

# GENERAL FRAMEWORK FOR DATA MINING PROJECTS

*Guideline for Implementation of Crisp-DM Framework On Data Mining Projects*

**TEAM B:**

ALAIN ARTURO GRULLÓN GONZALEZ

ALBERTO DE RONI

AYŞE UMUT VAROL

FRANCISCO MANSILLA NAVARRO

ROSAMARIA MEJIA GARCIA

TIMO BACHMANN

## Table of Contents

Abbreviations.....	4
List of Figures.....	5
List of Tables.....	6
Document Objectives .....	7
1. Introduction.....	8
2. Methodology .....	9
2.1 Business Understanding.....	10
2.1.1 Determine Business Objectives.....	10
Output 1: Background.....	10
Output 2: Business Objectives .....	10
Output 3: Business Success criteria .....	11
2.1.2 Assess Situation.....	13
Output 4: Inventory of resources.....	13
Output 5: Requirements, Assumptions, and Constraints .....	15
Output 6: Risks and Contingencies .....	16
Output 7: Terminology.....	16
Output 8: Costs and benefits .....	17
2.1.3 Determine Data Mining Goals.....	19
Output 9: Data Mining Goals .....	19
Output 10: Data Mining Success Criteria .....	20
2.1.4 Produce Project Plan.....	21
Output 11: Project Plan.....	21
Output 12: Initial Assessment of Tools and Techniques .....	24
2.2 Data Understanding.....	25
2.2.1 Collect Initial Data .....	25
Output 1: Initial Data Collection Report .....	26
2.2.2 Describe Data .....	27
Output 2: Data Description Report .....	27
2.2.3 Explore Data.....	28
Output 3: Data Exploration Report .....	28
2.2.4 Verify Data Quality.....	28
Output 4: Data Quality Report.....	29
2.3 Data Preparation.....	30

2.3.1 Select Data .....	31
Output 1: Rationale for inclusion/exclusion .....	31
2.3.2 Clean Data .....	31
Output 2: Data Cleaning Report.....	33
2.3.3 Construct Data .....	33
Output 3: Derived attributes.....	33
Output 4: Generated records.....	34
2.3.4 Integrate Data .....	34
Output 5: Merged Data.....	35
2.3.5 Format Data .....	36
Output 6: Reformatted Data .....	36
2.4 Modelling .....	36
2.4.1 Select Modelling Techniques .....	36
2.4.2 Generate Test Design .....	37
2.4.3 Build Model .....	38
2.4.4 Assess Model.....	39
2.5 Evaluation and Presentation of results .....	40
2.5.1 Evaluate Results .....	40
Output 1: Assessment of data mining results with respect to business success criteria.....	40
Output 2: Approved models .....	42
2.5.2 Review Process.....	42
Output 3: Review of process.....	42
Output 4: Presentation .....	43
2.5.3 Determine Next Steps .....	44
Output 5: List of possible actions.....	44
Output 6: Decision .....	45
2.6 Deployment .....	45
2.6.1 Plan Deployment.....	45
Output 1: Deployment Plan .....	45
2.6.2 Plan Monitoring and Maintenance .....	47
Output 2: Monitoring and Maintenance Plan.....	47
2.6.3 Produce Final Report.....	48
Output 3: Final Report .....	48
Output 4: Final Presentation.....	48

2.6.4 Review Final Project .....	49
Output 5: Experience Documentation .....	49
Bibliography .....	50

## Abbreviations

<b>ABT</b>	: Analytical Base Table
<b>AI</b>	: Artificial Intelligence
<b>CRISP-DM</b>	: Cross-Industry Standard Process for Data Mining
<b>DB</b>	: Data Base
<b>Dept.</b>	: Department
<b>DS</b>	: Data Scientist
<b>Etc.</b>	: Et cetera
<b>GDPR</b>	: General Data Protection Regulation
<b>HR</b>	: Human Resources
<b>Mgmt.</b>	: Management
<b>Ops.</b>	: Operations
<b>Recomm.</b>	: Recommendation
<b>Vs.</b>	: Versus

## List of Figures

Figure 1: CRISP-DM Data Mining Lifecycle .....	8
Figure 2: Generic tasks (bold) and outputs (italic) of the CRISP-DM reference model .....	8
Figure 3: On-Premise Computing vs. IaaS vs PaaS vs SaaS.....	14
Figure 4: Data mining success criteria example .....	21
Figure 5: Hierarchical breakdown of the WBS.....	22
Figure 6: Sample responsibility matrix for a data mining project .....	23
Figure 7: Sample and cut project timeline.....	23
Figure 8: Sample project communication plan.....	24
Figure 9: Attributes and key relationships.....	28
Figure 10: Data preparation stage tasks and outputs .....	30
Figure 11: Methods of cleaning data .....	32
Figure 12: Missing and outlier data (red) illustration example in Dataiku .....	32
Figure 13: Common attributes on where to join data tables .....	34
Figure 14: Clean joined data set from various data tables.....	35
Figure 15: Example of formatting data features such as date.....	36
Figure 16: Review of process questioner.....	43
Figure 17: Suggestion of percentile brackets for threshold analysis and corresponding actions .....	46

## List of Tables

Table 1: Step-by-step examination of business situation.....	10
Table 2: Illustration of exemplary SMART goals for BEES Airlines.....	11
Table 3: Sample success criteria for BEES Airlines.....	12
Table 4: Basic business profiles needed for a generic analytical question.....	13
Table 5: Template for availability of glossaries .....	16
Table 6: Template for understanding the terminology .....	16
Table 7: Costs of data collection .....	17
Table 8: Benefits of the project .....	18
Table 9: Sample fields from ROI Calculation Excel .....	18
Table 10: Examples for possible business objectives and corresponding data mining goals.....	19
Table 11: Necessary data and cross-validation parties / related resources.....	26
Table 12: Sample guide for choosing modeling technique.....	37
Table 13: Sample metrics for different algorithm classes .....	38
Table 14: Status of models .....	40
Table 15: Visual understanding guidelines .....	41
Table 16: Recommendation spreadsheet.....	42
Table 17: Approved models.....	42
Table 18: Decision based on list of actions.....	45

## **Document Objectives**

This internal handbook includes guidelines, information and examples on the implementation of widely used and successful framework of CRISP-DM (The CRISP-DM Consortium, 1999) on data mining projects within BEES Airlines by the data professionals.



## 1. Introduction

CRISP-DM which stands for “Cross-Industry Standard Process for Data Mining” is a complete framework including Data Mining Lifecycle for guiding data mining projects (IBM, n.d.).

CRISP-DM was chosen over the other methodologies such as KDD and SEMMA since:

- ✓ it incorporates business perspective successfully,
- ✓ enables aligning efforts around a technical goal that serves the business objective,
- ✓ it is flexible and can be customized easily, (IBM, n.d.)

CRISP-DM has 6 main phases displaying the data mining lifecycle as:

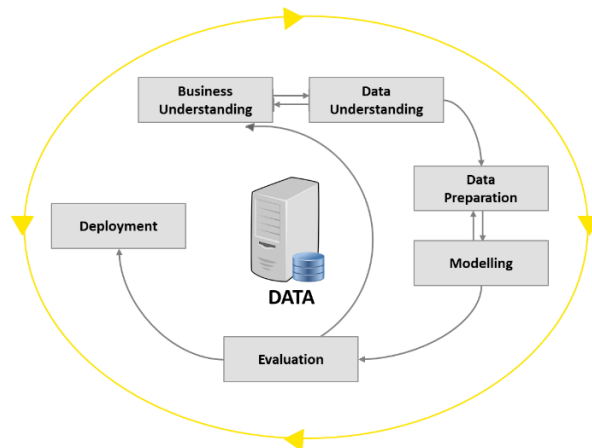


Figure 1: CRISP-DM Data Mining Lifecycle  
(The CRISP-DM consortium, 2000, p. 10)

Note that the arrows indicate the most critical dependencies between phases yet, these phases are not strictly sequential which means that teams can move forward and back between the phases based on project’s needs (IBM, n.d.).

Phases, main tasks and their outputs to be detailed in this document are structured in the following way:

Business Understanding	Data Understanding	Data Preparation	Modelling	Evaluation	Deployment
<b>Determine Business Objectives</b> <ul style="list-style-type: none"> <li>Background</li> <li>Business Objectives</li> <li>Business Success Criteria</li> </ul> <b>Assess Situation</b> <ul style="list-style-type: none"> <li>Inventory of Resources</li> <li>Requirements, Assumptions, and Constraints</li> <li>Risks and Contingencies</li> <li>Terminology</li> <li>Costs and Benefits</li> </ul> <b>Determine Data Mining Goals</b> <ul style="list-style-type: none"> <li>Data Mining Goals</li> <li>Data Mining Success Criteria</li> </ul> <b>Produce Project Plan</b> <ul style="list-style-type: none"> <li>Project Plan</li> <li>Initial Assessment of Tools and Techniques</li> </ul>	<b>Collect Initial Data</b> <ul style="list-style-type: none"> <li>Initial Data Collection Report</li> </ul> <b>Describe Data</b> <ul style="list-style-type: none"> <li>Data Description Report</li> </ul> <b>Explore Data</b> <ul style="list-style-type: none"> <li>Data Exploration Report</li> </ul> <b>Verify Data Quality</b> <ul style="list-style-type: none"> <li>Data Quality Report</li> </ul>	<b>Select Data</b> <ul style="list-style-type: none"> <li>Rationale for Inclusion / Exclusion</li> </ul> <b>Clean Data</b> <ul style="list-style-type: none"> <li>Data Cleaning Report</li> </ul> <b>Construct Data</b> <ul style="list-style-type: none"> <li>Derived Attributes</li> <li>Generated Records</li> </ul> <b>Integrate Data</b> <ul style="list-style-type: none"> <li>Merged Data</li> </ul> <b>Format Data</b> <ul style="list-style-type: none"> <li>Reformatted Data</li> <li>Dataset</li> <li>Dataset Description</li> </ul>	<b>Select Modeling Techniques</b> <ul style="list-style-type: none"> <li>Modeling Technique</li> <li>Modeling Assumptions</li> </ul> <b>Generate Test Design</b> <ul style="list-style-type: none"> <li>Test Design</li> </ul> <b>Build Model</b> <ul style="list-style-type: none"> <li>Parameter Settings</li> <li>Models</li> <li>Model Descriptions</li> </ul> <b>Assess Model</b> <ul style="list-style-type: none"> <li>Model Assessment</li> <li>Revised Parameter Settings</li> </ul>	<b>Evaluate Results</b> <ul style="list-style-type: none"> <li>Assessment of Data Mining Results with respect to Business Criteria</li> <li>Approved Models</li> </ul> <b>Review Process</b> <ul style="list-style-type: none"> <li>Review of Process</li> </ul> <b>Determine Next Steps</b> <ul style="list-style-type: none"> <li>List of Possible Actions</li> <li>Decision</li> </ul>	<b>Plan Deployment</b> <ul style="list-style-type: none"> <li>Deployment Plan</li> </ul> <b>Plan Monitoring and Maintenance</b> <ul style="list-style-type: none"> <li>Monitoring and Maintenance Plan</li> </ul> <b>Produce Final Report</b> <ul style="list-style-type: none"> <li>Final Report</li> <li>Final Presentation</li> </ul> <b>Review Project</b> <ul style="list-style-type: none"> <li>Experience Documentation</li> </ul>

Figure 2: Generic tasks (bold) and outputs (italic) of the CRISP-DM reference model  
(The CRISP-DM consortium, 2000)

## 2. Methodology

This section includes detailed information about the major phases and sub-processes, tasks, activities, and outputs of the CRISP-DM framework;

- Business Understanding
- Data Understanding
- Data Repreparation
- Modelling
- Evaluation
- Deployment

## 2.1 Business Understanding

The business understanding phase intends to set a foundation for the data mining project in which it proposes professionals to follow a bottom-up approach. The methodology begins with a basic understanding of the business setting and the problem to solve. The conversion of such knowledge into data mining goals forms another vital component (Wirth & Hipp, 2000). The first version of the project plan is to be drafted at the end of this phase.

### 2.1.1 Determine Business Objectives

The first task of determining the business objectives is all about the quest for the right questions to assess customers' needs. With the primary aim to comprehend the fundamental business problem, this handbook suggests evaluating the company's background, listing explicit business objectives, and pinning down some business success criteria.

#### Output 1: Background

To begin with, we must lay down a basis for the data mining project. In order to provide the first overview, current business situation should be examined in the areas suggested in Table 1.

Organizational Charts	Key Positions	Internal Sponsors	Steering Committee	Affected Business Units
-----------------------	---------------	-------------------	--------------------	-------------------------

*Table 1: Step-by-step examination of business situation  
(The CRISP-DM consortium, 2000, p. 32)*

First, we create organizational charts to recognize the company-internal structures and to assign responsibilities within it. Then, we list key individuals in the business and determine both primary users and financial sponsors for better understanding of needs and expectations for the project. In case there exists a steering committee, its members should be clearly specified. Lastly, all business units affected by the project need to be highlighted. The steps explained above intend to establish a status-quo and underline the actual purpose of the data mining project (The CRISP-DM consortium, 2000).

Following this setup, we can narrow down the pain points and related business areas in the organization. After the concerned division has been identified, the problem should be briefly characterized and expected benefits should be listed. It should become evident which project requirements are already satisfied and what further resources must be gathered. Moreover, it will clarify the hierarchical level of the future reports to delivered, namely the target groups. If there is already an existing solution in place, we need to define the solution and evaluate its benefits and drawbacks (The CRISP-DM consortium, 2000).

#### Output 2: Business Objectives

Building upon the outlined background; effective business objectives, related questions, and additional business conditions are to be defined. Business objectives are normally split into primary business objectives and complementary secondary objectives (The CRISP-DM consortium, 2000). For instance, a primary goal could be to "employ adaptive pricing in times of higher demand", whereas a secondary aim could be to "identify customer groups that are relatively more strongly affected by such dynamic price changes". Mentioned business conditions, for example, could be "to generally maintain market share throughout the project's implementation".

At this stage, we suggest coming up with SMART goals, i.e. specific-measurable-attainable-relevant-timely goals (Rubin, 2002). As we see a particular value in shaping the main goals in a realistic and assignable manner, the acronym was amended to our cause. A draft table could look as follows:

SMART goals	Specific	Measurable	Assign-able	Realistic (budget)	Time-related (deadline)
1	Employ adaptive pricing to balance out changes in demand	KPI: Sales more stable; define a range of +/- 50% for weekly average sales fluctuation rate	IT Dept.	\$50,000	Beginning of August
2	Reduce missing luggage on transfer flights through improved tracking	KPI: Lost Luggage per Traveler < 1%	Project Task Force	\$30,000	End of Q3
3	...	...	...	...	...

Table 2: Illustration of exemplary SMART goals for BEES Airlines

### Output 3: Business Success criteria

Any analytical project must undergo a scrutinous evaluation before being launched and used in day-to-day operations. At this stage, we should develop a list of success/failure criteria to assess the ability of our model(s) to answer our analytical question(s). Mind that, even though CRISP-DM is a revolving process that allows professionals to circle back to a previous stage if needed, this list should be given special attention as its metrics are the ones that is going to determine whether the project is “working” or not.

The first thing to keep in mind when thinking of these criteria, is whether our analytical questions need an objective or subjective answer. While the first are rather straightforward and can be assigned quantitative metrics to be evaluated, the latter will generally require a degree of domain knowledge, meaning we have to pick carefully the most appropriate evaluators and evaluation criteria.

#### 3.1 - Objectively measurable criteria

The first kind of evaluation criteria are rather easy to shape up: They fall into a quasi-pass/fail kind of realm which are mathematically measurable. Many analytical questions will tend to be a good fit for these criteria. If something is quantifiable, it can be assigned a metric to later be evaluated such as ratios, percentage shares, etc. Naturally, to assess whether a given value is to be deemed as ‘good’ or not, there needs to be a pre-set threshold for any given criteria. The following table shows three airline-related examples to provide context:

Case	Good Metric	Decent Metric	Bad Metric	General Notes
<b>1. Personalized marketing campaign</b>	% change in spending of targets after contact	Yes/No binary metric – Did subjects book a flight after contact?	% change of time spent looking at tickets (may be useful if used along others for context)	There might have been non-marketing related factors influencing people's buying decision
<b>2. Adaptive ticket pricing algorithm</b>	% change in ticket revenue	% change in # of tickets sold	Passenger satisfaction survey	Best is not to let passengers know that you are systematically trying to have them pay more
<b>3. Luggage Handling (Missing Bag Rates)</b>	Plane hold to conveyor belt time – find the weak spot	% of lost bags per route	# of lost bags per day	Just knowing you are losing a lot of bags without context won't help you solve anything

Table 3: Sample success criteria for BEES Airlines

In short, there are several right answers to the same question. The key is finding the best combination of metrics to make sure we can act upon our answer, therefore, by connecting the dots we will get the best insights and evaluation possible. The important things to keep in mind “What do we want to achieve with this project?” and “How can we measure that?” when setting your criteria. The metrics needed will start to shape up once we mix those these point views.

It would be beneficial to have some quantitatively oriented people to check the numbers for this kind of criteria. Notwithstanding the need for a solid base of domain knowledge, once the metrics and thresholds are agreed upon by the team, the evaluation part is rather straightforward.

### 3.2 - Subjectively measurable criteria

There are things that cannot be quantified as being positive or negative by means of calculations. If the scope of an analytical process is to develop new insight; how is one supposed to “measure insight”? Imagine our company was scouting for new destinations by analytical means, would the number of passengers passing through a given airport every year be meaningful for the decision-making process? Wouldn't it make sense to, first, think of kinds of customers who would want to fly to that destination, kinds of customers that are included our base, or the purpose of people's visit for that particular destination. For this kind of situations, the key to success is in choosing the right person(s) to evaluate the output.

Here, domain knowledge would play a big part in allowing us to determine whether a given piece of information can be acted upon and made profitable.

In practice, it would make sense to have both domain experts and data scientists working together to develop a meaningful metric. However, Data Scientist's scope would fall more into the ‘following the orders’ part of the scale in this situation. The domain expert would have to clearly convey to DS what kind of information would be ideal to evaluate a given solution. The DS on the other hand, would have to deliver a meaningful output that is easy for the domain expert to understand and evaluate. Developing this kind of metrics would highly require shared effort among the team members. Nonetheless, there is great potential in developing insightful models able to give winning edge in problems that may not otherwise (by means of raw, objective calculations) be identifiable.

### 2.1.2 Assess Situation

Once the direction of the project has been clearly identified, we need to understand “how we plan to get there”. In this phase, we need to clearly outline the resources available for the project and the ones needed to be acquired in order to succeed. These resources fall into two main categories: personnel and technology. We need to ask:

- ✓ Do we have all the manpower needed?
- ✓ Do we have the right tools?
- ✓ What may be the biggest barriers to the project in terms of resources?
- ✓ Do we have a contingency plan for the problems that may arise?

Also, it is vital to ensure all team members are aware of the importance of cross-department communication and cooperation at this stage.

### Output 4: Inventory of resources

In this first sub-section, we need to identify the needs of the team may arise throughout the project: people, data, hardware and software.

**Human resources:** Generally speaking, a small team of around 5 people could take care of the whole analytical process but, what are the key profiles needed to succeed as a team? Table 4 aims to outline the requirements of a team for a generic analytical question.

Bus. Profile	Reason for Necessity	Task	Contextual Examples	General Notes
<b>Business Experts (various company dept.s)</b>	Domain knowledge is needed as a base for analytics. Finding the right department match for analytical question is important	Collaborate with data scientists in drawing up evaluation metrics and suggest actions upon the newly generated information	<ul style="list-style-type: none"><li>· Marketing dept. for personalized ads</li><li>· Ops. dept. for luggage handling</li><li>· Finance dept. for adaptive pricing</li></ul>	In most cases, you will find these profiles within the firm
<b>Data Analysts</b>	Very specific domain, not an easy to find skill. There is no analytics without an analyst.	Generating insight from sheer data and cross-department collaboration to ensure success of analytical quest + Choosing the languages/tools to run analytics with	<ul style="list-style-type: none"><li>· Data manipulation and presentation</li><li>· Building the analytical model</li></ul>	Training may take a long time, generally best to acquire new talent – mind outsourcing though, these roles may very well become permanent
<b>Technical Support (IT dept.)</b>	General support for technical problems	Hardware/software integration + Collaboration with/support to data experts and project team as a whole and making sure hardware is functional	<ul style="list-style-type: none"><li>· Help data miners in structuring databases</li><li>· Making sure non-technical profiles understand what they need to</li></ul>	As for business expertise, IT skills will generally not be the rarest inside the company, meaning the talent just needs to be picked for the project
<b>Data Miners</b>	Working along with analysts to develop data architecture, structure DBs, ensure data quality	Making sure analysts are always working with clean, up-to-date, insightful, quality data, and that databases are running smoothly	<ul style="list-style-type: none"><li>· Data cleaning and preparation</li><li>· DB structuring</li><li>· Overseeing data value chain along with analysts</li></ul>	Similar case to data analysts: Training vs. Outsourcing

Table 4: Basic business profiles needed for a generic analytical question

**Data Necessities:** Once we know who will be needed to form the parts of the project, it is time to figure out what kind of data we are after and in which format, and how do we plan to have it reached to the analysts.

There are four main kinds of data sources (public/private data, social network data, internal company generated data and data from official organizations) and we will generally find ourselves in one of two situations: either the data will be available to us, or it will not. In the first case, we just have to start developing the data infrastructure; and in the second, we will first have to draw up a *data acquisition plan*.

For the scope of our guide (analytics within an airline), the data will generally be available to us as internal data (Prada, 2020). However, it will mostly come in the form of excel spreadsheets, which are not ideal for analytics, or SQL databases, which would be better suited, but with little to no pre-processing. This means we will come across missing values, redundant columns, useless logs tables, etc. Here, a joint effort by analysts and miners, along with the rest of the team when needed, will eventually generate an analytics-dedicated database.

**Software/Hardware necessities:** *(For this section, we interviewed a member of an analytics team within a Spanish regional airline to be able to give some real-world examples)*

There are a few concepts to keep in mind when developing the architecture planned to use for the analytical process. Firstly, keep in mind that the company probably already uses web structures for some processes outside of analytics. This means that to have everything run smoothly, we might want to consider “locking” ourselves with a single vendor. To give an example: Amazon Web Services is commonly viewed as one of the best providers of cloud computing, but if the whole system runs on Microsoft products, it would make more sense to go for Azure instead. Additionally, a way to smoothly handle large amounts of data will be needed. Hadoop will generally hold the answers to the data structure problems.

Once we have “holding facilities” set up, we must think of how we are planning to run the analytical process. Rstudio is widely considered a powerful analytical workbench with a huge potential for any quantitative task and should be able to accommodate most if not all our needs. Also, for the sake of ease of communication and streamlining our analytics process, we might want to consider using software such as Jupyter Notebook. Doing this would allow our data experts to more seamlessly share, review and edit each other's codes.

Naturally, all of this has no chance of happening without some powerful hardware behind it.

Here, once the vendor locking issues are solved, we will have to consider what option suits our solution the best. Would we want all the hardware to be on premise, or perhaps would it be best to go for an all-cloud setup? The image below gives an idea of the differences.

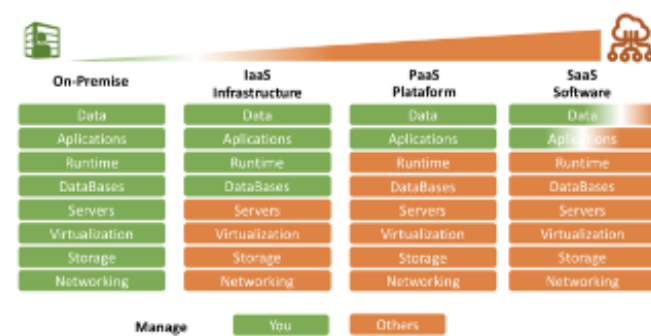


Figure 3: On-Premise Computing vs. IaaS vs PaaS vs SaaS  
(Lozano, 2020, p. 142)

While going for a SaaS solution may have its advantages, company officers may be worried of data privacy and cybersecurity issues. While data can be anonymized, cybersecurity will always be an issue, whether our servers are on the cloud or right next to us.

Nonetheless, it is worth keeping in mind that, at the very start of our analytics journey, we will likely not be able to use data to its full potential. This means that we are going to have a lot of extra computing power not being used (i.e. having 2/5 servers running, due to low necessities). In this situation, we may want to consider moving more of our resources on cloud (Prada, 2020).

All in all, the key to success is finding the solution that best fits the skills of our personnel and our analytical necessities.

## **Output 5: Requirements, Assumptions, and Constraints**

---

### **Requirements:**

For business objectives to be carried out successfully, the following requirements must serve as a guide. First, the target group profile must be identified and specified with the objective of further understanding of the purposes and analysis of the data. The schedule of completion must guide the team in terms of time available for completion of every task in this handbook, including the building of said schedule (i.e. a Gantt chart). The human resources mentioned above, including the business experts, IT department, data miners, data analysts must work together to ensure that the business objectives, data mining and outputs are comprehensible for highest quality results. They will all be held accountable for said comprehensibility and quality of results. Security will be non-negotiable, as all these human resources above mentioned, are expected to work under and comply with the IT Departments guidelines regarding cybersecurity. Above all, legal aspects regarding data usage must be in line with the company's strict compliance with the General Data Protection Regulation (GDPR). The data usage must be legally cleared by both the IT and Legal Departments before any work is performed.

### **Assumptions:**

Any assumption necessary to carry out the project must be clearly stated at every stage of the project, before and after data mining if the assumption pertains to the data, as well as before and after identifying the business objectives and criteria. Usually there will be assumptions pertaining to the data and other assumptions pertaining to the business aspects. Specifically, for the data assumptions, there will usually be some of them that can be verifiable during the data mining or data analysis, all verified assumptions as well as assumed verified assumptions must be documented as such. The data miners, analysts and business experts are accountable for identifying and documenting any assumptions, assumed verified assumptions, or actual verified assumptions made during the process. The list of these must be documented, as they may enter in the evaluation stage due to their nature of affecting the validity of results.

### **Constraints:**

All constraints must also be listed accordingly, they may affect: the outputs, comprehensibility, amount of assumptions, and quality of analysis and results. These constraints may be related to the project budget, human resources, technological resources, data sizes, etc. Regardless of the nature of the constraint, these must be communicated to the IT Department and the higher executive in charge of the departments related to the needed resource acting as the constraint, before executing the project plan. For example, the financial



department must be involved closely with the team to define and follow budgeting constraints so that the project has the monetary resources necessary for completion. The involved department may be able to provide access to the required resource that is acting as a constraint and a better project plan may be achievable. Regardless of the outcome, the constraints and related department must be documented for future use (The CRISP-DM consortium, 2000).

## Output 6: Risks and Contingencies

Risks or any event that might take the project on an undesired path to failure, must be brainstormed and analyzed thoroughly by all parties involved, business experts, data science experts, and the IT Department before determining mining goals and the project plan. Understanding and documenting any possible event that could cause failure will allow the team to plan accordingly and avoid possible mistakes. A list of risks, along with possible actions or a “Plan B” that could serve to mitigate the situation caused by any possible event, must be stated, communicated and documented, and further analyzed along the next all steps of the project. Hope for the best, but plan for the worst.

## Output 7: Terminology

With this output we will try to make a glossary relevant to the project, by checking the availability of previous glossaries. Check last updates and draft new ones if necessary.

- **Activity 1: Check availability of glossaries**

Check how many glossaries we have regarding business understanding; how can we obtain them. If there are glossaries available, complete the following spread sheet with all available attributes (columns).

Glossary	Author	Digital/ Printed	Departments of interest	Where is it stored	Date of 1 <sup>st</sup> publication	Last update	Lang.
Name	Name and dept., if available	Digital	Check organigram to specify the departments that can be included	Database and accessibility	dd.mm.yyyy	Important to check for new concepts	EN FR ES

Table 5: Template for availability of glossaries

If no glossaries are available, start to draft one as shown in next activity

- **Activity 2: Understanding of glossaries and terminology**

Understand the content of each glossary: Read the terminology of the glossary. Whenever a concept is not clear, talk to the domain experts and add a business example of that definition. The glossary will be updated by the domain expert.

Terminology	Definition	Other related concepts	Business Example
Name of the concept	Explanation of the concept	<ul style="list-style-type: none"> <li>· Concept 1</li> <li>· Concept 2</li> </ul>	Short explanatory example related to the terminology

Table 6: Template for understanding the terminology

- **Activity 3: Repeat activities (1) and (2) this time with data mining terminology. In total we will have 4 components.**

## Output 8: Costs and benefits

The aim of this output is to clearly describe the costs of the project and predicted business benefits in a successful scenario to help with decision-making (The CRISP-DM consortium, 2000, p. 35).



The key of developing this kind of analysis is its scalability to other projects and the advantages are numerous such as developing benchmarks for comparing projects, final decision of pursuing a proposed project, measuring social benefits or even quantifying effects on stakeholders among others (Smartsheet Inc., n.d.)

- **Activity 1: Estimate costs for data collection**

A detailed estimation of costs must be included into the following spreadsheet, marking with an x the category to which each cost corresponds. The total will be automatically computed, and statistical analysis will be automatically performed for each category.

Cost Name*	Cost Category		Year1	Year 2	Year3	TOTAL
	Fixed					
	Direct	Indirect				
Name1	x		Value1	Value11	Value13	SumValues
Name2		x	Value2	Value22	Value23	SumValues
	Variable					
	Direct	Indirect				
Name3		x	Value3	Value32	Value33	SumValues
Name4		x	Value4	Value42	Value43	SumValues
	Sunk / Hidden					
Name5			Value5	Value52	Value53	SumValues

Table 7: Costs of data collection  
(Utrecht University, n.d.)

\*The most important considerations for data collection costs are the following:

- Acquiring external datasets: Research repositories
- Granularity of data: What level of detail we want from data
- Format of data: Is the format uniform across all datasets?
- Transcription time cost: Additional software, transcription guidelines
- Transferring data: Encryption software, use of different devices, etc.
- Documentation: Description of data gathering, creation and quality control
- Data back up: How often do we need backups and how much storage do we need?
- Data security (Trusted Third party)
- Preservation of data: Conversion to standard or open format
- Legal issues: Anonymization of data. Anonymization done before data collection will result in lower costs; Copyright of data
- Data cleaning: In time with big datasets
- Digitization: Transformation of analogue data

- Hidden costs: New data requirements after iteration of the methodology, changes in models, data types, etc. (Utrecht University, n.d.)

- **Activity 2: Estimate costs of developing and implementing the solution**

We need to identify the costs and categorize them. We need to estimate the rest of the costs of the project. We need to fill the spreadsheet with the suitable costs of the project (same format as the one used on the previous activity). A more detailed calculation of costs must be done in separate spreadsheets.

\*The most important considerations for other costs are the following: Inventory, materials, manufacturing, direct labor personnel, including airplanes, cabin crew, etc.

- Maintenance and supervision costs: Hardware, software
- Downtime costs, crucial for our airline, as is one of the critical costs.
- Indirect costs: Electricity, rent, etc.
- Hidden costs: Training time during learning, hiring new specialists, etc.
- Devaluation costs
- Opportunity costs: alternative investments, build instead of buy
- Cost of potential risks such as regulatory risks, competition, and environmental impacts (Will, 2019).

- **Activity 3: Estimate benefits for our solution**

Identify the benefits and translate them into quantitative measures. For this purpose, the level of assumptions must be as detailed as possible. The spreadsheet will allow to have different scenarios defined in risks and contingency plans.

Benefit Name	Year1	Year 2	Year3	TOTAL
Cost Reduction				
Revenue Increase				
Customer Experience				

Table 8: Benefits of the project

- **Activity 4: Determine the impact that the project will have on the company**

Review the following spreadsheet, that automatically will give us the ROI and the Payback Time. If other metrics were to be included during the initial project plan, update the automatic spreadsheets with the metrics required.

Year	Investment	Revenue	Costs	Profit	Cum. Profit	Cash Flow	Cum. Cash Flow
Year1							
Year2							
Year3							

ROI		Payback Time	
-----	--	--------------	--

Table 9: Sample fields from ROI Calculation Excel

### 2.1.3 Determine Data Mining Goals

The **aim** of this process step is to;

- ✓ **re-interpret the business objective in technical, data mining terms** considering the resources in hand, constraints, assumptions and other factors that may affect the project **to come up with “data mining goals”**;
- ✓ and **specify “data mining success criteria”** which can be considered as a set of parameters to measure success of the data mining practices supporting the data mining goals aligned with the business objective.

The **outputs** of this process step to be documented are:

- **Data Mining Goals**
- **Data Mining Success Criteria**

#### Output 9: Data Mining Goals

We need to determine a “technical vision statement” for the data mining project which “enables the achievement of the business objective” (Lozano, 2020).

Although this process step is often overlooked by project managers, study including more than 1,400 project managers have shown that “50% of the planning problems relate to unclear definition of scope and goals” (Larson & Gray, 2011, p. 102).

To be able to develop a solid project plan based on the right goal, following activities should be performed to produce the output properly:

- **Activity 1: Determine the data mining goals referred by the business questions**  
We need to think about the projection of analytical questions asked while discovering the business objective on the data mining studies in order to find the path that will take us to the business objective.

Business Objective	Data Mining Goal
Decrease the number of lost luggage in transfer flights	Calculate risk rate of being mis-handled for the luggage based on time, route, airport, connection times and handing process based on 6 years of past data (Bonthu & Bindu, 2017)
Employ adaptive pricing to balance out changes in demand	Calculate customized price points for each potential customer in real time based on route, location, seasonality, market demand and price level, customer segment, personal information (alexsoft, 2019)
Reduce missing luggage on short-distance flights through improved tracking	Calculate an index for ranking the routes and airports according to probability of bags being lost

Table 10: Examples for possible business objectives and corresponding data mining goals

- **Activity 2: Classify the data mining problem**

We need to determine the class our data mining problem belong to in order to get a better glimpse of the methods we can use and the requirements related to these methods (e.g. data requirements, technological requirements, time estimation for project planning, etc.)

Namely, data mining functionality can be broken into 4 main areas:

- **Classification:** “Classification involves finding a model which describes distinct finite data classes, which can then be used to classify instances of unknown data” (Mayo, 2016)
- **Regression:** “Involves building a model to predict a continuous numeric data” (Mayo, 2016)
- **Clustering:** “Clustering involve analyzing and grouping data which does not include pre-labeled classes by maximizing intraclass similarity and minimizing the similarity between differing classes” (Mayo, 2016)
- **Frequent Pattern Mining:** “Involves applying statistical methods to find interesting and previously-unknown patterns within said set of data. We might investigate which patterns emerge frequently, which items are associated, and which items correlate with others” (Mayo, 2016)
- **Outlier Detection:** “Involves determination of the data points that do not seem to readily fit the behavior of the remaining data or a resulting model. This method can be useful for fraud detection, fault detection, etc.” (Mayo, 2016)

## Output 10: Data Mining Success Criteria

We need to determine a set of key performance indicators in order to assess the level of success from data mining perspective, like in the business success criteria.

These criteria should define the measures and their corresponding levels such that “a project member might consider the project successful when the level was achieved” (Nemati & Barko, 2003).

Following activities should be performed during this sub-process:

- **Activity 1: Determine the success criteria for the data mining output**

We need to set right assessment parameter to assess the performance of the data mining solution: Although data mining success criteria is defined for technical area, the outcome / solution provided with this technical practice may affect related stakeholders in different terms such as time consumption (velocity) and complexity (easiness of using) (Nemati & Barko, 2003). Thus, it is recommended to take main objectives and stakeholders’ expectations into account to be able to reflect them in the data mining success criteria.

In this step, we also need to lay out the calculation / measurement methodology for each data mining success criteria determined.

- **Activity 2: Define benchmark for each business criteria**

We need a baseline or/and a target level across the criteria in order to make a judgement about the level of success: In simple terms, to be able to answer the question of “Compared to what this project is not / - successful?”. We might obtain target level / baseline from different channels:

- Direct imputation of a pre-determined target level by the top management
- Within and cross-industry benchmarks obtained from external resources
- Comparison versus the baseline (current level, if exists)

- **Activity 3: Specify subjective data mining success criteria and identify the related responsible**  
There might be few business criteria that require subjective assessment such as “insight provided by the model” or “easiness of using the interface”. In this case, we need to determine the corresponding assessors and agree on the assessment methodology for each subjective data mining success criteria.

Data Mining Success Criteria: Prediction Model Accuracy and Deviation		
Definition and Methodology		
Deviation is a measure of model's tendency to systematically over/under-predict; where, WAPE is a measure for the magnitude of prediction error in percentages.		
$\text{Deviation} = \frac{\sum_i (\text{Realized Value} - \text{Predicted Value})}{\# \text{ of Data Points}}$	$\text{Weighted Absolute Percentage Error (WAPE)} = \frac{\sum_i \text{Realized Value (i)} * \text{MAPE(i)}}{\sum_i \text{Realized Value (i)}}$	
$\text{Mean Absolute Percentage Error MAPE(i)} = \frac{ \text{Realized Value} - \text{Predicted Value} }{\text{Predicted Value}}$		
Responsible	Assessor	Stakeholders
<ul style="list-style-type: none"> <li>Project Team</li> <li>Sales &amp; Distribution</li> <li>Information Technology &amp; Services</li> </ul>	<ul style="list-style-type: none"> <li>Steering Committee</li> <li>Board of Management</li> </ul>	<ul style="list-style-type: none"> <li>Sales &amp; Distribution</li> <li>Product &amp; Service Development</li> <li>Strategy &amp; Business Development</li> <li>Finance &amp; Accounting</li> <li>Information Technology &amp; Services</li> </ul>
Reporting and Assessment Frequency	Reporting Level	Respective Target Levels
<b>Reporting:</b> Monthly <b>Assessment:</b> Monthly	<ul style="list-style-type: none"> <li>SKU</li> <li>Product segment</li> <li>Customer segment</li> <li>Sales channel</li> </ul>	<ul style="list-style-type: none"> <li>81%</li> <li>85%</li> <li>90%</li> <li>95%</li> </ul>

Figure 4: Data mining success criteria example

#### 2.1.4 Produce Project Plan

The **aim** of this process step is **to lay out a detailed, structured roadmap** that would lead the team towards the common business goal, within the defined project scope.

The **outputs** of this process step to be documented are:

- **Project Plan**
- **Initial Assessment of Tools and Techniques**

#### Output 11: Project Plan

An ideal project plan should;

- ✓ involve deliverables, milestones, tasks and responsibility allocations, effort / time and cost estimates, resources and stakeholder communication plan;
- ✓ and, take time, budget and technical constraints into account.

Following activities should be performed to build an extensive project plan:

- **Activity 1: Define work breakdown structure**  
Following the definition of scope and deliverables, the project can be divided into smaller, hierarchical work packages named “work breakdown structure (WBS)”. These work packages constitute “the basic unit used for planning, scheduling, and controlling the project” (Larson & Gray, 2011, p. 112).

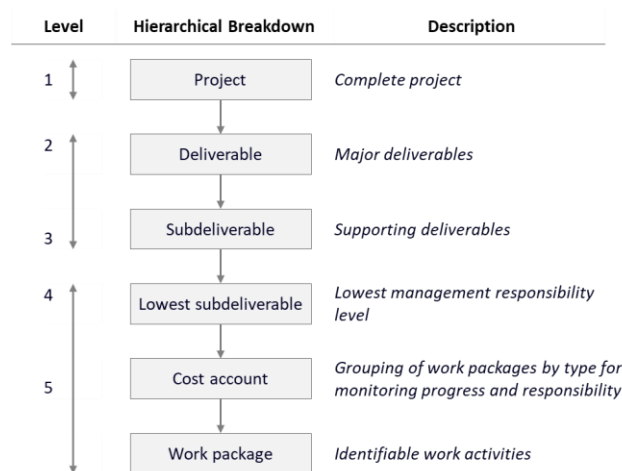


Figure 5: Hierarchical breakdown of the WBS  
(Larson & Gray, 2011, p. 108)

The hierarchy draws a guide to analyze the project (level 1) in terms of deliverables (level 2, 3) with a output-oriented approach, and creates manageable work elements (level 4) including set of related activities (level 5) which are assignable specifically to team members.

Work packages are the “short duration of tasks with defined start-end points, consume resources and represent cost” (Larson & Gray, 2011). The work packages defined should be mutually exclusive and collectively exhaustive (MECE) and should be as independent as possible. They should clearly define (Larson & Gray, 2011):

- The work (what?)
- Time for completion (How long?)
- Resources (both internal and external) needed (How much?)
- A single accountable person and responsible (Who?)
- Monitoring points for measuring progress (How well?)

- **Activity 2: Perform responsibility allocation**

It is important to clearly define the responsibility allocation of the activities including:

- **Responsible:** “The person who completes (performs) and coordinates the work to complete the activity” (Harned, 2019)
- **Accountable:** “The person who delegates the work and is the last one to review the task or deliverable before it’s deemed complete, with “Yes/No/Veto” rights for the related activity” (Harned, 2019)
- **Consulted:** “The person who provides input based on either how it will impact their future work or their domain of expertise on the deliverable itself” (Harned, 2019)
- **Informed:** “The person who will be updated on decisions and actions during the process step” (Harned, 2019)

Hierarchy	Task	Steering Committee	Project Manager	Cost Account Resp.	Team Member 1	Team Member 2	Team Member 3	Team Member 4	Team Member 5	Sales & Distribution	IT
1.1.3.4.2.1	Obtain historical sales data	A	A	R	R	I	I	I	I	C	C
1.1.3.4.2.2	Analyze properties of different attributes of historical sales data	I	I	A	R	R	C	C	C	C	C
1.1.3.4.2.3	Form hypothesis to feed / shape data mining goals	A	A	C	R	R	-	-	-	C	C
1.1.3.4.2.4	Perform analysis for hypothesis testing and report findings	A	A	C	R	R	I	I	I	C	C
1.1.3.4.2.5	Prepare Data Exploration Report	I	I	A	R	R	-	-	-	-	-
1.1.3.4.2.6	Discuss the results of Data Exploration Report and reshape data mining goals or set direction for further steps	C	R	R	R	R	R	R	R	C	C
1.1.3.4.2.7	Obtain historical sales data	A	A	R	R	I	I	I	I	C	C

Figure 6: Sample responsibility matrix for a data mining project

- **Activity 3: Visualize Detailed Project Time Plan**

After building up the WBS in detail, overall project time plan should be prepared;

- stating activities and durations, precedence / succession relationships between the activities, deliverable due dates, milestones, decision points;
- taking risks and constraints into account;
- including the whole project team throughout the planning process.

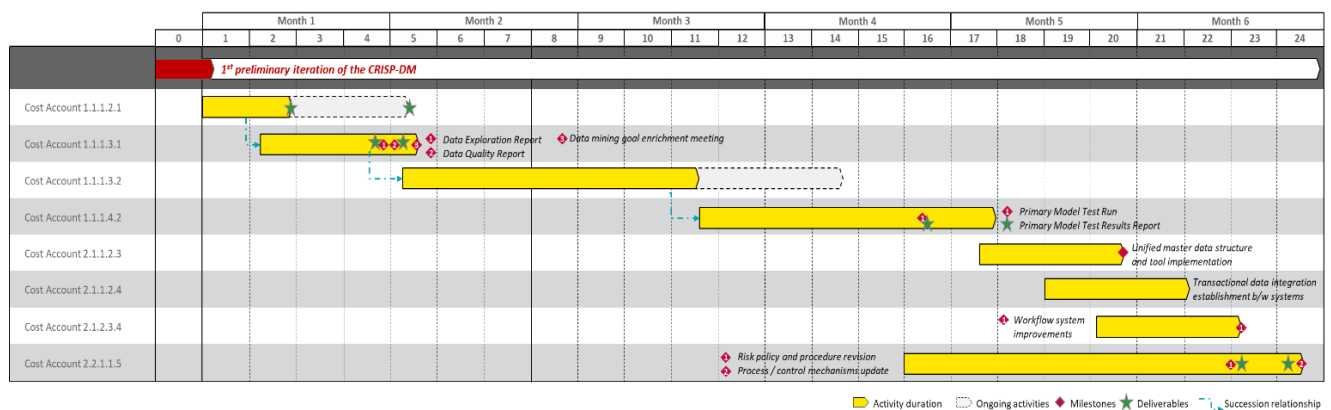


Figure 7: Sample and cut project timeline

- **Activity 4: Prepare stakeholder communication plan**

Once the project deliverables and work plan are prepared, we need to set up a communication plan to:

- “mitigate project problems”,
- “ensure that team members and stakeholders have the information to perform necessary tasks” (Larson & Gray, 2011),
- We need to analyze the stakeholders, their jurisdiction areas and decide: “WHAT, WHO, HOW and WHEN information will be transmitted to project stakeholders so schedules, issues, and action items can be tracked” (Larson & Gray, 2011)



Project communication plan is often prepared by the project manager and addresses:

- Which information should be collected and when?
- Who is going to receive the information?
- Which methods should be used for gathering and storing information?
- Are there any access restrictions / confidentiality for the information for some departments / people?
- When the information should be transmitted?
- How should the information be communicated?
- Who is going to send out the information? (Larson & Gray, 2011)

Deliverable Code	Information	Target Audience	Frequency	Mean of Communication	Provider
1	Milestone Report	Senior Mgmt. and Project Manager	Bimonthly	E-mail and hardcopy	Project office
2	Project Status Report & Agenda	All stakeholders	Weekly	E-mail and hardcopy	Project manager
3	Team Status Report	Project manager and project office	Weekly	E-mail	Team recorder
4	Issues Report	All stakeholders	Weekly	E-mail	Team recorder
5	Escalation Report	All stakeholders	When needed	Meeting and hardcopy	Project manager
6	Outsourcing Performance Report	All stakeholders	Bimonthly	Meeting	Project manager
7	Oversight Gate Decisions	Senior Mgmt. and Project Manager	As required	E-mail meeting report	Overseight group and project office

*Figure 8: Sample project communication plan  
(Larson & Gray, 2011, p. 117)*

## Output 12: Initial Assessment of Tools and Techniques

After the preparation of the detailed project plan, we need to assess the main and supportive tools and techniques to be used during the project.

Following activities should be performed for a healthy assessment of tools and techniques:

- **Activity 1: Create list of potential selection criteria for tools and techniques**

We need to identify the right potential tools and techniques for further assessment and comparison since we need to come up with the optimal one subject to our constraints.

If there exists a list available, it might be used directly, or after modification according to data mining goals.

Some assessment criteria might involve:

- **Functionality alignment with the data mining goal**
- **Security**
- **Hardware resources requirements**
- **Integrability**
- License requirements
- Scalability
- Level of commercial support

- Continuity / sustainability
- Language used for development
- Easiness of use
- Level of expertise (Bonthu & Bindu, 2017)

Despite listing the detailed assessment criteria, some of the points that can be characterized as primary filter, can be used for selection of the potential tools and techniques for detailed assessment.

It also might be helpful to define weights for each criterion at this stage by asking “Which criteria should be prioritized for the decision making for the success of the project and its sustainability?”.

- **Activity 2: Choose potential tools and techniques and perform assessment of tools and techniques**  
Following the determination of potential tools and techniques, we need to assess each tool and technique in terms of pre-determined assessment criteria.

At this stage, it is crucial to list the advantages, disadvantages and information regarding each criterion correctly. During this process, involvement of the project stakeholders (especially the IT and other technical teams) can be crucial and beneficial for appropriate assessment “since the selection of tools and techniques may influence the entire project” (Lozano, 2020, p. 80).

- **Activity 3: Prioritize tools and techniques**

In this step, we need to prioritize the tools and techniques based on the assessment performed.

It would be beneficial to ensure alignment with the stakeholders about the prioritization.

## 2.2 Data Understanding

The **aim** of this process step is to;

- ✓ **gather** initial **data**,
- ✓ **understand** the **data structure, types, attributes** and **relationships**,
- ✓ **identify** data **quality problems**,
- ✓ **explore** it to extract information and meaningful insights (patterns, subsets with specific characteristics etc.),
- ✓ **form** initial **hypothesis** regarding the underlying information (IBM, n.d.)

Note that Business Understanding and Data Understanding phases should be seen as a whole because of the data mining goals setting and hypothesis building. Because it would be cumbersome to focus on data overall without a goal in mind, and vice versa, it would be hard to build up detailed data mining goals without forming hypothesis carrying out some preliminary analysis. Thus, the going back and forth between these phases are expected (Hofmann & Tierney, 2009).

Data understanding consist of the following steps:

### 2.2.1 Collect Initial Data

We need to acquire or access data from the sources that are listed as project resources and load the acquired data if there are any tools to be used for data understanding (IBM).

As an output of this process step, we need to prepare the **Initial Data Collection Report**.

### Output 1: Initial Data Collection Report

Following activities should be performed for proper collection of the data and documented:

- **Activity 1: Plan for data requirements**

We need to plan explicitly the information needed for solving the data mining goal and check the availability of the information from the related sources.

Necessary Data	Type	Cross-Validation Party
Sales Data	Internal	Commercial
Customer Data	Internal	Commercial
Inventory of Routes	Internal	Operations
Weather Data	External & Internal	FAOSTAT & Operations
⋮	⋮	⋮

Table 11: Necessary data and cross-validation parties / related resources

- **Activity 2: Develop data selection criteria**

After the clarification of the available data, we need to specify the selection criteria to be able to filter the necessary data considering:

- Which attributes are serving the data mining goal?
- Which attributes are irrelevant?
- How many attributes can selected techniques handle? (The CRISP-DM consortium, 2000)

We need to identify files / tables that contain necessary information and filter out the data of interest within the historic time period needed (e.g. 8 years of sales data might be available yet, 4 years of sales data might be needed for utilization in the process) (The CRISP-DM consortium, 2000).

- **Activity 3: Insertion of data**

We need to identify different types of data available in sources and optimal ways to extract data.

We might have to think:

- “Do the data contain free text entries, do we need to encode them for modeling?”
- “Do we want to group specific entries?”
- “How can missing attributes be acquired?”
- “How can we best extract the data?” (IBM)

- **Activity 4: Put Initial Data Collection Report together**

We need to prepare a report for supporting further re-iterations, assessments and “future execution of similar projects” pointing out:

- Various data used for the project with sources
- Data selection criteria

- Specifically more important attributes (IBM)

### 2.2.2 Describe Data

---

After the acquisition of the data needed, we need to examine the data by its high level / general / “surface” properties better and report the results (IBM).

As an **output** of this process step, we need to prepare the **Data Description Report**.

### Output 2: Data Description Report

---

Following activities should be performed for proper collection of the data and documented:

- **Activity 1: Perform volumetric analysis of data**

We need to identify the data, decide the extraction method based on the Initial Data Collection Report and access the data source. After gaining access to the data, we need to use basic statistical analysis (summary statistics, if suitable), check the data volume, number of multiples and complexity.

Additionally, we must identify and take note of the free text entries for further consideration about the means of standardizing / capturing information from them, if necessary.

- **Activity 2: Analyze attributes types and values**

In this step, we should try to assess the availability of attributes needed, described in the Functional Data Representation. We need to check the attribute types (e.g., numeric, string, etc.) and basic statistics (min – max, average, standard deviation, median, mode, etc.) for each attribute, and try to understand their meaning in business term. Based on the availability of attributes, and their relations, we need to decide if the attribute is relevant for our data mining problem.

Before proceeding further, it is important to:



- check if the meaning of the attributes have been used consistently across different data sets,
- validate our understanding and relevance of each attribute for the data mining goal with the domain experts

We need to analyze the data in a critical way, i.e. “we need to check to ensure that they make sense” (EMC Education Services, 2015).

**\*\* To exemplify;** if we are analyzing departure times (hours) of the planes for a project, we would not expect to see 25:00 or negative values as departure times (EMC Education Services, 2015).

After selecting the relevant data for problem solving, we need to identify the key relationships, overlaps between the tables in primary / foreign key attributes.

Following the completion of high-level analysis and understanding of the data, we would review our assumptions / goals and update them if necessary. It is good practice to align with corresponding stakeholders in case of an update in the goals (The CRISP-DM consortium, 2000).

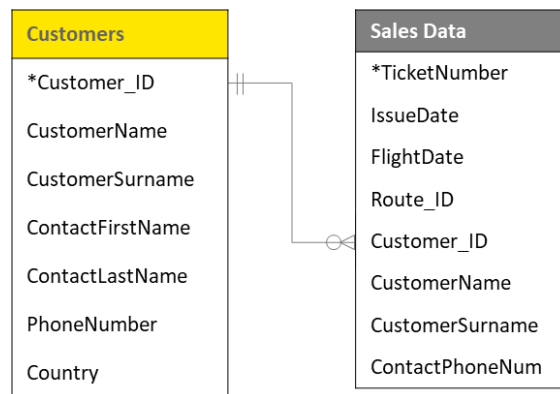


Figure 9: Attributes and key relationships

### 2.2.3 Explore Data

The aim of this stage is to perform some data analysis tasks such as querying, visualization and use some reporting techniques to “explore” data to set a ground for further data preparation phases. Note that analysis to be performed in this point could also feed / require update of Data Description and Data Quality reports (The CRISP-DM consortium, 2000).

We need to record our work in the **Data Exploration Report**.

### Output 3: Data Exploration Report

Although the analysis would go a bit further and deeper than the data description stage, the activities that we need to perform are quite parallel to the other steps:

- Activity 1: Perform analysis to explore data**  
 We have started analyzing the basic statistics of the attributes already in the previous step. On top of these statistics, we can build some plots that could point to behavior of data; and, we can look for different sub-populations with distinct patterns and report their characteristics (The CRISP-DM consortium, 2000).
- Activity 2: Form initial hypothesis**  
 After data exploration, assess the results with Data Description Report and form the initial hypothesis and identify related actions for further investigation. It is necessary to check the hypothesis against the data mining goal, update / enrich / specify the data mining goal if necessary, for providing better direction to further analysis. Following, we can perform basic analysis to verify the hypothesis, accept / reject based on analysis (The CRISP-DM consortium, 2000).

### 2.2.4 Verify Data Quality

The aim of this step is to analyze the quality of the data further and report on the results within the output report: **Data Quality Report**.

Note that the activities of this step would be more as complementary ones to the previous ones. Because, we are taking notes, and exploring further when we see an obvious problem in the data. That’s why especially Data Exploration and Data Quality Verification steps go hand-in-hand.

## Output 4: Data Quality Report

---

Following activities should be performed to enrich the findings on data quality from previous steps, if we have any, and to prepare the Data Quality Report:

- **Activity 1: Identify special values and catalogue their meaning**

We need to check if we have different interpretations for specific cases by attribute, and record their meaning.

- **Activity 2: Review keys, attributes and files**

We need to **check**:

- for the **data coverage** to see if we have all possible values presented in the data,
- the **keys across tables**,
  - We need to assess the ability and level of match since merging operations are done based on keys. We need to identify the problems / inconsistencies to be able to minimize the data loss caused by merging.
- **alignment** of the **attributes' values and meaning**,
  - As stated in the example **\*\* in page 27** of this document, we need to make sure that the values "make sense".
- **missing values** in the data,
  - We need to investigate meaning and the source of the missing information. There might be several cases regarding this:
    - **Errors when integrating data:** We need to check if we have made a mistake while collecting and integrating data into our system or our tool. If that is the case, we need to take corrective actions before proceeding further.
    - **Structural missing:** Data might be missing because it really needs to be missing. For example; if we were to look at a data obtained from a survey, having missing values as an answer to "# of flights taken with BEES Airlines before this flight" for the customers answered "Yes" to "Is this your first flight with BEES Airlines?" (Kuhn & Kjell, 2013)
    - **No data available:** If the data is not available, we need to check the reason behind it. There might be a problem with the process / hardware / software that is feeding this particular data (Méndez López, 2020).
      - It is important to make this distinction because different cases would lead us to different actions; in integration errors, we would go to data collection step, and with data unavailability, we would investigate the source and some treatment of our choice during the data manipulation phase if we are not dropping the attribute due to excessive unavailability.
- **outliers in the data**,
  - We need to define thresholds for outlier detection "What is an outlier for us given our data set and the business perspective?". Percentage of outliers are important for data consistency and reliability, and their source should be investigated further in case of anomaly (Méndez López, 2020).
  - Also, we need to pay attention if these outliers belong to a specific sub-population of interest since that might direct us in the further phases (Méndez López, 2020).

- Do not forget that we need to also take a look at the atypical entries / outliers of the categorical variables (Anomaly Detection) (Méndez López, 2020).
- **“delimiters’ consistency among files** if the data is stored in flat files” (The CRISP-DM consortium, 2000, p. 41)
- **measuring inconsistencies,**
  - To exemplify, we can have “M” and “male” both for indicating sex information of men (IBM)
- **“plausibility of values”**
  - We need to check if nearly all our data fields have the same or nearly the same value. Yet, we need to check this against our data understanding because some data would indeed could have similar values (The CRISP-DM consortium, 2000).
- **data inconsistencies among sources,** and how the noise would affect the study (The CRISP-DM consortium, 2000).

## 2.3 Data Preparation

The data preparation stage has two main goals:

- to produce a dataset,
- and a thorough description of that dataset that will serve for the modelling and analysis of the work.

These goals are properly performed by following five steps containing at least one output per step. The following are the steps to follow.

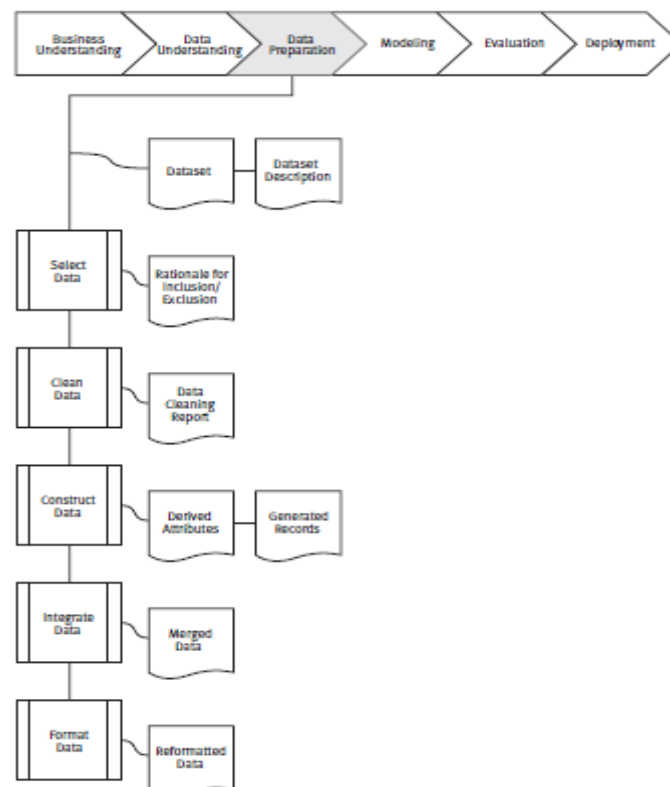


Figure 10: Data preparation stage tasks and outputs  
(The CRISP-DM consortium, 2000, p. 22)

The final output will serve as the final merged modelling dataset, along with a clear description that will serve to define parameters and inputs for the modelling, analysis, and evaluation phases.

### 2.3.1 Select Data

---

Based on the Business Understanding and Data Understanding phases, after defining the project plan, data mining goals, quality and quantity of data available, the team must decide as a whole which data sets will be used for the analysis. The proper, informed, and detailed selection of the data to be analyzed is crucial to the success of the project, this is a decision that must be taken with priority by the team, including:

- business experts to ensure the validity and correspondence with the business objectives/plan,
- data analysts to ensure that the data mining goals are to be met and the modelling will yield an important and productive output for the purposes of the project,
- data miners to check and inform any technical constraints regarding the data volume or data types ensure that the quality of the raw data can be transformed into useful information,
- the IT department to ensure that the team has the proper software installed for the type and volume of data to be selected,
- and the Legal department to ensure that the datasets to be used are compliant with all data privacy laws.

### Output 1: Rationale for inclusion/exclusion

---

The team mentioned above must prepare a list of the datasets to be used and excluded, stating the reasons for each decision. The data miners will be in charge of searching for relevant data both available inside the company as well as outside sources. Then data analysts will perform analysis of significance and correlation on the data sets for inclusion/exclusion. At this point, the data selection criteria defined in the Data Understanding phase is reviewed by the data miners taking into consideration the experiences in data quality and exploration, and then reviewed by the data analysts taking into consideration the past experiences in modelling. Sampling techniques must be taken into consideration if the data sizes are not handleable by the software available.

Finally, these tasks will yield all datasets that could be useful to the project to be reviewed by the rest of the team. Then the team runs the appropriate selection once more on the long list of possible datasets to include. Once the selection is performed, datasets are included into and excluded from the project, this information must be documented thoroughly for record-keeping and later stages of analysis and evaluation.

### 2.3.2 Clean Data

---

This task involves transforming the raw data into a clean dataset by raising the quality level to whatever is required by the data analysts and the software to be used. This will yield a data cleaning report, that is guided by the data quality problems described in the Verify Data Quality Task, that yield any correcting, removing, or ignoring of noise to perform the data cleaning task. The team members involved and their roles will be:

- data miners to perform the cleaning of data,
- data analysts to ensure that the data quality is raised to the level required by their analysis,
- the IT department to ensure that the data quality is raised to the level required by the software available.



Depending on the requirements and quality of the raw data and its subsets, the data miner will be selecting clean subsets to include into the dataset where needed, inserting defaults and nulls where needed, estimating missing data using statistics, as well as other techniques that the data miner deems necessary. For example, if a flight's fuel use is missing, the data miner can calculate the fuel use into the missing value with other available information and statistics.

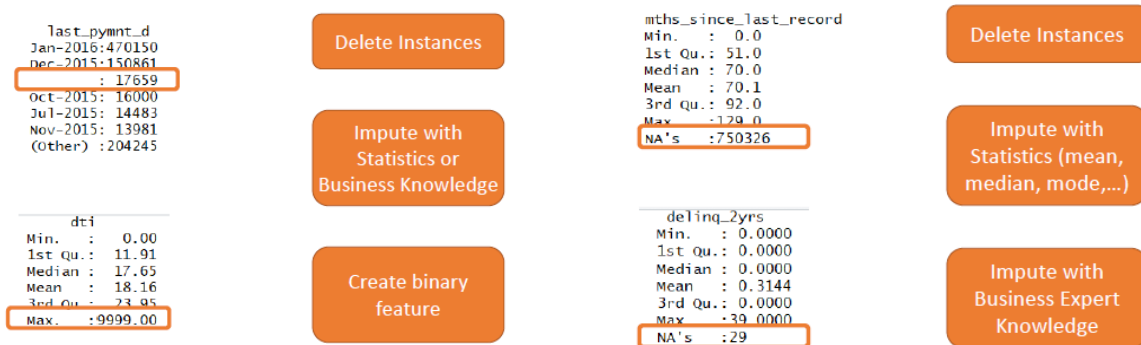


Figure 11: Methods of cleaning data  
(Lozano, 2020, p. 154)

These can serve as example of missing attributes or records that must be filled in or deleted with the objective of having clean and complete datasets.

sessionID	sex	device	age	channel	event_1_go_blog	event_2_product_details	event_3_change_product_color	event_4_show_shipping_cost
S0000000000196	Male	Mobile	+65	Display	0	0	0	0
S0000000000236	Female	Desktop	[45-54]	Organic search	0	0	0	0
S0000000000326	Female	Desktop	[35-44]	Organic search	0	0	0	0
S0000000000362	Male	Mobile	[45-54]	Organic search	1	0	0	0
S0000000000367	Female	Desktop	[35-44]	Direct	0	0	0	0
S0000000000463	Female	Desktop	[45-54]	Organic search	1	0	0	0
S0000000000481	Female	Mobile	[55-64]	Organic search	0	0	0	0
S0000000000608	Male	Mobile	[45-54]	Organic search	0	0	0	1
S0000000000654	Female	Mobile	+65	Display	1	0	0	0
S0000000000717	Male	Desktop	[35-44]	Organic search	1	0	0	0
S0000000000739	Male	Mobile	[45-54]	Organic search	0	0	0	0
S0000000000769	Female	Mobile	+65	Organic search	1	0	0	0
S0000000000773	Male	Desktop	[45-54]	Organic search	0	0	0	0
S0000000000834	Female	Mobile	[55-64]	Organic search	0	0	0	0

Figure 12: Missing and outlier data (red) illustration example in Dataiku  
(Lozano, 2020, p. 111)

As we can see in this illustration of Dataiku, there are missing values or values that are very much outside the limits (outliers). These outliers and missing values can affect the results and must be dealt with in a proper manner.

## Output 2: Data Cleaning Report

---

The output will be both a clean data set ready for construction, and a report detailing the techniques and decisions made by the data miner during the data cleaning. It starts by describing what was done to mitigate the data quality problems that arose during the Verify Data Quality task of the previous phase (Data Understanding). The data miner will document every single decision and action taken, as well as a description of the technique performed and the specific performance of the technique in this case.

Any transformation for cleaning purposes must be detailed in this report, as it will serve as a footprint to ensure the validity of the source of analysis of this project. Every output is only as good as its input; therefore, the analysis and modelling will only be as valid and useful as the quality of its data.

### 2.3.3 Construct Data

---

Much like the data cleaning task, this task has the objective of raising the data quality level but this time by creating new attributes (called derived attributes) that can add value to the database, as well as completing new records that can be inferred from existing ones with the guidance of the data analysts. However, this stage also invites the business experts to weigh in on any derived attribute that the analysis could benefit from. The following team members must be involved in this task:

- data miners to perform the constructing of derived attributes, completing new records, or transforming values for existing attributes,
- data analysts to ensure that the activities are adding to the data overall quality in terms of what is needed to improve the analysis and modelling,
- business experts to consult with and determine if the data construction activities add value to the project or objectives.

In the airline industry a derived attribute can be, for example if the data contains flight times (in HR) and distances (in km), the flight speed in km/hr. It is up to the team members to decide if a derived attribute is valuable to the analysis and modelling. Many times, we have two pieces of data that can build a third piece, but the valuable one would be the third, to make things uniform and comparable to one another.

This task has two outputs described below.

### Output 3: Derived attributes

---

This output will yield attributes inside the dataset that adds clarity and value to the modelling phase. First of all, any attribute must be normalized if need be, take for example currencies: if the airline is European but flies internationally, all data points must be normalized to the euro currency in order for it to be understandable, comparable, and useful. Second, the data miner and analyst can add a new attribute to add information about the derived attribute if needed and if the team members think it adds value, most of the time, the more information the better. Third, the members involved can determine a way to fill in the missing values of the derived attribute by computing an average or median, much like it was done during the cleaning task, but for the newly constructed data. The derived attributes added to the datasets must be documented accordingly.

## Output 4: Generated records

Newly generated or completed records can be crucial to an analysis or modelling evaluation. Say, for example, that the airline has lost some recurring customers to other airlines, but that is not reflected in the database since it is simply missing, completing records by representing lost sales and can add value during the modelling phase. The newly completed records also must be documented in order to understand the outputs of the modelling phase and evaluation phase, as these are crucial details that everyone involved must know about.

### 2.3.4 Integrate Data

Integrating data involves joining two or more data tables with different information, but on a common attribute, with the objective of creating organized merged tables that better serve the purpose of the project, while eliminating redundancy. The task also involves the aggregation of data, which means summarizing information in a way that better serves the purposes of the project. This task produces a summarized dataset from multiple data tables, called Merged data, which is the main output of this task. Examples can be seen in the output section. Given the importance of the decisions to take in this step the team members involved in this task are the following:

- data analysts to ensure that the merging activities are adding value to the data set and eliminating redundancy given the data mining goals, as well as consulting with the business experts,
- business experts to determine if the common attribute on where to join the tables, and where to perform aggregation is optimal given the business objectives,
- data miners to perform the joining and aggregation operations, storing and delivering the merged tables,
- IT Department to ensure the software provides the integration facilities required by the team.

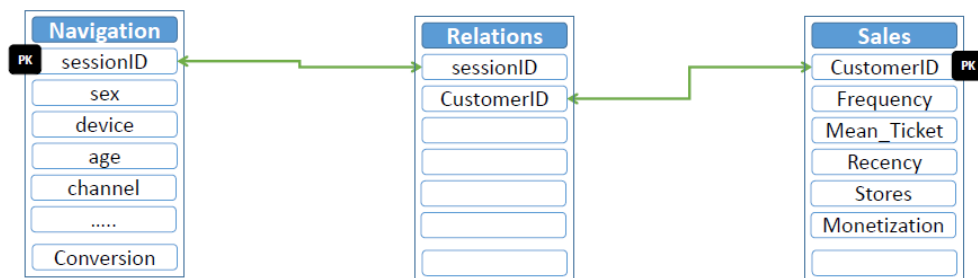


Figure 13: Common attributes on where to join data tables  
(Lozano, 2020, p. 161)

These data tables have common attributes, they can be joined into one data set with new values.

micro_conversion	step_1	step_2	step_3	conversion	pageviews	count	Customerid	Frequency	Mean_Ticket	Recency	Stores	Money
1	1	0	0	0	12	1	C12684234336844	2	185.05	19	2	2
0	0	0	0	0	4	1	C12624234266227	2	280.38	8	1	1
0	0	0	0	0	10	1	C24364634138018	10	84.91	1	2	2
0	0	0	0	0	8	1	C13714234176969	18	134.17	0	5	5
0	0	0	0	0	6	1	C32624234317945	5	288.58	48	2	2
0	0	0	0	0	8	1	C14054234144132	1	336.45	60	1	1
0	0	0	0	0	7	1	C13474234192545	40	80.45	2	5	5
0	0	0	0	0	6	1	C13254234236499	1	22.43	55	1	1
1	1	1	0	0	13	1	C13554234174126	20	253.17	4	8	8

Figure 14: Clean joined data set from various data tables  
(Lozano, 2020, p. 166)

The merged data set will contain all the information collected, that was previously on three different data tables.

Generating new records by using aggregation to summarize data tables into a more useful dataset is highly recommended with the purpose of shifting the focus towards what is being analyzed or studied. During this task, it is important to reconsider the data selection criteria, as some data tables will be deemed unnecessary and the team will find that other data tables are missing or could add value to the analysis. Therefore, as part of the data integration exercises, the team could find value in including or excluding data from the project.

## Output 5: Merged Data

The output of the data integration task is the actual merged data and ready for formatting into the dataset for Modelling, as well as the documentation providing every action taken during this step. Everything stated above (in the Data Integration task) must be considered to optimally produce this output. Let us take some examples to illustrate expectations during this stage.

Take for example a given flight that departs every Friday at noon, if the airline is collecting data from two or more different sources, the most probable scenario is that multiple data tables are collected and stored separately. The most logical solution is to join these tables together on the common attribute that could be customer ID (if ticket sales is the objective) or airplane ID (if operations and maintenance is the objective). The team members must decide on which common attributes joining the tables would make most sense from a business objective perspective.

Let us now take an example of in-flight entertainment for the aggregation task. If there is a data table with information about movie reproductions, showing how many times each movie was reproduced on a certain flight year round with a certain customer ID attribute attached to it and one record per movie, but the objective is to learn consumer preference targeting specific customers. Then it would make sense to transform the data table into being oriented towards providing what movies were watched by the customers, or one movie record per customer. That would be an example of aggregation.

The team must follow these general guidelines to produce a data set that is prepared for formatting into the data set that will be used for Modelling. Given the examples above, the team shall brainstorm the best ways to proceed given the purposes of the project.

### 2.3.5 Format Data

Formatting data is the last step in the data preparation stage and involves simply making modifications that are required by the modelling tool, such as date formatting (if it is stored as a string or character, transform into a date or factor), again depending on the modelling tool requirements. These changes will not change the meaning of the information whatsoever, it will simply change the data's format as the task's name suggests. For the formatting task the following team members should be involved:

- Data analysts to determine and communicate the modelling tool requirements regarding formatting,
- data miners to perform the formatting given the instructions of IT and data analysts,
- IT Department to ensure that the output is fit for the modelling tool.

The output of this task is the reformatted data, completely prepared for the modelling tool.



Figure 15: Example of formatting data features such as date  
(Lozano, 2020, p. 151)

### Output 6: Reformatted Data

This output will ensure that the modelling tool receives exactly what it needs to perform. This involves rearranging attributes or reordering records depending on the requirements. For example, if the field the model is to predict the amount of flights remaining until the next maintenance, and the tool requires that attribute to be in a certain location, then the data must be formatted in that way. Any changes should be documented to keep track of the changes made to the data set for future reference. Finally, the team must again reconsider dropping or adding other data sets given the outcomes and learnings of this task, there is always value in evaluating the data sources after each step.

## 2.4 Modelling

### 2.4.1 Select Modelling Techniques

Before starting to work on the actual model, the first step must be choosing the appropriate modelling technique. This task should be given special attention because every dataset will be a great fit to a given model, but we will not have a clear view of which model would be better until we explore our options. This idea can be best summarized by Wolpert and Macready's 'No Free Lunch' Theorem, stating that "any two optimization algorithms are equivalent when their performance is averaged across all possible problems". Regardless of the final algorithm of choice, we need to keep in mind that we will likely have to try out multiple alternatives.

We may want to start from thinking of the kind of analysis we want to run. That will generally depend on the kind of data in hand and what we want to do with it. Generally, a simple rule of thumb can be to use supervised learning for prediction and estimation and unsupervised techniques for pattern observation. Basically, we must understand the data available, the goal we want to achieve with it and the needs of the clients (internal clients) are in terms of analytical answers.

Kind of Algorithm	Aim of Use	Examples
<u>Classification algorithms (supervised)</u>	Purchase decision prediction - association by looking at similarities between instances	Logistic Regression, SVM, Random Forest Bayesian networks, Trees Etc.
<u>Regression algorithms (supervised)</u>	Price estimation – estimating a value for each instance based on information available	Linear Regression, SVM, Neural Networks, Random Forest Etc.
<u>Clustering algorithms (unsupervised)</u>	Business vs. leisure travelers vs ? - Dividing elements into groups sharing common features	K-Means, T-SNE, Bi-Clustering, DBSCAN Etc.

Table 12: Sample guide for choosing modeling technique

As suggested above, an important note when choosing the modelling technique would be to keep in mind the people we are choosing it for and the level of complexity that they are willing to allow the model to have. To put it simply, let's refer to what are commonly known as White or Black Box models. While in a white box model we have a clear view of how the computations lead to specific answer; yet, black boxes will generally be more complex and opaquer. So much so in fact, that white box models are occasionally needed to explain how black box ones work. This means that some business leaders will be weary of black box models, not trusting what they cannot see or understand.

In these stages, it is useful to have a check list available, something on the lines of:

- Have we understood the model's data requirements? (Data splits for training purposes, etc.)
- Do we have enough data at disposal? Is its quality acceptable? Is it the right type for our model?

Once we have chosen the kind of model(s) we want to use for the analytical question, it is time to select our modelling elements and outline the underlying assumptions. Different modelling techniques will feature different data needs. We may have to eliminate all missing values or evaluate the distribution of our records and so on.

In general, we will have to start from our data and see what we can formulate from it. Keep in mind that once we have decided on the function(s), having taken care of all of the model's requirements, we will have to evaluate the results. At this stage, it would be wise to list of all these needs, or assumptions. First and foremost, to help we keep track of things for the coming evaluation stage, as well as to have the clearest possible idea of the inner workings of the model. We will eventually have to explain our analysis to a likely less technically savvy person, say a director, that person will want clarity and ease of explanation to be able to evaluate the value of our analytical answer.

## 2.4.2 Generate Test Design

Before starting to build the model, it is important to have a premade list of the ways we plan to evaluate it. This stage is known as test design and it consists of clearly describing how we plan to 'feed' our data to the

model and later evaluate its results. By ‘feeding’, we mean how much of our data will be used to build the train and test datasets and whether or not we will also build a validation set.

Different models will need different evaluation metrics, and a model’s quality can be evaluated in multiple ways. Generally speaking, we can group algorithms by their scope and associate these to the best respective metrics. Mind that these are generalizations and that every metric will always have its strengths and limitations.

Algorithm Class	Best Metrics
Classification ( <i>supervised</i> )	-Accuracy/percentage of error -Area Under Curve (AUC) -Kolmogorov-Smirnov (SV) -Akaike Information Criterion (AIC) -Baesian Information Criterion (BIC)
Regression ( <i>supervised</i> )	-R-square/Adjusted R-square -Root Mean Square of Error (RMSE) -Mean Absolute Error (MAE) -Weighted Mean Absolut Percent Error (WMAPE)
Clustering ( <i>unsupervised</i> )	- Within/Between Cluster Sum of Squares -Purity, Rand Index, Mutual Information, F-measure <b>Note: evaluating pattern observation is harder due to the natural lack of a target variable</b>

Table 13: Sample metrics for different algorithm classes

Before setting off to the next step, we need to ask:

- Did we decide upon the data we plan to use? Did we split it into a train and test sets? Are we going to need a validation set too?
- Which evaluation metrics will we use? Which fit best our data and proposed model(s)?
- Do we have a settings adjustment threshold to discard the model if it passes the limit?

### 2.4.3 Build Model

Parameter setting is the starting point of keeping track of things. List all of our internal configuration variables and give a rational of why we think they work best in that scenario. Once that’s done, we must also keep a distinct track of the multiple models we created and what exactly they are supposed to do, mind this is not a full report, we just need to be able to tell one from the other.

It is now time to start using the tool of our choice to build the functioning model(s). At this stage, we should have the datasets ready. As said and repeated above, our best option is to try out different models to see which ones give the best results; we may even discover now insight at this stage! Here, it is pivotal to keep track of our actions. We will be testing in lots of different ways and getting all sorts of results, we need to be sure to have a way to make sense of it all and most importantly to know how to discuss or present it later.

Lastly, we need to give a quick interpretation of the models, talk about what the results can mean and list any technical difficulty or issues with data or performance that we may have encountered. This will be useful in the coming sections when having to identify the best options.

#### 2.4.4 Assess Model

---

Now that we have tested out our data, we need to identify which models gave us the best outcomes. We will do so by comparing our results to the set of evaluation criteria that we have stipulated in the previous steps of this stage.

We need to compare our models, rank them, check their plausibility and reliability. Are our results feasible and logical? What parameters do we see the need to change, add or delete? Do we have any insightful results in business terms? These are the questions we must answer in order to select the best of our outputs. We should keep in mind that all of this evaluation must be done with the validation set previously made; using the training set(s) would inevitably bias our conclusions.

Many problems, difficulties and complications may arise at this stage. A common and noteworthy issue would be that of overfitting. The best way to describe this issue is thinking of our analytical solution as getting too comfortable with its training set. If this happens, the model's capabilities to accurately give answers from other, real, datasets will be hindered. This will be due to the fact that the model paid too much attention to the specific case of its training data, looking at peculiarity instead of generality. In short, the model will be worthless due to the fact that it will base its analysis on the wrong factors. To prevent this data science nightmare from happening, we must put extra care into building our training ABTs and cross-validate (Amazon Web Services, Inc., 2020) our models.

Once we have clearly evaluated and ranked the models, we will have a clearer idea of what works and what does not. At this point, we will be able to make an informed decision on which parameters need changing in order to add value to our analysis and which are working just fine.

To reiterate, this is the phase where we have to construct our analytical model from scratch. The main decision we will have to take is related to the complexity of the model: should we go for a white or a black box model? That being decided, we should keep in mind the two main challenges will be: making sure we have enough data and computing power to successfully answer our question and avoiding overfitting as it will jeopardize the success of the work.

The main steps to successfully move on to the next stage are (Lozano, 2020):

- Designing a meaningful test, able to outline the strengths and limitations of the model
- Carefully keep track of detailed descriptions of the models, in order to be able to compare and contrast them
- Critically evaluate our model, summarizing its results and getting a good idea of its value to the business question



## 2.5 Evaluation and Presentation of results

Once the model is completed, the results must be interpreted. If the model does not meet the business objectives, this phase helps to explain why it is deficient. The total outputs of a data mining project will be defined by the equation (The CRISP-DM consortium, 2000, pp. 51-54):

$$Results = Models + Findings$$

Having a good model, from which we cannot extract useful insights or findings, is a lost model. It is important to set on the calendar enough time to be able to discover as much findings as possible and consult experts when necessary.

### 2.5.1 Evaluate Results

This step helps to assess the degree to which the model has fulfilled the business objectives set on the first phase. Additional actions such as testing the models in real applications will be performed only when we are ahead of our schedule. We need to consider if the budget constraints took into account the testing on the application.

#### Output 1: Assessment of data mining results with respect to business success criteria

The results will be summarized in terms of business success criteria, including a final statement whether the project meets the initial business objectives and the corresponding traffic light for presentations.






	All business objectives have been fulfilled
	Primary objective fulfilled. 1-2 secondary objectives not fulfilled
	Primary objective fulfilled. None secondary objectives fulfilled
	Primary objective not fulfilled. Secondary objectives fulfilled
	Primary and secondary objectives not fulfilled

Table 14: Status of models

As an example in our BEES Airlines, if the primary objective of reducing the missing luggage on transfer flights through improved tracking is successful but the secondary objective of improving the speed of luggage transfer between airplanes is not fulfilled; then the status of the model will be on “Yellow”.

Secondly, other data mining results need to be assessed as well. We will evaluate and present results of models that are related to other findings such as new challenges or hints. (Lozano, 2020)

- **Activity 1: Understand the data mining results**

The key of success is the presentation of results. Within this activity we need to define which visuals are we choosing to explain the results to the rest of the team. Prior to a meeting with the members of the evaluation team, a short document must be drawn up and sent to all members including:

- All visuals to be used in presentation, labeled (e.g. Visual 1)
- Short explanation of results within visuals, labeled (e.g. Exp.Visual1)

Each member will indicate which visual he/she did not understand at first sight and send back the document to the presentation moderator who will decide whether to exclude certain non-understandable visuals within the following guidelines<sup>1</sup>:

Visual Label	% of members no understanding visual	Recommended action	Moderator Decision
Visual1	70-100 %	Exclude	⚙️ Check Labels
Visual2	40-70 %	Reduce dimensions of graph / Check labels and legends / Think about another graph type / Reformulate the explanation	⚙️ Reduce dimensions
Visual3	10-40 %	Reformulate the explanation / Select as potential visual	⚙️ Exclude
Visual4	< 10 %	Include in presentation	⚙️ Include

Table 15: Visual understanding guidelines

- **Activity 2: Interpret the results in terms of the application**
- **Activity 3: Check effect on data mining goal**
- **Activity 4: Check data mining result against the given knowledge base**
- **Activity 5: Evaluate and assess results with respect to business success criteria**  
The assessment involves the performance of the models, but also aspects related to the deployment such as the training and prediction velocity, the robustness and maintainability, the use of tools, the dependence on other subsystems or the use of library code vs. homegrown code (Lopuszynski, 2016).
- **Activity 6: Compare evaluation results and interpretation**  
One meaningful activity is to interpret the results in a short-written explanation of them. This text should be drafted by different members of the analytical team to agree on the interpretation of the results.
- **Activity 7: Rank results with respect to business success criteria**  
The ranking of the results of the models will serve as baseline to approve or reject the different models. Some models will be discarded for next iterations and some others will be considered as potential candidates for further iterations.
- **Activity 8: Check effect of result on initial application goal**  
Does the result fulfill our application goal? Which things need to be improved to obtain the desired effect? It is highly recommended to define the risks and new requirements for the development of the application

<sup>1</sup> If the visual requires dynamic actions (e.g. Displaying change of variables dynamically) send link to video instead of visual.

- **Activity 9: Determine if there are new business objectives**

In case that new business objectives arise, we need to check if the results of our models fulfill all requirements and expectations. Ranking of results may vary and must be updated.

- **Activity 10: State recommendations for future data mining projects**

One of the advantages of this handbook is to have an idea of how to enhance the performance of data mining projects after gaining some experience with the methodology. Therefore, a list of recommendations must be included in the following spreadsheet. Each phase will have a list of categories defined. In case new categories need to be included, then select “Others” and write the new category.

Date	Phase	Generic Task	Output	Activity	Category	Recomm.
dd.mm.yyyy	Data preparation	Clean Data	Outliers	Empty names	Time saving / Data quality	Use of X tool

Table 16: Recommendation spreadsheet

## Output 2: Approved models

Select and approve the generated models that meet the business success criteria. In order to avoid misunderstandings with the choice of models, fill the following spreadsheet. The date of approval is a must on the spreadsheet.





Model Name	Status	Approval	Date	Ranking	Things to improve
SVM1		Yes	dd.mm	1	e.g. Accuracy, complexity, overfitting, ...
Bayes		Yes	dd.mm	2	
SVM2		Yes	dd.mm	3	
Class1		No	dd.mm	4	

Table 17: Approved models

## 2.5.2 Review Process

This task considers whether some aspects or tasks have been overlooked during the whole process. In order to avoid a deployment of a model that did not take into account some crucial factors, we need to perform a quality assurance review up to this point.

## Output 3: Review of process

The success of an iteration or a project does not mean that some phases or outputs do not show some weaknesses. (IBM, n.d., p. 36) Future projects will take into consideration these inflexion points to improve their performance and efficiency. There are two main activities within this output.

- **Activity 1: Definition of questioner**

Each member involved on the project will receive a questioner, which will be crucial for moving into further steps. The following template shows the default design of the questioner<sup>2</sup>. It consists of 35 questions, 5 for each phase. Other questions, both generic or phase-specific, could be addressed by the evaluation assessment members and included in the questioner after the presentation of the results.

QUESTIONER REVIEW OF PROCESS		Date	Respondent
<b>BUSINESS UNDERSTANDING</b>			
<ul style="list-style-type: none"> <li>Was this stage necessary for the final result? Rank it [1-5]</li> </ul>	<input checked="" type="radio"/> Yes <input type="radio"/> No		
<input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 (Mandatory)			
<ul style="list-style-type: none"> <li>Was it executed optimally?</li> </ul>	<input type="radio"/> Yes <input checked="" type="radio"/> No		
<ul style="list-style-type: none"> <li>Did you encounter failures during this phase? If so, which ones?</li> </ul> <hr/>	<input checked="" type="radio"/> Yes <input type="radio"/> No		
<ul style="list-style-type: none"> <li>Could you identify misleading steps or dead ends?</li> </ul>	<input type="radio"/> Yes <input checked="" type="radio"/> No		
<ul style="list-style-type: none"> <li>Did you encounter during this phase? Which ones?</li> </ul> <hr/>	<input checked="" type="radio"/> Yes <input type="radio"/> No		

Figure 16: Review of process questioner

- **Activity 2: Assessment of questioner. Report of recommendations**

Some useful insights can be extracted from the questioner. The opinion of the members involved in the process will determine the course of actions and new approaches in new iterations and projects. To do so, we will write a report with recommendations done by all members for each phase and some statistics to show where we should focus our attention (e.g. Average of question 1 to define where we allocated more efforts).

#### Output 4: Presentation

The presentation of the results will define further courses of actions and will help with new iterations of the methodology. The presentations must be performed within the standards and guidelines of the company. The following steps regarding the content of the presentation must be read carefully to avoid missing meaningful insights of the results:

- 1<sup>st</sup> slide defining business objectives: Highlight Primary and success criteria. It is important to define the date of the initiation of the project, as well as the updates in the business objectives
- 2<sup>nd</sup> slide: Definition of main assumptions, requirements and constraints focusing on data quality and technical accessibility to data
- 3<sup>rd</sup> slide: Main challenges not solved on the data preparation phase.

<sup>2</sup> Example for the Business Understanding phase. The questioner will include all phases.

- 4<sup>th</sup> slide: List of approved models and status, matching the business objectives defined on the 1st slide (See **Error! Reference source not found.**)
- 5<sup>th</sup> – nth slide: Each slide must contain a detailed view of the approved models. The ranking of the models will determine which ones must be presented. In case the moderator wants to use an interactive tool to present some results (e.g. Power BI), it is highly recommended to record the actions and present directly the video recorded. The tool can be left on the background in case some questions arise during the presentation. It is important to include the time frame used to develop each model.
- 6<sup>th</sup> slide: Further steps. Each phase must contain time estimations of course of actions as well as indicating critical steps on each phase.
- Backup slide: Terminology
- Backup slide: Main insights of initial data collection report and data quality report, focusing on missing attributes, coverage or degree of completion of datasets and mention possible actions
- Backup slide: Technology platform, with architecture design, big data component selection and technology management strategy
- Backup slide: Data preparation rationale inclusion/exclusion decisions



Avoid **overwhelming with too much data**  
**Double Check:** Legends, Axes, Colors, Scales, Labels, etc.

Once the presentation is ready, we need to make sure that everybody is on the same page in terms of terminology. The terms that we are using have to be defined on the glossary (see **Error! Reference source not found.**). In case that we need to use new ones, update the glossaries.

The feedback of the presentation will be crucial to determine whether the explanation of results was successful or not.

### 2.5.3 Determine Next Steps

The assessment of results, the process of review and the final presentation will help the project team on how to proceed. The first decision to be done is to define the status of the project: if the project is finished and we can move on to deployment; if we need to perform further iterations to polish certain aspects of the process; or if we need to set up new data mining projects.

### Output 5: List of possible actions

A list of possible actions analyzing the results along with reasons for and against the different options. The first possible action would be a new presentation of results depending on the feedback, as this output is a critical point of extracting useful insights from our data. The following criterial will serve as a guideline for decision making from this point on:

- Status and ranking of model
- Potential gains: Accuracy in model, low effort to fulfill secondary business objectives
- Time effort and deadlines
- Recommendation report
- Availability of new data

- Scalability: Hardware resources requirements, budget
- Re-evaluation of additional costs with new iterations

## Output 6: Decision

The list of actions will be ranked and well documented to determine which one will be more valuable.




Action	Rank	Main criteria	Status	Time-effort	Recomm. Report	Scalability	Additional Costs
New iteration1	1	Potential benefits in Accuracy of model X, Low time effort		Low	Focus on data preparation: Outliers handling	Difficult	+ 2%
New iteration2	2	Easy solution to achieve secondary business objective					
Deployment	3	Model Y ROC, Status					

Table 18: Decision based on list of actions

## 2.6 Deployment

To ensure an organized deployment of the chosen data mining approaches and to facilitate the repeatability of processes, we need to come up with a written deployment strategy. Hereby, a crucial shift in perspective occurs as the main agent and executor of this stage will be the eventual user and not the data mining team as in the previous generic tasks (Wirth & Hipp, 2000). This final job intends to clarify the entire process structure and make it accessible to all stakeholders.

### 2.6.1 Plan Deployment

As with any significant stage, we must carefully approach the deployment phase. It is especially critical to plan the following tasks since knowledge generated throughout the project may have shifted the focus of our deployment strategy (IBM, n.d.). For instance, finding a certain customer segment appeal more strongly to dynamic pricing schedules, we need to make our findings available to responsible departments like marketing that can leverage on such insights.

## Output 1: Deployment Plan

To begin planning out the deployment, we suggest the following list of activities according to IBM (n.d.).

- **Activity 1: Summarize the results - Models and findings**
- **Activity 2: Per model - Develop deployment plan including integration with systems**
- **Activity 3: Per finding - Draft a plan how to distribute knowledge**
- **Activity 4: Consider alternative deployment plans**

- **Activity 5: Consider a first monitoring and maintenance strategy**
- **Activity 6: Identify possible problems during deployment and respective contingency plans**

For the technical deployment, where we will deploy the data mining model in the IT environment, we can follow the above instructions emphasizing on the model. Major challenges have been identified not only as the actual integration in applications and process planning (extracting, transforming, loading of data) but also to answer questions like what will effectively be deployed and who is operating the task. In general, a model launched within existing analytical tools will be easier to deploy, whereas a model outside of current applications should be easier to integrate. Since there is a trade-off between the described approaches, it is advisable to run one first iteration at an early stage in the project to identify possible compatibility and stability issues from a technical point of view. The technical deployment plan should consider these mentioned points (Lozano, 2020).

From another perspective, the business deployment seeks to integrate the new model in the organization's operations. Thereby, we need to establish a link to the business understanding and aim to solve the problems we originally raised. During this phase of business deployment, we focus on decision making so it is required to select a threshold that will initiate business action (Lozano, 2020). Instead of maximizing some statistical measures that might not apply to the varying circumstances and natures of our projects, we opt for a threshold selection based on business activity. The rationale for this threshold analysis should be some cost-benefit conversion metric, which is subject to a case-by-case evaluation. It is crucial that these thresholds are adaptive as it facilitates the versatility of our model. For instance, we will be able to work with the same model regardless of our data mining strategy. While we could focus on recall in one problem task, we might want to achieve maximized precision in another one (Lozano, 2020). Linking back to our case, we propose a flexible framework that considers the significance of observations to the organization, splitting them into percentile buckets. To illustrate this analysis, we show a possible diamond structure to classify the percentile ranges into 5 distinct brackets. In Figure 16, the exemplified bin widths vary from 15 to 30%. Depending on the business problem, we must determine the effective thresholds and assign specific action strategies to each bucket. In the example of customer segments, where the most valuable customer groups are identified as percentiles 85-100%, those actions could state:

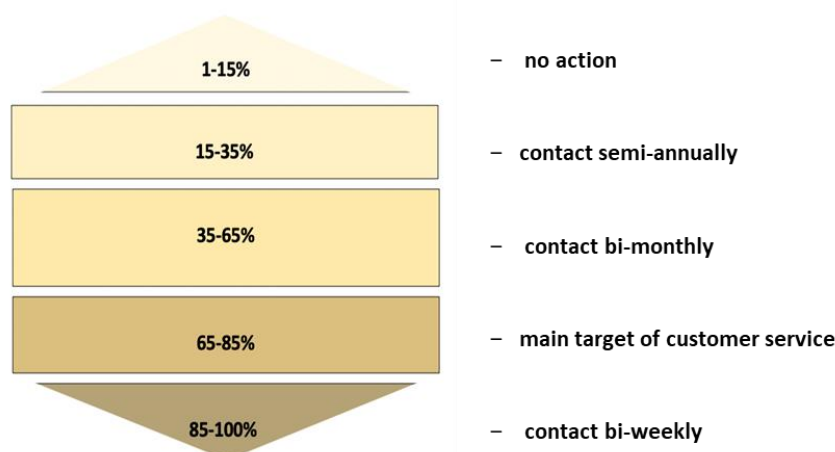


Figure 17: Suggestion of percentile brackets for threshold analysis and corresponding actions (Lozano, 2020, pp. 305-309)

## 2.6.2 Plan Monitoring and Maintenance

---

In a setting where the data mining process is continuous and is running daily, we must have a profound monitoring and maintenance plan in place that allows us to unceasingly supervise such processes.

### Output 2: Monitoring and Maintenance Plan

---

The central output here is the bespoke plan which will focus on the maintenance and monitoring of deployed results. It should set up and summarize an updating process for the data mining model: when it will be initiated (frequency, trigger event, performance measuring) and how it will be executed (The CRISP-DM consortium, 2000).

- **Activity 1: Check for dynamic impacts on results**  
As a first assessment, we want to raise awareness for external sources that may influence the results or models employed. Possible examples have been found as seasonal changes or market value fluctuations that might weaken an established model (IBM, n.d.).
- **Activity 2: Define accuracy monitoring measures**  
We then must unambiguously determine how we will monitor the model accuracy. We need consistency here to ensure an objective and unbiased evaluation of project components over time. For instance, we might want to look at the stability of either the entire model or the variables themselves. It is advisable to evaluate the model capability with a capability metric over time (Lozano, 2020).
- **Activity 3: Determine objective criteria for expiration of data mining results and set up action plan**  
In this key activity, we specify further objective thresholds and requirements that will decide when a data mining model is not valid anymore. To link to the said capability metric, an example would be to set a minimum requirement that is to be fulfilled. On top, we put forward to formulate an action plan that details what next steps are taken upon a model expiring (The CRISP-DM consortium, 2000). In case of the capability metric, this plan will provide guidance should the metric fall below its limit. Potential actions could include a guideline that initiates an independent data mining project or, more simply, a structured update plan for the respective model.
- **Activity 4: Assess dynamics of business objectives**  
After assessing external factors in Activity 1, we now have a closer look at internal forces. More specifically, it is the business objectives that require thorough examination. In case they evolve over time, we need to be aware of it and accordingly update our maintenance plan. Optimally, we journalize every project version with its distinct business problem description. An additional benefit of proper documentation is demonstrated whenever facing similar business challenges. Having stored each set of business problems accurately, we can always refer to it and have the previous solution readily available (IBM, n.d.).
- **Activity 5: Develop the monitoring and maintenance plan**  
To sum up, the listed components of such an updating scheme can eventually be formulated as the monitoring and maintenance plan.



### 2.6.3 Produce Final Report

---

When the project comes to an end, the respective team should finalize their work with a final report. In many cases and additional presentation will be required as indicated below.

#### Output 3: Final Report

---

The final report will serve as a summary of project and should state the obtained results for future references. Since it covers all aspects of the data mining project, we would like to refer to the CRISP-DM Data Mining Lifecycle in Figure 1. In order to reach the desired level of detail, we suggest following the below list of activities proposed by the CRISP-DM consortium (The CRISP-DM consortium, 2000):

- **Activity 1: Identify what reports are needed**  
First, we must conclude about what reports to work with. While this guide has introduced numerous relevant documents, we want to point out some selected elements. As key components the original business problem, the data mining process together with cost incurrence, possible variations from both, model and finding results from data mining, the deployment strategy, and conclusions for future data mining should be discussed (IBM, n.d.).
- **Activity 2: Analyze to what extent data mining goals have been reached**  
In a next step, we explore how well the given objectives for data mining were met. This analysis should follow from generic task 2.4.4 “Assessment” of the model.
- **Activity 3: Identify target groups for reports**  
At this stage, we must actively decide what our final target audience for the report will be. It forms a critical decision as the following activities will build upon it. We suggest considering several groups, however, bear in mind that different groups mean heterogenous knowledge base and therefore will ask for several tailored report versions.
- **Activity 4: Outline structure and contents of reports**  
Following, the reports should be customized in that we cherry-pick the actual content of the selected reports. This will allow us to serve all the previously mentioned demands.
- **Activity 5: Select findings to be included in the reports**  
Once the basic structures for the relevant reports are defined, we need to choose some key take-aways as our main argument in the deployment. The selection will feed both our final report and the final presentation as in Output number 2.
- **Activity 6: Write the report**  
After defining the scope of the report in the named levels of depth, we produce the output “final report”. With help of the step-by-step logic of breaking down what reports to include, what was reached, for whom they are intended, their individual structure and focus, we set the stage for a well-defined report.

#### Output 4: Final Presentation

---

Analogous to the report, we find it useful to prepare a presentation that will summarize the data mining project, giving a brief glimpse at some selected findings from the report. It will be presented to the responsible target group as means of a final communication.

The final presentation therefore requires us to both specify key take-aways from the final report and define the target group to which it will be reported to as described in the above activities (The CRISP-DM consortium, 2000). We stress that the information provided should be dependent on the target group, e.g. when presenting

to the head of data science we want to highlight technical aspects raised in the deployment strategy, whereas for the operating department business aspects will be of higher relevance.

#### 2.6.4 Review Final Project

---

As a last generic task, a review of the project is to be carried out that specifies how it was rolled out. Taking note of strong and weak points, this will allow to gather and generalize the observations made throughout the course of the project.

#### Output 5: Experience Documentation

---

As a summary of experiences from the data mining project, we follow the steps outlined below. It is worth mentioning that this document is considered a living document, i.e. it is subject to constant change and will be updated on a running basis (Lozano, 2020). In working towards this final output, the CRISP-DM guidelines suggest the following list of activities (The CRISP-DM consortium, 2000).

- **Activity 1: Interview all significant people involved in project**  
It is important to see how all members of the involved workforce feel about the data mining project and what their experiences look like. This way, we ensure no perspective is left out when discussing the general impact on the company.
- **Activity 2: If applicable – Interview end user that works with the data mining results**  
This activity should investigate the perception of the active end user, evaluate their level of satisfaction, assess the need for further support and reveal where they see potential for improvement. The take-aways from this interview bring particular value as they highlight the practical business perspective on things in the CRISP-DM approach.
- **Activity 3: Summarize feedback and initiate the effective output “experience documentation”**  
Equipped with the inputs from within the organization, it is now time to summarize the impressions and to put them to paper. The result will be the eventual document called “experience documentation”. In order to ensure a transparent and approachable structure, we propose to follow a simple layout that includes a section for each strength, weaknesses, areas of potential, and areas of improvement.
- **Activity 4: Analyze the process**  
Here, it is our aim to discuss and process successes, struggles, challenges, dead ends, and ultimately some general learnings. This will serve as a basis for the subsequent activity.
- **Activity 5: Document the specific data mining process**  
The question arises as to how the gained experience and results can be reapplied to our particular process. Therefore, we suggest writing down and documenting the key findings.
- **Activity 6: Generalize from details**  
Finally, to conserve the value generated from applying the findings to future data mining projects, we need to generalize the detailed information collected above.

## Bibliography

alexsoft. (2019, April 24). *Dynamic Pricing Explained: Machine Learning in Revenue Management and Pricing Optimization*. Retrieved from alexsoft: software r&d engineering:  
<https://www.altexsoft.com/blog/datascience/dynamic-pricing-explained-use-in-revenue-management-and-pricing-optimization/>

Amazon Web Services, Inc. (2020). *AWS: Amazon Machine Learning Developer Guide - Evaluating Machine Learning Models*. Retrieved from AWS: <https://docs.aws.amazon.com/machine-learning/latest/dg/cross-validation.html>

Bonthu, S., & Bindu, K. H. (2017). Review of Leading Data Analytics Tools. *International Journal of Engineering & Technology*, 10-15.

EMC Education Services. (2015). *Data Science and Big Data Analytics : Discovering, Analyzing, Visualizing and Presenting Data*. N/A: Wiley. Retrieved from <https://ebookcentral.proquest.com/lib/bibliotecaie-ebooks/detail.action?docID=4548030>

Harned, B. (2019, September 16). *How to Clear Project Confusion with a RACI Chart*. Retrieved from teamgantt: <https://www.teamgantt.com/blog/raci-chart-definition-tips-and-example>

Hofmann, M., & Tierney, B. (2009). Development of an Enhanced Generic Data Mining Life Cycle (DMLC). *The ITB Journal*, 10(1), 50-71. doi:10.21427/D75R0B

IBM. (n.d.). *IBM SPSS Modeler CRISP-DM GUIDE*. Retrieved from <ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/17.0/en/ModelerCRISPDm.pdf>

IBM. (n.d.). *IBM SPSS Modeler CRISP-DM Guide: CRISP-DM Help Overview*. Retrieved from IBM Knowledge Center: [https://www.ibm.com/support/knowledgecenter/SS3RA7\\_sub/modeler\\_crispdm\\_ddita/clementine/crisp\\_help/crisp\\_overview.html](https://www.ibm.com/support/knowledgecenter/SS3RA7_sub/modeler_crispdm_ddita/clementine/crisp_help/crisp_overview.html)

IBM. (n.d.). *IBM SPSS Modeler CRISP-DM Guide: Data Understanding*. Retrieved from IBM Knowledge Center: [https://www.ibm.com/support/knowledgecenter/SS3RA7\\_sub/modeler\\_crispdm\\_ddita/clementine/crisp\\_help/crisp\\_verify\\_data\\_quality.html](https://www.ibm.com/support/knowledgecenter/SS3RA7_sub/modeler_crispdm_ddita/clementine/crisp_help/crisp_verify_data_quality.html)

IBM. (n.d.). *IBM SPSS Modeler CRISP-DM Guide: Deployment*. Retrieved from IBM Knowledge Center: [https://www.ibm.com/support/knowledgecenter/SS3RA7\\_sub/modeler\\_crispdm\\_ddita/clementine/crisp\\_help/crisp\\_deployment\\_phase.html#crisp\\_deployment\\_phase](https://www.ibm.com/support/knowledgecenter/SS3RA7_sub/modeler_crispdm_ddita/clementine/crisp_help/crisp_deployment_phase.html#crisp_deployment_phase)

Kuhn, M., & Kjell, J. (2013). *Applied Predictive Modeling*. New York: Applied Predictive Modeling.

Larson, E. W., & Gray, C. F. (2011). *Project Management: The Project Management*. New York: McGraw-Hill Companies.

Lopuszynski, M. (2016, 06 07). (W. D. Meetup, Producer) Retrieved from Slidehare.net: <https://www.slideshare.net/lopusz/crispdm-agile-approach-to-data-mining-projects>

Lozano, A. P. (2020). *The Knowledge Discovery Process*. IE.

Mayo, M. (2016). *Data Science Basics: What Types of Patterns Can Be Mined From Data?* Retrieved from KDnuggets: <https://www.kdnuggets.com/2016/12/data-science-basics-types-patterns-mined-data.html>

Méndez López, A. J. (2020). *Machine Learning I: Data Manipulation with Dataiku DSS*. IE.

Nemati, H. R., & Barko, C. D. (2003). Key factors for achieving organizational data-mining success. *Industrial Management & Data Systems*, 282-292. doi:10.1108/02635570310470692

Prada, J. (2020, May 17). E-mail interview about hardware and software requirements. (A. D. Roni, Interviewer)

Rubin, R. S. (2002). Will the real SMART goals please stand up. *The Industrial-Organizational Psychologist*, 39(4), 26-27.

Smartsheet Inc. (n.d.). Retrieved from <https://www.smartsheet.com/expert-guide-cost-benefit-analysis>:  
<https://www.smartsheet.com/expert-guide-cost-benefit-analysis>

The CRISP-DM consortium. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. SPSS Inc.

Utrecht University. (n.d.). *Guides: Costs of data management*. Retrieved from Utrecht University Research Data Management Support: <https://www.uu.nl/en/research/research-data-management/guides/costs-of-data-management>

Will, K. (2019, June 23). *Business Essentials: Cost-Benefit Analysis*. Retrieved from Investopedia:  
<https://www.investopedia.com/terms/c/cost-benefitanalysis.asp>

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (pp. 29-39). London, UK: Springer-Verlag.