Data Mining with Microsoft® SQL Server® 2008

Jamie MacLennan
ZhaoHui Tang
Bogdan Crivat

WILEY

Wiley Publishing, Inc.

About the Authors

Jamie MacLennan is the principal development manager of SQL Server Analysis Services at Microsoft. In addition to being responsible for the development and delivery of the Data Mining and OLAP technologies for SQL Server, MacLennan is a proud husband and father of four. He has more than 25 patents and patents pending for his work on SQL Server Data Mining. MacLennan has written extensively on the data mining technology in SQL Server, including many articles in *MSDN Magazine*, *SQL Server Magazine*, and postings on `SQLServerDataMining.com` and his blog at `http://blogs.msdn.com/jamiemac`. This is his second edition of *Data Mining with SQL Server*. MacLennan has been a featured and invited speaker at conferences worldwide, including Microsoft TechEd, Microsoft TechEd Europe, SQL PASS, the Knowledge Discovery and Data Mining (KDD) conference, the Americas Conference on Information Systems (AMCIS), and the Data Mining Cup conference.

ZhaoHui Tang is a group program manager at Microsoft adCenter Labs, where he manages a number of research projects related to paid search and content ads. He is the inventor of Microsoft Keyword Services Platform. Prior to adCenter, he spent six years as a lead program manager in the SQL Server Business Intelligence (BI) group, mainly focusing on data mining development. He has written numerous articles for both academic and industrial publications, such as *The VLDB Journal* and *SQL Server Magazine*. He is a frequent speaker at business intelligence conferences. He was also a co-author of the previous edition of this book, *Data Mining with SQL Server 2005*.

Bogdan Crivat is a senior software design engineer in SQL Server Analysis Services at Microsoft, working primarily on the Data Mining platform.

Crivat has written various articles on data mining for *MSDN Magazine* and *Access/VB/SQL Advisor Magazine*, as well as numerous postings on the `SQLServerDataMining.com` website and on the MSDN Forums. He presented at various Microsoft and data mining professional conferences. Crivat also blogs about SQL Server Data Mining at `www.bogdancrivat.net/dm`.

Acknowledgments

First of all we would like to acknowledge the help from our data mining team members and other colleagues in the Microsoft SQL Server Business Intelligence (BI) organization. In addition to creating the best data mining package on the planet, most of them gave up some of their free time to review the text and sample code. Direct thanks go to Shuvro Mitra, Raman Iyer, Dana Cristofor, Jeanine Nelson-Takaki, and Niketan Pansare for helping review our text to ensure that it makes sense and that our samples work. Thanks also to the rest of the data mining team, including Donald Farmer, Tatyana Yakushev, Yimin Wu, Fernando Godinez Delgado, Gang Xiao, Liu Tang, and Bo Simmons for building such a great product. In addition, we would like to thank the SQL BI management of Kamal Hathi and Tom Casey for supporting data mining in SQL Server.

SQL Server 2008 Data Mining (including the Data Mining Add-Ins) is a product jointly developed by the SQL Server Analysis Services team and other teams inside Microsoft. We would like to thank colleagues from Excel —notably Rob Collie, Howie Dickerman, and Dan Battagin, whose valuable input into the design of the Data Mining Add-Ins guaranteed their success. Also thanks to those in the Machine Learning and Applied Statistics (MLAS) Group, headed by Research Manager David Heckerman, who continue to advise us on deep algorithmic issues in our product. We would like to thank David Heckerman, Jesper Lind, Alexei Bocharov, Chris Meek, Bo Thiesson, and Max Chickering for their contributions.

We would like to give special thanks to Kevin Shafer for his close editing of our text, which has greatly improved the quality of this manuscript. Also thanks to Wiley Publications acquisitions editor Bob Elliot for his support and patience.

Special thanks from Jamie to his wife, April, who yet again supported him through the ups and downs of authoring a book, particularly during painful rewrites and recaptures of screen shots, while taking care of our kids and the world around me. Elalu, honey.

Bogdan would like to thank his wife, Irinel, for supporting him, reviewing his chapters, and some really helpful hints for capturing screen shots.

Contents at a Glance

Foreword xxix

Introduction xxxi

Chapter 1 Introduction to Data Mining in SQL Server 2008 1

Chapter 2 Applied Data Mining Using Microsoft Excel 2007 15

Chapter 3 Data Mining Concepts and DMX 83

Chapter 4 Using SQL Server Data Mining 127

Chapter 5 Implementing a Data Mining Process Using Office 2007 187

Chapter 6 Microsoft Naïve Bayes 215

Chapter 7 Microsoft Decision Trees Algorithm 235

Chapter 8 Microsoft Time Series Algorithm 263

Chapter 9 Microsoft Clustering 291

Chapter 10 Microsoft Sequence Clustering 319

Chapter 11 Microsoft Association Rules 343

Chapter 12 Microsoft Neural Network and Logistic Regression 371

Chapter 13 Mining OLAP Cubes 399

Chapter 14 Data Mining with SQL Server Integration Services 439

Chapter 15 SQL Server Data Mining Architecture 475

Chapter 16 Programming SQL Server Data Mining 497

Chapter 17 Extending SQL Server Data Mining 541

Chapter 18 Implementing a Web Cross-Selling Application 563
Chapter 19 Conclusion and Additional Resources 581
Appendix A Data Sets 589
Appendix B Supported Functions 595
Index 607

Contents

Foreword **xxix**

Introduction **xxxi**

Chapter 1 Introduction to Data Mining in SQL Server 2008 1

Business Problems for Data Mining 4

Data Mining Tasks 6

Classification 6

Clustering 6

Association 7

Regression 8

Forecasting 8

Sequence Analysis 9

Deviation Analysis 9

Data Mining Project Cycle 9

Business Problem Formation 10

Data Collection 10

Data Cleaning and Transformation 10

Model Building 12

Model Assessment 12

Reporting and Prediction 12

Application Integration 13

Model Management 13

Summary 13

Chapter 2 Applied Data Mining Using Microsoft Excel 2007 15

Setting Up the Table Analysis Tools 16

Configuring Analysis Services with Administrative Privileges 17

Configuring Analysis Services without Administrative Privileges 18
What the Add-Ins Expect 19
What to Do If You Need Help 22
The Analyze Key Influencers Tool 22
The Main Influencers Report 24
The Discrimination Report 26
Summary of the Analyze Key Influencers Task 28
The Detect Categories Tool 28
Launching the Tool 29
The Categories Report 30
Categories and the Number of Rows in Each 30
Characteristics of Each Category 31
The Category Profiles Chart 32
Summary of the Detect Categories Tool 34
The Fill From Example Tool 35
Running the Tool and Interpreting the Results 36
Refining the Results 38
Summary of the Fill From Example Tool 39
The Forecasting Tool 39
Launching the Tool and Specifying Options 40
Interpreting the Results 42
Summary of the Forecast Tool 44
The Highlight Exceptions Tool 44
Using the Tool 45
More Complex Interactions 48
Limitations and Troubleshooting 50
Summary of the Highlight Exceptions Tool 51
The Scenario Analysis Tool 51
The Goal Seek Tool 53
Using Goal Seek for a Numeric Goal 56
Using Goal Seek for the Whole Table 57
The What-If Tool 58
Using What-If for the Whole Table 61
Summary of the Scenario Analysis Tool 62
The Prediction Calculator Tool 62
Running the Tool 64
The Prediction Calculator Spreadsheet 65
The Printable Calculator Spreadsheet 67
Refining the Results 68
Using the Results 73
Summary of the Prediction Calculator Tool 73

The Shopping Basket Analysis Tool 74

Using the Tool 75

The Bundled Item Report 76

The Recommendations Report 77

Tweaking the Tool 79

Summary of the Shopping Basket Analysis Tool 81

Technical Overview of the Table Analysis Tools 81

Summary 82

Chapter 3 Data Mining Concepts and DMX 83

History of DMX 83

Why DMX? 84

The Data Mining Process 85

Key Concepts 86

Attribute 86

State 87

Case 88

Keys 89

Inputs and Outputs 91

DMX Objects 93

Mining Structure 93

Mining Model 94

DMX Query Syntax 95

Creating Mining Structures 96

Discretized Columns 97

Nested Tables 98

Partitioning into Testing and Training Sets 99

Creating Mining Models 100

Nested Tables 101

Complex Nesting Scenarios 104

Filters 107

Populating Mining Structures 108

Populating Nested Tables 110

Querying Structure Data 112

Querying Model Data 112

Prediction 115

Prediction Join 116

Prediction Query Syntax 116

Nested Source Data 117

Real-Time Prediction 118

Degenerate Predictions 119

Prediction Functions 120
PredictNodeID 122
External and User-Defined Functions 123
Predictions on Nested Tables 123
Predicting Nested Value Columns 124
Summary 125

Chapter 4 Using SQL Server Data Mining 127
Introducing the Business Intelligence Development Studio 128
Understanding the User Interface 128
Offline Mode and Immediate Mode 130
Immediate Mode 131
Getting Started in Immediate Mode 131
Offline Mode 132
Getting Started in Offline Mode 133
Switching Project Modes 135
Creating Data Mining Objects 135
Setting Up Your Data Sources 135
Understanding Data Sources 136
Creating the MovieClick Data Source 137
Using the Data Source View 137
Creating the MovieClick Data Source View 138
Working with Named Calculations 140
Creating a Named Calculation on the Customers Table 142
Working with Named Queries 142
Creating a Named Query Based on the Customers Table 143
Organizing the DSV 144
Exploring Data 145
Creating and Editing Models 148
Structures and Models 148
Using the Data Mining Wizard 148
Creating the MovieClick Mining Structure and Model 155
Using Data Mining Designer 157
Working with the Mining Structure Editor 157
Adding the Genre Column to the Movies Nested Table 159
Working with the Mining Models Editor 160
Creating and Modifying Additional Models 163
Processing 164
Processing the MovieClick Mining Structure 165
Using Your Models 166
Understanding the Model Viewers 166
Using the Mining Accuracy Chart 167
Selecting Test Data 168

Understanding the Accuracy Charts 169
Using the Profit Chart 171
Multiple Target Accuracy Charts 172
Using the Classification Matrix 173
Scatter Accuracy Charts 173
Creating a Lift Chart on MovieClick 174
Using CrossValidation 174
Using the Mining Model Prediction Builder 178
Executing a Query on the MovieClick Model 179
Creating Data Mining Reports 180
Using SQL Server Management Studio 181
Understanding the Management Studio User Interface 182
Using Server Explorer 182
Using Object Explorer 183
Using the Query Editor 184
Summary 185

Chapter 5 Implementing a Data Mining Process Using Office 2007 187
Introducing the Data Mining Client 188
Importing Data Using the Data Mining Client 189
Data Exploration and Preparation 190
Discretizing Data with the Explore Data Tool 191
Chopping Off the Long Tail 191
Consolidating Meaning 192
Eliminating Spurious Values 194
Rebalancing Data 195
Modeling 196
Task-Based Modeling 196
Introduction 198
Select Data 198
Select Columns and Options 198
Split Data 200
Finishing the Task 200
Advanced Modeling in the Data Mining Client 200
Accuracy and Validation 203
Model Usage 204
Browsing Models 204
Viewing Models with Visio 205
Querying Models 208
Query Wizard 208
Data Mining Cell Functions 211
DMPREDICT 211
DMPREDICTTABLEROW 212

DMCONTENTQUERY 212
Model Management 213
Trace 213
Summary 213

Chapter 6 Microsoft Naïve Bayes 215
Introducing the Naïve Bayes Algorithm 216
Using the Naïve Bayes Algorithm 216
Creating a Predictive Model 217
Data Exploration 219
Analysis of Key Influencers 219
Document Classification 220
DMX 222
Drill-through 222
Understanding Naïve Bayes Content 223
Exploring a Naïve Bayes Model 225
Dependency Network 225
Attribute Profiles 226
Attribute Characteristics 227
Attribute Discrimination 228
Understanding Naïve Bayes Principles 229
Limitations of the Naïve Bayes Algorithm 231
Naïve Bayes Parameters 233
MAXIMUM_INPUT_ATTRIBUTES 233
MAXIMUM_OUTPUT_ATTRIBUTES 233
MAXIMUM_STATES 233
MINIMUM_DEPENDENCY_PROBABILITY 234
Summary 234

Chapter 7 Microsoft Decision Trees Algorithm 235
Introducing Decision Trees 236
Using Decision Trees 237
Creating a Decision Tree Model 237
DMX Queries 237
Classification Model 237
Regression Model 239
Association 241
Model Content 244
Interpreting the Model 244
Decision Tree Principles 248
Basic Concepts of Tree Growth 248
Working with Many States in an Attribute 251
Avoiding Overtraining 252
Incorporating Prior Knowledge 252
Feature Selection 253

Using Continuous Inputs 253
Regression 254
Association Analysis with Microsoft Decision Trees 255
Parameters 256
COMPLEXITY_PENALTY 257
MINIMUM_SUPPORT 257
SCORE_METHOD 257
SPLIT_METHOD 258
MAXIMUM_INPUT_ATTRIBUTES 258
MAXIMUM_OUTPUT_ATTRIBUTES 258
FORCE_REGRESSOR 258
Stored Procedures 259
Summary 260

Chapter 8 Microsoft Time Series Algorithm 263
Overview 264
Usage 265
Time Series Scenarios 267
Performing a Simple Forecast 267
Predicting Interdependent Series 268
Understanding Your Time Series 268
What-If Scenarios 269
Predicting New Series 269
DMX 270
Model Creation 270
Model Processing 272
Forecasting 274
Returning Supplemental Statistics 275
Changing the Future — Executing a What-If Forecast 276
Forecasting with Little Data — Applying Models to New Data 277
Drill-Through 280
Principles of Time Series 280
Autoregression 281
Periodicity 281
Autoregression Trees 282
Prediction 284
Parameters 285
MISSING_VALUE_SUBSTITUTION 285
PERIODICITY_HINT 286
AUTO_DETECT_PERIODICITY 286
MINIMUM and MAXIMUM_SERIES_VALUE 286
FORECAST_METHOD 286
PREDICTION_SMOOTHING 287

INSTABILITY_SENSITIVITY 287
HISTORIC_MODEL_COUNT and HISTORIC_MODEL_GAP 287
COMPLEXITY_PENALTY and MINIMUM_SUPPORT 288
Model Content 289
Summary 289

Chapter 9 Microsoft Clustering 291
Overview 292
Usage of Clustering 294
Performing a Clustering 295
Clustering as an Analytical Step 297
Anomaly Detection Using Clustering 297
DMX 299
Model Creation 300
Drill-Through 301
Cluster 301
ClusterProbability 301
PredictHistogram 302
PredictCaseLikelihood 302
Model Content 303
Understanding Your Cluster Models 304
Get a High-Level Overview 305
Pick a Cluster and Determine how It Is Different from the General Population 307
Determine how the Cluster Is Different from Nearby Clusters 308
Verify that Your Assertions Are True 309
Label the Cluster 309
Principles of Clustering 309
Hard Clustering versus Soft Clustering 311
Discrete Clustering 312
Scalable Clustering 313
Clustering Prediction 314
Parameters 314
CLUSTERING_METHOD 314
CLUSTER_COUNT 315
MINIMUM_CLUSTER_CASES 315
MODELLING_CARDINALITY 316
STOPPING_TOLERANCE 316
SAMPLE_SIZE 316
CLUSTER_SEED 317
MAXIMUM_INPUT_ATTRIBUTES 317
MAXIMUM_STATES 318

Summary 318

Chapter 10 Microsoft Sequence Clustering 319

Introducing the Microsoft Sequence Clustering Algorithm 320

Using the Microsoft Sequence Clustering Algorithm 320

Creating a Sequence Clustering Model 321

DMX Queries 322

Executing Cluster Predictions 323

Executing Sequence Predictions 323

Extracting the Probability for the Sequence Predictions 325

Using the Histogram of the Sequence Predictions 326

Detecting Unusual Sequence Patterns 329

Interpreting the Model 329

Cluster Diagram 330

Cluster Profiles 331

Cluster Characteristics 331

Cluster Discrimination 333

State Transitions 333

Microsoft Sequence Clustering Algorithm Principles 334

Understanding a Markov Chain 334

Order of a Markov Chain 335

State Transition Matrix 336

Clustering with a Markov Chain 337

Cluster Decomposition 339

Model Content 339

Algorithm Parameters 340

CLUSTER_COUNT 340

MINIMUM_SUPPORT 340

MAXIMUM_STATES 341

MAXIMUM_SEQUENCE_STATES 341

Summary 341

Chapter 11 Microsoft Association Rules 343

Introducing Microsoft Association Rules 344

Using the Association Rules Algorithm 344

Data Exploration Models 345

A Simple Recommendation Engine 346

Advanced Cross-Sales Analysis 349

DMX 351

Model Content 355

Interpreting the Model 357

Association Algorithm Principles 359

Understanding Basic Association Algorithm Terms and Concepts 359

Itemset 360

Support 360
Probability (Confidence) 361
Importance 361
Finding Frequent Itemsets 363
Generating Association Rules 366
Prediction 367
Algorithm Parameters 368
MINIMUM_SUPPORT 368
MAXIMUM_SUPPORT 368
MINIMUM_PROBABILITY 368
MINIMUM_IMPORTANCE 368
MAXIMUM_ITEMSET_SIZE 369
MINIMUM_ITEMSET_SIZE 369
MAXIMUM_ITEMSET_COUNT 369
OPTIMIZED_PREDICTION_COUNT 369
AUTODETECT_MINIMUM_SUPPORT 369
Summary 370

Chapter 12 Microsoft Neural Network and Logistic Regression 371
Same Principle, Two Algorithms 372
Using the Microsoft Neural Network 373
Text Classification Models 373
Utility Models 378
DMX Queries 378
Model Content 381
Interpreting the Model 382
Principles of the Microsoft Neural Network Algorithm 384
What Is a Neural Network? 385
Combination and Activation 387
Backpropagation, Error Function, and Conjugate Gradient 389
A Simple Example of Processing a Neural Network 390
Normalization and Mapping 392
Topology of the Network 393
Training the Ending Condition 394
Nonlinearly Separable Classes 395
Algorithm Parameters 396
MAXIMUM_INPUT_ATTRIBUTES 396
MAXIMUM_OUTPUT_ATTRIBUTES 396
MAXIMUM_STATES 396
HOLDOUT_PERCENTAGE 397
HOLDOUT_SEED 397
HIDDEN_NODE_RATIO 397
SAMPLE_SIZE 397
Summary 397

Chapter 13 Mining OLAP Cubes 399

Introducing OLAP 400
- Understanding Star and Snowflake Schemas 401
- Understanding Dimension and Hierarchy 402
- Understanding Measures and Measure Groups 404
- Understanding Cube Processing and Storage 404
- Using Proactive Caching 405
- Querying a Cube 406

Performing Calculations 407

Browsing a Cube 408

Understanding Unified Dimension Modeling 408

Understanding the Relationship between OLAP and Data Mining 413
- Mining Aggregated Data 414
- OLAP Pattern Discovery Needs 415
- OLAP Mining versus Relational Mining 415

Building OLAP Mining Models Using Wizards and Editors 417
- Using the Data Mining Wizard 417
  - Building the Customer Segmentation Model 417
  - Creating a Market Basket Model 420
  - Creating a Sales Forecast Model 424
- Using the Data Mining Designer 428

Understanding Data Mining Dimensions 429

Using MDX within DMX Queries 432

Using Analysis Management Objects for the OLAP Mining Model 434

Summary 438

Chapter 14 Data Mining with SQL Server Integration Services 439

An Overview of SSIS 440
- Understanding SSIS Packages 442
- Task Flow 442
  - Standard Tasks in SSIS 442
  - Containers 443
  - Debugging 444
  - Exploring a Control Flow Example 444
- Data Flow 444
  - Transformations 445
  - Viewers 446
  - Exploring a Data Flow Example 447

Working with SSIS in Data Mining 447
- Data Mining Tasks 448
  - Data Mining Query Task 449

Analysis Services Processing Task 452
Analysis Services Execute DDL Task 453
Data Mining Transformations 455
Data Mining Model Training Destination 455
Data Mining Query Transformation 458
Example Data Flows 462
Using Non-Predictive Data Mining Queries in an Integration Services Pipeline 463
Text Mining Transformations 464
Term Extraction Transformation 465
Term Lookup Transformation 467
More Details on the Text Mining Process 470
Summary 472
Chapter 15 SQL Server Data Mining Architecture 475
Introducing Analysis Services Architecture 476
XML for Analysis 476
XMLA APIs 477
Discover 478
Execute 479
XMLA and Analysis Services 480
Processing Architecture 482
Predictions 486
Data Mining Administration 487
Server Configuration 488
Data Mining Security 489
Security Requirements for Creating and Training Mining Objects 491
Security for Various Deployment Scenarios 491
Local Database and Analysis Services 492
Local Analysis Services and a Remote Database 493
Intranet Analysis Services and Databases on the Same Server 493
Analysis Services and Databases behind an HTTP Endpoint in an Internet Deployment 494
Configuring Analysis Services for Use with Data Mining Excel Add-Ins over HTTP 495
Summary 496
Chapter 16 Programming SQL Server Data Mining 497
Data Mining APIs 498
ADO 498
ADO.NET 500
ADOMD.NET 501
Server ADOMD.NET 501
AMO 501

Using Analysis Services APIs 502
Using Microsoft.AnalysisServices to Create and Manage Mining Models 502
AMO Basics 503
AMO Applications and Security 505
Object Creation 506
Creating Data Access Objects 507
Creating the Mining Structure 510
Creating the Mining Models 512
Processing Mining Models 513
Deploying Mining Models 515
Setting Mining Permissions 516
Browsing and Querying Mining Models 517
Predicting with ADOMD.NET 517
More on Table-Valued Parameters in ADOMD.NET 522
Browsing Models 525
Stored Procedures 527
Writing Stored Procedures 529
Stored Procedures and Prepare Invocations 530
A Stored Procedure Example 530
Executing Queries inside Stored Procedures 533
Returning Data Sets from Stored Procedures 534
Deploying and Debugging Stored Procedure Assemblies 537
Summary 538

Chapter 17 Extending SQL Server Data Mining 541
Plug-in Algorithms 542
Plug-in Algorithm Framework 543
Lifetime of a Plug-in Algorithm Instance 543
Conceptual Overview 545
Model Creation and Processing 547
Prediction 553
Content Navigation 554
Custom Functions 555
PMML 557
Managed vs. Native Plug-ins 557
Installing Plug-in Algorithms 558
Where to Find Out More about Plug-in Algorithms 558
Data Mining Viewers 558
Interfaces to Be Implemented 559
Rendering the Information 559
Retrieving Information from Analysis Services 560
Registering the Viewer 561
Where to Find Out More about Plug-in Viewers 561
Summary 562

Chapter 18 Implementing a Web Cross-Selling Application 563
Source Data Description 564
Building Your Model 564
Identifying the Data Mining Task 564
Using Decision Trees for Association 565
Using the Association Rules Algorithm 567
Comparing the Two Models 568
Making Predictions 570
Making Batch Prediction Queries 570
Using Singleton Prediction Queries 572
Integrating Predictions with Web Applications 573
Understanding Web Application Architecture 573
Setting the Permissions 574
Examining Sample Code for the Web Recommendation Application 575
Summary 578

Chapter 19 Conclusion and Additional Resources 581
Recapping the Highlights of SQL Server 2008 Data Mining 581
State-of-the-Art Algorithms 582
Easy-to-Use Tools 583
Simple-Yet-Powerful API 584
Integration with Sibling BI Technologies 584
Exploring New Data Mining Frontiers and Opportunities 585
Further Reference 586
Microsoft Data Mining 586
General Data Mining 586

Appendix A: Data Sets 589
MovieClick Data Set 589
Voting Records Data Set 591
Wine Sales 591
Foodmart 593
College Plans Data Set 593

Appendix B: Supported Functions 595
DMX Language Functions 595
VBA Functions 595
Excel Functions 595
ASSprocs Stored Procedures 605

Index 607