



## STATISTICA Success Stories STATISTICA Data Miner in der analytischen Praxis

StatSoft-Konferenz, 29. November 2007 in Hamburg

**Hintergrund:** Auf unserer Konferenz der Reihe „STATISTICA Success Stories“ am 29. November 2007 in Hamburg berichteten Anwender aus unterschiedlichen Branchen über ihre Praxiserfahrungen. Referenten wie Teilnehmer nutzten die Gelegenheit, sich auszutauschen. In diesem Dokument finden Sie Abstracts der gehaltenen Vorträge aus den Bereichen Medien, Strategisches Marketing, Industrielle Fertigung und Biotechnologie.

**Dr. Gerald Musiol Consulting – Medien**

### Data Mining im interaktiven Fernsehen: Optimierung von Call-In-Shows

Seit einigen Jahren stellen Call-In-Shows, d.h. Sendungen, in denen die Zuschauer per Anruf, SMS oder Internet an Gewinnspielen partizipieren können, eine weitere wichtige Einnahmequelle für eine Vielzahl europäischer und asiatischer Fernsehsender dar.

Durch den zunehmenden Konkurrenzdruck und die veränderten Rahmenbedingungen ist es erforderlich geworden, dass diese Call-In-Shows weiter optimiert werden, um auch in Zukunft das Anrufvolumen stabil halten zu können. Die Referenten stellen an praxisnahen Beispielen einige Möglichkeiten des Data Mining in diesem Kontext vor.

Anhand von Analysen, die mit dem *STATISTICA Data Miner* für einen Fernsehsender in Großbritannien durchgeführt worden sind, werden aber auch mögliche Probleme, die die Umsetzung der Analyseergebnisse im Fernsehprogramm betreffen, diskutiert.

**psychonomics AG – Strategisches Marketing**

### Psychologische Zielgruppensegmentierung durch Verknüpfung von Bestandsdaten und Marktforschungsergebnissen

Gesamtziel des vorgestellten Projektes war eine Optimierung von Dialogmarketing-Maßnahmen für den Gesamtkundenbestand eines deutschlandweit tätigen Versicherungsunternehmens. Im Gesamtprojekt wurden verschiedene Teilziele verfolgt, u.a. a) eine psychologische Segmentierung des Kundenbestandes in verschiedene Kommunikationstypen, b) die Qualifikation der Bestandskunden nach zukünftig geplanten Produktabschlüssen und c) die Qualifikation des Bestandes nach Präferenzen für bestimmte Ansprachekanäle. Sämtliche Informationen zur Qualifikation des Kundenbestandes wurden im CRM-System des Auftraggebers in Form einfach zu interpretierender numerischer Scorewerte hinterlegt und dienen dort zur Steuerung von Prozessen des Dialogmarketings.



Die Sonderstellung des Projektes bestand darin, dass eine Vielzahl der vorherzusagenden Größen nicht unmittelbar aus einer reinen Bestandsdatenanalyse abgeleitet werden konnten. Die notwendigen Informationen wurden daher auf Basis einer repräsentativen Kundenbefragung generiert. Die Antworten der Befragten wurden anhand der vorgegeben Ziele voranalysiert. Im zweiten Schritt erfolgte eine Verknüpfung der Befragungsergebnisse mit den Bestandsdaten der befragten Kunden.

Nach der Bestandsdatenaufbereitung für die eigentlichen Analysen erfolgten unter Einsatz verschiedener Module des *Data Miners* die Entwicklung geeigneter Algorithmen zur Prognose der Zielgrößen (a) bis (c). Die Herausforderung bestand dabei darin, unter Ausschöpfung möglichst vieler Informationen bzw. Informationskombinationen, bei einer gleichzeitig aus Kostengründen begrenzten Stichprobe (n=4.000) sowohl treffsichere als auch stabile Prognosemodelle für die einzelnen Zielgrößen zu generieren.

Der Vortrag zeigt auf, wie sich unter Berücksichtigung des Moduls Feature Selection mit Prüfung von Interaktionseffekten und des Moduls Boosted Trees bei Kombinationen verschiedener multivariater Techniken unter den geschilderten Rahmenbedingungen zufriedenstellende Prognoseergebnisse erzielen lassen.

### DEVK Versicherungen – Strategisches Marketing

#### Kundenauswahl und Erfolgsanalyse im Direktmarketing

Die DEVK Versicherungen setzen seit mehreren Jahren *STATISTICA*-Software im Direktmarketing ein. Die Basisstatistiken geben schnell Auskunft über die Zusammensetzung der Kundengruppen, die für eine geplante Mailingaktion zur Verfügung stehen. Mit den Data-Mining-Techniken wird dann untersucht, ob es Merkmale gibt, mit denen die Abschlusswahrscheinlichkeit für ein spezielles Mailing prognostiziert werden kann. Den einzelnen Merkmalen werden ihrer Trennschärfe nach Werte zugeordnet und diese zu einem Scorewert je Kunde addiert. Die Kunden mit den höchsten Scorewerten werden angeschrieben. Die Abschlussquote liegt bei Einsatz dieses Verfahrens deutlich höher als sonst; wird der "Cut" richtig gesetzt ist sie bis zu 50 Prozent höher.

**„Die schnelle Einsatzbereitschaft und der Betrieb von *STATISTICA Data Miner* durch die Fachabteilung ermöglichte uns ein schlankes Projektvorgehen.“**

*Bruno Küpper, Referent für Kommunikationstechnik im Bereich Kundenbindung und Dialog-Marketing*

Weitere Analysen mit *STATISTICA Data Miner* haben zum Ziel, die Affinität der Kunden zu bestimmten Produkten und Vertriebswegen allgemein zu bestimmen. Für zentrale Aktionen sind diese Informationen immer dann interessant, wenn speziellere Analysen aufgrund von Termindruck oder wegen fehlender Daten nicht möglich sind.

Der Einsatz von *STATISTICA Data Miner* zahlt sich auch in betriebswirtschaftlicher Sicht aus. So konnten kürzlich bei einem Mailing rund 100.000 € durch Reduzierung der Versandmenge eingespart werden. Die Responsemenge lag aber nur unwesentlich unter der eines vergleichbaren Mailings im Vorjahr.

In der Zukunft soll auch den Vertriebspartnern eine Kombination aus Bonitäts- und Affinitätsscore für ihre dezentralen Vertriebsaktivitäten zur Verfügung gestellt werden.



### FactWorks GmbH – Strategisches Marketing

#### Welcher Kunde gehört in welches Segment?

#### Predictive Modeling zur operativen Umsetzung einer strategischen Marktsegmentierung

Ein wesentlicher Erfolgsfaktor für die Akzeptanz und Nutzung strategischer Segmentierungen ist die Klassifikation von Kunden eines Unternehmens in ein entwickeltes Segmentierungsschema. Dabei werden Kunden in der Regel mit den im Data Warehouse zur Verfügung stehenden Informationen mittels Scoring-Algorithmen den verschiedenen Segmenten einer Segmentierung zugeordnet. Darüber hinaus kann es in bestimmten Anwendungsfällen (z. B. Marktuntersuchungen) auch wünschenswert sein, Nichtkunden oder Neukunden in Segmente zu klassifizieren. Da für diese Zielgruppen naturgemäß keine oder nur wenige Daten vorliegen, muss ein Set von möglichst wenigen Fragen, so genannte ‚Magic Questions‘, entwickelt werden, das eine Zuordnung mit ausreichender Genauigkeit ermöglicht.

Der Vortrag zeigt auf:

- wie alternative Klassifikationsmodelle zu bewerten sind (Accuracy ist nicht alles!),
- welche Schwierigkeiten es bei der Klassifikation auf Basis von Data-Warehouse-Informationen insbesondere bei multidimensionalen Segmentierungen (bspw. Attitudes/Needs/Behaviour) gibt,
- wie die Klassifikationsgenauigkeit durch Ergänzung weniger externer Daten deutlich erhöht werden kann,
- wie optimale Sets von ‚Magic Questions‘ für verschiedene Anwendungsfälle entwickelt und implementiert werden können.

### Procter & Gamble Service GmbH – Fertigung

#### Einfluss verborgener Faktoren auf die Produktqualität –

#### Modellierung von Qualitätsdaten in Abhängigkeit von Prozess- und Störgrößen

Ein umfassendes Prozessverständnis ist die Grundlage zur Herstellung von Qualitätsprodukten. Der Einfluss aller Hauptfaktoren ist typischerweise bereits während der Entwicklung komplett bekannt. Schwieriger wird es jedoch wenn man den Einfluss von sekundären Faktoren und Störgrößen, die oftmals erst unter regulären Produktionsbedingungen aktiv werden, bestimmen möchte. Da es meistens nicht möglich ist diese Faktoren experimentell zu untersuchen, muss auf Produktionsdaten zurückgegriffen werden, die sich häufig nur mühsam mit Methoden der klassischen Statistik modellieren lassen. Nicht-parametrische Modellierungsverfahren, wie sie im Data Mining Anwendung finden, sind hier die Methoden der Wahl.

Im vorgetragenen Praxisbeispiel wird gezeigt wie Data Mining mit dem *STATISTICA QC Miner* dazu genutzt werden kann, den Einfluss verborgener Faktoren auf die Produktqualität zu modellieren. Hierbei werden alle Schritte vom Verstehen der Daten bis zur Anwendung der Modelle durch die Ingenieure Schritt für Schritt aufgezeigt.



### Voestalpine Stahl GmbH – Fertigung

#### Anwendung statistischer Methoden zur Klassifizierung von Stahlsorten

In der voestalpine werden verschiedene Stahlsorten erzeugt, die sich durch unterschiedliche Analysezusammensetzungen und unterschiedliche Eigenschaften auszeichnen. Ausgangspunkt für die Klassifizierung der Stahlsorten war die Entwicklung eines Prognosemodells für den Arbeitswalzenverschleiß der Breitbandstraße, wo sich die Bandqualität als signifikanter Einfluss auf den Walzenverschleiß herausgestellt hat.

Die bisher verwendeten Stahlgruppeneinteilungen haben erhebliche Nachteile aufgewiesen, z.B. das Vorkommen von Restklassen. Ausgehend von der Analysezusammensetzung der einzelnen Stahlsorten wurde daher versucht unter Anwendung verschiedenster statistischer Methoden eine neue Klassifizierung vorzunehmen, die einerseits mit den bisherigen Einteilungen in etwa konform bleibt, andererseits die Nachteile der bisherigen Gruppierungen eliminiert.

Zur Bewertung der Ergebnisse der einzelnen Analysen wurde neben der Fehlklassifikationsrate auch auf die Erklärbarkeit und Stabilität der Modelle Rücksicht genommen.

Als beste Lösung hat sich ein interaktiver Entscheidungsbaum herausgestellt, in dem auch metallurgische Hintergründe berücksichtigt wurden. Die neu gewonnene Stahlgruppenklassifizierung lässt sich künftig in Zusammenhang mit den verschiedensten Fragestellungen anwenden.

### INDIVUMED GmbH – Biotechnologie

#### Krebsdiagnostik mit Hilfe von **STATISTICA**: Analyse von Patienten- und Genexpressionsdaten

Die Verknüpfung und Analyse von klinischen Patientendaten mit experimentell gewonnenen molekularen Signaturen von Tumoren eröffnen neue Möglichkeiten für die frühe Diagnose von Krebspatienten. Die Nutzung der Daten im Verbund sollte es außerdem ermöglichen bessere Prognosen über den Krankheitsverlauf der Patienten zu stellen.

Um für die Diagnose geeignete biologische Marker als Werkzeuge in Form von Tumor-spezifischen Genprofilen zu identifizieren, wurden in einer Studie mit Hilfe von *STATISTICA Data Miner* zwei unabhängige Patientenkohorten (Training- und Testsatz) gebildet. Beide Gruppen bestanden jeweils aus 30 Patienten mit gutartigen und bösartigen Erkrankungen, die sich durch hohe Intra- und Inter-Homogenität bezüglich ihrer klinischen Daten auszeichneten.

Von allen Patienten wurden Gewebeproben (Normal- und Tumorgewebe) standardisiert gewonnen und die Expression aller Gene in den Proben mittels Microarray-Technologie gemessen (Affymetrix® Plattform und Chips). Die Qualität und Spezifität der experimentellen Datensätze wurden mit verschiedenen statistischen Verfahren untersucht und Patienten- und Gewebespezifische Gensignale unter Nutzung von *STATISTICA Data Miner* als Plattform analysiert und mit den vorhandenen klinischen Daten abgeglichen. Die Anwendung multivariater Verfahren sowie explorativer Datenanalyse führte zur Identifikation von tumorspezifischen Genen, die sich als Biomarker für die Patientendiagnostik und Krankheitsprognose eignen sollten.