Data Mining Techniques

For Marketing, Sales, and Customer Relationship Management

Third Edition

Gordon S. Linoff
Michael J. A. Berry

WILEY
Wiley Publishing, Inc.

Contents

Introduction xxxvii

Chapter 1 What Is Data Mining and Why Do It? 1

What Is Data Mining? 2
Data Mining Is a Business Process 2
Large Amounts of Data 3
Meaningful Patterns and Rules 3
Data Mining and Customer Relationship Management 4
Why Now? 6
Data Is Being Produced 6
Data Is Being Warehoused 6
Computing Power Is Affordable 7
Interest in Customer Relationship Management Is Strong 7
Every Business Is a Service Business 7
Information Is a Product 7
Commercial Data Mining Software Products Have Become Available 8
Skills for the Data Miner 9
The Virtuous Cycle of Data Mining 9
A Case Study in Business Data Mining 11
Identifying BofA's Business Challenge 12
Applying Data Mining 12
Acting on the Results 13
Measuring the Effects of Data Mining 14
Steps of the Virtuous Cycle 15
Identify Business Opportunities 16
Transform Data into Information 17
Act on the Information 19
Measure the Results 20

Data Mining in the Context of the Virtuous Cycle 23
Lessons Learned 26

Chapter 2 Data Mining Applications in Marketing and Customer Relationship Management 27
Two Customer Lifecycles 27
The Customer's Lifecycle 28
The Customer Lifecycle 28
Subscription Relationships versus Event-Based Relationships 30
Event-Based Relationships 30
Subscription-Based Relationships 31
Organize Business Processes Around the Customer Lifecycle 32
Customer Acquisition 33
Who Are the Prospects? 33
When Is a Customer Acquired? 34
What Is the Role of Data Mining? 35
Customer Activation 36
Customer Relationship Management 37
Winback 38
Data Mining Applications for Customer Acquisition 38
Identifying Good Prospects 39
Choosing a Communication Channel 39
Picking Appropriate Messages 40
A Data Mining Example: Choosing the Right Place to Advertise 40
Who Fits the Profile? 41
Measuring Fitness for Groups of Readers 44
Data Mining to Improve Direct Marketing Campaigns 45
Response Modeling 46
Optimizing Response for a Fixed Budget 47
Optimizing Campaign Profitability 49
Reaching the People Most Influenced by the Message 53
Using Current Customers to Learn About Prospects 54
Start Tracking Customers Before They Become "Customers" 55
Gather Information from New Customers 55
Acquisition-Time Variables Can Predict Future Outcomes 56
Data Mining Applications for Customer Relationship Management 56
Matching Campaigns to Customers 56
Reducing Exposure to Credit Risk 58
Predicting Who Will Default 58
Improving Collections 59
Determining Customer Value 59
Cross-selling, Up-selling, and Making Recommendations 60

Finding the Right Time for an Offer 60
Making Recommendations 60
Retention 60
Recognizing Attrition 60
Why Attrition Matters 61
Different Kinds of Attrition 62
Different Kinds of Attrition Model 63
Predicting Who Will Leave 63
Predicting How Long Customers Will Stay 64
Beyond the Customer Lifecycle 64
Lessons Learned 65

Chapter 3 The Data Mining Process 67

What Can Go Wrong? 68
Learning Things That Aren't True 68
Patterns May Not Represent Any Underlying Rule 69
The Model Set May Not Reflect the Relevant Population 70
Data May Be at the Wrong Level of Detail 71
Learning Things That Are True, but Not Useful 73
Learning Things That Are Already Known (or Should Be Known) 73
Learning Things That Can't Be Used 74
Data Mining Styles 74
Hypothesis Testing 75
Generating Hypotheses 75
Testing Hypotheses Using Existing Data 76
Hypothesis Testing and Experimentation 77
Case Study in Hypothesis Testing: Measuring the Wrong Thing 79
Directed Data Mining 81
Undirected Data Mining 81
Goals, Tasks, and Techniques 82
Data Mining Business Goals 82
Data Mining Tasks 83
Preparing Data for Mining 83
Exploratory Data Analysis 84
Binary Response Modeling (Binary Classification) 85
Classification 85
Estimation 86
Finding Clusters, Associations, and Affinity Groups 86
Applying a Model to New Data 87
Data Mining Techniques 88
Formulating Data Mining Problems: From Goals to Tasks to Techniques 88
Choosing the Best Places to Advertise 89
Determining the Best Product to Offer a Customer 89
Finding the Best Locations for Branches or Stores 90

Segmenting Customers on Future Profitability 91
Decreasing Exposure to Risk of Default 92
Improving Customer Retention 93
Detecting Fraudulent Claims . 95
What Techniques for Which Tasks? 9§
Is There a Target or Targets? 96
What Is the Target Data Like? 96
What Is the Input Data Like? 96
How Important Is Ease of Use? 97
How Important Is Model Explicability? 97
Lessons Learned 98

Chapter 4 Statistics 101: What You Should Know About Data < .101
Occam's Razor 10$
Skepticism and Simpson's Paradox 103
The Null Hypothesis 104
P-Values 105
Looking At and Measuring Data 106
Categorical Values 106
Histograms • 107
Time Series 107
Standardized Values (Z-Scores) 109
From Z-Scores to Probabilities , 114
Cross-Tabulations 116
Numeric Variables 117'
Statistical Measures for Continuous Variables 117
Interesting Properties of the Average and Median 118
Variance and Standard Deviation • 119
A Couple More Statistical Ideas 120
Measuring Response 123
Standard Error of a Proportion 121
Comparing Results Using Confidence Bounds 123
Comparing Results Using Difference of Proportions 124
Size of Sample . 125
What the Confidence Interval Really Means 126
Size of Test and Control for an Experiment 127
Multiple Comparisons 129
The Confidence Level with Multiple Comparisons 129
Bonferroni's Correction 129
Chi-Square Test . 130
Expected Values . 130
Chi-Square Value 132
Comparison of Chi-Square to Difference of Proportions 134
An Example: Chi-Square for Regions and Starts 134
Case Study: Comparing Two Recommendation Systems with an A/B Test 138
First Metric: Participating Sessions . 140

Second Metric: Daily Revenue Per Session 141
Third Metric: Who Wins on Each Day? 142
Fourth Metric: Average Revenue Per Session 143
Fifth Metric: Incremental Revenue Per Customer 143
Data Mining and Statistics 144
No Measurement Error in Basic Data 145
A Lot of Data 146
Time Dependency Pops Up Everywhere 146
Experimentation Is Hard 147
Data Is Censored and Truncated 147
Lessons Learned 148

Chapter 5 Descriptions and Prediction: Profiling and Predictive Modeling 151
Directed Data Mining Models 152
Defining the Model Structure and Target 152
Incremental Response Modeling 154
Model Stability 156
Time-Frames in the Model Set 157
Prediction Models 158
Profiling Models 158
Directed Data Mining Methodology 159
Step 1: Translate the Business Problem into a Data Mining Problem 161
How Will Results Be Used? 163
How Will Results Be Delivered? 163
The Role of Domain Experts and Information Technology 164
Step 2: Select Appropriate Data 165
What Data Is Available? 166
How Much Data Is Enough? 167
How Much History Is Required? 167
How Many Variables? 168
What Must the Data Contain? 168
Step 3: Get to Know the Data 169
Examine Distributions 169
Compare Values with Descriptions 170
Validate Assumptions 170
Ask Lots of Questions 171
Step 4: Create a Model Set 172
Assembling Customer Signatures 172
Creating a Balanced Sample 172
Including Multiple Timeframes 174
Creating a Model Set for Prediction 174
Creating a Model Set for Profiling 176
Partitioning the Model Set 176
Step 5: Fix Problems with the Data 177
Categorical Variables with Too Many Values 177

Numeric Variables with Skewed Distributions and Outliers 178
Missing Values 178
Values with Meanings That Change over Time 179
Inconsistent Data Encoding 179
Step 6: Transform Data to Bring Information to the Surface 180
Step 7: Build Models 180
Step 8: Assess Models 180
Assessing Binary Response Models and Classifiers 181
Assessing Binary Response Models Using Lift 182
Assessing Binary Response Model Scores Using Lift Charts 184
Assessing Binary Response Model Scores Using Profitability Models 185
Assessing Binary Response Models Using ROC Charts 186
Assessing Estimators 188
Assessing Estimators Using Score Rankings 189
Step 9: Deploy Models 190
Practical Issues in Deploying Models 190
Optimizing Models for Deployment 191
Step 10: Assess Results 191
Step 11: Begin Again 193
Lessons Learned 193

Chapter 6 Data Mining Using Classic Statistical Techniques 195
Similarity Models 196
Similarity and Distance 196
Example: A Similarity Model for Product Penetration 197
The Business Problem 197
Data Used for Similarity Model 197
Steps for Building a Similarity Model 199
Step 1: What Distinguishes High Penetration Towns from Low Penetration Towns? 199
Step 2: What Would the Ideal Town Look Like? 201
Step 3: How Far Is Each Town from the Ideal? 201
Evaluating the Similarity Model 202
Table Lookup Models 203
Choosing Dimensions 204
Partitioning the Dimensions 205
From Training Data to Scores 205
Handling Sparse and Missing Data by Removing Dimensions 205
RFM: A Widely Used Lookup Model 206
RFM Cell Migration 207
RFM and the Test-and-Measure Methodology 208
Every Campaign Is an Experiment 208

New Types of Campaigns Should Be Tested Before Being Rolled Out 209
RFM and Incremental Response Modeling 209
Naive Bayesian Models 210
Some Ideas from Probability 210
Probability, Odds, and Likelihood 210
Converting for Convenience 212
The Naive Bayesian Calculation 212
Comparison with Table Lookup Models 213
Linear Regression 213
The Best-fit Line 215
Goodness of Fit 217
Residuals 217
R^2 218
Global Effects 219
Multiple Regression 220
The Equation 220
The Range of the Target Variable 221
Interpreting Coefficients of Linear Regression Equations 221
Capturing Local Effects with Linear Regression 223
Additional Considerations with Multiple Regression 224
Linear Independence 224
Interactions 224
Adding Variables Can Change the Coefficients of Variables Already in the Model 225
Variable Selection for Multiple Regression 225
Forward Selection 226
Stepwise Selection 226
Backward Elimination 227
Logistic Regression 227
Modeling Binary Outcomes 227
Estimating Probabilities with Linear Regression 227
Bending the Regression Line into Shape 228
The Logistic Function 229
Fixed Effects and Hierarchical Effects 231
Hierarchical Effects 232
Within and Between Effects 232
Fixed Effects 233
Lessons Learned 234

Chapter 7 Decision Trees 237
What Is a Decision Tree and How Is It Used? 238
A Typical Decision Tree 238
Using the Tree to Learn About Churn 240
Using the Tree to Learn About Data and Select Variables 241
Using the Tree to Produce Rankings 243
Usine the Tree to Estimate Class Probabilities 243

Using the Tree to Classify Records 244
Using the Tree to Estimate Numeric Values 244
Decision Trees Are Local Models 245
Growing Decision Trees 247
Finding the Initial Split 248
Splitting on a Numeric Input Variable 249
Splitting on a Categorical Input Variable 249
Splitting in the Presence of Missing Values 250
Growing the Full Tree 251
Finding the Best Split 252
Gini (Population Diversity) as a Splitting Criterion 253
Entropy Reduction or Information Gain as a Splitting Criterion 254
Information Gain Ratio 256
Chi-Square Test as a Splitting Criterion 256
Incremental Response as a Splitting Criterion 258
Reduction in Variance as a Splitting Criterion for Numeric Targets 259
FTest 262
Pruning 262
The CART Pruning Algorithm 263
Creating Candidate Subtrees 263
Picking the Best Subtree 266
Pessimistic Pruning: The C5.0 Pruning Algorithm 267
Stability-Based Pruning 268
Extracting Rules from Trees 269
Decision Tree Variations 270
Multiway Splits 270
Splitting on More Than One Field at a Time 271
Creating Nonrectangular Boxes 271
Assessing the Quality of a Decision Tree 275
When Are Decision Trees Appropriate? 276
Case Study: Process Control in a Coffee Roasting Plant 277
Goals for the Simulator 277
Building a Roaster Simulation 278
Evaluation of the Roaster Simulation 278
Lessons Learned 279

Chapter 8 Artificial Neural Networks 281
A Bit of History 282
The Biological Model 283
The Biological Neuron 285
The Biological Input Layer 286
The Biological Output Layer 287
Neural Networks and Artificial Intelligence 287
Artificial Neural Networks 288

The Artificial Neuron 288
Combination Function 288
Transfer Function 288
The Multi-Layer Perceptron 291
A Network Example 292
Network Topologies 293
A Sample Application: Real Estate Appraisal 295
Training Neural Networks 299
How Does a Neural Network Learn Using Back Propagation? 299
Pruning a Neural Network 300
Radial Basis Function Networks 303
Overview of RBF Networks 303
Choosing the Locations of the Radial Basis Functions 305
Universal Approximators 305
Neural Networks in Practice 308
Choosing the Training Set 309
Coverage of Values for All Features 309
Number of Features 310
Size of Training Set 310
Number and Range of Outputs 310
Rules of Thumb for Using MLPs 310
Preparing the Data 311
Interpreting the Output from a Neural Network 313
Neural Networks for Time Series 315
Time Series Modeling 315
A Neural Network Time Series Example 316
Can Neural Network Models Be Explained? 317
Sensitivity Analysis 318
Using Rules to Describe the Scores 318
Lessons Learned 319

Chapter 9 Nearest Neighbor Approaches: Memory-Based Reasoning and Collaborative Filtering 321
Memory-Based Reasoning 322
Look-Alike Models 323
Training and Scoring a Look-Alike Model 323
Look-Alike Models and Paired Tests 324
Example: Using MBR to Estimate Rents in Tuxedo, New York 324
Challenges of MBR 327
Choosing a Balanced Set of Historical Records 328
Representing the Training Data 328
Exhaustive Comparisons 328
R-Tree 329
Reducing the Training Set Size 329
Determining the Distance Function, Combination Function, and Number of Neighbors 331

Case Study: Using MBR for Classifying Anomalies in Mammograms 331
The Business Problem: Identifying Abnormal Mammograms 332
Applying MBR to the Problem 332
The Total Solution **334**
Measuring Distance and Similarity ^®
What Is a Distance Function? **33S**
Building a Distance Function One Field at a Time **337**
Distance Functions for Other Data Types **340**
When a Distance Metric Already Exists **311**
The Combination Function: Asking the Neighbors for Advice **342**
The Simplest Approach: One Neighbor **341**
The Basic Approach for Categorical Targets: Democracy 342
Weighted Voting for Categorical Targets **344**
Numeric Targets **344**
Case Study: Shazam — Finding Nearest Neighbors for Audio Files **345**
Why This Feat Is Challenging **346**
The Audio Signature **347**
Measuring Similarity **348**
Simple Distance Between Constellations 348
Time Slice Similarity 349
Anchor Point Distance 349
Shazam Implementation 350
Collaborative Filtering: A Nearest-Neighbor Approach to Making Recommendations 351
Building Profiles 352
Comparing Profiles 352
Making Predictions 353
Lessons Learned 354

Chapter 10 Knowing When to Worry: Using Survival Analysis to Understand Customers 357
Customer Survival 360
What Survival Curves Reveal 360
Finding the Average Tenure from a Survival Curve 362
Customer Retention Using Survival 364
Looking at Survival as Decay 365
Hazard Probabilities 367
The Basic Idea 368
Examples of Hazard Functions 369
Constant Hazard 369
Bathtub-Shaped Hazard 370
A Real-World Example 370
Censoring 371
The Hazard Calculation 372

Other Types of Censoring] 375
From Hazards to Survival 376
Retention 376
Survival 378
Comparison of Retention and Survival 378
Proportional Hazards 380
Examples of Proportional Hazards 381
Stratification: Measuring Initial Effects on Survival 382
Cox Proportional Hazards 382
The Basic Idea 383
Using Proportional Hazards 384
Limitations of Proportional Hazards 384
Survival Analysis in Practice 385
Handling Different Types of Attrition 385
When Will a Customer Come Back? 387
Understanding Customer Value 389
Doing the Basic Calculation 390
Extending the Ideas to the Money Side 391
Including Customer Migration 391
Forecasting 392
Hazards Changing over Time 393
Lessons Learned 394

Chapter 11 Genetic Algorithms and Swarm Intelligence 397
Optimization 398
What Is an Optimization Problem? 398
An Optimization Problem in Ant World 399
E Pluribus Unum 400
A Smarter Ant 401
Genetic Algorithms 403
A Bit of History 404
Genetics on Computers 404
Selection 409
Crossover 410
Mutation 412
Representing the Genome 413
Schemata: The Building Blocks of Genetic Algorithms 414
What Is a Schema? 414
Schemata as Building Blocks 415
Why Fit Schemata Survive 416
Beyond the Simple Algorithm 417
The Traveling Salesman Problem 418
Exhaustive Search 419
A Simple Greedy Algorithm 419
The Genetic Algorithms Approach 419
The Swarm Intelligence Approach 420
Real Ants 420

Artificial Ants 421
Case Study: Using Genetic Algorithms for Resource Optimization 421
Case Study: Evolving a Solution for Classifying Complaints 423
Business Context **424**
Data **425**
The Comment Signature **425**
The Genomes **426**
The Fitness Function **427**
The Results **427**
Lessons Learned **427**

Chapter 12 Tell Me Something New: Pattern Discovery and Data Mining 429
Undirected Techniques, Undirected Data Mining 431
Undirected versus Directed Techniques **431**
Undirected versus Directed Data Mining **431**
Case Study: Undirected Data Mining Using Directed Techniques **432**
What is Undirected Data Mining? **435**
Data Exploration **435**
Segmentation and Clustering **436**
Distance Between Records **437**
Segmentation **437**
Directed Segmentation **438**
Target Variable Definition, When the Target Is Not Explicit **438**
Simulation, Forecasting, and Agent-Based Modeling **443**
Monte Carlo Simulation **443**
Customer-Centric Forecasting **449**
Agent-Based Modeling **454**
Methodology for Undirected Data Mining **455**
There Is No Methodology **456**
Things to Keep in Mind **456**
Lessons Learned **457**

Chapter 13 Finding Islands of Similarity: Automatic Cluster Detection **459**
Searching for Islands of Simplicity **461**
Customer Segmentation and Clustering **461**
Similarity Clusters **463**
Tracking Campaigns by Cluster-Based Segments **464**
Clustering Reveals an Overlooked Market Segment **466**
Fitting the Troops **467**
The K-Means Clustering Algorithm **468**
Two Steps of the K-Means Algorithm 468
Voronoi Diagrams and K-Means Clusters 471
Choosing the Cluster Seeds 473

Choosing K i 473
Using K-Means to Detect Outliers 474
Semi-Directed Clustering 475
Interpreting Clusters 475
Characterizing Clusters by Their Centroids 476
Characterizing Clusters by What Differentiates Them 477
Using Decision Trees to Describe Clusters 478
Evaluating Clusters 479
Cluster Measurements and Terminology 480
Cluster Silhouettes 480
Limiting Cluster Diameter for Scoring 483
Case Study: Clustering Towns 484
Creating Town Signatures 484
Creating Clusters 486
Determining the Right Number of Clusters 486
Evaluating the Clusters 487
Using Demographic Clusters to Adjust Zone Boundaries 488
Business Success 490
Variations on K-Means 490
K-Medians, K-Medoids, and K-Modes 490
K-Medians 490
K-Medoids 493
The Soft Side of K-Means 494
Data Preparation for Clustering 495
Scaling for Consistency 496
Use Weights to Encode Outside Information 496
Selecting Variables for Clustering 497
Lessons Learned 497
Chapter 14 Alternative Approaches to Cluster Detection 499
Shortcomings of K-Means 500
Reasonableness 500
An Intuitive Example 501
Fixing the Problem by Changing the Scales 503
What This Means in Practice 504
Gaussian Mixture Models 505
Adding "Gaussians" to K-Means 505
Multi-Dimensional Gaussians 505
Applying Gaussians to a K-Means Centroid 506
Using Gaussians for K-Means Soft Clustering 507
Back to Gaussian Mixture Models 508
Different Shapes 508
The GMM Approach 509
Expectation Maximization 509
Scoring GMMs 510
Applying GMMs 511
Divisive Clustering 513
A Decision Tree-Like Method for Clustering 513

When All Fields Are Numeric 513
Divisive Clustering When All Fields Are Categorical 514
A General Approach 514
Scoring Divisive Clusters 515
Clusters and Trees 515
Agglomerative (Hierarchical) Clustering 516
Overview of Agglomerative Clustering Methods 516
Clustering People by Age: An Example of Agglomerative Clustering 516
Visualizing the Clusters 518
Combining Nearby Clusters 519
An Agglomerative Clustering Algorithm 520
Scoring Agglomerative Clusters 522
Limitations of Agglomerative Clustering 523
Computationally Expensive 524
Difficult to Visualize Clusters 524
Sensitivity to Outliers 524
Agglomerative Clustering in Practice 525
Clustering Products by Customer Preferences 525
Clustering Direct Marketing Campaigns by Customer Response 526
Combining Agglomerative Clustering and K-Means 526
Self-Organizing Maps 527
What Is a Self-Organizing Map? 527
Training an SOM 530
Scoring an SOM 531
The Search Continues for Islands of Simplicity 532
Lessons Learned 533

Chapter 15 Market Basket Analysis and Association Rules 535
Defining Market Basket Analysis 536
Four Levels of Market Basket Data 537
The Foundation of Market Basket Analysis: Basic Measures 539
Order Characteristics 540
Item (Product) Popularity 541
Tracking Marketing Interventions 542
Case Study: Spanish or English 543
The Business Problem 543
The Data 544
Defining "Hispanicity" Preference 545
The Solution 546
Association Analysis 547
Rules Are Not Always Useful 548
Actionable Rules 548
Trivial Rules 550
Inexplicable Rules 550

Item Sets to Association Rules 551
How Good Is an Association Rule? 553
Support 553
Confidence 553
Lift 554
Chi-Square Value 554
Building Association Rules 555
Choosing the Right Set of Items 556
Product Hierarchies Help to Generalize Items 557
Virtual Items Go Beyond the Product Hierarchy 559
Data Quality 560
Anonymous Versus Identified 561
Generating Rules from All This Data 561
Calculating Support 562
Calculating Confidence 562
Calculating Lift 563
The Negative Rule 564
Calculating Chi-Square 564
Overcoming Practical Limits 565
The Problem of Big Data 567
Extending the Ideas 569
Different Items on the Right- and Left-Hand Sides 569
Using Association Rules to Compare Stores 570
Association Rules and Cross-Selling 572
A Typical Cross-Sell Model 572
A More Confident Approach to Product Propensities 573
Results from Using Confidence 574
Sequential Pattern Analysis 574
Finding the Sequences 575
Sequential Patterns with Just One Item 575
Sequential Patterns to Visualize Switching Behavior 577
Working with Sequential Patterns 578
Sequential Association Rules 578
Sequential Analysis Using Other Data Mining Techniques 579
Lessons Learned 579

Chapter 16 Link Analysis 581
Basic Graph Theory 582
What Is a Graph? 582
Directed Graphs 584
Weighted Graphs 585
Seven Bridges of Konigsberg 585
Detecting Cycles in a Graph 588
The Traveling Salesman Problem Revisited 589
The Traveling Salesman Problem Is Difficult 590
The Exhaustive Solution 590
The Greedy Algorithm 592

Social Network Analysis 593
Six Degrees of Separation 593
What Your Friends Say About You 595
Finding Childcare Benefits Fraud **596**
Who Responds to Whom on Dating Sites **597**
Social Marketing **598**
Mining Call Graphs **598**
Case Study: Tracking Down the Leader of the Pack **601**
The Business Goal **601**
The Data Processing Challenge **601**
Finding Social Networks in Call Data **602**
How the Results Are Used for Marketing **602**
Estimating Customer Age **603**
Case Study: Who Is Using Fax Machines from Home? **604**
Why Finding Fax Machines Is Useful **604**
How Do Fax Machines Behave? **604**
A Graph Coloring Algorithm **605**
"Coloring" the Graph to Identify Fax Machines **606**
— How Google Came to Rule the World **607**
Hubs and Authorities **608**
The Details **609**
Creating the Root Set **609**
Identifying the Candidates **610**.
Ranking Hubs and Authorities **610**
Hubs and Authorities in Practice 611
Lessons Learned 612

Chapter 17 Data Warehousing, OLAP, Analytic Sandboxes, and Data Mining 613
The Architecture of Data 615
Transaction Data, the Base Level 616
Operational Summary Data 617
Decision-Support Summary Data 617
Database Schema/Data Models 618
Metadata 623
Business Rules 623
A General Architecture for Data Warehousing 624
Source Systems 624
Extraction, Transformation, and Load 626
Central Repository 627
Metadata Repository 630
Data Marts 630
Operational Feedback 631
Users and Desktop Tools 631
Analysts 631

Application Developers 632
Business Users 633
Analytic Sandboxes 633
Why Are Analytic Sandboxes Needed? 634
Too Much Data 634
More Advanced Techniques 635
Complex Data 635
Technology to Support Analytic Sandboxes 636
Faster Databases 636
In Database Analytics 637
Statistical Tools 638
Hadoop/MapReduce 638
Where Does OLAP Fit In? 639
What's in a Cube? 641
Facts 643
Dimensions and Their Hierarchies 644
Conformed Dimensions 646
Star Schema 646
OLAP and Data Mining 648
Where Data Mining Fits in with Data Warehousing 650
Lots of Data 651
Consistent, Clean Data 651
Hypothesis Testing and Measurement 652
Scalable Hardware and RDBMS Support 653
Lessons Learned 653

Chapter 18 Building Customer Signatures 655

Finding Customers in Data 656
What Is a Customer? 657
Accounts? Customers? Households? 658
Anonymous Transactions 658
Transactions Linked to a Card 659
Transactions Linked to a Cookie 659
Transactions Linked to an Account 660
Transactions Linked to a Customer 661
Designing Signatures 661
Is a Customer Signature Necessary? 666
What Does a Row Represent? 666
Householding 667
Households Change Over Time 671
Will the Signature Be Used for Predictive Modeling? 671
Has a Target Been Defined? 672
Are There Constraints Imposed by the Particular Data Mining Techniques to be Employed? 672
Which Customers Will Be Included? 673
What Might Be Interesting to Know About Customers? 673

What a Signature Looks Like
Process for Creating Signatures 677
Some Data Is Already at the Right Level of Granularity 678
Pivoting a Regular Time Series *ffl9*
Aggregating Time-Stamped Transactions 680
Creating a Regular Time Series 681
When Transactions Are Irregular and Rare 681
Creating an Overall Transaction Summary 681
Dealing with Missing Values 685
Missing Values in Source Data 685
Unknown or Non-Existent? 687
What Not to Do 68?
Don't Throw Records Away 687
Don't Replace with a "Special" Numeric Value
Don't Replace with Average, Median, or Mode
Things to Consider
Consider Doing Nothing
Consider Multiple Models .690
Consider Imputation 6§0
Imputed Values Should Never Be Surprising 691
Lessons Learned 691

Chapter 19 Derived Variables: Making the Data Mean More 693
Handset Churn Rate as a Predictor of Churn 694
Single-Variable Transformations 696
Standardizing Numeric Variables 696
Centering #6
Rescaling 697
Turning Numeric Values into Percentiles 697
Turning Counts into Rates 6H8
Relative Measures 699
Replacing Categorical Variables with Numeric Ones *70Q*
What Not to Do 701
Using Indicator Variables 701
Replacing Categorical Variables with Numeric Descriptors 702
Binning Numeric Variables 702
Combining Variables *Tffi*
Classic Combinations 707
Body Mass Index (BMI) 707
On-Base Percentage and Slugging Percentage 708
Wind Chill Index 709
Combining Highly Correlated Variables Tjl0
Differences between Nearly Synonymous Variables Tjll
Ratios of Highly Correlated Variables 7^2
Example: Deriving a Variable for the Relationship of Rent to Home Value 712

The Degree of Correlation] 716
Extracting Features from Time Series 718
Trend - 719
Seasonality 721
Extracting Features from Geography 722
Geocoding 722
Mapping 723
Using Geography to Create Relative Measures 724
Using Past Values of the Target Variable 725
Using Model Scores as Inputs 725
Handling Sparse Data 726
Account Set Patterns . 726
Binning Sparse Values 727
Capturing Customer Behavior from Transactions 727
Widening Narrow Data 728
Sphere of Influence as a Predictor of Good Customers 728
An Example: Ratings to Rater Profile 730
Sample Fields from the Rater Signature 730
The Rating Signature and Derived Variables 732
Lessons Learned 733

Chapter 20 Too Much of a Good Thing? Techniques for Reducing the Number of Variables 735
Problems with Too Many Variables - 736
Risk of Correlation Among Input Variables 736
Risk of Overfitting 738
The Sparse Data Problem 738
Visualizing Sparseness 739
Independence 740
Exhaustive Feature Selection . 743
Flavors of Variable Reduction Techniques 744
Using the Target 744
Original versus New Variables 744
Sequential Selection of Features ' 745
The Traditional Forward Selection Methodology 745
Forward Selection Using a Validation Set 747
Stepwise Selection 748
Forward Selection Using Non-Regression Techniques 748
Backward Selection 748
Undirected Forward Selection 749
Other Directed Variable Selection Methods 749
Using Decision Trees to Select Variables 750
Why Is This Different From Forward Selection? 750
Categorical versus Numeric Inputs 752
Variable Reduction Using Neural Networks 752
, Principal Components 753

What Are Principal Components? 753
Geometric Definition of the First Principal Component 753
Properties of the First Principal Component 755
From the First Principal Component to the Next Principal Component 756
Properties of Principal Components 756
Comment About Categorical Input Variables 758
Principal Components Example
Principal Component Analysis
Reducing the Number of Variables •
Understanding Input Variables •
Data Visualization
Factor Analysis
Variable Clustering
Example of Variable Clusters .
Using Variable Clusters ; **770**
Hierarchical Variable Clustering **770**
Correlation Variable Clustering **770**
Example of Correlation Variable Clustering **771**
Variations on Hierarchical Method **773**
Divisive Variable Clustering ***77%***
Lessons Learned ***Tflk***

Chapter 21 Listen Carefully to What Your Customers Say: Text Mining 775
What Is Text Mining? ***77b***
Text Mining for Derived Columns ***7%***
. Beyond Derived Features ***777***
Text Analysis Applications ***77%***
Spell Checking and Grammar Checking **778**
Translation from One Human Language to Another **778**
Search **77f**
Summarizing Documents **780**
Turing Test and Natural Language Processing **781**
Working with Text Data **781**
Sources of Text **781**
Language Effects **781**
Basic Approaches to Representing Documents **783**
Bag of Words **783**
Natural Language Processing **784**
Representing Documents in Practice **784**
Stop Words • ***T***
Stemming ***m***
Word Pairs and Phrases 785
Using a Lexicon 786
Documents and the Corpus 786

Case Study: Ad Hoc Text Mining | 786
The Boycott 787
Business as Usual 787
Combining Text Mining and Hypothesis Testing 787
The Results 788
Classifying News Stories Using MBR 789
What Are the Codes? 789
Applying MBR 790
Choosing the Training Set 791
Choosing the Distance Function 791
Choosing the Combination Function 791
Choosing the Number of Neighbors 793
The Results 793
From Text to Numbers 794
Starting with a "Bag of Words" 794
Words with Multiple Meanings 795
Words with Many Forms 795
Ignoring Grammar 796
Term-Document Matrix 796
Corpus Effects 797
Singular Value Decomposition (SVD) 798
What Does SVD Have to Do with Text Mining? 799
Latent Semantic Indexing 800
Applying SVD to Text Mining 800
Text Mining and Naive Bayesian Models 800
Naive Bayesian in the Text World 801
Identifying Spam Using Naive Bayesian 801
The Complexities of Spam Detection 802
Content-Only Spam Analysis 802
Finding Important Terms 803
Building a Naive Bayesian Model 803
The Results 803
Sentiment Analysis 806
What Sentiment Analysis Does 807
Basic Approach 807
Using a "Psycho-Social" Dictionary 808
DIRECTV: A Case Study in Customer Service 809
Background 809
The Call Center Interface 810
What Happened Over Time 811
Applying Text Mining 811
Determining the Usefulness of the RV Segment 811
Acting on the Results 812
Continuing Clustering 813

Taking the Technical Approach 814
Parsing the Comments 814
Fixing Misspellings 815
Stemming 815
Applying Synonym Lists 816
Using a Stop List 81F
Converting Text to Numbers 81?
Clustering 818
Not an Iterative Process 818
Continuing to Benefit W8
Lessons Learned 819

Index **821**