Customer and Business Analytics
Applied Data Mining for Business Decision Making Using R

Daniel S. Putler
Robert E. Krider

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the Taylor & Francis Group an Informa business
A CHAPMAN & HALL BOOK
Contents

List of Figures xiii

List of Tables xxvi

Preface xxvii

I Purpose and Process 1

1 Database Marketing and Data Mining 3

1.1 Database Marketing 4

1.1.1 Common Database Marketing Applications 5

1.1.2 Obstacles to Implementing a Database Marketing Program 8

1.1.3 Who Stands to Benefit the Most from the Use of Database Marketing? 9

1.2 Data Mining 9

1.2.1 Two Definitions of Data Mining 9

1.2.2 Classes of Data Mining Methods 10

1.2.2.1 Grouping Methods 10

1.2.2.2 Predictive Modeling Methods 11

1.3 Linking Methods to Marketing Applications 14

2 A Process Model for Data Mining—CRISP-DM 17

2.1 History and Background 17

2.2 The Basic Structure of CRISP-DM 19
4.2.1 The Probability of a Relationship between the Variables 89
4.2.2 Outliers ........................................ 91
4.3 Multiple Regression .................................. 91
  4.3.1 Categorical Predictors ....................... 92
  4.3.2 Nonlinear Relationships and Variable Transformations 94
  4.3.3 Too Many Predictor Variables: Overfitting and Adjusted R^2 .................................. 97
4.4 Summary ............................................ 98
4.5 Data Visualization and Linear Regression Tutorial ..... 99

5 Logistic Regression ................................. 117
  5.1 A Graphical Illustration of the Problem ........ 118
  5.2 The Generalized Linear Model .................. 121
  5.3 Logistic Regression Details ..................... 124
  5.4 Logistic Regression Tutorial .................... 126
    5.4.1 Highly Targeted Database Marketing ........ 126
    5.4.2 Oversampling .................................. 127
    5.4.3 Overfitting and Model Validation ........... 128

6 Lift Charts ........................................... 147
  6.1 Constructing Lift Charts ....................... 147
    6.1.1 Predict, Sort, and Compare to Actual Behavior .... 147
    6.1.2 Correcting Lift Charts for Oversampling .... 151
  6.2 Using Lift Charts .................................. 154
  6.3 Lift Chart Tutorial ................................ 159

7 Tree Models ......................................... 165
  7.1 The Tree Algorithm .............................. 166
    7.1.1 Calibrating the Tree on an Estimation Sample .... 167
    7.1.2 Stopping Rules and Controlling Overfitting .... 170
  7.2 Trees Models Tutorial ........................... 172
## 8 Neural Network Models

8.1 The Biological Inspiration for Artificial Neural Networks ... 187
8.2 Artificial Neural Networks as Predictive Models ............... 192
8.3 Neural Network Models Tutorial .............................. 194

## 9 Putting It All Together

9.1 Stepwise Variable Selection ................................. 201
9.2 The Rapid Model Development Framework .................... 204
  9.2.1 Up-Selling Using the Wesbrook Database ............... 204
  9.2.2 Think about the Behavior That You Are Trying to Predict .......... 205
  9.2.3 Carefully Examine the Variables Contained in the Data Set .......... 205
  9.2.4 Use Decision Trees and Regression to Find the Important Predictor Variables .......... 207
  9.2.5 Use a Neural Network to Examine Whether Nonlinear Relationships Are Present .......... 208
  9.2.6 If There Are Nonlinear Relationships, Use Visualization to Find and Understand Them .......... 209
9.3 Applying the Rapid Development Framework Tutorial ........... 210

## III Grouping Methods

10 Ward’s Method of Cluster Analysis and Principal Components

10.1 Summarizing Data Sets ...................................... 235
10.2 Ward’s Method of Cluster Analysis .......................... 236
  10.2.1 A Single Variable Example ............................. 238
  10.2.2 Extension to Two or More Variables .................... 240
10.3 Principal Components ........................................... 242
10.4 Ward’s Method Tutorial ......................................... 248
11 K-Centroids Partitioning Cluster Analysis 259

11.1 How K-Centroid Clustering Works .......................... 260
  11.1.1 The Basic Algorithm to Find K-Centroids Clusters .. 260
  11.1.2 Specific K-Centroid Clustering Algorithms ......... 261
11.2 Cluster Types and the Nature of Customer Segments .... 264
11.3 Methods to Assess Cluster Structure .................... 267
  11.3.1 The Adjusted Rand Index to Assess Cluster Structure Reproducibility ........................................ 268
  11.3.2 The Calinski-Harabasz Index to Assess within Cluster Homogeneity and between Cluster Separation ....... 274
11.4 K-Centroids Clustering Tutorial .......................... 275

Bibliography 283

Index 287