SAS Data Mining Fellowship Programm “Data Mining im Direktmarketing”

Uwe Steinlein

WEKA Management Fachverlage
86438 Kissing
Tel.: 08233/23-378
Fax: 08233/23-114
E-Mail: Uwe.steinlein@mfv.weka.de

In dieser Arbeit wurde unter Einsatz des SAS Enterprise Miners eine Responseoptimierung und Cross-Selling-Analyse für Direktmarketingaktionen in Kooperation mit der Firma WEKA Management Fachverlage GmbH durchgeführt.

Der zeitliche Rahmen betrug 4 Monate, wobei ca. 2 Monate allein für die Datenvorverarbeitung verwendet werden mussten (siehe Abbildung 1).

Monat	April				Mai				
KW	14	15	16	17	18	19	20	21	22
Tätigkeit	Enter-prise Miner Schulung	Daten-manage-ment	Daten-manage-ment	Daten-manage-ment	Variablen durch-forsten	Datentransfor-mationen	Daten-trans-forma-tionen	Kunden segmen-tieren	Segmente für Response-Test auswählen

Monat	Juni				Juli				August
KW	23	24	25	26	27	28	29	30	31
Tätigkeit	Re-sponse-Modell 1	Modell-verbess-erungen	Response-Modell 2	Modell-verbess-erungen	Trans-forma-tionen für Cross-Selling	Produkte identifizieren	Zusam-menhän-ge	Cluster	Fazit

Abbildung 1 : Projektplan

Dabei wurden die Daten aus dem operativen System in SAS geladen und verschiedene Verdichtungen ausgeführt, um 4 Basistabellen zu erhalten (siehe Abbildung 2).

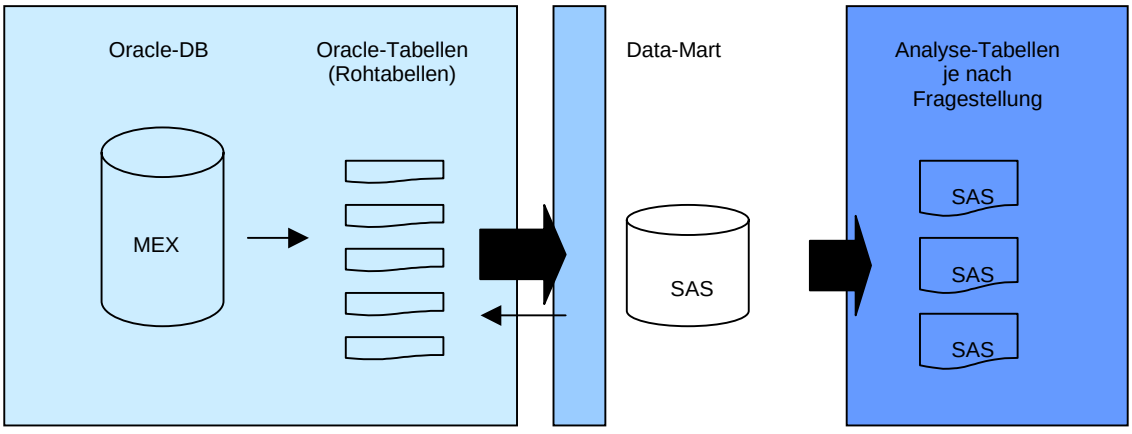


Abbildung 2: Ladeprozess

Somit konnten nun aus dem aus 4 Tabellen (Kundendaten, Umsatzdaten, Werbeaktionsdaten, Zuordnung Werbeaktionen-Kunde) bestehenden Data Mart je nach Fragestellung die entsprechende Analysetabelle gebildet werden.
Im ersten Schritt wurde eine Tabelle zur **Responseoptimierung** benötigt. Es sollte eine Bestellwahrscheinlichkeit für ein bestimmtes Produkt berechnet werden. Dazu wurden nun verschiedene bereits ausgelaufene Werbeaktionen zu diesem Produkt als Trainingsdaten aggregiert, wobei ein Datensatz dieser Tabelle eindeutig durch die Werbeaktionsnummer und Kundennummer definiert war. Zu jedem dieser ID's wurden adressbezogene Daten, Umsatzaggregationen und die Werbehistorie abbildende Merkmale hinzugefügt jeweils relativ zum entsprechenden Werbezeitpunkt gesehen, um aus diesen Merkmalen einen möglichst guten Classifier zu erhalten.

Ein Problem war vor allem die ungleiche Verteilung Besteller / Nichtbesteller. In diesem Fall wurde ein Downsizing durchgeführt, d.h. es wurden alle Besteller selektiert und eine ebenso große Zahl an Nichtbestellern über eine Zufallsauswahl beigemischt. Da eine ausreichend große Menge Daten verfügbar war, wurde der Datensatz in Trainings- und Validationsdaten im Verhältnis 70/30 aufgeteilt. Die fehlenden Werte hatten zum Teil eine echte Bedeutung, andere sollten bestmöglich ersetzt werden. Dazu bietet der „Replacement" Knoten viele Möglichkeiten, beispielsweise feste Werte einsetzen oder über Mittelwert, verteilungsabhängige Werte oder über einen Entscheidungsbaum die fehlenden Werte ersetzen. Im nächsten Schritt wurde über den „Transformation" Knoten des Enterprise Miners neue Variablen gebildet (z.B. Verhältniszahlen, aber auch andere Aggregationstufen, bspw. bei der Branche eine allgemeinere Stufe) und alte Variablen z.T. über binning in Gruppen eingeteilt.

Zur Modellerstellung wurde eine logistische Regression (backward selection), ein CHAID-Entscheidungsbaum und ein neuronales Netz verwendet. Da die Interpretierbarkeit der Ergebnisse vorerst im Vordergrund stand, wurde das Neuronale Netz nur zum Vergleich eingesetzt. Zur Veranschaulichung der Ergebnisse gibt es den „Assessment" Knoten, der je nach Einstellung verschiedene Grafiken erstellt. Die Ergebnisse bei Verwendung der Validationsdaten zeigen, daß der Entscheidungsbaum und das Neuronale Netz die besten Quoten liefern, wobei beide einen sehr guten Lift erzielen (siehe Abbildung 3).

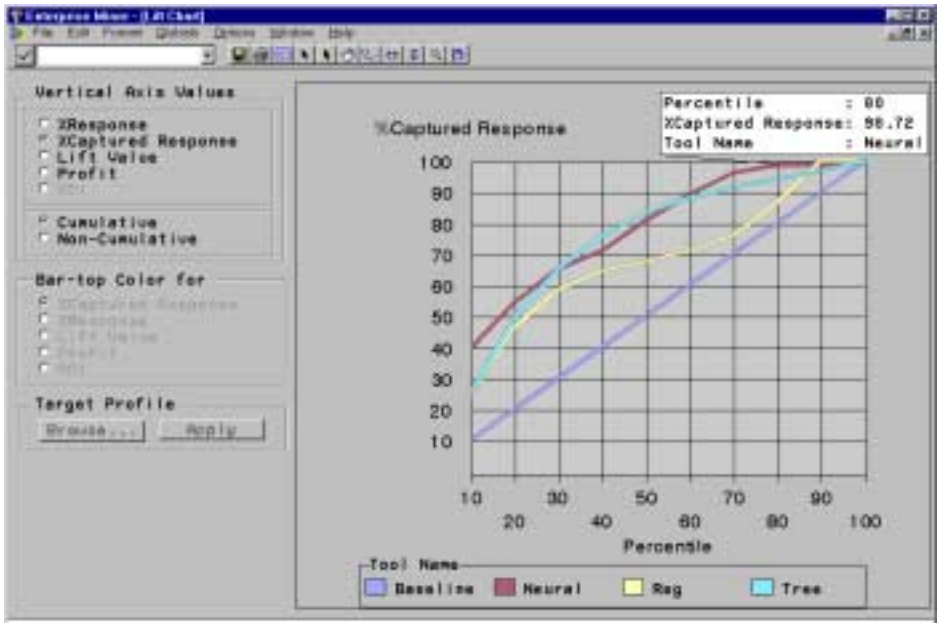
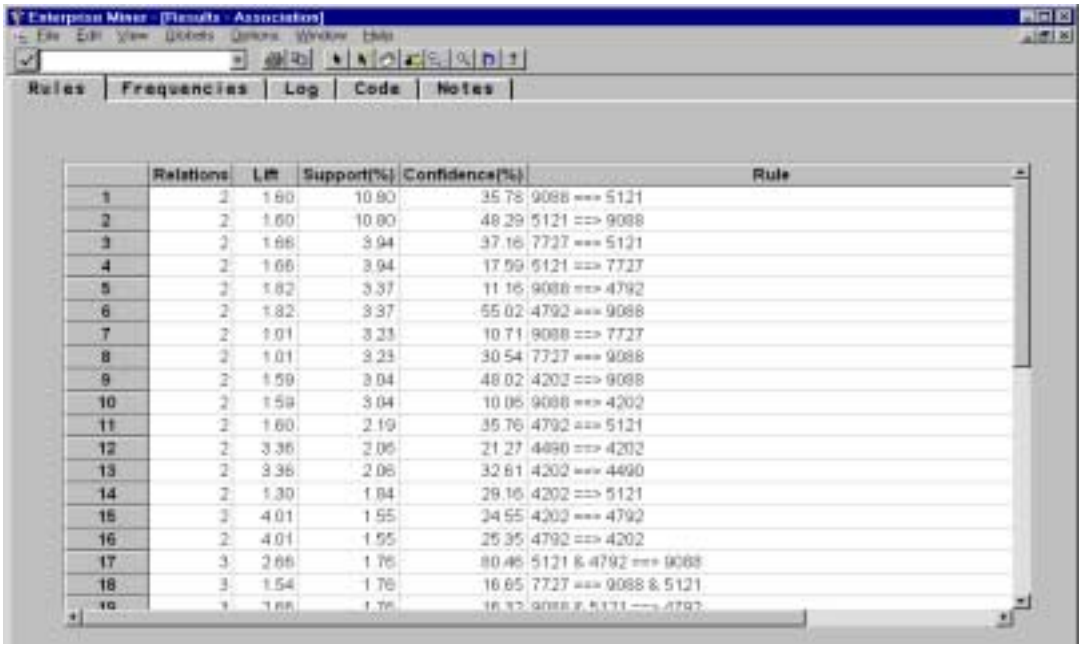


Abbildung 3: Model-Assessment

Auch die Erkenntnis, welche Variablen eine Bestellung stark beeinflussen haben weitere Einsicht in das Kundenverhalten gegeben, die wiederum gewinnbringend für zukünftige Werbeaktionen und Kundenprogramme verwendet werden kann.

Die zweite Aufgabenstellung betraf eine **Cross-Selling Analyse**. Dazu musste wieder ein neuer Datensatz gebildet werden, der nur aus 2 Spalten besteht – Kundennummer und Artikelnummer.
Die Assoziationsanalyse liefert 3 Kennzahlen: Lift, Support und Confidence.
Support: Wie häufig kommt die Kombination AB im Verhältnis zur GG vor ? ➔ *Wichtigkeit*
Confidence: Wieviele Personen, die A besitzen, haben auch B ? ➔ *Richtigkeit*
Lift: die Personen, die bereits A bestellt haben, haben eine x-mal so große Wahrscheinlichkeit als die GG, auch B zu bestellen ➔ *„Auffälligkeit“*

Die Ergebnisse sind Regeln, die Aufschluss über den Zusammenhand der verschiedenen Produkte liefern. Abbildung 4 zeigt einen Screenshot des Association Knotens.



	Relations	Lift	Support(%)	Confidence(%)	Rule
1	2	1.60	10.90	35.78	9088 ==> 5121
2	2	1.60	10.90	48.29	5121 ==> 9088
3	2	1.66	3.94	37.16	7727 ==> 5121
4	2	1.66	3.94	17.59	5121 ==> 7727
5	2	1.82	3.37	11.16	9088 ==> 4792
6	2	1.82	3.37	55.02	4792 ==> 9088
7	2	1.01	3.23	10.71	9088 ==> 7727
8	2	1.01	3.23	30.54	7727 ==> 9088
9	2	1.59	3.04	48.02	4202 ==> 9088
10	2	1.59	3.04	10.06	9088 ==> 4202
11	2	1.60	2.19	35.76	4792 ==> 5121
12	2	3.36	2.06	21.27	4490 ==> 4202
13	2	3.36	2.06	32.61	4202 ==> 4490
14	2	1.30	1.84	28.16	4202 ==> 5121
15	3	4.01	1.55	24.55	4202 ==> 4792
16	2	4.01	1.55	25.35	4792 ==> 4202
17	3	2.66	1.76	80.46	5121 & 4792 ==> 9088
18	3	1.54	1.76	16.65	7727 ==> 9088 & 5121

Abbildung 4: Ergebnisse Assoziationsanalyse

Diese Kenntnisse kann man direkt in die Produktpolitik und Bewerbung einfließen lassen, um beispielsweise Produktkombinationen zu einem Vorteilspreis anzubieten oder bestimmte Cross-Selling Aktionen zu initiieren.

Ein weiteres Vorgehen im Bereich der Cross-Selling Analyse ist das Auffinden von bestimmten Produktportfolios über eine Clusteranalyse.

Der hier benötigte Datensatz muss pro Kunde eine Zeile aufweisen, wobei in den Spalten jeweils das Merkmal Produkt vorhanden/nicht vorhanden abgebildet sein muss.

Im Enterprise Miner ist auch ein Clustering Knoten vorhanden. Da der Eingangsdatensatz ein einheitliches Skalenniveau aufweist (alles binäre Daten), müssen auch keine Normierungen durchgeführt werden und es kann der Algorithmus direkt gestartet werden (es wird hier die Least-Squares (Fast) Methode angewandt). Das Ergebnis enthält einen Vorschlag, welche Anzahl von Klassen verwendet werden soll (nach dem Cubic Clustering Criterion), man kann jedoch auch eine andere Klassenzahl wählen (falls beispielsweise ein

weiterer Knick in der Kurve zu beobachten ist). Bei diesem Datensatz zeigte sich folgendes Ergebnis:

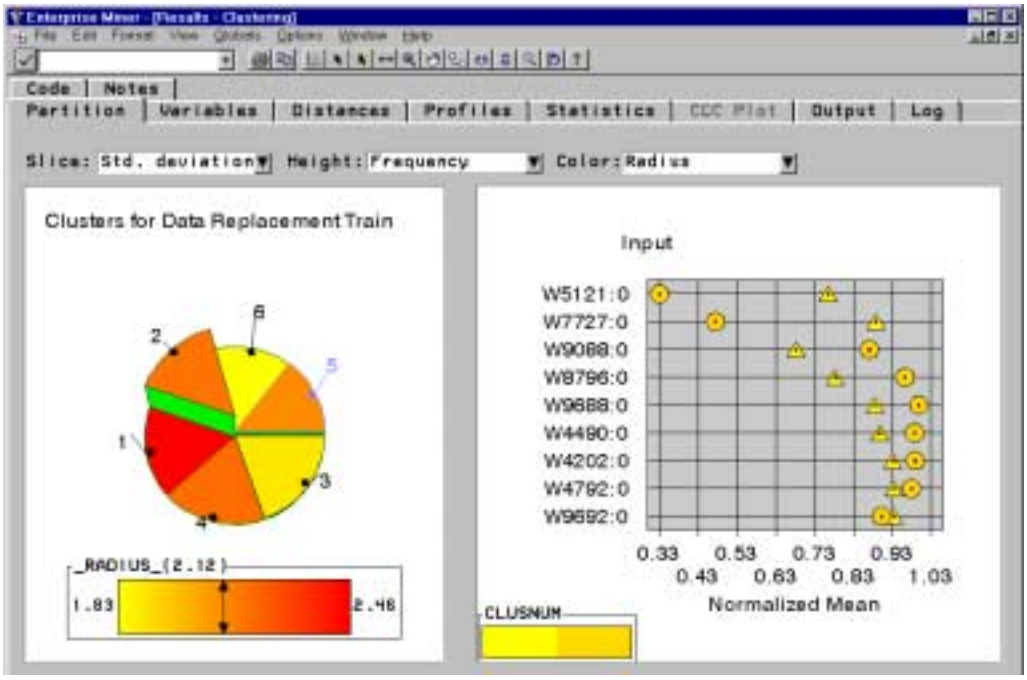


Abbildung 5 : Clustering

Es zeigte sich beispielsweise, dass im Cluster 5 die Produkte 5521 und 7727 ein typisches Produktportfolio bilden.

Anhand der Methodenvielfalt und Datenverarbeitungshilfen, die der SAS Enterprise Miner zur Verfügung stellt, lassen sich sehr effizient Analysen durchführen und dokumentieren (Reporter Knoten). Die angenehme Oberfläche macht das Tool auch Anwendern zugänglich, die noch keine Programmiererfahrung im SAS System haben. Um allerdings gezielte Einstellungen vornehmen zu können oder bei bestimmten Analysen Feintuning betreiben zu wollen (bspw. Clustering) ist die Eingabe der Prozedur/des Makros mit allen gewünschten Optionen unumgänglich.

Fazit

Die durchgeführten Analysen haben gezeigt, daß durch Einsatz von Data Mining Verfahren ein erhebliches Einsparpotenzial bei Direktmarketingaktionen vorhanden ist und darüber hinaus auch wichtiges Wissen aus den Daten über den Kunden und sein Verhalten gewonnen werden kann. Die Analyse lässt sich mit Hilfe des SAS Enterprise Miners in erheblich kürzerer Zeit bei großer Methodenvielfalt durchführen und dokumentieren.