Data Mining with SPSS Modeler

Tilo Wendler • Sören Gröttrup

Data Mining with SPSS Modeler

Theory, Exercises and Solutions

Springer

Tilo Wendler
HTW Berlin
University of Applied Sciences
Berlin
Germany

Sören Gröttrup
Stuttgart
Germany

ISBN 978-3-319-28707-2 ISBN 978-3-319-28709-6 (eBook)
DOI 10.1007/978-3-319-28709-6

Library of Congress Control Number: 2016941509

Mathematics Subject Classification (2010): 62-07; 62Hxx; 62Pxx; 62-01

© Springer International Publishing Switzerland 2016
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

Preface

Data Analytics, Data Mining and Big Data are terms often used in everyday business. Companies collect more and more data and store it in databases, with the hope of finding helpful patterns that can improve business. Shortly after deciding to more use of such data, managers often confess that analysing these datasets is resource-consuming and anything but easy. Involving the firm's IT-experts leads to a discussion regarding which tools to use. Very few applications are available in the marketplace that are appropriate for handling large datasets in a professional way. Two commercial products worth mentioning are 'Enterprise Miner' by SAS and 'SPSS Modeler' by IBM.

At first glance, these applications are easy to use. After a while, however, many users realize that more difficult questions require deeper statistical knowledge. Many people are interested in gaining such statistical skills and applying them, using one of the data mining tools offered by the industry.

This book will help users to become familiar with a wide range of statistical concepts or algorithms and apply them to concrete datasets. After a short statistical overview of how the procedures work and what assumptions to keep in mind, step-by-step procedures show how to find the solutions with the SPSS Modeler.

Features of This Book

- Easy to read
- Standardised chapter structure, including exercises and solutions
- All datasets are provided as downloads and explained in detail
- Template streams help the reader focus on the interesting parts of the stream and leave out recurring tasks
- Complete solution streams are ready to use
- Each example includes step-by-step explanations
- Short explanations of the most important statistical assumptions used when applying the algorithms are included
- Hundreds of screenshots are included, to ensure successful application of the algorithms to the datasets
- Exercises teach how to secure and systematise this knowledge
- Explanations and solutions are provided for all exercises
- Skills acquired through solving the exercises allow the user to create his/her own streams

The authors of the book, Tilo Wendler and Sören Gröttrup, want to thank all the people who supported the writing process. These include IBM support experts who dealt with some of the more difficult tasks, discovering more efficient ways to handle the data. Furthermore, the authors want to express gratitude to Jeni Ringland, Katrin Minor and Maria Sabottke for their outstanding efforts and their help in professionalising the text, layout, figures and tables.

Berlin, Germany — Tilo Wendler
Stuttgart, Germany — Sören Gröttrup

Contents

1 Introduction 1

1.1 The Concept of the SPSS Modeler 2

1.2 Structure and Features of This Book 5

1.2.1 Prerequisites for Using This Book 5

1.2.2 Structure of the Book and the Exercise/Solution Concept 6

1.2.3 Using the Data and Streams Provided with the Book 8

1.2.4 Datasets Provided with This Book 9

1.2.5 Template Concept of This Book 10

1.3 Introducing the Modeling Process 13

1.3.1 Exercises 16

1.3.2 Solutions 18

Literature 22

2 Basic Functions of the SPSS Modeler 25

2.1 Defining Streams and Scrolling Through a Dataset 25

2.2 Switching Between Different Streams 32

2.3 Defining or Modifying Value Labels 34

2.4 Adding Comments to a Stream 40

2.5 Exercises 43

2.6 Solutions 44

2.7 Data Handling and Sampling Methods 49

2.7.1 Theory 49

2.7.2 Calculations 50

2.7.3 String Functions 56

2.7.4 Extracting/Selecting Records 61

2.7.5 Filtering Data 65

2.7.6 Data Standardization: Z-Transformation 73

2.7.7 Partitioning Datasets 82

2.7.8 Sampling Methods 88

2.7.9 Merge Datasets 111

2.7.10 Append Datasets 124
2.7.11 Exercises 132
2.7.12 Solutions 147
Literature 184

3 Univariate Statistics 185
3.1 Theory 185
3.1.1 Discrete Versus Continuous Variables 185
3.1.2 Scales of Measurement 187
3.1.3 Exercises 188
3.1.4 Solutions 191
3.2 Simple Data Examination Tasks 194
3.2.1 Theory 194
3.2.2 Frequency Distribution of Discrete Variables 194
3.2.3 Frequency Distribution of Continuous Variables 199
3.2.4 Distribution Analysis with the Data Audit Node 202
3.2.5 Concept of “SuperNodes” and Transforming a Variable to Normality 207
3.2.6 Reclassifying Values 224
3.2.7 Binning Continuous Data 236
3.2.8 Exercises 248
3.2.9 Solutions 259
Literature 286

4 Multivariate Statistics 287
4.1 Theory 287
4.2 Scatterplot 290
4.3 Scatterplot Matrix 296
4.4 Correlation 302
4.5 Correlation Matrix 310
4.6 Exclusion of Spurious Correlations 314
4.7 Contingency Tables 315
4.8 Exercises 323
4.9 Solutions 325
Literature 345

5 Regression Models 347
5.1 Introduction to Regression Models 348
5.1.1 Motivating Examples 348
5.1.2 Concept of the Modeling Process and Cross-Validation 350
5.2 Simple Linear Regression 353
5.2.1 Theory 353
5.2.2 Building the Stream in SPSS Modeler 356
5.2.3 Identification and Interpretation of the Model Parameters 360
5.2.4 Assessment of the Goodness of Fit 362

5.2.5 Predicting Unknown Values 365
5.2.6 Exercises 367
5.2.7 Solutions 369
5.3 Multiple Linear Regression 390
5.3.1 Theory 390
5.3.2 Building the Model in SPSS Modeler 392
5.3.3 Final MLR Model and Its Goodness of Fit 397
5.3.4 Prediction of Unknown Values 404
5.3.5 Cross-Validation of the Model 404
5.3.6 Boosting and Bagging (for Regression Models) 406
5.3.7 Exercises 415
5.3.8 Solutions 418
5.4 Generalized Linear (Mixed) Model 448
5.4.1 Theory 448
5.4.2 Building a Model with the GLMM Node 450
5.4.3 The Model Nugget 455
5.4.4 Cross-Validation and Fitting a Quadric Regression Model 458
5.4.5 Exercises 468
5.4.6 Solutions 469
5.5 The Auto Numeric Node 488
5.5.1 Building a Stream with the Auto Numeric Node 490
5.5.2 The Auto Numeric Model Nugget 497
5.5.3 Exercises 500
5.5.4 Solutions 500
Literature 511

6 Factor Analysis 513
6.1 Motivating Example 513
6.2 General Theory of Factor Analysis 515
6.3 Principal Component Analysis 519
6.3.1 Theory 519
6.3.2 Building a Model in SPSS Modeler 520
6.3.3 Exercises 547
6.3.4 Solutions 550
6.4 Principal Factor Analysis 569
6.4.1 Theory 569
6.4.2 Building a Model 573
6.4.3 Exercises 579
6.4.4 Solutions 579
Literature 584

7 Cluster Analysis 587
7.1 Motivating Examples 587
7.2 General Theory of Cluster Analysis 589

7.2.1 Exercises 596
7.2.2 Solutions 598
7.3 TwoStep Hierarchical Agglomerative Clustering 601
7.3.1 Theory of Hierarchical Clustering 601
7.3.2 Characteristics of the TwoStep Algorithm 614
7.3.3 Building a Model in SPSS Modeler 615
7.3.4 Exercises 627
7.3.5 Solutions 629
7.4 K-Means Partitioning Clustering 640
7.4.1 Theory 640
7.4.2 Building a Model in SPSS Modeler 642
7.4.3 Exercises 659
7.4.4 Solutions 662
7.5 Auto Clustering 685
7.5.1 Motivation and Implementation of the Auto Cluster Node 685
7.5.2 Building a Model in SPSS Modeler 687
7.5.3 Exercises 699
7.5.4 Solutions 700
7.6 Summary 710
Literature 711

8 Classification Models 713
8.1 Motivating Examples 714
8.2 General Theory of Classification Models 716
8.2.1 Process of Training and Using a Classification Model 716
8.2.2 Classification Algorithms 718
8.2.3 Classification vs. Clustering 720
8.2.4 Making a Decision and the Decision Boundary 721
8.2.5 Performance Measures of Classification Models 723
8.2.6 The Analysis Node 725
8.2.7 Exercises 727
8.2.8 Solutions 730
8.3 Logistic Regression 733
8.3.1 Theory 734
8.3.2 Building the Model in SPSS Modeler 736
8.3.3 Optional: Model Types and Variable Interactions . . . 743
8.3.4 Final Model and Its Goodness of Fit 746
8.3.5 Classification of Unknown Values 750
8.3.6 Cross-Validation of the Model 751
8.3.7 Exercises 756
8.3.8 Solutions 758

8.4 Linear Discriminate Classification 776
8.4.1 Theory 776
8.4.2 Building the Model with SPSS Modeler 779
8.4.3 The Model Nugget and the Estimated Model Parameters 785
8.4.4 Exercises 788
8.4.5 Solutions 789
8.5 Support Vector Machine 808
8.5.1 Theory 809
8.5.2 Building the Model with SPSS Modeler 810
8.5.3 The Model Nugget 820
8.5.4 Exercises 821
8.5.5 Solutions 822
8.6 Neuronal Networks 843
8.6.1 Theory 844
8.6.2 Building a Network with SPSS Modeler 846
8.6.3 The Model Nugget 856
8.6.4 Exercises 860
8.6.5 Solutions 862
8.7 k-Nearest Neighbor 878
8.7.1 Theory 878
8.7.2 Building the Model with SPSS Modeler 882
8.7.3 The Model Nugget 891
8.7.4 Dimensional Reduction with PCA for Data Preprocessing 893
8.7.5 Exercises 901
8.7.6 Solutions 903
8.8 Decision Trees 917
8.8.1 Theory 917
8.8.2 Building a Decision Tree with the C5.0 Node 925
8.8.3 The Model Nugget 929
8.8.4 Building a decision tree with the CHAID node 932
8.8.5 Exercises 938
8.8.6 Solutions 939
8.9 The Auto Classifier Node 960
8.9.1 Building a Stream with the Auto Classifier Node . . . 961
8.9.2 The Auto Classifier Model Nugget 971
8.9.3 Exercises 973
8.9.4 Solutions 974
Literature 983

9 Using R with the Modeler 985
9.1 Advantages of R with the Modeler 985
9.2 Connecting with R 986
9.3 Test the SPSS Modeler Connection to R 990
9.4 Calculating New Variables in R 994

9.5 Model Building in R . . . 999
9.6 Exercises . . . 1008
9.7 Solutions . . . 1018
Literature . . . 1035

10 Appendix . . . 1037
10.1 Data Sets Used in This Book . . . 1037
10.1.1 adult_income_data.txt . . . 1037
10.1.2 beer.sav . . . 1037
10.1.3 benchmark.xlsx . . . 1037
10.1.4 car_simple.sav . . . 1039
10.1.5 car_sales_modified.sav . . . 1039
10.1.6 chess_endgame_data.txt . . . 1039
10.1.7 customer_bank_data.csv . . . 1040
10.1.8 diabetes_data_reduced.sav . . . 1040
10.1.9 DRUG1n.sav . . . 1041
10.1.10 EEG_Sleep_Signals.csv . . . 1042
10.1.11 employee_dataset_001 and employee_dataset_002 . . . 1042
10.1.12 England Payment Datasets . . . 1042
10.1.13 Features_eeg_signals.csv . . . 1044
10.1.14 gene_expression_leukemia.csv . . . 1044
10.1.15 gene_expression_leukemia_short.csv . . . 1045
10.1.16 gravity_constant_data.csv . . . 1045
10.1.17 Housing.data.txt . . . 1046
10.1.18 Iris.csv . . . 1046
10.1.19 IT-projects.txt . . . 1047
10.1.20 IT user satisfaction.sav . . . 1047
10.1.21 longley.csv . . . 1047
10.1.22 LPGA2009.csv . . . 1049
10.1.23 Mtcars.csv . . . 1050
10.1.24 nutrition_habites.sav . . . 1051
10.1.25 optdigits_training.txt, optdigits_test.txt . . . 1051
10.1.26 Orthodont.csv . . . 1052
10.1.27 Ozone.csv . . . 1052
10.1.28 pisa2012_math_q45.sav . . . 1052
10.1.29 sales_list.sav . . . 1054
10.1.30 ships.csv . . . 1054
10.1.31 test_scores.sav . . . 1054
10.1.32 Titanic.xlsx . . . 1055
10.1.33 tree_credit.sav . . . 1055
10.1.34 wine_data.txt . . . 1056
10.1.35 WisconsinBreastCancerData.csv . . . 1056
10.1.36 z_pm_customer1.sav . . . 1057
Literature . . . 1057