

# Data Mining and Statistics for Decision Making

**Stéphane Tufféry**

*University of Rennes, France*

Translated by Rod Riesco



A John Wiley & Sons, Ltd., Publication

# Contents

<b>Preface</b>	<b>xvii</b>
<b>Foreword</b>	<b>xxi</b>
<b>Foreword from the French language edition</b>	<b>xxiii</b>
<b>List of trademarks</b>	<b>xxv</b>
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	$k$ -means and dynamic clouds	248
9.9.3	Processing qualitative data	249
9.9.4	$k$ -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 $\chi^2$ test	659
A.2.11	Example of use of the $\chi^2$ 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
<b>Index</b>		<b>685</b>