John MacGregor

Predictive Analysis with SAP®

The Comprehensive Guide

Galileo Press

Bonn • Boston

Contents

Introduction 17
Acknowledgments 21

PART I Predictive Analysis Overview

1 An Introduction to Predictive Analysis 25

1.1 Definitions of Predictive Analysis 25
1.2 The Value of Predictive Analysis 28
1.3 User Personas 31
1.4 Applications of Predictive Analysis 33
1.5 Classes of Applications 37
1.5.1 Time Series Analysis 37
1.5.2 Classification Analysis 37
1.5.3 Cluster Analysis 38
1.5.4 Association Analysis 38
1.5.5 Outlier Analysis 38
1.6 Algorithms for Predictive Analysis 39
1.7 The Predictive Analysis Process 41
1.8 Hot Topics and Trends 44
1.9 Challenges and Criteria for Success 45
1.10 Summary 47

2 An Overview of the Predictive Analysis Products in SAP 49

2.1 The Predictive Analysis Library in SAP HANA 53
2.1.1 PAL Workflow and Business Example 55
2.2 The R Integration for SAP HANA 59
2.2.1 R Integration Worked Business Example 60
2.3 SAP Predictive Analysis 63
2.4 SAP Business Solutions with Predictive Analysis 73
2.5 Summary 77

PART II Predictive Analysis Applied

3 Initial Data Exploration 81

3.1 Data Types 83
3.2 Data Visualization for Data Exploration 86
3.3 Sampling 92
3.4 Scaling 97
3.5 Binning 101
3.6 Outliers 104
3.7 Summary 105

4 Which Algorithm When? 107

4.1 The Main Factors When Selecting an Algorithm 107
4.2 Classes of Applications and Algorithms 109
4.3 Matrix of Application Tasks, Variable Types and Output 113
4.4 Which Algorithm Is the Best? 115
4.5 A Set of Rules for Which Algorithm When 116
4.6 Summary 118

5 When Mining, Beware of Mines 119

5.1 Data Mining Heaven and Hell 119
5.2 Five Myths 121
5.2.1 Myth No.1. Predictive Analysis is all about Algorithms ... 121
5.2.2 Myth No. 2. Predictive Analysis is all about Accuracy 122
5.2.3 Myth No. 3. Predictive Analysis Requires a Data Warehouse 122
5.2.4 Myth No. 4. Predictive Analysis is all about Vast Quantities of Data 123
5.2.5 Myth No. 5. Predictive Analysis is done by Predictive Experts 123
5.3 Five Pitfalls 124
5.3.1 Pitfall No. 1: Throwing in Data without Thinking 125
5.3.2 Pitfall No. 2: Lack of Business Knowledge 125
5.3.3 Pitfall No. 3: Lack of Data Knowledge 125
5.3.4 Pitfall No. 4: Erroneous Assumptions 126

5.3.5 Pitfall No. 5: Disorganized Project 126
5.4 Further Challenges and Resolution 126
5.4.1 Cause and Effect 127
5.4.2 Lies, Damned Lies, and Statistics 128
5.4.3 Model Overfitting 132
5.4.4 Correlation between the Independent Variables 135
5.5 Key Factors for Success 137
5.6 Summary 138

6 Applications in SAP 139

6.1 SAP Smart Meter Analytics 139
6.1.1 Application Description 140
6.1.2 Current and Planned Use of Predictive Analysis 141
6.1.3 Benefits 142
6.2 SAP Customer Engagement Intelligence 142
6.2.1 Application Description 143
6.2.2 Current and Planned Use of Predictive Analysis 146
6.2.3 Benefits 149
6.3 SAP Enterprise Inventory & Service-Level Optimization 149
6.3.1 Application Description 150
6.3.2 Current and Planned Use of Predictive Analysis 156
6.3.3 Benefits 157
6.4 SAP Precision Gaming 158
6.4.1 Application Description 158
6.4.2 Current and Planned Use of Predictive Analysis 160
6.4.3 Benefits 160
6.5 SAP Affinity Insight 161
6.5.1 Application Description 161
6.5.2 Current and Planned Use of Predictive Analysis 164
6.5.3 Benefits 165
6.6 SAP Demand Signal Management 166
6.6.1 Application Description 166
6.6.2 Current and Planned Use of Predictive Analysis 167
6.6.3 Benefits 171
6.7 SAP On-Shelf Availability 172
6.7.1 Application Description 172
6.7.2 Current and Planned Use of Predictive Analysis 175
6.7.3 Benefits 176

6.8 SAP Product Recommendation Intelligence 177
6.8.1 Application Description 177
6.8.2 Current and Planned Use of Predictive Analysis 180
6.8.3 Benefits 182
6.9 SAP Credit Insight 182
6.9.1 Application Description 182
6.9.2 Current and Planned Use of Predictive Analysis 183
6.9.3 Benefits 184
6.10 SAP Convergent Pricing Simulation 184
6.10.1 Application Description 184
6.10.2 Current and Planned Use of Predictive Analysis 187
6.10.3 Benefits 187
6.11 Summary 187

7 SAP Predictive Analysis 189

7.1 Getting Started in PA 189
7.2 Accessing and Viewing the Data Source 195
7.3 Preparing Data for Analysis 199
7.4 Applying Algorithms to Analyze the Data 202
7.4.1 In-Database Analysis using an SAP HANA Table and the PAL 203
7.4.2 In-Process Analysis using a CSV File and R Integration in PA 205
7.5 Running the Model and Viewing the Results 209
7.6 Deploying the Model in a Business Application 213
7.6.1 Exporting the Model as PMML 216
7.6.2 Sharing the Analysis in the Share View in PA 217
7.6.3 Exporting and Importing Analyses between PA Users 218
7.6.4 Exporting an SAP HANA PAL Model from PA as a Stored Procedure 218
7.7 Summary 219

PART III Predictive Analysis Categories

8 Outlier Analysis 223

8.1 Introduction to Outlier Analysis 223
8.2 Applications of Outlier Analysis 225

8.3 The Inter-Quartile Range Test 227
8.3.1 The Inter-Quartile Range Test in the PAL 227
8.3.2 An Example of the IQR Test in the PAL 228
8.3.3 An Example of the Inter-Quartile Range Test in PA 231
8.4 The Variance Test 232
8.4.1 An Example of the Variance Test in the PAL 233
8.5 K Nearest Neighbor Outlier 235
8.6 Anomaly Detection using Cluster Analysis 238
8.6.1 An Example of the Anomaly Detection Algorithm in the PAL 239
8.6.2 An Example of Anomaly Detection in PA 241
8.7 The Business Case for Outlier Analysis 243
8.8 Strengths and Weaknesses of Outlier Analysis 244
8.9 Summary 245

9 Association Analysis 247

9.1 Applications of Association Analysis 248
9.2 Apriori Association Analysis 250
9.3 Apriori Association Analysis in the PAL 255
9.4 An Example of Apriori Association Analysis in the PAL 257
9.5 An Example of Apriori in SAP Predictive Analysis 260
9.6 Apriori Lite Association Analysis 262
9.6.1 Example 1: Use All the Data to Calculate the Single Items Pre-Rule and Post-Rule 264
9.6.2 Example 2: 70% Sample Single Items Pre-Rule and Post-Rule 264
9.6.3 Example 3: Using All the Available Data to Sample and Calculate Single Items 265
9.7 Strengths and Weaknesses of Association Analysis 266
9.8 Business Case for Association Analysis 266
9.9 Summary 267

10 Cluster Analysis 269

10.1 Introduction to Cluster Analysis 269
10.2 Applications of Cluster Analysis 270
10.3 ABC Analysis 271

10.3.1 ABC Analysis in the PAL ... 273
10.3.2 An Example of ABC Analysis in the PAL ... 274
10.4 K-Means Cluster Analysis ... 275
10.4.1 A Visualization of K-Means ... 275
10.4.2 A Simple Example of K-Means in Excel ... 276
10.4.3 K-Means in the PAL ... 278
10.4.4 An Example of K-Means in the PAL ... 281
10.4.5 Choosing the Value of K ... 288
10.5 Silhouette ... 290
10.6 An Example of the Silhouette in the PAL ... 291
10.7 An Example of Validate K-Means in the PAL ... 292
10.8 Choosing the Initial Cluster Centers ... 294
10.9 Categorical Data and Numeric Cluster Analysis ... 296
10.10 Self-Organizing Maps ... 298
10.10.1 Self-Organizing Maps in the PAL ... 302
10.10.2 An Example of Self-Organizing Maps in the PAL ... 303
10.11 The Business Case for Cluster Analysis ... 309
10.12 Strengths and Weaknesses of Cluster Analysis ... 310
10.13 Summary ... 311

11 Classification Analysis ... 313

11.1 Introduction to Classification Analysis ... 313
11.2 Applications of Classification Analysis ... 314
11.3 An Introduction to Regression Analysis ... 315
11.4 An Introduction to Decision Trees ... 317
11.5 An Introduction to Nearest Neighbors ... 321
11.6 Summary ... 324

PART IV Classification Analysis

12 Classification Analysis—Regression ... 327

12.1 Bi-Variate Linear Regression ... 327
12.1.1 Bi-Variate Linear Regression in the PAL ... 332
12.1.2 An Example of Bi-Variate Linear Regression in the PAL ... 334
12.1.3 Predicting or Scoring the Model in the PAL ... 336
12.1.4 Bi-Variate Linear Regression in PA ... 339

12.1.5 Predicting or Scoring the Model in the PA 342
12.1.6 PMML and Exporting the Model 343
12.2 Bi-Variate Geometric, Exponential, and Logarithmic Regression ... 345
12.2.1 Bi-Variate Geometric Regression in the PAL 345
12.2.2 An Example of Bi-Variate Geometric Regression in the PAL 346
12.2.3 Using the Bi-Variate Geometric Regression Model to Predict 349
12.2.4 Bi-Variate Exponential Regression in PA using R 350
12.2.5 Bi-Variate Logarithmic Regression using the PA Native Algorithm 354
12.3 Multiple Linear Regression 357
12.3.1 An Example of Multiple Linear Regression in the PAL ... 357
12.3.2 An Example of Multiple Linear Regression in PA using the PAL 361
12.3.3 Predicting or Scoring the Model in the PAL 361
12.4 Multiple Exponential Regression 363
12.4.1 An Example of Multiple Exponential Regression in the PAL 363
12.4.2 An Example of Multiple Exponential Regression in PA using the PAL 366
12.4.3 Predicting or Scoring the Model in the PAL 366
12.5 Polynomial Regression 368
12.5.1 An Example of Polynomial Regression in the PAL 368
12.6 Logistic Regression 373
12.6.1 Logistic Regression in the PAL 373
12.6.2 An Example of Logistic Regression in the PAL 375
12.7 The Business Case for Regression Analysis 384
12.8 Strengths and Weaknesses of Regression Analysis 384
12.9 Summary 385

13 Classification Analysis—Decision Trees 387

13.1 Introduction to the Decision Trees Algorithm 387
13.2 CHAID Analysis 390
13.2.1 Worked Example of CHAID Analysis 390
13.2.2 CHAID Analysis in the PAL 396
13.2.3 CHAID Analysis in PA 399

13.2.4 Binning of Numeric Variables 402
13.2.5 Predicting using CHAID Analysis in the PAL 403
13.3 The C4.5 Algorithm 406
13.3.1 C4.5 in the PAL 410
13.3.2 C4.5 in PA 412
13.4 CNR Tree—Classification and Regression Trees 415
13.5 Decision Trees and Business Rules 424
13.6 Strengths and Weaknesses of Decision Trees 426
13.7 Summary 426

14 Classification Analysis—K Nearest Neighbor 427

14.1 Introduction 427
14.2 Worked Example 428
14.2.1 K Nearest Neighbor Analysis in the PAL 430
14.2.2 KNN Analysis in PA using the PAL KNN Algorithm 432
14.2.3 Categorical Target or Class Variable 436
14.3 Strengths and Weaknesses of the KNN Algorithm 437
14.4 Summary 438

PART V Advanced Predictive Analysis

15 Time Series Analysis 441

15.1 Introduction to Time Series Analysis 441
15.2 Time Series Patterns 443
15.3 Naïve Methods 445
15.4 Single Exponential Smoothing 446
15.4.1 Worked Example 447
15.4.2 Single Exponential Smoothing in the PAL 448
15.4.3 Single Exponential Smoothing in PA using the PAL 451
15.5 Double Exponential Smoothing 453
15.5.1 Worked Example 454
15.5.2 Double Exponential Smoothing in the PAL 455
15.5.3 Double Exponential Smoothing in PA using the PAL 457
15.6 Triple Exponential Smoothing 460
15.6.1 Worked Example 461

15.6.2 Triple Exponential Smoothing in the PAL 462
15.6.3 Triple Exponential Smoothing in PA using the PAL 464
15.7 Bi-Variate Linear Regression 467
15.8 The Business Case for Time Series Analysis 470
15.9 Strengths and Weaknesses of Time Series Analysis 470
15.10 Summary 471

16 Text Analysis and Text Mining 473

16.1 Introduction 473
16.2 Applications 474
16.3 Full Text Search 475
16.4 Fuzzy Search 481
16.5 Text Mining and Text Analysis 484
16.5.1 Examples 487
16.6 The Business Case for Text Analysis and Text Mining 496
16.7 Summary 496

17 Customer Applications 497

17.1 eBay 497
17.2 MKI Japan 499
17.3 CISCO 499
17.4 CIR Foods 500
17.5 Home Shopping Europe 24 501
17.6 Bigpoint 502
17.7 Other Customer Use Cases 503
17.7.1 Retail 503
17.7.2 Manufacturing 505
17.7.3 Transport and Logistics 506
17.7.4 Banking 507
17.7.5 Public Sector 509
17.7.6 High Tech 510
17.7.7 Oil and Gas 511
17.7.8 Utilities 511
17.8 Summary 513

Appendices 515

A References and Resources 515
A.1 References 515
A.2 Additional Resources 516
B The Author 519

Index 521