

Applying the CRISP-DM Framework for Teaching Business Analytics



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



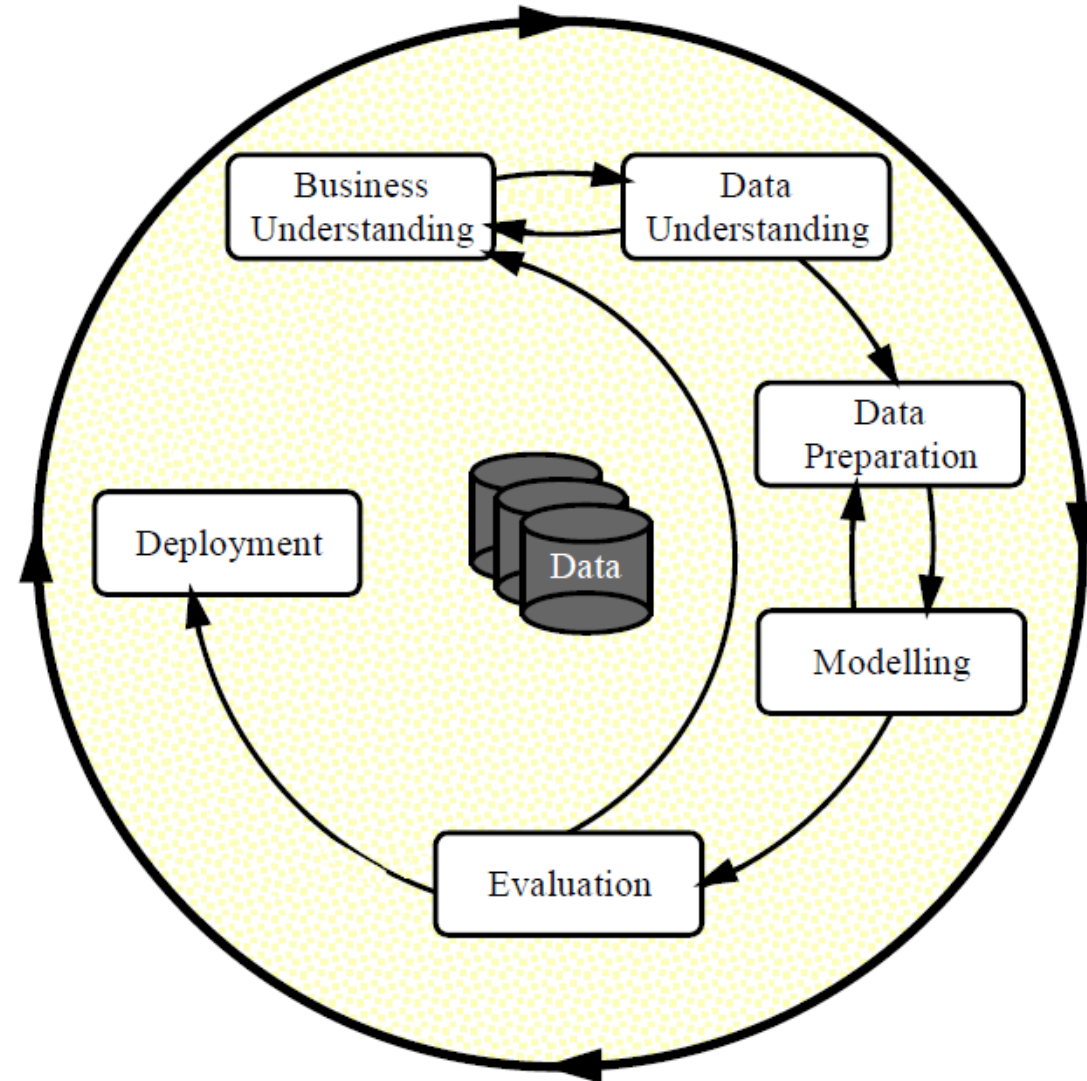
Leida Chen
Professor of
Information Systems
California Polytechnic
State University

Innovative Methods for Teaching Business Analytics

- 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

The CRISP-DM Framework

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



The Data

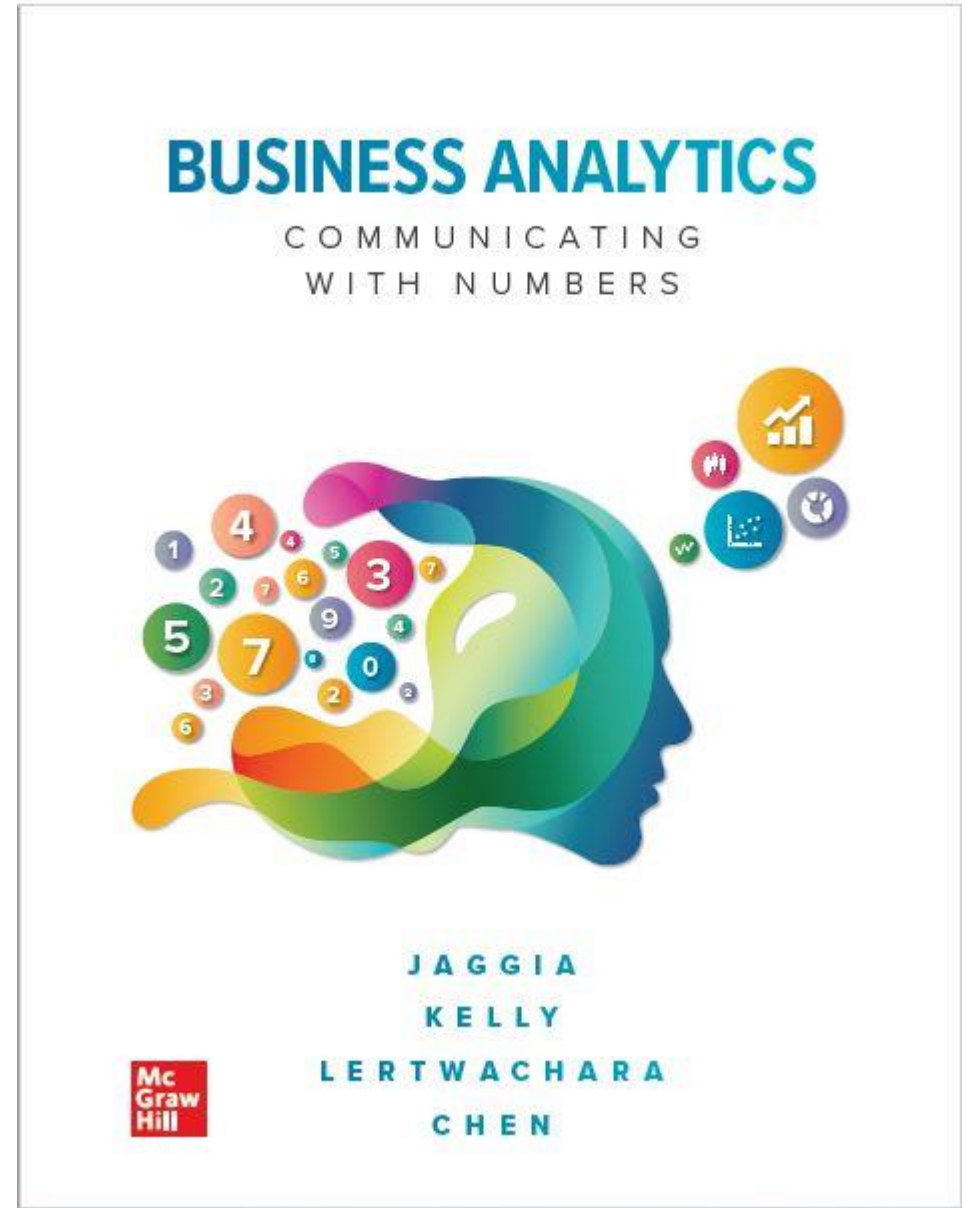
- ERIM data set provided by the James M. Kilts Center of the University of Chicago's Booth School of Business (<https://www.chicagobooth.edu/research/kilts/datasets/erim>)
- 3,189 households in two Midwestern cities and their purchases in a number of product categories (e.g. frozen dinners, yogurt, ketchup, margarine, etc.)
- Project scope:
 - One product category per team
 - Competition among teams on the same product category or categories

Project Learning Objectives

CRISP-DM Phases	Project 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 predictive models to inform decision-making and select the best predictive model(s).
Evaluation	Evaluate model performance from the business perspective.
Deployment	Communicate key findings through storytelling.

A Textbook that Applies CRISP-DM

- **Business Understanding** (Chapter 1 plus Intro Case and Big Data Cases in all chapters)
- **Data Preparation and Understanding** (Data Wrangling, Data Visualization, and Summary Measures; Chapters 2, 3)
- **Modeling** (Predictive (Chapters 6-12) and Prescriptive (Chapter 13))
- **Evaluation** (Intro Case and Big Data Cases; all chapters)
- **Storytelling** (Intro Case Synopsis and Writing with Big Data; all chapters)





Emphasis on Communication

- **Integrated Introductory Case and Synopsis**
 - ✓ Each chapter opens with a real-life case study that forms the basis for several examples within the chapter.
 - ✓ A synopsis is presented once the questions pertaining to the case have been answered.

- **Writing with Big Data**
 - ✓ Throughout the text, we use big data sets to help introduce problems, formulate possible solutions, and communicate the findings, based on the concepts introduced in the chapters.



Big Data Used

- Data 1: Car Crash Data (112,000 accidents)
- Data 2: College Admissions Data (18,000 applicants)
- Data 3: House Price Data (11,000 homes)
- Data 4: Longitudinal Survey Data (12,000 individuals)
- Data 5: NBA Data (30 teams, 457 players)
- Data 6: Tech Sales Reps Data (22,000 sales reps)

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 to open house attendees 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 9.1.

TABLE 9.1 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

Janet would like to use the data to accomplish the following tasks.

1. Develop a data-driven classification model for predicting whether or not a potential gym member will purchase a gym membership.
2. Identify which of the 23 new open house attendees are likely to purchase a gym membership.

A synopsis of this case is provided in Section 9.2.



FIGURE 9.10
Confusion matrix for KNN

Confusion Matrix and Statistics

```

Reference
prediction 0 1
0 221 17
1 17 144

Accuracy : 0.9148
95% CI : (0.883, 0.9403)
No Information Rate : 0.5965
P-Value [Acc > NIR] : <2e-16

Kappa : 0.823

McNemar's Test P-Value : 1

Sensitivity : 0.8944
Specificity : 0.9286
Pos Pred Value : 0.8944
Neg Pred Value : 0.9286
Prevalence : 0.4035
Detection Rate : 0.3609
Detection Prevalence : 0.4035
Balanced Accuracy : 0.9115

'Positive' Class : 1
    
```

SYNOPSIS OF INTRODUCTORY CASE

Gyms and exercise facilities usually have a high turnover rate among their members. Like other gyms, 24/7 Fitness Center relies on recruiting new members on a regular basis in order to sustain its business and financial well-being. Completely familiar with data analytics techniques, Janet Williams, a manager at 24/7 Fitness Center, uses the KNN method to analyze data from the gym's past open house. She wants to gain a better insight into which attendees are likely to purchase a gym membership after attending this event.

Overall, Janet finds that the KNN analysis provides reasonably high accuracy in predicting whether or not potential gym members will purchase a membership. The accuracy, sensitivity, and specificity rates from the test data set are well above 80 percent. More importantly, the KNN analysis identifies individual open house attendees who are likely to purchase a gym membership. For example, the analysis results indicate that open house attendees who are 50 years or older with a relatively high annual income and those in the same age group who spend at least nine hours on weekly exercise are more likely to enroll after attending the open house. With these types of actionable insights, Janet decides to train her staff to regularly analyze the monthly open house data in order to help 24/7 Fitness Center grow its membership base.



Detailed Software Instructions

Using R

(As discussed in Appendix C, we note that the following instructions are based on **R version 3.5.3**. They may not work for different versions of R.)

- Import the data from the *Gym_Data* worksheet into a data frame (table) and label it *myData*.
- For KNN estimation and the resulting performance measures and diagrams, install and load the *caret*, *gains*, and *pROC* packages. Enter:

```
> install.packages(c("caret", "gains", "pROC"))
> library(caret)
```

```
> library(gains)
> library(pROC)
```

On some computers, you might also need to install other packages that support the *caret* package using the command `> install.packages("caret", dependencies = c("Depends", "Suggests"))`. Also, if prompted by R Studio, install and load the *car* package.

- We use the **scale** function to standardize the Age, Income, and Hours variables; store the standardized values in a new data frame called *myData1*; and append the original *Enroll* variable back to *myData1*. We use the **as.factor** function to convert the target variable (*Enroll*) into a categorical data type. To simplify the R code, we use the **colnames** function to rename `myData1$myData.Enroll` (in column 4) to `myData1$Enroll`. Enter:

```
> myData1 <- scale(myData[2:4])
> myData1 <- data.frame(myData1, myData$Enroll)
> colnames(myData1)[4] <- 'Enroll'
> myData1$Enroll <- as.factor(myData1$Enroll)
```

SOLUTION:

Using Analytic Solver

As discussed in Chapters 7 and 8, to develop and evaluate a classification model, we generally perform cross-validation using either the hold-out method or the *k*-fold method. Analytic Solver provides a procedure for the hold-out method and allows for partitioning a data set into training, validation, and test data sets. In this chapter, we partition our data set in Analytic Solver as follows: 50% for training, 30% for validation, and 20% for test. Here, an independent assessment of the predictive performance of the KNN model is conducted with the test data set that is not used in the model development.

- Open the *Gym* data file and go to the *Gym_Data* worksheet.
- Choose **Data Mining > Partition** (under the *Data Mining* group) > **Standard Partition**.

FIGURE 9.2

Analytic Solver's standard data partition

Data Source	
Worksheet:	Gym
Workbook:	Gym.xlsx
Data range:	\$A\$1:\$D\$1001
#Rows:	1000
#Cols:	4

Variables	
<input checked="" type="checkbox"/> First Row Contains Headers	
Variables In Input Data	Selected Variables
	Enroll
	Age
	Income
	Hours



Dedicated Chapter on Data Wrangling

Topics include:

- Key concepts related to data management
- Data inspection
- Binning, subsetting, and treatment of missing values and outliers
- Transformation of numeric variables
- Transformation of categorical variables

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

The manager suspects that data about the gig employees are sometimes incomplete, perhaps due to the short engagement and the transient nature of the employees. She would like to find the number of missing observations for the HourlyWage, Industry, and Job variables. In addition, she would like information on the number of employees who (1) worked in the automotive industry, (2) earned more than \$30 per hour, and (3) worked in the automotive industry and earned more than \$30 per hour. Finally, the manager would like to know the hourly wage of the lowest- and the highest-paid employees at the company as a whole and the hourly wage of the lowest- and the highest-paid accountants who worked in the automotive and the tech industries.

Use counting and sorting functions in Excel and R to find the relevant information requested by the manager, and then summarize the results.



Emphasis on Data Mining

- In addition to two chapters on linear and logistic regression and a chapter on forecasting, there are four exclusive chapters on data mining, including
 - ✓ Introduction to data mining: distance measures, performance evaluation, and principal component analysis (PCA)
 - ✓ Supervised learning: k-nearest neighbors (KNN) and naïve Bayes
 - ✓ Supervised learning: classification and regression trees, and ensemble trees
 - ✓ Unsupervised learning: hierarchical and k-means clustering, and association rules
- Over 200 exercises in these four exclusive chapters.



Different ways to use the Text

- One Term

- ✓ Business Statistics with Analytics Flavor (Chapters 1, 2, 3, 4, 5, 6, 7, 12)
- ✓ Holistic Approach to Business Analytics (Chapters 1, 2, 3, 6, 7, 8, 9, 10, and 11)

- Two Terms

- ✓ Term 1: Chapters 1, 2, 3, 4, 5, 6, 12
- ✓ Term 2: Chapters 7, 8, 9, 10, 11, 13

- Applicable for Undergraduate and Graduate Courses



Using the Text for Online Instruction

- **Example: undergraduate business analytics course**
 - ✓ Coverage: Chapters 1, 2, 3, 6, 7, 8, 9, 10, and 11
 - ✓ Students use asynchronous resources (e.g., reading SmartBook, watching recorded lectures) and take quizzes prior to synchronous online meetings on Zoom
 - ✓ With detailed software instructions in the text, synchronous meetings can focus more on coaching, consultation, and troubleshooting
 - ✓ During synchronous meetings, students can discuss introductory cases and examples beyond interpretation of results described in the Text



Asynchronous Resources

- McGraw Hill's SmartBook and ReadAnywhere mobile app (in addition to printed text)
- Recorded lectures based on concepts and examples in the Text
- All data sets and software instructions are available online
- Homework, quizzes, and practice problems on McGraw Hill's Connect
 - ✓ Instructor-defined options (e.g., time limit, attempts, score reduction, access to hints/eBook and solutions,)
 - ✓ Performance reports by assignment and by student on Connect



McGraw Hill's Connect

- Fully engaged with the development of the Connect product
- Separate questions for Analytic Solver and R, when needed
- Careful with the choice of:
 - Tolerance limits
 - Algorithmic exercises

Questions, Comments, and Suggestions?

