



From *Predictive Modeling with SAS[®] Enterprise Miner[™]*,
Second Edition. Full book available for purchase [here](#).

Contents

Preface	xi
About This Book	xv
About The Author	xix
Acknowledgments	xxi
Chapter 1: Research Strategy	1
1.1 Introduction	1
1.2 Measurement Scales for Variables	1
1.3 Defining the Target	2
1.3.1 Predicting Response to Direct Mail	2
1.3.2 Predicting Risk in the Auto Insurance Industry	3
1.3.3 Predicting Rate Sensitivity of Bank Deposit Products	4
1.3.4 Predicting Customer Attrition	6
1.3.5 Predicting a Nominal Categorical (Unordered Polychotomous) Target	7
1.4 Sources of Modeling Data	8
1.4.1 Comparability between the Sample and Target Universe	8
1.4.2 Observation Weights	8
1.5 Pre-Processing the Data	8
1.5.1 Data Cleaning Before Launching SAS Enterprise Miner	9
1.5.2 Data Cleaning After Launching SAS Enterprise Miner	9
1.6 Alternative Modeling Strategies	10
1.6.1 Regression with a Moderate Number of Input Variables	10
1.6.2 Regression with a Large Number of Input Variables	11
1.7 Notes	11
Chapter 2: Getting Started with Predictive Modeling	13
2.1 Introduction	14
2.2 Opening SAS Enterprise Miner 12.1	14
2.3 Creating a New Project in SAS Enterprise Miner 12.1	14
2.4 The SAS Enterprise Miner Window	15
2.5 Creating a SAS Data Source	16
2.6 Creating a Process Flow Diagram	25
2.7 Sample Nodes	26
2.7.1 Input Data Node	26
2.7.2 Data Partition Node	27
2.7.3 Filter Node	28

2.7.4 File Import Node	32
2.7.5 Time Series Node	35
2.7.6 Merge Node.....	45
2.7.7 Append Node	48
2.8 Tools for Initial Data Exploration.....	50
2.8.1 Stat Explore Node.....	51
2.8.2 MultiPlot Node	56
2.8.3 Graph Explore Node.....	58
2.8.4 Variable Clustering Node.....	61
2.8.5 Cluster Node	69
2.8.6 Variable Selection Node.....	72
2.9 Tools for Data Modification	79
2.9.1 Drop Node	79
2.9.2 Replacement Node.....	80
2.9.3 Impute Node.....	83
2.9.4 Interactive Binning Node	83
2.9.5 Principal Components Node	90
2.9.6 Transform Variables Node.....	95
2.10 Utility Nodes	101
2.10.1 SAS Code Node	101
2.11 Appendix to Chapter 2.....	107
2.11.1 The Type, the Measurement Scale, and the Number of Levels of a Variable	107
2.11.2 Eigenvalues, Eigenvectors, and Principal Components.....	110
2.11.3 Cramer's V.....	113
2.11.4 Calculation of Chi-Square Statistic and Cramer's V for a Continuous Input.....	113
2.12 Exercises.....	115
Chapter 3: Variable Selection and Transformation of Variables.....	117
3.1 Introduction	117
3.2 Variable Selection	118
3.2.1 Continuous Target with Numeric Interval-scaled Inputs (Case 1)	119
3.2.2 Continuous Target with Nominal-Categorical Inputs (Case 2).....	124
3.2.3 Binary Target with Numeric Interval-scaled Inputs (Case 3)	129
3.2.4 Binary Target with Nominal-scaled Categorical Inputs (Case 4)	135
3.3 Variable Selection Using the Variable Clustering Node.....	138
3.3.1 Selection of the Best Variable from Each Cluster.....	140
3.3.2 Selecting the Cluster Components.....	148
3.4 Variable Selection Using the Decision Tree Node.....	150
3.5 Transformation of Variables	153
3.5.1 Transform Variables Node.....	153
3.5.2 Transformation before Variable Selection	155
3.5.3 Transformation after Variable Selection	157
3.5.4 Passing More Than One Type of Transformation for Each Interval Input to the Next Node.....	159
3.5.5 Saving and Exporting the Code Generated by the Transform Variables Node.....	163

3.6 Summary	163
3.7 Appendix to Chapter 3.....	164
3.7.1 Changing the Measurement Scale of a Variable in a Data Source	164
3.7.2 SAS Code for Comparing Grouped Categorical Variables with the Ungrouped Variables	165
Exercises.....	166
Note	167
Chapter 4: Building Decision Tree Models to Predict Response and Risk.....	169
4.1 Introduction	170
4.2 An Overview of the Tree Methodology in SAS Enterprise Miner	170
4.2.1 Decision Trees	170
4.2.2 Decision Tree Models	170
4.2.3 Decision Tree Models vs. Logistic Regression Models.....	172
4.2.4 Applying the Decision Tree Model to Prospect Data.....	173
4.2.5 Calculation of the Worth of a Tree.....	173
4.2.6 Roles of the Training and Validation Data in the Development of a Decision Tree.....	175
4.2.7 Regression Tree	176
4.3 Development of the Tree in SAS Enterprise Miner.....	176
4.3.1 Growing an Initial Tree	176
4.3.2 P-value Adjustment Options	183
4.3.3 Controlling Tree Growth: Stopping Rules	185
4.3.4 Pruning: Selecting the Right-Sized Tree Using Validation Data.....	185
4.3.5 Step-by-Step Illustration of Growing and Pruning a Tree.....	188
4.3.6 Average Profit vs. Total Profit for Comparing Trees of Different Sizes.....	192
4.3.7 Accuracy /Misclassification Criterion in Selecting the Right-sized Tree (Classification of Records and Nodes by Maximizing Accuracy).....	193
4.3.8 Assessment of a Tree or Sub-tree Using Average Square Error.....	194
4.3.9 Selection of the Right-sized Tree	194
4.4 A Decision Tree Model to Predict Response to Direct Marketing.....	195
4.4.1 Testing Model Performance with a Test Data Set	204
4.4.2 Applying the Decision Tree Model to Score a Data Set	205
4.5 Developing a Regression Tree Model to Predict Risk	208
4.5.1 Summary of the Regression Tree Model to Predict Risk.....	214
4.6 Developing Decision Trees Interactively	215
4.6.1 Interactively Modifying an Existing Decision Tree	215
4.6.2 Growing a Tree Interactively Starting from the Root Node	225
4.6.3 Developing the Maximal Tree in Interactive Mode	231
4.7 Summary	233
4.8 Appendix to Chapter 4.....	234
4.8.1 Pearson's Chi-Square Test.....	234
4.8.2 Adjusting the Predicted Probabilities for Over-sampling	235
4.8.3 Expected Profits Using Unadjusted Probabilities	236
4.8.4 Expected Profits Using Adjusted Probabilities	236
4.9 Exercises.....	236

Chapter 5: Neural Network Models to Predict Response and Risk.....	239
5.1 Introduction	240
5.1.1 Target Variables for the Models.....	240
5.1.2 Neural Network Node Details.....	240
5.2 A General Example of a Neural Network Model	241
5.2.1 Input Layer	242
5.2.2 Hidden Layers	242
5.2.3 Output Layer or Target Layer	246
5.2.4 Activation Function of the Output Layer	247
5.3 Estimation of Weights in a Neural Network Model.....	247
5.4 A Neural Network Model to Predict Response	249
5.4.1 Setting the Neural Network Node Properties	250
5.4.2 Assessing the Predictive Performance of the Estimated Model	254
5.4.3 Receiver Operating Characteristic (ROC) Charts	258
5.4.4 How Did the Neural Network Node Pick the Optimum Weights for This Model?.....	261
5.4.5 Scoring a Data Set Using the Neural Network Model	263
5.4.6 Score Code.....	266
5.5 A Neural Network Model to Predict Loss Frequency in Auto Insurance.....	266
5.5.1 Loss Frequency as an Ordinal Target	267
5.5.3 Classification of Risks for Rate Setting in Auto Insurance with Predicted Probabilities	279
5.6 Alternative Specifications of the Neural Networks	279
5.6.1 A Multilayer Perceptron (MLP) Neural Network	279
5.6.2 A Radial Basis Function (RBF) Neural Network	281
5.7 Comparison of Alternative Built-in Architectures of the Neural Network Node	286
5.7.1 Multilayer Perceptron (MLP) Network.....	287
5.7.2 Ordinary Radial Basis Function with Equal Heights and Widths (ORBFEQ)	288
5.7.3 Ordinary Radial Basis Function with Equal Heights and Unequal Widths (ORBFUN).....	291
5.7.4 Normalized Radial Basis Function with Equal Widths and Heights (NRBFEQ).....	292
5.7.5 Normalized Radial Basis Function with Equal Heights and Unequal Widths (NRBFEH). 295	
5.7.6 Normalized Radial Basis Function with Equal Widths and Unequal Heights (NRBFEW) 297	
5.7.7 Normalized Radial Basis Function with Equal Volumes (NRBFEV)	300
5.7.8 Normalized Radial Basis Function with Unequal Widths and Heights (NRBFUN).....	302
5.7.9 User-Specified Architectures	305
5.8 AutoNeural Node.....	307
5.9 DMNeural Node	309
5.10 Dmine Regression Node	312
5.11 Comparing the Models Generated by DMNeural, AutoNeural, and Dmine Regression Nodes	314
5.12 Summary	316
5.13 Appendix to Chapter 5.....	317
5.14 Exercises.....	318
Chapter 6: Regression Models.....	321
6.1 Introduction	321

6.2 What Types of Models Can Be Developed Using the Regression Node?	321
6.2.1 Models with a Binary Target	321
6.2.2 Models with an Ordinal Target	324
6.2.3 Models with a Nominal (Unordered) Target	329
6.2.4 Models with Continuous Targets	333
6.3 An Overview of Some Properties of the Regression Node	333
6.3.1 Regression Type Property	333
6.3.2 Link Function Property	333
6.3.3 Selection Model Property	335
6.3.4 Selection Criterion Property	348
6.4 Business Applications	358
6.4.1 Logistic Regression for Predicting Response to a Mail Campaign	359
6.4.2 Regression for a Continuous Target	371
6.5 Summary	379
6.6 Appendix to Chapter 6	380
6.6 Exercises	382
Chapter 7: Comparison and Combination of Different Models	383
7.1 Introduction	383
7.2 Models for Binary Targets: An Example of Predicting Attrition	384
7.2.1 Logistic Regression for Predicting Attrition	386
7.2.2 Decision Tree Model for Predicting Attrition	387
7.2.3 A Neural Network Model for Predicting Attrition	389
7.3 Models for Ordinal Targets: An Example of Predicting the Risk of Accident Risk	392
7.3.1 Lift Charts and Capture Rates for Models with Ordinal Targets	393
7.3.2 Logistic Regression with Proportional Odds for Predicting Risk in Auto Insurance	394
7.3.3 Decision Tree Model for Predicting Risk in Auto Insurance	396
7.3.4 Neural Network Model for Predicting Risk in Auto Insurance	400
7.4 Comparison of All Three Accident Risk Models	401
7.5 Boosting and Combining Predictive Models	402
7.5.1 Gradient Boosting	402
7.5.2 Stochastic Gradient Boosting	404
7.5.3 An Illustration of Boosting Using the Gradient Boosting Node	404
7.5.4 The Ensemble Node	407
7.5.5 Comparing the Gradient Boosting and Ensemble Methods of Combining Models	410
7.6 Appendix to Chapter 7	411
7.6.1 Least Squares Loss	411
7.6.2 Least Absolute Deviation Loss	411
7.6.3 Huber-M Loss	411
7.6.4 Logit Loss	412
7.7 Exercises	412
Chapter 8: Customer Profitability	415
8.1 Introduction	415
8.2 Acquisition Cost	417
8.3 Cost of Default	418

8.4 Revenue	419
8.5 Profit	419
8.6 The Optimum Cut-off Point.....	421
8.7 Alternative Scenarios of Response and Risk.....	422
8.8 Customer Lifetime Value	422
8.9 Suggestions for Extending Results	423
Chapter 9: Introduction to Predictive Modeling with Textual Data	425
9.1 Introduction	425
9.1.1 Quantifying Textual Data: A Simplified Example.....	426
9.1.2 Dimension Reduction and Latent Semantic Indexing	429
9.1.3 Summary of the Steps in Quantifying Textual Information	431
9.2 Retrieving Documents from the World Wide Web.....	432
9.2.1 The %TMFILTER Macro.....	432
9.3 Creating a SAS Data Set from Text Files.....	433
9.4 The Text Import Node.....	436
9.5 Creating a Data Source for Text Mining	436
9.6 Text Parsing Node.....	436
9.7 Text Filter Node.....	440
9.7.1 Frequency Weighting	440
9.7.2 Term Weighting.....	440
9.7.3 Adjusted Frequencies	441
9.7.4 Frequency Weighting Methods	441
9.7.5 Term Weighting Methods	441
9.8 Text Topic Node	445
9.8.1 Developing a Predictive Equation Using the Output Data Set Created by the Text Topic Node.....	449
9.9 Text Cluster Node	450
9.9.1 Hierarchical Clustering	451
9.9.2 Expectation-Maximization (EM) Clustering	452
9.9.3 Using the Text Cluster Node	458
9.10 Exercises.....	461
Index.....	463