# Data Mining mit Rapidminer im Direktmarketing – ein erster Versuch

Hasan Tercan und Hans-Peter Weih

Standard Life

# Motivation und Ziele des Projekts

- Anwendung von Data Mining im Versicherungssektor
  - Unternehmen: Standard Life Versicherung
  - Produkte für Altersvorsorge und Investment

- Themengebiet: jährliche Direktmarketing-Kampagne
  - Ausgewählte Kunden bekommen Anschreiben mit dem Angebot einer steuerbegünstigten Zuzahlung in ihren Vertrag

- Motivation: Optimierung der Kundenselektion für Kampagne anhand Data Mining (bisher anhand Expertenwissen und „Trial-and-Error“)

# Erfolg der Aktionen 2010 bis 2013

| Jahr | Anschreiben | Zuzahlungen | Quote |
|---|---|---|---|
| 2013 | 27.508 | 1.645 | 5,98% |
| 2012 | 20.836 | 1.393 | 6,69% |
| 2011 | 22.059 | 1.250 | 5,67% |
| 2010 | 15.994 | 1.295 | 8,10% |

- Ziele des Projekts:
  - Implementierung eines erfolgreichen Data Mining Prozesses
  - Reduktion Kosten der Kampagne (Anzahl der Briefe)
  - Erhöhung der Rücklaufquote
  - Erhöhung des Profits

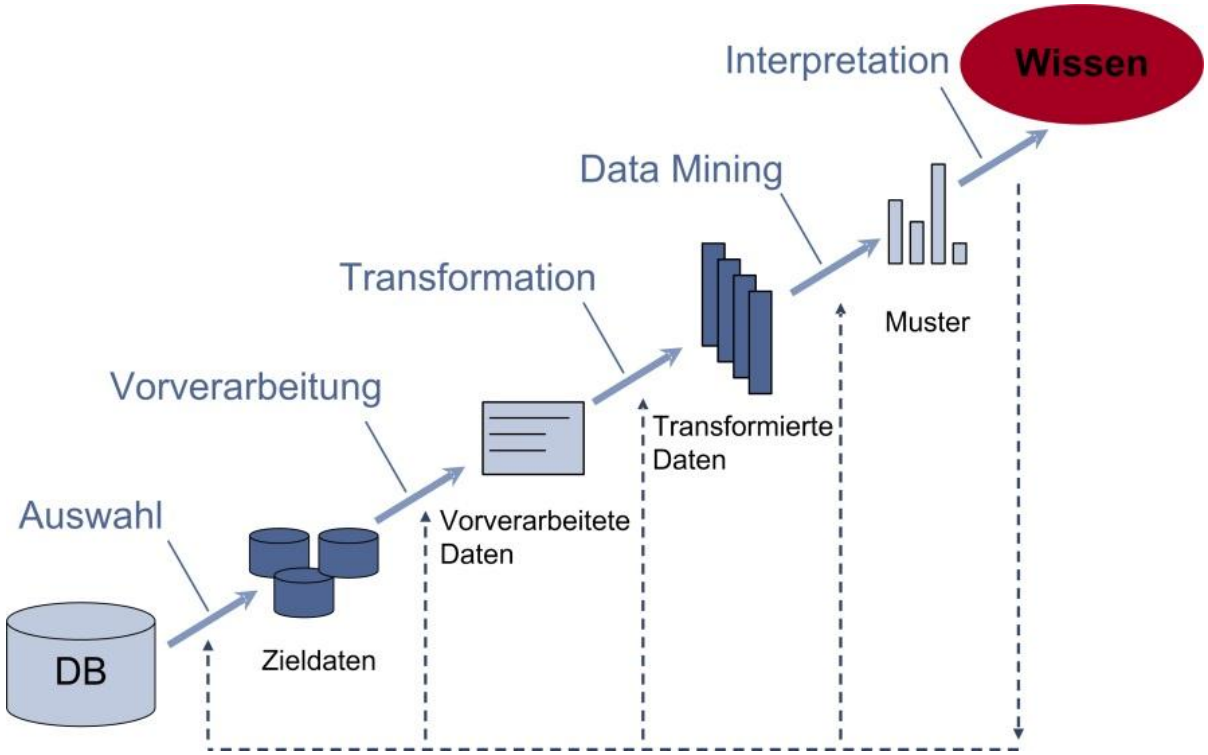
# Data Mining

- Rechnergestützte Verfahren zur Analyse von großen Datenbeständen

- Ziel: verstecktes Wissen aus Datenbeständen zu extrahieren. Finden von Mustern, die
  - bislang unbekannt,
  - potenziell nützlich und
  - leicht verständlich sind

- Data Mining als Prozess der Wissensentdeckung in Datenbanken (inkl. Aufbereitung der Daten und Evaluation der Resultate)

# Data Mining Prozess (KDD)




1. Auswahl des relevanten Datenbestandes (z.B. Vertragsdaten, Kundendaten)
2. Datenbereinigung und Vorverarbeitung (z.B. durch Korrektur und Ergänzung)
3. Datentransformation und Datenreduktion: Konstruktion des Finalen Datensatzes (z.B. durch Typ-Konversion, Filtern)
4. Anwendung von Data Mining Algorithmen
5. Interpretation und Evaluation des gefundenen Modells.

# Data Mining Tools - Rapidminer

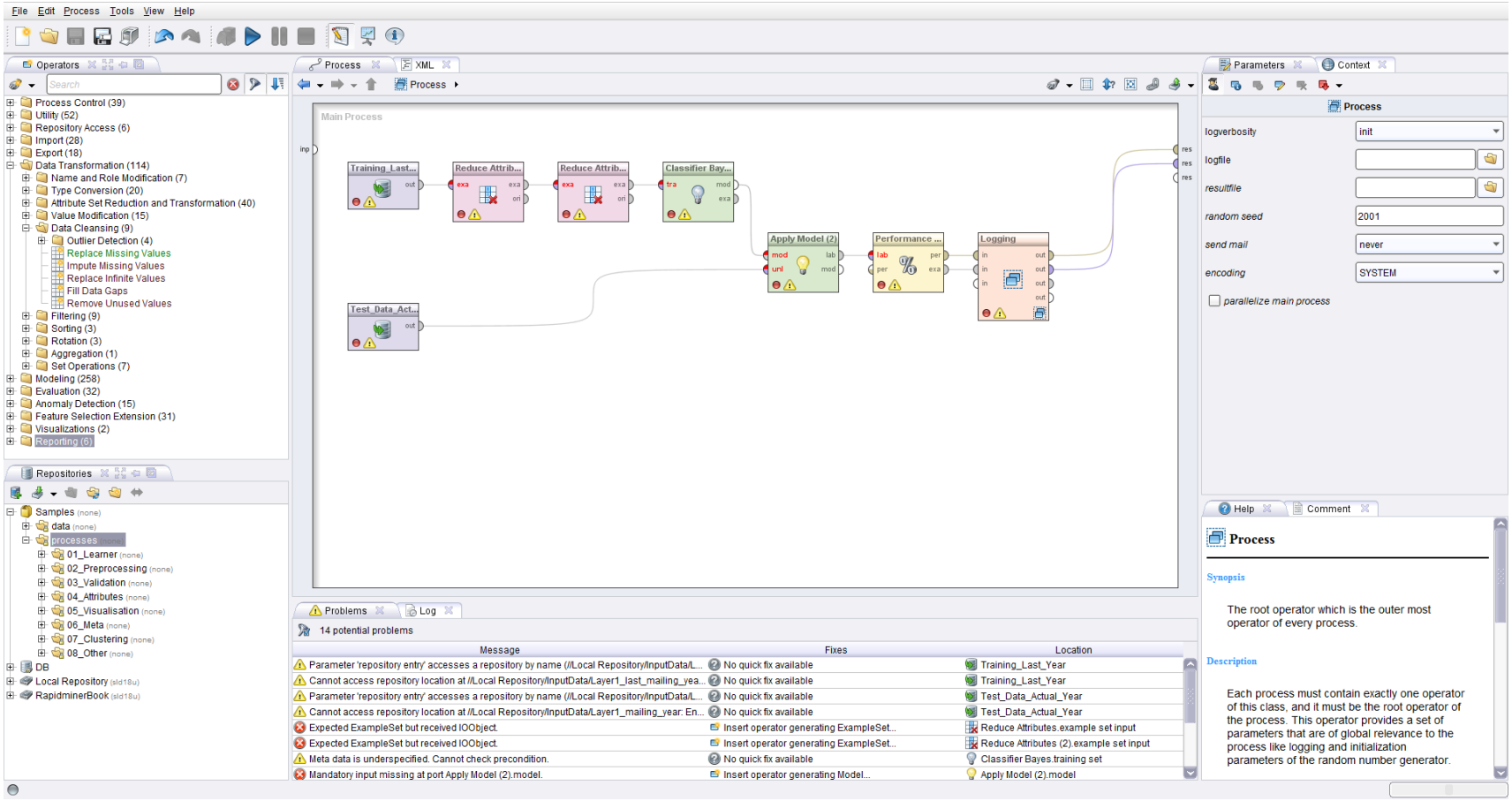
- Open Source Tools: KNIME, WEKA, Rapidminer

- Rapidminer Studio:
  - Version 5.3 (Open-Source)
  - Version 6 (Kommerziell)

- Rapidminer unterstützt den Data Mining Prozess und implementiert viele Methoden der einzelnen Phasen.

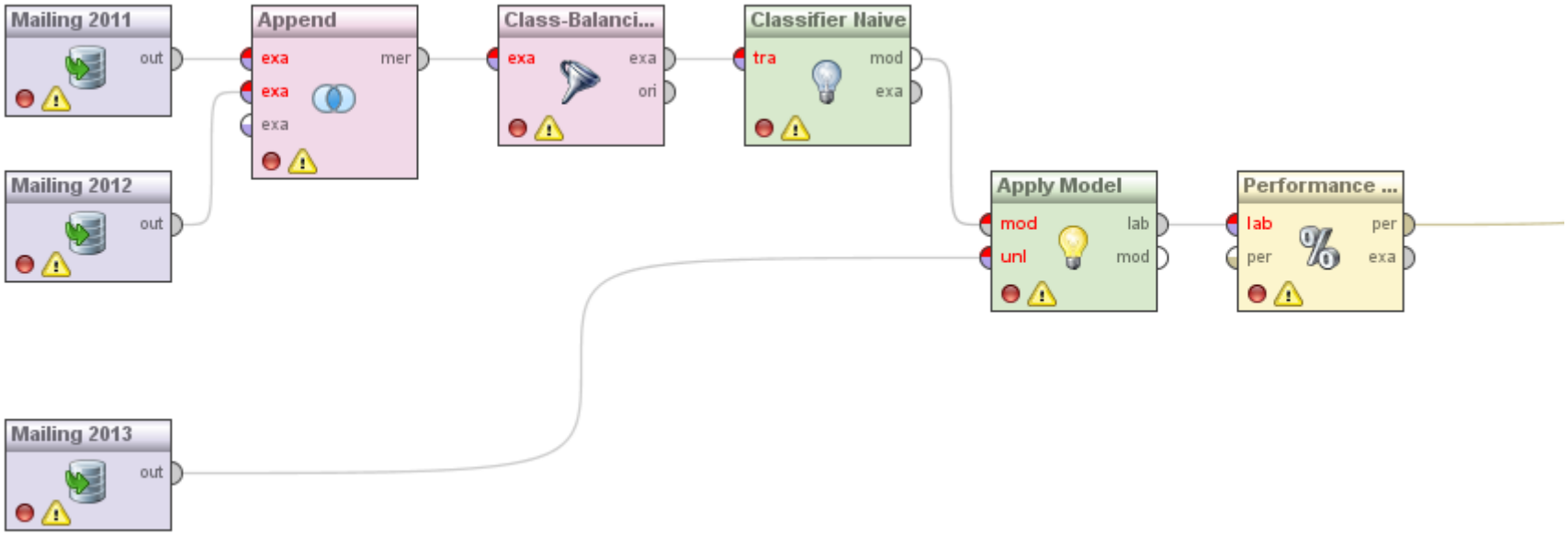
- Anwendung eins WEKA Plug-Ins in Rapidminer möglich

# Rapidminer - Ansicht

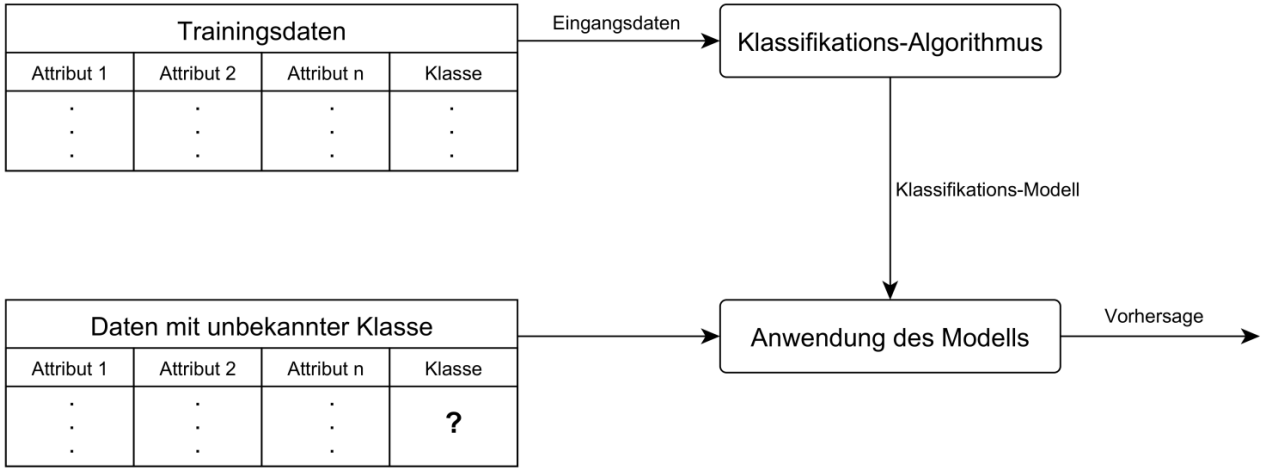


# Rapidminer Prozess

- Operatoren für Datenverarbeitung, Modellentwicklung und Evaluation
- Bilden eines Knowledge-Flows



# Klassifikationsproblem Marketingaktion




- Bilden von Klassifikationsmodellen auf folgenden Daten
  - Vorherige Aktionsdaten (2011-2013)
  - Kunden- und Vertragsdaten
  - 2 Klassen von Kunden: Zuzahlung & Nicht-Zuzahlung

- Benutze das beste Modell, um Kunden für künftige Aktion (Jahr 2014) zu klassifizieren (Prognose ja/nein)

# Schritt 1: Auswahl Datenbestand

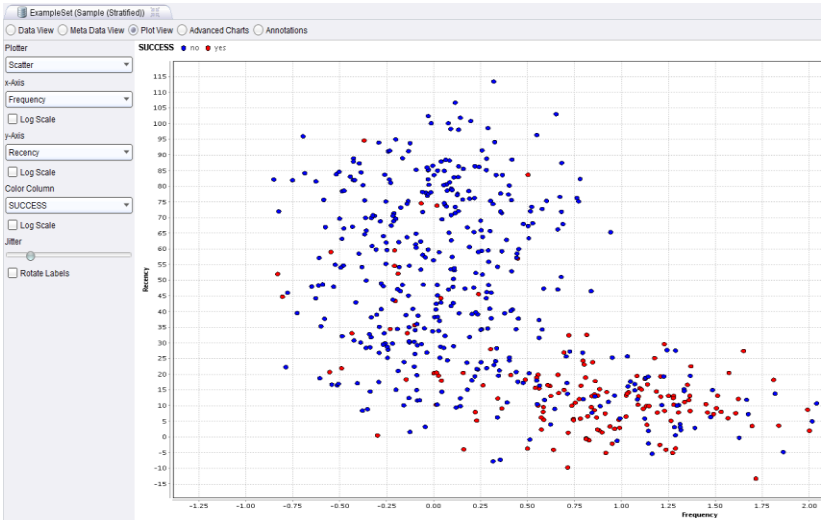
- Sammlung und Verknüpfen geeigneter Daten aus dem Bestand
  - Kunde (sozio-geographische Daten)
  - Vertrag
  - Transaktionen (z.B. Historie der getätigten Zuzahlungen)
  - Externe Daten

- Ausführung mit ETL-Tool in unserem Data Warehouse

- Resultat: eine große Datenmenge
  - Tausende Datensätze für jeweils angeschriebene Kunden
  - 33 Attribute: 8 binäre, 13 nominelle, 12 numerische
  - Eine binäre Klasse „Response“ (Zuzahlung/nicht-Zuzahlung)

# Voranalyse der Daten

- Initiale Analyse des Datensatzes mit Rapidminer
  - Statistiken der Attribute
  - Verteilung der Klasseninstanzen
  - Scatter-Plots
  - Erweiterte Charts




ExampleSet (Replace Missing Values (8))

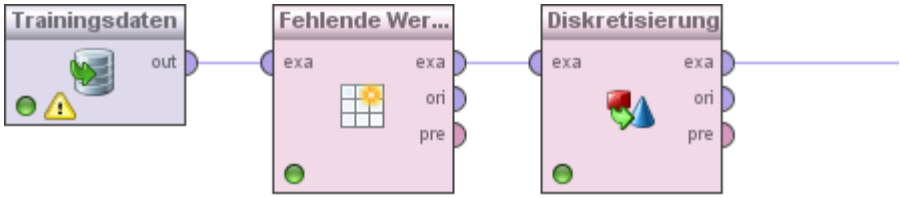
Data View | Meta Data View | Plot View | Advanced Charts | Annotations

ExampleSet (28989 examples, 2 special attributes, 41 regular attributes)

| Name | Type | Statistics | Range |
|---|---|---|---|
| Alter | integer | avg = 44.278 +/- 7.965 | [19.000 ; 70.000] |
| BU-Schutz | binominal | mode = no (24138), least = yes (4851) | no (24138), yes (4851) |
| Berufskategorie | polynominal | mode = ANDERE (9043), least = BANKER (3) | ANDERE (9043), KAUFMANN (824), ME |
| Broker_Account | polynominal | mode = Andere (12388), least = Key_Account_ | MLP (6996), Andere (12388), Key_Accou |
| Dynamik | polynominal | mode = -- (13489), least = 3,5% (2) | -- (13489), 10% (3576), 2% (1009), 3% |
| Einmalbeitrag | real | avg = 688.029 +/- 3739.490 | [0.000 ; 98155.740] |
| Frequency | integer | avg = 0.075 +/- 0.336 | [0.000 ; 12.000] |
| Geschlecht | binominal | mode = M (19626), least = F (9363) | F (9363), M (19626) |
| Jahresbeitrag | real | avg = 2083.063 +/- 2159.253 | [0.000 ; 19020.000] |
| KAUFKRAFT | real | avg = 107.552 +/- 19.025 | [61.400 ; 252.400] |

# Schritte 2-3: Aufarbeitung und Transformation

- Säubern fehlerhafter Werte und Ergänzen fehlender Werte
  - Hilfreich: hohe Datenqualität im Data Warehouse von Standard Life
- Konvertierung von Datentypen (z.B. Diskretisierung)
- Datenreduktion
  - Entfernung von Datensätzen (z.B. Sampling)
  - Entfernung irrelevanter Attribute

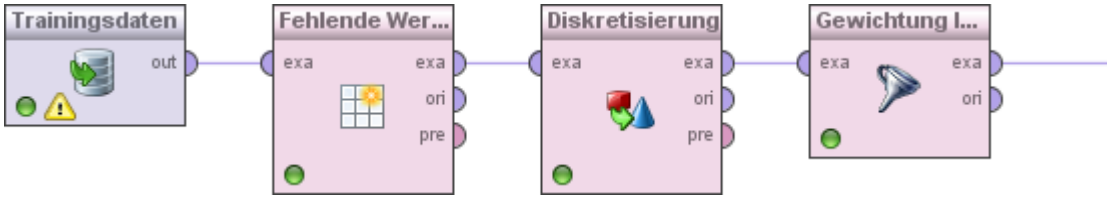


# Balancierung der Klassen in Trainingsdaten

- Under-Sampling
  - Instanzen der häufigeren Klasse werden von Trainingsmenge entfernt
  - ➢ Anzahl der Instanzen insgesamt reduziert

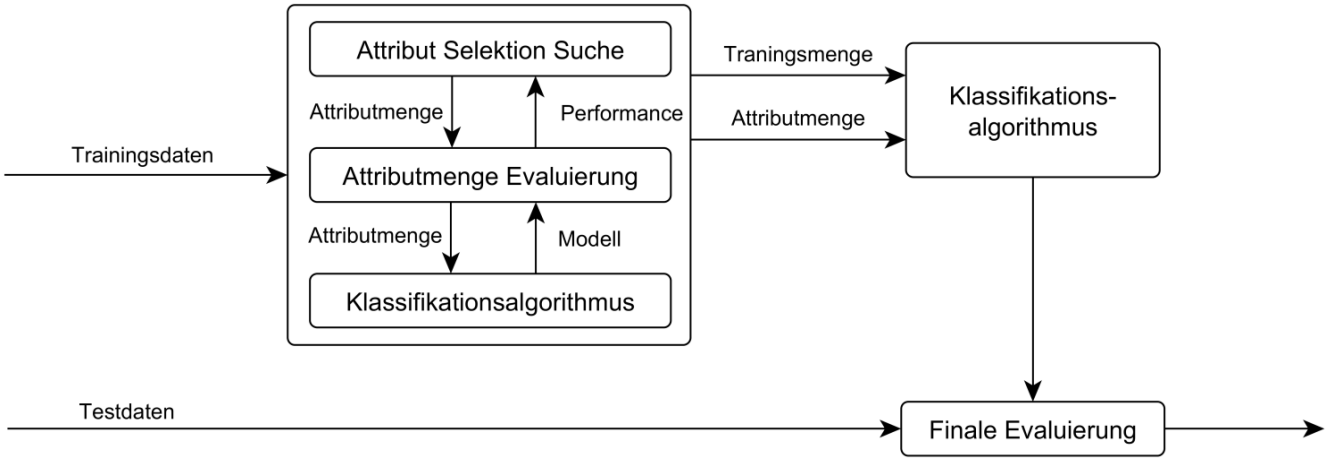
- Over-Sampling
  - Instanzen der selteneren Klasse werden vervielfacht
  - ➢ Anzahl der Instanzen insgesamt erhöht

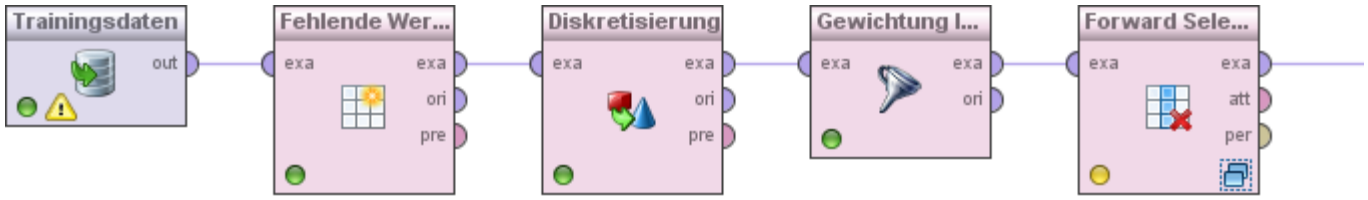
- Gewichtung der Instanzen
  - Jede Instanz bekommt eine Gewichtung
  - Seltenere Klasse bekommt höheres Gewicht wie Häufige



# Attribut-Selektion (Wrapper Ansatz)

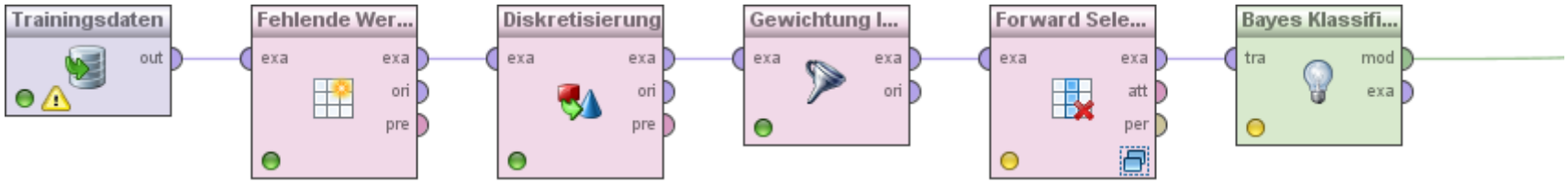
- Einbindung des Klassifikationsalgorithmus in die Attribut-Auswahl
- Auswahl der optimalen Teilmenge von Attributen für Modell
- Zwei Verfahren getestet: Backward Elimination und Forward Selection



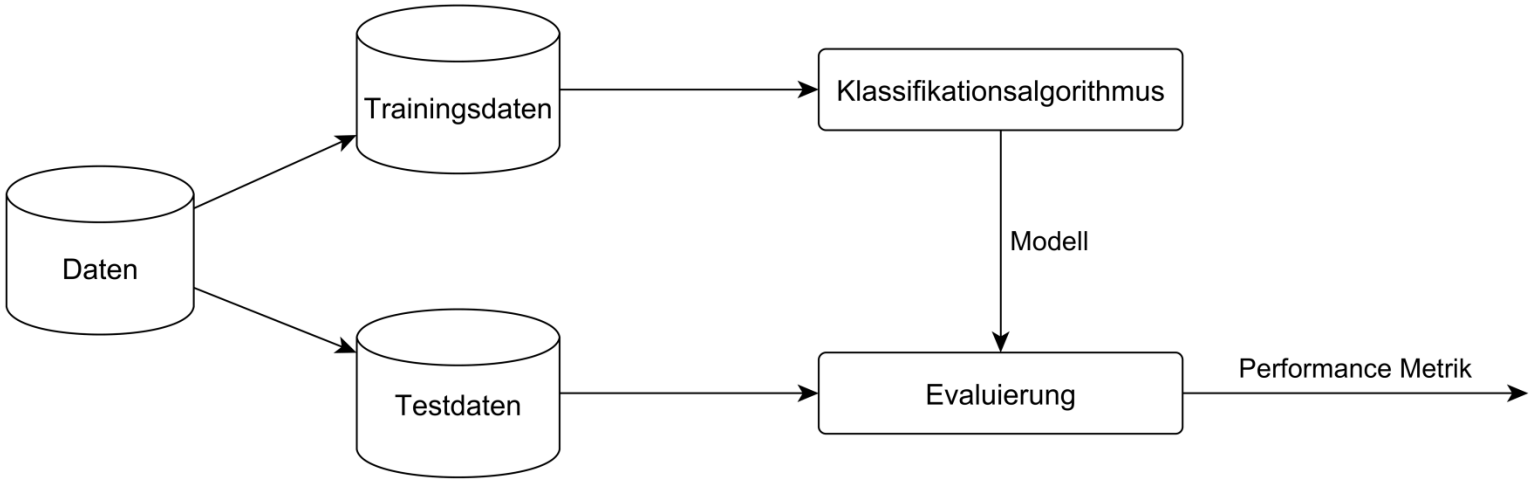
# Schritt 4: Klassifikationsmodelle

- Entscheidungsbaum
- Regelmenge
- Random Forest
- Lineare und Nicht-lineare Modelle
  - Logistische Regression
  - Support Vector Machine
  - Neuronales Netze (ANN)
- Statistische Modelle
  - Naive-Bayes
  - Bayesian Netzwerk



# Schritt 5: Evaluierung

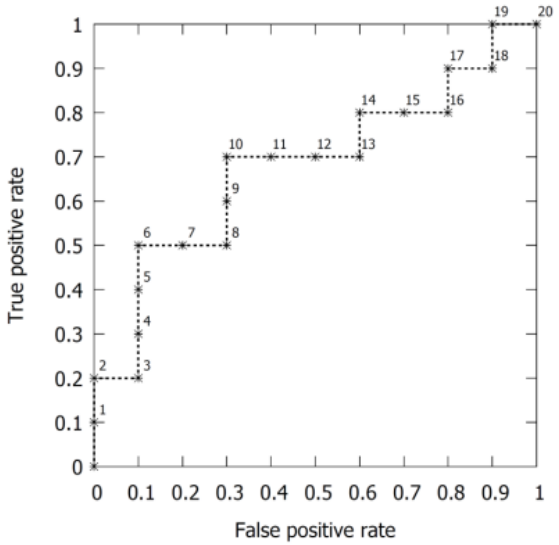
- Wie gut funktioniert die Vorhersage?
- Gelerntes Modell wird auf einer Testmenge evaluiert.
- Testmenge <> Trainingsmenge!!!
- Modell klassifiziert jede Instanz der Testmenge
- Vergleich der Klassifikation mit „realem“ Ergebnis in Testmenge



# Metriken zur Evaluierung: ROC/AUC

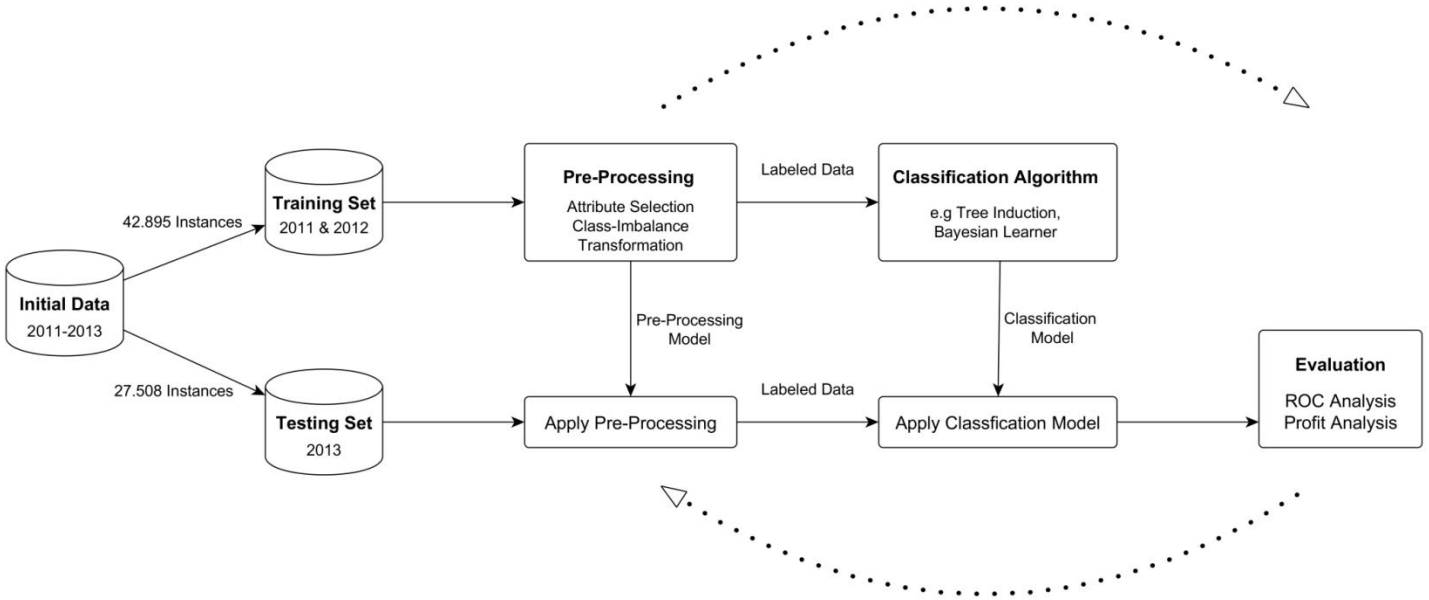
- Modellvorhersage: Wahrscheinlichkeit für Zuzahlung eines Kunden
- ROC-Performance
  - Wie gut unterscheidet Modell zwischen Zuzahlern und Nicht-Zuzahlern? (Wahr-Positiv-Rate zu Falsch-Positiv-Rate)

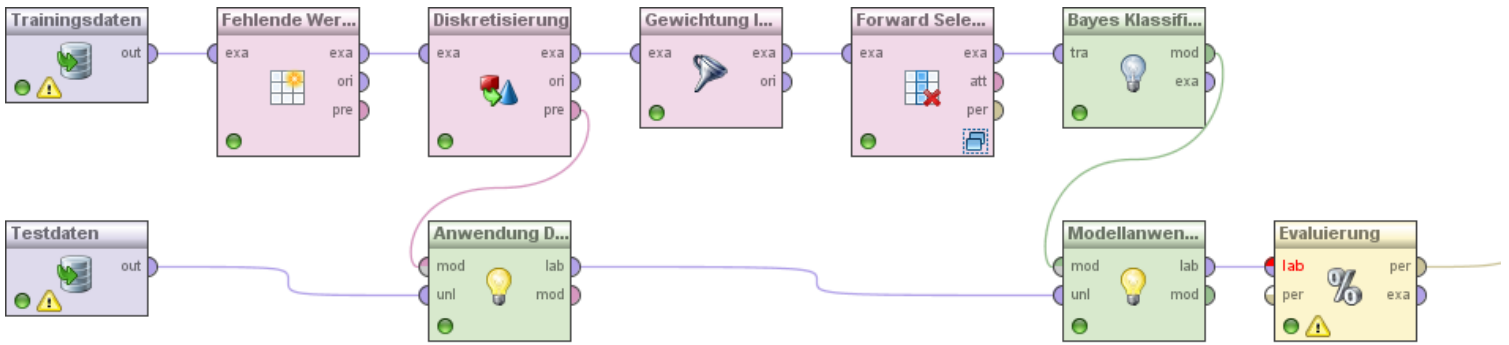
| Rank | Class | Score | Rank | Class | Score |
|---|---|---|---|---|---|
| 1 | p | 0.95 | 11 | n | 0.5 |
| 2 | p | 0.91 | 12 | n | 0.49 |
| 3 | n | 0.8 | 13 | n | 0.44 |
| 4 | p | 0.75 | 14 | p | 0.35 |
| 5 | p | 0.7 | 15 | n | 0.3 |
| 6 | p | 0.66 | 16 | n | 0.25 |
| 7 | n | 0.6 | 17 | p | 0.18 |
| 8 | n | 0.58 | 18 | n | 0.15 |
| 9 | p | 0.56 | 19 | p | 0.10 |
| 10 | p | 0.55 | 20 | n | 0.05 |




- Alternative Metrik: Area Under the Curve (**AUC**)

# Experimentelle Vorgehensweise



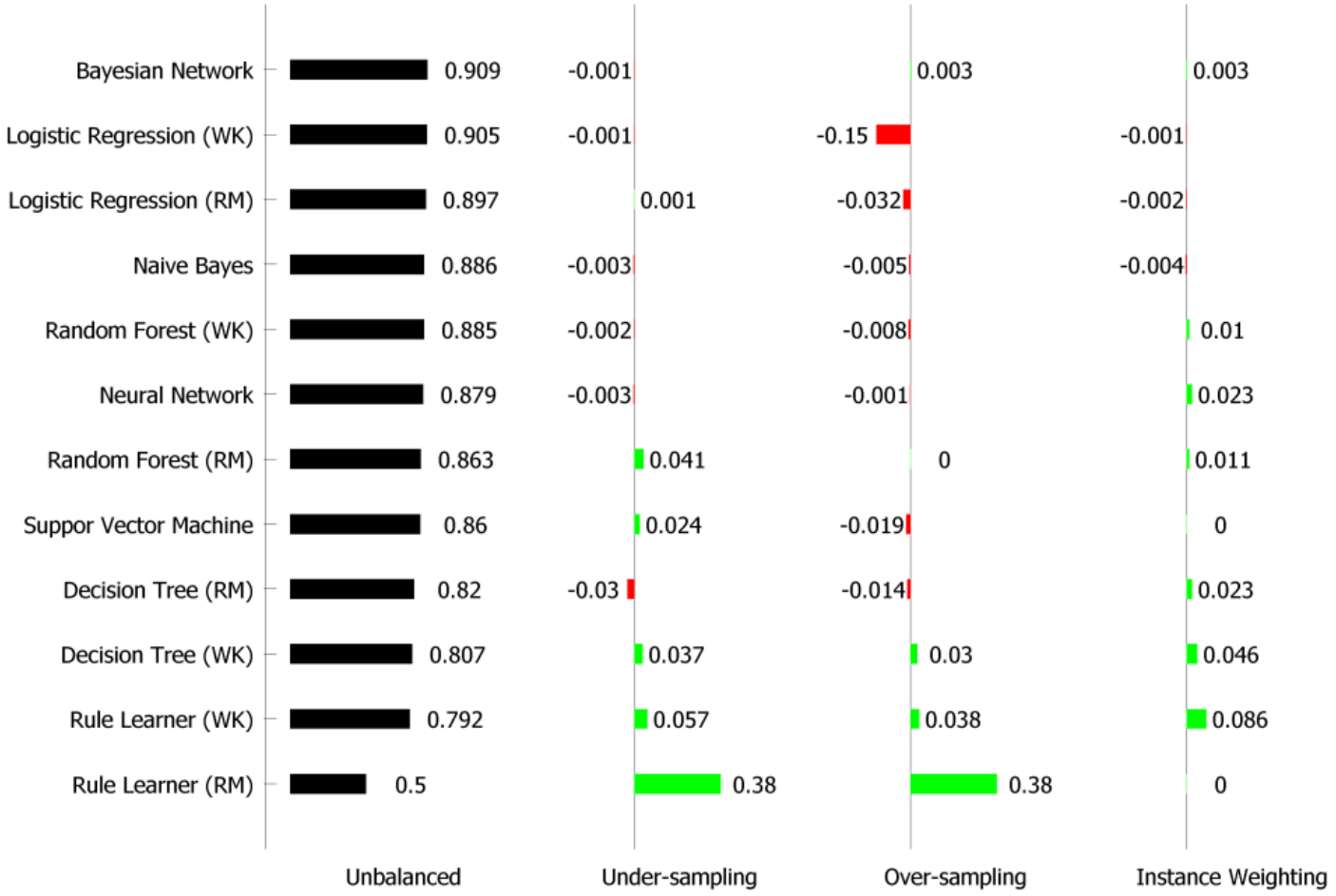



# Erste Ergebnisse ohne Datenreduktion

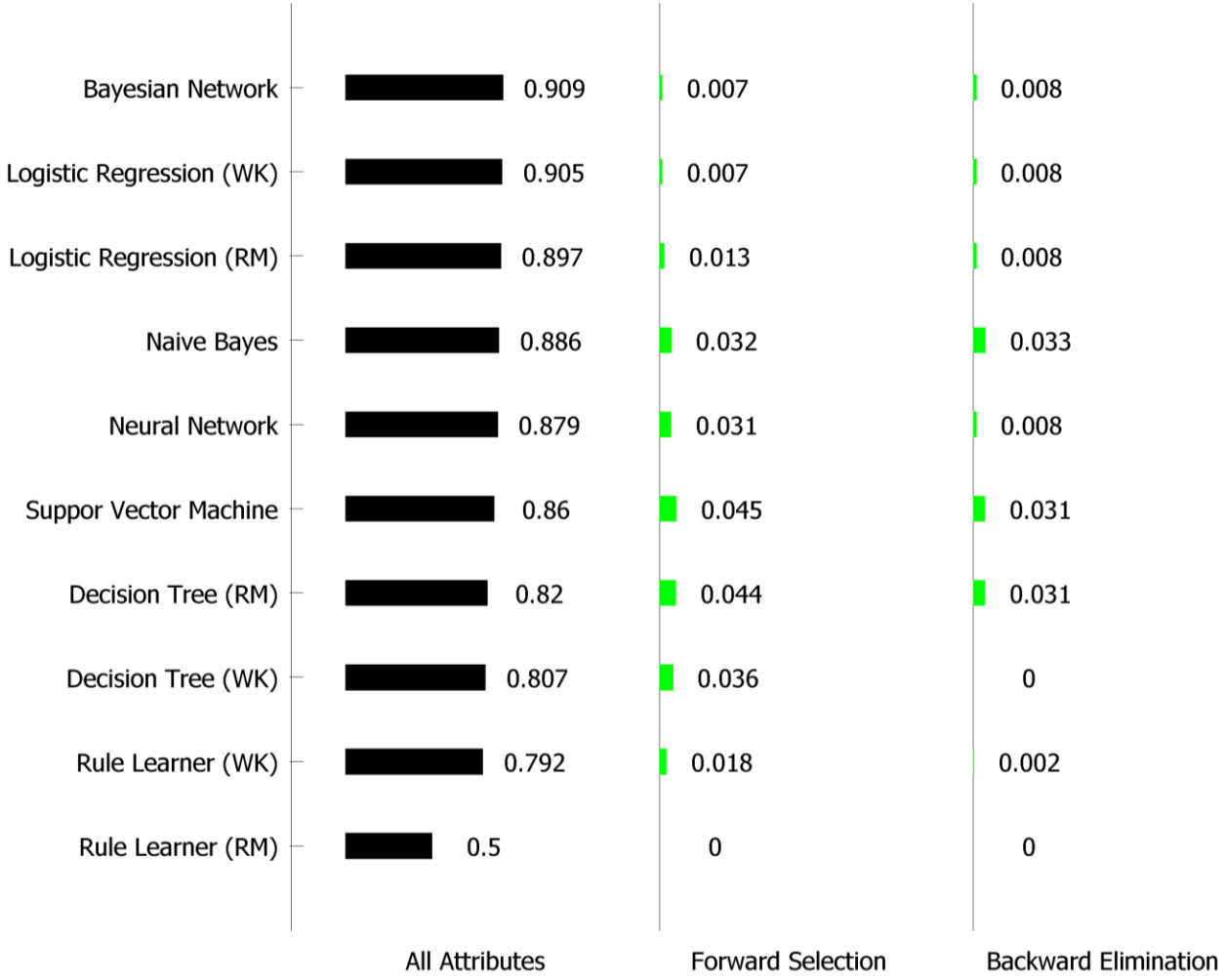
- Modelle für Wahrscheinlichkeit funktionieren besser
- Weka Algorithmen führen zu besseren Resultaten
- Regelmenge leidet unter unausgewogenen Klassen

| Model | Tool | AUC |
|---|---|---|
| Bayesian Netz | WK | 0,909 |
| Logistische Regression | WK | 0,905 |
| Logistische Regression | RM | 0,897 |
| Naiver Bayes | RM | 0,886 |
| Random Forest | WK | 0,885 |
| Neuronales Netz | RM | 0,879 |
| Random Forest | RM | 0,863 |
| Support Vector Machine | RM | 0,86 |
| Entscheidungsbaum | RM | 0,82 |
| Entscheidungsbaum | WK | 0,807 |
| Regelmenge | WK | 0,792 |
| Regelmenge | RM | 0,5 |

# Einfluss von Balancierung der Klassen (AUC)



# Einfluss von Attributselektion (AUC)



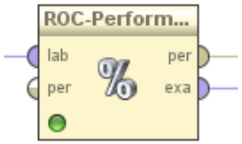
# Ergebnisse optimierter Modelle

- Statistische Modelle am Besten
- Bestes Modell
  - ➢ **Naive Bayes**
- Unterschiedliche Vorverarbeitung für die Modelle hilfreich

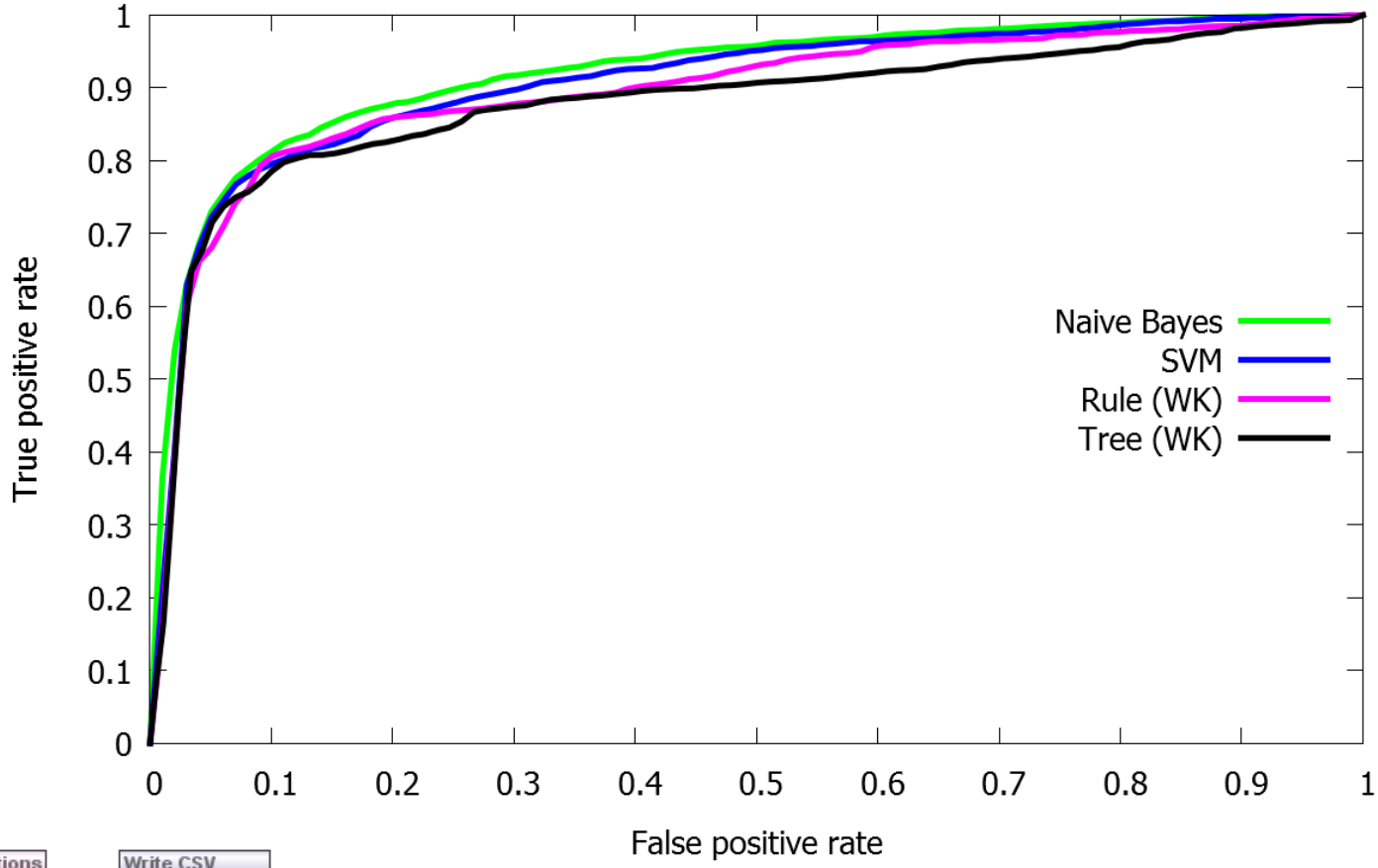
| Model | Tool | Attribut Selektion | Balance | AUC |
|---|---|---|---|---|
| Naiver Bayes | RM | BE | - | 0.919 |
| Bayesian Netz | WK | BE | - | 0,917 |
| Log. Regression | WK | BE | - | 0,913 |
| Log. Regression | RM | FS | US | 0,913 |
| ANN | RM | FS, | IW | 0,91 |
| SVM | RM | FS | US | 0.906 |
| Random Forest | RM | - | US | 0,904 |
| Regelmenge | RM | FS | US | 0,897 |
| Random Forest | RM | - | IW | 0,895 |
| Regelmenge | WK | BE | IW | 0,886 |
| Entscheidungsbaum | WK | FS | IW | 0,881 |
| Entscheidungsbaum | RM | FS | IW | 0,88 |

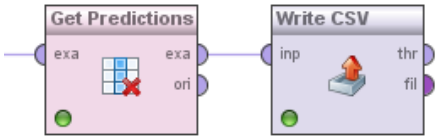
BE = Backward Elimination, FS = Forward Selection
US = Undersampling, IW = Instance Weighting

# ROC-Kurve des naiven Bayes

# ROC-Kurven ausgewählter Modelle



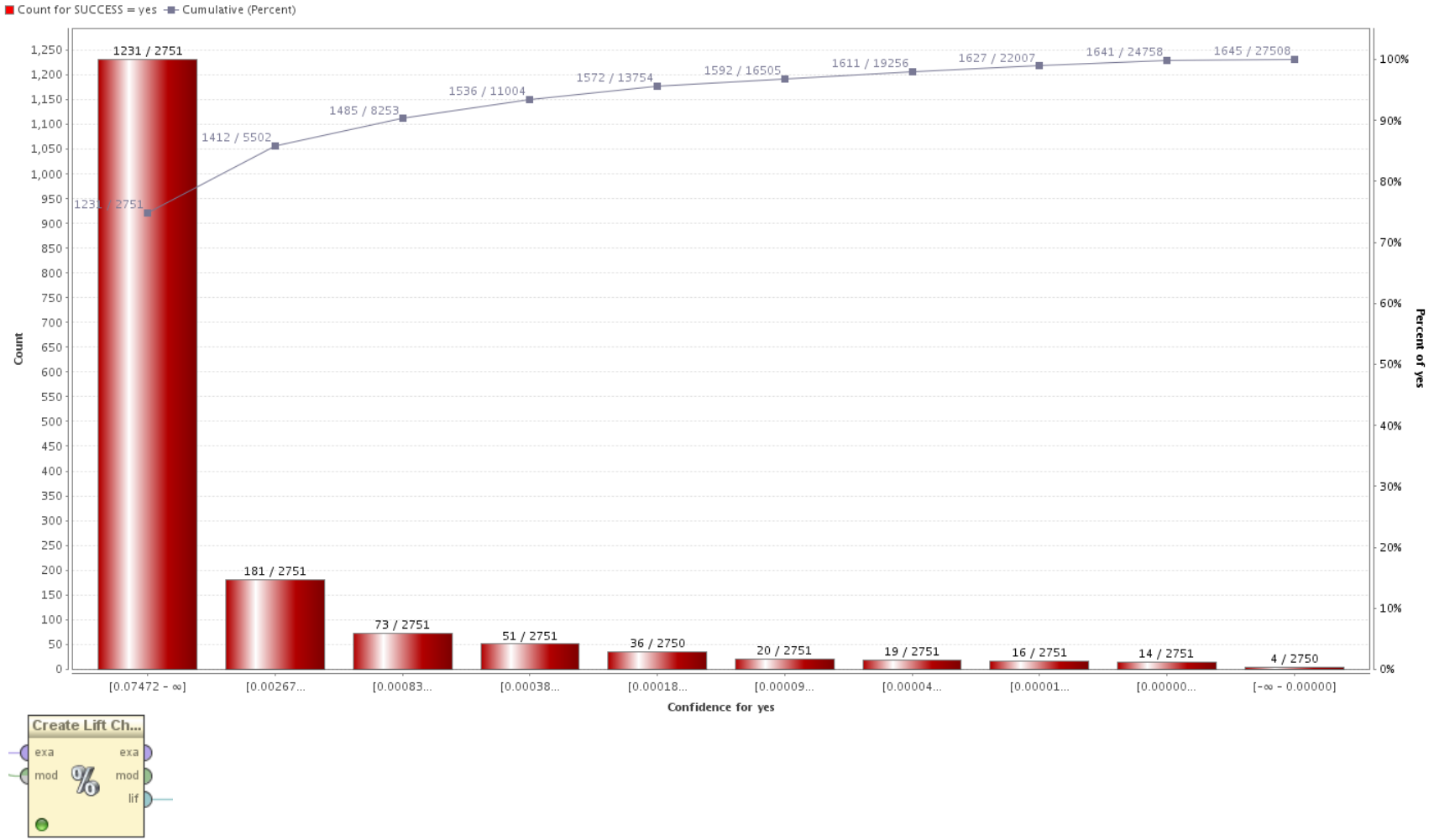
# Ökonomischer Nutzen

- Frage: bringt das naive-Bayes Modell einen ökonomischen Nutzen?
  - ➢ Kosten-Nutzen-Analyse notwendig!

- Wirtschaftliche Daten zur Aktion 2013 (ohne Modell)
  - Kosten der Aktion
    - Fixkosten: ca. 5.300 €
    - Variable Kosten: ca. 1 €/Brief
  - Umsatz der Aktion
    - ca. 100 €/Zuzahlung
  - Reiner Gewinn: 131.692 €

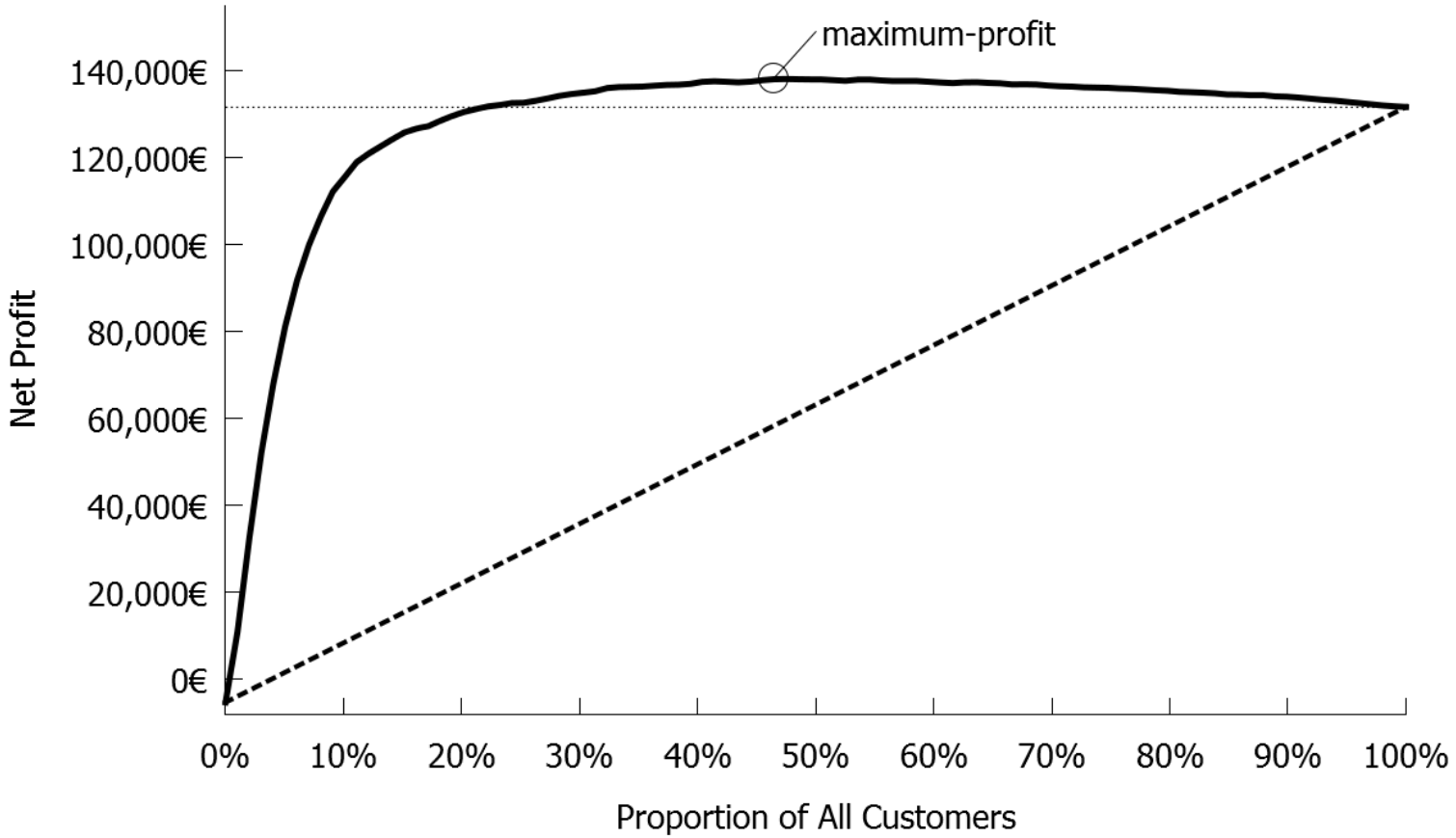
- Beachte: diese Zahlen entsprechen nicht dem realen Kosten-Nutzen-Verhältnis

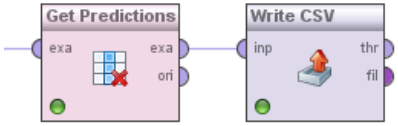
# Lift Chart vom Modell in Rapidminer



# Kosten-Nutzen Analyse

## Gewinnkurve mit und ohne naivem Bayes Klassifizierer



# Erfolg von Data Mining anhand Aktion 2013

| | Naiver Bayes | Ohne Modell |
|---|---|---|
| **Anschreiben** | 12.773 | 27.508 |
| **Zuzahlungen** | 1.563 | 1.645 |
| **Rücklaufquote** | 12,24% | 5,98% |
| **Fixe Kosten** | 5.300 € | 5.300 € |
| **Variable Kosten** | 12.773 € | 27.508 € |
| **Umsatz** | 156.300 € | 164.500 € |
| **Reiner Gewinn** | 138.227 € | 131.692 € |

- Mit naive Bayes: Kosten reduziert, Rücklaufquote erhöht und Gewinn erhöht

# Erkenntnisse

- Data Mining kann erfolgreich im Unternehmen eingesetzt werden
  - Voraussetzung: Data Mining Prozess mit geeigneten Methoden zur Datenaufbereitung und Evaluierung
- Daten über Kundenverhalten und Vertrag wichtiger als sozio-geographische Informationen
- Attribut Selektion und Balancierung bringen große Verbesserungen
- Open-Source Tool (Rapidminer) unterstützt den Data Mining Prozess
  - Umfangreiche und flexible Gestaltung
  - Keine Programmierung notwendig
  - Dennoch: tiefes Verständnis für Data Mining nötig

# Kontakte & Literatur


**Hans-Peter Weih**

- Dipl.-Informatiker
- Leiter Management Information
- Bei Standard Life seit 1998
- hanspeter.weih@standardlife.de

**Hasan Tercan**

- M. Sc. Informatik
- Seit 2015 Business Specialist bei Standard Life
- hasan.tercan@standardlife.de


Literatur


- Rapidminer Buch mit vielen Use-Cases (von Markus Hofmann und Ralf Klinkenberg): http://rapidminerbook.com/
- Rapidminer Vortrag auf der letztjährigen OSBI-Konferenz (von Ralf Klinkenberg): http://www.osbi-workshop.de/wordpress/wp-content/uploads/2014/05/RapidMiner__Predictive_Big_Data_Analytics_for_Extracting_Actionable_Insights__Applications_in_Manufacturing.pdf

# Rechenleistung & Laufzeiten

- Hardware Spezifikation
  - Office PC mit Windows 7 Enterprise
  - CPU: Intel i7 (Quad-Core) 3,40 GHz
  - RAM: 16,0 GB

- Min. und max. Laufzeiten für Algorithmen auf vollem Datensatz
  - Naiver-Bayes: ca. 5 Sekunden
  - Neuronales Netz (ANN) : ca. 7 Minuten
  - ➢ Bei Attributselektion mit hunderten Läufen eines ANN: mehrere Stunden Laufzeit