

WILEY SERIES IN COMPUTATIONAL STATISTICS



Stéphane Tufféry

DATA MINING AND STATISTICS FOR DECISION MAKING

 **WILEY**

Data Mining and Statistics for Decision Making

Wiley Series in Computational Statistics

Consulting Editors:

Paolo Giudici
University of Pavia, Italy

Geof H. Givens
Colorado State University, USA

Bani K. Mallick
Texas A&M University, USA

Wiley Series in Computational Statistics is comprised of practical guides and cutting edge research books on new developments in computational statistics. It features quality authors with a strong applications focus. The texts in the series provide detailed coverage of statistical concepts, methods and case studies in areas at the interface of statistics, computing, and numerics.

With sound motivation and a wealth of practical examples, the books show in concrete terms how to select and to use appropriate ranges of statistical computing techniques in particular fields of study. Readers are assumed to have a basic understanding of introductory terminology.

The series concentrates on applications of computational methods in statistics to fields of bioinformatics, genomics, epidemiology, business, engineering, finance and applied statistics.

Titles in the Series

Biegler, Biros, Ghattas, Heinkenschloss, Keyes, Mallick, Marzouk, Tenorio, Waanders, Willcox – Large-Scale Inverse Problems and Quantification of Uncertainty
Billard and Diday – Symbolic Data Analysis: Conceptual Statistics and Data Mining
Bolstad – Understanding Computational Bayesian Statistics
Borgelt, Steinbrecher and Kruse – Graphical Models, 2e
Dunne – A Statistical Approach to Neutral Networks for Pattern Recognition
Liang, Liu and Carroll – Advanced Markov Chain Monte Carlo Methods
Ntzoufras – Bayesian Modeling Using WinBUGS

Data Mining and Statistics for Decision Making

Stéphane Tufféry

University of Rennes, France

Translated by Rod Riesco



A John Wiley & Sons, Ltd., Publication

First published under the title 'Data Mining et Statistique Decisionnelle' by Editions Technip
© Editions Technip 2008
All rights reserved.
Authorised translation from French language edition published by Editions Technip, 2008

This edition first published 2011
© 2011 John Wiley & Sons, Ltd

Registered office

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com.

The right of the author to be identified as the author of this work has been asserted in accordance with the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by the UK Copyright, Designs and Patents Act 1988, without the prior permission of the publisher.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The publisher is not associated with any product or vendor mentioned in this book. This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

Library of Congress Cataloging-in-Publication Data

Tufféry, Stéphane

Data mining and statistics for decision making / Stéphane Tufféry.
p. cm. – (Wiley series in computational statistics)

Includes bibliographical references and index.

ISBN 978-0-470-68829-8 (hardback)

1. Data mining. 2. Statistical decision. I. Title.

QA76.9.D343T84 2011

006.3'12–dc22

2010039789

A catalogue record for this book is available from the British Library.

Print ISBN: 978-0-470-68829-8

ePDF ISBN: 978-0-470-97916-7

oBook ISBN: 978-0-470-97917-4

ePub ISBN: 978-0-470-97928-0

Typeset in by 10/12pt Times Roman by Thomson Digital, Noida, India

*to Paul and Nicole Tufféry,
with gratitude and affection*

Contents

Preface	xvii
Foreword	xxi
Foreword from the French language edition	xxiii
List of trademarks	xxv
1 Overview of data mining	1
1.1 What is data mining?	1
1.2 What is data mining used for?	4
1.2.1 Data mining in different sectors	4
1.2.2 Data mining in different applications	8
1.3 Data mining and statistics	11
1.4 Data mining and information technology	12
1.5 Data mining and protection of personal data	16
1.6 Implementation of data mining	23
2 The development of a data mining study	25
2.1 Defining the aims	26
2.2 Listing the existing data	26
2.3 Collecting the data	27
2.4 Exploring and preparing the data	30
2.5 Population segmentation	33
2.6 Drawing up and validating predictive models	35
2.7 Synthesizing predictive models of different segments	36
2.8 Iteration of the preceding steps	37
2.9 Deploying the models	37
2.10 Training the model users	38
2.11 Monitoring the models	38
2.12 Enriching the models	40
2.13 Remarks	41
2.14 Life cycle of a model	41
2.15 Costs of a pilot project	41
3 Data exploration and preparation	43
3.1 The different types of data	43
3.2 Examining the distribution of variables	44
3.3 Detection of rare or missing values	45
3.4 Detection of aberrant values	49
3.5 Detection of extreme values	52

3.6	Tests of normality	52
3.7	Homoscedasticity and heteroscedasticity	58
3.8	Detection of the most discriminating variables	59
3.8.1	Qualitative, discrete or binned independent variables	60
3.8.2	Continuous independent variables	62
3.8.3	Details of single-factor non-parametric tests	65
3.8.4	ODS and automated selection of discriminating variables	70
3.9	Transformation of variables	73
3.10	Choosing ranges of values of binned variables	74
3.11	Creating new variables	81
3.12	Detecting interactions	82
3.13	Automatic variable selection	85
3.14	Detection of collinearity	86
3.15	Sampling	89
3.15.1	Using sampling	89
3.15.2	Random sampling methods	90
4	Using commercial data	93
4.1	Data used in commercial applications	93
4.1.1	Data on transactions and RFM data	93
4.1.2	Data on products and contracts	94
4.1.3	Lifetimes	94
4.1.4	Data on channels	96
4.1.5	Relational, attitudinal and psychographic data	96
4.1.6	Sociodemographic data	97
4.1.7	When data are unavailable	97
4.1.8	Technical data	98
4.2	Special data	98
4.2.1	Geodemographic data	98
4.2.2	Profitability	105
4.3	Data used by business sector	106
4.3.1	Data used in banking	106
4.3.2	Data used in insurance	108
4.3.3	Data used in telephony	108
4.3.4	Data used in mail order	109
5	Statistical and data mining software	111
5.1	Types of data mining and statistical software	111
5.2	Essential characteristics of the software	114
5.2.1	Points of comparison	114
5.2.2	Methods implemented	115
5.2.3	Data preparation functions	116
5.2.4	Other functions	116
5.2.5	Technical characteristics	117
5.3	The main software packages	117
5.3.1	Overview	117

5.3.2	IBM SPSS	119
5.3.3	SAS	122
5.3.4	R	124
5.3.5	Some elements of the R language	133
5.4	Comparison of R, SAS and IBM SPSS	136
5.5	How to reduce processing time	164
6	An outline of data mining methods	167
6.1	Classification of the methods	167
6.2	Comparison of the methods	174
7	Factor analysis	175
7.1	Principal component analysis	175
7.1.1	Introduction	175
7.1.2	Representation of variables	181
7.1.3	Representation of individuals	185
7.1.4	Use of PCA	187
7.1.5	Choosing the number of factor axes	189
7.1.6	Summary	192
7.2	Variants of principal component analysis	192
7.2.1	PCA with rotation	192
7.2.2	PCA of ranks	193
7.2.3	PCA on qualitative variables	194
7.3	Correspondence analysis	194
7.3.1	Introduction	194
7.3.2	Implementing CA with IBM SPSS Statistics	197
7.4	Multiple correspondence analysis	201
7.4.1	Introduction	201
7.4.2	Review of CA and MCA	205
7.4.3	Implementing MCA and CA with SAS	207
8	Neural networks	217
8.1	General information on neural networks	217
8.2	Structure of a neural network	220
8.3	Choosing the learning sample	221
8.4	Some empirical rules for network design	222
8.5	Data normalization	223
8.5.1	Continuous variables	223
8.5.2	Discrete variables	223
8.5.3	Qualitative variables	224
8.6	Learning algorithms	224
8.7	The main neural networks	224
8.7.1	The multilayer perceptron	225
8.7.2	The radial basis function network	227
8.7.3	The Kohonen network	231

9	Cluster analysis	235
9.1	Definition of clustering	235
9.2	Applications of clustering	236
9.3	Complexity of clustering	236
9.4	Clustering structures	237
9.4.1	Structure of the data to be clustered	237
9.4.2	Structure of the resulting clusters	237
9.5	Some methodological considerations	238
9.5.1	The optimum number of clusters	238
9.5.2	The use of certain types of variables	238
9.5.3	The use of illustrative variables	239
9.5.4	Evaluating the quality of clustering	239
9.5.5	Interpreting the resulting clusters	240
9.5.6	The criteria for correct clustering	242
9.6	Comparison of factor analysis and clustering	242
9.7	Within-cluster and between-cluster sum of squares	243
9.8	Measurements of clustering quality	244
9.8.1	All types of clustering	245
9.8.2	Agglomerative hierarchical clustering	246
9.9	Partitioning methods	247
9.9.1	The moving centres method	247
9.9.2	<i>k</i> -means and dynamic clouds	248
9.9.3	Processing qualitative data	249
9.9.4	<i>k</i> -medoids and their variants	249
9.9.5	Advantages of the partitioning methods	250
9.9.6	Disadvantages of the partitioning methods	251
9.9.7	Sensitivity to the choice of initial centres	252
9.10	Agglomerative hierarchical clustering	253
9.10.1	Introduction	253
9.10.2	The main distances used	254
9.10.3	Density estimation methods	258
9.10.4	Advantages of agglomerative hierarchical clustering	259
9.10.5	Disadvantages of agglomerative hierarchical clustering	261
9.11	Hybrid clustering methods	261
9.11.1	Introduction	261
9.11.2	Illustration using SAS Software	262
9.12	Neural clustering	272
9.12.1	Advantages	272
9.12.2	Disadvantages	272
9.13	Clustering by similarity aggregation	273
9.13.1	Principle of relational analysis	273
9.13.2	Implementing clustering by similarity aggregation	274
9.13.3	Example of use of the R <i>amap</i> package	275
9.13.4	Advantages of clustering by similarity aggregation	277
9.13.5	Disadvantages of clustering by similarity aggregation	278
9.14	Clustering of numeric variables	278
9.15	Overview of clustering methods	286

10	Association analysis	287
10.1	Principles	287
10.2	Using taxonomy	291
10.3	Using supplementary variables	292
10.4	Applications	292
10.5	Example of use	294
11	Classification and prediction methods	301
11.1	Introduction	301
11.2	Inductive and transductive methods	302
11.3	Overview of classification and prediction methods	304
11.3.1	The qualities expected from a classification and prediction method	304
11.3.2	Generalizability	305
11.3.3	Vapnik's learning theory	308
11.3.4	Overfitting	310
11.4	Classification by decision tree	313
11.4.1	Principle of the decision trees	313
11.4.2	Definitions – the first step in creating the tree	313
11.4.3	Splitting criterion	316
11.4.4	Distribution among nodes – the second step in creating the tree	318
11.4.5	Pruning – the third step in creating the tree	319
11.4.6	A pitfall to avoid	320
11.4.7	The CART, C5.0 and CHAID trees	321
11.4.8	Advantages of decision trees	327
11.4.9	Disadvantages of decision trees	328
11.5	Prediction by decision tree	330
11.6	Classification by discriminant analysis	332
11.6.1	The problem	332
11.6.2	Geometric descriptive discriminant analysis (discriminant factor analysis)	333
11.6.3	Geometric predictive discriminant analysis	338
11.6.4	Probabilistic discriminant analysis	342
11.6.5	Measurements of the quality of the model	345
11.6.6	Syntax of discriminant analysis in SAS	350
11.6.7	Discriminant analysis on qualitative variables (DISQUAL Method)	352
11.6.8	Advantages of discriminant analysis	354
11.6.9	Disadvantages of discriminant analysis	354
11.7	Prediction by linear regression	355
11.7.1	Simple linear regression	356
11.7.2	Multiple linear regression and regularized regression	359
11.7.3	Tests in linear regression	365
11.7.4	Tests on residuals	371
11.7.5	The influence of observations	375
11.7.6	Example of linear regression	377

11.7.7	Further details of the SAS linear regression syntax	383
11.7.8	Problems of collinearity in linear regression: an example using R	387
11.7.9	Problems of collinearity in linear regression: diagnosis and solutions	394
11.7.10	PLS regression	397
11.7.11	Handling regularized regression with SAS and R	400
11.7.12	Robust regression	430
11.7.13	The general linear model	434
11.8	Classification by logistic regression	437
11.8.1	Principles of binary logistic regression	437
11.8.2	Logit, probit and log-log logistic regressions	441
11.8.3	Odds ratios	443
11.8.4	Illustration of division into categories	445
11.8.5	Estimating the parameters	446
11.8.6	Deviance and quality measurement in a model	449
11.8.7	Complete separation in logistic regression	453
11.8.8	Statistical tests in logistic regression	454
11.8.9	Effect of division into categories and choice of the reference category	458
11.8.10	Effect of collinearity	459
11.8.11	The effect of sampling on <i>logit</i> regression	460
11.8.12	The syntax of logistic regression in SAS Software	461
11.8.13	An example of modelling by logistic regression	463
11.8.14	Logistic regression with R	474
11.8.15	Advantages of logistic regression	477
11.8.16	Advantages of the logit model compared with probit	478
11.8.17	Disadvantages of logistic regression	478
11.9	Developments in logistic regression	479
11.9.1	Logistic regression on individuals with different weights	479
11.9.2	Logistic regression with correlated data	479
11.9.3	Ordinal logistic regression	482
11.9.4	Multinomial logistic regression	482
11.9.5	PLS logistic regression	483
11.9.6	The generalized linear model	484
11.9.7	Poisson regression	487
11.9.8	The generalized additive model	491
11.10	Bayesian methods	492
11.10.1	The naive Bayesian classifier	492
11.10.2	Bayesian networks	497
11.11	Classification and prediction by neural networks	499
11.11.1	Advantages of neural networks	499
11.11.2	Disadvantages of neural networks	500
11.12	Classification by support vector machines	501
11.12.1	Introduction to SVMs	501
11.12.2	Example	506
11.12.3	Advantages of SVMs	508
11.12.4	Disadvantages of SVMs	508

11.13	Prediction by genetic algorithms	510
11.13.1	Random generation of initial rules	511
11.13.2	Selecting the best rules	512
11.13.3	Generating new rules	512
11.13.4	End of the algorithm	513
11.13.5	Applications of genetic algorithms	513
11.13.6	Disadvantages of genetic algorithms	514
11.14	Improving the performance of a predictive model	514
11.15	Bootstrapping and ensemble methods	516
11.15.1	Bootstrapping	516
11.15.2	Bagging	518
11.15.3	Boosting	521
11.15.4	Some applications	528
11.15.5	Conclusion	532
11.16	Using classification and prediction methods	534
11.16.1	Choosing the modelling methods	534
11.16.2	The training phase of a model	537
11.16.3	Reject inference	539
11.16.4	The test phase of a model	540
11.16.5	The ROC curve, the lift curve and the Gini index	542
11.16.6	The classification table of a model	551
11.16.7	The validation phase of a model	553
11.16.8	The application phase of a model	553
12	An application of data mining: scoring	555
12.1	The different types of score	555
12.2	Using propensity scores and risk scores	556
12.3	Methodology	558
12.3.1	Determining the objectives	558
12.3.2	Data inventory and preparation	559
12.3.3	Creating the analysis base	559
12.3.4	Developing a predictive model	561
12.3.5	Using the score	561
12.3.6	Deploying the score	562
12.3.7	Monitoring the available tools	562
12.4	Implementing a strategic score	562
12.5	Implementing an operational score	563
12.6	Scoring solutions used in a business	564
12.6.1	In-house or outsourced?	564
12.6.2	Generic or personalized score	567
12.6.3	Summary of the possible solutions	567
12.7	An example of credit scoring (data preparation)	567
12.8	An example of credit scoring (modelling by logistic regression)	594
12.9	An example of credit scoring (modelling by DISQUAL discriminant analysis)	604
12.10	A brief history of credit scoring	615
	References	616

13	Factors for success in a data mining project	617
13.1	The subject	617
13.2	The people	618
13.3	The data	618
13.4	The IT systems	619
13.5	The business culture	620
13.6	Data mining: eight common misconceptions	621
13.6.1	No <i>a priori</i> knowledge is needed	621
13.6.2	No specialist staff are needed	621
13.6.3	No statisticians are needed ('you can just press a button')	622
13.6.4	Data mining will reveal unbelievable wonders	622
13.6.5	Data mining is revolutionary	623
13.6.6	You must use all the available data	623
13.6.7	You must always sample	623
13.6.8	You must never sample	623
13.7	Return on investment	624
14	Text mining	627
14.1	Definition of text mining	627
14.2	Text sources used	629
14.3	Using text mining	629
14.4	Information retrieval	630
14.4.1	Linguistic analysis	630
14.4.2	Application of statistics and data mining	633
14.4.3	Suitable methods	633
14.5	Information extraction	635
14.5.1	Principles of information extraction	635
14.5.2	Example of application: transcription of business interviews	635
14.6	Multi-type data mining	636
15	Web mining	637
15.1	The aims of web mining	637
15.2	Global analyses	638
15.2.1	What can they be used for?	638
15.2.2	The structure of the log file	638
15.2.3	Using the log file	639
15.3	Individual analyses	641
15.4	Personal analysis	642
Appendix A	Elements of statistics	645
A.1	A brief history	645
A.1.1	A few dates	645
A.1.2	From statistics . . . to data mining	645
A.2	Elements of statistics	648
A.2.1	Statistical characteristics	648

A.2.2	Box and whisker plot	649
A.2.3	Hypothesis testing	649
A.2.4	Asymptotic, exact, parametric and non-parametric tests	652
A.2.5	Confidence interval for a mean: student's t test	652
A.2.6	Confidence interval of a frequency (or proportion)	654
A.2.7	The relationship between two continuous variables: the linear correlation coefficient	656
A.2.8	The relationship between two numeric or ordinal variables: Spearman's rank correlation coefficient and Kendall's tau	657
A.2.9	The relationship between n sets of several continuous or binary variables: canonical correlation analysis	658
A.2.10	The relationship between two nominal variables: the χ^2 test	659
A.2.11	Example of use of the χ^2 test	660
A.2.12	The relationship between two nominal variables: Cramér's coefficient	661
A.2.13	The relationship between a nominal variable and a numeric variable: the variance test (one-way ANOVA test)	662
A.2.14	The cox semi-parametric survival model	664
A.3	Statistical tables	665
A.3.1	Table of the standard normal distribution	665
A.3.2	Table of student's t distribution	665
A.3.3	Chi-Square table	666
A.3.4	Table of the Fisher–Snedecor distribution at the 0.05 significance level	667
A.3.5	Table of the Fisher–Snedecor distribution at the 0.10 significance level	673
Appendix B	Further reading	675
B.1	Statistics and data analysis	675
B.2	Data mining and statistical learning	678
B.3	Text mining	680
B.4	Web mining	680
B.5	R software	680
B.6	SAS software	681
B.7	IBM SPSS software	682
B.8	Websites	682
Index		685

Preface

All models are wrong but some are useful.
George E. P. Box¹

[Data analysis] is a tool for extracting the jewel of truth from the slurry of data.
Jean-Paul Benzécri²

This book is concerned with data mining, which is the application of the methods of statistics, data analysis and machine learning to the exploration and analysis of large data sets, with the aim of extracting new and useful information for the benefit of the owner of these data.

An essential component of decision assistance systems in many economic, industrial, scientific and medical fields, data mining is being applied in an increasing variety of areas. The most familiar applications include market basket analysis in the retail and distribution industry (to find out which products are bought at the same time, enabling shelf arrangements and promotions to be planned accordingly), scoring in financial establishments (to predict the risk of default by an applicant for credit), consumer propensity studies (to target mailshots and telephone calls at customers most likely to respond favourably), prediction of attrition (loss of a customer to a competing supplier) in the mobile telephone industry, automatic fraud detection, the search for the causes of manufacturing defects, analysis of road accidents, assistance to medical prognosis, decoding of the genome, sensory analysis in the food industry, and others.

The present expansion of data mining in industry and also in the academic sphere, where research into this subject is rapidly developing, is ample justification for providing an accessible general introduction to this technology, which promises to be a rich source of future employment and which was presented by the Massachusetts Institute of Technology in 2001 as one of the ten emerging technologies expected to ‘change the world’ in the twenty-first century.³

This book aims to provide an introduction to data mining and its contribution to organizations and businesses, supplementing the description with a variety of examples. It details the methods and algorithms, together with the procedures and principles, for implementing data mining. I will demonstrate how the methods of data mining incorporate and extend the conventional methods of statistics and data analysis, which will be described reasonably thoroughly. I will therefore cover conventional methods (clustering, factor analysis, linear regression, ridge regression, partial least squares regression, discriminant

¹ Box, G.E.P. (1979) Robustness in the strategy of scientific model building. In R.L. Launer and G.N. Wilkinson (eds), *Robustness in Statistics*. New York: Academic Press.

² Benzécri, J.-P. (1976) *Histoire et Préhistoire de l'Analyse des Données*. Paris: Dunod.

³ In addition to data mining, the other nine major technologies of the twenty-first century according to MIT are: biometrics, voice recognition, brain interfaces, digital copyright management, aspect-oriented programming, microfluidics, optoelectronics, flexible electronics and robotics.

analysis, logistic regression, the generalized linear model) as well as the latest techniques (decision trees, neural networks, support vector machines and genetic algorithms). We will take a look at recent and increasingly sophisticated methods such as model aggregation by bagging and boosting, the lasso and the ‘elastic net’. The methods will be compared with each other, revealing their advantages, their drawbacks, the constraints on their use and the best areas for their application. Particular attention will be paid to scoring, which is still the most widespread application of predictive data mining methods in the service sector (banking, insurance, telecommunications), and fifty pages of the book are concerned with a comprehensive credit scoring case study. Of course, I also discuss other predictive techniques, as well as descriptive techniques, ranging from market basket analysis, in other words the detection of association rules, to the automatic clustering method known in marketing as ‘customer segmentation’. The theoretical descriptions will be illustrated by numerous examples using SAS, IBM SPSS and R software, while the statistical basics required are set out in an appendix at the end of the book.

The methodological part of the book sets out all the stages of a project, from target setting to the use of models and evaluation of the results. I will indicate the requirements for the success of a project, the expected return on investment in a business setting, and the errors to be avoided.

This survey of new data analysis methods is completed by an introduction to text mining and web mining.

The criteria for choosing a statistical or data mining program and the leading programs available will be mentioned, and I will then introduce and provide a detailed comparison of the three major products, namely the free R software and the two market leaders, SAS and SPSS.

Finally, the book is rounded off with suggestions for further reading and an index.

This is intended to be both a reference book and a practical manual, containing more technical explanations and a greater degree of theoretical underpinning than works oriented towards ‘business intelligence’ or ‘database marketing’, and including more examples and advice on implementation than a volume dealing purely with statistical methods.

The book has been written with the following facts in mind. Pure statisticians may be reluctant to use data mining techniques in a context extending beyond that of conventional statistics because of its methods and philosophy and the nature of its data, which are frequently voluminous and imperfect (see Section A.1.2 in Appendix A). For their part, database specialists and analysts do not always make the best use of the data mining tools available to them, because they are unaware of their principles and operation. This book is aimed at these two groups of readers, approaching technical matters in a sufficiently accessible way to be usable with a minimum of mathematical baggage, while being sufficiently precise and rigorous to enable the user of these methods to master them and exploit them fully, without disregarding the problems encountered in the daily use of statistics. Thus, being based on both theoretical and practical knowledge, this book is aimed at a wide range of readers, including:

- statisticians working in private and public businesses, who will use it as a reference work alongside their statistical or data mining software manuals;
- students and teachers of statistics, econometrics or engineering, who can use it as a source of real applications of their statistical learning;

- analysts and researchers in the relevant departments of companies, who will discover what data mining can do for them and what they can expect from data miners and other statisticians;
- chief executive and IT managers which may use it a source of ideas for productive investment in the analysis of their databases, together with the conditions for success in data mining projects;
- any interested reader, who will be able to look behind the scenes of the computerized world in which we live, and discover how our personal data are used.

It is the aim of this book to be useful to the expert and yet accessible to the newcomer.

My thanks are due, in the first place, to David Hand, who found the time to carefully read my manuscript, give me his precious advice on several points and write a very interesting and kind foreword for the English edition, and to Gilbert Saporta, who has done me the honour of writing the foreword of the original French edition, for his support and the enlightening discussions I have had with him. I sincerely thank Jean-Pierre Nakache for his many kind suggestions and constant encouragement. I also wish to thank Olivier Decourt for his useful comments on statistics in general and SAS in particular. I am grateful to Hervé Abdi for his advice on some points of the manuscript. I must thank Hervé Mignot and Grégoire de Lassence, who reviewed the manuscript and made many useful detailed comments. Thanks are due to Julien Fournel for his kind and always relevant contributions. I have not forgotten my friends in the field of statistics and my students, although there are too many of them to be listed in the space available. Finally, a special thought for my wife and children, for their invaluable patience and support during the writing of this book.

This book includes an accompanying website. Please visit www.wiley.com/go/decision_making for more information.

Foreword

It is a real pleasure to be invited to write the foreword to the English translation of Stéphane Tufféry's book *Data Mining and Statistics for Decision Making*.

Data mining represents the merger of a number of other disciplines, most notably statistics and machine learning, applied to the problem of squeezing illumination from large databases. Although also widely used in scientific applications – for example bioinformatics, astrophysics, and particle physics – perhaps the major driver behind its development has been the commercial potential. This is simply because commercial organisations have recognised the competitive edge that expertise in this area can give – that is, the business intelligence it provides - enabling such organisation to make better-informed and superior decisions.

Data mining, as a unique discipline, is relatively young, and as with other youngsters, it is developing rapidly. Although originally it was secondary analysis, focusing solely on large databases which had been collated for some other purpose, nowadays we find more such databases being collected with the specific aim of subjecting them to a data mining exercise. Moreover, we also see formal experimental design being used to decide what data to collect (for example, as with supermarket loyalty cards or bank credit card operations, where different customers receive different cards or coupons).

This book presents a comprehensive view of the modern discipline, and how it can be used by businesses and other organizations. It describes the special characteristics of commercial data from a range of application areas, serving to illustrate the extraordinary breadth of potential applications. Of course, different application domains are characterised by data with different properties, and the author's extensive practical experience is evident in his detailed and revealing discussion of a range of data, including transactional data, lifetime data, sociodemographic data, contract data, and other kinds.

As with any area of data analysis, the initial steps of cleaning, transforming, and generally preparing the data for analysis are vital to a successful outcome, and yet many books gloss over this fundamental step. I hate to think how many mistaken conclusions have been drawn simply because analysts ignored the fact that the data had missing values! This book gives details of these necessary first steps, examining incomplete data, aberrant values, extreme values, and other data distortion issues.

In terms of methodology, as well as the more standard and traditional tools, the book comes up to date with extensive discussions of neural networks, support vector machines, bagging and boosting, and other tools.

The discussion of eight common misconceptions in Chapter 13 will be particularly useful to newcomers to the area, especially business users who are uncertain about the legitimacy of their analyses. And I was struck by the observation, also in this chapter, that for a successful business data mining exercise, the whole company has to buy into the exercise. It is not something to be undertaken by geeks in a back room. Neither is it a one-off exercise, which can be undertaken and then forgotten about. Rather it is an ongoing process, requiring commitment from a wide range of people in an organisation. More generally, data mining is

not a magic wand, which can be waved over a miscellaneous and disorganised pile of data, to miraculously extract understanding and insight. It is an advanced technology of painstaking analysis and careful probing, using highly sophisticated software tools. As with any other advanced technology, it needs to be applied with care and skill if meaningful results are to be obtained. This book very nicely illustrates this in its mix of high level coverage of general issues, deep discussions of methodology, and detailed explorations of particular application areas.

An attractive feature of the book is its discussion of some of the most important data mining software tools and its illustrations of these tools in practice. Other data mining books tend to focus either on the technical methodological aspects, or on a more superficial presentation of the results, often in the form of screen shots, from a particular software package. This book nicely intertwines the two levels, in a way which I am sure will be attractive to readers and potential users of the technology.

The detailed case study of scoring methods in Chapter 12 is excellent, as are the other two application areas discussed in some depth – text mining and web mining. Both of these have become very important areas in their own right, and hold out great promise for knowledge discovery.

This book will be an eye-opener to anyone approaching data mining for the first time. It outlines the methods and tools, and also illustrates very nicely how they are applied, to very good effect, in a variety of areas. It shows how data mining is an essential tool for the data based businesses of today. More than that, however, it also shows how data mining is the equivalent of past centuries' voyages of discovery.

David J. Hand
Imperial College, London, and Winton Capital Management

Foreword from the French language edition

It is a pleasure for me to write the foreword to the third edition of this book, whose popularity shows no sign of diminishing. It is most unusual for a book of this kind to go through three editions in such a short time. It is a clear indication of the quality of the writing and the urgency of the subject matter.

Once again, Stéphane Tufféry has made some important additions: there are now almost two hundred pages more than in the second edition, which itself was practically twice as long as the first. More than ever, this book covers all the essentials (and more) needed for a clear understanding and proper application of data mining and statistics for decision making. Among the new features in this edition, I note that more space has been given to the free R software, developments in support vector machines and new methodological comparisons.

Data mining and statistics for decision making are developing rapidly in the research and business fields, and are being used in many different sectors. In the twenty-first century we are swimming in a flood of statistical information (economic performance indicators, polls, forecasts of climate, population, resources, etc.), seeing only the surface froth and unaware of the nature of the underlying currents.

Data mining is a response to the need to make use of the contents of huge business databases; its aim is to analyse and predict the individual behaviour of consumers. This aspect is of great concern to us as citizens. Fortunately, the risks of abuse are limited by the law. As in other fields, such as the pharmaceutical industry (in the development of new medicines, for example), regulation does not simply rein in the efforts of statisticians; it also stimulates their activity, as in banking engineering (the new Basel II solvency ratio). It should be noted that this activity is one of those which is still creating employment and that the recent financial crisis has shown the necessity for greater regulation and better risk evaluation.

So it is particularly useful that the specialist literature is now supplemented by a clear, concise and comprehensive treatise on this subject. This book is the fruit of reflection, teaching and professional experience acquired over many years.

Technical matters are tackled with the necessary rigour, but without excessive use of mathematics, enabling any reader to find both pleasure and instruction here. The chapters are also illustrated with numerous examples, usually processed with SAS software (the author provides the syntax for each example), or in some cases with SPSS and R.

Although there is an emphasis on established methods such as factor analysis, linear regression, Fisher's discriminant analysis, logistic regression, decision trees, hierarchical or partitioning clustering, the latest methods are also covered, including robust regression, neural networks, support vector machines, genetic algorithms, boosting, arcing, and the like. Association detection, a data mining method widely used in the retail and distribution industry for market basket analysis, is also described. The book also touches on some less

familiar, but proven, methods such as the clustering of qualitative data by similarity aggregation. There is also a detailed explanation of the evaluation and comparison of scoring models, using the ROC curve and the lift curve. In every case, the book provides exactly the right amount of theoretical underpinning (the details are given in an appendix) to enable the reader to understand the methods, use them in the best way, and interpret the results correctly.

While all these methods are exciting, we should not forget that exploration, examination and preparation of data are the essential prerequisites for any satisfactory modelling. One advantage of this book is that it investigates these matters thoroughly, making use of all the statistical tests available to the user.

An essential contribution of this book, as compared with conventional courses in statistics, is that it provides detailed examples of how data mining forms part of a business strategy, and how it relates to information technology and the marketing of databases or other partners. Where customer relationship management is concerned, the author correctly points out that data mining is only one element, and the harmonious operation of the whole system is a vital requirement. Thus he touches on questions that are seldom raised, such as: What do we do if there are not enough data (there is an entertaining section on ‘forename scoring’)? What is a generic score? What are the conditions for correct deployment in a business? How do we evaluate the return on investment? To guide the reader, Chapter 2 also provides a summary of the development of a data mining project.

Another useful chapter deals with software; in addition to its practical usefulness, this contains an interesting comparison of the three major competitors, namely R, SAS and SPSS.

Finally, the reader may be interested in two new data mining applications: text mining and web mining.

In conclusion, I am sure that this very readable and instructive book will be valued by all practitioners in the field of statistics for decision making and data mining.

Gilbert Saporta

Chair of Applied Statistics

National Conservatory of Arts and Industries, Paris

List of trademarks

SAS®, SAS/STAT®, SAS/GRAPH®, SAS/Insight®, SAS/OR®, SAS/IML®, SAS/ETS®, SAS® High-Performance Forecasting, SAS® Enterprise Guide, SAS® Enterprise Miner™, SAS® Text Miner and SAS® Web Analytics are trademarks of SAS Institute Inc., Cary, NC, USA.

IBM® SPSS® Statistics, IBM® SPSS® Modeler, IBM® SPSS® Text Analytics, IBM® SPSS® Modeler Web Mining and IBM® SPSS® AnswerTree® are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide.

SPAD® is a trademark of Coheris-SPAD, Suresnes, France.

DATALAB® is a trademark of COMPLEX SYSTEMS, Paris, France.

1

Overview of data mining

This first chapter defines data mining and sets out its main applications and contributions to database marketing, customer relationship management and other financial, industrial, medical and scientific fields. It also considers the position of data mining in relation to statistics, which provides it with many of its methods and theoretical concepts, and in relation to information technology, which provides the raw material (data), the computing resources and the communication channels (the output of the results) to other computer applications and to the users. We will also look at the legal constraints on personal data processing; these constraints have been established to protect the individual liberties of people whose data are being processed. The chapter concludes with an outline of the main factors in the success of a project.

1.1 What is data mining?

Data mining and statistics, formerly confined to the fields of laboratory research, clinical trials, actuarial studies and risk analysis, are now spreading to numerous areas of investigation, ranging from the infinitely small (genomics) to the infinitely large (astrophysics), from the most general (customer relationship management) to the most specialized (assistance to pilots in aviation), from the most open (e-commerce) to the most secret (prevention of terrorism, fraud detection in mobile telephony and bank card applications), from the most practical (quality control, production management) to the most theoretical (human sciences, biology, medicine and pharmacology), and from the most basic (agricultural and food science) to the most entertaining (audience prediction for television). From this list alone, it is clear that the applications of data mining and statistics cover a very wide spectrum. The most relevant fields are those where large volumes of data have to be analysed, sometimes with the aim of rapid decision making, as in the case of some of the examples given above. Decision assistance is becoming an objective of data mining and statistics; we now expect these techniques to do more than simply provide a model of reality to help us to understand it. This approach is not completely new, and is already established in medicine, where some treatments have been developed on the basis of statistical analysis, even though the biological mechanism of the disease is little understood because of its

complexity, as in the case of some cancers. Data mining enables us to limit human subjectivity in decision-making processes, and to handle large numbers of files with increasing speed, thanks to the growing power of computers.

A survey on the www.kdnuggets.com portal in July 2005 revealed the main fields where data mining is used: banking (12%), customer relationship management (12%), direct marketing (8%), fraud detection (7%), insurance (6%), retail (6%), telecommunications (5%), scientific research (4%), and health (4%).

In view of the number of economic and commercial applications of data mining, let us look more closely at its contribution to ‘customer relationship management’.

In today’s world, the wealth of a business is to be found in its customers (and its employees, of course). Customer share has replaced market share. Leading businesses have been valued in terms of their customer file, on the basis that each customer is worth a certain (large) amount of euros or dollars. In this context, understanding the expectations of customers and anticipating their needs becomes a major objective of many businesses that wish to increase profitability and customer loyalty while controlling risk and using the right channels to sell the right product at the right time. To achieve this, control of the information provided by customers, or information about them held by the company, is fundamental. This is the aim of what is known as customer relationship management (CRM). CRM is composed of two main elements: operational CRM and analytical CRM.

The aim of analytical CRM is to extract, store, analyse and output the relevant information to provide a comprehensive, integrated view of the customer in the business, in order to understand his profile and needs more fully. The raw material of analytical CRM is the data, and its components are the data warehouse, the data mart, multidimensional analysis (online analytical processing¹), data mining and reporting tools.

For its part, operational CRM is concerned with managing the various channels (sales force, call centres, voice servers, interactive terminals, mobile telephones, Internet, etc.) and marketing campaigns for the best implementation of the strategies identified by the analytical CRM. Operational CRM tools are increasingly being interfaced with back office applications, integrated management software, and tools for managing workflow, agendas and business alerts. Operational CRM is based on the results of analytical CRM, but it also supplies analytical CRM with data for analysis. Thus there is a data ‘loop’ between operational and analytical CRM (see Figure 1.1), reinforced by the fact that the multiplication of communication channels means that customer information of increasing richness and complexity has to be captured and analysed.

The increase in surveys and technical advances make it necessary to store ever-greater amounts of data to meet the operational requirements of everyday management, and the global view of the customer can be lost as a result. There is an explosive growth of reports and charts, but ‘too much information means no information’, and we find that we have less and less knowledge of our customers. The aim of data mining is to help us to make the most of this complexity.

It makes use of databases, or, increasingly, data warehouses,² which store the profile of each customer, in other words the totality of his characteristics, and the totality of his past and

¹ Data storage in a cube with n dimensions (a ‘hypercube’) in which all the intersections are calculated in advance, so as to provide a very rapid response to questions relating to several axes, such as the turnover by type of customer and by product line.

² A *data warehouse* is a set of databases with suitable properties for decision making: the data are thematic, consolidated from different production information systems, user-oriented, non-volatile, documented and possibly aggregated.