Section 1.4. The Cross-Industry Standard Process for Data Mining: CRISP-DM

Note. In this section, we outline "CRISP-DM" by presenting the six phases of it graphically and in outline form.

Note. The Cross-Industry Standard Process for Data Mining (CRISP-DM) is an "industry-neutral" data mining process; that is, it is not specific to any specific type of data (sales data, political poll data, health-related information, etc.) but is a model that applies to non-industry-specific data. It is the most widely-used analytics model (according to What IT Needs To Know About The Data Mining Process in *Forbes*, July 29, 2015 [accessed 6/16/2021]). The CRISP-DM (or simply "CRISP") consists of six phases. It is *adaptive* in the sense that each phase depends on the output of the previous phase. See Figure 1.1.

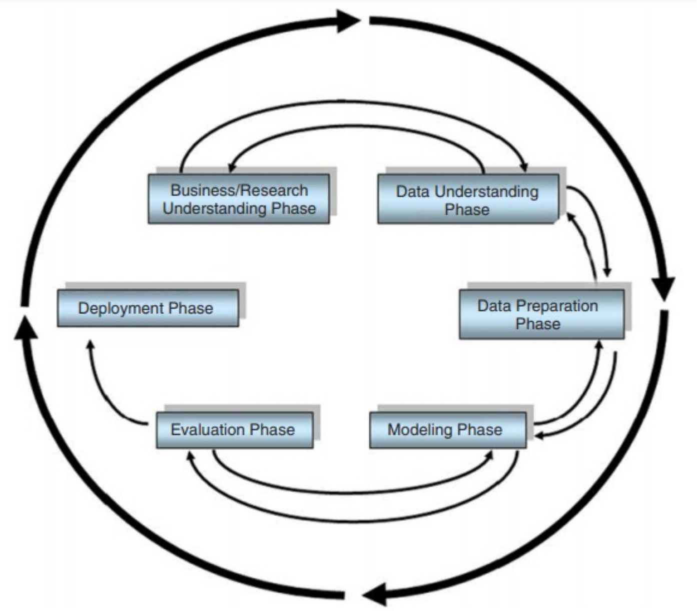

Figure 1.1. CRISP-DM is an iterative, adaptive process.

Notice that we may move "backwards" to a previous phase for "refinement" before moving forward. The outer path of Figure 1.1 reflects the iterative nature of CRISP; the solution of one problem may lead to additional questions which may also be addressed CRISP.

Note. The six phases of CRISP-DM, with some details, are (quoting from the text, pages 7 and 8):

1. Business/Research Understanding Phase

- **a.** First, clearly enunciate the project objectives and requirements in terms of the business or research unit as a whole.
- **b.** Then, translate these goals and restrictions into the formulation of a data mining problem definition.
- **c.** Finally, prepare a preliminary strategy for achieving these objective.

2. Data Understanding Phase

- **a.** First, collect the data.
- **b.** Then, use exploratory data analysis to familiarize yourself with the data, and discover initial insights.
- **c.** Evaluate the quality of the data.
- **d.** Finally, if desired, select interesting subsets that may contain actionable patterns.

3. Data Preparation Phase

- **a.** This labor-intensive phase covers all aspects of preparing the final data set, which shall be used for subsequent phases, from the initial, raw, dirty data.
- **b.** Select the cases and variables you want to analyze, and that are appropriate for your analysis.
- **c.** Perform transformations on certain variables, if needed.

   d. Clean the raw data so that it is ready for the modeling tools.

4. Modeling Phase

   a. Select and apply appropriate modeling techniques.

   b. Calibrate model settings to optimize results.

   c. Often, several different techniques may be applied for the same data mining problem.

   d. May require looping back to data preparation phase, in order to bring the form of the data into line with the specific requirements of a particular data mining technique.

5. Evaluation Phase

   a. The modeling phase has delivered one or more models. These models must be evaluated for quality and effectiveness, before we deploy them for use in the field.

   b. Also, determine whether the model in fact achieves the objectives set for it in phase 1.

   c. Establish whether some important facet of the business or research problem has not been sufficiently accounted for.

   d. Finally, come to a decision regarding the use of the data mining results.

6. Deployment Phase

   a. Model creation does not signify the completion of the project. Need to make use of created models.

   b. Example of a simple deployment: Generate a report.

   c. Example of a more complex deployment: Implement a parallel data mining process in another department.

   d. For businesses, the customer often carries out the deployment based on your model.

The text book largely follows CRISP-DM, with some small modifications to be introduced later.

Revised: 6/16/2021