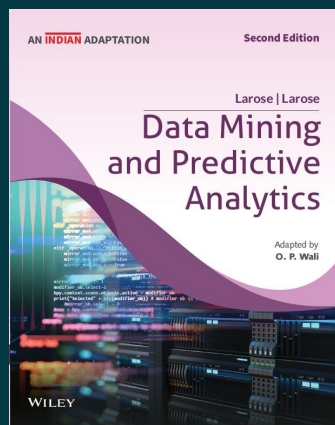




Data Mining and Predictive Analytics, 2ed (An Indian Adaptation)

By [Daniel T. Larose](#), [Chantal D. Larose](#), [O.P. Wali](#)

Paperback

9789354247255

₹1,219.00

Pages: 908

Description

Data Mining and Predictive Analytics serves as an introduction to data mining methods and models, including association rules, clustering, neural networks, logistic regression, and multivariate analysis. The authors apply a unified “white box” approach to data mining methods and models. This approach is designed to walk readers through the operations and nuances of the various methods, using small data sets, so readers can gain an insight into the inner workings of the method under review.

About the Author

Daniel T. Larose, Chantal D. Larose, O.P. Wali

Daniel T. Larose is Professor of Mathematical Sciences and Director of the Data Mining programs at Central Connecticut State University. He has published several books

Table of Contents

Part I: Data Preparation

Chapter 1: An Introduction to Data Mining and Predictive Analytics

1.1 What is Data Mining? What Is Predictive Analytics?

1.2 Wanted: Data Miners

1.3 The Need For Human Direction of Data Mining

1.4 The Cross-Industry Standard Process for Data Mining: CRISP-DM

1.5 Fallacies of Data Mining

1.6 What Tasks can Data Mining Accomplish

Chapter 2: Data Preprocessing

2.1 Why do We Need to Preprocess the Data?

2.2 Data Cleaning

2.3 Handling Missing Data

2.4 Identifying Misclassifications

2.5 Graphical Methods for Identifying Outliers

2.6 Measures of Center and Spread

2.7 Data Transformation

2.8 Min-Max Normalization

2.9 Z-Score Standardization

2.10 Decimal Scaling

2.11 Transformations to Achieve Normality

2.12 Numerical Methods for Identifying Outliers

2.13 Flag Variables

2.14 Transforming Categorical Variables into Numerical Variables

2.15 Binning Numerical Variables

2.16 Reclassifying Categorical Variables

2.17 Adding an Index Field

2.18 Removing Variables that are not Useful

2.19 Variables that Should Probably not be Removed

2.20 Removal of Duplicate Records

2.21 A Word About ID Fields

Chapter 3: Exploratory Data Analysis

3.1 Hypothesis Testing Versus Exploratory Data Analysis

3.2 Churn Example – Getting to Know the Data Set

3.3 IPL Example – Getting to Know the Data Set

Chapter 4: Dimension-Reduction Methods

4.1 Need for Dimension-Reduction in Data Mining

4.2 Principal Components Analysis

4.3 Applying PCA to the Houses Data Set

4.4 How Many Components Should We Extract?

4.5 Profiling the Principal Components

4.6 Communalities

4.7 Validation of the Principal Components

4.8 Factor Analysis

4.9 Applying Factor Analysis to the Adult Data Set

4.10 Factor Rotation

4.11 User-Defined Composites

4.12 An Example of a User-Defined Composite

Part II: Statistical Analysis

Chapter 5: Univariate Statistical Analysis

5.1 Data Mining Tasks in Discovering Knowledge in Data

5.2 Statistical Approaches to Estimation and Prediction

5.3 Statistical Inference

5.4 How Confident are We in Our Estimates?

5.5 Confidence Interval Estimation of the Mean

5.6 How to Reduce the Margin of Error

5.7 Confidence Interval Estimation of the Proportion

5.8 Hypothesis Testing for the Mean

5.9 Assessing The Strength of Evidence Against The Null Hypothesis

5.10 Using Confidence Intervals to Perform Hypothesis Tests

5.11 Hypothesis Testing for The Proportion

Chapter 6: Multivariate Statistics

6.1 Two-Sample t-Test for Difference in Means

6.2 Two-Sample Z-Test for Difference in Proportions

6.3 Test for the Homogeneity of Proportions

6.4 Chi-Square Test for Goodness of Fit of Multinomial Data

6.5 Analysis of Variance

Chapter 7: Preparing to Model the Data

7.1 Supervised Versus Unsupervised Methods

7.2 Statistical Methodology and Data Mining Methodology

7.3 Cross-Validation

7.4 Overfitting

7.5 Bias-Variance Trade-Off

7.6 Balancing The Training Data Set

7.7 Establishing Baseline Performance

Chapter 8: Simple Linear Regression

8.1 An Example of Simple Linear Regression

8.2 Dangers of Extrapolation

8.3 How Useful is the Regression? The Coefficient of Determination

8.4 Standard Error of the Estimate

8.5 Correlation Coefficient

8.6 ANOVA Table for Simple Linear Regression

8.7 Outliers, High Leverage Points, and Influential Observations

8.8 Population Regression Equation

8.9 Verifying The Regression Assumptions

8.10 Inference in Regression

8.11 t-Test for the Relationship Between x and y

8.12 Confidence Interval for the Slope of the Regression Line

8.13 Confidence Interval for the Correlation Coefficient ρ

8.14 Confidence Interval for the Mean Value of Given

8.15 Prediction Interval for a Randomly Chosen Value of Given

8.16 Transformations to Achieve Linearity

8.17 Box-Cox Transformations

Chapter 9: Multiple Regression and Model Building

- 9.1 An Example of Multiple Regression
- 9.2 The Population Multiple Regression Equation
- 9.3 Inference in Multiple Regression
- 9.4 Regression With Categorical Predictors, Using Indicator Variables
- 9.5 Adjusting R²: Penalizing Models For Including Predictors That Are Not Useful
- 9.6 Sequential Sums of Squares
- 9.7 Multicollinearity
- 9.8 Variable Selection Methods
- 9.9 Gas Mileage Data Set
- 9.10 An Application of Variable Selection Methods
- 9.11 Using the Principal Components as Predictors in Multiple Regression

Part III: Classification

Chapter 10: k-Nearest Neighbor Algorithm

- 10.1 Classification Task
- 10.2 k-Nearest Neighbor Algorithm
- 10.3 Distance Function
- 10.4 Combination Function
- 10.5 Quantifying Attribute Relevance: Stretching the Axes
- 10.6 Database Considerations
- 10.7 k-Nearest Neighbor Algorithm for Estimation and Prediction
- 10.8 Choosing k
- 10.9 Application of k-Nearest Neighbor Algorithm Using IBM/SPSS Modeler

Chapter 11: Decision Trees

- 11.1 What is a Decision Tree?
- 11.2 Requirements for Using Decision Trees
- 11.3 Classification and Regression Trees
- 11.4 C4.5 Algorithm
- 11.5 Decision Rules
- 11.6 Comparison of the C5.0 and CART Algorithms Applied to Real Data

Chapter 12: Neural Networks

- 12.1 Input and Output Encoding
- 12.2 Neural Networks for Estimation and Prediction
- 12.3 Simple Example of a Neural Network
- 12.4 Sigmoid Activation Function
- 12.5 Back-Propagation
- 12.6 Gradient-Descent Method
- 12.7 Back-Propagation Rules
- 12.8 Example of Back-Propagation

12.9 Termination Criteria

12.10 Learning Rate

12.11 Momentum Term

12.12 Sensitivity Analysis

12.13 Application of Neural Network Modeling

Chapter 13: Logistic Regression

13.1 Simple Example of Logistic Regression

13.2 Maximum Likelihood Estimation

13.3 Interpreting Logistic Regression Output

13.4 Inference: Are the Predictors Significant?

13.5 Odds Ratio and Relative Risk

13.6 Interpreting Logistic Regression for a Dichotomous Predictor

13.7 Interpreting Logistic Regression for a Polychotomous Predictor

13.8 Interpreting Logistic Regression for a Continuous Predictor

13.9 Assumption of Linearity

13.10 Zero-Cell Problem

13.11 Multiple Logistic Regression

13.12 Introducing Higher Order Terms to Handle Nonlinearity

13.13 Validating the Logistic Regression Model

13.14 WEKA: Hands-On Analysis Using Logistic Regression

Chapter 14: Naïve Bayes and Bayesian Networks

14.1 Bayesian Approach

14.2 Maximum A Posteriori (MAP) Classification

14.3 Posterior Odds Ratio

14.4 Balancing The Data

14.5 Naïve Bayes Classification

14.6 Interpreting The Log Posterior Odds Ratio

14.7 Zero-Cell Problem

14.8 Numeric Predictors for Naïve Bayes Classification

14.9 WEKA: Hands-on Analysis Using Naïve Bayes

14.10 Bayesian Belief Networks

14.11 Clothing Purchase Example

14.12 Using The Bayesian Network to Find Probabilities

Chapter 15: Model Evaluation Techniques

15.1 Model Evaluation Techniques for the Description Task

15.2 Model Evaluation Techniques for the Estimation and Prediction Tasks

15.3 Model Evaluation Measures for the Classification Task

15.4 Accuracy and Overall Error Rate

15.5 Sensitivity and Specificity

15.6 False-Positive Rate and False-Negative Rate

15.7 Proportions of True Positives, True Negatives, False Positives, and False Negatives

15.8 Misclassification Cost Adjustment to Reflect Real-World Concerns

15.9 Decision Cost/Benefit Analysis

15.10 Lift Charts and Gains Charts

15.11 Interweaving Model Evaluation with Model Building

15.12 Confluence of Results: Applying a Suite of Models

Chapter 16: Cost-Benefit Analysis Using Data-Driven Costs

16.1 Decision Invariance Under Row Adjustment

16.2 Positive Classification Criterion

16.3 Demonstration Of The Positive Classification Criterion

16.4 Constructing The Cost Matrix

16.5 Decision Invariance Under Scaling

16.6 Direct Costs and Opportunity Costs

16.7 Case Study: Cost-Benefit Analysis Using Data-Driven Misclassification Costs

16.8 Rebalancing as a Surrogate for Misclassification Costs

Chapter 17: Cost-Benefit Analysis for Trinary and -Nary Classification Models

17.1 Classification Evaluation Measures for a Generic Trinary Target

17.2 Application of Evaluation Measures for Trinary Classification to the Loan Approval Problem

17.3 Data-Driven Cost-Benefit Analysis for Trinary Loan Classification Problem

17.4 Comparing Cart Models With and Without Data-Driven Misclassification Costs

17.5 Classification Evaluation Measures for a Generic k-Nary Target

17.6 Example of Evaluation Measures and Data-Driven Misclassification Costs for k-Nary Classification

Chapter 18: Graphical Evaluation of Classification Models

18.1 Review of Lift Charts and Gains Charts

18.2 Lift Charts and Gains Charts Using Misclassification Costs

18.3 Response Charts

18.4 Profits Charts

18.5 Return on Investment (ROI) Charts

Part IV: Clustering

Chapter 19: Hierarchical and k-Means Clustering

19.1 The Clustering Task

19.2 Hierarchical Clustering Methods

19.3 Single-Linkage Clustering

19.4 Complete-Linkage Clustering

19.5 k-Means Clustering

19.6 Example of k-Means Clustering at Work

19.7 Behavior of MSB, MSE, and Pseudo-F as the k-Means Algorithm Proceeds

19.8 Application of k-Means Clustering Using SAS Enterprise Miner

19.9 Using Cluster Membership to Predict Churn

Chapter 20: Kohonen Networks

20.1 Self-Organizing Maps

20.2 Kohonen Networks

20.3 Example of a Kohonen Network Study

20.4 Cluster Validity

20.5 Application of Clustering Using Kohonen Networks

20.6 Interpreting The Clusters

20.7 Using Cluster Membership as Input to Downstream Data Mining Models

Chapter 21: BIRCH Clustering

21.1 Rationale for BIRCH Clustering

21.2 Cluster Features

21.3 Cluster Feature TREE

21.4 Phase 1: Building The CF Tree

21.5 Phase 2: Clustering The Sub-Clusters

21.6 Example of Birch Clustering, Phase 1: Building The CF Tree

21.7 Example of BIRCH Clustering, Phase 2: Clustering The Sub-Clusters

21.8 Evaluating The Candidate Cluster Solutions

21.9 Case Study: Applying BIRCH Clustering to The Bank Loans Data Set

Chapter 22: Measuring Cluster Goodness

22.1 Rationale for Measuring Cluster Goodness

22.2 The Silhouette Method

22.3 Silhouette Example

22.4 Silhouette Analysis of the IRIS Data Set

22.5 The Pseudo-F Statistic

22.6 Example of the Pseudo-F Statistic

22.7 Pseudo-F Statistic Applied to the IRIS Data Set

22.8 Cluster Validation

22.9 Cluster Validation Applied to the Loans Data Set

Part V: Association Rules

Chapter 23: Association Rules

23.1 Affinity Analysis and Market Basket Analysis

23.2 Support, Confidence, Frequent Itemsets, and the A Priori Property

23.3 How Does The A Priori Algorithm Work (Part 1)? Generating Frequent Itemsets

23.4 How Does The A Priori Algorithm Work (Part 2)? Generating Association Rules

23.5 Extension From Flag Data to General Categorical Data

- 23.6 Information-Theoretic Approach: Generalized Rule Induction Method
- 23.7 Association Rules are Easy to do Badly
- 23.8 How Can We Measure the Usefulness of Association Rules?
- 23.9 Do Association Rules Represent Supervised or Unsupervised Learning?
- 23.10 Local Patterns Versus Global Models

Part VI: Enhancing Model Performance

Chapter 24: Segmentation Models

- 24.1 The Segmentation Modeling Process
- 24.2 Segmentation Modeling Using EDA to Identify the Segments
- 24.3 Segmentation Modeling using Clustering to Identify the Segments

Chapter 25: Ensemble Methods: Bagging and Boosting

- 25.1 Rationale for Using an Ensemble of Classification Models
- 25.2 Bias, Variance, and Noise
- 25.3 When to Apply, and not to apply, Bagging
- 25.4 Bagging
- 25.5 Boosting
- 25.6 Application of Bagging and Boosting Using IBM/SPSS Modeler

Chapter 26: Model Voting and Propensity Averaging

- 26.1 Simple Model Voting
- 26.2 Alternative Voting Methods
- 26.3 Model Voting Process
- 26.4 An Application of Model Voting
- 26.5 What is Propensity Averaging?
- 26.6 Propensity Averaging Process
- 26.7 An Application of Propensity Averaging

Part VII: Further Topics

Chapter 27: Genetic Algorithms

- 27.1 Introduction To Genetic Algorithms
- 27.2 Basic Framework of a Genetic Algorithm
- 27.3 Simple Example of a Genetic Algorithm at Work
- 27.4 Modifications and Enhancements: Selection
- 27.5 Modifications and Enhancements: Crossover
- 27.6 Genetic Algorithms for Real-Valued Variables
- 27.7 Using Genetic Algorithms to Train a Neural Network
- 27.8 WEKA: Hands-On Analysis Using Genetic Algorithms

Chapter 28: Imputation of Missing Data

- 28.1 Need for Imputation of Missing Data
- 28.2 Imputation of Missing Data: Continuous Variables

28.3 Standard Error of the Imputation

28.4 Imputation of Missing Data: Categorical Variables

28.5 Handling Patterns in Missingness

Part VIII: Case Study: Predicting Response to Direct-Mail Marketing

Chapter 29: Case Study, Part 1: Business Understanding, Data Preparation, and EDA

29.1 Cross-Industry Standard Practice for Data Mining

29.2 Business Understanding Phase

29.3 Data Understanding Phase, Part 1: Getting a Feel for the Data Set

29.4 Data Preparation Phase

29.5 Data Understanding Phase, Part 2: Exploratory Data Analysis

Chapter 30: Case Study, Part 2: Clustering and Principal Components Analysis

30.1 Partitioning the Data

30.2 Developing the Principal Components

30.3 Validating the Principal Components

30.4 Profiling the Principal Components

30.5 Choosing the Optimal Number of Clusters Using Birch Clustering

30.6 Choosing the Optimal Number of Clusters Using k-Means Clustering

30.7 Application of k-Means Clustering

30.8 Validating the Clusters

30.9 Profiling the Clusters

Chapter 31: Case Study, Part 3: Modeling And Evaluation For Performance And Interpretability

31.1 Do You Prefer The Best Model Performance, Or A Combination Of Performance And Interpretability?

31.2 Modeling And Evaluation Overview

31.3 Cost-Benefit Analysis Using Data-Driven Costs

31.4 Variables to be Input To The Models

31.5 Establishing The Baseline Model Performance

31.6 Models That Use Misclassification Costs

31.7 Models That Need Rebalancing as a Surrogate for Misclassification Costs

31.8 Combining Models Using Voting and Propensity Averaging

31.9 Interpreting The Most Profitable Model

Chapter 32: Case Study, Part 4: Modeling and Evaluation for High Performance Only

32.1 Variables to be Input to the Models

32.2 Models that use Misclassification Costs

32.3 Models that Need Rebalancing as a Surrogate for Misclassification Costs

32.4 Combining Models using Voting and Propensity Averaging

32.5 Lessons Learned

32.6 Conclusions

Appendix A: Data Summarization and Visualization

Part 1: Summarization 1: Building Blocks of Data Analysis

Part 2: Visualization: Graphs and Tables for Summarizing and Organizing Data

Part 3: Summarization 2: Measures of Center, Variability, and Position

Part 4: Summarization and Visualization of Bivariate Relationships

Index

To purchase this product, please visit:

<https://www.wileyindia.com/data-mining-and-predictive-analytics-2ed-an-indian-adaptation.html>



Scan to buy