



An Introduction to Predictive Modeling with SAS® Enterprise Miner™: Practical Solutions for Business Applications, Third Edition. Full book available for purchase [here](#).

## Contents

<b>About This Book .....</b>	<b>xi</b>
<b>About The Author .....</b>	<b>xiii</b>
<b>Chapter 1: Research Strategy.....</b>	<b>1</b>
1.1 Introduction .....	1
1.2 Types of Inputs.....	2
1.2.1 Measurement Scales for Variables.....	2
1.2.2 Predictive Models with Textual Data.....	2
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.....	4
1.3.3 Predicting Rate Sensitivity of Bank Deposit Products .....	5
1.3.4 Predicting Customer Attrition .....	7
1.3.5 Predicting a Nominal Categorical (Unordered Polychotomous) Target.....	8
1.4 Sources of Modeling Data.....	10
1.4.1 Comparability between the Sample and Target Universe.....	10
1.4.2 Observation Weights.....	10
1.5 Pre-Processing the Data .....	10
1.5.1 Data Cleaning Before Launching SAS Enterprise Miner .....	11
1.5.2 Data Cleaning After Launching SAS Enterprise Miner .....	11
1.6 Alternative Modeling Strategies .....	12
1.6.1 Regression with a Moderate Number of Input Variables .....	12
1.6.2 Regression with a Large Number of Input Variables.....	13
1.7 Notes .....	13
<b>Chapter 2: Getting Started with Predictive Modeling.....</b>	<b>15</b>
2.1 Introduction .....	16
2.2 Opening SAS Enterprise Miner 14.1 .....	16
2.3 Creating a New Project in SAS Enterprise Miner 14.1 .....	16
2.4 The SAS Enterprise Miner Window .....	17
2.5 Creating a SAS Data Source.....	18
2.6 Creating a Process Flow Diagram .....	27
2.7 Sample Nodes .....	27
2.7.1 Input Data Node.....	27
2.7.2 Data Partition Node .....	29
2.7.3 Filter Node.....	29
2.7.4 File Import Node .....	33

2.7.5 Time Series Nodes .....	36
2.7.6 Merge Node.....	50
2.7.7 Append Node .....	53
2.8 Tools for Initial Data Exploration.....	56
2.8.1 Stat Explore Node.....	57
2.8.2 MultiPlot Node .....	64
2.8.3 Graph Explore Node.....	67
2.8.4 Variable Clustering Node.....	73
2.8.5 Cluster Node .....	82
2.8.6 Variable Selection Node.....	85
2.9 Tools for Data Modification .....	94
2.9.1 Drop Node .....	94
2.9.2 Replacement Node.....	95
2.9.3 Impute Node.....	98
2.9.4 Interactive Binning Node .....	99
2.9.5 Principal Components Node .....	106
2.9.6 Transform Variables Node.....	112
2.10 Utility Nodes .....	120
2.10.1 SAS Code Node .....	120
2.11 Appendix to Chapter 2.....	126
2.11.1 The Type, the Measurement Scale, and the Number of Levels of a Variable .....	126
2.11.2 Eigenvalues, Eigenvectors, and Principal Components.....	129
2.11.3 Cramer's V.....	132
2.11.4 Calculation of Chi-Square Statistic and Cramer's V for a Continuous Input.....	133
2.12 Exercises.....	135
Notes .....	137
<b>Chapter 3: Variable Selection and Transformation of Variables.....</b>	<b>139</b>
3.1 Introduction .....	139
3.2 Variable Selection .....	140
3.2.1 Continuous Target with Numeric Interval-scaled Inputs (Case 1) .....	140
3.2.2 Continuous Target with Nominal-Categorical Inputs (Case 2).....	147
3.2.3 Binary Target with Numeric Interval-scaled Inputs (Case 3) .....	153
3.2.4 Binary Target with Nominal-scaled Categorical Inputs (Case 4) .....	158
3.3 Variable Selection Using the Variable Clustering Node.....	162
3.3.1 Selection of the Best Variable from Each Cluster.....	164
3.3.2 Selecting the Cluster Components.....	174
3.4 Variable Selection Using the Decision Tree Node.....	176
3.5 Transformation of Variables .....	179
3.5.1 Transform Variables Node.....	179
3.5.2 Transformation before Variable Selection .....	181
3.5.3 Transformation after Variable Selection .....	183
3.5.4 Passing More Than One Type of Transformation for Each Interval Input to the Next Node.....	185

3.5.5 Saving and Exporting the Code Generated by the Transform Variables Node.....	189
3.6 Summary .....	190
3.7 Appendix to Chapter 3.....	190
3.7.1 Changing the Measurement Scale of a Variable in a Data Source .....	190
3.7.2 SAS Code for Comparing Grouped Categorical Variables with the Ungrouped Variables.....	192
Exercises.....	192
Note .....	193
<b>Chapter 4: Building Decision Tree Models to Predict Response and Risk....</b>	<b>195</b>
4.1 Introduction .....	196
4.2 An Overview of the Tree Methodology in SAS® Enterprise Miner™ .....	196
4.2.1 Decision Trees .....	196
4.2.2 Decision Tree Models .....	196
4.2.3 Decision Tree Models vs. Logistic Regression Models .....	198
4.2.4 Applying the Decision Tree Model to Prospect Data .....	198
4.2.5 Calculation of the Worth of a Tree.....	199
4.2.6 Roles of the Training and Validation Data in the Development of a Decision Tree....	201
4.2.7 Regression Tree .....	202
4.3 Development of the Tree in SAS Enterprise Miner .....	202
4.3.1 Growing an Initial Tree.....	202
4.3.2 P-value Adjustment Options .....	209
4.3.3 Controlling Tree Growth: Stopping Rules.....	211
4.3.3.1 Controlling Tree Growth through the Split Size Property .....	211
4.3.4 Pruning: Selecting the Right-Sized Tree Using Validation Data.....	211
4.3.5 Step-by-Step Illustration of Growing and Pruning a Tree.....	213
4.3.6 Average Profit vs. Total Profit for Comparing Trees of Different Sizes.....	218
4.3.7 Accuracy /Misclassification Criterion in Selecting the Right-sized Tree (Classification of Records and Nodes by Maximizing Accuracy) .....	218
4.3.8 Assessment of a Tree or Sub-tree Using Average Square Error .....	220
4.3.9 Selection of the Right-sized Tree .....	220
4.4 Decision Tree Model to Predict Response to Direct Marketing .....	221
4.4.1 Testing Model Performance with a Test Data Set .....	230
4.4.2 Applying the Decision Tree Model to Score a Data Set .....	231
4.5 Developing a Regression Tree Model to Predict Risk .....	236
4.5.1 Summary of the Regression Tree Model to Predict Risk.....	243
4.6 Developing Decision Trees Interactively .....	244
4.6.1 Interactively Modifying an Existing Decision Tree.....	244
4.6.3 Developing the Maximal Tree in Interactive Mode .....	266
4.7 Summary .....	269
4.8 Appendix to Chapter 4.....	270
4.8.1 Pearson’s Chi-Square Test.....	270
4.8.2 Calculation of Impurity Reduction using Gini Index .....	271
4.8.3 Calculation of Impurity Reduction/Information Gain using Entropy.....	272
4.8.4 Adjusting the Predicted Probabilities for Over-sampling .....	274

4.8.5 Expected Profits Using Unadjusted Probabilities .....	275
4.8.6 Expected Profits Using Adjusted Probabilities .....	275
4.9 Exercises.....	275
Notes .....	277
<b>Chapter 5: Neural Network Models to Predict Response and Risk.....</b>	<b>279</b>
5.1 Introduction .....	280
5.1.1 Target Variables for the Models.....	280
5.1.2 Neural Network Node Details.....	281
5.2 General Example of a Neural Network Model.....	281
5.2.1 Input Layer .....	282
5.2.2 Hidden Layers .....	283
5.2.3 Output Layer or Target Layer.....	288
5.2.4 Activation Function of the Output Layer .....	289
5.3 Estimation of Weights in a Neural Network Model.....	290
5.4 Neural Network Model to Predict Response .....	291
5.4.1 Setting the Neural Network Node Properties.....	293
5.4.2 Assessing the Predictive Performance of the Estimated Model.....	297
5.4.3 Receiver Operating Characteristic (ROC) Charts .....	300
5.4.4 How Did the Neural Network Node Pick the Optimum Weights for This Model?.....	303
5.4.5 Scoring a Data Set Using the Neural Network Model .....	305
5.4.6 Score Code .....	308
5.5 Neural Network Model to Predict Loss Frequency in Auto Insurance .....	308
5.5.1 Loss Frequency as an Ordinal Target .....	309
5.5.1.1 Target Layer Combination and Activation Functions .....	311
5.5.3 Classification of Risks for Rate Setting in Auto Insurance with Predicted Probabilities .....	321
5.6 Alternative Specifications of the Neural Networks .....	322
5.6.1 A Multilayer Perceptron (MLP) Neural Network.....	322
5.6.2 Radial Basis Function (RBF) Neural Network.....	324
5.7 Comparison of Alternative Built-in Architectures of the Neural Network Node .....	330
5.7.1 Multilayer Perceptron (MLP) Network.....	332
5.7.2 Ordinary Radial Basis Function with Equal Heights and Widths (ORBFEQ) .....	333
5.7.3 Ordinary Radial Basis Function with Equal Heights and Unequal Widths (ORBFUN).....	335
5.7.4 Normalized Radial Basis Function with Equal Widths and Heights (NRBFEQ).....	338
5.7.5 Normalized Radial Basis Function with Equal Heights and Unequal Widths (NRBFEH).....	340
5.7.6 Normalized Radial Basis Function with Equal Widths and Unequal Heights (NRBFEW).....	343
5.7.7 Normalized Radial Basis Function with Equal Volumes (NRBFEV) .....	346
5.7.8 Normalized Radial Basis Function with Unequal Widths and Heights (NRBFUN).....	348
5.7.9 User-Specified Architectures.....	351
5.8 AutoNeural Node.....	354
5.9 DMNeural Node.....	356
5.10 Dmine Regression Node .....	358

5.11 Comparing the Models Generated by DMNeural, AutoNeural, and Dmine Regression Nodes .....	360
5.12 Summary .....	362
5.13 Appendix to Chapter 5.....	363
5.14 Exercises.....	365
Notes .....	367
<b>Chapter 6: Regression Models.....</b>	<b>369</b>
6.1 Introduction .....	369
6.2 What Types of Models Can Be Developed Using the Regression Node? .....	369
6.2.1 Models with a Binary Target .....	369
6.2.2 Models with an Ordinal Target.....	373
6.2.3 Models with a Nominal (Unordered) Target.....	379
6.2.4 Models with Continuous Targets.....	383
6.3 An Overview of Some Properties of the Regression Node.....	383
6.3.1 Regression Type Property .....	384
6.3.2 Link Function Property.....	384
6.3.3 Selection Model Property .....	386
6.3.4 Selection Criterion Property.....	403
6.4 Business Applications .....	415
6.4.1 Logistic Regression for Predicting Response to a Mail Campaign .....	417
6.4.2 Regression for a Continuous Target .....	431
6.5 Summary .....	442
6.6 Appendix to Chapter 6.....	443
6.6.1 SAS Code .....	443
6.6.2 Examples of the selection criteria when the Model Selection property set to Forward. ....	447
6.7 Exercises.....	451
Notes .....	452
<b>Chapter 7: Comparison and Combination of Different Models .....</b>	<b>453</b>
7.1 Introduction .....	453
7.2 Models for Binary Targets: An Example of Predicting Attrition .....	454
7.2.1 Logistic Regression for Predicting Attrition.....	456
7.2.2 Decision Tree Model for Predicting Attrition.....	458
7.2.3 A Neural Network Model for Predicting Attrition .....	460
7.3 Models for Ordinal Targets: An Example of Predicting the Risk of Accident Risk .....	464
7.3.1 Lift Charts and Capture Rates for Models with Ordinal Targets.....	465
7.3.2 Logistic Regression with Proportional Odds for Predicting Risk in Auto Insurance .	466
7.3.3 Decision Tree Model for Predicting Risk in Auto Insurance .....	469
7.3.4 Neural Network Model for Predicting Risk in Auto Insurance.....	473
7.4 Comparison of All Three Accident Risk Models .....	476
7.5 Boosting and Combining Predictive Models.....	476
7.5.1 Gradient Boosting .....	477
7.5.2 Stochastic Gradient Boosting .....	479
7.5.3 An Illustration of Boosting Using the Gradient Boosting Node.....	479

7.5.4 The Ensemble Node .....	482
7.5.5 Comparing the Gradient Boosting and Ensemble Methods of Combining Models ...	485
7.6 Appendix to Chapter 7 .....	486
7.6.1 Least Squares Loss .....	486
7.6.2 Least Absolute Deviation Loss.....	486
7.6.3 Huber-M Loss .....	487
7.6.4 Logit Loss .....	487
7.7 Exercises.....	488
Note .....	488
<b>Chapter 8: Customer Profitability .....</b>	<b>489</b>
8.1 Introduction .....	489
8.2 Acquisition Cost.....	491
8.3 Cost of Default .....	492
8.5 Profit.....	493
8.6 The Optimum Cutoff Point .....	495
8.7 Alternative Scenarios of Response and Risk.....	496
8.8 Customer Lifetime Value .....	496
8.9 Suggestions for Extending Results.....	497
Note .....	497
<b>Chapter 9: Introduction to Predictive Modeling with Textual Data .....</b>	<b>499</b>
9.1 Introduction .....	499
9.1.1 Quantifying Textual Data: A Simplified Example.....	500
9.1.2 Dimension Reduction and Latent Semantic Indexing .....	503
9.1.3 Summary of the Steps in Quantifying Textual Information .....	506
9.2 Retrieving Documents from the World Wide Web.....	507
9.2.1 The %TMFILTER Macro.....	507
9.3 Creating a SAS Data Set from Text Files.....	509
9.4 The Text Import Node.....	512
9.5 Creating a Data Source for Text Mining .....	514
9.6 Text Parsing Node.....	516
9.7 Text Filter Node.....	521
9.7.1 Frequency Weighting .....	521
9.7.2 Term Weighting .....	521
9.7.3 Adjusted Frequencies .....	521
9.7.4 Frequency Weighting Methods .....	521
9.7.5 Term Weighting Methods .....	523
9.8 Text Topic Node .....	528
9.8.1 Developing a Predictive Equation Using the Output Data Set Created by the Text Topic Node .....	533
9.9 Text Cluster Node .....	534
9.9.1 Hierarchical Clustering .....	535
9.9.2 Expectation-Maximization (EM) Clustering .....	536
9.9.3 Using the Text Cluster Node .....	542

<b>9.10 Exercises</b> .....	<b>546</b>
<b>Notes</b> .....	<b>546</b>
<b>Index</b> .....	<b>547</b>