

**Chapman & Hall/CRC**  
**Data Mining and Knowledge Discovery Series**

# **Data Mining with R**

## **Learning with Case Studies**

™

**Luis Torgo**

**CRC Press**

Taylor & Francis Group  
Boca Raton London New York

CRC Press is an imprint of the  
Taylor & Francis Group, an **informa** business

A CHAPMAN & HALL BOOK

V.

# Contents

<b>Preface</b>	<b>ix</b>
<b>Acknowledgments</b>	<b>xi</b>
<b>List of Figures</b>	<b>xiii</b>
<b>List of Tables</b>	<b>xv</b>
<b>1 Introduction</b>	<b>1</b>
1.1 How to Read This Book?	2
1.2 A Short Introduction to R	3
1.2.1 Starting with R	3
1.2.2 R Objects	5
1.2.3 Vectors	7
1.2.4 Vectorization	10
1.2.5 Factors	11
1.2.6 Generating Sequences	14
1.2.7 Sub-Setting	16
1.2.8 Matrices and Arrays	19
1.2.9s, Lists	23
1.2.10 Data Frames	26
1.2.11 Creating New Functions	30
1.2.12 Objects, Classes, and Methods	33
1.2.13 Managing Your Sessions	34
1.3 A Short Introduction to MySQL	35
<b>2 Predicting Algae Blooms</b>	<b>39</b>
2.1 Problem Description and Objectives	39
2.2 Data Description	40
2.3 Loading the Delta into R	41
2.4 Data Visualisation and Summarization	43
2.5 Unknown Values	52
2.5.1 Removing the Observations with Unknown Values	53
2.5.2 Filling in the Unknowns with the Most Frequent Values	55
2.5.3 Filling in the Unknown Values by Exploring Correlations	56

2.5.4	Filling in the Unknown Values by Exploring Similarities between Cases	60
2.6	Obtaining Prediction Models	63
2.6.1	Multiple Linear Regression	64
2.6.2	Regression Trees	71
2.7	Model Evaluation and Selection . . . T	77
2.8	Predictions for the Seven Algae	91
2.9	Summary	94
<b>Predicting Stock Market Returns</b>		<b>95</b>
3.1	Problem Description and Objectives	95
3.2	The Available Data	96
3.2.1	Handling Time-Dependent Data in R	97
3.2.2	Reading the Data from the CSV File	101
3.2.3	Getting the Data from the Web	102
3.2.4	Reading the Data from a MySQL Database	104
3.2.4.1	Loading the Data into R Ruffrning on Windows	105
3.2.4.2	Loading the Data into R Running on Linux .	107
3.3	Defining the Prediction Tasks	108
3.3.1	What to Predict?	108
3.3.2	Which Predictors?	<b>III</b>
3.3.3	The Prediction Tasks	117
3.3.4	Evaluation Criteria	118
3.4	The Prediction Models	120
3.4.1	How Will the Training Data Be Used?	121
3.4.2	The Modeling Tools	123
3.4.2.1	Artificial Neural Networks	123
3.4.2.2	Support Vector Machines	126
3.4.2.3	Multivariate Adaptive Regression Splines . .	129
3.5	From Predictions into Actions	130
3.5.1	How Will the Predictions Be Used?	130
3.5.2	Trading-Related Evaluation Criteria	132
3.5.3	Putting Everything Together: A Simulated Trader . .	133
3.6	Model Evaluation and Selection	141
3.6.1	Monte Carlo Estimates	141
3.6.2	Experimental Comparisons	143
3.6.3	Results Analysis	148
3.7	The Trading System	156
3.7.1	Evaluation of the Final Test Data	156
3.7.2	An Online Trading System	162
3.8	Summary	163

<b>4</b>	<b>Detecting Fraudulent Transactions</b>	<b>165</b>
4.1	Problem Description and Objectives	165
4.2	The Available Data	166
4.2.1	Loading the Data into R	166
4.2.2	Exploring the Dataset	167
4.2.3	Data Problems	174
	4.2.3.1 Unknown Values	175
	4.2.3.2 Few Transactions of Some Products	179
4.3	Defining the Data Mining Tasks	183
4.3.1	Different Approaches to the Problem	183
	4.3.1.1 Unsupervised Techniques	184
	4.3.1.2 Supervised Techniques	185
	4.3.1.3 Semi-Supervised Techniques	186
4.3.2	Evaluation Criteria	187
	4.3.2.1 Precision and Recall	188
	4.3.2.2 Lift Charts and Precision/Recall Curves . . .	188
	4.3.2.3 Normalized Distance to Typical Price - - - -	193
4.3.3	Experimental Methodology	194
4.4	Obtaining Outlier Rankings	195
4.4.1	Unsupervised Approaches	196
	4.4.1.1 The Modified Box Plot Rule	196
	4.4.1.2 Local Outlier Factors (LOF)	201
	4.4.1.3 Clustering-Based Outlier Rankings (OR,h)	205
4.4.2	Supervised Approaches	208
	4.4.2.1 The Class Imbalance Problem	209
	4.4.2.2 Naive Bayes	211
	4.4.2.3 AdaBoost	217
4.4.3	Semi-Supervised Approaches	223
4.5	Sumniary	230
<b>5</b>	<b>Classifying Microarray Samples</b>	<b>233</b>
5.1	Problem Description and Objectives	233
	5.1.1 Brief Background on Microarray Experiments	233
	5.1.2 The ALL Dataset	234
5.2	The Available Data	235
	5.2.1 Exploring the Dataset	238'
5.3	Gene (Feature) Selection	241
	5.3.1 Simple Filters Based on Distribution Properties - - - -	241
	5.3.2 ANOVA Filters	244
	5.3.3 Filtering Using Random Forests	246
	5.3.4 Filtering Using Feature Clustering Ensembles	248
5.4	Predicting Cytogenetic Abnormalities	251
	5.4.1 Defining the Prediction Task	251
	5.4.2 The Evaluation Metric	252
	5.4.3 The Experimental Procedure	253

5.4.4	The Modeling Techniques	254
5.4.4.1	Random Forests	254
5.4.4.2	$k$ -Nearest Neighbors	255
5.4.5	Comparing the Models	258
5.5	Summary	267

TV

<b>Bibliography</b>	<b>269</b>
<b>Subject Index</b>	<b>279</b>
<b>Index of Data Mining Topics</b>	<b>285</b>
<b>Index of R Functions</b>	<b>287</b>