GenIQ, Ltd. **BRUCE RATNER, PhD**

574 Flanders Drive North Woodmere, NY 11581 br@dmstat1.com
516.791.3544 ~ fax 516.791.5075 www.GenIQ.net

Statistical and Machine-Learning Data Mining:
Techniques for Better Predictive Modeling and Analysis of Big Data

Ratner, B. CRC Press/Taylor & Francis, January 9, 2012

Preface

This book is unique. It is the only book, to date, that distinguishes between statistical data mining and machine-learning data mining. I was an orthodox statistician until I resolved my struggles with the weaknesses of statistics within the big data setting of today. Now, as a reform statistician who is free of the statistical rigors of yesterday, with many degrees of freedom to exercise, I have composed by intellectual might the original and practical statistical data mining techniques in the first part of the book. The GenIQ Model, a machine-learning alternative to statistical regression, led to the creative and useful machine-learning data mining techniques in the remaining part of the book.

This book is a compilation of essays that offer detailed background, discussion, and illustration of specific methods for solving the most commonly experienced problems in predictive modeling and analysis of big data. 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 predictive modeling and analysis. 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, with one exception, 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 is covered in but one chapter that is tied in the dependency model. Because interdependency models comprise a minimal proportion of the data analyst's workload, I humbly suggest that the focus of this book will prove utilitarian.

Therefore, these essays have been organized in the following fashion. Chapter 1 reveals the two most influential factors in my professional life: John W. Tukey and the personal computer (PC). The PC has changed everything in the world of statistics. The PC can effortlessly produce precise calculations and eliminate the computational burden associated with statistics. One need only provide the right questions. Unfortunately, the confluence of the PC and the world of statistics has turned generalists with minimal statistical backgrounds into quasi statisticians and affords them a false sense of confidence.

In 1962, in his influential article, "The Future of Data Analysis" [1], John

Tukey predicted a movement to unlock the rigidities that characterize statistics. It was not until the publication of *Exploratory Data Analysis* [2] in 1977 that Tukey led statistics away from the rigors that defined it into a new area, known as EDA (from the first initials of the title of his seminal work). At its core, EDA, known presently as data mining or formally as statistical data mining, is an unending effort of numerical, counting, and graphical detective work.

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 unfamiliar to many, as well as introduce two useful methods for variable assessment. Building on the concept of smooth scatterplot presented in Chapter 2, I introduce in Chapter 3 the smoother scatterplot based on CHAID (chi-squared automatic interaction detection). The new method has the potential of exposing a more reliable depiction of the unmasked relationship for paired-variable assessment than that of the smoothed scatterplot.

In Chapter 4, I show the importance of straight data for the simplicity and desirability it brings for good model building. In Chapter 5, I introduce the method of symmetrizing ranked data and add it to the paradigm of simplicity and desirability presented in Chapter 4.

Principal component analysis, the popular data reduction technique invented in 1901, is repositioned in Chapter 6 as a data mining method for many-variable assessment. In Chapter 7, I readdress the correlation coefficient. I discuss the effects the distributions of the two variables under consideration have on the correlation coefficient interval. Consequently, I provide a procedure for calculating an adjusted correlation coefficient.

In Chapter 8, 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 9. On the heels of discussing the workhorses of statistical regression in Chapters 8 and 9, I resurface the scope of literature on the weaknesses of variable selection methods, and I enliven anew a notable solution for specifying a well-defined regression model in Chapter 10. Chapter 11 focuses on the interpretation of the logistic regression model with the use of CHAID as a data mining tool. Chapter 12 refocuses on the regression coefficient and offers common misinterpretations of the coefficient that point to its weaknesses. Extending the concept of coefficient, I introduce the average correlation coefficient in Chapter 13 to provide a quantitative criterion for assessing competing predictive models and the importance of the predictor variables.

In Chapter 14, I demonstrate how to increase the predictive power of a model 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 the 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 15, where I solve the problem of market segment classification modeling using not only logistic regression but also CHAID. In Chapter 16, CHAID is yet again utilized in a somewhat unconventional manner—as a method for filling in missing values in one's data. To bring an interesting real-life problem into the picture, I wrote Chapter 17 to describe profiling techniques for the marketer who wants a method for identifying his or her best customers. The benefits of the predictive profiling approach is demon-

strated and expanded to a discussion of look-alike profiling.

I take a detour in Chapter 18 to discuss how marketers assess the accuracy of a model. Three concepts of model assessment are discussed: the traditional decile analysis, as well as two additional concepts, precision and separability. In Chapter 19, continuing in this mode, I point to the weaknesses in the way the decile analysis is used and offer a new approach known as the bootstrap for measuring the efficiency of marketing models.

The purpose of Chapter 20 is to introduce the principal features of a bootstrap validation method for the ever-popular logistic regression model. Chapter 21 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.

I close the statistical data mining part of the book with Chapter 22, in which I offer a data-mining alternative measure, the predictive contribution coefficient, to the standardized coefficient.

With the discussions just described behind us, we are ready to venture to new ground. In Chapter 1, I elaborated on the concept of machine-learning data mining and defined it as PC learning without the EDA/statistics component. In Chapter 23, I use a metrical modelogue, "To Fit or Not to Fit Data to a Model," to introduce the machine-learning method of GenIQ and its favorable data mining offshoots.

In Chapter 24, I maintain that the machine-learning paradigm, which lets the data define the model, is especially effective with big data. Consequently, I present an exemplar illustration of genetic logistic regression outperforming statistical logistic regression, whose paradigm, in contrast, is to fit the data to a predefined model. In Chapter 25, I introduce and illustrate brightly, perhaps, the quintessential data mining concept: data reuse. Data reuse is appending new variables, which are found when building a GenIQ Model, to the original dataset. The benefit of data reuse is apparent: The original dataset is enhanced with the addition of new, predictive-full GenIQ data-mined variables.

In Chapters 26–28, I address everyday statistics problems with solutions stemming from the data mining features of the GenIQ Model. In statistics, an outlier is an observation whose position falls outside the overall pattern of the data. Outliers are problematic: Statistical regression models are quite sensitive to outliers, which render an estimated regression model with questionable predictions. The common remedy for handling outliers is "determine and discard" them. In Chapter 26, I present an alternative method of moderating outliers instead of discarding them. In Chapter 27, I introduce a new solution to the old problem of overfitting. I illustrate how the GenIQ Model identifies a structural source (complexity) of overfitting, and subsequently instructs for deletion of the individuals who contribute to the complexity, from the dataset under consideration. Chapter 28 revisits the examples (the importance of straight data) discussed in Chapters 4 and 9, in which I posited the solutions without explanation as the material needed to understand the solution was not introduced at that point. At this point, the background required has been covered. Thus, for completeness, I detail the posited solutions in this chapter.

GenIQ is now presented in Chapter 29 as such a nonstatistical machine-learning model. Moreover, in Chapter 30, GenIQ serves as an effective method for finding the best possible subset of variables for a model. Because GenIQ has no coefficients—and coefficients furnish the key to prediction—Chapter 31 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.

References

1. Tukey, J.W., The future of data analysis, *Annals of Mathematical Statistics*, 33, 1–67, 1962.
2. Tukey, J.W., *Exploratory Data Analysis*, Addison-Wesley, Reading, MA, 1977.

Contents