MGS 3701 / MKT 3950: Data Mining

Chapter1: Introduction

Chungil Chae

温州肯恩大学
WENZHOU-KEAN UNIVERSITY

2022/01/01 (updated: 2022-02-09)

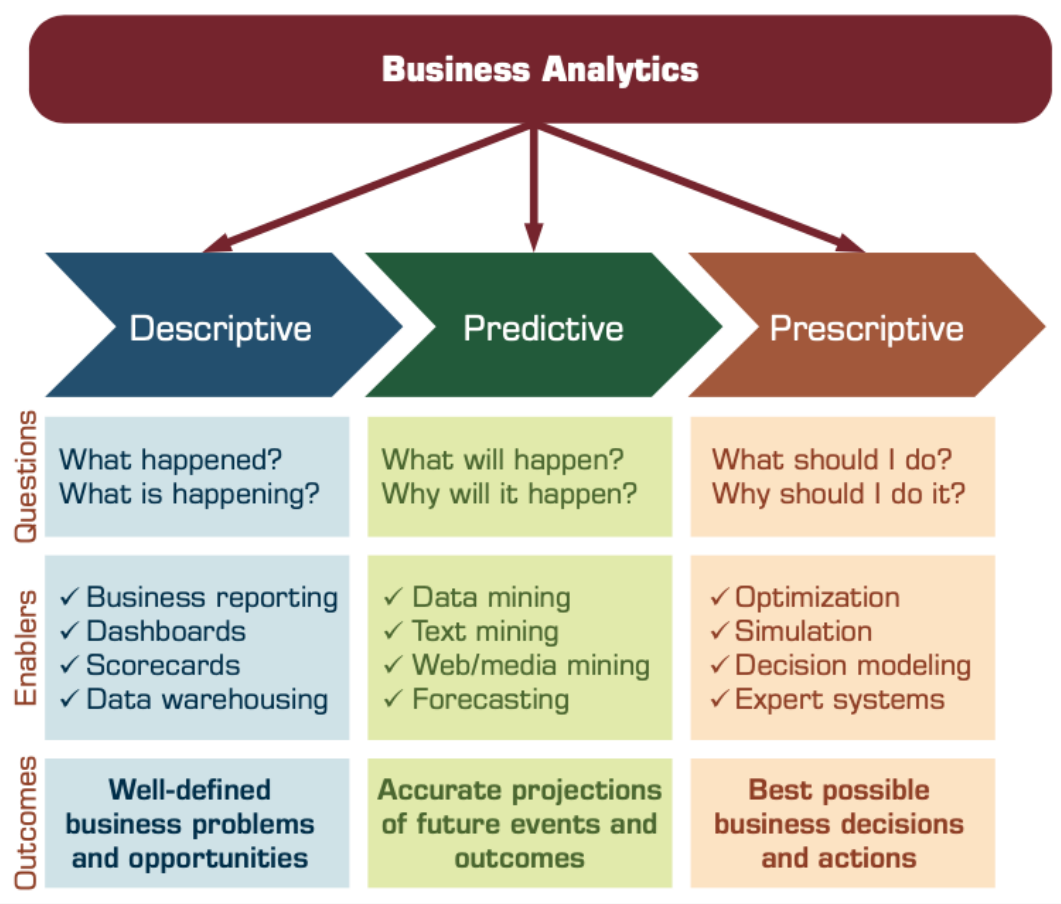
1. What Is Business Analytics?

Business Analytics

- Aanalytics represents the combination of computer technology, management science techniques, and statistics to solve real problems.

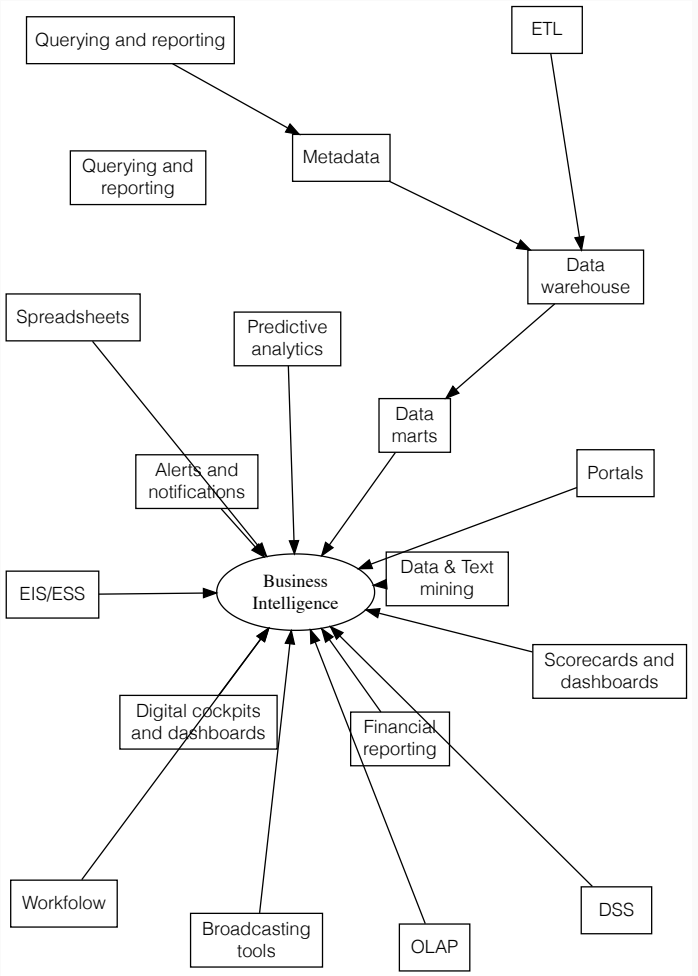
Thus,

- Business Analytics (BA) is the practice and art of bringing quantitative data to bear on besuiness related decision-making.



Business Intelligence

- Business Intelligence (BI), refers to data visualization and reporting for understanding “what happened and what is happening.”
- BI, which earlier consisted mainly of generating static reports, has evolved into more user-friendly and effective tools and practices,
  - interactive dashboards
  - displays customer orders in a single two-dimensional display,
  - using color and bubble size as added variables, showing customer name, type of product, etc



BA vs BI

- Business Analytics now typically includes BI
  - statistical models
  - data mining algorithms
  - exploring data,
  - quantifying and explaining relationships between measurements
  - predicting new records.

Methods like regression models are used to describe and quantify “on average” relationships (e.g., between advertising and sales), to predict new records and to forecast future values

- Business analytics is a solution in search of a problem: A manager, knowing that business analytics and data mining are hot areas
    - deploy them
    - to capture that hidden value
- Successful use of analytics and data mining requires
    - understanding of the business context where value is to be captured,
    - understanding of exactly what the data mining methods do.

2. What Is Data Mining?

- Data mining refers to business analytics methods that go beyond counts, descriptive techniques, reporting, and methods based on business rules.
  - statistical methods
  - machine-learning methods
  - inform decision-making
  - in an automated fashion.

- Prediction is typically an important component, often at the individual level.
  - Rather than *"what is the relationship between advertising and sales"*
  - we might be interested in **"what specific advertisement"**.
- The era of Big Data has accelerated the use of data mining.
  - Data mining methods, with their power and automaticity, have the ability to cope with huge amounts of data and extract value.

3. Data Mining and Related Terms

Overlap and inconsistency of definitions.

- Different things to different people.
  - To the general public, it may have a general
- Major consulting firm has a “data mining department,”
  - but its responsibilities are in the area of studying and graphing past data in search of general trends.
  - More advanced predictive models are the responsibility of an “advanced analytics department.”

- Other terms that organizations use are
  - predictive analytics
  - predictive modeling
  - machine learning.
- Data mining stands at the confluence of the fields of statistics and machine learning (also known as artificial intelligence).
  - A variety of techniques for exploring data in statistics:
  - linear regression,
  - logistic regression,
  - discriminant analysis,
  - principal components analysis.
  - But the core tenets of classical statistics—computing is difficult
  - Data are scarce—do not apply in data mining applications where both data and computing power are plentiful.

Data mining as “statistics at scale and speed” (Pregibon, 1999).

- Major difference between the fields of statistics and machine learning
  - focus in statistics on inference from a sample to the population regarding an “average effect”
- The emphasis that classical statistics places on inference
- Data mining deals with large datasets in an open-ended fashion

> the general approach to data mining is vulnerable to the danger of overfitting(the model is fitting the noise, not just the signal)

In this course

- machine learning
  - to refer to algorithms that learn directly from data, especially local patterns, often in layered or iterative fashion.
- statistical models
  - to refer to methods that apply global structure to the data.

> Lastly, many practitioners, particularly those from the IT and computer science communities, use the term machine learning to refer to all the methods discussed in this course.

4. Big Data

Big Data is a relative term—data today are big by reference to the past, and to the methods and devices available to deal with them.

- The challenge Big Data presents is often characterized by the four V's—volume, velocity, variety, and veracity.
  - **Volume** refers to the amount of data.
  - **Velocity** refers to the flow rate—the speed at which it is being generated and changed.
  - **Variety** refers to the different types of data being generated (currency, dates, numbers, text, etc.).
  - **Veracity** refers to the fact that data is being generated by organic distributed processes.

> Most large organizations face both the challenge and the opportunity of Big Data.

- If the analytical challenge is substantial, so can be the reward:
  - **OKCupid**, the online dating site, uses statistical models with their data to predict what forms of message content are most likely to produce a response.
  - **Telenor**, a Norwegian mobile phone service company, was able to reduce subscriber turnover 37% by using models to predict which customers were most likely to leave, and then lavishing attention on them.
  - **Allstate**, the insurance company, tripled the accuracy of predicting injury liability in auto claims by incorporating more information about vehicle type.

5. Data Science

- The ubiquity, size, value, and importance of Big Data has given rise to a new profession: the data scientist.
  - Data science is a mix of skills in the areas of
  - statistics
  - machine learning
  - math
  - programming
  - business
  - IT

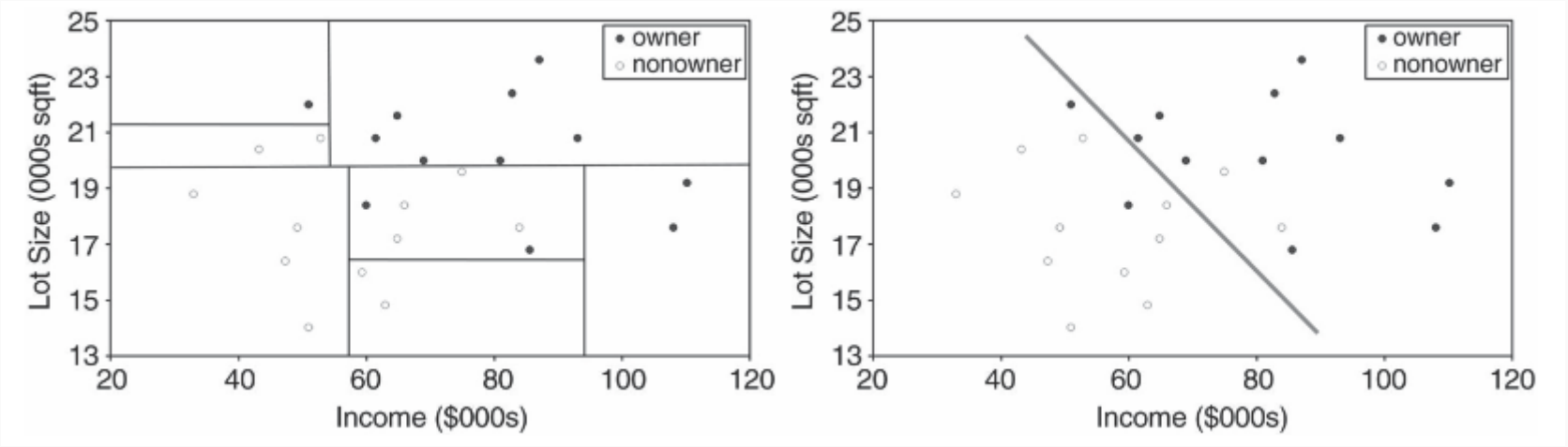
- The term itself is thus broader than the other concepts we discussed above, and it is a rare individual who combines deep skills in all the constituent areas.

> The skill sets of most data scientists as resembling a ‘T’—deep in one area (the vertical bar of the T), and shallower in other areas (the top of the T) (Harris et al., 2013).

- Although Big Data is the motivating power behind the growth of data science, most data scientists do not actually spend most of their time working with terabyte-size or larger data.
    - Data of the terabyte or larger size would be involved at the deployment stage of a model.
    - There are manifold challenges at that stage, most of them IT and programming issues related to data-handling and tying together different components of a system.
- It is that earlier piloting and prototyping phase on which this book focuses—developing the statistical and machine learning models that will eventually be plugged into a deployed system.
    - What methods do you use with what sorts of data and problems?
    - How do the methods work?
    - What are their requirements, their strengths, their weaknesses?
    - How do you assess their performance?

6. Why Are There So Many Different

- Many different methods for prediction and classification.
    - why they coexist
    - whether some are better than others.
- The answer is that each method has advantages and disadvantages.
    - The usefulness of a method can depend on factors such as
    - the size of the dataset,
    - the types of patterns that exist in the data,
    - whether the data meet some underlying assumptions of the method,
    - how noisy the data are, and
    - the particular goal of the analysis.



- A small illustration is shown in Figure 1.1, where the goal is to find a combination of household income level and household lot size that separates buyers (solid circles) from nonbuyers (hollow circles) of riding mowers
  - The first method (left panel) looks only for horizontal and vertical lines to separate buyers from nonbuyers,
  - The second method (right panel) looks for a single diagonal line

> Different methods can lead to different results, and their performance can vary. It is therefore customary in data mining to apply several different methods and select the one that appears most useful for the goal at hand.

7. Terminology and Notation

- Because of the hybrid parentry of data mining, its practitioners often use multiple terms to refer to the same thing.
  - In the machine learning (artificial intelligence) field, the variable being predicted is the output variable or target variable.
  - To a statistician, it is the dependent variable or the response.

- **Algorithm**
  - A specific procedure used to implement a particular data mining technique: classification tree, discriminant analysis, and the like.
- **Attribute**
  - see Predictor.
- **Case**
  - see Observation.
- **Confidence**
  - A performance measure in association rules of the type “IF A and B are purchased, THEN C is also purchased.” Confidence is the conditional probability that C will be purchased IF A and B are purchased.
  - Confidence also has a broader meaning in statistics (confidence interval), concerning the degree of error in an estimate that results from selecting one sample as opposed to another.
- **Dependent Variable**
  - see Response.
- **Estimation**
  - see Prediction.

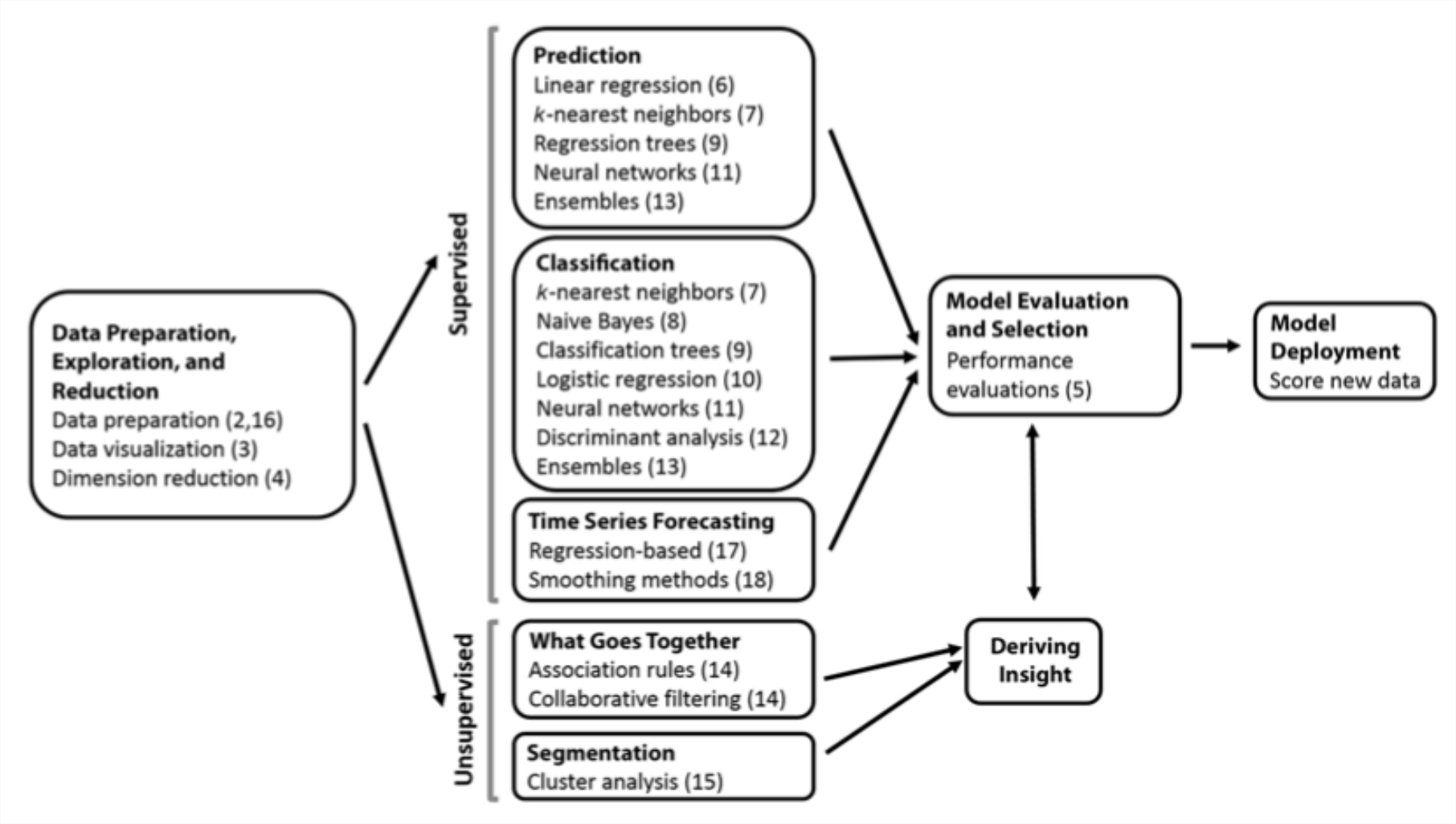
- **Feature**
  - see Predictor.
- **Holdout Data (or holdout set)**
  - A sample of data not used in fitting a model, but instead used to assess the performance of that model. This book uses the terms validation set and test set instead of holdout set.
- **Input Variable**
  - see Predictor.
- **Model**
  - An algorithm as applied to a dataset, complete with its settings (many of the algorithms have parameters that the user can adjust).
- **Observation**
  - The unit of analysis on which the measurements are taken (a customer, a transaction, etc.), also called instance, sample, example, case, record, pattern, or row. In spreadsheets, each row typically represents a record; each column, a variable. Note that the use of the term “sample” here is different from its usual meaning in statistics, where it refers to a collection of observations.

- **Outcome Variable**
  - see Response.
- **Output Variable**
  - see Response.
- **P (A | B)**
  - The conditional probability of event A occurring given that event B has occurred, read as “the probability that A will occur given that B has occurred.”
- **Prediction**
  - The prediction of the numerical value of a continuous output variable; also called estimation.
- **Predictor**
  - A variable, usually denoted by X, used as an input into a predictive model, also called a feature, input variable, independent variable, or from a database perspective, a field.
- **Profile**
  - A set of measurements on an observation (e.g., the height, weight, and age of a person).

- **Record**
  - see Observation.
- **Response**
  - A variable, usually denoted by Y , which is the variable being predicted in supervised learning, also called dependent variable, output variable, target variable, or outcome variable.
- **Sample**
  - In the statistical community, “sample” means a collection of observations. In the machine learning community, “sample” means a single observation.
- **Score**
  - A predicted value or class. Scoring new data means using a model developed with training data to predict output values in new data.
- **Success Class**
  - The class of interest in a binary outcome (e.g., purchasers in the outcome purchase/no purchase).
- **Supervised Learning**
  - The process of providing an algorithm (logistic regression, regression tree, etc.) with records in which an output variable of interest is known and the algorithm “learns” how to predict this value with new records where the output is unknown.

- **Target**
  - see Response.
- **Test Data (or test set)**
  - The portion of the data used only at the end of the model building and selection process to assess how well the final model might perform on new data.
- **Training Data (or training set)**
  - The portion of the data used to fit a model. Unsupervised Learning An analysis in which one attempts to learn patterns in the data other than predicting an output value of interest.
- **Validation Data (or validation set)**
  - The portion of the data used to assess how well the model fits, to adjust models, and to select the best model from among those that have been tried.
- **Variable**
  - Any measurement on the records, including both the input (X) variables and the output (Y) variable.

8. Road Maps to This Book



	Supervised		Unsupervised
	Continuous Response	**Categorical Response**	**No Response**
Continuous predictors	Linear regression (6) Neural nets (11) k-Nearest neighbors (7) Ensembles (13)	Logistic regression (10) Neural nets (11) Discriminant analysis (12) k-Nearest neighbors (7) Ensembles (13)	Principal components (4) Cluster analysis (15) Collaborative filtering (14)
Categorical predictors	Linear regression (6) Neural nets (11) Regression trees (9) Ensembles (13)	Neural nets (11) Classification trees (9) Logistic regression (10) Naive Bayes (8) Ensembles (13)	Association rules (14) Collaborative filtering (14)

Order of Topics

- Part I (Chapters 1–2) gives a general overview of data mining and its components.
- Part II (Chapters 3–4) focuses on the early stages of data exploration and dimension reduction.
- Part III (Chapter 5) discusses performance evaluation. The principles covered in this part are crucial for the proper evaluation and comparison of supervised learning methods.
- Part IV includes eight chapters (Chapters 6–13), covering a variety of popular supervised learning methods (for classification and/or prediction).
- Part V focuses on unsupervised mining of relationships. It presents association rules and collaborative filtering (Chapter 14) and cluster analysis (Chapter 15).
- Part VI includes three chapters (Chapters 16–18), with the focus on fore- casting time series: regression-based forecasting and smoothing methods.
- Part VII (Chapters 19–20) presents two broad data analytics topics: social network analysis and text mining.
- Part VIII includes a set of cases.

> Although the topics in the book can be covered in the order of the chapters, each chapter stands alone. We advise, however, to read parts I–III before proceeding to chapters in parts IV–V. Similarly, Chapter 16 should precede other chapters in part VI.

Chapter Video

References

Shmueli, G., P. C. Bruce, P. Gedeck, and N. R. Patel (2019). Data mining for business analytics: concepts, techniques and applications in Python. John Wiley & Sons.

Shmueli, G., P. C. Bruce, I. Yahav, N. R. Patel, and K. C. Lichtendahl Jr (2017). Data mining for business analytics: concepts, techniques, and applications in R. John Wiley & Sons.