STATISTICAL MODELING and ANALYSIS for DATABASE MARKETING

Effective Techniques for Mining Big Data

Bruce Ratner

CHAPMAN & HALL/CRC

A CRC Press Company

Boca Raton London New York Washington, D.C.

Library of Congress Cataloging-in-Publication Data

Ratner, Bruce.
Statistical modeling and analysis for database marketing : effective techniques for mining big data / Bruce Ratner.
p. cm.
Includes bibliographical references and index.
ISBN 1-57444-344-5 (alk. paper)
1. Database marketing--Statistical methods. I. Title

HF5415.126.R38 2003
658.8′--dc21 2002041714

Direct all inquiries to CRC Press LLC, 2000 N.W. Corporate Blvd., Boca Raton, Florida 33431.

Visit the CRC Press Web site at www.crcpress.com

International Standard Book Number 1-57444-344-5
Library of Congress Card Number 2002041714
Printed in the United States of America 1 2 3 4 5 6 7 8 9 0
Printed on acid-free paper

Dedication

This book is dedicated to:

My father Isaac — my role model who taught me by doing, not saying.

My mother Leah — my friend who taught me to love love and hate hate.

My daughter Amanda — my crowning and most significant result.

Preface

This book is a compilation of essays that offer succinct, specific methods for solving the most commonly experienced problems in database marketing. The common theme among these essays is to address each methodology and assign its application to a specific type of problem. To better ground the reader, I spend considerable time discussing the basic methodologies of database analysis and modeling. While this type of overview has been attempted before, my approach offers a truly nitty-gritty, step-by-step approach that both tyros and experts in the field can enjoy playing with. The job of the data analyst is overwhelmingly to predict and explain the result of the target variable, such as RESPONSE or PROFIT. Within that task, the target variable is either a binary variable (RESPONSE is one such example) or a continuous variable (of which PROFIT is a good example). The scope of this book is purposely limited to dependency models, for which the target variable is often referred to as the "left-hand" side of an equation, and the variables that predict and/or explain the target variable is the "right-hand" side. This is in contrast to interdependency models that have no left- or right-hand side, and are not covered in this book. Since interdependency models comprise a minimal proportion of the analyst's work load, the author humbly suggests that the focus of this book will prove utilitarian.

Therefore, these essays have been organized in the following fashion. To provide a springboard into more esoteric methodologies, Chapter 2 covers the correlation coefficient. While reviewing the correlation coefficient, I bring to light several issues that many are unfamiliar with, as well as introducing two useful methods for variable assessment. In Chapter 3, I deal with logistic regression, a classification technique familiar to everyone, yet in this book, one that serves as the underlying rationale for a case study in building a response model for an investment product. In doing so, I introduce a variety of new data mining techniques. The continuous side of this target variable is covered in Chapter 4. Chapters 5 and 6 focus on the regression coefficient, and offer several common misinterpretations of the concept that point to the weaknesses in the method. Thus, in Chapter 7, I offer an alternative measure — the predictive contribution coefficient — which offers greater utility than the standardized coefficient.

Up to this juncture, I have dealt solely with the variables in a model. Beginning with Chapter 8, I demonstrate how to increase a model's predictive power beyond that provided by its variable components. This is accomplished by creating an interaction variable, which is the product of two or more component variables. To test the significance of my interaction variable, I make what I feel to be a compelling case for a rather unconventional use of CHAID. Creative use of well-known techniques is further carried out in Chapter 9, where I solve the problem of market segment classification modeling using not only logistic regression but CHAID as well. In Chapter 10, CHAID is yet again utilized in a somewhat unconventional manner — as a method for filling in missing values in one's data. In order to bring an interesting real-life problem into the picture, I wrote Chapter 11 to describe profiling techniques for the database marketer who wants a method for identifying his or her best customers. The benefits of the predictive profiling approach is demonstrated and expanded to a discussion of look-alike profiling.

I take a detour in Chapter 12 to discuss how database marketers assess the accuracy of the models. Three concepts of model assessment are discussed — the traditional decile analysis, as well as two additional concepts I introduce: precision and separability. Continuing in this mode, Chapter 13 points to the weaknesses in the way decile analysis is used, and instead offers a new approach known as the bootstrap for measuring the efficiency of database models. Chapter 14 offers a pair of graphics or visual displays that have value beyond the commonly used exploratory phase of analysis. In this chapter, I demonstrate the hitherto untapped potential for visual displays to describe the functionality of the final model once it has been implemented for prediction.

With the discussions described above behind us, we are ready to venture to new ground. In Chapter 1, I talked about statistics and machine learning, and I defined that statistical learning is the ability to solve statistical problems using nonstatistical machine learning. GenIQ is now presented in Chapter 15 as such a nonstatistical machine learning model. Moreover, in Chapter 16, GenIQ serves as an effective method for finding the best possible subset of variables for a model. Since GenIQ has no coefficients — and coefficients are the paradigm for prediction — Chapter 17 presents a method for calculating a quasi-regression coefficient, thereby providing a reliable, assumption-free alternative to the regression coefficient. Such an alternative provides a frame of reference for evaluating and using coefficient-free models, thus allowing the data analyst a comfort level for exploring new ideas, such as GenIQ.

About the Author

Bruce Ratner, Ph.D., is President and Founder of DM STAT-1 CONSULTING, the leading firm for analysis and modeling in the database marketing industry, specializing in statistical methods and knowledge discovery and data mining tools. Since 1986, Bruce has applied his expertise in the areas of marketing research, banking, insurance, finance, retail, telecommunications, mass and direct advertising, business-to-business, catalog marketing, E-commerce and Web-mining.

Bruce is active in the database marketing community as the instructor of the advanced statistics course sponsored by the Direct Marketing Association, and as a frequent speaker at industry conferences. Bruce is the author of the *DM STAT-1 Newsletter* on the Web, and many articles on modeling techniques and software tools. He is a co-author of the popular text book *The New Direct Marketing*, and is on the editorial board of *The Journal of Database Marketing*.

Bruce holds a doctorate in mathematics and statistics, with a concentration in multivariate statistics and response model simulation. His research interests include developing hybrid modeling techniques, which combine traditional statistics and machine learning methods. He holds a patent for a unique application in solving the two-group classification problem with genetic programming.

Acknowledgment

This book like all books — except the Bible — was written with the assistance of others. First and foremost, I acknowledge Hashem who has kept me alive, sustained me, and brought me to this season.

I am significantly grateful to my friend and personal editor Lynda Spiegel, who untwisted my sentences into free-flowing statistical prose. I am appreciative of the hard work of Paul Moskowitz, who put the tables and figures into a useable format.

I am indebted to the staff of CRC Press for their excellent work (in alphabetical order): Gerry Jaffe, Project Editor; Allyson Kline, Proofreader; Suzanne Lassandro, Copy Editor; Shayna Murry, Designer; Rich O'Hanley, Publisher; Will Palmer, Prepress Technician; Pat Roberson, Project Coordinator; and Carol Shields, Typesetter.

Last and assuredly not least, it is with pride that I acknowledge my daughter Amanda's finishing touch by providing the artwork image for the book cover.

Contents

1

Introduction

Whatever you are able to do with your might, do it.

— Koheles 9:10

1.1 The Personal Computer and Statistics

The personal computer (PC) has changed everything — both for better and for worse — in the world of statistics. It can effortlessly produce precise calculations, eliminating the computational burden associated with statistics; one need only provide the right questions. With the minimal knowledge required to program it, which entails telling it where the input data reside, which statistical procedures and calculations are desired, and where the output should go, tasks such as testing and analysis, the tabulation of raw data into summary measures, as well as many others are fairly rote. The PC has advanced statistical thinking in the decision making process, as evidenced by visual displays, such as bar charts and line graphs, animated three-dimensional rotating plots, and interactive marketing models found in management presentations. The PC also facilitates support documentation, which includes the calculations for measures such as the mean profitability across market segments from a marketing database; statistical output is copied from the statistical software, then pasted into the presentation application. Interpreting the output and drawing conclusions still require human intervention.

Unfortunately, the confluence of the PC and the world of statistics have turned generalists with minimal statistical backgrounds into quasi-statisticians, and affords them a false sense of confidence because they can now produce statistical output. For instance, calculating the mean profitability is standard fare in business. However, mean profitability is not a valid summary measure if the individual profit values are not bell-shaped; this is not uncommon in marketing databases. The quasi-statistician would doubtless not know to check this assumption, thus rendering the interpretation of the mean value questionable.

Another example of how the PC fosters a quick and dirty approach to statistical analysis can be found in the ubiquitous correlation coefficient, which is the measure of association between two variables and is second in popularity to the mean as a summary measure. There is an assumption (which is the underlying straight-line relationship between the two variables) that must be met for the proper interpretation of the correlation coefficient. Rare is the quasi-statistician who is actually aware of the assumption. Meanwhile, well-trained statisticians often do not check this assumption, a habit developed by the uncritical use of statistics with the PC.

The professional statistician has also been empowered by the PC's computational strength; without it, the natural seven-step cycle of statistical analysis would not be practical. [1] The PC and the analytical cycle comprise the perfect pairing as long as the steps are followed in order and the information obtained from one step is used in the next step. Unfortunately, statisticians are human and succumb to taking shortcuts through the seven-step cycle. They ignore the cycle and focus solely on the sixth step listed below. However, careful statistical endeavor requires additional procedures, as described in the seven-step cycle[1] that follows:

1. *Definition of the problem:* Determining the best way to tackle the problem is not always obvious. Management objectives are often expressed qualitatively, in which case the selection of the outcome or target (dependent) variable is subjectively biased. When the objectives are clearly stated, the appropriate dependent variable is often not available, in which case a surrogate must be used.
2. *Determining technique*: The technique first selected is often the one with which the data analyst is most comfortable; it is not necessarily the best technique for solving the problem.
3. *Use of competing techniques:* Applying alternative techniques increases the odds that a thorough analysis is conducted.
4. *Rough comparisons of efficacy:* Comparing variability of results across techniques can suggest additional techniques or the deletion of alternative techniques.
5. *Comparison in terms of a precise (and thereby inadequate) criterion:* Explicit criterion is difficult to define; therefore, precise surrogates are often used.
6. *Optimization in terms of a precise, and similarly inadequate criterion:* Explicit criterion is difficult to define; therefore, precise surrogates are often used.
7. *Comparison in terms of several optimization criteria:* This constitutes the final step in determining the best solution.

[1] The seven steps are Tukey's. The annotations are the author's.

The founding fathers of classical statistics — Karl Pearson and Sir Ronald Fisher — would have delighted in the PC's ability to free them from time-consuming empirical validations of their concepts. Pearson, whose contributions include regression analysis, the correlation coefficient, the standard deviation (a term he coined), and the chi-square test of statistical significance, would have likely developed even more concepts with the free time afforded by the PC. One can further speculate that the PC's functionality would have allowed Fisher's methods of maximum likelihood estimation, hypothesis testing, and analysis of variance to have immediate, practical applications.

The PC took the classical statistics of Pearson and Fisher from their theoretical blackboards into the practical classroom and boardroom.[2,3] In the 1970s statisticians were starting to acknowledge that their methodologies had potential for wider applications. First, a computing device was required to perform any multivariate statistical analysis with an acceptable accuracy and within a reasonable time frame. Although techniques had been developed for a *small-data setting* consisting of one or two handfuls of variables and up to hundreds of records, the tabulation of data was computationally demanding, and almost insurmountable. With the inception of the microprocessor in the mid-1970s, statisticians immediately found their computing device in the PC. As the PC grew in its capacity to store bigger data and perform knotty calculations at greater speeds, statisticians started replacing hand-held calculators with desktop PCs in the classrooms. From the 1990s to the present, the PC offered statisticians advantages that were imponderable decades earlier.

1.2 Statistics and Data Analysis

As early as in 1957, Roy believed that the classical statistical analysis was to a large extent likely to be supplanted by assumption-free, nonparametric approaches, which were more realistic and meaningful. [2] It was an onerous task to understand the robustness of the classical (parametric) techniques to violations of the restrictive and unrealistic assumptions underlying their use. In practical applications, the primary assumption of "a random sample from a multivariate normal population" is virtually untenable. The effects of violating this assumption and additional model-specific assumptions (e.g., linearity between predictor and dependent variables, constant variance among errors, and uncorrelated errors) are difficult to determine with any exactitude. It is difficult to encourage the use of the statistical techniques, given that their limitations are not fully understood.

[2] Karl Person (1900s) contributions include regression analysis, the correlation coefficient, and the chi-square test of statistical significance. He coined the term 'standard deviation' in 1893.
[3] Sir Ronald Fisher (1920s) invented the methods of maximum likelihood estimation, hypothesis testing, and analysis of variance.

In 1962, in his influential paper *The Future of Statistics*, John Tukey expressed concern that the field of statistics was not advancing. [1] He felt there was too much focus on the mathematics of statistics and not enough on the analysis of data, and predicted a movement to unlock the rigidities that characterize the discipline. In an act of statistical heresy, Tukey took the first step toward revolutionizing statistics by referring to himself not as a statistician, but a data analyst. However, it was not until the publication of his seminal masterpiece *Exploratory Data Analysis* in 1977 when Tukey lead the discipline away from the rigors of statistical inference into a new area, known as EDA. [3] For his part, Tukey tried to advance EDA as a separate and distinct discipline from statistics, an idea that is not universally accepted today. EDA offered a fresh, assumption-free, nonparametric approach to problem solving, in which the analysis is guided by the data itself, and utilizes self-educating techniques, such as iteratively testing and modifying the analysis as the evaluation of feedback, to improve the final analysis for reliable results.

The essence of EDA is best described in Tukey's own words: "Exploratory data analysis is detective work — numerical detective work — or counting detective work — or graphical detective work… [It is] about looking at data to see what it seems to say. It concentrates on simple arithmetic and easy-to-draw pictures. It regards whatever appearances we have recognized as partial descriptions, and tries to look beneath them for new insights." [3, page 1.] EDA includes the following characteristics:

1. *Flexibility* — techniques with greater flexibility to delve into the data
2. *Practicality* — advice for procedures of analyzing data
3. *Innovation* — techniques for interpreting results
4. *Universality* — use all of statistics that apply to analyzing data
5. *Simplicity* – above all, the belief that simplicity is the golden rule

On a personal note, when I learned that Tukey preferred to be called a data analyst, I felt both validated and liberated because many of my own analyses fell outside the realm of the classical statistical framework. Furthermore, I had virtually eliminated the mathematical machinery, such as the calculus of maximum likelihood. In homage to Tukey, I will use the terms data analyst and data analysis rather than statistical analysis and statistician throughout the book.

1.3 EDA

Tukey's book is more than a collection of new and creative rules and operations; it defines EDA as a discipline that holds that data analysts fail only if they fail to try many things. It further espouses the belief that data analysts are especially successful if their detective work forces them to notice the unexpected. In other words, the philosophy of EDA is a trinity of *attitude* and *flexibility* to do whatever it takes to refine the analysis, and *sharp-sightedness* to observe the unexpected when it does appear. EDA is thus a self-propagating theory; each data analyst adds his or her own contribution, thereby contributing to the discipline, as I hope to accomplish with this book.

The sharp-sightedness of EDA warrants more attention, as it is a very important feature of the EDA approach. The data analyst should be a keen observer of those indicators that are capable of being dealt with successfully, and use them to paint an analytical picture of the data. In addition to the ever-ready visual graphical displays as an *indicator* of what the data reveal, there are numerical indicators, such as counts, percentages, averages and the other classical descriptive statistics (e.g., standard deviation, minimum, maximum and missing values). The data analyst's personal judgment and interpretation of indictors are not considered a bad thing, as the goal is to draw informal inferences, rather than those statistically significance inferences that are the hallmark of statistical formality.

In addition to visual and numerical indicators, there are the *indirect messages* in the data that force the data analyst to take notice, prompting responses such as "the data look like...," or, "it appears to be... ." Indirect messages may be vague; but their importance is to help the data analyst draw informal inferences. Thus, indicators do not include any of the hard statistical apparatus, such as confidence limits, significance test, or standard errors.

With EDA, a new trend in statistics was born. Tukey and Mosteller quickly followed up in 1977 with the second EDA book (commonly referred to EDA II), *Data Analysis and Regression*, which recasts the basics of classical inferential procedures of data analysis and regression as an assumption-free, nonparametic approach guided by "(a) a sequence of philosophical attitudes... for effective data analysis, and (b) a flow of useful and adaptable techniques that make it possible to put these attitudes to work." [4, page vii.]

Hoaglin, Mosteller and Tukey in 1983 succeeded in advancing EDA with *Understanding Robust and Exploratory Analysis*, which provides an understanding of how badly the classical methods behave when their restrictive assumptions do not hold, and offers alternative robust and exploratory methods to broaden the effectiveness of statistical analysis. [5] It includes a collection of methods to cope with data in an informal way, guiding the identification of data structures relatively quickly and easily, and trading off optimization of objective for stability of results.

Hoaglin et al. in 1991 continued their fruitful EDA efforts with *Fundamentals of Exploratory Analysis of Variance*. [6] They recast the basics of the analysis of variance with the classical statistical apparatus (e.g., degrees of freedom, F ratios and p-values) in a host of numerical and graphical displays, which often give insight into the structure of the data, such as sizes effects, patterns and interaction and behavior of residuals.

EDA set off a burst of activity in the visual portrayal of data. Published in 1983, *Graphical Methods for Data Analysis* (Chambers et al.) presents new and old methods — some which require a computer, while others only paper and pencil — but all are powerful data analysis tools to learn more about data structure. [7] In 1986 du Toit et al. came out with *Graphical Exploratory Data Analysis,* providing a comprehensive, yet simple presentation of the topic. [8] Jacoby with *Statistical Graphics for Visualizing Univariate and Bivariate Data* (1997), and *Statistical Graphics for Visualizing Multivariate Data* (*1998*) carries out his objective to obtain pictorial representations of quantitative information by elucidating histograms, one-dimensional and enhanced scatterplots and nonparametric smoothing. [9,10] In addition, he successfully transfers graphical displays of multivariate data on a single sheet of paper, a two-dimensional space.

1.4 The EDA Paradigm

EDA presents a major paradigm shift in the ways models are built. With the mantra "Let your data be your guide," EDA offers a view that is a complete reversal of the classical principles that govern the usual steps of model building. The EDA declares the model must always follow the data, not the other way around, as in the classical approach.

In the classical approach, the problem is stated and formulated in terms of an outcome variable Y. It is assumed that the *true* model explaining all the variation in Y is known. Specifically, it is assumed that all the structures (predictor variables, X_i s) affecting Y and their forms are known and present in the model. For example, if Age effects Y, but the log of Age reflects the true relationship with Y, then log of Age must be present in the model. Once the model is specified, the data are taken through the model-specific analysis, which provides the results in terms of numerical values associated with the structures, or estimates of the true predictor variables' coefficients. Then, interpretation is made for declaring X_i an important predictor, assessing how Xi affects the prediction of Y, and ranking X_i in order of predictive importance.

Of course, the data analyst never knows the true model. So, familiarity with the content domain of the problem is used to explicitly put forth the true *surrogate* model, from which good predictions of Y can be made. According to Box, "all models are wrong, but some are useful." [11] In this case, the model selected provides serviceable predictions of Y. Regardless of the model used, the assumption of knowing the truth about Y sets the statistical logic in motion to cause likely bias in the analysis, results and interpretation.

In the EDA approach, not much is assumed beyond having some prior experience with content domain of the problem. The right attitude, flexibility and sharp-sightedness are the forces behind the data analyst, who assesses the problem and lets the data guide the analysis, which then suggests the structures and their forms of the model. If the model passes the validity check, then it is considered final and ready for results and interpretation to be made. If not, with the force still behind the data analyst, the analysis and/or data are revisited until new structures produce a sound and validated model, after which final results and interpretation are made (see Figure 1.1). Without exposure to assumption violations, the EDA paradigm offers a degree of confidence that its prescribed exploratory efforts are not biased, at least in the manner of classical approach. Of course, no analysis is bias-free, as all analysts admit their own bias into the equation.

1.5 EDA Weaknesses

With all its strengths and determination, EDA as originally developed had two minor weaknesses that could have hindered its wide acceptance and great success. One is of a subjective or psychological nature, and the other is a misconceived notion. Data analysts know that failure to look into a multitude of possibilities can result in a flawed analysis, thus finding themselves in a competitive struggle against the data itself. Thus, EDA can foster in data analysts an insecurity that their work is never done. The PC can assist the data analysts in being thorough with their analytical due diligence, but bears no responsibility for the arrogance EDA engenders.

The belief that EDA, which was originally developed for the small-data setting, does not work as well with large samples is a misconception. Indeed, some of the graphical methods, such as the stem-and-leaf plots, and some of the numerical and counting methods, such as folding and binning, do breakdown with large samples. However, the majority of the EDA methodology is unaffected by data size. Neither the manner in which the methods are carried out, nor the reliability of the results is changed. In fact, some of the most powerful EDA techniques scale up quite nicely, but do require the PC to do the serious number crunching of the *big data*.[4] [12] For example, techniques such as ladder of powers, re-expressing and smoothing are very valuable tools for large sample or big data applications.

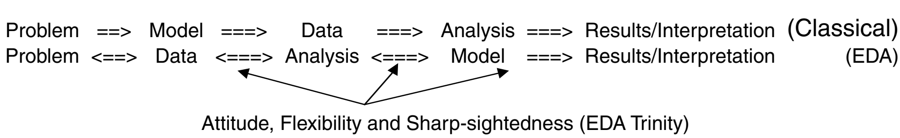

FIGURE 1.1
EDA Paradigm

[4] Authors Weiss and Indurkhya and I use the general concept of "big" data. However, we stress different characteristics of the concept.

1.6 Small and Big Data

I would like to clarify the general concept of small and big data, as size, like beauty, is in the mind of the data analyst. In the past, small data fit the conceptual structure of classical statistics. Small always referred to the sample size, not the number of variables which were always kept to a handful. Depending on the method employed, small was seldom less than 5, was sometimes between 5 and 20, frequently between 30 and 50, and 50 and 100, and rarely between 100 and 200 individuals. In contrast to today's big data, small data are a tabular display of rows (observations or individuals) and columns (variables or features) that fits on a few sheets of paper.

In addition to the compact area they occupy, small data are neat and tidy. They are "clean," in that they contain no unexpected values, except for those due to primal data entry error. They do not include the statistical outliers and influential points, or the EDA far-out and outside points. They are in the "ready-to-run" condition required by classical statistical methods.

There are two sides to big data. On one side is classical statistics which considers big as simply being *not small*. Theoretically, big is the sample size after which asymptotic properties of the method "kick in" for valid results. On the other side is contemporary statistics which considers big in terms of *lifting observations* and *learning from the variables*. Although it depends on who is analyzing the data, a sample size greater than 50,000 individuals can be considered "big." Thus, calculating the average income from a database of 2 million individuals requires heavy-duty lifting (number crunching). In terms of learning or uncovering the structure among the variables, big can be considered 50 variables or more. Regardless of which side the data analyst is working, EDA scales up for both rows and columns of the data table.

1.6.1 Data Size Characteristics

There are three distinguishable characteristics of data size: *condition*, *location*, and *population*. Condition refers to the state of readiness of the data for analysis. Data that require minimal time and cost to clean before reliable analysis can be performed are well conditioned; data that involve a substantial amount of time and cost are ill conditioned. Small data are typically clean, and thus well conditioned.

Big data are an outgrowth of today's digital environment, which generates data flowing continuously from all directions at unprecedented speed and volume, and which almost always require cleansing. They are considered "dirty" mainly because of the merging of multiple sources. The merging process is inherently a time-intensive process, as multiple passes of the sources must be made to get a sense of how the combined sources fit together. Because of the iterative nature of the process, the logic of matching individual records across sources is at first "fuzzy," then fine-tuned to soundness; until that point unexplainable, seemingly random, nonsensical values result. Thus, big data are ill conditioned.

Location refers to where the data reside. Unlike the rectangular sheet for small data, big data reside in relational databases consisting of a *set of data tables*. The link among the data tables can be hierarchical (rank- or level-dependent) and/or sequential (time- or event-dependent). Merging of multiple data sources, each consisting of many rows and columns, produces data of even greater number of rows and columns, clearly suggesting bigness.

Population refers to the group of individuals having qualities or characteristics in common and related to the study under consideration. Small data ideally represent a random sample of a known population, which is not expected to encounter changes in its composition in the foreseeable future. The data are collected to answer a specific problem, permitting straightforward answers from a given problem-specific method. In contrast, big data often represent multiple, nonrandom samples of unknown populations, shifting in composition within the short-term. Big data are "secondary" in nature; that is, they are not collected for an intended purpose. They are available from the hydra of marketing information, for use on any post hoc query, and may not have a straightforward solution.

It is interesting to note that Tukey never talked specifically about the big data per se. However, he did predict that the cost of computing, both in time and dollars, would be cheap, which arguably suggests that he knew big data were coming. Regarding the cost, clearly today's PC bears this out.

1.6.2 Data Size: Personal Observation of One

The data size discussion raises the following question: how large should a sample be? Sample size can be anywhere from folds of 10,000 up to 100,000.

In my experience as a database marketing consultant and a teacher of statistics and data analysis, I have observed that the less experienced and trained statistician/data analyst uses sample sizes that are unnecessarily large. I see analyses performed on and models built from samples too large by factors ranging from 20 to 50. Although the PC can perform the heavy calculations, the extra time and cost in getting the larger data out of the data warehouse and then processing it and thinking about it are almost never justified. Of course, the only way a data analyst learns that extra big data are a waste of resources is by performing small vs. big data comparisons, a step I recommend.

1.7 Data Mining Paradigm

The term *data mining* emerged from the database marketing community sometime between the late 1970s and early 1980s. Statisticians did not understand the excitement and activity caused by this new technique, since the discovery of patterns and relationships (structure) in the data is not new to them. They had known about data mining for a long time, albeit under various names such as data fishing, snooping, and dredging, and most disparaging, "ransacking" the data. Because any discovery process inherently exploits the data, producing spurious findings, statisticians did not view data mining in a positive light.

Simply looking for something increases the odds that it will be found; therefore looking for structure typically results in finding structure. All data have spurious structures, which are formed by the "forces" that makes things come together, such as chance. The bigger the data, the greater are the odds that spurious structures abound. Thus, an expectation of data mining is that it produces structures, both real and spurious, without distinction between them.

Today, statisticians accept data mining only if it embodies the EDA paradigm. They define data mining as *any process that finds unexpected structures in data and uses the EDA framework to insure that the process explores the data, not exploits it* (see Figure 1.1). Note the word "unexpected," which suggests that the process is exploratory, rather than a confirmation that an expected structure has been found. By finding what one expects to find, there is no longer uncertainty as to the existence of the structure.

Statisticians are mindful of the inherent nature of data mining and try to make adjustments to minimize the number of spurious structures identified. In classical statistical analysis, statisticians have explicitly modified most analyses that search for interesting structure, such as adjusting the overall alpha level/type I error rate, or inflating the degrees of freedom [13,14]. In data mining, the statistician has no explicit analytical adjustments available, only the implicit adjustments affected by using the EDA paradigm itself. The following steps outline the data mining/EDA paradigm. As expected from EDA, the steps are defined by *soft* rules.

Suppose the objective is to find structure to help make good predictions of response to a future mail campaign. The following represent the steps that need to be taken:

Obtain the database that has similar mailings to the future mail campaign.

Draw of sample from the database. Size can be several folds of 10,000, up to 100,000.

Perform many exploratory passes of the sample. That is, do all desired calculations to determine the interesting or noticeable structure.

Stop the calculations that are used for finding the noticeable structure.

Count the number of noticeable structures that emerge. The structures are not final results and should not be declared significant findings.

Seek out indicators, visual and numerical, and the indirect messages.

React or respond to all indicators and indirect messages.

Ask questions. Does each structure make sense by itself? Do any of the structures form natural groups? Do the groups make sense; is there consistency among the structures within a group?

Try more techniques. Repeat the many exploratory passes with several fresh samples drawn from the database. Check for consistency across the multiple passes. If results do not behave in a similar way, there may be no structure to predict response to a future mailing, as chance may have infected your data. If results behave similarly, then assess the variability of each structure and each group.

Choose the most stable structures and groups of structures for predicting response to a future mailing.

1.8 Statistics and Machine Learning

Coined by Samuel in 1959, the term "machine learning" (ML) was given to the field of study that assigns computers the ability to learn without being explicitly programmed. [15] In other words, ML investigates ways in which the computer can acquire knowledge directly from data and thus learn to solve problems. It would not be long before machine learning would influence the statistical community.

In 1963, Morgan and Sonquist led a rebellion against the restrictive assumptions of classical statistics. [16] They developed the Automatic Interaction Detection (AID) regression tree, a methodology without assumptions. AID is a computer-intensive technique that finds or learns multidimensional patterns and relationships in data, and serves as an assumption-free, nonparametric alternative to regression analysis. AID marked the beginning of a nonstatistical machine learning approach to solving statistical problems. There have been many improvements and extensions of AID: THAID, MAID, CHAID, and CART, which are now considered valid data mining tools. CHAID and CART have emerged as the most popular today.

Independent from the work of Morgan and Sonquist, machine-learning researchers had been developing algorithms to automate the induction process, which provided another alternative to regression analysis. In 1979 Quinlan used the well-known Concept Learning System developed by Hunt et al. to implement one of the first intelligent systems — ID3 — which was succeeded by C4.5 and C5.0. [17,18] These algorithms are also considered data mining tools, but have not successfully crossed over to the statistical community.

The interface of statistics and machine learning began in earnest in the 1980s. Machine learning researchers became familiar with the three classical problems facing statisticians: regression (predicting a continuous outcome variable), classification (predicting a categorical outcome variable), and clustering (generating a few composite variables that carry a large percentage of the information in the original variables). They started using their machinery (algorithms and the PC) for a nonstatistical, or assumption-free, nonparametric approach to the three problem areas. At the same time, statisticians began harnessing the power of the desktop PC to influence the classical problems they know so well, thus relieving themselves from the starchy parametric road.

The machine-learning community has many specialty groups working on data mining: neural networks, fuzzy logic, genetic algorithms and programming, information retrieval, knowledge acquisition, text processing, inductive logic programming, expert systems, and dynamic programming. All areas have the same objective in mind, but accomplish it with their own tools and techniques. Unfortunately, the statistics community and the machine learning subgroups have no real exchanges of ideas or best practices. They create distinctions of no distinction.

1.9 Statistical Learning

In the spirit of EDA, it is incumbent on data analysts to try something new, and retry something old. They can benefit not only from the computational power of the PC in doing the heavy lifting of big data, but from the machine-learning ability of the PC in uncovering structure nestled in the big data. In the spirit of trying something old, statistics still has a lot to offer.

Thus, today's data mining can be defined in terms of three easy concepts:

1. *Statistics with emphasis on EDA proper:* This includes using the descriptive noninferential parts of classical statistical machinery as indicators. The parts include sum-of-squares, degrees of freedom, F ratios, chi-square values and p-values, but exclude inferential conclusions.
2. *Big data:* Big data are given special mention because of today's digital environment. However, because small data are a component of big data, it is not excluded.
3. *Machine-learning:* The PC is the learning machine, the *essential processing unit*, having the ability to learn without being explicitly programmed, and the intelligence to find structure in the data. Moreover, the PC is essential for big data, as it can always do what it is explicitly programmed to do.

In view of the above, the following data-mining mnemonic can be formed:

Data Mining = Statistics + Big data + Machine Learning and Lifting

Thus, data mining is defined today as *all of statistics and EDA for big and small data with the power of PC for the lifting of data and learning the structures within the data.* Explicitly referring to big and small data implies the process works equally well on both.

Again, in the spirit of EDA, it is prudent to parse the mnemonic equation. Lifting and learning require two different aspects of the data table and the PC. Lifting involves the rows of the data table and the capacity the PC in terms of MIPS (million instructions per second), the speed in which explicitly programmed steps are executed. An exception would be when there are many populations or clusters, in which case the PC will have to study the rows and their relationships to each other to identify the structure within the rows.

Learning focuses on the columns of the data table and the ability of the PC to find the structure within the columns without being explicitly programmed. Learning is more demanding on the PC than lifting, especially when the number of columns increases, in the same ways that learning from books is always more demanding than merely lifting the books. When lifting and learning of the rows are required in addition to learning within the columns, the PC must work exceptionally hard, but can yield extraordinary results.

Further parsing of the mnemonic equation reveals a branch of data mining referred to as *statistical learning*. Contracting the right-hand side of the mnemonic demonstrates that data mining *includes* statistical learning, given that the process makes no distinction between big and small data. *Statistical* connotes statistics and its classical problems of classification, prediction and clustering. *Learning* connotes the capacity of artificial or machine intelligence to exert its influence directly or indirectly on solving (statistical) problems. In other words, *statistical learning is solving statistical problems with nonstatistical machine learning.*

References

1. Tukey, J.W., The future of statistics, *Annals of Mathematical Statistics*, 33, 1–67, 1962.
2. Roy, S.N., *Some Aspects of Multivariate Analysis*, Wiley, New York, 1957.
3. Tukey, J.W., *The Exploratory Data Analysis*, Addison-Wesley, Reading, MA, 1977.
4. Mosteller, F. and Tukey, J.W., *Data Analysis and Regression*, Addison-Wesley, Reading, MA, 1977.
5. Hoaglin, D.C., Mosteller, F., and Tukey, J.W., *Understanding Robust and Exploratory Data Analysis*, John Wiley & Sons, New York, 1983.
6. Hoaglin, D.C., Mosteller, F., and Tukey, J.W., *Fundamentals of Exploratory Analysis of Variance*, John Wiley & Sons, New York, 1991.

7. Chambers, M.J., Cleveland, W.S., Kleiner, B., and Tukey, P.A., *Graphical Methods for Data Analysis*, Wadsworth & Brooks/Cole Publishing Company, CA, 1983.
8. du Toit, S.H.C., Steyn, A.G.W., and Stumpf, R.H., *Graphical Exploratory Data Analysis,* Springer-Verlag, New York, 1986.
9. Jacoby, W.G., *Statistical Graphics for Visualizing Univariate and Bivariate Data*, Sage Publication, Thousand Oaks, CA, 1997.
10. Jacoby, W.G., *Statistical Graphics for Visualizing Multivariate Data*, Sage Publication, Thousand Oaks, CA, 1998.
11. Box, G.E.P., Science and statistics, *Journal of the American Statistical Association,* 71, 791–799, 1976.
12. Weiss, S.M. and Indurkhya, N., *Predictive Data Mining,* Morgan Kaufman Publishers Inc., San Francisco, CA, 1998.
13. Dun, O.J., Multiple comparison among means, *Journal of the American Statistical Association*, 54, 52–64, 1961.
14. Ye, J., On measuring and correcting the effects of data mining and model selection, *Journal of the American Statistical Association*, 93, 120–131, 1998.
15. Samuel, A., Some studies in machine learning using the game of checkers, In Feigenbaum, E. and Feldman, J., Eds., *Computers and Thought.* McGraw-Hill, New York, 1963.
16. Morgan, J.N. and Sonquist, J.A., Problems in the analysis of survey data, and a proposal, *Journal of the American Statistical Association*, 58, 415–435, 1963.
17. Hunt, E., Marin, J., and Stone, P., *Experiments in Induction*, Academic Press, New York, 1966.
18. Quinlan, J.R., Discovering rules by induction from large collections of examples, In Mite, D., Ed., *Expert Systems in the Micro Electronic Age*, Edinburgh University Press, Edinburgh, U.K., 1979.

References

1 1. Introduction

1. Tukey, J.W., The future of statistics, Annals of Mathematical Statistics , 33, 1-67, 1962.

2. Roy, S.N., Some Aspects of Multivariate Analysis , Wiley, New York, 1957.

3. Tukey, J.W., The Exploratory Data Analysis , Addison-Wesley, Reading, MA, 1977.

4. Mosteller, F. and Tukey, J.W., Data Analysis and Regression , Addison-Wesley, Reading, MA, 1977.

5. Hoaglin, D.C., Mosteller, F., and Tukey, J.W., Understanding Robust and Exploratory Data Analysis , John Wiley & Sons, New York, 1983.

6. Hoaglin, D.C., Mosteller, F., and Tukey, J.W., Fundamentals of Exploratory Analysis of Variance , John Wiley & Sons, New York, 1991.

7. Chambers, M.J., Cleveland, W.S., Kleiner, B., and Tukey, P.A., Graphical Methods for Data Analysis , Wadsworth & Brooks/Cole Publishing Company, CA, 1983.

8. du Toit, S.H.C., Steyn, A.G.W., and Stumpf, R.H., Graphical Exploratory Data Analysis, Springer-Verlag, New York, 1986.

9. Jacoby, W.G., Statistical Graphics for Visualizing Univariate and Bivariate Data, Sage Publication, Thousand Oaks, CA, 1997.

10. Jacoby, W.G., Statistical Graphics for Visualizing Multivariate Data, Sage Publication, Thousand Oaks, CA, 1998.

11. Box, G.E.P., Science and statistics, Journal of the American Statistical Association, 71, 791-799, 1976.

12. Weiss, S.M. and Indurkhya, N., Predictive Data Mining, Morgan Kaufman Publishers Inc., San Francisco, CA, 1998.

13. Dun, O.J., Multiple comparison among means, Journal of the American Statistical Association, 54, 52-64, 1961.

14. Ye, J., On measuring and correcting the effects of data

mining and model selection, Journal of the American Statistical Association, 93, 120-131, 1998.

15. Samuel, A., Some studies in machine learning using the game of checkers, In Feigenbaum, E. and Feldman, J., Eds., Computers and Thought. McGraw-Hill, New York, 1963.

16. Morgan, J.N. and Sonquist, J.A., Problems in the analysis of survey data, and a proposal, Journal of the American Statistical Association, 58, 415-435, 1963.

17. Hunt, E., Marin, J., and Stone, P., Experiments in Induction, Academic Press, New York, 1966.

18. Quinlan, J.R., Discovering rules by induction from large collections of examples, In Mite, D., Ed., Expert Systems in the Micro Electronic Age, Edinburgh University Press, Edinburgh, U.K., 1979.

2 2. Two Simple Data Mining Methods for Variable Assessment

1. Anscombe, F.J., Graphs in statistical analysis, American Statistician , 27, 17–22, 1973.

2. Tukey, J.W., Exploratory Data Analysis , Addison-Wesley, Reading, MA, 1997.

3. Hardle, W., Smoothing Techniques , Springer-Verlag, New York, 1990.

4. Simonoff, J.S., Smoothing Methods in Statistics , Springer-Verlag, New York, 1996 .

5. Quenouille, M.H., Rapid Statistical Calculations , Hafner Publishing, New York, 1959.

4 4. Ordinary Regression: The Workhorse of Database Profit Modeling

1. Horst, P., The role of predictor variables which are independent of the criterion. Social Science Research Bulletin, 48, 431-436, 1941.

2. Conger, A.J., A revised definition for suppressor variables: A guide to their identification and interpretation, Educational and Psychological Measurement, 34, 35-46, 1974.

6 6. The Importance of the Regression Coefficient

1. Kraemer, H.C. and Thiemann, S. How Many Subjects? Sage Publications, Thousand Oaks, CA, 1987.

2. Dash, M. and Liu, H., Feature Selection for Classification, Intelligent Data Analysis, Elsevier Science, New York, 1997.

3. Mosteller, F. and Tukey, J., Data Analysis and Regression, Addison-Wesley, Reading, MA, 1977.

4. Hayes, W.L., Statistics for the Social Sciences, Holt, Rinehart and Winston, Austin, TX, 1972.

8 8. CHAID for Specifying a Model with Interaction Variables

1. Marquardt, D.W., You should standardize the predictor variables in your regression model. Journal of the American Statistical Association, 75, 87–91, 1980.

2. Aiken, L.S. and West, S.G., Multiple Regression: Testing and Interpreting Interactions, Sage Publications, Thousand Oaks, CA ,1991.

3. Chipman, H., Bayesian variable selection with related predictors, Canadian Journal of Statistics, 24, 17–36, 1996.

4. Peixoto, J.L., Hierarchical variable selection in polynomial regression models, The American Statistician, 41, 311–313, 1987.

5. Nelder, J.A., Functional marginality is important (letter to editor), Applied Statistics, 46, 281–282, 1997.

6. McCullagh, P.M. and Nelder, J.A., Generalized Linear Models, Chapman & Hall, London, 1989.

7. Fox, J., Applied Regression Analysis, Linear Models, and Related Methods, Sage Publications, Thousand Oaks, CA, 1997.

8. Nelder, J.A., The selection of terms in response-surface models — how strong is the weak-heredity principle? The American Statistician, 52, 4, 1998.

10 10. CHAID as a Method for Filling in Missing Values

1. Cohen, J. and Cohen, P., Applied Multiple Regression and Correlation Analysis for the Behavioral Sciences , Erlbaum, Hillsdale, NJ, 1987.

2. Rubin, D.B., Inference and missing data, Biometrika, 63, 581–592, 1976.

3. Allison, P.D., Missing Data , Sage Publication, Thousand Oaks, CA, 2002, 2.

13 13. Bootstrapping in Database Marketing: A New Approach for Validating Models

1. Noreen, E.W., Computer Intensive Methods for Testing Hypotheses, John Wiley & Sons, New York, 1989.

2. Efron, B., The Jackknife, the Bootsrap and Other Resampling Plans, SIAM, Philadelphia, 1982.

3. Draper, N.R. and Smith, H., Applied Regression Analysis, John Wiley & Sons, New York, 1966.

4. Neter, J. and Wasserman, W., Applied Linear Statistical Models, Irwin, Inc., 1974.

5. Hayes, W.L., Statistics for the Social Sciences, Holt, Rinehart and Winston, Austin, TX, 1973.

1. Descartes, R., The Geometry of Rene Descartes, Dover, New York, 1954.

2. Costigan-Eaves, P., Data Graphics in the 20th Century: A Comparative and Analytical Survey. Ph.D. thesis, Rutgers University, New Jersey, 1984.

3. Funkhouser, H.G., Historical development of the graphical representation of statistical data, Osiris, 3, 269–404, 1937.

4. Du Toit, S.H.C., Steyn, A.G.W., and Stumpf, R.H., Graphical Exploratory Data Analysis, Springer-Verlag, New York, 1986, p.2.

5. Tukey, J.W., The Exploratory Data Analysis, Addison-Wesley, Reading, MA, 1977.

6. Salsburg, D., The Lady Tasting Tea, Freeman, New York, 2001.

7. Snee, R.D., Hare, L.B., and Trout, J. R., Experiments in Industry. Design, Analysis and Interpretation of Results, American Society for Quality, Milwaukee, WI, 1985.

8. Friedman, H.P., Farrell, E.S., Goldwyn, R.M., Miller, M., and Siegel, J., A graphic way of describing changing multivariate patterns, Proceedings of the Sixth Interface Symposium on Computer Science and Statistics, University of California Press, Berkeley, 1972.

9. Andrews, D.F., Plots of high-dimensional data, Biometrics, 28, 1972.

15 15. Genetic Modeling in Database Marketing: The GenIQ Model

1. Koza, J., Genetic Programming: On the Programming of Computers by Means of Natural Selection, The MIT Press, Cambridge, MA, 1992.

16 16. Finding the Best Variables for Database Marketing Models

1. Dash, M. and Liu, H., Feature selection for classification, Intelligent Data Analysis, Elsevier Science, New York, 1997.

2. Ryan, T.P., Modern Regression Methods, Wiley, New York, 1997.

3. Miller, A.J., Subset Selection in Regression, Chapman and Hall, London, 1990.

4. Fox, J., Applied Regression Analysis, Linear Models, and Related Methods, Sage Publications, Thousand Oaks, CA, 1997.