CRISP-DM: Towards a Standard Process Model for Data

Introduction:

Data mining is a process used to extract usable data from a larger set of any raw data. This paper emphasizes on how important it is for the data mining industries to use a standard process model.

Literature Review:

CRISP-DM stands for cross-industry process for data mining. The CRISP-DM methodology provides a structured approach to planning a data mining project. It is a robust and well-proven methodology.

CRISP-DM breaks the process of data mining into six major phases:

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

Business understanding is crucial because it helps to clarify the goal of the customer. It basically means to develop an understanding from a business perspective and design a plan accordingly in order to achieve the goals.

The process of Data Understanding starts by collecting data, then getting familiar with the data, to identify the problems, to gain firsthand knowledge and to develop understanding about hidden information. At the end of this phase, a decision on the use of the data mining results is reached.

Data preparation is the process of cleaning and transforming raw data into the form suitable to processing and analysis. It is an important step in data mining and often involves making corrections to data and combining the collected data sets to enrich data.

Data modelling consists of a group of processes in which multiple sets of data are combined and analysed to find out the relationships or patterns. In data mining you search for valuable and relevant data to solve the marketing question. You use that data as a basis to build a model to predict future patterns.

Evaluating the performance of a data mining technique is a fundamental aspect of Crisp-DM Model. However, Evaluation measures may differ from model to model. Evaluation is the determination of how certain attributes will behave in the future. For example, how much sales a store will generate in a given period.

The last step in the data mining process is to deploy the models to a production environment. Deployment is important because it makes the models available to users so that you can use the models to create predictions and take business decisions.

Conclusion:

The Crisp-DM Model has made it much simpler to write process models by following the check lists.

Crisp-DM is and ideal approach when it come to repeating certain processes and for projects that are planned on a large scale with several people involved.