Lecture 3: Data Mining Process

BITS Pilani

Pilani | Dubai | Goa | Hyderabad

Data Mining Process

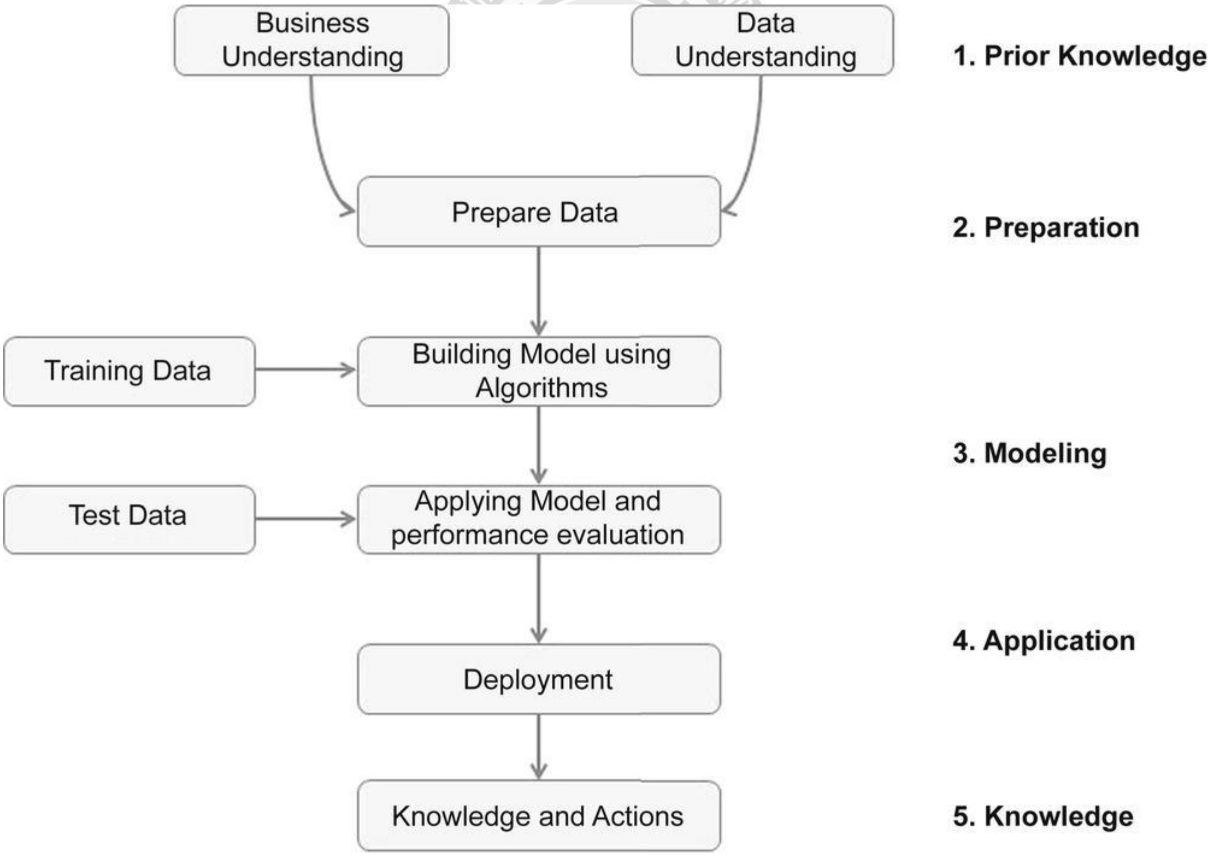
Standard data mining process involves following stages.

1. Understanding the problem - We need to have a good in-depth understanding of the business domain if we want to build a useful data mining system.
2. Preparing the data (samples) - Data which is available in the real world (raw data) is not ready for use. So we need to do pre-processing of data and make it ready for use.
3. Developing the model – This involves constructing the data model.
4. Applying the model - Applying the model on a data set to see how the model may work in real world.
5. Production deployment - After we are confident that model works properly.

Standard process framework for data mining work is called as CRISP (Cross Industry Standard Process for Data Mining). This framework was developed by a consortium of companies involved in data mining.

Generic Data Mining Process

Generic data mining process stages are as shown below.



- ✓ First step is to have an understanding of both business and data. Then,
- ✓ Pre-processing of data.
- ✓ Building Model. This may require the use of training data and test data.
- ✓ Apply and evaluate the model. Once the model is evaluated for its response and behaviour, it is deployed.
- ✓ Post deployment, start generating the knowledge from the data and use that knowledge.

Prior Knowledge

Below is some of the prior knowledge that is required for data mining.

During data mining process, tools used may receive numerous patterns and may analyse multiple alternatives. In such situation, one may get some false or trivial patterns. So it is important to instruct the machine or mechanism to filter and generate useful patterns. We need to know how the data is being collected, that may give us a perspective on how to handle it. For example: If we know that it is manually entered data and somebody has entered 'year of birth' instead of 'age'. Instead of entering the age of say 30 years, the person has entered 1980 as the age. Such type of errors is quite possible when it is being entered manually. Some of the systems may have been existing for a long time and have high-quality data. In some other systems data might be collected through a very simple process where there is a significant possibility of the noise, errors, etc.

We need to know how the data is being collected, that may give us a perspective on how to handle it.

One needs to understand that among the patterns, some of the patterns are not correct from the business perspective. For example: Bank offers the interest rate to the customers for a loan depending on the credit rating of the customer. Here credit rating determines the interest. But when you provide the data that includes both interest and credit rating to the data mining algorithms, it may come to a conclusion that interest rate determines the credit rating. But that is not the business logic. Understanding of data will help in determining useful pattern and pattern you should discard.

Data Preparation

Following needs to be taken care during data preparation stage.

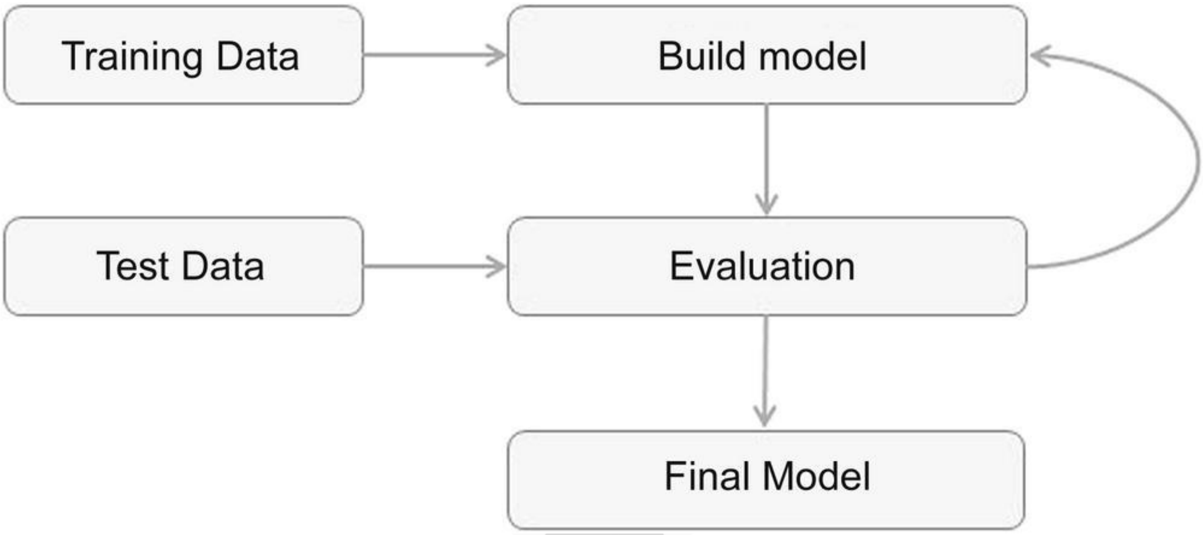
1. Understand the data - Understand the description of the data such as the mean, median, mode, standard deviation and range for each attribute. This would help to determine if it requires some normalization or some other form of restructuring.
2. Data quality - There may be duplicate records, lot of noise, missing values etc.
3. Data Types and Conversion - May require to do the conversion of the data because, data is in numerous units and currency formats.
4. Transformation
   a. Dimensionality reduction - Some dimensions may have strong correlation that it is redundant to incorporate other dimensions. In such cases, we may remove some dimensions to reduce the data.
   b. Numerosity reduction -Instead of having the entire data we may work with data at aggregate level or on some other form of sampling for reasonable understanding of the data. In that case we may have to identify the outliers

upfront. For example: In place of age, year of birth has been entered, that may be an outlier and we may need a mechanism to detect it and eliminate it.

These kind of activities put together will constitute the data pre-processing or the data preparation that needs to be done before data mining.

Modeling

A model is the abstract representation of the data and its relationships in a given data set.



- ✓ Training data –Identifies consequences with respect to particular events in the past. Based on this, the model may predict for the future.
- ✓ Test data – Sometimes test data is based on past experience. In such cases, test data would help in evaluating how good the model is.

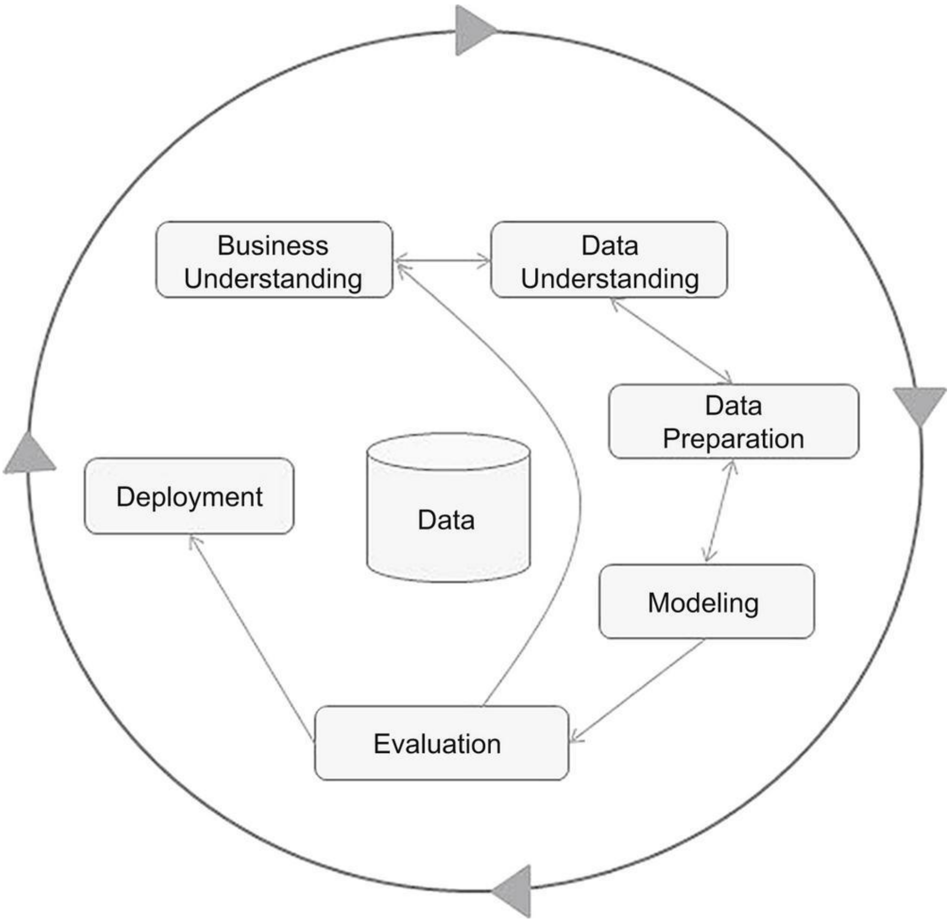
In a way, both the training data and test data assists in refining the model for accurate predictions or accurate descriptions of the data. Thus one can arrive at the final model.

Application of Model

Following needs to be taken care during model deployment stage.

- ✓ Production Readiness
  - o Need to ensure model is ready to be integrated with current environment
  - o Model must be capable to handle diverse situations
- ✓ Timeliness - Need to respond real-time. For example: In Managing furnace temperature, may need to respond within split seconds, so that you can take quick action and be safe.
- ✓ Re-modeling– For new data, does it mean that we need to create a different model or existing model will work sufficiently well with an incremental data.
- ✓ Assimilation–One should be able to assimilate the knowledge gained from the model and use it for the benefit of organization.

CRISP Data Mining Framework



CRISP is considered to be one of the most popular data mining methodology. Majority of the data mining or data science projects use this methodology as the framework. It considers data mining as a cycle. That means, it iterates data mining process stages starting from business understanding to evaluation till it meets the expectation. Finally, data mining model is deployed.

Issues/Challenges in Mining Methodology

Below are some of the issues and challenges with respect to the data mining methodology.

- ✓ Mining new kinds of knowledge – Mining should help in identifying new kinds of knowledge different from what is used earlier.
- ✓ Mining knowledge in multidimensional space - It should be possible to work within the multi-dimensional space with a very large number of dimensions.
- ✓ Data mining an interdisciplinary effort – Need to look at the information from the perspective of domain knowledge, technical knowledge, Mathematics, hardware, algorithms etc. All the interdisciplinary knowledge has to come together to create data mining.
- ✓ Boosting the power of discovery in a networked environment - Most data objects reside in a linked or interconnected environment, whether it be the web, database relations, files or documents. Semantic links across multiple data objects can be used to its advantage in data mining.
- ✓ Handling uncertainty, noise, or incompleteness of data - Errors and noise may confuse the data mining process, leading to the derivation of erroneous patterns. Data cleaning, data preprocessing, outlier detection and removal, and uncertainty reasoning are examples of techniques that need to be integrated with the data mining process.
- ✓ Pattern evaluation – Need to identify the good patterns and remove the noisy patterns.

Issues/Challenges in User Interaction

Below are some of the issues and challenges with respect to user interaction in data mining process.

- ✓ Interactive mining
  - One would like to have the data mining solutions to be interactive so that user can try with sample data.
  - It should facilitate user interaction with the system.
  - It should help the user in achieving the results that he or she is looking for.
- ✓ Incorporation of background knowledge
  - Data mining should incorporate the domain knowledge.
- ✓ Ad hoc data mining and data mining query languages
  - Data mining should allow ad hoc approach. Typically when it comes to the business problems, they will not be the same. Operational problems likely to be repetitive. When it comes to knowledge related problems, which is meant for the strategic purposes, it requires an ad hoc approach to solve the new problem. For example: Business which was five years back is not the same today due to change in competition level. Expectations from the customer have changed. In the new context, one should not expect to completely change the system. It should take the inputs in an ad hoc manner and should continue to work.
- ✓ Presentation and visualization of data mining results
  - Data mining results should be in a good visual format so that, the customer or user can quickly recognize the patterns.

Issues/Challenges in Efficiency and Scalability

Below are some of the issues and challenges with respect to the efficiency and scalability of data mining process.

- ✓ Efficiency and scalability of data mining algorithms
  - Data mining deals with large volumes of data. Hence efficiency and scalability of data mining algorithm is important.
- ✓ Parallel, distributed, and incremental mining algorithms
  - Data mining requires parallel, distributed and incremental capabilities, since data is large and widely distributed. If algorithm is also distributed and can work on the parallel large number of machines, it can solve the problems in a required timeline. For that reason the distribution is important.
- ✓ Cloud computing and cluster computing
  - It is necessary that the data mining solutions also work in environments such as clusters and cloud models. It is better to use distributed power rather than building a very powerful, single machine. Trend is that it is becoming easier to create a large number of machines of lower cost than single machine with a huge power.

Issues/Challenges in Handling Diversity of Data Types

Below are some of the issues and challenges in handling diversity of data types in data mining process.

- ✓ Handling complex types of data
  - Data mining deals with wide varieties of data. Both structured and unstructured data can be generated by various machines and devices. Hence, data mining process should handle all the complex types of data.
- ✓ Mining dynamic, networked and global data repositories
  - Data mining process should be able to work with multiple repositories distributed across multiple locations all over the world through the network.

Issues/Challenges to Society

Below are some of the data mining issues and challenges to the society.

- ✓ Social impacts of data mining
  - Data is sometimes collected without the individual's knowledge and may get misused. Hence, data mining should provide the necessary transparency.
- ✓ Privacy-preserving data mining
  - Privacy and rights of the individuals should be protected.
  - Data mining should help you in preventing the intrusions and cyber-attacks or such undesirable events.
- ✓ Invisible data mining
  - Data mining has to be applied in a positive way with respect to the society. Data mining many a times used without the knowledge of the people. For example: Data related to the people's behaviour, way the transactions they carry out, purchasing habits or even their personal data getting captured without the knowledge of people. Hence it is necessary that the data mining community should be alert and handling data in a responsible way.