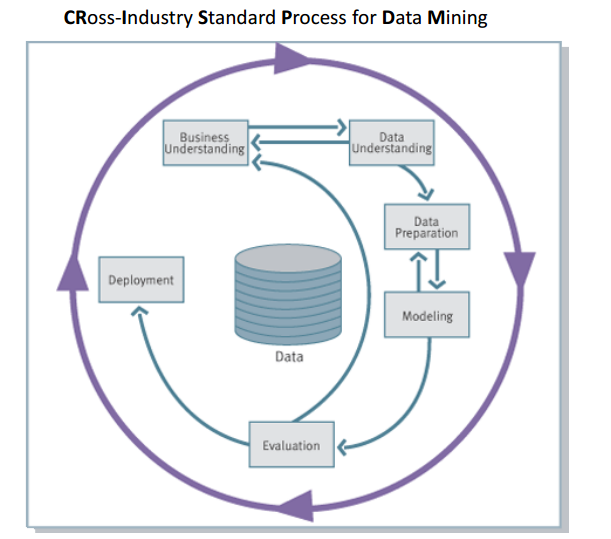
**CRISP-DM**

**Introduction:**

CRISP-DM stands for cross industry standard process for data mining. It is a comprehensive data mining methodology and process model that provides anyone from novices to data mining experts with a complete blueprint for conducting a data mining project. CRISP-DM breaks down the life cycle of a data mining project into six phases. These 6 high-level phases of CRISP-DM are still a good description for the analytics process.



**Phases of CRISP-DM:**

1. **Business Understanding**

First and foremost step in Analytics project is understanding client’s business to formulate problem statement. Once we define problem statement then we can drive data accordingly.

Focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the below objectives

• What the client really wants to accomplish?

• Uncover important factors (constraints, competing objectives)

1. **Data Understanding**

We start with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover initial level of insights into the data or to detect interesting subsets to form hypotheses for hidden information.

Collect data

* List the datasets acquired (locations, methods used to acquire, problems encountered and solutions achieved).
* Describe data
* Check data volume and examine its gross properties.
* Accessibility and availability of attributes. Attribute types, range, correlations, identities.
* Understand the meaning of each attribute and attribute value in business terms.
* For each attribute, compute basic statistics (e.g., distribution, average, max, min, standard deviation, variance, mode, skewness).

Explore data

* Analyze properties of interesting attributes in detail
* Distribution, relations between pairs or small numbers of attributes, properties of significant sub-populations, simple statistical analyses.

Verify data quality

* Identify special values and catalogue their meaning.
* Does it cover all the cases required? Does it contain errors and how common are they?
* Identify missing attributes and blank fields. Meaning of missing data.
* Do the meanings of attributes and contained values fit together?
* Check spelling of values (e.g., same value but sometime beginning with a lower case letter, sometimes with an upper case letter).
* Check for plausibility of values, e.g. all fields has the same or nearly the same values.

1. **Data Preparation**

It covers all activities to construct the final dataset from the initial raw data. Converting raw data into analytical dataset is very important. Quality of cleaned data will impact on model performance. Data preparation tasks are likely to be performed multiple times and not in any prescribed order. Tasks include table, record and attribute selection as well as transformation and cleaning of data for modeling tools.

Select data

1. Reconsider data selection criteria
2. Decide which dataset will be used
3. Collect appropriate additional data (internal or external)
4. Consider use of sampling techniques
5. Explain why certain data was included or excluded

Clean raw data using pre-processing techniques like

1. Missing value Analysis
2. Outlier Analysis
3. Feature Engineering
4. Scaling numerical variables
5. Sampling
6. Smooting
7. **Modeling**

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

You can use Machine learning algorithms or statistical methods to build analytical model. Some of the famous algorithms are Decision tree, random forest, KNN, naïve Bayes, K means, linear regression and logistics regression.

1. **Evaluation**

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

Points to consider:

* 1. Understand data mining result. Check impact for data mining goal.
  2. Check result against knowledge base to see if it is novel and useful.
  3. Evaluate and assess result with respect to business success criteria
  4. Rank results according to business success criteria. Check result impact on

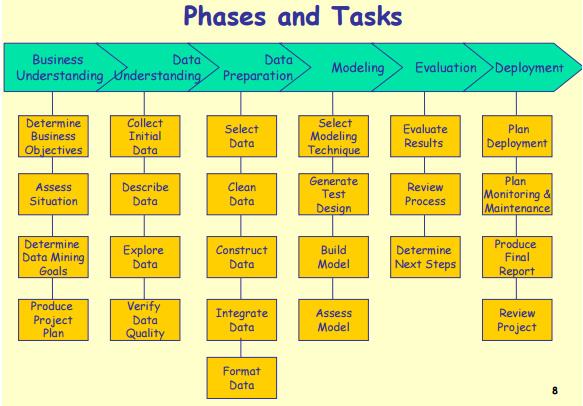
1. **Deployment**

In this phase we need to determine how the results to be utilized. The knowledge gained will need to be organized and presented in a way that the customer can use it. However, depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.

There are multiple methods of model deployment

1. Traditional method using schedulers
2. Online method using AWS, online Tools

Pictorial representation of each of the phases and its functionality



**Interview Questions:**

1. Describe each phase and its implementation in real time projects?
2. Explain few statistical tests used in data preparation?