

CONTENTS

Preface to Second Edition(xv)

Chapter 1: Data Science: Key Concepts

Five Disruptive Products..... 1
The Japanese App..... 1
The Healthcare App 2
Coursera..... 3
Sensory Device in Agriculture Sector 3
Autonomous Car..... 4
Data Science Vs Traditional Methods..... 5
Difference in Architecture..... 5
Demystifying Machine Learning 6
Technique 1: Segmentation 6
Technique 2: Unstructured Text Mining 10
Technique 3: Scoring for Signal Processing..... 11
Technique 4: Forecasting..... 11
Technique 5: Recommenders..... 11
Summary..... 16

Chapter 2: Data Wrangling

Data Collection and Data Types 17
Data Treatment 19
Data Transformation..... 25
Summary..... 30

Chapter 3: Spotting Signals: An Overview

Signals Across Verticals..... 32

(vi) CONTENTS

Analyzing and Application of a Signal Pattern 32
Signals: A Few Key Concepts 34
Signal Extraction Methodology – Simplistic View 35
Simplistic Nine Step Process 36
Summary..... 38

Chapter 4.1: Introduction to R

Analytical Tool – Tool R 39

Chapter 4.2: Business Storytelling Using R

10 Basic Commands in R..... 56
Summary..... 58

Chapter 5.1: Problem based Analysis

Problem based Analysis 59
Univariate Analysis Across Sectors 61
Banking Industry 61
Measures of Central Tendency..... 62
Measure of Dispersion 65
Scope of Data Analyzed – EDA (Fleet Industry)..... 68
Summary of Key Data Quality Related Observations 69
Univariate Best Practice 74
Summary..... 76

Chapter 5.2: Model

Importance of Model 77
An Example of a Model 79
Three Common Things in a Predictive Model and Caricature 79
Summary..... 80

Chapter 6.1: Bivariate Analysis

Correlation Analysis 81
Types of Correlation..... 82
Summary..... 89

Chapter 6.2: Cross Tabs

<i>Cross Tab Analysis – Core Philosophy</i>	92
<i>Correlation across Industries</i>	92
<i>Retail Industry</i>	92
<i>Telecom Industry</i>	93
<i>Banking Industry</i>	94
<i>Crosstab across Industries</i>	94
<i>Retail Industry</i>	94
<i>Telecom Industry</i>	94
<i>Banking Industry</i>	94
<i>Correlation and Crosstabs in R</i>	95
<i>Summary</i>	98

Chapter 7: Correlation Matrix

<i>Full Correlation Matrix</i>	99
<i>Correlation Analysis Methods</i>	102
<i>Stability Checks</i>	102
<i>Partial Correlation</i>	103
<i>Conditional Correlation</i>	104
<i>GARCH Process</i>	104
<i>Summary</i>	105

Chapter 8.1: Visualization and Visual Constructs

<i>Visual Constructs</i>	107
<i>Detecting Patterns using Visual Constructs</i>	107
<i>Demistifying Advanced Visualization</i>	108
<i>Box Plot</i>	108
<i>Scatter Plot</i>	109
<i>Runcharts</i>	109
<i>Interpret a Run Chart</i>	110
<i>Pareto Charts</i>	112
<i>Geospatial Map</i>	116
<i>Heat Maps</i>	116
<i>Spider Chart</i>	118

Chapter 8.2: Advance Visualization

<i>Core Concepts in Advanced Visualization</i>	119
<i>Visualization Consumers</i>	119
<i>Creating Dashboards</i>	120
<i>Connecting the Dots to Create a Dashboard</i>	120
<i>Best Practices in Designing Dashboards and Scoreboards</i>	121
<i>Domestic Loan Analysis Example</i>	122
<i>Visualization Commands in R</i>	123
<i>Summary</i>	126

Chapter 9.1: Machine Learning in Action

<i>Logistic Regression</i>	127
<i>Evaluating the Model</i>	128
<i>Confusion Matrix</i>	128
<i>Lift Charts</i>	129
<i>ROC Curve</i>	129
<i>AIC</i>	130
<i>Null Deviance and Residual Deviance</i>	130
<i>Logistic Regression in R</i>	130
<i>Create a Baseline Model</i>	130
<i>Split Train and Test Data</i>	130
<i>Logistic Regression Model</i>	131
<i>Confusion Matrix</i>	133
<i>ROC Curve</i>	134
<i>Prediction on Test Set</i>	135
<i>Summary</i>	135

Chapter 9.2: Decision Trees

<i>Terminology in Decision Trees</i>	137
<i>Decision Tree Algorithm</i>	138
<i>Gini Index</i>	138
<i>Chi – Square</i>	139
<i>Steps to Calculate</i>	139

<i>Information Gain</i>	140
<i>Steps to Calculate</i>	141
<i>Reduction in Variance</i>	141
<i>Summary</i>	142

Chapter 9.3:Support Vector Machines

<i>SVM-Standardization</i>	146
<i>SVM in R</i>	146
<i>Tuning Parameters for Support Vector Machine Algorithm</i>	149
<i>Regularization</i>	149
<i>Gamma</i>	149
<i>Margin</i>	150
<i>Summary</i>	150

Chapter 9.4:Naive Bayes

<i>Feature Engineering</i>	151
<i>Calculating Probabilities</i>	152
<i>Naïve Bayes Programming in R</i>	155
<i>Validation Observations</i>	157
<i>Summary</i>	158

Chapter 9.5:Linear Regression

<i>Requirements for Linear Regression</i>	159
<i>The Least Square Regression Line</i>	159
<i>Properties of Regression Line</i>	160
<i>The Coefficient of Determination</i>	160
<i>The p Value and T – Value</i>	160
<i>Linear Regression in R</i>	161
<i>Correlation</i>	163
<i>Build a Linear Model</i>	163
<i>Predicting a Linear Model</i>	166
<i>Summary</i>	167

Chapter 9.6:Regression

<i>5 Powerful Unanswered Questions by Regression – Unknown Unknowns</i>	170
<i>Regression Across Sectors</i>	171
<i>Scenario 1 Cost of Insurance</i>	171
<i>Scenario 2 Model Building for Property Design</i>	172
<i>Scenario 3 Estimating Patient Stay at Hospital</i>	172
<i>Scenario 4 Estimate Defect Density</i>	173
<i>Population and Sample Regression Models</i>	173
<i>Deterministic</i>	174
<i>Probabilistic</i>	174
<i>Regression Assumptions</i>	174
<i>Specify the Deterministic Component</i>	175
<i>R –Squared</i>	175
<i>P value</i>	175
<i>Regression in R</i>	176
<i>Some Key Points</i>	177
<i>Summary</i>	178

Chapter 9.7:A/B Testing

<i>Collaborative Filtering</i>	181
<i>Application for Collaborative Filtering</i>	182
<i>Fixed Size Neighborhood</i>	182
<i>Threshold Based Neighborhood</i>	183
<i>Graph</i>	186
<i>Page Rank Algorithm</i>	187
<i>Summary</i>	187

Chapter 9.8:Classification

<i>Clustering</i>	188
<i>Hierarchical Clustering</i>	189
<i>Linkage Criteria</i>	190

<i>K-Means Clustering</i>	193
<i>K means Clustering in R</i>	194
<i>Summary</i>	197

Chapter 9.9: Introduction to Gradient Boosting

<i>Summary</i>	201
----------------------	-----

Chapter 10.1: Sample Preparation

<i>Population and Sample</i>	203
<i>Population and Simple Random Sample</i>	203
<i>Sample with Replacement</i>	203
<i>Sample without Replacement</i>	204
<i>Combination</i>	204
<i>Permutation</i>	204
<i>Stratified Random Sampling</i>	204
<i>Types of Stratified Random Sampling</i>	206
<i>Multistage Sampling</i>	207
<i>Systematic Sampling</i>	208
<i>Summary</i>	209

Chapter 10.2: Data Train and Test Data

<i>Over Fitting</i>	210
<i>Underfitting</i>	211
<i>Cross Validation</i>	211
<i>K-Folds Cross Validation</i>	211
<i>Leave One Out Cross Validation (LOOCV)</i>	212
<i>Coding in R using k-fold Cross Validation</i>	213
<i>Model Performance Metrics</i>	214
<i>K-fold Cross-Validation</i>	214
<i>Summary</i>	215

Chapter 11.1: Multivariate Analysis Topics

<i>Dimensionality Reduction Technique</i>	217
<i>Feature Engineering</i>	217

(xii) CONTENTS

Feature Engineering – Key Point..... 219
Feature Selection – Definition 219
Goals of Feature Extraction..... 219
Correlation and Variables..... 220
Ranking Criteria - Correlation..... 221
Feature Subset Selection 223
Model a Retail Shopper: A Problem Statement..... 225
Summary..... 226

Chapter 11.2: Principal Component Analysis

PCA in R..... 230
Summary..... 232

Chapter 11.3: Factor Analysis

Variance in Factor Analysis..... 233
FA – The Process 234
Factor Analysis in R..... 235
Summary..... 237

Chapter 11.4: ANOVA

Interpret ANOVA Test 238
Significance Testing 239
Summary of ANOVA..... 239
ANOVA Models 239
ANOVA in R 240
MANOVA..... 242
MANOVA in R..... 244
Summary..... 246

Chapter 12.1: Additional Topics in Analytics

Survival Analysis..... 247
Life Table Analysis 247
Censoring..... 249
Survival Function 249

Hazard Curve 250
Compute the Cox Model in R 250
Compute the Cox Model..... 251

Chapter 12.2: Exploratory Data Analysis Case Study – Business Perspective

Exploratory Data Analysis 255
Scenario 1: Survival Analysis..... 255
Scenario 2: Attrition Analysis 257
Ratio Analysis of New Customers to Existing Customers 257
Scenario 3: Difference between Active and Inactive customers:
Valuable Vulnerable..... 260
Scenario 4: Days to Repeat Purchases 262
Scenario 5: Sales Trends Seasonality in Sales/Identifying Patterns 264
Scenario 6: Segmenting Watch Company’s Customers Region Wise 266
Scenario 7: Customer Lifetime Value 268
Summary..... 270

Chapter 13: Text Mining

Text Analytics Understanding Taking an Example 273
Some of the Steps that are followed for Text Mining in R..... 274
Steps to Create Word Cloud in R 275
Summary..... 280

Case Studies281

References307

Index309