

Applied Analytics through Case Studies Using SAS and R

**Implementing Predictive Models
and Machine Learning Techniques**

Deepti Gupta

Apress®

Applied Analytics through Case Studies Using SAS and R

Deepti Gupta
Boston, Massachusetts, USA

ISBN-13 (pbk): 978-1-4842-3524-9
<https://doi.org/10.1007/978-1-4842-3525-6>

ISBN-13 (electronic): 978-1-4842-3525-6

Library of Congress Control Number: 2018952360

Copyright © 2018 by Deepti Gupta

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.

Trademarked names, logos, and images may appear in this book. Rather than use a trademark symbol with every occurrence of a trademarked name, logo, or image we use the names, logos, and images only in an editorial fashion and to the benefit of the trademark owner, with no intention of infringement of the trademark.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Managing Director, Apress Media LLC: Welmoed Spahr
Acquisitions Editor: Celestin John
Development Editor: James Markham
Coordinating Editor: Divya Modi

Cover designed by eStudioCalamar

Cover image designed by Freepik (www.freepik.com)

Distributed to the book trade worldwide by Springer Science+Business Media New York, 233 Spring Street, 6th Floor, New York, NY 10013. Phone 1-800-SPRINGER, fax (201) 348-4505, e-mail orders-ny@springer-sbm.com, or visit www.springeronline.com. Apress Media, LLC is a California LLC and the sole member (owner) is Springer Science + Business Media Finance Inc (SSBM Finance Inc). SSBM Finance Inc is a **Delaware** corporation.

For information on translations, please e-mail rights@apress.com, or visit <http://www.apress.com/rights-permissions>.

Apress titles may be purchased in bulk for academic, corporate, or promotional use. eBook versions and licenses are also available for most titles. For more information, reference our Print and eBook Bulk Sales web page at <http://www.apress.com/bulk-sales>.

Any source code or other supplementary material referenced by the author in this book is available to readers on GitHub via the book's product page, located at www.apress.com/978-1-4842-3524-9. For more detailed information, please visit <http://www.apress.com/source-code>.

Printed on acid-free paper

I am dedicating this book to my Family.

Table of Contents

About the Author	xi
About the Contributor	xiii
About the Technical Reviewer	xv
Acknowledgments	xvii
Introduction	xix
Chapter 1: Data Analytics and Its Application in Various Industries.....	1
What Is Data Analytics?	2
Data Collection	3
Data Preparation.....	4
Data Analysis	4
Model Building.....	5
Results.....	5
Put into Use	5
Types of Analytics	6
Understanding Data and Its Types.....	7
What Is Big Data Analytics?	8
Big Data Analytics Challenges	10
Data Analytics and Big Data Tools	11
Role of Analytics in Various Industries.....	14
Who Are Analytical Competitors?	18
Key Models and Their Applications in Various Industries.....	18
Summary.....	21
References.....	21

TABLE OF CONTENTS

- Chapter 2: Banking Case Study 27**
- Applications of Analytics in the Banking Sector..... 28
- Increasing Revenue by Cross-Selling and Up-Selling 29
- Minimizing Customer Churn 30
- Increase in Customer Acquisition 30
- Predicting Bank-Loan Default..... 31
- Predicting Fraudulent Activity..... 32
- Case Study: Predicting Bank-Loan Defaults with Logistic Regression Model..... 34
- Logistic Regression Equation 35
- Odds 36
- Logistic Regression Curve 37
- Logistic Regression Assumptions..... 38
- Logistic Regression Model Fitting and Evaluation 39
- Statistical Test for Individual Independent Variable in Logistic 40
- Regression Model..... 40
- Predictive Value Validation in Logistic Regression Model..... 41
- Logistic Regression Model Using R..... 46
- About Data..... 47
- Performing Data Exploration 47
- Model Building and Interpretation of Full Data..... 52
- Model Building and Interpretation of Training and Testing Data..... 56
- Predictive Value Validation..... 61
- Logistic Regression Model Using SAS 65
- Model Building and Interpretation of Full Data..... 74
- Summary..... 92
- References..... 92
- Chapter 3: Retail Case Study 97**
- Supply Chain in the Retail Industry 98
- Types of Retail Stores 99

Role of Analytics in the Retail Sector	100
Customer Engagement	100
Supply Chain Optimization	101
Price Optimization	103
Space Optimization and Assortment Planning.....	103
Case Study: Sales Forecasting for Gen Retailers with SARIMA Model.....	105
Overview of ARIMA Model	107
Three Steps of ARIMA Modeling.....	111
Identification Stage	111
Estimation and Diagnostic Checking Stage.....	113
Forecasting Stage.....	114
Seasonal ARIMA Models or SARIMA.....	115
Evaluating Predictive Accuracy of Time Series Model	117
Seasonal ARIMA Model Using R	118
About Data	119
Performing Data Exploration for Time Series Data	119
Seasonal ARIMA Model Using SAS.....	133
Summary.....	158
References.....	159
Chapter 4: Telecommunication Case Study	161
Types of Telecommunications Networks.....	162
Role of Analytics in the Telecommunications Industry.....	163
Predicting Customer Churn	163
Network Analysis and Optimization.....	165
Fraud Detection and Prevention	166
Price Optimization	166
Case Study: Predicting Customer Churn with Decision Tree Model	168
Advantages and Limitations of the Decision Tree.....	169
Handling Missing Values in the Decision Tree	170

TABLE OF CONTENTS

- Handling Model Overfitting in Decision Tree..... 170
- How the Decision Tree Works 171
- Measures of Choosing the Best Split Criteria in Decision Tree..... 172
- Decision Tree Model Using R..... 179
 - About Data 179
 - Performing Data Exploration 180
 - Splitting Data Set into Training and Testing 183
 - Model Building & Interpretation on Training and Testing Data..... 184
- Decision Tree Model Using SAS 193
 - Model Building and Interpretation of Full Data..... 200
 - Model Building and Interpretation on Training and Testing Data 208
- Summary..... 217
- References..... 217
- Chapter 5: Healthcare Case Study 221**
 - Application of Analytics in the Healthcare Industry 224
 - Predicting the Outbreak of Disease and Preventative Management 225
 - Predicting the Readmission Rate of the Patients 225
 - Healthcare Fraud Detection..... 227
 - Improve Patient Outcomes & Lower Costs 228
 - Case Study: Predicting Probability of Malignant and Benign Breast Cancer with Random Forest Model..... 230
 - Working of Random Forest Algorithm..... 230
 - Random Forests Model Using R 238
 - Random Forests Model Using SAS 249
 - Summary..... 271
 - References..... 271
- Chapter 6: Airline Case Study 277**
 - Application of Analytics in the Airline Industry..... 280
 - Personalized Offers and Passenger Experience 281
 - Safer Flights 282

Airline Fraud Detection	283
Predicting Flight Delays.....	284
Case Study: Predicting Flight Delays with Multiple Linear Regression Model	286
Multiple Linear Regression Equation.....	287
Multiple Linear Regression Assumptions and Checking for Violation of Model Assumptions	287
Variables Selection in Multiple Linear Regression Model.....	290
Evaluating the Multiple Linear Regression Model	290
Multiple Linear Regression Model Using R	292
About Data	293
Performing Data Exploration	293
Model Building & Interpretation on Training and Testing Data.....	299
Multiple Linear Regression Model Using SAS	311
Summary.....	340
References.....	340
Chapter 7: FMCG Case Study	345
Application of Analytics in FMCG Industry	346
Customer Experience & Engagement.....	347
Sales and Marketing.....	347
Logistics Management	348
Markdown Optimization	349
Case Study: Customer Segmentation with RFM Model and K-means Clustering	350
Overview of RFM Model	351
Overview of K-means Clustering.....	355
RFM Model & K-means Clustering Using R.....	358
About Data	358
Performing Data Exploration	359
RFM Model & K-means Clustering Using SAS.....	376
Summary.....	393
References.....	394
Index.....	397

About the Author



Deepti Gupta completed her MBA in Finance & PGPM in Operation Research in 2010. She has worked with KPMG and IBM private limited as a Data Scientist and is currently working as a data science freelancer. Deepti has extensive experience in predictive modeling and machine learning and her expertise is in SAS and R. Deepti has developed data science courses and delivered data science trainings and conducted workshops in both corporate and academic institutions. She has written multiple blogs and white papers. Deepti has a passion for mentoring budding data scientists.

About the Contributor



Dr. Akshat Gupta is currently working as a Senior Applications Engineer at MilliporeSigma in Applications Engineering, Global Manufacturing Sciences and Technology (MSAT) group. He authored the health-care case study (Chapter5) of this book. His focal area of research is cell culture clarification and tangential flow filtration. Dr. Gupta has extensive experience in Design of Experiments (DOE) and statistical analysis. He holds a Bachelor of Technology (B.Tech) degree in Chemical

Engineering from the Vellore Institute of Technology, and a Master of Science (MS) and Doctor of Philosophy (Ph.D.) in Chemical Engineering from the University of Massachusetts Lowell. He also has graduate certificates in Modeling and Simulation, and Nanotechnology.

About the Technical Reviewer



Preeti Pandhu has a Master of Science degree in Applied (Industrial) Statistics from the University of Pune. She is SAS certified as a base and advanced programmer for SAS 9 as well as a predictive modeler using SAS Enterprise Miner 7. Preeti has more than 18 years of experience in analytics and training.

She started her career as a lecturer in statistics and began her journey into the corporate world with IDEaS (now a SAS company), where she managed a team of business analysts in the optimization and forecasting domain. She joined SAS as a corporate trainer before stepping back into the analytics domain to contribute to a solution-testing team and research/consulting team. She was with SAS for 9 years. Preeti is currently passionately building her analytics training firm, DataScienceLab (www.datasciencelab.in).

Acknowledgments

Book writing is one of the most interesting and challenging attempt one can take up. This book could not have been completed without the encouragement, guidance, and support of my family. I would like to thank Dr. Akshat Gupta, Ved Prakash Garg, Col. Atul Gupta, Dr. Anvita Garg, Ayush Gupta, RS Miyan, Ansi Miyan, Dr. James Chrostowski, and my colleagues and friends for their productive discussions and suggestions. My special thanks to Celestin John who provided great help on everything ranging from technical support to answering my queries. I appreciate the thoughtful and insightful comments from the editor and the reviewers. Thanks to the Apress team, especially to Divya Modi for all the patience, support, and guidance in completing this project.

Introduction

Analytics is a big buzz and a need for today's industries to solve their business problems. Analytics helps in mining the structured and unstructured data in order to withdraw the effective insights from the data, which will help to make effective business decisions. SAS and R are highly used tools in analytics across the globe by all industries for data mining and building machine learning and predictive models. This book focuses on industrial business problems and a practical analytical approach to solve those problems by implementing predictive models and machine learning techniques using SAS and R analytical languages.

The primary objective of this book is to help statisticians, developers, engineers, and data analysts who are well versed in writing codes; have a basic understanding of data and statistics; and are planning to transition to a data scientist profile. The most challenging part is practical and hands-on knowledge of building predictive models and machine learning algorithms and deploying them in industries to address industrial business problems. This book will benefit the reader in solving the business problems in various industrial domains by sharpening their analytical skills in getting practical exposure to various predictive model and machine learning algorithms in six industrial domains.

What's in This Book

This book focuses on industrial business problems and practical analytical approaches to solve those problems by implementing predictive models and machine learning techniques using SAS Studio and R analytical languages. **This book contains six industrial case studies of various domains with data and all the codes in SAS Studio and R languages, which would benefit all readers to practice and implement these models in their own business cases.**

In Chapter 1 the general outline about analytics, the role of analytics in various industries, and a few popular data science and analytical tools are discussed. Chapter 2 describes the role of analytics in the banking industry with a detailed explanation of predicting a bank loan default case study in R and SAS. Chapter 3

INTRODUCTION

describes how analytics contribute in the retail industry and offers a detailed explanation of forecasting a case study in R and SAS. Chapter 4 describes how analytics is reshaping the telecommunications industry and gives a detailed explanation of a case study on predicting customer churn in R and SAS. Chapter 5 describes the application of analytics in the healthcare industry and gives a clear explanation of a case study on predicting the probability of benign and malignant breast cancer using R and SAS. Chapter 6 describes the role of analytics in the airline industry and provides a case study on predicting flight arrival delays (minutes) in R and SAS. Chapter 7 describes the application of analytics in the FMCG industry with a detailed explanation of a business case study on customer segmentation based on their purchasing history using R and SAS.

Who's the Target Audience?

- Data Scientists who would like to implement machine learning techniques with a practical analytical approach toward a particular industrial problem.
- Statistician, Engineers, and Researchers with a great theoretical understanding of data and statistics and would like to enhance their skills by getting practical exposure to data modeling.
- Data analysts who know about data mining but would like to implement predictive models and machine learning techniques.
- Developers who are well versed with coding but would like to transition to a career in data science.

What You Will Learn

- Introduction to analytics and data understanding.
- How to approach industrial business problems with an analytical approach.
- Practical and hands-on knowledge in building predictive model and machine learning techniques.
- Building the analytical strategies.