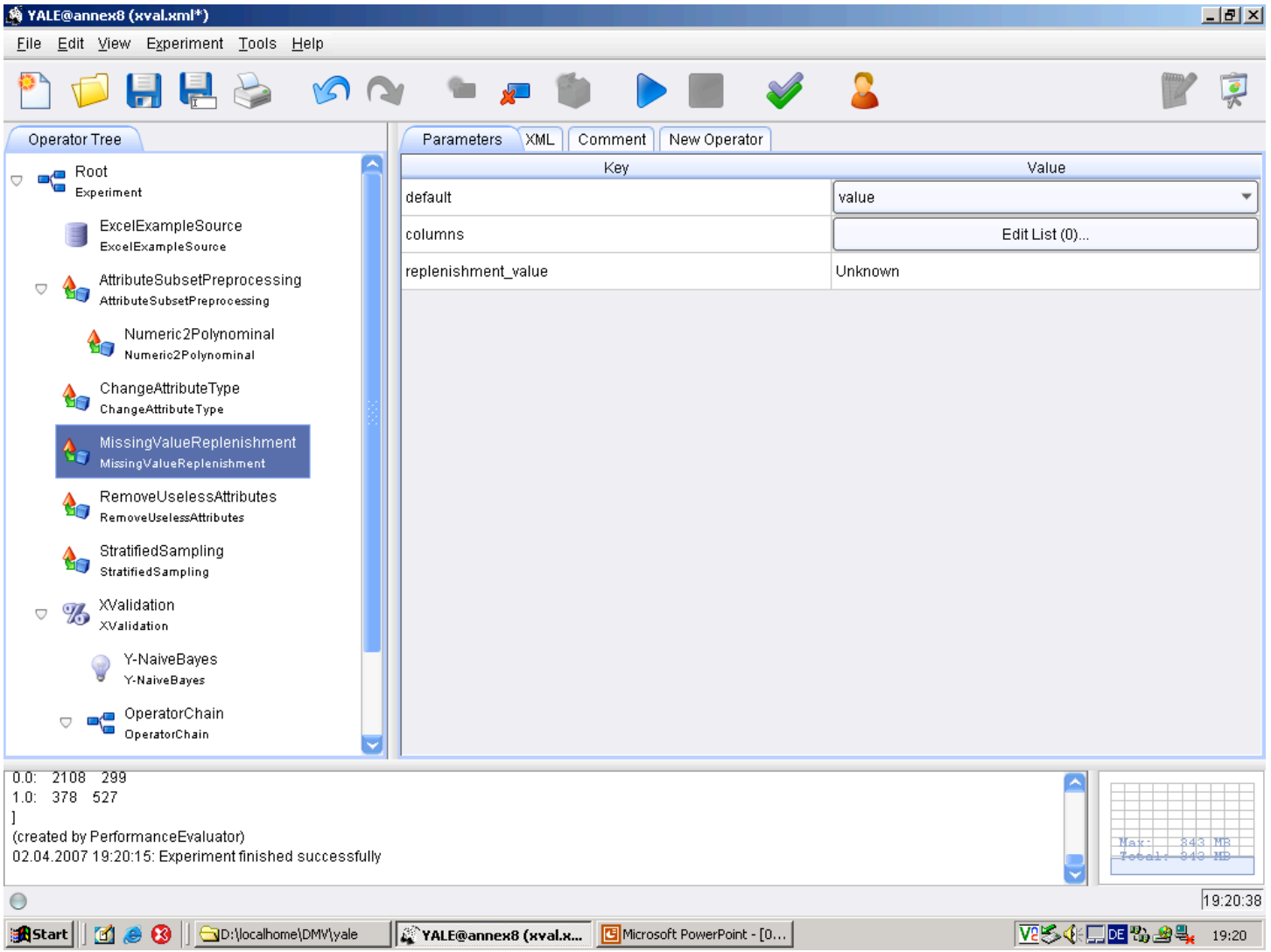


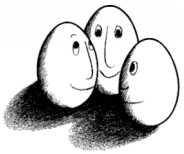
Data Preprocessing (3)

- 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

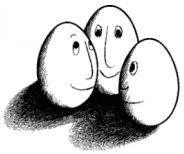


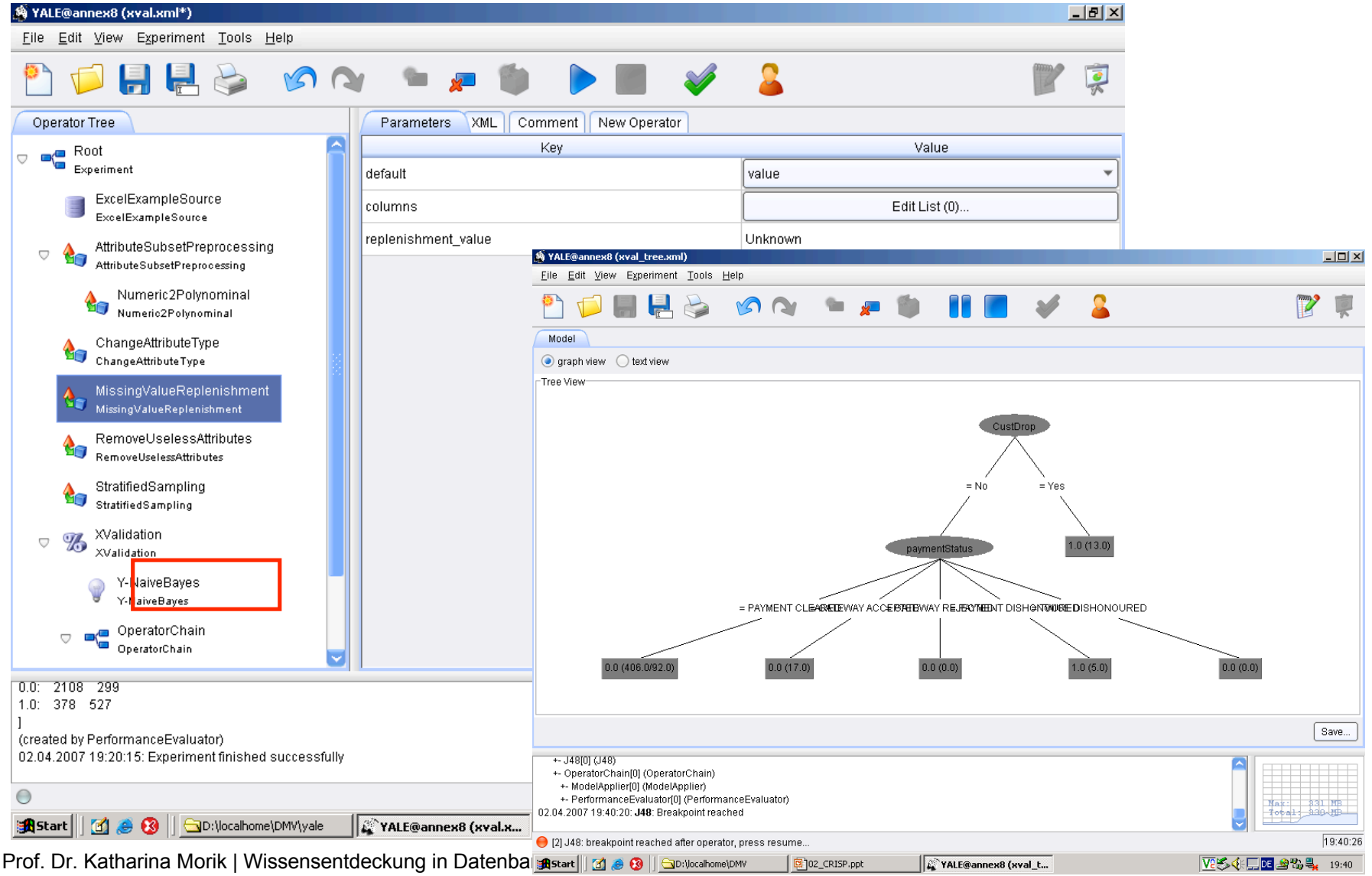

YALE@annex8 (xval.xml*)
File Edit View Experiment Tools Help
Operator Tree
Root
Experiment
ExcelExampleSource
ExcelExampleSource
AttributeSubsetPreprocessing
AttributeSubsetPreprocessing
Numeric2Polynominal
Numeric2Polynominal
ChangeAttributeType
ChangeAttributeType
MissingValueReplenishment
MissingValueReplenishment
RemoveUselessAttributes
RemoveUselessAttributes
StratifiedSampling
StratifiedSampling
XValidation
XValidation
Y-NaiveBayes
Y-NaiveBayes
OperatorChain
OperatorChain
Parameters
XML
Comment
New Operator
Key
Value
default
value
columns
Edit List (0)...
replenishment_value
Unknown
0.0: 2108 299
1.0: 378 527
]
(created by PerformanceEvaluator)
02.04.2007 19:20:15: Experiment finished successfully
Max: 343 MB
Total: 343 MB
19:20:38
Start
D:\localhome\DMV\yale
YALE@annex8 (xval.x...
Microsoft PowerPoint - [0...
19:20

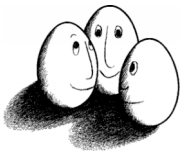


Model Building (4)

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




YALE@annex8 (xval.xml*)
File Edit View Experiment Tools Help
Operator Tree
Root
Experiment
ExcelExampleSource
ExcelExampleSource
AttributeSubsetPreprocessing
AttributeSubsetPreprocessing
Numeric2Polynominal
Numeric2Polynominal
ChangeAttributeType
ChangeAttributeType
MissingValueReplenishment
MissingValueReplenishment
RemoveUselessAttributes
RemoveUselessAttributes
StratifiedSampling
StratifiedSampling
XValidation
XValidation
Y-NaiveBayes
Y-NaiveBayes
OperatorChain
OperatorChain
Parameters
XML
Comment
New Operator
Key
Value
default
value
columns
Edit List (0)...
replenishment_value
Unknown
0.0: 2108 299
1.0: 378 527
]
(created by PerformanceEvaluator)
02.04.2007 19:20:15: Experiment finished successfully
YALE@annex8 (xval_tree.xml)
File Edit View Experiment Tools Help
Model
graph view
text view
Tree View
CustDrop
= No
= Yes
paymentStatus
1.0 (13.0)
0.0 (406.0/92.0)
0.0 (17.0)
0.0 (0.0)
1.0 (5.0)
0.0 (0.0)
Save...
+- J48[0] (J48)
+- OperatorChain[0] (OperatorChain)
+- ModelApplier[0] (ModelApplier)
+- PerformanceEvaluator[0] (PerformanceEvaluator)
02.04.2007 19:40:20: J48: Breakpoint reached
[2] J48: breakpoint reached after operator, press resume...
19:40:26
Start
D:\localhome\DMV\yale
YALE@annex8 (xval.x...
D:\localhome\DMV
02_CRISP.ppt
YALE@annex8 (xval_t...
19:40



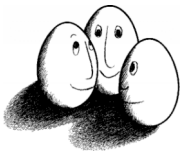
Model Evaluation (5)

- 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!

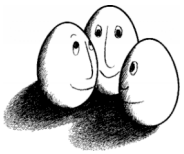
classification_error: 10.76% +/- 1.50% (mikro: 10.76%)

	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