Contents

1 The Scorecard Development Process **1**

1.1 The Standard Scorecard 1

1.1.1 Types and Objectives of Scorecards 1

1.1.2 The `Good`, the `Bad`, and the Undefined . . . 2

1.1.3 The Standard Scorecard Format 4

1.2 The Scorecard Development Process 6

1.3 Problem Setup 7

1.4 Data Acquisition and Integration 8

1.5 EDA and Data Profiling 8

1.6 Data Preparation 9

1.7 Variable Selection 10

1.8 Model Development 11

1.9 Model Evaluation 11

1.10 Scorecard Creation and Scaling 12

1.11 Scorecard Deployment 12

1.12 Reject Inference 13

1.13 Monitoring and Reporting 13

1.14 Notes on the SAS Code 14

2 Data Acquisition and Integration **19**

2.1 Introduction 19

2.2 Variable Types 20

2.3 The Modeling (Mining) View 21

2.4 Sources of Data 23

2.5 Modeling and Deployment Windows 24
2.6 Data Alignment 25
2.7 Data Roll-up 28
2.7.1 Continuous Variables 29
2.7.2 Nominal and Ordinal Variables 31
2.8 Data Integration 33
2.8.1 Merging 33
2.8.2 Concatenation 35
2.8.3 Data Step or PROC SQL? 36
2.9 Integrity Checks 36

3 EDA and Data Profiling **41**
3.1 Introduction 41
3.2 Univariate Statistics 42
3.3 Variable Distribution 46
3.4 Characteristics Analysis 47
3.5 Cross-tabulation 51
3.6 Detection of Outliers 55
3.6.1 Identification of Outliers Using Ranges . . . 58
3.6.2 Identification of Outliers Using Clustering . . 59
3.6.3 Multistage Identification of Outliers 62
3.6.4 Treatment of Outliers 63

4 Measures of Predictive Power **65**
4.1 Introduction 65
4.2 Notation 68
4.2.1 Two Continuous Variables 68
4.2.2 Two Nominal Variables 68
4.2.3 Nominal x and Continuous y 70
4.3 Pearson Correlation Coefficient 72
4.4 Spearman Correlation Coefficient 76
4.5 The Pearson Chi-square Statistic 79
4.6 The Likelihood Ratio Test Statistic 82
4.7 Odds Ratio 85
4.8 The F-Test 92

4.9 The Gini Variance 93
4.10 The Entropy Variance 99
4.11 The Information Value 101
4.12 Automation of Variable Selection 105

5 Data Preparation **113**
5.1 Introduction 113
5.2 Reduction of Cardinality 115
5.3 Binning of Continuous Variables 120
5.4 Sampling and Calculation of Weights 124
5.4.1 Sampling Methods 124
5.4.2 Sample Size 125
5.4.3 Random Sampling 126
5.4.4 Balanced Sampling and Weights 128
5.4.5 Calculation of Sample Weights 130

6 The Credit Card Sample Dataset **135**
6.1 Introduction 135
6.2 The Data Dictionary 135

7 Logistic Regression **139**
7.1 Introduction 139
7.2 Basic Formulations 140
7.3 The Likelihood Equations 144
7.4 Information Matrix 148
7.5 Parameter Estimation 150
7.6 Model Fit Statistics 154
7.7 The Hosmer-Lemeshow Test 158
7.8 Testing the Global Null Hypothesis 161
7.9 The Score Statistic 163
7.10 Interpretation of Model Parameters 164
7.11 Confidence Intervals for Odds Ratios 166
7.12 Prior Probabilities and Weights 167

8 Coarse Classing and WOE **169**
8.1 Introduction 169
8.2 Definition of WOE 170
8.3 The Meaning of WOE 171
8.4 WOE and the Standard Scorecard 174
8.5 SAS Implementation 177
8.6 WOE for Continuous Variables 178

9 Methods of Variable Selection **185**
9.1 Introduction 185
9.2 Overview of Selection Methods 186
9.3 `Stepwise` Variables Selection 191
9.4 Forcing Variables into the Model 197
9.5 Controlling Selection Sequence 199
9.6 Logistic Regression Output 200

10 Model Evaluation **205**
10.1 Introduction 205
10.2 Validation and Confusion Matrix 206
10.3 The Lift and Lorenz Curve 213
10.4 The Gini Coefficient 219
10.5 The K-S Curve and Statistic 221
10.6 The ROC Curve and the c-Statistic 224
10.7 Overall Model Evaluation 229

11 Scorecard Scaling and Deployment **231**
11.1 The Standard Format 231
11.2 Scaling the Scorecard 233
11.3 Assigning the Score Points 235
11.4 SAS Implementation 238
11.5 Setting the Cut-off Levels 249

12 Monitoring and Reporting **253**
12.1 Objectives of Reporting 253
12.2 Stability Report 254

12.3 Scorecard Characteristics Analysis 259

13 Reject Inference **261**
13.1 Definition and Justification 261
13.2 Methods of Reject Inference 262
13.3 Simple Assignment Methods 263
13.3.1 Ignore Rejects 263
13.3.2 Assign All Rejects a `Bad` Status 263
13.3.3 Proportional Assignment 263
13.4 Augmentation Methods 266
13.4.1 Simple Augmentation 266
13.4.2 Fuzzy Augmentation 268
13.4.3 Parceling 270
13.5 Using Inferred Rejects 274

A Obtaining the Source Code **277**

B SAS Macros **279**

C Listing of SAS Macros **283**
C.1 Copyright and Software License 283
C.2 `%ApplyMap1` 284
C.3 `%ApplyMap2` 285
C.4 `%AVRollup` 286
C.5 `%B2Partitions` 288
C.6 `%BinContVar` 291
C.7 `%BinVar` 306
C.8 `%BSWeight` 306
C.9 `%CalcGrF` 307
C.10 `%CalcWOE` 309
C.11 `%ChcAnalysis` 311
C.12 `%ChiLike` 314
C.13 `%ClustOL` 315
C.14 `%ConfMat` 316
C.15 `%DummyGrpn` 317

C.16 %DummyGrps 319
C.17 %ENomNom 320
C.18 %EqWBinn 323
C.19 %ExtractCorr 324
C.20 %ExtrctTop 326
C.21 %Extremes 327
C.22 %ExtUnique 328
C.23 %FAugment 330
C.24 %GenSCDS 332
C.25 %GiniStat 338
C.26 %GNomNom 340
C.27 %GrFBinDV 343
C.28 %InfValue 345
C.29 %KSStat 347
C.30 %LiftChart 349
C.31 %OddsRatio 351
C.32 %Parceling 352
C.33 %PlotKS 357
C.34 %PlotLift 358
C.35 %PlotLorenz 358
C.36 %PlotROC 359
C.37 %PowerChiL 359
C.38 %PowerFG 361
C.39 %PowerGini 362
C.40 %PowerIV 363
C.41 %PowerOdds 364
C.42 %PropAssign 366
C.43 %R2Partitions 369
C.44 %RandomSample 371
C.45 %ReduceCats 371
C.46 %ROC 385
C.47 %SAugment 388
C.48 %SCCCode 390
C.49 %SCCSV 393

C.50 %SCSQLCode 395
C.51 %SCSSasCode 399
C.52 %TRollup 403
C.53 %VarMode 404

About the Author **407**

Bibliography **409**

Index **411**