

Contents

Preface	vii
---------	-----

I Explorations	1
-----------------------	----------

1 Introduction	3
-----------------------	----------

1.1 Data Mining Beginnings	5
1.2 The Data Mining Team	5
1.3 Agile Data Mining	6
1.4 The Data Mining Process	7
1.5 A Typical Journey	8
1.6 Insights for Data Mining	9
1.7 Documenting Data Mining	10
1.8 Tools for Data Mining: R	10
1.9 Tools for Data Mining: Rattle	11
1.10 Why R and Rattle?	13
1.11 Privacy	17
1.12 Resources	18

2 Getting Started	21
--------------------------	-----------

2.1 Starting R	22
2.2 Quitting Rattle and R	24
2.3 First Contact	25
2.4 Loading a Dataset	26
2.5 Building a Model	28
2.6 Understanding Our Data	31
2.7 Evaluating the Model: Confusion Matrix	35
2.8 Interacting with Rattle	39
2.9 Interacting with R	43

2.10	Summary	54
2.11	Command Summary	55
3	Working with Data	57
3.1	Data Nomenclature	58
3.2	Sourcing Data for Mining	61
3.3	Data Quality	62
3.4	Data Matching	63
3.5	Data Warehousing	65
3.6	Interacting with Data Using R	68
3.7	Documenting the Data	71
3.8	Summary	73
3.9	Command Summary	74
4	Loading Data	75
4.1	CSV Data	76
4.2	ARFF Data	82
4.3	ODBC Sourced Data	84
4.4	R Dataset—Other Data Sources	87
4.5	R Data	90
4.6	Library	91
4.7	Data Options	93
4.8	Command Summary	97
5	Exploring Data	99
5.1	Summarising Data	100
5.1.1	Basic Summaries	101
5.1.2	Detailed Numeric Summaries	103
5.1.3	Distribution	105
5.1.4	Skewness	105
5.1.5	Kurtosis	106
5.1.6	Missing Values	106
5.2	Visualising Distributions	108
5.2.1	Box Plot	110
5.2.2	Histogram	114
5.2.3	Cumulative Distribution Plot	116
5.2.4	Benford's Law	119
5.2.5	Bar Plot	120
5.2.6	Dot Plot	121

5.2.7	Mosaic Plot	122
5.2.8	Pairs and Scatter Plots	123
5.2.9	Plots with Groups	127
5.3	Correlation Analysis	128
5.3.1	Correlation Plot	128
5.3.2	Missing Value Correlations	132
5.3.3	Hierarchical Correlation	133
5.4	Command Summary	135
6	Interactive Graphics	137
6.1	Latticist	138
6.2	GGobi	141
6.3	Command Summary	148
7	Transforming Data	149
7.1	Data Issues	149
7.2	Transforming Data	153
7.3	Rescaling Data	154
7.4	Imputation	161
7.5	Recoding	164
7.6	Cleanup	167
7.7	Command Summary	167
II	Building Models	169
8	Descriptive and Predictive Analytics	171
8.1	Model Nomenclature	172
8.2	A Framework for Modelling	172
8.3	Descriptive Analytics	175
8.4	Predictive Analytics	175
8.5	Model Builders	176
9	Cluster Analysis	179
9.1	Knowledge Representation	180
9.2	Search Heuristic	181
9.3	Measures	182
9.4	Tutorial Example	185
9.5	Discussion	189
9.6	Command Summary	191

10 Association Analysis	193
10.1 Knowledge Representation	194
10.2 Search Heuristic	195
10.3 Measures	196
10.4 Tutorial Example	197
10.5 Command Summary	203
11 Decision Trees	205
11.1 Knowledge Representation	206
11.2 Algorithm	208
11.3 Measures	212
11.4 Tutorial Example	215
11.5 Tuning Parameters	230
11.6 Discussion	241
11.7 Summary	242
11.8 Command Summary	243
12 Random Forests	245
12.1 Overview	246
12.2 Knowledge Representation	247
12.3 Algorithm	248
12.4 Tutorial Example	249
12.5 Tuning Parameters	261
12.6 Discussion	264
12.7 Summary	267
12.8 Command Summary	268
13 Boosting	269
13.1 Knowledge Representation	270
13.2 Algorithm	270
13.3 Tutorial Example	272
13.4 Tuning Parameters	285
13.5 Discussion	285
13.6 Summary	290
13.7 Command Summary	291
14 Support Vector Machines	293
14.1 Knowledge Representation	294
14.2 Algorithm	297

14.3 Tutorial Example	299
14.4 Tuning Parameters	302
14.5 Command Summary	304

III Delivering Performance 305

15 Model Performance Evaluation 307

15.1 The Evaluate Tab: Evaluation Datasets	308
15.2 Measure of Performance	312
15.3 Confusion Matrix	314
15.4 Risk Charts	315
15.5 ROC Charts	320
15.6 Other Charts	320
15.7 Scoring	321

16 Deployment 323

16.1 Deploying an R Model	323
16.2 Converting to PMML	325
16.3 Command Summary	327

IV Appendices 329

A Installing Rattle 331

B Sample Datasets 335

B.1 Weather	336
B.1.1 Obtaining Data	336
B.1.2 Data Preprocessing	339
B.1.3 Data Cleaning	339
B.1.4 Missing Values	341
B.1.5 Data Transforms	343
B.1.6 Using the Data	345
B.2 Audit	347
B.2.1 The Adult Survey Dataset	347
B.2.2 From Survey to Audit	348
B.2.3 Generating Targets	349
B.2.4 Finalising the Data	354
B.2.5 Using the Data	354

B.3 Command Summary	354
References	357
Index	365

<http://www.springer.com/978-1-4419-9889-7>

Data Mining with Rattle and R

The Art of Excavating Data for Knowledge Discovery

Williams, G.

2011, XX, 374 p. 95 illus., 80 illus. in color., Softcover

ISBN: 978-1-4419-9889-7