CRISP-DM

Vanshita Arya

Volunteer – Data Science Research

Indian Institute of Artificial Intelligence and Accelerated Computing

vanshita@iiaiac.org

CRISP-DM (Cross-Industry Standard Process for Data Mining) is a widely used methodology that provides a structured approach to data mining and analytics projects. Developed in the late 1990s by a consortium of companies and organizations, it outlines a comprehensive process model with six key phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. This iterative framework helps ensure that data mining projects are aligned with business objectives, produce reliable and valid results, and can be deployed effectively. CRISP-DM's versatility and industry-agnostic nature make it a popular choice for data professionals seeking to systematically approach data-driven problem-solving.

CRISP-DM, which stands for Cross-Industry Standard Process for Data Mining, is a methodology or a process that helps us or provides us with the planning or blueprint for how to carry forward our data mining projects. Here is what it involves:

- A Step-by-Step Guide: structure the industry-driven projects stepwise organised into phases.
- A Big Picture View: as a process model, it provides an overview of the data mining life cycle.

It provides a fluid framework for devising, creating, building, testing and deploying ML solutions.

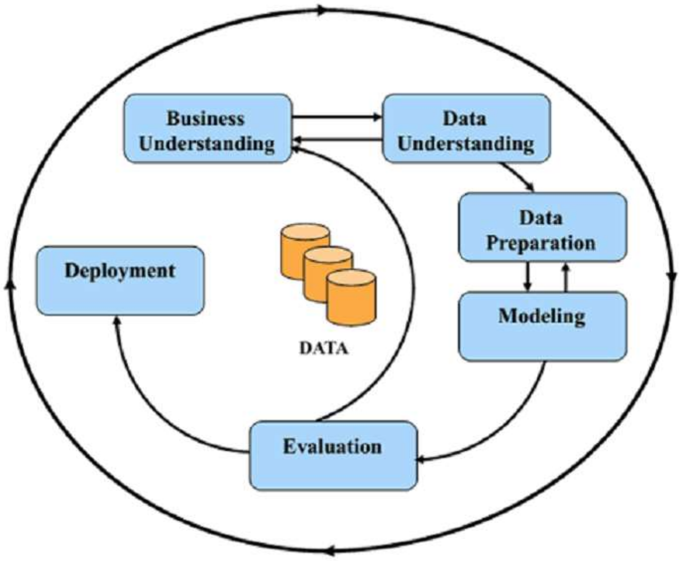


Fig. 1: Process flow of CRISP-DM

Background and Context

- ***Early 1990s:***

  As the volume of data increased, more studies and practices were conducted to extract valuable insights from the piles of data. As a result, advanced tools and techniques came into existence, rising in the world of the Data-driving Industry. The Business world and Intelligence agencies recognized the potential of Data Mining techniques for decision-making & predictions.

- ***1996:***

  Due to the lack of any standardized process to follow up on data mining steps, an initiative was taken by the group of companies and institutions to address this issue. They called it the CRISP-DM initiative. The goal was to have a consistent framework that could be used across different industries. This collaboration involved organizations from various countries, including NCR Systems Engineering Copenhagen, Daimler-Benz, and SPSS Inc., with funding from the European Union.

- ***1997-1999:***

  The consortium put in a lot of effort. They did extensive research, had consultations, and even ran pilot projects. Their goal? To create a CRISP-DM methodology. In 1999, they created and published their first version of CRISP-DM, which gained acceptance and huge popularity due to its comprehensive and practical approach.

In this article, we will cover what CRISPR-DM is, how it helps, why to use it, how to use it when to use & practical usage of this framework in a project.

What is CRISP-DM?

Key features:

The key features of CRISP-DM involve flexibility, neutrality, commonality, iterability, and objective focus, which are discussed below:

1. *Flexibility:* CRISP-DM models are flexible and can be customized easily. Imagine you are investigating a money laundering case, so instead of applying models directly you may first explore and visualize data to reveal suspicious patterns and anomalies. So CRISP-DM lets us adapt our approach based on our specific goals. We can focus on what matters most, whether it's exploration, modelling, or other steps.

2. *Iterative Process*: It implies the process is not linear. It may be possible to revisit the steps based on new issues or insights. Assume while making a model, you might find problems with the data or missing information. This means you have to go back and fix the data. This back-and-forth process helps improve the project, making the results more accurate and useful for the business.
3. *Focus on Business Objective*: The focus stays on the goal defined at the earliest stage. Therefore, data mining efforts align with business needs only. If the target is to understand customer satisfaction towards the provided service, the mining process will be directed at sentimental analysis.
4. *Neutrality*: CRISP-DM is an agnostic framework that can be used with any data mining tools or software such as Python, R, SAS, or any other without giving any favour. *Standardization*: Imagine a team working on predicting customer churn for an e-commerce company. They follow CRISP-DM's standardized steps: understanding the business problem, exploring data, building models, evaluating performance, and deploying the best model. This common language ensures everyone is on the same page, even if they use different tools. It's like having a shared map for a journey—everyone knows the route, even if they drive different cars!

CRISP-DM : Implementation

Following up the Fig. 1, CRISP-DM outlines a six-phase process for data mining projects:

1. ***Business Understanding*:** This is the initial phase of data mining where the subject is to understand the problem to be solved. Here we go with the questions like what is our business about? what is the goal of it? what is the background of the problem to work on? and then define the proper project plan. It can be understood better by taking a real case of the financial company HSBC. The bank aimed to identify fraudulent transactions in real-time, to reduce its financial loss and enhance security. Therefore, the aim is to develop a robust fraud detection system to protect customers' accounts. By understanding this particular problem, we would able to establish an interactive process of discovery with the data understanding phase. The accomplishment of this phase is to figure out what we are trying to do. In this, creativity often plays a massive role.

   The crucial steps to follow up are:

- To determine the business objective: Clarify business goals and success criteria.
- Compile business background: Gather information about the current situation, resources, and problem areas.
- Define Objectives: Convert business goals into data mining tasks.
- Assess the Situation: Evaluate the available data and resources, identify risks, and plan accordingly.

2. *Data understanding*: It is a data-acquiring phase made to fulfil the sufficiency of business objectives. This phase requires the data science team to acquire the data from internal/external/other sources and look over it cautiously. The first task is to understand the data properly including its strengths and limitations, and then to analyse if the data exactly matches the problem we are trying to find out. In the case of HSBC, to target the detection of fraudulent transactions, it can analyse the transaction data, customer profiles and historical fraud cases. In simple words, gather and look up the data you have.

External or other data-gathering sources cost money, and some have even had availability issues. The Data scientist needs to evaluate the cost and benefits of various potential data sources. To understand customer buying behaviour, their purchase history is to be collected and analysed for patterns. This is an example of customer segmentation.

Some specific steps to follow up are:

- *Collect Initial Data:* Gather all relevant data. This could be databases, spreadsheets, APIs, or other data repositories.
- *Describe Data:* Summarize the main characteristics of the data.
- *Explore Data:* Analyse the data to find patterns and insights.
- *Verify Data Quality:* Ensure data is accurate and complete.

3. *Data Preparation*: It is the data-preprocessing step in advancing towards the data mining lifecycle. Here, the raw data is to be converted into finished data to move further for model training. Analytics techniques are required to make datasets in the forms from how the data was initially provided. Some conversions may be necessary to maintain data consistency. For example, for instance, fix any typos in the customer names and make sure the dates are in the same format.

Common data preparation methods may include:

i. Data conversion:
    - convert data into a structured tabular format, such as a spreadsheet or database table.

ii. Data Cleaning and Preprocessing:
    - Handling Missing Values: Identify missing data and decide how to handle it (e.g., impute values or remove rows).
    - Outlier Detection: Detect and address outliers to handle skewness.
    - Data Transformation: Convert data types, normalize, or scale features.
    - Feature Engineering: Create new features or derive meaningful ones from existing data.

iii. Exploratory Data Analysis (EDA):

- Descriptive Statistics: Calculate summary statistics (mean, median, etc.).
- Visualization: Plot histograms, scatter plots, and other visualizations to understand data distributions and relationships.

iv. Feature Selection:

- Identify Relevant Features: Choose features that contribute most to the problem.
- Dimensionality Reduction: Reduce the number of features (e.g., using PCA).

v. Data Splitting:

- Divide data into training, validation, and test sets for model evaluation.

In short, select relevant data, clean it, construct new data if needed, and format it properly.

It is the most time-consuming of all the six stages models. A whole book can be written on this very topic.

Some specific steps include:

- Select Data: Choose the data that will be used for analysis.
- Clean Data: Correct errors and handle missing values.
- Construct Data: Create new attributes or variables if needed.
- Integrate Data: Combine data from different sources.
- Format Data: Ensure data is in the right format for analysis.

4. ***Modelling*:** At this stage, data is ready to delve into the zoo of data mining simple and advanced mining techniques. It is possible to get various patterns but not all the patterns are valid in the evaluation stage. If the creation of a baseline model does not fit as far as our expectation on the predefined business objective goals, the cycle goes back to the data preparation step. So, it is an iterative process.

Choosing modelling techniques, building models, training & testing them and adjusting parameters are the core objectives of this model phase. Example: Use clustering to segment customers based on buying behaviour. In the HBSC, machine learning algorithms to detect anomalies indicative of fraud.

Steps for modelling include:

- Select Techniques: Choose appropriate modelling techniques.
- Generate Test Design: Plan how to test the models for robustness.
- Build Models: Create and train the models.
- Assess Models: Evaluate model performance.

5. ***Evaluation***: The purpose of this stage is to assess data mining results both from a qualitative & quantitative perspective and the key aspects to determine the results for both are justiciable and feasible. It is encouraged to check if the patterns we found have actually helped to solve our problem. For example, does our model really help to predict future sales? It usually ends up being cheaper and safer than simply skipping the stage and going directly for model deployment.

   Once we realise that the evaluation stage results, we get versus what we are trying to find do not match, there can be a possibility of a lack of communication between stakeholders. Let's discuss this situation in detail. Imagine a junior data scientist who has created a model to predict the monthly churn in a startup. The sales team found that the recommendations were irrelevant. This issue happens because of these two reasons:

   - Misaligned Goals: The data scientist built a model to predict churn within the next month, focusing on short-term risk.
   - Business Reality Ignored: The sales team knew customers signed long-term contracts (3-5 years).

   Therefore, the Monthly churn wasn't relevant. Hence, we can say that the model wasn't wrong, but it addressed the wrong question. This situation makes us understand two major aspects:

   i) Business understanding is the key means to proper dialogue between different stakeholders (Sales team and Data Science team) to take place for filling the information gap. "Will this client renew their contract?" should have been the question.

ii) Focus must be on the right metrics implies models should predict metrics relevant to the business. In this case, long-term contract renewal was more important than monthly churn.

Many times, just because the model passes the lab test doesn't mean that it is going to work in real life from a practical standpoint. The challenges such as budget limitations, data limitations, overfitting on the training data, and sometimes missing context, external factors (sudden outbreak of lethal disease can surpass the general public health prediction which lab results give) etc. can vividly impact the performance of the model. So it must also be thinkable at this stage.

Major steps included in this phase are:

- Evaluate Results: Review model results and compare them against the business objectives.
- Review Process: Check the entire process for any issues or improvements needed.
- Determine Next Steps: Decide on the next actions based on the evaluation.

6. *Deployment*: In simpler words, putting the findings to use in the real world. Whether it is about integrating the model into the bank's transaction monitoring system or integrating insights into customer service strategies and product development, an improvised and updated model is brought to the public domain and accessible to all. Necessary resources, such as the time, personnel, and tools needed for this implementation. Detailed documentation is prepared to ensure that users understand how to utilize the model. Some key performance indicators (KPIs) help in monitoring and evaluating the model's performance effectively. For ex. Scikit learn documentation, IBM SPSS modalar CRISP-DM guide etc. After the deployment stage, teams go back to the business understanding stage to really internalize all of the insights that were drawn from the stage process. To keep the model effective and accurate, we need a plan to update it regularly. We should also create a simple report that sums up the project's goals, methods, results, and how it was implemented, along with suggestions for improvements. This overall strategy helps ensure the model works well and stays up-to-date.

We need a plan to update the model regularly to keep it accurate and useful. Additionally, we should prepare a simple report that covers the project's goals, methods, results, and how it was implemented. This report should also include ideas for future improvements. This approach ensures that the model is effectively integrated, maintained, and continuously improved for optimal performance.

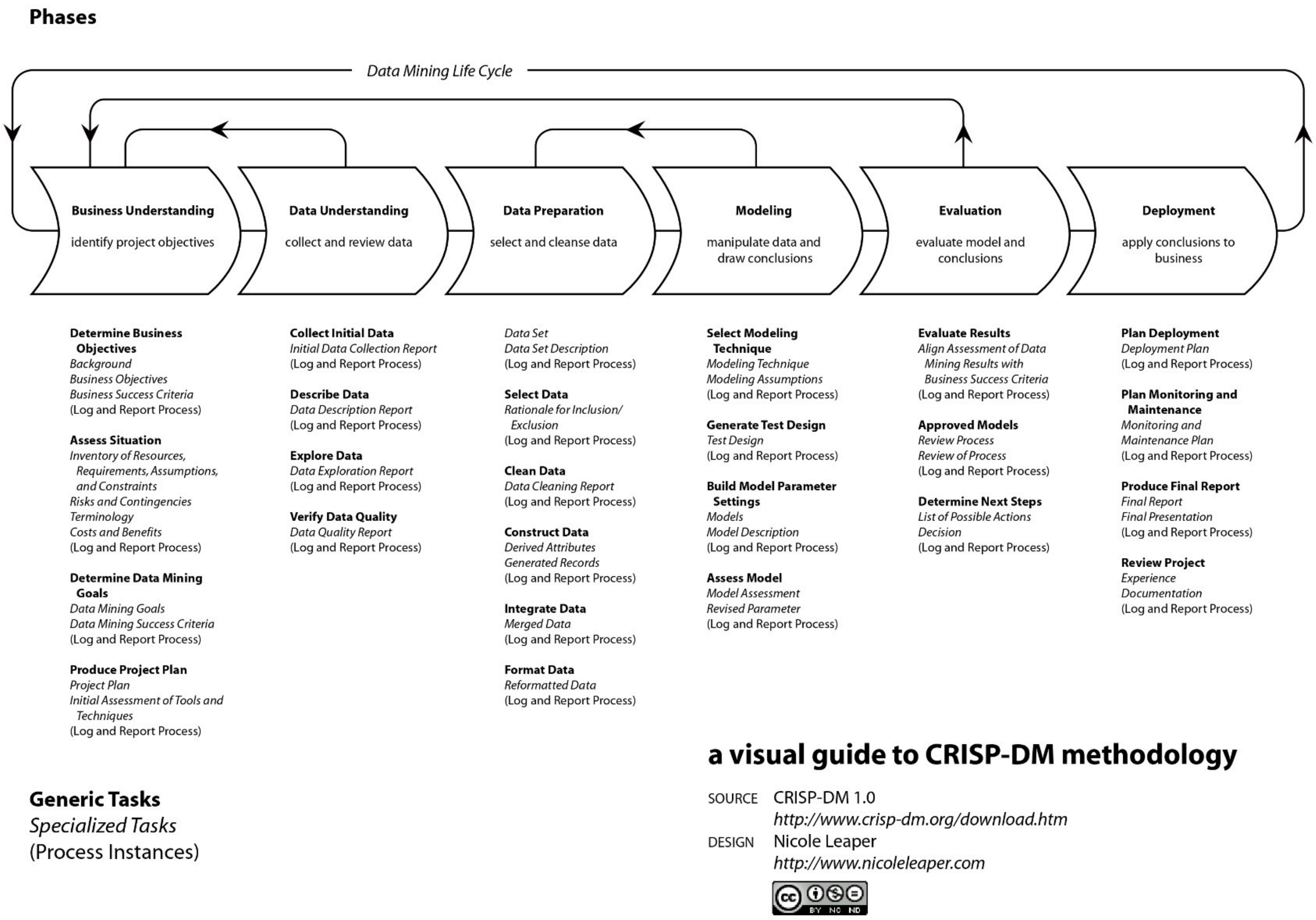
The crucial steps in this phase are:

- Plan Deployment: Prepare for implementing the model.
- Monitor and Maintain: Set up processes to monitor model performance and maintain it.
- Produce Final Report: Document the results and the process.
- Final Presentation: Prepare a presentation of the findings and recommendations.

Current Relevance:

There are other data mining processes in the market used such as SAS institutes SEMMA (Sample, Explore, Modify, Model and Assess) which emphasizes sampling, EDA, feature engineering, modelling and model assessment but due to the CRISP-DM's robustness, flexibility, and practicality approach, it acquires more acceptance and popularity than others. It has made planning, plotting and working environment of data mining easy.

CRISP-DM in a Nutshell[1]:



Phases

Data Mining Life Cycle

Business Understanding
identify project objectives

Data Understanding
collect and review data

Data Preparation
select and cleanse data

Modeling
manipulate data and draw conclusions

Evaluation
evaluate model and conclusions

Deployment
apply conclusions to business

Determine Business Objectives
Background
Business Objectives
Business Success Criteria
(Log and Report Process)

Assess Situation
Inventory of Resources, Requirements, Assumptions, and Constraints
Risks and Contingencies
Terminology
Costs and Benefits
(Log and Report Process)

Determine Data Mining Goals
Data Mining Goals
Data Mining Success Criteria
(Log and Report Process)

Produce Project Plan
Project Plan
Initial Assessment of Tools and Techniques
(Log and Report Process)

Collect Initial Data
Initial Data Collection Report
(Log and Report Process)

Describe Data
Data Description Report
(Log and Report Process)

Explore Data
Data Exploration Report
(Log and Report Process)

Verify Data Quality
Data Quality Report
(Log and Report Process)

Data Set
Data Set Description
(Log and Report Process)

Select Data
Rationale for Inclusion/ Exclusion
(Log and Report Process)

Clean Data
Data Cleaning Report
(Log and Report Process)

Construct Data
Derived Attributes
Generated Records
(Log and Report Process)

Integrate Data
Merged Data
(Log and Report Process)

Format Data
Reformatted Data
(Log and Report Process)

Select Modeling Technique
Modeling Technique
Modeling Assumptions
(Log and Report Process)

Generate Test Design
Test Design
(Log and Report Process)

Build Model Parameter Settings
Models
Model Description
(Log and Report Process)

Assess Model
Model Assessment
Revised Parameter
(Log and Report Process)

Evaluate Results
Align Assessment of Data Mining Results with Business Success Criteria
(Log and Report Process)

Approved Models
Review Process
Review of Process
(Log and Report Process)

Determine Next Steps
List of Possible Actions
Decision
(Log and Report Process)

Plan Deployment
Deployment Plan
(Log and Report Process)

Plan Monitoring and Maintenance
Monitoring and Maintenance Plan
(Log and Report Process)

Produce Final Report
Final Report
Final Presentation
(Log and Report Process)

Review Project
Experience
Documentation
(Log and Report Process)

Generic Tasks
Specialized Tasks
(Process Instances)

a visual guide to CRISP-DM methodology

SOURCE CRISP-DM 1.0
http://www.crisp-dm.org/download.htm
DESIGN Nicole Leaper
http://www.nicoleleaper.com

[1] https://exde.wordpress.com/wp-content/uploads/2009/03/crisp_visualguide.pdf