II. The Data Science Process

Introduction to Data Science

Patricio Mallea[1]

[1]Data Scientist

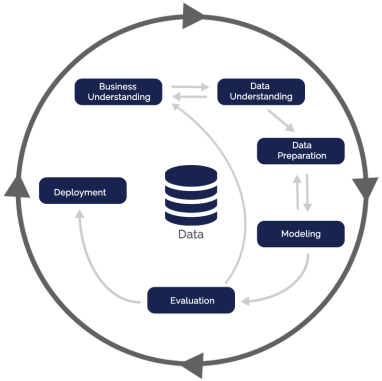
May 11, 2020

Overview I

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment

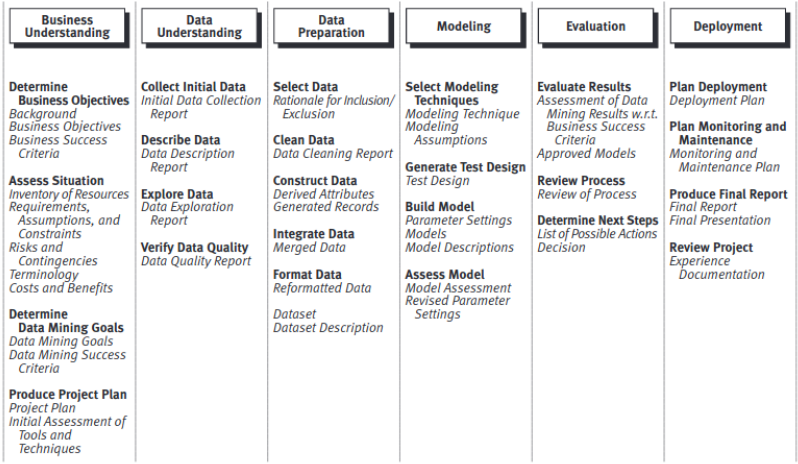
The Data Science Procces: An approaching

DataOwl



The **CRISP-DM** methodology (*Cross-Industry Standard Process for Data Mining*) is a lifecycle process. Consists of six phases with arrows indicating the most important and frequent dependencies between phases. The sequence of the phases is not strict. In fact, most projects move back and forth between phases as necessary.

CRISP-DM Methodology



1. Business Understanding

1. Business Understanding

DataOwl

This initial phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

1. Determine business objectives.
2. Assess situation.
3. Determine data mining goals.
4. Produce project plan.

The results of the Business Understanding phase can be summarized in one report.

2. Data Understanding

2. Data Understanding

DataOwl

The data understanding phase starts with initial data collection and proceeds with activities that enable you to become familiar with the data, identify data quality problems, discover first insights into the data, and/or detect interesting subsets to form hypotheses regarding hidden information.

1. Collect initial data.
2. Describe data.
3. Explore data.
4. Verify data quality.

The results of the Data Understanding phase are usually documented in several reports. Ideally, these reports should the written while performing the respective tasks. The reports describe the datasets that are explored during data understanding. For the final report, a summary of the most relevant parts is sufficient.

3. Data Preparation

3. Data Preparation

The data preparation phase covers all activities needed to construct the final dataset [data that will be fed into the modeling tool(s)] from the initial raw data. Data preparation tasks are likely to be performed multiple times and not in any prescribed order. Tasks include table, record, and attribute selection, as well as transformation and cleaning of data for modeling tools.

1. Select data.
2. Clean data.
3. Construct data.
4. Integrate data.
5. Format data.

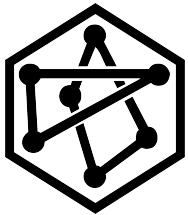
The reports in the data preparation phase focus on the pre-processing steps that produce the data to be mined. Dataset description report. This report provides a description of the dataset (after pre-processing) and the process by which it was produced.

4. Modeling

4. Modeling

In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, going back to the data preparation phase is often necessary.

1. Select modeling technique.
2. Generate test design.
3. Build model.
4. Assess model.



Also, the outputs produced during the Modeling phase can be combined into one report.

5. Evaluation

5. Evaluation

It's important to thoroughly evaluate it and review the steps executed to create it, to be certain the model properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.

1. Evaluate results.
2. Review process.
3. Determine next steps.

In the Assessment of data mining results with respect to business success criteria, the report compares the data mining results with the business objectives and the business success criteria.

6. Deployment

6. Deployment

The deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise. In many cases, it is the customer, not the data analyst, who carries out the deployment steps. However, even if the analyst will carry out the deployment effort, it is important for the customer to understand up front what actions need to be carried out in order to actually make use of the created models.

1. Plan deployment.
2. Plan monitoring and maintenance.
3. Produce final report.
4. Review project.

This report specifies the deployment of the data mining results. The final report is used to summarize the project and its.

II. The Data Science Process

Introduction to Data Science

Patricio Mallea[1]

[1]Data Scientist

May 11, 2020