

Acknowledgments	xv
Introduction	xvii
Part One Setting the Focus	1
Chapter 1 Data Mining in Context	5
What Is Data Mining?	7
What Can Data Mining Do?	8
Classification	8
Estimation	9
Prediction	10
Affinity Grouping or Association Rules	10
Clustering	10
Description and Visualization	11
The Business Context for Data Mining	11
Data Mining as a Research Tool	12
Data Mining for Process Improvement	13
Data Mining for Marketing	13
Data Mining for Customer Relationship Management	14
The Technical Context for Data Mining	14
Data Mining and Machine Learning	15
Data Mining and Statistics	15
Data Mining and Decision Support	16
Data Mining and Computer Technology	19
The Societal Context for Data Mining	19
Chapter 2 Why Master the Art?	21
Four Approaches to Data Mining	23
Purchasing Scores	24
Purchasing Software	24
Hiring Outside Experts	32
Developing In-House Expertise	35
Lessons Learned	36
Chapter 3 Data Mining Methodology: The Virtuous Cycle Revisited	39
Two Styles of Data Mining	40
Directed Data Mining	40
Undirected Data Mining	42
The Virtuous Cycle of Data Mining	43
Identifying the Right Business Problem	44
Is the Data Mining Effort Necessary?	45

Is There a Particular Segment or Subgroup That Is Most Interesting?	46
What Are the Relevant Business Rules?	46
What about the Data?	47
Verifying the Opinion of Domain Experts	47
Transforming Data into Actionable Results	48
Identify and Obtain Data	49
Validate, Explore, and Clean the Data	50
Transpose the Data to the Right Granularity	51
Add Derived Variables	52
Prepare the Model Set	53
Choose the Modeling Technique and Train the Model	54
Check Performance of the Models	54
Acting on the Results	57
Measuring the Model's Effectiveness	58
What Makes Predictive Modeling Successful?	59
Time Frames of Predictive Modeling	59
Modeling Shelf-Life	60
Assumption 1: The Past Is a Good Predictor of the Future	61
Assumption 2: The Data Is Available	62
Assumption 3: The Data Contains What We Want to Predict	63
Lessons Learned	64
Chapter 4 Customers and Their Lifecycles	65
Who Is the Customer?	66
Consumers	66
Business Customers	67
Customer Segments	70
The Customer Lifecycle	72
Stages of the Lifecycle	73
Major Lifecycle Events	75
Data Appears at Different Times in the Lifecycle	78
The Customer's Lifecycle	79
Targeting the Right Customers at the Right Time	80
Budget Optimization	80
Campaign Optimization	82
Customer Optimization	86
Lessons Learned	90
Part Two The Three Pillars of Data Mining	91
Chapter 5 Data Mining Techniques and Algorithms	99
Different Goals Call for Different Techniques	100
Different Data Types Call for Different Techniques	102
Three Data Mining Techniques	102

Automatic Cluster Detection	103
How K-Means Cluster Detection Works	104
Consequences of Choosing Clustering	107
Decision Trees	111
How Decision Trees Work	112
How Decision Trees Are Built	113
Consequences of Choosing Decision Trees	119
Neural Networks	121
Training Neural Networks	125
Consequences of Choosing Neural Networks	127
Lessons Learned	129
Chapter 6 Data, Data Everywhere...	131
What Should Data Look Like?	132
The Rows	132
The Columns	134
Roles of Columns in Data Mining	138
Data for Data Mining	140
What Does Data Really Look Like?	141
Where Data Comes From	141
The Right Level of Granularity	150
Different Ways to Measure Data Values	153
How Much Data Is Enough?	157
Derived Variables	158
Issues in Working with Derived Variables	159
Handling Outliers	160
Combinations of Columns	162
Summarizations	163
Extracting Features from Single Columns	164
Time Series	167
Case Study: Defining Customer Behavior	169
Dirty Data	177
Missing Data	177
Fuzzy Definitions	179
Incorrect Values	180
Lessons Learned	181
Chapter 7 Building Effective Predictive Models	183
Building Good Predictive Models	184
A Process for Building Predictive Models	184
Lift as a Measure of Performance	186
Model Stability	191
The Challenge of Model Stability	192
Working with the Model Set	193
Divide and Conquer: Training, Test, and Evaluation Sets	193
How the Size of the Model Set Affects Results	194

How the Density of the Model Set Affects Results	195
Sampling	196
What Is Oversampling?	197
Modeling Time-Dependent Data	201
Model Inputs and Model Outputs	203
Latency: Taking Model Deployment into Account	206
Time and Missing Data	209
Building Models that Easily Shift in Time	210
Naming Fields	212
Using Multiple Models	213
Multiple Model Voting	213
Segmenting the Input	218
Other Reasons to Combine Models	219
Experiment!	221
The Model Set	221
Different Types of Models and Model Parameters	223
Time Frame	223
Lessons Learned	224
Chapter 8 Taking Control: Setting Up a Data Mining Environment	227
Getting Started	228
What Is a Data Mining Environment?	228
Four Case Studies	229
What Makes a Data Mining Environment Successful?	229
Case 1: Building Up a Core Competency Internally	230
Data Mining in the Insurance Industry	231
Getting Started	231
Case 2: Building a New Line of Business	235
Going Online	235
The Environment	236
The Prospect Data Warehouse	236
The Next Step	237
Case 3: Building Data Mining Skills on Data Warehouse Efforts	240
A Special Kind of Data Warehouse	240
The Plan for Data Mining	240
Data Mining in IT	241
Case 4: Data Mining Using Tessera RME	242
Requirements for an Advanced Data Mining Environment	242
What Is RME?	243
How RME Works	244
How RME Helps Prepare Data	246
How RME Supports Sampling	247
How RME Helps in Model Development	248
How RME Helps in Scoring and Managing Models	249
Lessons Learned	251

Part Three Case Studies**253**

Chapter 9	Who Needs Bag Balm and Pants Stretchers	261
	The Vermont Country Store	262
	How Vermont Country Store Got Where It Is Today	263
	Predictive Modeling at Vermont Country Store	265
	The Business Problem	265
	The Data	267
	The Technical Approach	270
	Choice of Software Package	270
	The Baseline—RFM and Segmentation	271
	The Challengers—Neural Networks, Decision Trees, and Regression	273
	Determining What Would Have Happened	276
	Calculating the Return on Investment	276
	The Future	276
	Expected Benefits	277
	Lessons Learned	277
Chapter 10	Who Gets What? Building a Best Next Offer Model for an Online Bank	279
	Gaining Wallet-Share	279
	The Business Problem	281
	The Data	282
	From Accounts to Customers	284
	Defining the Products to Be Offered	287
	Approach to the Problem	290
	Comparable Scores	291
	If It Walks Like a Duck, ...	292
	Pitfalls of This Approach	292
	Building the Models	294
	Building a Decision Tree Model for Brokerage	296
	Building the Rest of the Models	306
	Getting to a Cross-Sell Model	307
	In a More Perfect World	308
	Lessons Learned	308
Chapter 11	Please Don't Go! Churn Modeling in Wireless Communication	311
	The Wireless Telephone Industry	312
	A Rapidly Maturing Industry	313
	Some Differences from Other Industries	314
	The Business Problem	316
	Project Background	316
	Specifics about This Market	317

What Is Churn?	318
Why Is Churn Modeling Useful?	319
Three Goals	320
Approach to Building the Churn Model	322
The Project Itself	325
Building a Churn Model: A Real-Life Application	326
The Choice of Tool	326
Segmenting the Model Set	326
The Final Four (Models)	327
Choice of Modeling Algorithm	331
The Size and Density of the Model Set	338
The Effect of Latency (or Taking Deployment into Account)	339
Translating Models in Time	340
The Data	341
The Basic Customer Model	341
From Telephone Calls to Data	343
Historical Churn Rates	344
Data at the Customer and Account Level	345
Data at the Service Level	345
Data Billing History	346
Rejecting Some Variables	346
Derived Variables	347
Lessons about Building Churn Models	349
Finding the Most Significant Variables	349
Listening to the Business Users	349
Listening to the Data	350
Including Historical Churn Rates	351
Composing the Model Set	352
Building a Model for the Churn Management Application	354
Listening to the Data to Determine Model Parameters	354
Understanding the Algorithm and the Tool	355
Lessons Learned	355
Chapter 12 Converging on the Customer: Understanding Customer Behavior in the Telecommunications Industry	357
Dataflows	358
What Is a Dataflow?	359
Basic Operations	360
Dataflows in a Parallel Environment	361
Why Are Dataflows Efficient?	363
The Business Problem	364
Project Background	364
Important Marketing Questions	365
The Data	366
Call Detail Data	366
Customer Data	368
Auxiliary Files	372

A Voyage of Discovery	372
What Is in a Call Duration?	372
Calls by Time of Day	374
Calls by Market Segment	378
International Calling Patterns	382
When Are Customers at Home?	384
Internet Service Providers	387
Private Networks	388
Concurrent Calls	391
Lessons Learned	393
Chapter 13 Who Is Buying What? Getting to Know Supermarket Shoppers	395
An Industry in Transition	396
Supermarkets as Information Brokers	396
Shifting the Focus from Products to Customers	399
Three Case Studies	401
Analyzing Ethnic Purchasing Patterns	402
Business Background	402
The Data	402
A Triumph for Visualization	405
A Failed Approach	407
Just the Facts	407
Who Buys Yogurt at the Supermarket?	410
Business Background	411
The Data	411
From Groceries to Customers	416
Finding Clusters of Customers	419
Putting the Clusters to Work	421
Who Buys Meat at the Health Food Store?	424
Association Rules for Market Basket Analysis	424
People are More Interesting Than Groceries	429
Lessons Learned	432
Chapter 14 Waste Not, Want Not: Improving Manufacturing Processes	435
Data Mining to Reduce Cost at R. R. Donneley	436
The Technical Problem	436
The Business Problem	437
The Data	438
Inducing Rules for Cylinder Bands	442
Change on the Shop Floor	444
Long-Term Impact	445
Reducing Paper Wastage at Time Inc.	445
The Business Problem	446
The Data	449
Approach to the Problem	452

Types of Waste	453
Addressable Waste	455
Inducing Rules for Addressable Waste	456
Data Transformation	456
Data Characterization and Profiling	459
Decision Trees	459
Association Rules	462
Putting It All Together	463
Lessons Learned	464
Chapter 15 The Societal Context: Data Mining and Privacy	465
The Privacy Prism	466
Is Data Mining a Threat?	468
The Expectation of Privacy	470
The Importance of Privacy	471
Information in the Material World	476
Information in the Electronic World	477
Identifying the Customer	478
Putting It All Together	480
The Promise of Data Mining	483
Index	485