Karlsruher Institut für Technologie

Praktikum: Data Warehousing und Mining

Zwischenpräsentation
Gruppe 3

Institut IPD Böhm, Karlsruher Institut für Technologie (KIT)

Data Preparation

Data Preparation: Nicht verwendete Spalten

Spalte	Verwendung	S	dung
customernumber	**nicht verwenden**		
date		numberitems	
salu		**gift**	**nicht verwenden**
title	**nicht verwenden**	entry	
domai		**points**	**nicht verwenden**
model			
paymenttype		weight	
deliverytype		remi	
invoicepostcode		cancel	
delivpostcode		used	
voucher		w0-w10	
advertisingdatacode			

Versuch Auffälligkeiten zu finden schlug fehl.

Nur 148 Einträge haben eine 1

Nur 228 Einträge besitzen 1, Rest 0
Beide Klassen liefern im Schnitt
18,x% 1er bei target90

Hier liegt nur eine Ausprägung vor (0)

Data Preparation: Übernommene Spalten

Spalte	Verwendung	Spalte	Verwendung
customernumber	nicht verwenden	**case**	**verwenden**
date		**numberitems**	**verwenden**
salutation	**verwenden**	gift	nicht verwenden
title	nicht verwenden	**entry**	**verwenden**
domain	verwenden	points	nicht verwenden
datecreated		**shippingcosts**	**verwenden**
newsletter	**verwenden**	deliverydatepromised	
model	**verwenden**	deliverydatereal	
paymenttype	**verwenden**	weight	
deliverytype	**verwenden**	**remi**	**verwenden**
invoicepostcode		**cancel**	**verwenden**
delivpostcode		**used**	**verwenden**
voucher	**verwenden**	**w0-w10**	**z.T. verwenden**
advertisingdatacode			

Data Preparation: Modifizierte Spalten

Spalte	Verwendung	Spalte	Verwendung
customernumber	nicht verwenden		
date	**Monat verwenden**		
		gift	nicht verwenden
datecreated	**Monat verwenden**		
		deliverydatepromised	**Jahr verwenden**
		deliverydatereal	**Diff zu promised**
		weight	**Klassifikation**
	verwenden		
invoicepostcode	**zusammenfassen**		
delivpostcode	**erstens Zeichen**		verwenden
voucher	verwenden	**w0-w10**	**Klassifikation**
advertisingdatacode	**erstes Zeichen**	**date-datecreated**	**erstellen**

Jahr und Tag ohne große Bedeutung

Jahr und Tag ohne große Bedeutung

Monat korreliert sehr stark mit date und datecreated, Jahr ist vielversprechender

Reduzierung auf eine Spalte und Verringerung der Ausprägungen

Zusammenfassung

Differenz wird betrachtet

Bsp.: alle Bücher bilden eine Spalte

Verringerung der Ausprägungen

neue Spalte

Data Preparation: Übersicht

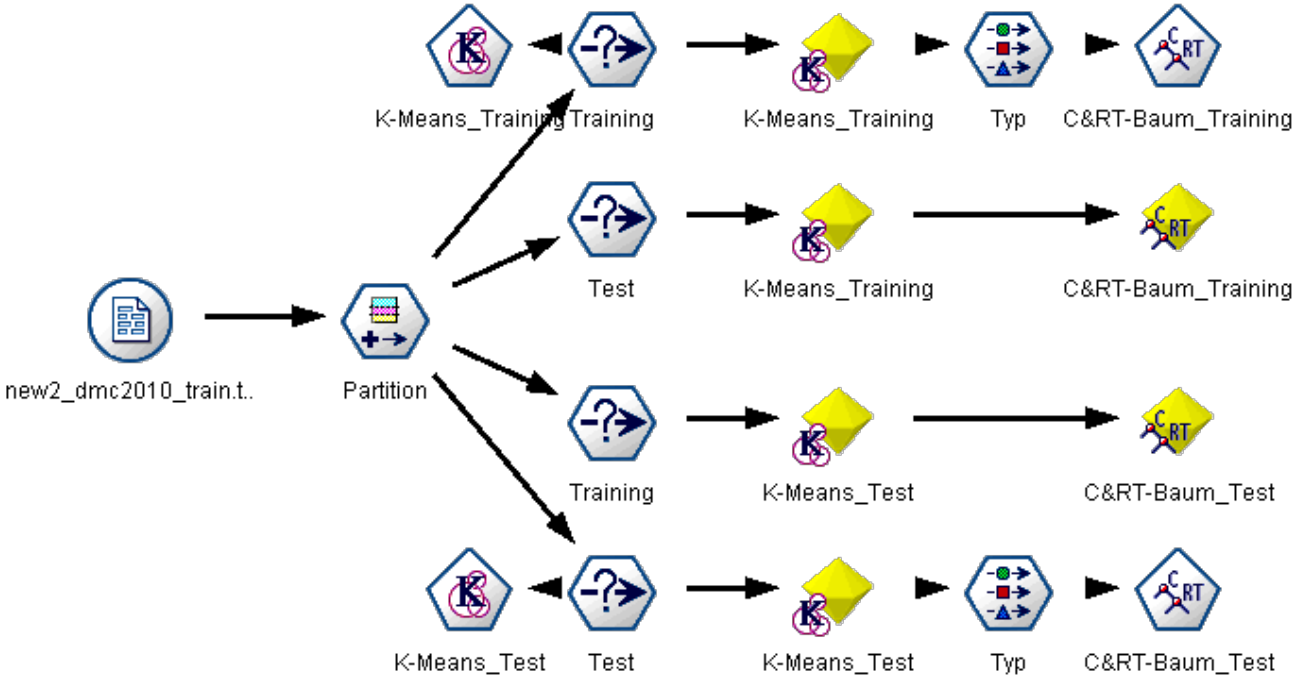
Spalte	Verwendung	Spalte	Verwendung
customernumber	nicht verwenden	case	verwenden
date	Monat verwenden	numberitems	verwenden
salutation	verwenden	gift	nicht verwenden
title	nicht verwenden	entry	verwenden
domain	verwenden	points	nicht verwenden
datecreated	Monat verwenden	shippingcosts	verwenden
newsletter	verwenden	deliverydatepromised	Jahr verwenden
model	verwenden	deliverydatereal	Diff zu promised
paymenttype	verwenden	weight	Klassifikation
deliverytype	verwenden	remi	verwenden
invoicepostcode	zusammenfassen	cancel	verwenden
delivpostcode	erstens Zeichen	used	verwenden
voucher	verwenden	w0-w10	Klassifikation
advertisingdatacode	erstes Zeichen	date-datecreated	erstellen

Modeling & Evaluation

Entscheidungsbaum: Modellierung

- Zuerst werden die Daten klassifiziert (mit K-Means)
- Danach wird auf dem neuen Datenbestand ein Entscheidungsbaum (C&RT) aufgebaut

- Dieses Verfahren wurde mit drei verschiedenen Partitionierungen sowohl auf der Training- als auch auf der Testpartition durchgeführt.



Entscheidungsbaum: Aufbau

- Im Entscheidungsbaum häufig enthaltene Spalten
  - Klassifikationsergebnis
  - remi
  - paymenttype
  - newsletter
  - date_month
  - weight
  - model
  - invoices_deliv_postcode
  - numberitems

Entscheidungsbaum: Ergebnisse

- Ergebnis:
  - Kostenmatrix (Durchschnitt):

		Wirkliche Wert	
		0	1
Gutschein	Nein	17.7 %	30.0 %
	Ja	82.3 %	70.0 %

Großer Verbesserungsbedarf!

  - Durchschnittlicher Gesamtgewinn: 11.384
  - Maximaler Gewinn: 12.296
  - Minimaler Gewinn: 10.136
  - Gewinn, wenn jeder einen Gutschein bekommt: 9.311
  - Gewinn, wenn alles richtig klassifiziert wird: 39.566

Vielen Dank für eure Aufmerksamkeit!