IT Solutions Series

Managing Data Mining
Advice from Experts

Stephan Kudyba, Ph.D.
New Jersey Institute of Technology/Null Sigma Inc., USA
Published in the United States of America by
CyberTech Publishing (an imprint of Idea Group Inc.)
701 E. Chocolate Avenue, Suite 200
Hershey PA 17033 USA
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@idea-group.com
Web site: http://www.idea-group.com

and in the United Kingdom by
CyberTech Publishing (an imprint of Idea Group Inc.)
3 Henrietta Street
Covent Garden
London WC2E 8LU
Tel: 44 20 7240 0856
Fax: 44 20 7379 3313
Web site: http://www.eurospan.co.uk

Copyright © 2004 by Idea Group Inc. All rights reserved. No part of this book may be reproduced in any form or by any means, electronic or mechanical, including photocopying, without written permission from the publisher.

Library of Congress Cataloging-in-Publication Data

eISBN

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.
IT Solutions Series: 
Managing Data Mining
Advice from Experts

Preface ............................................................................................................................ vi

Chapter I: Introduction: Data Mining and the World of 
Commerce: A Background of Data Mining, Business 
Intelligence and Business Process Management ............... 1
Stephan Kudyba, PhD, New Jersey Institute of Technology/
Null Sigma Inc., USA

Chapter II: Data Mining and the Banking Sector: Managing 
Risk in Lending and Credit Card Activities ..................... 17
Ákos Felsóvályi, Citigroup, USA
Jennifer Courant, Citigroup, USA

Chapter III: Credit Scoring and Risk Management for Small 
Business Lending .................................................................................................. 38
Vernon Gerety, PhD, PredictiveMetrics Inc., USA
CHAPTER IV: THE UTILIZATION OF BUSINESS INTELLIGENCE AND DATA MINING IN THE INSURANCE MARKETPLACE .......... 77
Jeffrey Hoffman, Chubb & Sons Insurance, USA

CHAPTER V: SALES FORCE MANAGEMENT IN THE CONSUMER PRODUCTS INDUSTRY ........................................................ 102
Navin Sharma, PhD, MarketWell Inc., USA

CHAPTER VI: THE HEALTHCARE COST DILEMMA: WHAT HEALTH INSURANCE COMPANIES CAN DO TO MITIGATE UNSUSTAINABLE PREMIUM INCREASES ................................. 115
Russ Danstrom, Anthem Blue Cross Blue Shield, USA
Jeff Nicola, Anthem Blue Cross Blue Shield, USA

CHAPTER VII: THE APPLICATION OF DATA MINING TECHNIQUES IN HEALTH PLAN POPULATION MANAGEMENT: A DISEASE MANAGEMENT APPROACH .................................................. 135
Theodore L. Perry, PhD, American Healthways Corp., USA
Travis Tucker, American Healthways Corp., USA
Laurel R. Hudson, American Healthways Corp., USA
William Gandy, American Healthways Corp., USA
Amy L. Neftzger, American Healthways Corp., USA
Guy B. Hamar, American Healthways Corp., USA

CHAPTER VIII: TOILING IN THE MEDIA DATA MINES ...................... 154
Robert Young, PHD Canada, Omnicom Group Inc., Canada

CHAPTER IX: AUDIENCE MEASUREMENT APPLICATIONS FOR ONLINE ADVERTISING AND E-COMMERCE ........................................ 175
David Martin, Nielsen//NetRatings, USA
CHAPTER X: TAKING A CLOSE LOOK AT THE EVOLVING UTILITIES

INDUSTRY: FACTORS THAT DRIVE VOLATILITY AND

METHODS TO HELP MANAGE IT ............................................. 215
Nicholas Galletti, Con Edison Energy, USA

ABOUT THE EDITOR ........................................................................ 230

ABOUT THE AUTHORS ..................................................................... 231

INDEX ........................................................................................ 235
Foreword

Dr. Jim Goodnight
CEO & President, SAS

In response to increasingly high customer demands and formidable competition, today’s organizations have implemented new technologies that generate and collect massive volumes of data. However, a large majority of the data collected goes to waste or becomes obsolete. Inconsistency, repetition, time-sensitivity, disparate data sources and data overkill are just a few of the problems overwhelming many decision makers.

Fortunately, the need for digestible and useful data has created a new generation of business intelligence technologies. This includes software that sifts through mountain ranges of data until the most useful nuggets of information are extracted and identified, much the same way that California prospectors of the 19th century refined their techniques to strike gold more quickly.

This new technology, commonly known as Data Mining, has taken on a life of its own and continues to evolve.

Data Mining is the science of revealing useful patterns and relationships in vast data stores. Data Mining — considered by many as the new name for statistics — draws on many disciplines. With statistics as its foundation, it incorporates computer science, machine learning, artificial intelligence and other specialties. Data Mining ultimately provides a
framework for dealing with uncertainty. As organizations and the global economy become more complex, sources of uncertainty become more plentiful. To make decisions confidently, new and sophisticated approaches are required.

The vast quantities of data that businesses produce, gather and store fuel the demand for more effective means of deriving value from them. Technologies that distill value from data have been available for some time. We have long been able to ask simple questions of the data: What were sales last quarter? How did the western region do compared to the eastern region? This book is not about such query and reporting or OLAP. These tools, while necessary, do not generate the big returns, nor do they provide answers to the really challenging questions. Business managers have often shied away from high-end data analysis, perhaps fearing it to be overly complex. With ever-increasing quantities of data, and lots of questions needing better answers, a growing number of decision makers want to know what Data Mining can do for them.

Today’s powerful multivariate modeling techniques provide answers to the hard questions: Which patterns and relationships are truly significant and which are merely chance? What are the really important factors affecting quality and sales? Which customers are likely to leave? Why and when? Given the probabilities of default, what is the exposure for a given portfolio of loans? Analytical rigor is required to effectively address the wide range of growing business needs, such as assessing risk more accurately, improving price optimization, and anticipating customer demand and preferences. Such applications of Data Mining are increasingly possible—and valuable—in business, as the examples in this book clearly demonstrate.

Data Mining delivers on the promise of helping executives make, save and spend money more effectively. Enlightened decision makers now enjoy new peace of mind. They can chart their courses more strategically and with greater precision over longer periods, rather than simply reacting,
in fire-drill fashion, to either tactical or strategic issues that flare up. Savvy executives harness the power of Data Mining in new areas of business for superior performance and sustained competitive advantage. No longer is high-end data analysis considered a rare, specialized activity with limited use in business. Instead, corporate survival depends on how well Data Mining is applied to vast and growing data stores.

Data Mining has led to highly refined processes, especially where the stakes are high. Minimizing the downside often piques management’s interest first — reducing write-offs, losses and fraud — since these improvements are typically easy to measure. Significant returns can also come from the upside where the potential may be less obvious — increasing customer response, sales and market share. Success breeds success, and the interest in high-end data analysis continues to grow. Only high-end data analysis keeps delivering big returns across increasing areas of business; ROI in excess of 1,000% is not uncommon. Data Mining has become integral to many aspects of business with returns that are significant, yet often incalculable.

In today’s data-driven economy, Data Mining is an essential tool in the pursuit of enhanced productivity, reduced uncertainty, delighted customers, mitigated risk, maximized returns, refined processes and optimally allocated resources. To thrive, performance must be measured along the way so that the factors that contributed to success — or failure — can be understood. By applying a more scientific approach to business decision making, the stage is set for continuous learning and improvement.

This book highlights the value of making sound decisions in the face of uncertainty. Real-world examples span industry sectors and illustrate how Data Mining drives optimal decision making for a variety of business processes. The business leaders and experts contributing to this book provide unique insights on a variety of business processes, challenges, solutions and best practices. As Data Mining becomes more ubiquitous in
business, let this book serve as an invitation to reap the benefits for those who have not yet begun. And for those who have already embarked, may this book allow greater success.

Dr. James H. Goodnight is CEO, chairman, co-founder and president of SAS Institute, the world’s largest privately held software company. Chief executive since the company’s incorporation in 1976, Goodnight continues to focus on strategic planning for the global business, which provides software and services that enable customers to transform data from all areas of their business into intelligence. An accomplished programmer, Goodnight has authored many of the procedures that comprise SAS® software.

SAS passed the $1 billion revenue mark in 1999 and revenues continue to grow. According to Goodnight, the key to the company’s success has been its ability to listen to more than 3.5 million software users and respond to their needs. SAS customers represent numerous industries and can be found in more than 100 countries. More than 200 SAS offices around the globe support this large customer base. SAS responds to customer needs by staying near the top of the software industry in the percentage of revenue reinvested in research and development, devoting over a quarter of total revenue to R&D.

In addition to this significant investment in technology, Goodnight also invests in people — SAS employees and their families. The company’s work environment is designed to nurture and encourage creativity, innovation and quality. Since the early 1980s, Goodnight has supported on-site child care, health care and recreation and fitness centers. His commitment to these progressive work-life
programs has earned SAS national recognition in publications such as *The Wall Street Journal* as well as *Fortune, Fast Company, Business Week* and *Working Mother* magazines.

Goodnight’s passion, and the focus of SAS’ philanthropic efforts, is education. In August 1997, the doors opened at Cary Academy, an independent college preparatory day school for students in grades 6 through 12. Goodnight co-founded the school in 1996 as a model school — one that integrates technology into all facets of education. Shortly after Cary Academy opened, SAS launched SAS inSchool, which develops content-based educational software that is helping move schools into the next millennium. The software contains the framework for a new generation of teaching courseware that will further the use of technology as a learning tool.

A native of Wilmington, N.C., Goodnight holds bachelor’s and master’s degrees as well as a doctorate in statistics from North Carolina State University. He served on the faculty of NCSU from 1972 to 1976, and continues to serve as an adjunct professor. Goodnight is a Fellow of the American Statistical Association, and has authored numerous papers on statistical computing.
The term Data Mining continues to be an elusive concept to many business managers as it is often imbedded in the idea of searching through vast amounts of data records that reside in organizations across industry sectors. The purpose of this book is to help reduce the uncertainty as to what data mining really is and more importantly, illustrate how prominent organizations incorporate it in order to enhance their operational efficiency. Without going into great detail at this stage, data mining can help augment operational efficiency or enable decision makers to better manage business processes by empowering users to gain a greater understanding of the critical factors that drive corresponding operational activities. Through the combination of data, data management, business modeling and finally applying mining methodologies such as logit regression, segmentation or neural networks, managers can identify not only which variables impact such operational measures such as Sales, Customer Response Rates, Default Propensities but also estimate a quantifiable connection between them. The resulting models provide decision makers with ability to perform “what if” simulations or simply forecast into the future.

Data Mining is not limited to just a few prominent business applications but can provide a value add to a variety of operational activities of organizations across industry sectors. The information in the following chapters will provide a much clearer understanding of some prominent business processes in which it can be utilized as a strategic component to
enhancing productivity. More specifically, Chapter I provides a more detailed description of what data mining is, what is entailed in conducting a robust mining analysis, why mining has become so much more popular over the past decade, what are some prevalent business processes that mining is utilized and how Data Mining complements and augments existing strategic initiatives such as Six Sigma.

Chapter II entails our first detailed analysis of a prominent business process as it addresses how mining and multivariate modeling can help augment the decision making process in the world of finance, more specifically, how decision makers can better manage the risk associated with lending activities. The chapter was written by two modeling experts from Citigroup who describe the activities entailed in generating multivariate models to help measure risk of both corporate and consumer borrowers. This high level concept is addressed in more detail in Chapter III, where a leading Data Mining consultant describes the process by which quantitative modeling techniques are used to help manage risk in lending to small businesses. The section provides an overview of the lending industry, inherent risks that are involved in it and the types of models (Credit Scoring, New Application Scoring and Behavior Scoring) that can reduce risk. It ends off with a small case study which helps drive home the concepts highlighted in the section.

The book then turns its focus towards the realm of Customer Relationship Management, initially in the Insurance industry. Chapter IV is written by a senior business strategist at Chubb Insurance. This section describes key concepts that are involved in applying analytics to better understand customer behavior and preferences and how to best make a connection with them through effectively managing the sales force. Through the effective incorporation of Data, Data Mining & BI, along with sound management policies, organizations can better understand the various needs that correspond to particular consumer groups and with this
information, can better direct sales force initiatives in order to make a
clearer connection with them. Chapter V extends the description of sales
force management but takes a more detailed look at the characteristics of
sales representatives. An expert consultant in the field of Data Mining
provides a thorough illustration of how to utilize detailed data that describe
the activities of sales representatives along with leading quantitative tech-
niques in order to better identify the type of sales reps that are likely to
perform well in selling a company’s products or services.

Chapter VI diverges from sales force applications but continues to
address Customer Relationship Management and organizational operating
efficiency. Two experts from one of the largest Health Insurance compa-
nies (Blue Cross Blue Shield) describe the process by which firms can
incorporate Data Mining and the “high-level” strategic methodology of Six
Sigma to help manage product and service costs and corresponding prices
in a rising cost industry. Chapter VII continues to focus on the Health Care
industry and Customer Relationship Management but concentrates on a
different type of business application and consumer. A group of Health
Care experts from American Healthways Corp. describe the process by
which decision makers can utilize the power of Data Mining to enhance the
process of managing potential diseases/illnesses that exist in patients. With
this information, health care providers can better allocate resources to
reduce potential development of chronic illnesses and better manage
overall operational costs.

Chapter VIII takes on a whole new focus as it addresses the world of
advertising. A senior partner of a subsidiary of one of the worlds largest
advertising organizations (Omnicom Group Inc.) offers some expert
insights on how Data Mining and Econometric modeling can better estimate
the returns to different strategic initiatives involved in the advertising
industry. These include such topics as promotional effectiveness, medium
effectiveness and brand awareness. We all know that an organization can
have one of the best products on the market, but if no one is aware of it, its sales may suffer. Our next chapter extends the analysis of advertising but shifts towards the world of e-commerce. An expert analyst at Neilsen’s/Netratings provides some background on the Internet marketplace and then addresses such topics as online audience measurement, tracking advertising activities and then describes a model of analyzing consumer online activities. The section then provides a few case studies to drive home concepts more clearly.

Our last chapter delves into the complex task of managing operational risk in an evolving and volatile industry. The recently deregulated Utilities industry, which involves the process of supplying power to a variety of consumers, many times incorporates sophisticated quantitative methods in order to help mitigate the uncertainty in allocating enough power to the marketplace. An expert from Con Edison provides some insightful background information describing the evolution of the industry and then illustrates how decision makers can use quantitative techniques to better manage such variables as price, supply and demand of electricity over various time horizons. This last chapter concludes the content for “Data Mining: Advice from Experts” but highlights an important point to remember regarding the utilization and power of Data Mining techniques, and that is, the quantitative (mathematical and algorithmic) methodologies that comprise this analytic space can be applied to a variety of business applications. Through the use of Data Mining, decision makers can help reduce the uncertainty as to what drives particular business processes, and with this enhanced knowledge, can more efficiently allocate available resources to bring a good or service to market.
Chapter I

Data Mining and the World of Commerce

Stephan Kudyba, PhD
New Jersey Institute of Technology/Null Sigma Inc., USA

Introduction

Despite the research written, the software developed and the business applications that can be enhanced by it, the terms data mining and multivariate modeling continue to stoke uncertainty, complexity and sometimes fear in business managers and strategic decision-makers across industry sectors. Why is this? There are a number of reasons to cite, but probably the most common involves the complex nature of the methodologies incorporated in this analytic technique. The complexity we refer to involves the use of mathematical equations, sophisticated algorithms and advanced search and query techniques, not to mention statistical applications that are utilized in analyzing data. If that is not enough to throw management back on their heels, how about data acquisition, normalization, and model optimization, which are often involved in the process? Let’s add one more attribute to the list, and that is the ability to not only understand these complex methods, but more importantly, to understand when and where they can be used to enhance operational efficiency. Now
is there any wonder why data mining continues to be this mysterious phenomenon in the world of commerce? No doubt; however, to dispel some of the uncertainties regarding this issue, the following book will provide the reader with expert input on how these quantitative methods are being used in prominent organizations in a variety of industry sectors to help enhance productivity, efficiency and to some extent, profitability. Before we get into the details of the applied material, the following chapter will provide some general information on what data mining and multivariate modeling is, where it came from, and how it can be used in a corporate setting to enhance operational efficiency.

Data Mining and Multivariate Modeling
(What Is It?)

Before we begin a book entitled IT Solutions: Data Mining, Advice from Experts, we will need to provide some basic background of what the term data mining refers to. Without committing to a formal definition, let’s initially focus on the big picture. In order to do this we need to begin with the first term, data, a word that should not be underestimated. Given the evolution of the Information Economy, where innovations in information technology have facilitated the ability to store, manipulate, organize and generally process data, organizations have increasingly become aware of the valuable information that is available to them. Data exist both within a given organization (e.g., operational, customer activity based) and also outside the boundaries of corporate entities (e.g., industry and market specific descriptive data). By combining elements of these resources and analyzing data variables with appropriate methodologies, companies can increase their understanding of how effectively their strategic initiatives perform in the market in which they operate.
This previous sentence, which mentions “analyzing with appropriate methodologies”, provides a logical lead into the overall terms of data mining and multivariate modeling. These concepts involve the processing of data that have been collected, normalized, corrected and organized, which incorporate key variables that describe the essential underpinnings of a business process. Processing generally refers to the application of mathematical equations or algorithms, sometimes in conjunction with statistical verification techniques that help identify relationships between these variables. Prominent data mining and multivariate modeling techniques include:

- Regression
- Logistic Regression
- Clustering
- Segmentation Classification
- Neural Networks
- Affinity Analysis

The quantitative and statistical techniques incorporated in these methodologies help identify reliable relationships that may exist between the variables that describe a business process. With this information, managers increase their understanding of how certain variables generally relate to each other in describing a business process, and the resulting models also empower them to perform simulations (e.g., “what-ifs”) to gain a greater understanding of future expectations of business performance measures given a set of strategic variable inputs.

**Business Intelligence Technology**

**(A Complement to Data Mining)**

A complementary data processing activity to the data mining spectrum involves the activity of querying databases and organizing data variables
into report form or into cubes of information that can be analyzed in a basic context. OLAP (Online Analytic Processing) provides users with the ability to view variables that describe a business process and perform data manipulation to view variables according to different formats in a numeric and graphical context. OLAP cubes give users a preliminary look at what drives particular business processes across an organization, as it presents a multidimensional view of the various factors that explain a process. What do we mean by a multidimensional view? Well, we can almost compare the term “multidimensional” with “multivariate”, mentioned above. Simply put, multidimensional refers to the variety of characteristics or variables that describe/explain the operational activities of an organization.

For example, if you are a sales executive and wish to gain a greater understanding of your sales activities over the past year, you could access an OLAP-based “sales cube”. This cube of information gives the user the ability to analyze sales performance measures such as sales/revenue, unit sales, profit metrics, and so forth, corresponding to descriptive variables of their sales activities such as particular products or product lines in a particular region (East Coast) over a particular time period (quarterly, monthly), in which each variable mentioned is a dimension. The cube enables users to quickly change dimensions of the analysis with just a click of a mouse (see Figure 1).

![Figure 1](image-url)

**Figure 1**
The process of creating a cube from data in a data warehouse or mart to enable users to analyze a particular operational activity of a firm is a major technological component of the business intelligence spectrum and is addressed in a number of chapters throughout this book.

Growing Utilization of Data Mining

Despite the power complementary business intelligence methodologies such as OLAP provide for business users, analysts and managers at all levels often need to go beyond gaining a greater understanding of what their operational activities have looked like in the past. They need to answer more complex questions such as, “What can I expect in the future if I take specific strategic actions?” As we mentioned, because of the mathematical and quantitative techniques incorporated by data mining, users can build models that can help them answer these “what if” scenarios. But when considering the more prominent reasons behind the increase in popularity and utilization of data mining over the past few years, one needs to turn to some factors that are related to the general evolution of the information economy.

Some of the prominent reasons why data mining has become a common strategic tool of corporations across industry sectors include such areas as:

- Enhancements in computer processing and increased ease of use of data mining software technology.
- The ability to capture and store data that describe the operational activities of organizations.
- The pulse of business process management which entails strategic utilization of available resources to enhance operational efficiency.