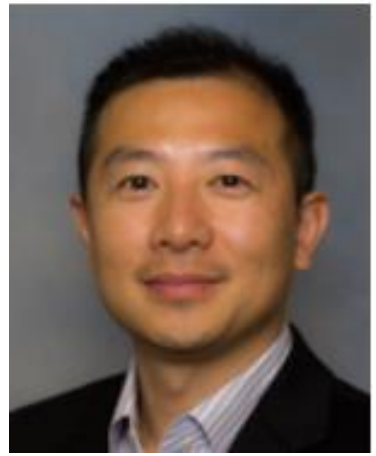
Preparing Students for Careers in Business Analytics: Innovative Teaching Approach

Sanjiv Jaggia
Professor of
Economics and
Finance
California Polytechnic
State University

Alison Kelly
Professor of
Economics
Suffolk University

Kevin Lertwachara
Professor of
Information Systems
California Polytechnic
State University



Leida Chen
Professor of
Information Systems
California Polytechnic
State University

Innovative Methods for Teaching Business Analytics

- Key career readiness skills: Mastering the technology and storytelling

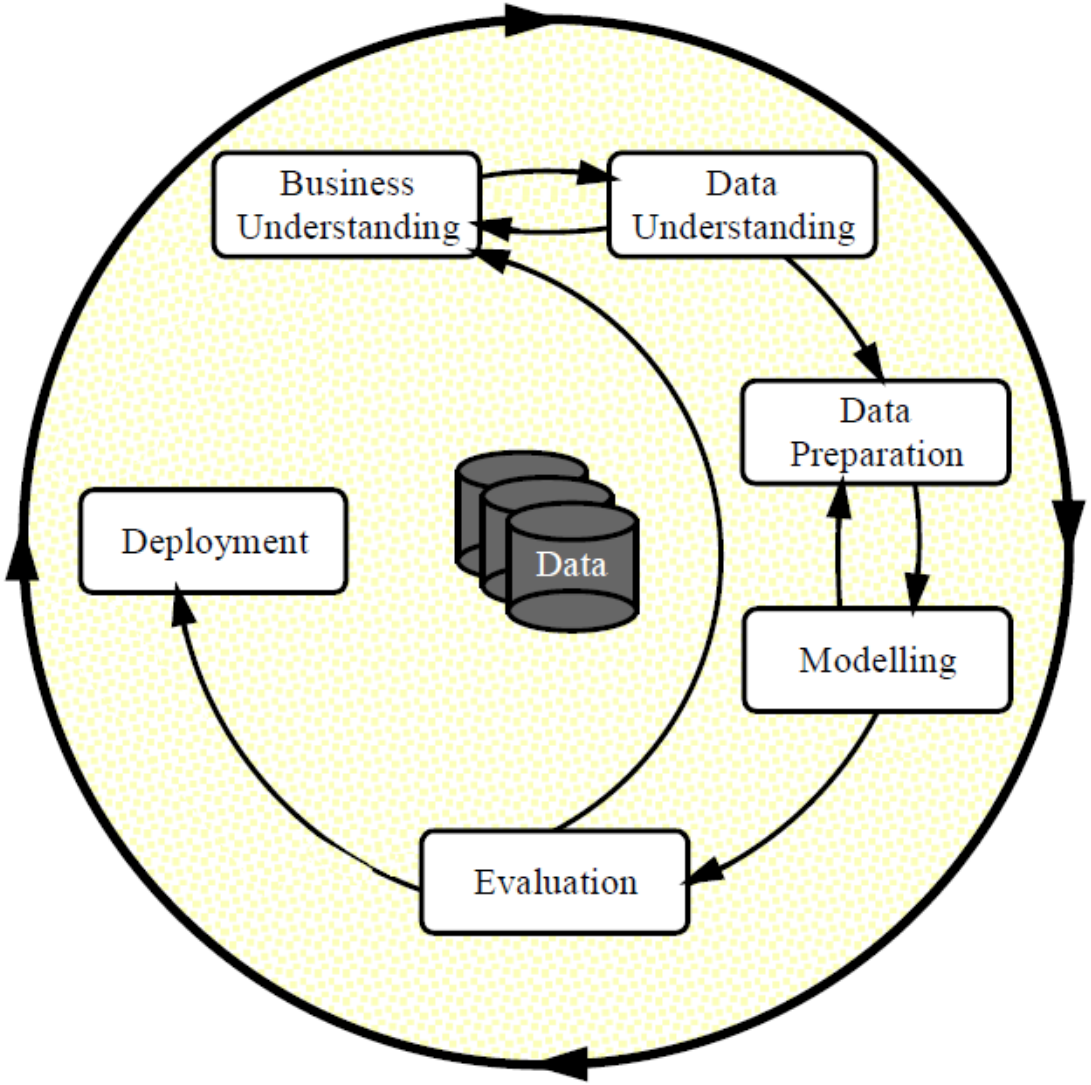
- Business analytics
  - The scientific process of transforming data into insights for the purpose of making better decisions.

- Limitations of current business analytics pedagogy
  - Heavy focus on the modeling phase only
  - Emphasis on technical analytical skillsets at the expense of storytelling
  - Students not adequately trained to deal with real life data and projects

- Infusion of the CRISP-DM framework in business analytics pedagogy
  (S. Jaggia, A. Kelly, K. Lertwachara, & L. Chen, 2020, "Applying the CRISP-DM Framework for Teaching Business Analytics," *Decision Sciences Journal of Innovative Education*, 18(4), pp. 612 – 634)

The CRISP-DM Framework

- CRISP-DM Phases:
  - Business understanding
  - Data understanding
  - Data preparation
  - Modeling
  - Evaluation
  - Deployment



Mastering Technology

CRISP-DM Phases	**Learning Objectives**
Business Understanding	Formulate business questions that lead to business strategies or actions.
Data Understanding	Describe the data in terms of the business context.
Data Preparation	Perform data wrangling to prepare the data for subsequent analyses.
Modeling	Develop analytical models to inform decision-making and select the best model(s).
Evaluation	Evaluate model performance from the business perspective.
Deployment	Communicate key findings through storytelling.

Communications and Storytelling

CRISP-DM Phases	**Learning Objectives**
Business Understanding	Formulate business questions that lead to business strategies or actions.
Data Understanding	Describe the data in terms of the business context.
Data Preparation	Perform data wrangling to prepare the data for subsequent analyses.
Modeling	Develop analytical models to inform decision-making and select the best model(s).
Evaluation	Evaluate model performance from the business perspective.
Deployment	Communicate key findings through storytelling.

Business Understanding, Data Understanding and Data Wrangling

EXAMPLE 2.1

BalanceGig is a company that matches independent workers for short-term engagements with businesses in the construction, automotive, and high-tech industries. The 'gig' employees work only for a short period of time, often on a particular project or a specific task. A manager at BalanceGig extracts the employee data from their most recent work engagement, including the hourly wage (HourlyWage), the client's industry (Industry), and the employee's job classification (Job). A portion of the *Gig* data set is shown in Table 2.3.

TABLE 2.3 Gig Employee Data

EmployeeID	HourlyWage	Industry	Job
1	32.81	Construction	Analyst
2	46	Automotive	Engineer
⋮	⋮	⋮	⋮
604	26.09	Construction	Other

- Describing business context and defining specific objectives
- Communicating project objectives that lead to business strategies and actions
- Data understanding and preparation phases tend to receive minimal coverage
- Mastering the technology: Students need expertise in using computer software such as Excel, R, and Tableau

Using Excel and R

SOLUTION:

Using Excel

a. Open the ***Gig*** data file. Note that the employee data are currently sorted by their employee ID in column A. Scroll to the end of the data set and note that the last record is in row 605. With the column heading in row 1, the data set has a total of 604 records.

b. We use two Excel functions, **COUNT** and **COUNTA**, to inspect the number of values in each column. The **COUNT** function counts the number of cells that contain numeric values and, therefore, can only apply to the EmployeeID and HourlyWage variables. The **COUNTA** function counts the number of cells that are not empty and is applicable to all four variables. Because HourlyWage is a numerical variable, we can enter either =COUNT(B2:B605) or =COUNTA(B2:B605) in an empty cell to count the number of values for HourlyWage. We get 604 values, implying that there are no missing values. Similarly, we enter =COUNTA(C2:C605) and =COUNTA(D2:D605) in empty cells to count the number of values for the Industry (column C) and Job (column D) variables. Because these two variables are non-numerical, we use **COUNTA** instead of **COUNT**. Verify that the number of records for Industry and Job are 594 and 588, respectively, indicating that there are 10 and 16 blank or missing values, respectively, in these two variables.

Using R

Before following all R instructions, make sure that you have read Appendix C ("Getting Started with R"). We assume that you have downloaded R and RStudio and that you know how to import an Excel file. Throughout the text, our goal is to provide the simplest way to obtain the relevant output. We denote all function names in **boldface** and all options within a function in *italics*.

a. Import the ***Gig*** data file into a data frame (table) and label it myData. Keep in mind that the R language is case sensitive.

b. We use the **dim** function in R to count the number of observations and variables. Verify that the R output shows 604 observations and four variables. Enter:

```
dim(myData)
```

.
.
.

g. To identify and count the number of employees with multiple selection criteria, we use the **which** and **length** functions. In the following command, we identify which employees worked in the automotive industry with the **which** function and count the number of these employees using the **length** function. The double equal sign (==), also called the equality operator, is used to check whether the industry is automotive. In R, text characters such as 'Automotive' are enclosed in quotation marks. Enter:

```
length(which(myData$Industry=='Automotive'))
```

Communicating Results

Summary

- There are a total of 604 records in the data set. There are no missing values in the HourlyWage variable. The Industry and Job variables have 10 and 16 missing values, respectively.
- 190 employees worked in the automotive industry, 536 employees earned more than $30 per hour, and 181 employees worked in the automotive industry and earned more than $30 per hour.
- The lowest and the highest hourly wages in the data set are $24.28 and $51.00, respectively. The three employees who had the lowest hourly wage of $24.28 all worked in the construction industry and were hired as Engineer, Sales Rep, and Accountant, respectively. Interestingly, the employee with the highest hourly wage of $51.00 also worked in the construction industry in a job type classified as Other.
- The lowest- and the highest-paid accountants who worked in the automotive industry made $28.74 and $49.32 per hour, respectively. In the technology industry, the lowest- and the highest-paid accountants made $36.13 and $49.49 per hour, respectively. Note that the lowest hourly wage for an accountant is considerably higher in the technology industry compared to the automotive industry ($36.13 > $28.74).

Descriptive Statistics

EXAMPLE 3.2

The marketing analyst of an online retail company is trying to understand spending behavior of customers during the holiday season. She has compiled information on 130 existing customers that includes the customer's sex and spending (in $) in the following categories: clothing (Clothing), health and beauty (Health), technology (Tech), and miscellaneous items (Misc). A portion of the data is shown in Table 3.5.

TABLE 3.5 Online Spending by Females versus Males

Customer	Sex	Clothing	Health	Tech	Misc
1	Female	246	185	64	75
2	Male	171	78	345	10
⋮	⋮	⋮	⋮	⋮	⋮
130	Male	52	73	542	58

Use Excel and R to find the average spending for each of the product categories for female customers and for male customers. Then, help the manager determine whether it seems appropriate to target females or males for the different product categories.

If we just calculated the means for each category, we would find the following:

	Clothing	Health	Tech	Misc
Mean	168.7	100.4	164.8	126.8

Using Excel and R

SOLUTION:

Using Excel:

a. Open the ***Online*** data file.

b. We use the **AVERAGEIF** function. The inputs for the function are (1) the range of cells that are to meet a certain criteria, (2) the criteria, and (3) the cells that are to be averaged. For instance, to find the average amount that females spend on clothing, we enter =AVERAGEIF(B2:B131, "Female", C2:C131), and Excel returns 225.67. To find the average amount that males spend on clothing, we enter =AVERAGEIF(B2:B131, "Male", C2:C131), and Excel returns 97.93. The averages for the other categories can be found in a similar manner. Table 3.6 summarizes the results.

Using R

a. Import the ***Online*** data file into a data frame (table) and label it myData.

b. The **tapply** function is useful for finding means (or standard deviations—discussed in Section 3.2) of subgroups. The inputs for the function are (1) the outcome variable, (2) the categorical variable for subsetting, and (3) the function to be performed (**mean** or **sd**). In order to find the average amount that females and males spend on clothing, enter:

```
tapply(myData$Clothing, myData$Sex, mean)
```

And R returns:

```
Female     Male
225.66667  97.93103
```

The averages for the other categories can be found in a similar manner.

Communicating Results

TABLE 3.6 Average Amount Spent (in $) by Females versus Males

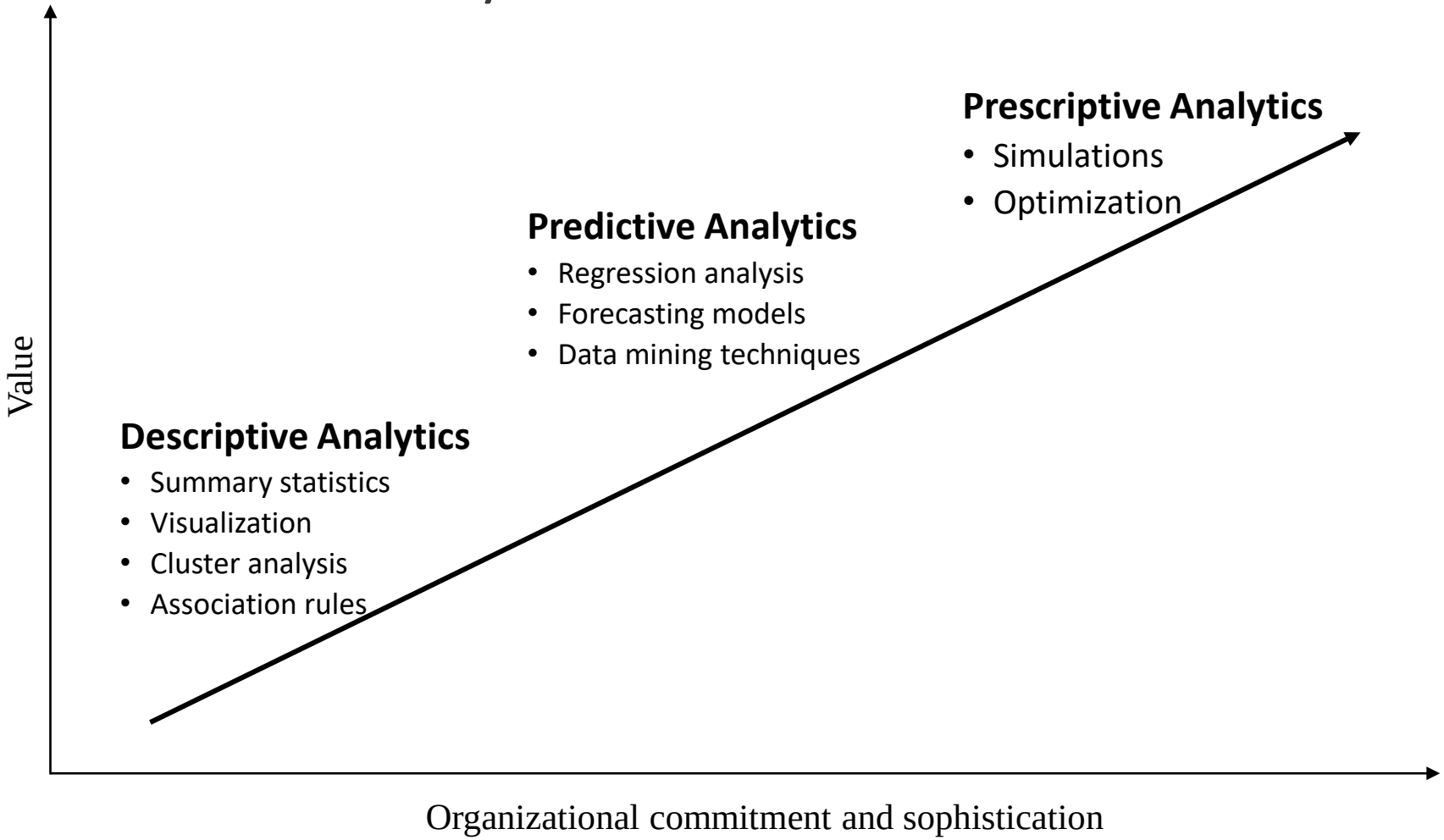
Sex	Clothing	Health	Tech	Misc
Female	225.67	100.25	47.10	159.88
Male	97.93	100.64	310.97	85.84

Summary:

Given the means for the two groups, the manager should target females for clothing and miscellaneous products and males for technology products. Because females and males spend approximately the same on health products, the manager need not differentiate this market.

Analysis Models

- Stages of Business Analytics



Analysis Models & Evaluation

INTRODUCTORY CASE

24/7 Fitness Center Annual Membership

24/7 Fitness Center is a high-end full-service gym and recruits its members through advertisements and monthly open house events. Each open house attendee is given a tour and a one-day pass. Potential members register for the open house event by answering a few questions about themselves and their exercise routine. The fitness center staff places a follow-up phone call with the potential member and sends information by mail in the hopes of signing the potential member up for an annual membership.

Janet Williams, a manager at 24/7 Fitness Center, wants to develop a data-driven strategy for selecting which new open house attendees to contact. She has compiled information from 1,000 past open house attendees in the Gym_Data worksheet of the ***Gym*** data file. The data include whether or not the attendee purchases a club membership (Enroll equals 1 if purchase, 0 otherwise), the age and the annual income of the attendee, and the average number of hours that the attendee exercises per week. Janet also collects the age, income, and number of hours spent on weekly exercise from 23 new open house attendees and maintains a separate worksheet called Gym_Score in the ***Gym*** data file. Because these are new open house attendees, there is no enrollment information on this worksheet. A portion of the two worksheets is shown in Table 12.1.

TABLE 12.1 24/7 Fitness Data

a. The *Gym_Data* Worksheet

Enroll	Age	Income	Hours
1	26	18000	14
0	43	13000	9
⋮	⋮	⋮	⋮
0	48	67000	18

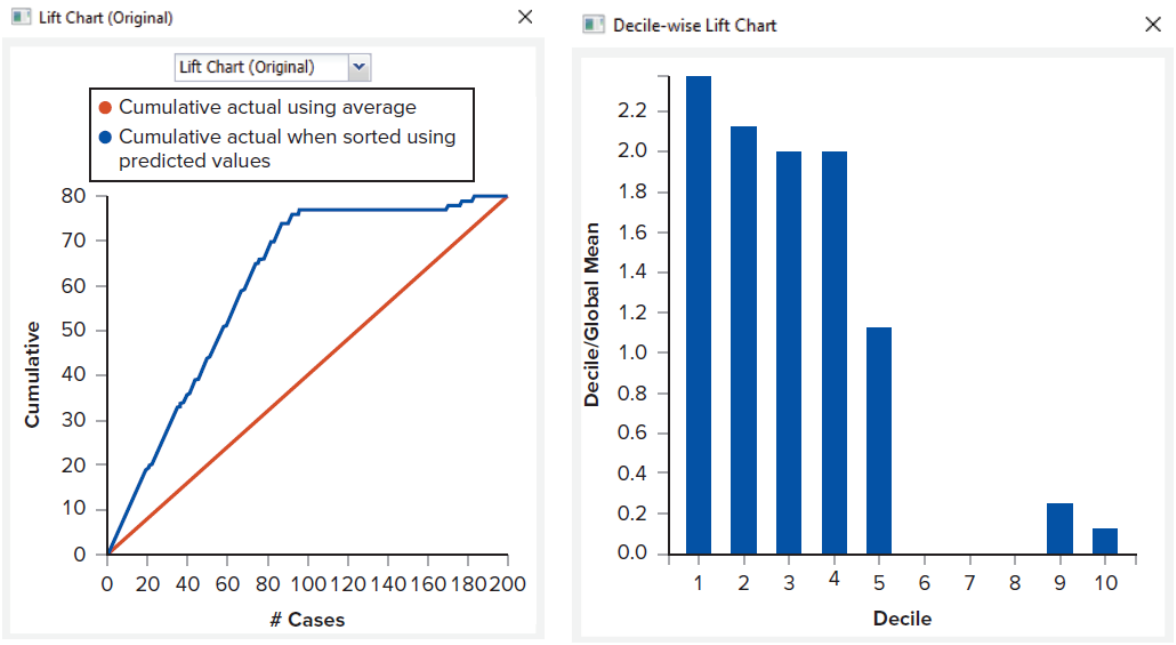
b. The *Gym_Score* Worksheet

Age	Income	Hours
22	33000	5
23	65000	9
⋮	⋮	⋮
51	88000	6

- Critical thinking required for selecting analytics techniques and models
- Mastering the computer software (e.g., Excel's Analytics Solver and R)
- Data partitioning for model assessment

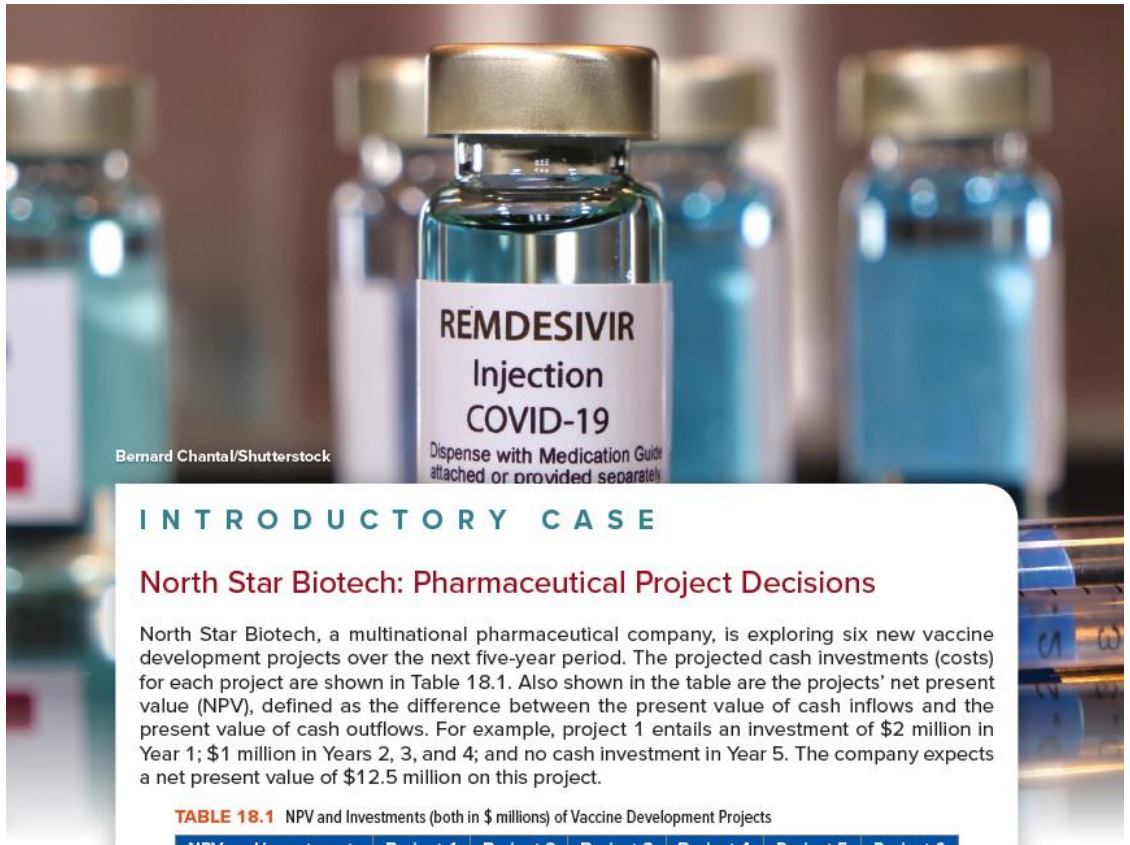
Analysis Models & Evaluation

FIGURE 12.8 Performance charts for KNN



- Reviewing and interpreting the analysis results in the context of the business objectives and success criteria
- Mastering the software for visualizing and communicating the results

Analysis Models & Evaluation



INTRODUCTORY CASE

North Star Biotech: Pharmaceutical Project Decisions

North Star Biotech, a multinational pharmaceutical company, is exploring six new vaccine development projects over the next five-year period. The projected cash investments (costs) for each project are shown in Table 18.1. Also shown in the table are the projects' net present value (NPV), defined as the difference between the present value of cash inflows and the present value of cash outflows. For example, project 1 entails an investment of $2 million in Year 1; $1 million in Years 2, 3, and 4; and no cash investment in Year 5. The company expects a net present value of $12.5 million on this project.

TABLE 18.1 NPV and Investments (both in $ millions) of Vaccine Development Projects

NPV and investments	Project 1	Project 2	Project 3	Project 4	Project 5	Project 6
Net present value	12.5	13.5	13	14.5	15	15.5
Year 1 investment	2	2	1	4	2	4
Year 2 investment	1	2	2	3	0	1
Year 3 investment	1	1	1	2	4	1
Year 4 investment	1	1	2	1	1	1
Year 5 investment	0	1	0	0	0	1

- Prescriptive analytics: Optimization
- Applications based on current events
- Managerial implications of the analysis results

Deployment

- Effective communication and storytelling play a key role
- Written and oral communications
- Critical reflection is an essential component of the experiential learning

Storytelling

- Storytelling refers to crafting and delivering compelling data-driven stories to decision makers for the purpose of converting insights into actions
- Three key questions:
  1. Why should the decision maker care about the findings?
  2. How do these findings affect the business?
  3. What actions do you recommend to the decision maker?

Self-reflection

- Critical reflection helps enhance students' understanding of the experiential activities in the context of the learning objectives.
- A self-reflective model based on three simple steps "What?," "So what?," and "Now what?" can reinforce students' learning experience.
  - What happened in the project? What was the problem being solved?
  - What other issues and opportunities arose from the project? What conclusions did you draw from the project?
  - How will you apply what you learned from the project? If you need to complete a similar project again, what would you do differently? What other skills might be beneficial to learn before you proceed to the next project?

Questions, Comments, and Suggestions?