

Contents

List of Tutorials on the Elsevier Companion Web Page	xi
Foreword 1 for 1st Edition	xiii
Foreword 2 for 1st Edition	xv
Preface	xvii
Introduction	xxi
Frontispiece	xxv
Biographies of the Primary Authors of This Book	xxvii

I

HISTORY OF PHASES OF DATA ANALYSIS, BASIC THEORY, AND THE DATA MINING PROCESS

1. The Background for Data Mining Practice

Preamble	3
Data Mining or Predictive Analytics?	4
A Short History of Statistics and Predictive Analytics	6
Modern Statistics: A Duality?	6
Two Views of Reality	11
The Rise of Modern Statistical Analysis: The Second Generation	13
Machine Learning Methods: The Third Generation	15
Statistical Learning Theory: The Fourth Generation	16
Reinforced and Deep Learning	18
Current Trends of Development in Predictive Analytics	18
Postscript	19
References	19

2. Theoretical Considerations for Data Mining

Preamble	21
The Scientific Method	21
What Is Data Mining?	22
A Theoretical Framework for the Data Mining Process	24
Strengths of the Data Mining Process	25
Customer-Centric Versus Account-Centric: A New Way to Look at Your Data	25
The Data Paradigm Shift	27
Creation of the Car	28
Major Activities of Data Mining	28
Major Challenges of Data Mining	30
General Examples of Data Mining Applications	31
Major Issues in Data Mining	31
General Requirements for Success in a Data Mining Project	33
Example of a Data Mining Project: Classify a Bat's Species by Its Sound	33
The Importance of Domain Knowledge	35
Postscript	35
References	36
Further Reading	37

3. The Data Mining and Predictive Analytic Process

Preamble	39
The Science of Data Mining/Predictive Analytics	39
The Approach to Understanding and Problem Solving	40
CRISP-DM	40
Business Understanding (Mostly Art)	42
Data Understanding (Mostly Science)	44
Data Preparation (A Mixture of Art and Science)	47
Modeling (A Mixture of Art and Science)	47
Deployment (Mostly Art)	52

Closing the Information Loop (Art)	52
The Art of Data Mining	52
Postscript	53
References	54

4. Data Understanding and Preparation

Preamble	55
Activities of Data Understanding and Preparation	55
Issues That Should Be Resolved	56
Data Understanding	57
Postscript	81
References	82
Further Reading	82

5. Feature Selection

Preamble	83
Variables as Features	83
Types of Feature Selection	84
Feature Ranking Methods	84
Subset Selection Methods	88
Postscript	97
References	97

6. Accessory Tools for Doing Data Mining

Preamble	99
Data Access Tools	100
Data Exploration Tools	102
Modeling Management Tools	108
Modeling Analysis Tools	110
In-place Data Processing (IDP)	113
Rapid Deployment of Predictive Models	115
Model Monitors	117
Postscript	117
Further Reading	117

II

THE ALGORITHMS AND METHODS IN DATA MINING AND PREDICTIVE ANALYTICS AND SOME DOMAIN AREAS

7. Basic Algorithms for Data Mining:

A Brief Overview

Preamble	121
Introduction	121

Generalized Additive Models (GAM)	136
Classification and Regression Trees (CART)	138
Generalized EM and <i>k</i> -Means Cluster Analysis—An Overview	145
Postscript	147
References	147
Further Reading	147

8. Advanced Algorithms for Data Mining

Preamble	149
Introduction	150
Advanced Data Mining Algorithms	151
Quality Control Data Mining and Root Cause Analysis	166
Postscript	167
References	167
Further Reading	167

9. Classification

Preamble	169
What Is Classification?	169
Initial Operations in Classification	169
Major Issues With Classification	170
Assumptions of Classification Procedures	171
Analyzing Imbalanced Data Sets With Machine Learning Programs	172
Phases in the Operation of Classification Algorithms	172
Advantages and Disadvantages of Common Classification Algorithms	174
CHAID	177
Which Algorithm Is Best for Classification?	185
Postscript	186
References	186
Further Reading	186

10. Numerical Prediction

Preamble	187
Linear Response Analysis and the Assumptions of the Parametric Model	188
Parametric Statistical Analysis	188
Assumptions of the Parametric Model	189
Linear Regression	192
Generalized Linear Model (GLM)	195
Methods for Analyzing Nonlinear Relationships	198
Nonlinear Regression and Estimation	198
Data Mining and Machine Learning Algorithms Used in Numerical Prediction	201
Advantages of Classification and Regression Trees (CART) Methods	205

Application to Mixed Models 207
 Neural Nets for Prediction 208
 Support Vector Machines (SVMS) and Other Kernel
 Learning Algorithms 211
 Postscript 212
 References 213

11. Model Evaluation and Enhancement

Preamble 215
 Evaluation and Enhancement: Part of the Modeling
 Process 215
 Types of Errors in Analytical Models 216
 Model Enhancement Techniques 227
 Model Enhancement Checklist 231
 Postscript 232
 References 232

12. Predictive Analytics for Population Health and Care

Preamble 235
 The Future of Healthcare, and How Predictive
 Analytics Fits 235
 Predictive Analytics and Population Health 246
 Predictive Analytics and Precision
 Medicine 253
 Postscript 257
 References 257
 Further Reading 258

13. Big Data in Education: New Efficiencies for Recruitment, Learning, and Retention of Students and Donors

ANDY PETERSON

Preamble 259
 Introduction 259
 Industrial Integration of Educational Psychology and
 Big Data Analytics 274
 Postscript 275
 References 276
 Further Reading 277

14. Customer Response Modeling

Preamble 279
 Early CRM Issues in Business 279
 Knowing How Customers Behaved Before They
 Acted 280
 CRM in Business Ecosystems 281
 Conclusions 287

Postscript 288
 References 288

15. Fraud Detection

Preamble 289
 Issues With Fraud Detection 289
 How Do You Detect Fraud? 292
 Supervised Classification of Fraud 293
 How Do You Model Fraud? 294
 How Are Fraud Detection Systems Built? 295
 Intrusion Detection Modeling 296
 Comparison of Models With and Without
 Time-Based Features 297
 Building Profiles 301
 Deployment of Fraud Profiles 302
 Postscript 302
 References 302

III

TUTORIALS AND CASE STUDIES

Tutorial A Example of Data Mining Recipes Using Windows 10 and Statistica 13

LINDA A. MINER

Tutorial B Using the Statistica Data Mining Workspace Method for Analysis of Hurricane Data (Hurrdata.sta)

JEFF WONG

Tutorial C Case Study—Using SPSS Modeler and STATISTICA to Predict Student Success at High-Stakes Nursing Examinations (NCLEX)

GALINA BELOKUROVA, CHIARINA PIAZZA

Introduction 335
 Decision Management in Nursing Education 336
 Case Study 337
 Research Question 337
 Literature Review 337
 Dataset and Expected Strength of Predictors 338
 Data Mining With SPSS Modeler 339

Data Mining With STATISTICA 348
 Conclusion 355
 References 356
 Further Reading 357

Tutorial D Constructing a Histogram
 in KNIME Using MidWest Company
 Personality Data
 LINDA A. MINER

Tutorial E Feature Selection in KNIME
 BOB NISBET

Why Select Features? 377
 Occam's Razor—Simple, But Not
 Simplistic 377
 Local Minimum Error 378
 Moving Out of the Local Minimum 379
 Strategies for Reduction of Dimensionality
 in Predictive Analytics Available in
 KNIME 379

Tutorial F Medical/Business Tutorial
 LINDA A. MINER

Tutorial G A KNIME Exercise,
 Using Alzheimer's Training Data of
 Tutorial F
 LINDA A. MINER

Introduction 423
 KNIME Project 423
 Getting the Program to Open Microsoft
 Excel CSV File: Alzheimer Training
 Data 426
 Decision Trees Node 428
 Linear Correlation Node 432
 Conditional Box Plot Node 436
 Decision Trees Again 438
 End Note 442

Tutorial H Data Prep 1-1: Merging
 Data Sources
 ROBERTA BORTOLOTTI, MSIS, CBAP

Tutorial I Data Prep 1-2: Data
 Description
 ROBERTA BORTOLOTTI, MSIS, CBAP

Tutorial J Data Prep 2-1: Data
 Cleaning and Recoding
 ROBERTA BORTOLOTTI

Tutorial K Data Prep 2-2: Dummy
 Coding Category Variables
 ROBERTA BORTOLOTTI

Tutorial L Data Prep 2-3: Outlier
 Handling
 ROBERTA BORTOLOTTI

Tutorial M Data Prep 3-1: Filling
 Missing Values With Constants
 ROBERTA BORTOLOTTI

Tutorial N Data Prep 3-2: Filling Missing
 Values With Formulas
 ROBERTA BORTOLOTTI

Tutorial O Data Prep 3-3: Filling
 Missing Values With a Model
 ROBERTA BORTOLOTTI

Tutorial P City of Chicago Crime
 Map: A Case Study Predicting Certain
 Kinds of Crime Using Statistica
 Data Miner and Text Miner
 ENDRIN TUSHE

Data Analysis 599
 Text Mining 606
 Boosted Trees 614

Tutorial Q Using Customer Churn Data to Develop and Select a Best Predictive Model for Client Defection Using STATISTICA Data Miner 13 64-bit for Windows 10

RICHARD PORTER WITH ASSISTANCE OF ROBERT NISBET,
LINDA A. MINER, GARY MINER

About This Tutorial 627
Business Objectives 627
Data Preparation 630
Feature Selection 642
Building a Predictive Model With STATISTICA Data Miner DMRecipes 646
Model Evaluation 648

Tutorial R Example With C&RT to Predict and Display Possible Structural Relationships

GREG ROBINSON, LINDA A. MINER, MARY A. MILLIKIN

References 674

Tutorial S Clinical Psychology: Making Decisions About Best Therapy for a Client

LINDA A. MINER

IV

MODEL ENSEMBLES, MODEL COMPLEXITY; USING THE RIGHT MODEL FOR THE RIGHT USE, SIGNIFICANCE, ETHICS, AND THE FUTURE, AND ADVANCED PROCESSES

16. The Apparent Paradox of Complexity in Ensemble Modeling

JOHN ELDER, ANDY PETERSON

Preamble 705
Introduction 706
Model Ensembles 706
How Measure Model Complexity? 709

Generalized Degrees of Freedom 711
Examples: Decision Tree Surface With Noise 712
Summary and Discussion 715
Postscript 716
Acknowledgment 717
References 717
Further Reading 718

17. The “Right Model” for the “Right Purpose”: When Less Is Good Enough

Preamble 719
More Is Not Necessarily Better: Lessons From Nature and Engineering 720
Postscript 726
References 726

18. A Data Preparation Cookbook

Preamble 727
Introduction 727
CRISP-DM—Business Understanding Phase 728
CRISP-DM—Data Understanding Phase 729
CRISP-DM—Data Preparation Phase 732
CRISP-DM—Modeling Phase 736
18 Common Mistakes in Data Preparation in Predictive Analytics Projects 736
Postscript 739
References 740

19. Deep Learning

Preamble 741
The Guiding Concept of DL Technology—Human Cognition 742
Early Artificial Neural Networks (ANNs) 743
How ANNs Work 745
More Elaborate Architectures—DL Neural Networks 746
Postscript 750
References 751
Further Reading 751

20. Significance versus Luck in the Age of Mining: The Issues of *P*-Value “Significance” and “Ways to Test Significance of Our Predictive Analytic Models”

Preamble 753
Introduction 753

The Problem of Significance in Traditional P-Value Statistical Analysis	754
USUAL Data Mining/Predictive Analytic Performance Measures—Terminology	759
Unique Ways to Test Accuracy (“Significance”) of Machine Learning Predictive Models	760
Compare Predictive Model Performance Against Random Results With Lift Charts and Decile Tables	760
Evaluate the Validity of Your Discovery With Target Shuffling	762
Test Predictive Model Consistency With Bootstrap Sampling	763
Postscript	764
References	765

21. Ethics and Data Analytics

ANDY PETERSON

Preamble	767
The Birthday Party—A Practical Example for Ethical Action	767
Academic Secular Ethics	768
Ethics and Data Science for the Norms of Government (Deontological-Normative)	769
Ethics and Data Science for the Goals in Business (Situational-Teleological)	769

Ethics and Data Science for the Virtues of Personal Life (Existential-Motivational)	769
Combination: Right Standards, Right Goals, and Personal Virtue (Normative, Situational, Existential)	770
Michael Sandel on “Doing The Right Thing” With Data Analytics	770
Discovering Data Ethics in an “Alignment Methodology”	771
References	772
Further Reading	772

22. IBM Watson

Preamble	773
Introduction	773
What Exactly Is Watson?	773
Jeopardy!	774
Internal Features of Watson	774
Application Programming Interfaces (APIs)	776
Software Development Kits (SDKs)	778
Some Existing Applications of Watson Techology	778
Ushering in the Cognitive Era	780
Postscript	780
Reference	781

Index 783