Session Summary

CRISP Data Mining

Data analytics is typically aimed towards solving a problem to generate insights from data. Typically, you have to go through the following steps to obtain a solution for a data analytics problem:

- Start by developing a business understanding around the problem. This is the first step because understanding the problem and its impact on the business is very crucial.

- After developing the business understanding, you have to recognise and understand the various data sets or sources of data which can be leveraged to solve the problem at hand. This is the second step in the CRISP DM framework and is called data understanding.

- The third step is called data preparation. It is the most important and time-consuming step in the entire analysis. Once the data sets have been figured and understood, they need to be prepared in various ways for the analysis to happen.

- The fourth step – modelling – is the most interesting step of the entire CRISP DM process. After preparing the data, models are built on it to solve the business problem at hand and generate insights from the data.

- Before the model is deployed, it needs to be evaluated to check its accuracy, usefulness and to understand how well it is performing. Model evaluation is a continuous step and needs to be performed on the final model for the model to be valid over time.

- The final step in the framework is model deployment. Once the model passes the evaluation criterion, it is ready for deployment.

Business Understanding

The first stage of the framework is to develop a **business understanding.** For this, you have to carry out two steps:

1. Determine the business objective
2. Identify the goal of the data analysis

Determining the business objective is of utmost importance. Until the business objectives have been finalised, the data cannot be collected or worked upon. Then, you have to determine the goals of data analysis, and work towards achieving them in the CRISP-DM framework.

Consider IPL as a business. Here, the business objective could be either to win or to maximise profits. It is very important to have a well-defined business objective before you move on to the analysis. Then, you can identify the goals of the data analysis problem. If the business objective is to win, the goal of the data analysis could be to identify the most scoring players, or the bowlers with the most wicket. If, on the other hand, the business objective is to maximise profits, the goal of the data analysis could be to identify the popular players that attract funding. It is very important to define the business objectives clearly and then the goals of the data analysis problem.

Data Understanding

This stage comprises of four key steps to understand the available data, and identify new relevant data in order to solve the business problem.

- Collect relevant data: First, you need to identify and collect the right set of data sets that can be used for the analysis. They can be available within the firm, or you may have to collect the data from other sources such as open source repositories or government data sets. For example, to understand the investments across sectors globally, you can used the structured CrunchBase data set. You could also collect unstructured data, like news articles or twitter feeds about acquisitions, or semi-structured data such as annual reports, if it is relevant.

- Describe data – for explicit information: Once you have identified the data set, you need to describe its contents and explore insights to better understand the data and its business implications. To describe the CrunchBase data, we can create a data dictionary that lists down the types of variables (e.g. sectors, company names, etc.), the number of records, and the types of analysis.

- Explore data – for implicit insights: To explore data, you can plot simple graphs on Excel/R, e.g. to understand the range of funding received by companies.

- Verify data quality – to remove errors: Once you have understood the data structure, you can next examine the quality of data and address various factors:
  - Is the data complete, does it cover all the cases and records?
  - Is the data correct, or does it contain errors and, if there are errors, how common are they?
  - Are there missing values in the data? If so, how are they represented?
  - Where do the missing values occur, and how common are they?

  All the issues with the data are to be understood here so that they can be taken care of in the next steps. For example, in the CrunchBase data set, funding for a certain company is inaccurately reported as negative $-20K or $2Bn, when the normal reported funding ranges between 0 - $20M.

Data Preparation

Data preparation is the most important and time consuming step of the CRISP-DM framework and is carried out in the following steps:

- Data which is of interest is selected. It may be spread across different files or sources.

- Then, you integrate data from these multiple sources. For example, in the CrunchBase data, information is spread across different files and we integrate them together to solve a particular business problem.

- After data integration comes the data preparation stage. Missing value treatment, outlier treatment, and removing the erroneous values are a few major components of the data cleaning process.

- Constructing the data is the next step, which involves the creation of new features originally not present in the data set. The purpose of this step is to increase the information we get from the data and to reduce the number of variables originally present in the data set.

- Finally, format the data. In this step, no changes are made to the data in the data set, but to its structure. The variables present in the data set are set to the correct format as required by the analysis tool.

It is important to remember that any minute mistake in any of the following steps could lead to a waste of effort. It would amount to the classic problem of ‘garbage in, garbage out’.

Data analysts have always pointed out the time and effort that it takes to prepare data. By some estimates, an analyst spends around 70% of their total time in just preparing the data for analysis.

Data Modelling

Modelling activity in the CRISP-DM framework involves two major tasks. The first task is to understand the problem domain and select the appropriate family of models that is suitable for solving the problem at hand. The second task is to select appropriate algorithms for creating the model from the chosen family of models.

How will you teach a machine to choose an IPL team that maximises your chances of winning?

The algorithms identify patterns in data and learn which parameters are the most important in reliably predicting a team’s performance, like batting average, captaincy score, strike rate, and wickets. The chosen parameters are given as inputs to the model which gives the output we are interested in – whether a given team will win or lose. In fact, some models also use opinions of experts such as coaches and past players to include subjective details, such as leadership and team spirit, along with the hard statistics.

Model Evaluation and Deployment

Model Evaluation

The predictive models can be tested to assess their effectiveness in solving the problem. This is the fifth stage of the framework – model evaluation. Modelling and evaluation together is an iterative process in which the models are tweaked until satisfactory evaluation results are obtained.

Model Deployment

This is the last stage of the framework, where the model is translated into a business strategy. Business data is fed into the model and the model results are used to inform business decisions on an on-going basis.

The CRISP-DM framework does not end at the last stage of model deployment. The important thing to note is that CRISP-DM is an iterative process. For example, your data understanding can enhance your business understanding. Similarly, after model evaluation, if the model does not perform great, you will have to go back to the data preparation stage, and then develop the model again.