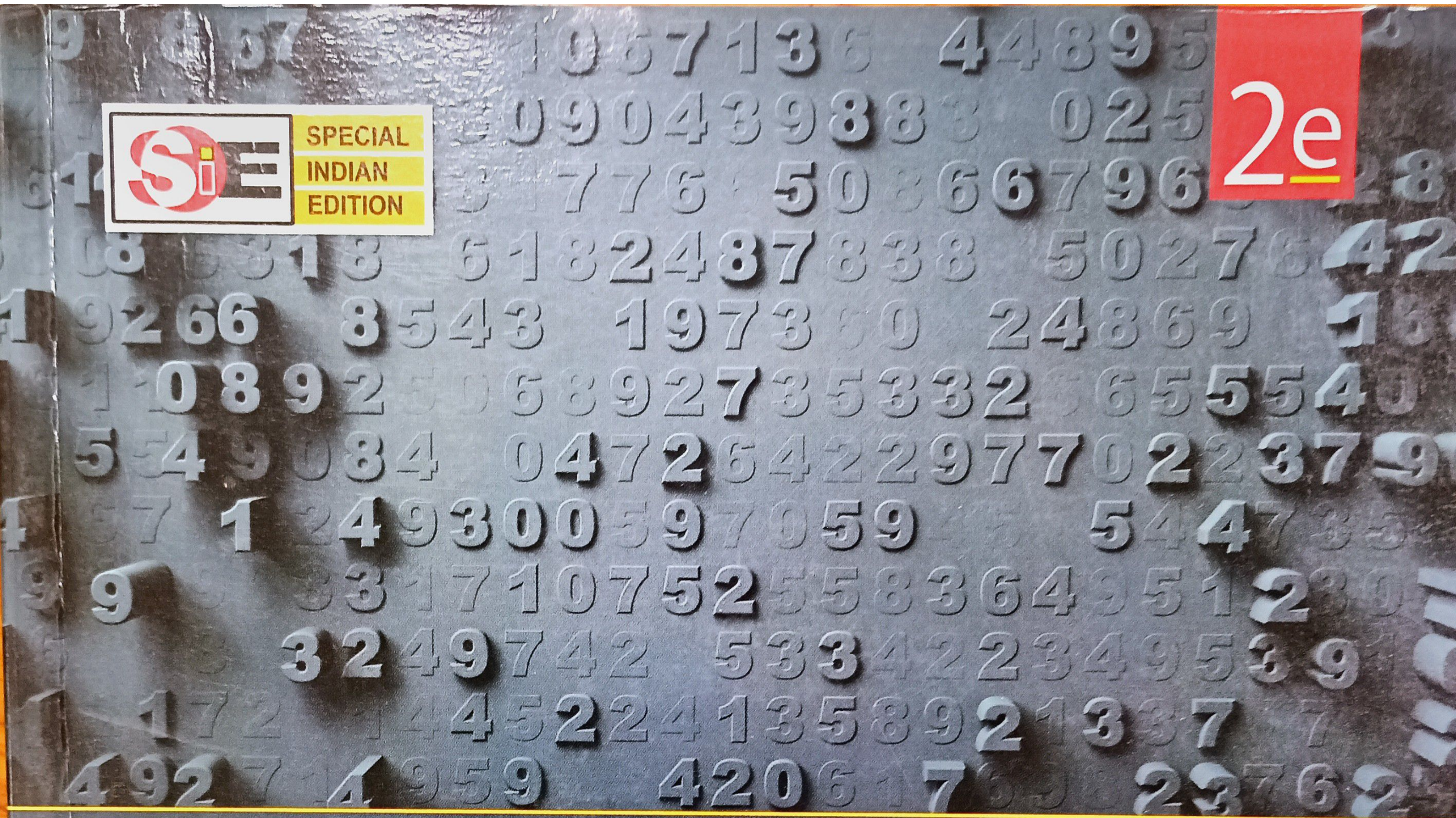

SPECIAL INDIAN EDITION

2e

BUSINESS ANALYTICS

COMMUNICATING WITH NUMBERS

Sanjiv Jaggia
Kevin Lertwachara
Alison Kelly
Leida Chen

Mc Graw Hill

CONTENTS

CHAPTER 1

INTRODUCTION TO BUSINESS ANALYTICS 2

1.1 Overview of Business Analytics 4
- Important Business Analytics Applications 5
- Data Privacy and Data Ethics 6

1.2 Types of Data 8
- Primary and Secondary Data 9
- Cross-Sectional and Time Series Data 9
- Structured and Unstructured Data 10
- Big Data 12

1.3 Variables and Scales of Measurement 15
- The Measurement Scales 16

1.4 Data Sources and File Formats 20
- Fixed-Width Format 21
- Delimited Format 21
- eXtensible Markup Language 22
- HyperText Markup Language 22
- JavaScript Object Notation 23

1.5 Writing with Big Data 25

CHAPTER 2

SUMMARY MEASURES 28

2.1 Measures of Location 31
- Measures of Central Location 31
- Other Measures of Location 36

2.2 Measures of Dispersion, Shape, and Association 38
- Measures of Dispersion 38
- Measures of Shape 42
- Measures of Association 43

2.3 Detecting Outliers 47

2.4 Writing with Big Data 54

CHAPTER 3

DATA VISUALIZATION 58

3.1 Methods to Visualize a Single Variable 60
- Methods to Visualize a Categorical Variable 60
- Methods to Visualize a Numerical Variable 64
- A Line Chart 69
- Cautionary Comments When Constructing or Interpreting Charts or Graphs 71

3.2 Methods to Visualize the Relationship between Two Variables 76
- Methods to Visualize the Relationship between Two Categorical Variables 76
- A Method to Visualize the Relationship between Two Numerical Variables 82

3.3 Other Data Visualization Methods 86
- A Scatterplot with a Categorical Variable 86
- A Bubble Plot 89
- A Heat Map 90
- Options for Advanced Visualizations 93

3.4 Writing with Big Data 94

Appendix 3.1: Getting Started with Tableau 96

CHAPTER 4

PROBABILITY AND PROBABILITY DISTRIBUTIONS 104

4.1 Probability Concepts and Probability Rules 106
- Events 106
- Assigning Probabilities 108
- Rules of Probability 108

4.2 The total Probability Rule, Bayes' Theorem, and Extensions 114
- The Total Probability Rule 114
- The Bayes' Theorem 114
- Extensions of the Total Probability Rule and Bayes' Theorem 116

4.3 Random Variables and Discrete Probability Distributions 119
- The Discrete Probability Distribution 120
- Summary Measures of a Discrete Random Variable 120

4.4 The Binomial and the Poisson Distributions 122
- The Binomial Distribution 122
- The Poisson Distribution 125

4.5 The Normal Distribution 129

4.6 Writing with Data 134

CHAPTER 5

STATISTICAL INFERENCE 138

5.1 Sampling Distributions 140
- The Sampling Distribution of the Sample Mean 140
- The Sampling Distribution of the Sample Proportion 144

5.2 Estimation 147
- Confidence Interval for the Population Mean μ 149
- Confidence Interval for the Population Proportion p 152

5.3 Hypothesis Testing 155
- Hypothesis Test for the Population Mean μ 158
- Hypothesis Test for the Population Proportion p 164

5.4 More Hypothesis Tests 167
- Hypothesis Test for $\mu_1 - \mu_2$ 167
- Hypothesis Test for μ_D 171

5.5 Writing with Data 177

CHAPTER 6

REGRESSION ANALYSIS 180

6.1 The Linear Regression Model 182
The Components of the Linear Regression Model 182
Estimating the Linear Regression Model 185
Interval Estimates for the Response Variable 189
Categorical Variables with Multiple Categories 192

6.2 Model Selection 197
Goodness-of-Fit Measures 197
Tests of Significance 201
Reporting Regression Results 207

6.3 Model Assumptions and Common Violations 213
Residual Plots 213

6.4 Writing with Big Data 225

CHAPTER 7

MORE TOPICS IN REGRESSION ANALYSIS 228

7.1 Regression Models with Interaction Variables 230
The Interaction of Two Dummy Variables 230
The Interaction of a Dummy Variable and a Numerical Variable 232
The Interaction of Two Numerical Variables 234

7.2 Regression Models for Nonlinear Relationships 241
The Quadratic Regression Model 241
Regression Models with Logarithms 244
Comparing Linear and Log-Transformed Regression Models 251

7.3 Cross-Validation Methods 255
The Holdout Method 256
The *k*-Fold Cross-Validation Method 259

7.4 Writing with Big Data 262

Appendix 7.1: The Caret Package in R for the *k*-fold Cross-Validation Method 265

CHAPTER 8

LOGISTIC REGRESSION 268

8.1 The Linear Probability Model and The Logistic Regression Model 270
The Linear Probability Model 270
The Logistic Regression Model 271

8.2 Odds and Accuracy Rate 277
The Odds of an Event 279
Accuracy of Binary Choice Models 281
Other Performance Measures and Imbalanced Data 283

8.3 Cross-Validation of Binary Choice Models 285
The Holdout Cross-Validation Method 285
The *k*-Fold Cross-Validation Method 289

8.4 Writing with Big Data 293

Appendix 8.1: The Caret Package in R for the *k*-fold Cross-Validation Method 295

CHAPTER 9

FORECASTING WITH TIME SERIES DATA 298

9.1 The Forecasting Process for Time Series 300
Forecasting Methods 301
Model Selection Criteria 301

9.2 Simple Smoothing Techniques 302
The Moving Average Technique 303
The Simple Exponential Smoothing Technique 305

9.3 Linear Regression Models for Trend and Seasonality 309
The Linear Trend Model 309
The Linear Trend Model with Seasonality 311
A Note on Causal Models for Forecasting 313

9.4 Nonlinear Regression Models for Trend and Seasonality 315
The Exponential Trend Model 315
The Polynomial Trend Model 318
Nonlinear Trend Models with Seasonality 319

9.5 Data Partitioning and Model Selection 323

9.6 Advanced Exponential Smoothing Methods 328
The Holt Exponential Smoothing Method 328
The Holt-Winters Exponential Smoothing Method 331

9.7 Writing with Data 334

CHAPTER 10

INTRODUCTION TO DATA MINING 338

10.1 Data Mining Overview 340
The Data Mining Process 340
Supervised and Unsupervised Data Mining 342

10.2 Similarity Measures 344
Similarity Measures for Numerical Variables 345
Similarity Measures for Categorical Variables 349

10.3 Performance Evaluation 353
Data Partitioning 353
Performance Measures for Classification Models 356
Selecting Cutoff Values 359
Performance Charts for Classification Models 361
Performance Evaluation for Prediction Models 367

10.4 Principal Component Analysis 372

10.5 Writing with Big Data 383

CHAPTER 11

SUPERVISED DATA MINING: *k*-NEAREST NEIGHBORS AND NAÏVE BAYES 386

11.1 Introduction to Supervised Data Mining 388
Comparison of Supervised Data Mining Techniques 389

11.2 The *k*-Nearest Neighbors Method 390

11.3 The Naïve Bayes Method 409
Transforming Numerical into Categorical Values 418

11.4 Writing with Big Data 424

CHAPTER 12

SUPERVISED DATA MINING: DECISION TREES 428

12.1 Introduction to Classification and Regression Trees (CART) 430
Classification and Regression Trees (CART) 430
12.2 Classification Trees 433
Developing Classification Tree Models 433
12.3 Regression Trees 458
Developing Regression Tree Models 459
12.4 Ensemble Tree Models 472
12.5 Writing with Big Data 488

CHAPTER 13

UNSUPERVISED DATA MINING 492

13.1 Hierarchical Cluster Analysis 494
An Overview of Cluster Analysis 494
Agglomerative Clustering 496
Agglomerative Clustering with Mixed Data 504
13.2 *k*-Means Cluster Analysis 510
13.3 Association Rule Analysis 519
13.4 Writing with Big Data 531

CHAPTER 14

SPREADSHEET MODELING 536

14.1 Spreadsheet Modeling 538
Developing a Spreadsheet Model 539
14.2 Spreadsheet Engineering and an Influence Diagram 543
14.3 What-If Analysis 551
A Data Table 551
Goal Seek 554
Scenario Manager 555
14.4 Avoiding and Detecting Errors in Spreadsheet Models 560
Common Spreadsheet Model Errors 560
Model Auditing 563
14 .5 Writing with Data 567

CHAPTER 15

RISK ANALYSIS AND SIMULATION 572

15 .1 Overview of Prescriptive Analytics 574
15.2 Risk and Risk Analysis 574
Conducting Risk Analysis 575
Using Random Variables to Model Risk and Uncertainty 578
15.3 Monte Carlo Simulation 592
15.4 Writing with Data 601

CHAPTER 16

OPTIMIZATION: LINEAR PROGRAMMING 604

16.1 Maximization Problems In Linear Programming 606
Formulating a Maximization Problem 606
Solving a Maximization Problem 608
16.2 Minimization Problems in Linear Programming 622
16.3 Special Cases in Linear Programming 628
Multiple Optimal Solutions 628
Infeasibility 629
Unboundedness 629
16.4 Writing with Data 630

CHAPTER 17

MORE APPLICATIONS IN OPTIMIZATION 634

17.1 Optimization with Integer Programming 636
17.2 Other Applications of Linear and Integer Programming 642
Capital Budgeting 642
Transportation Problems 645
Assignment Problems 649
Workforce Scheduling 652
Facility Locations 656
17.3 Optimization with Nonlinear Programming 663
17.4 Writing with Data 670

APPENDIXES

APPENDIX A Data Management and Wrangling 674
APPENDIX B Big Data Sets: Variable Description and Data Dictionary 721
APPENDIX C Getting Started with Excel and Excel Add-Ins 728
APPENDIX D Getting Started with R 733
APPENDIX E Answers to Selected Exercises 741

Index 763