

Data Mining Applications

Analyzing Business Data

Version 1.0.0

Gerald V. Post

University of the Pacific

- Business Applications of Data Mining, v
- Learning Assessment, v
- Organization, vi
- Features That Focus on Solving Problems, vii
- This Book Is Different from Other Texts, viii
- Instructional Support, viii
- Installing the Sample Databases, viii
 - Software*, ix
 - Database Installation* , x

1. Introduction, 1

- Introduction, 2
- Business Situation, 3
 - Finance, Risk, and Costs*, 3
 - Marketing*, 5
 - Production and Supply Chain Management*, 9
 - Human Resources Management*, 11
- Perspectives, 13
 - Probability and Statistics*, 14
 - Machine Learning*, 14
 - Computer Science: Challenges of Large Data Sets*, 14
 - Management Applications*, 15
- Database, 15
 - Traditional Transactions Processing*, 16
 - Data Warehouse and Analytical Processing*, 17
 - Data Sources*, 17
 - Data Extraction, Transformation, and Loading*, 17
- Software Tools, 18
 - Data Mining Techniques*, 19
 - Data Mining Tools*, 19
 - Statistical Tools*, 20
 - Production Systems and Scaling*, 21
- Potential Dangers, 21
 - Human Errors* , 23
 - Insufficient Data*, 23
 - Bad Data* , 24
 - Over Fitting*, 24
 - Random Chance*, 26
 - Estimation Instability*, 26
 - Model Instability*, 27
- Introduction to Cases, 28
 - Rolling Thunder Bicycle Company*, 29

- Diner*, 29
- Corner Med*, 29
- Basketball*, 30
- Bakery*, 30
- Cars*, 31
- Summary, 31
- Key Words, 32
- Review Questions, 32
- Exercises, 32
- Additional Reading, 35

I. Fundamental Tools, 37

2. Database Management Systems, 38

- Introduction, 39
- Relational Databases, 40
 - Tables*, 40
 - Data Types*, 42
- Four Questions to Retrieve Data, 44
 - What Output Do You Want to See?*, 45
 - What Do You Already Know?*, 45
 - What Tables Are Involved?*, 45
 - How Are the Tables Joined?*, 45
- Query Basics, 46
 - Single Tables*, 47
 - Introduction to SQL*, 49
 - Sorting the Output*, 49
 - Criteria*, 50
 - Useful WHERE Clauses*, 53
- Computations, 54
 - Basic Arithmetic Operators*, 54
 - Aggregation*, 55
 - Functions*, 57
- Subtotals and GROUP BY, 58
 - Conditions on Totals (HAVING)*, 60
 - WHERE versus HAVING*, 61
 - The Best and the Worst*, 62
- Multiple Tables, 63
 - Joining Tables*, 63
 - Identifying Columns in Different Tables*, 64
 - Joining Many Tables*, 65
 - Views: Saved Queries*, 66
 - LEFT JOIN*, 67
 - UNION*, 69
- Data Manipulation, 69

- UPDATE, 70
- INSERT, 71
- DELETE, 71
- SQL Server Reports, 71
 - Administration Configuration, 72
 - Creating a Report, 73
- Database Design Concepts, 78
 - Notation, 80
 - First Normal Form, 81
 - Second Normal Form, 83
 - Third Normal Form, 84
- Summary, 85
- Key Words, 86
- Review Questions, 86
- Exercises, 87
- Additional Reading, 89
- 3. OLAP Cubes, 90**
- Introduction, 91
- Challenges with the Relational Model, 93
 - Indexes, 94
 - Data Warehouse, 94
 - Extraction, Transformation, and Loading, 95
 - MOLAP, ROLAP, and HOLAP, 96
- OLAP Design, 97
 - Facts and Dimensions, 99
 - Star Design, 100
 - Snowflake Design, 100
 - Hierarchies, 102
- Creating a Cube with Microsoft Analysis Services, 103
 - Data Sources, 104
 - Data Source Views, 106
 - Cubes, 111
- Dimensions, 115
 - Hierarchies, 117
 - Time Dimensions, 118
 - Custom Geographic Hierarchy, 122
 - Attribute Relationships, 124
- Fine Tuning the Cube, 128
 - Calculations and Queries, 129
 - Perspectives, 132
 - Internationalization and Translations, 133
 - Performance: Partitions and Aggregations, 137
- Excel PivotTables, 138

- Actions, 140
- Key Performance Indicators, 143
 - Definition, 143
 - Creating KPIs, 144
 - Browsing a KPI, 147
- Summary, 148
- Key Words, 149
- Review Questions, 150
- Exercises, 150
- Additional Reading, 153
- 4. Probability and Statistics Summary, 154**
- Introduction, 155
- Probability Basics, 155
 - Discrete and Continuous Data, 156
 - Counting and Combinations, 157
 - Probability Rules, 159
- Interdependencies: Joint Probabilities, 162
 - Contingency Tables, 162
 - Tree Diagrams, 165
 - Bayes Theorem, 166
- Probability Distributions, 170
 - Discrete Data, 170
 - Continuous Data, 174
 - Joint and Conditional Probabilities, 175
 - Expected Value (Mean) and Variance, 177
 - Important Continuous Distributions, 181
- Statistics, 189
 - Samples, 189
 - Common Statistics, 190
 - Confidence Intervals, 191
 - Hypothesis Testing, 194
 - Chi-Square Hypothesis Tests, 196
 - Information Measure, 198
- Summary, 200
- Key Words, 201
- Review Questions, 201
- Exercises, 202
- Additional Reading, 207
- II. Business Analysis, 208**
- 5. Cluster Analysis, 209**
- Introduction, 210
- Business Situation, 211

- Model, 212
 - Distance or Dissimilarities*, 213
 - Combinatorial Searches with K-Means*, 216
 - Statistical Mixture Model with EM*, 217
 - Hierarchical Clusters*, 221
 - Other Statistical Methods*, 224
- Data, 227
 - Attributes and Observations*, 227
 - Continuous and Discrete Data*, 228
 - Missing Data*, 229
- Clustering on Products: Cars, 229
 - Goals*, 229
 - Data*, 230
 - Microsoft Clustering*, 232
 - Results from Microsoft Clustering*, 233
 - Prediction*, 237
 - Larger Model and Parameter Changes*, 238
- Traditional EM Clustering, 241
 - Goals and Data*, 242
 - Results*, 242
 - K-Means Clusters*, 244
- Comparison, 245
- Customer Clustering with Categorical Data, 247
 - Data*, 247
 - Microsoft Clustering Results*, 248
 - Weka Clustering Results*, 249
- Summary, 251
- Key Words, 252
- Review Questions, 252
- Exercises, 252
- Additional Reading, 255
- 6. Association and Market Baskets, 256**
 - Introduction, 257
 - Business Situation, 257
 - The Bakery*, 258
 - Product and Dimension Levels*, 260
 - Model, 261
 - Goal*, 261
 - Assigning Values to Rules*, 262
 - Problems with Dimensions, 268
 - The A Priori Algorithm*, 269
 - Issues in Setting Minimum Support and Confidence*, 270
 - Potential Problems, 272
 - Simpson's Paradox*, 272
 - Skewed Support Data*, 273
 - Continuous Data*, 274
 - Quantity*, 276
- Data, 276
 - Database Structure*, 276
 - Market Basket Structure*, 277
- Traditional Tools for Association Rules, 279
 - Goals*, 279
 - Data*, 279
 - Results*, 280
- Microsoft Association Rules, 281
 - Goals*, 281
 - Data*, 282
 - Results*, 283
 - Comparing Results*, 285
- Summary, 286
- Key Words, 287
- Review Questions, 287
- Exercises, 288
- Additional Reading, 291
- 7. Evaluation of Dimensions, 293**
 - Introduction, 294
 - Business Situation, 295
 - Model, 295
 - Data, 297
 - Attributes and Observations*, 298
 - Continuous and Discrete Data*, 298
 - Missing Data*, 299
 - Linear Regression, 299
 - Goals*, 300
 - Data*, 302
 - Tools*, 304
 - Results*, 308
 - Attribute Evaluation*, 313
 - Prediction*, 314
 - Logistic Regression, 316
 - Goals*, 316
 - Data*, 318
 - Tools*, 319
 - Results*, 320
 - Attribute Evaluation*, 324
 - Prediction*, 324
 - Naïve Bayes, 326
 - Goals*, 326
 - Data*, 330

- Tools, 331*
- Results, 332*
- Attribute Evaluation, 332*
- Prediction, 334*
- Decision Trees, 335
 - Goals, 336*
 - Data, 338*
 - Tools, 338*
 - Results, 339*
 - Attribute Evaluation, 341*
 - Prediction, 341*
- Neural Network, 344
 - Goals, 345*
 - Data, 347*
 - Tools, 347*
 - Results, 347*
 - Attribute Evaluation, 348*
 - Prediction, 349*
- Model Comparisons, 351
 - Prediction, 352*
 - Attribute Evaluation, 353*
 - Nonlinear Complications, 354*
- Summary, 355
- Key Words, 356
- Review Questions, 356
- Exercises, 357
- Additional Reading, 359
- 8. Time Series Analysis, 361**
- Introduction, 362
- Business Situation, 363
- Model, 364
 - Time Series Components, 364*
 - Auto Regression, 366*
 - Moving Average, 369*
 - Trends, 372*
 - ARIMA, 374*
 - Cross Correlations, 378*
 - Evaluating Models, 380*
- Data, 381
 - Attributes and Observations, 381*
 - Missing Data, 382*
- Traditional ARIMA Estimation, 382
 - Goals, 382*
 - Tools, 383*
 - Results, 385*
- Forecasts, 387*
- Seasonality Evaluation, 389*
- Microsoft Time Series Estimation, 390
 - Goals, 391*
 - Data, 392*
 - Tools, 393*
 - Results, 397*
 - Forecasts, 399*
 - Seasonality Evaluation, 401*
- Cross Correlation and Linear Regression, 402
 - Goals, 402*
 - Data, 402*
 - Tools, 403*
- Comparison, 410
- Summary, 410
- Key Words, 411
- Review Questions, 412
- Exercises, 412
- Additional Reading, 415