Chapter 1
Introduction

Sridhar Seshadri

Business analytics is the science of posing and answering data questions related to business. Business analytics has rapidly expanded in the last few years to include tools drawn from statistics, data management, data visualization, and machine learning. There is increasing emphasis on big data handling to assimilate the advances made in data sciences. As is often the case with applied methodologies, business analytics has to be soundly grounded in applications in various disciplines and business verticals to be valuable. The bridge between the tools and the applications are the modeling methods used by managers and researchers in disciplines such as finance, marketing, and operations. This book provides coverage of all three aspects: tools, modeling methods, and applications.

The purpose of the book is threefold: to fill the void in the graduate-level study materials for addressing business problems in order to pose data questions, obtain optimal business solutions via analytics theory, and ground the solution in practice. In order to make the material self-contained, we have endeavored to provide ample use of cases and data sets for practice and testing of tools. Each chapter comes with data, examples, and exercises showing students what questions to ask, how to apply the techniques using open source software, and how to interpret the results. In our approach, simple examples are followed with medium to large applications and solutions. The book can also serve as a self-study guide to professionals who wish to enhance their knowledge about the field.

The distinctive features of the book are as follows:

- The chapters are written by experts from universities and industry.
- The major software used are R, Python, MS Excel, and MYSQL. These are all topical and widely used in the industry.

S. Seshadri (✉)
Gies College of Business, University of Illinois at Urbana Champaign, Champaign, IL, USA
e-mail: sridhar@illinois.edu

B. Pochiraju, S. Seshadri (eds.), *Essentials of Business Analytics*, International
Series in Operations Research & Management Science 264,
https://doi.org/10.1007/978-3-319-68837-4_1

- Extreme care has been taken to ensure continuity from one chapter to the next. The editors have attempted to make sure that the content and flow are similar in every chapter.
- In Part A of the book, the tools and modeling methodology are developed in detail. Then this methodology is applied to solve business problems in various verticals in Part B. Part C contains larger case studies.
- The Appendices cover required material on Probability theory, R, and Python, as these serve as prerequisites for the main text.

The structure of each chapter is as follows:

- Each chapter has a business orientation. It starts with business problems, which are transformed into technological problems. Methodology is developed to solve the technological problems. Data analysis is done using suitable software and the output and results are clearly explained at each stage of development. Finally, the technological solution is transformed back to a business solution. The chapters conclude with suggestions for further reading and a list of references.
- Exercises (with real data sets when applicable) are at the end of each chapter and on the Web to test and enhance the understanding of the concepts and application.
- Caselets are used to illustrate the concepts in several chapters.

1 Detailed Description of Chapters

Data Collection: This chapter introduces the concepts of data collection and problem formulation. Firstly, it establishes the foundation upon which the fields of data sciences and analytics are based, and defines core concepts that will be used throughout the rest of the book. The chapter starts by discussing the types of data that can be gathered, and the common pitfalls that can occur when data analytics does not take into account the nature of the data being used. It distinguishes between primary and secondary data sources using examples, and provides a detailed explanation of the advantages and constraints of each type of data. Following this, the chapter details the types of data that can be collected and sorted. It discusses the difference between nominal-, ordinal-, interval-, and ratio-based data and the ways in which they can be used to obtain insights into the subject being studied.

The chapter then discusses problem formulation and its importance. It explains how and why formulating a problem will impact the data that is gathered, and thus affect the conclusions at which a research project may arrive. It describes a framework by which a messy real-world situation can be clarified so that a mathematical toolkit can be used to identify solutions. The chapter explains the idea of decision-problems, which can be used to understand the real world, and research-objectives, which can be used to analyze decision-problems.

The chapter also details the challenges faced when collecting and collating data. It discusses the importance of understanding what data to collect, how to collect it, how to assess its quality, and finally the most appropriate way of collating it so that it does not lose its value.

The chapter ends with an illustrative example of how the retailing industry might use various sources of data in order to better serve their customers and understand their preferences.

Data Management—Relational Database Management Systems: This chapter introduces the idea of data management and storage. The focus of the chapter is on relational database management systems or RDBMS. RDBMS is the most commonly used data organization system in enterprises. The chapter introduces and explains the ideas using MySQL, an open-source structural query language used by many of the largest data management systems in the world.

The chapter describes the basic functions of a MySQL server, such as creating databases, examining data tables, and performing functions and various operations on data sets. The first set of instructions the chapter discusses is about the rules, definition, and creation of relational databases. Then, the chapter describes how to create tables and add data to them using MySQL server commands. It explains how to examine the data present in the tables using the SELECT command.

Data Management—Big Data: This chapter builds on some of the concepts introduced in the previous chapter but focuses on big data tools. It describes what really constitutes big data and focuses on some of the big data tools. In this chapter, the basics of big data tools such as Hadoop, Spark, and surrounding ecosystem are presented.

The chapter begins by describing Hadoop's uses and key features, as well as the programs in its ecosystem that can also be used in conjunction with it. It also briefly visits the concepts of distributed and parallel computing and big data cloud.

The chapter describes the architecture of the Hadoop runtime environment. It starts by describing the cluster, which is the set of host machines, or nodes for facilitating data access. It then moves on to the YARN infrastructure, which is responsible for providing computational resources to the application. It describes two main elements of the YARN infrastructure—the Resource Manager and the Node Manager. It then details the HDFS Federation, which provides storage, and also discusses other storage solutions. Lastly, it discusses the MapReduce framework, which is the software layer.

The chapter then describes the functions of MapReduce in detail. MapReduce divides tasks into subtasks, which it runs in parallel in order to increase efficiency. It discusses the manner in which MapReduce takes lists of input data and transforms them into lists of output data, by implementing a "map" process and a "reduce" process, which it aggregates. It describes in detail the process steps that MapReduce takes in order to produce the output, and describes how Python can be used to create a MapReduce process for a word count program.

The chapter briefly describes Spark and an application using Spark. It concludes with a discussion about cloud storage. The chapter makes use of Cloudera virtual machine (VM) distributable to demonstrate different hands-on exercises.

Data Visualization: This chapter discusses how data is visualized and the way that visualization can be used to aid in analysis. It starts by explaining that humans use visuals to understand information, and that using visualizations incorrectly can lead to mistaken conclusions. It discusses the importance of visualization as a cognitive aid and the importance of working memory in the brain. It emphasizes the role of data visualization in reducing the load on the reader.

The chapter details the six meta-rules of data visualization, which are as follows: use the most appropriate chart, directly represent relationships between data, refrain from asking the viewer to compare differences in area, never use color on top of color, keep within the primal perceptions of the viewer, and chart with integrity.

Each rule is expanded upon in the chapter. The chapter discusses the kinds of graphs and tables available to a visualizer, the advantages and disadvantages of 3D visualization, and the best practices of color schemes.

Statistical Methods—Basic Inferences: This chapter introduces the fundamental concepts of statistical inferences, such as population and sample parameters, hypothesis testing, and analysis of variance. It begins by describing the differences between population and sample means and variance and the methods to calculate them. It explains the central limit theorem and its use in estimating the mean of a population.

Confidence intervals are explained for samples in which variance is both known and unknown. The concept of standard errors and the t- and Chi-squared distributions are introduced. The chapter introduces hypothesis testing and the use of statistical parameters to reject or fail to reject hypotheses. Type I and type II errors are discussed.

Methods to compare two different samples are explained. Analysis of variance between two samples and within samples is also covered. The use of the F-distribution in analyzing variance is explained. The chapter concludes with discussion of when we need to compare means of a number of populations. It explains how to use a technique called “Analysis of Variance (ANOVA)” instead of carrying out pairwise comparisons.

Statistical Methods—Linear Regression Analysis: This chapter explains the idea of linear regression in detail. It begins with some examples, such as predicting newspaper circulation. It uses the examples to discuss the methods by which linear regression obtains results. It describes a linear regression as a functional form that can be used to understand relationships between outcomes and input variables and perform statistical inference. It discusses the importance of linear regression and its popularity, and explains the basic assumptions underlying linear regression.

The modeling section begins by discussing a model in which there is only a single regressor. It explains why a scatter-plot can be useful in understanding single-regressor models, and the importance of visual representation in statistical inference. It explains the ordinary least squares method of estimating a parameter, and the use of the sum of squares of residuals as a measure of the fit of a model. The chapter then discusses the use of confidence intervals and hypothesis testing in a linear regression model. These concepts are used to describe a linear regression model in which there are multiple regressors, and the changes that are necessary to adjust a single linear regression model to a multiple linear regression model.

The chapter then describes the ways in which the basic assumptions of the linear regression model may be violated, and the need for further analysis and diagnostic tools. It uses the famous Anscombe data sets in order to demonstrate the existence of phenomena such as outliers and collinearity that necessitate further analysis. The methods needed to deal with such problems are explained. The chapter considers the ways in which the necessity for the use of such methods may be determined, such as tools to determine whether some data points should be deleted or excluded from the data set. The possible advantages and disadvantages of adding additional regressors to a model are described. Dummy variables and their use are explained. Examples are given for the case where there is only one category of dummy, and then multiple categories.

The chapter then discusses assumptions regarding the error term. The effect of the assumption that the error term is normally distributed is discussed, and the Q-Q plot method of examining the truth of this assumption for the data set is explained. The Box–Cox method of transforming the response variable in order to normalize the error term is discussed. The chapter then discusses the idea that the error terms may not have equal variance, that is, be homoscedastic. It explains possible reasons for heteroscedasticity, and the ways to adapt the analysis to those situations.

The chapter considers the methods in which the regression model can be validated. The root mean square error is introduced. Segmenting the data into training and validation sets is explained. Finally, some frequently asked questions are presented, along with exercises.

Statistical Methods—Advanced Regression: Three topics are covered in this chapter. In the main body of the chapter the tools for estimating the parameters of regression models when the response variable is binary or categorical is presented. The appendices to the chapter cover two other important techniques, namely, maximum likelihood estimate (MLE) and how to deal with missing data.

The chapter begins with a description of logistics regression models. It continues with diagnostics of logistics regression, including likelihood ratio tests, Wald's and the Hosmer–Lemeshow tests. It then discusses different R-squared tests, such as Cox and Snell, Nagelkerke, and McFadden. Then, it discusses how to choose the cutoff probability for classification, including discussion of discordant and concordant pairs, the ROC curve, and Youden's index. It concludes with a similar discussion of Multinomial Logistics Function and regression. The chapter contains a self-contained introduction to the maximum likelihood method and methods for treating missing data. The ideas introduced in this chapter are used in several following chapters in the book.

Text Analytics: This is the first of several chapters that introduce specialized analytics methods depending on the type of data and analysis. This chapter begins by considering various motivating examples for text analysis. It explains the need for a process by which unstructured text data can be analyzed, and the ways that it can be used to improve business outcomes. It describes in detail the manner in which Google used its text analytics software and its database of searches to identify vectors of H1N1 flu. It lists out the most common sources of text data, with social media platforms and blogs producing the vast majority.

The second section of the chapter concerns the ways in which text can be analyzed. It describes two approaches: a "bag-of-words" approach, in which the structure of the language is not considered important, and a "natural-language" approach, in which structure and phrases are also considered.

The example of a retail chain surveying responses to a potential ice-cream product is used to introduce some terminology. It uses this example to describe the problems of analyzing sentences due to the existence of grammatical rules, such as the abundance of articles or the different tense forms of verbs. Various methods of dealing with these problems are introduced. The term-document matrix (TDM) is introduced along with its uses, such as generation of wordclouds.

The third and fourth sections of the chapter describe how to run text analysis and some elementary applications. The text walks through a basic use of the program R to analyze text. It looks at two ways that the TDM can be used to run text analysis—using a text-base to cluster or segment documents, and elementary sentiment analysis.

Clustering documents is a method by which similar customers are sorted into the same group by analyzing their responses. Sentiment analysis is a method by which attempts are made to make value judgments and extract qualitative responses. The chapter describes the models for both processes in detail with regard to an example.

The fifth section of the chapter then describes the more advanced technique of latent topic mining. Latent topic mining aims to identify themes present in a corpus, or a collection of documents. The chapter uses the example of the mission statements of Fortune-1000 firms in order to identify some latent topics.

The sixth section of the chapter concerns natural-language processing (NLP). NLP is a set of techniques that enables computers to understand nuances in human languages. The method by which NLP programs detect data is discussed. The ideas of this chapter are further explored in the chapter on Deep Learning. The chapter ends with exercises for the student.

Simulation: This chapter introduces the uses of simulation as a tool for analytics, focusing on the example of a fashion retailer. It explains the use of Monte Carlo simulation in the presence of uncertainty as an aid to making decisions that have various trade-offs.

First, the chapter explains the purposes of simulation, and the ways it can be used to design an optimal intervention. It differentiates between computer simulation, which is the main aim of the chapter, and physical simulation. It discusses the advantages and disadvantages of simulations, and mentions various applications of simulation in real-world contexts.

The second part of the chapter discusses the steps that are followed in making a simulation model. It explains how to identify dependent and independent variables, and the manner in which the relationships between those variables can be modeled. It describes the method by which input variables can be randomly generated, and the output of the simulation can be interpreted. It illustrates these steps using the example of a fashion retailer that needs to make a decision about production.

The third part of the chapter describes decision-making under uncertainty and the ways that simulation can be used. It describes how to set out a range of possible interventions and how they can be modeled using a simulation. It discusses how to use simulation processes in order to optimize decision-making under constraints, by using the fashion retailer example in various contexts.

The chapter also contains a case study of a painting business deciding how much to bid for a contract to paint a factory, and describes the solution to making this decision. The concepts explained in this chapter are applied in different settings in the following chapters.

Optimization: Optimization techniques are used in almost every application in this book. This chapter presents some of the core concepts of constrained optimization. The basic ideas are illustrated using one broad class of optimization problems called linear optimization. Linear optimization covers the most widely used models in business. In addition, because linear models are easy to visualize in two dimensions, it offers a visual introduction to the basic concepts in optimization. Additionally, the chapter provides a brief introduction to other optimization models and techniques such as integer/discrete optimization, nonlinear optimization, search methods, and the use of optimization software.

The linear optimization part is conventionally developed by describing the decision variables, the objective function, constraints, and the assumptions underlying the linear models. Using geometric arguments, it illustrates the concept of feasibility and optimality. It then provides the basic theorems of linear programming. The chapter then develops the idea of shadow prices, reduced costs, and sensitivity analysis, which is the underpinning of any post-optimality business analysis. The solver function in Excel is used for illustrating these ideas. Then, the chapter explains how these ideas extend to integer programming and provides an outline of the branch and bound method with examples. The ideas are further extended to nonlinear optimization via examples of models for linear regression, maximum likelihood estimation, and logistic regression.

Forecasting Analytics: Forecasting is perhaps the most commonly used method in business analytics. This chapter introduces the idea of using analytics to predict the outcomes in the future, and focuses on applying analytics tools for business and operations. The chapter begins by explaining the difficulty of predicting the future with perfect accuracy, and the importance of accepting the uncertainty inherent in any predictive analysis.

The chapter begins by defining forecasting as estimating in unknown situations. It describes data that can be used to make forecasts, but focuses on time-series forecasting. It introduces the concepts of point-forecasts and prediction intervals, which are used in time-series analysis as part of predictions of future outcomes. It suggests reasons for the intervention of human judgment in the forecasts provided by computers. It describes the core method of time-series forecasting—identifying a model that forecasts the best.

The second part of the chapter describes quantitative approaches to forecasting. It begins by describing the various kinds of data that can be used to make forecasts, such as spoken, written, numbers, and so on. It explains some methods of dealing with outliers in the data set, which can affect the fit of the forecast, such as trimming and winsorizing.

The chapter discusses the effects of seasonal fluctuations on time-series data and how to adjust for them. It introduces the autocorrelation function and its use. It also explains the partial autocorrelation function.

A number of methods used in predictive forecasting are explained, including the naïve method, the average and moving average methods, Holt exponential smoothing, and the ARIMA framework. The chapter also discusses ways to predict stochastic intermittent demand, such as Croston's approach, and the Syntetos and Boylan approximation.

The third section of the chapter describes the process of applied forecasting analytics at the operational, tactical, and strategic levels. It propounds a seven-step forecasting process for operational tasks, and explains each step in detail.

The fourth section of the chapter concerns evaluating the accuracy of forecasts. It explains measures such as mean absolute error, mean squared error, and root mean squared error, and how to calculate them. Both Excel and R software use is explained.

Advanced Statistical Methods: Count Data: The chapter begins by introducing the idea of count variables and gives examples of where they are encountered, such as insurance applications and the amount of time taken off by persons that fall sick.

It first introduces the idea of the Poisson regression model, and explains why ordinary least squares are not suited to some situations for which the Poisson model is more appropriate. It illustrates the differences between the normal and Poisson distributions using conditional distribution graphs.

It defines the Poisson distribution model and its general use, as well as an example regarding insurance claims data. It walks through the interpretation of the regression's results, including the explanation of the regression coefficients, deviance, dispersion, and so on.

It discusses some of the problems with the Poisson regression, and how overdispersion can cause issues for the analysis. It introduces the negative binomial distribution as a method to counteract overdispersion. Zero-inflation models are discussed. The chapter ends with a case study on Canadian insurance data.

Advanced Statistical Methods—Survival Analysis: Like the previous chapter, this one deals with another specialized application. It involves techniques that analyze time-to-event data. It defines time-to-event data and the contexts in which it can be used, and provides a number of business situations in which survival analysis is important.

The chapter explains the idea of censored data, which refers to survival times in which the event in question has not yet occurred. It explains the differences between survival models and other types of analysis, and the fields in which it can be used. It defines the types of censoring: right-censoring, left-censoring, and interval-censoring, and the method to incorporate them into the data set.

The chapter then defines the survival analysis functions: the survival function and the hazard function. It describes some simple types of hazard functions. It describes some parametric and nonparametric methods of analysis, and defines the cases in which nonparametric methods must be used. It explains the Kaplan–Meier method in detail, along with an example. Semiparametric models are introduced for cases in which several covariate variables are believed to contribute to survival. Cox's proportional hazards model and its interpretation are discussed.

The chapter ends with a comparison between semiparametric and parametric models, and a case study regarding churn data.

Unsupervised Learning: The first of the three machine learning chapters sets out the philosophy of machine learning. This chapter explains why unsupervised learning—an important paradigm in machine learning—is akin to uncovering the proverbial needle in the haystack, discovering the grammar of the process that generated the data, and exaggerating the "signal" while ignoring the "noise" in it. The chapter covers methods of projection, clustering, and density estimation—three core unsupervised learning frameworks that help us perceive the data in different ways. In addition, the chapter describes collaborative filtering and applications of network analysis.

The chapter begins with drawing the distinction between supervised and unsupervised learning. It then presents a common approach to solving unsupervised learning problems by casting them into an optimization framework. In this framework, there are four steps:

- Intuition: to develop an intuition about how to approach the problem as an optimization problem
- Formulation: to write the precise mathematical objective function in terms of data using intuition
- Modification: to modify the objective function into something simpler or "more solvable"
- Optimization: to solve the final objective function using traditional optimization approaches

The chapter discusses principal components analysis (PCA), self-organizing maps (SOM), and multidimensional scaling (MDS) under projection algorithms. In clustering, it describes partitional and hierarchical clustering. Under density estimation, it describes nonparametric and parametric approaches. The chapter concludes with illustrations of collaborative filtering and network analysis.

Supervised Learning: In supervised learning, the aim is to learn from previously identified examples. The chapter covers the philosophical, theoretical, and practical aspects of one of the most common machine learning paradigms—supervised learning—that essentially learns to map from an observation (e.g., symptoms and test results of a patient) to a prediction (e.g., disease or medical condition), which in turn is used to make decisions (e.g., prescription). The chapter then explores the process, science, and art of building supervised learning models.

The first part explains the different paradigms in supervised learning: classification, regression, retrieval, recommendation, and how they differ by the nature of their input and output. It then describes the process of learning, from features description to feature engineering to models to algorithms that help make the learning happen.

Among algorithms, the chapter describes rule-based classifiers, decision trees, k-nearest neighbor, Parzen window, and Bayesian and naïve Bayes classifiers. Among discriminant functions that partition a region using an algorithm, linear (LDA) and quadratic discriminant analysis (QDA) are discussed. A section describes recommendation engines. Neural networks are then introduced followed by a succinct introduction to a key algorithm called support vector machines (SVM). The chapter concludes with a description of ensemble techniques, including bagging, random forest, boosting, mixture of experts, and hierarchical classifiers. The specialized neural networks for Deep Learning are explained in the next chapter.

Deep Learning: This chapter introduces the idea of deep learning as a part of machine learning. It aims to explain the idea of deep learning and various popular deep learning architectures. It has four main parts:

- Understand what is deep learning.
- Understand various popular deep learning architectures, and know when to use which architecture for solving a business problem.
- How to perform image analysis using deep learning.
- How to perform text analysis using deep learning.

The chapter explains the origins of learning, from a single perceptron to mimic the functioning of a neuron to the multilayered perceptron (MLP). It briefly recaps the backpropagation algorithm and introduces the learning rate and error functions. It then discusses the deep learning architectures applied to supervised, unsupervised, and reinforcement learning. An example of using an artificial neural network for recognizing handwritten digits (based on the MNIST data set) is presented.

The next section of the chapter describes Convolutional Neural Networks (CNN), which are aimed at solving vision-related problems. The ImageNet data set is introduced. The use of CNNs in the ImageNet Large Scale Visual Recognition Challenge is explained, along with a brief history of the challenge. The biological inspiration for CNNs is presented. Four layers of a typical CNN are introduced—the convolution layer, the rectified linear units layer, the pooling layers, and the fully connected layer. Each layer is explained, with examples. A unifying example using the same MNIST data set is presented.

The third section of the chapter discusses recurrent neural networks (RNNs). It begins by describing the motivation for sequence learning models, and their use in understanding language. Traditional language models and their functions in predicting words are explained. The chapter describes a basic RNN model with three units, aimed at predicting the next word in a sentence. It explains the detailed example by which an RNN can be built for next word prediction. It presents some uses of RNNs, such as image captioning and machine translation.

The next seven chapters contain descriptions of analytics usage in different domains and different contexts. These are described next.

Retail Analytics: The chapter begins by introducing the background and definition of retail analytics. It focuses on advanced analytics. It explains the use of four main categories of business decisions: consumer, product, human resources, and advertising. Several examples of retail analytics are presented, such as increasing book recommendations during periods of cold weather. Complications in retail analytics are discussed.

The second part of the chapter focuses on data collection in the retail sector. It describes the traditional sources of retail data, such as point-of-sale devices, and how they have been used in decision-making processes. It also discusses advances in technology and the way that new means of data collection have changed the field. These include the use of radio frequency identification technology, the Internet of things, and Bluetooth beacons.

The third section describes methodologies, focusing on inventory, assortment, and pricing decisions. It begins with modeling product-based demand in order to make predictions. The penalized L1 regression LASSO for retail demand forecasting is introduced. The use of regression trees and artificial neural networks is discussed in the same context. The chapter then discusses the use of such forecasts in decision-making. It presents evidence that machine learning approaches benefit revenue and profit in both price-setting and inventory-choice contexts.

Demand models into which consumer choice is incorporated are introduced. The multinomial logit, mixed multinomial logit, and nested logit models are described. Nonparametric choice models are also introduced as an alternative to logit models. Optimal assortment decisions using these models are presented. Attempts at learning customer preferences while optimizing assortment choices are described.

The fourth section of the chapter discusses business challenges and opportunities. The benefits of omnichannel retail are discussed, along with the need for retail analytics to change in order to fit an omnichannel shop. It also discusses some recent start-ups in the retail analytics space and their focuses.

Marketing Analytics: Marketing is one of the most important, historically the earliest, and fascinating areas for applying analytics to solve business problems. Due to the vast array of applications, only the most important ones are surveyed in this chapter. The chapter begins by explaining the importance of using marketing analytics for firms. It defines the various levels that marketing analytics can apply to: the firm, the brand or product, and the customer. It introduces a number of processes and models that can be used in analyzing and making marketing decisions, including statistical analysis, nonparametric tools, and customer analysis. The processes and tools discussed in this chapter will help in various aspects of marketing such as target marketing and segmentation, price and promotion, customer valuation, resource allocation, response analysis, demand assessment, and new product development.

The second section of the chapter explains the use of the interaction effect in regression models. Building on earlier chapters on regression, it explains the utility of a term that captures the effect of one or more interactions between other variables. It explains how to interpret new variables and their significance. The use of curvilinear relationships in order to identify the curvilinear effect is discussed. Mediation analysis is introduced, along with an example.

The third section describes data envelopment analysis (DEA), which is aimed at improving the performance of organizations. It describes the manner in which DEA works to present targets to managers and can be used to answer key operational questions in Marketing: sales force productivity, performance of sales regions, and effectiveness of geomarketing.

The next topic covered is conjoint analysis. It explains how knowing customers' preference provides invaluable information about how customers think and make their decisions before purchasing products. Thus, it helps firms devise their marketing strategies including advertising, promotion, and sales activities.

The fifth section of the chapter discusses customer analytics. Customer lifetime value (CLV), a measure of the value provided to firms by customers, is introduced, along with some other measures. A method to calculate CLV is presented, along with its limitations. The chapter also discusses two more measures of customer value: customer referral value and customer influence value, in detail. Additional topics are covered in the chapters on retail analytics and social media analytics.

Financial Analytics: Financial analytics like Marketing has been a big consumer of data. The topics chosen in this chapter provide one unified way of thinking about analytics in this domain—valuation. This chapter focuses on the two main branches of quantitative finance: the risk-neutral or "Q" world and the risk-averse or "P" world. It describes the constraints and aims of analysts in each world, along with their primary methodologies. It explains Q-quant theories such as the work of Black and Scholes, and Harrison and Pliska. P-quant theories such as net present value, capital asset pricing models, arbitrage pricing theory, and the efficient market hypothesis are presented.

The methodology of financial data analytics is explained via a three-stage process: asset price estimation, risk management, and portfolio analysis.

Asset price estimation is explained as a five-step process. It describes the use of the random walk in identifying the variable to be analyzed. Several methods of transforming the variable into one that is identical and independently distributed are presented. A maximum likelihood estimation method to model variance is explained. Monte Carlo simulations of projecting variables into the future are discussed, along with pricing projected variables.

Risk management is discussed as a three-step process. The first step is risk aggregation. Copula functions and their uses are explained. The second step, portfolio assessment, is explained by using metrics such as Value at Risk. The third step, attribution, is explained. Various types of capital at risk are listed.

Portfolio analysis is described as a two-stage process. Allocating risk for the entire portfolio is discussed. Executing trades in order to move the portfolio to a new risk/return level is explained.

A detailed example explaining each of the ten steps is presented, along with data and code in MATLAB. This example also serves as a stand-alone case study on financial analytics.

Social Media Analytics: Social-media-based analytics has been growing in importance and value to businesses. This chapter discusses the various tools available to gather and analyze data from social media and Internet-based sources, focusing on the use of advertisements. It begins by describing Web-based analytical tools and the information they can provide, such as cookies, sentiment analysis, and mobile analytics.

It introduces real-time advertising on online platforms, and the wealth of data generated by browsers visiting target websites. It lists the various kinds of advertising possible, including video and audio ads, map-based ads, and banner ads. It explains the various avenues in which these ads can be displayed, and details the reach of social media sites such as Facebook and Twitter. The various methods in which ads can be purchased are discussed. Programmatic advertising and its components are introduced. Real-time bidding on online advertising spaces is explained.

A/B experiments are defined and explained. The completely randomized design (CRD) experiment is discussed. The regression model for the CRD and an example are presented. The need for randomized complete block design experiments is introduced, and an example for such an experiment is shown. Analytics of multivariate experiments and their advantages are discussed. Orthogonal designs and their meanings are explained.

The chapter discusses the use of data-driven search engine advertising. The use of data in order to help companies better reach consumers and identify trends is discussed. The power of search engines in this regard is discussed. The problem of attribution, or identifying the influence of various ads across various platforms is introduced, and a number of models that aim to solve this problem are elucidated. Some models discussed are: the first click attribution model, the last click attribution model, the linear attribution model, and algorithmic attribution models.

Healthcare Analytics: Healthcare is once again an area where data, experiments, and research have coexisted within an analytical framework for hundreds of years. This chapter discusses analytical approaches to healthcare. It begins with an overview of the current field of healthcare analytics. It describes the latest innovations in the use of data to refine healthcare, including telemedicine, wearable technologies, and simulations of the human body. It describes some of the challenges that data analysts can face when attempting to use analytics to understand healthcare-related problems.

The main part of the chapter focuses on the use of analytics to improve operations. The context is patient flow in outpatient clinics. It uses Academic Medical Centers as an example to describe the processes that patients go through when visiting clinics that are also teaching centers. It describes the effects of the Affordable Care Act, an aging population, and changes in social healthcare systems on the public health infrastructure in the USA.

A five-step process map of a representative clinic is presented, along with a discrete event simulation of the clinic. The history of using operations research-based methods to improve healthcare processes is discussed. The chapter introduces a six-step process aimed at understanding complex systems, identifying potential improvements, and predicting the effects of changes, and describes each step in detail.

Lastly, the chapter discusses the various results of this process on some goals of the clinic, such as arrivals, processing times, and impact on teaching. Data regarding each goal and its change are presented and analyzed. The chapter contains a hands-on exercise based on the simulation models discussed. The chapter is a fine application of simulation concepts and modeling methodologies used in Operations Management to improve healthcare systems.

Pricing Analytics: This chapter discusses the various mechanisms available to companies in order to price their products. The topics pertain to revenue management, which constitutes perhaps the most successful and visible area of business analytics.

The chapter begins by introducing defining two factors that affect pricing: the nature of the product and its competition, and customers' preferences and values. It introduces the concept of a price optimization model, and the need to control capacity constraints when estimating customer demand.

The first type of model introduced is the independent class model. The underlying assumption behind the model is defined, as well as its implications for modeling customer choice. The EMSR heuristic and its use are explained.

The issue of overbooking in many service-related industries is introduced. The trade-off between an underutilized inventory and the risk of denying service to customers is discussed. A model for deciding an overbooking limit, given the physical capacity at the disposal of the company, is presented. Dynamic pricing is presented as a method to better utilize inventory.

Three main types of dynamic pricing are discussed: surge pricing, repricing, and markup/markdown pricing. Each type is comprehensively explained. Three models of forecasting and estimating customer demand are presented: additive, multiplicative, and choice.

A number of processes for capacity control, such as nested allocations, are presented. Network revenue management systems are introduced. A backward induction method of control is explained. The chapter ends with an example of a hotel that is planning allocation of rooms based on a demand forecast.

Supply Chain Analytics: This chapter discusses the use of data and analytical tools to increase value in the supply chain. It begins by defining the processes that constitute supply chains, and the goals of supply chain management. The uncertainty inherent in supply chains is discussed. Four applications of supply chain analytics are described: demand forecasting, inventory optimization, supply chain disruption, and commodity procurement.

A case study of VASTA, one of the largest wireless services carriers in the USA, is presented. The case study concerns the decision of whether the company should change its current inventory strategy from a "push" strategy to a "pull" strategy. The advantages and disadvantages of each strategy are discussed. A basic model to evaluate both strategies is introduced. An analysis of the results is presented. Following the analysis, a more advanced evaluation model is introduced. Customer satisfaction and implementation costs are added to the model.

The last three chapters of the book contain case studies. Each of the cases comes with a large data set upon which students can practice almost every technique and modeling approach covered in the book. The Info Media case study explains the use of viewership data to design promotional campaigns. The problem presented is to determine a multichannel ad spots allocation in order to maximize "reach" given a budget and campaign guidelines. The approach uses simulation to compute the viewership and then uses the simulated data to link promotional aspects to the total reach of a campaign. Finally, the model can be used to optimize the allocation of budgets across channels.

The AAA airline case study illustrates the use of choice models to design airline offerings. The main task is to develop a demand forecasting model, which predicts the passenger share for every origin–destination pair (O–D pair) given AAA, as well as competitors' offerings. The students are asked to explore different models including the MNL and machine learning algorithms. Once a demand model has been developed it can be used to diagnose the current performance and suggest various remedies, such as adding, dropping, or changing itineraries in specific city pairs. The third case study, Ideal Insurance, is on fraud detection. The problem faced by the firm is the growing cost of servicing and settling claims in their healthcare practice. The students learn about the industry and its intricate relationships with various stakeholders. They also get an introduction to rule-based decision support systems. The students are asked to create a system for detecting fraud, which should be superior to the current "rule-based" system.

2 The Intended Audience

This book is the first of its kind both in breadth and depth of coverage and serves as a textbook for students of first year graduate program in analytics and long duration (1-year part time) certificate programs in business analytics. It also serves as a perfect guide to practitioners.

The content is based on the curriculum of the Certificate Programme in Business Analytics (CBA), now renamed as Advanced Management Programme in Business Analytics (AMPBA) of Indian School of Business (ISB). The original curriculum was created by Galit Shmueli. The curriculum was further developed by the coeditors, Bhimasankaram Pochiraju and Sridhar Seshadri, who were responsible for starting and mentoring the CBA program in ISB. Bhimasankaram Pochiraju has been the Faculty Director of CBA since its inception and was a member of the Academic Board. Sridhar Seshadri managed the launch of the program and since then has chaired the academic development efforts. Based on the industry needs, the curriculum continues to be modified by the Academic Board of the Applied Statistics and Computing Lab (ASC Lab) at ISB.