

## Where's My Hard Hat? An Introduction To The Data Mining Process

Trevor Hodges demystifies the art of data mining, and suggests how you can shorten your analysis time by moving your data preparation to the Data Warehouse...



### Introduction

There I was in a previous life minding my own business when suddenly I was given line management responsibility for the Data Mining team in addition to the Business Intelligence (BI) team.

The data mining guys worked independently of the BI team very much as a standalone operation using unfamiliar toolsets and conjuring up 'customer scores' generated using complex algorithms. I had a vague understanding of data mining having seen the customer 'scores' loaded to the customer data warehouse; these were used by the marketing team to select prospects for outbound campaigns. How these scores were arrived at was a complete mystery, but there was one thing I did know: the development of these models took an age and didn't always deliver the expected business results.

Having little idea of what these data mining wizards with tongue twisting degree titles did, I decided to roll my sleeves up and get stuck in.

### Not as mysterious as you think!

The data mining process is not as mysterious as some make out and once you get over the 'you're not a statistician so you won't understand' mentality, a data mining project is not unlike any other IT project – although it is often seen (wrongly) as being apart from IT projects because of its academic origins.

The key difference between traditional query and reporting (Q&R) and data mining is the role of the database. In Q&R the database passively responds to questions posed by the user whereas in data mining the database automatically organises data into useful

information using pattern-finding algorithms. Data mining techniques drop into two categories:

#### Discovery methods

- Used to find patterns and relationships in data, for example segmenting your customer base. Clustering, associations and sequence are the typical techniques

#### Predictive methods

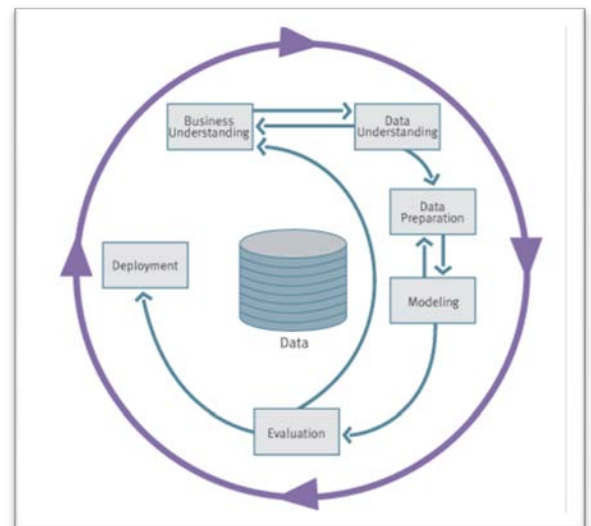
- Used to predict a value, for example the propensity of an individual to respond to a marketing campaign. The techniques used are classification and regression

But how do you go about delivering these data mining models in practice? A good place to start is to refer to the Cross Industry Standard Process for Data Mining (CRISP-DM for short) -

<http://www.crisp-dm.org/>

Figure 1 shows the CRISP DM lifecycle for a data mining project.

The methodology provides a non-proprietary process model based on real-world experiences of leading data mining experts.



The diagram in Figure 1 shows the six high level phases of the life cycle for a data mining project. In short:

#### Business Understanding

- Understand the business problem, the objectives and identify the analytic approach to deliver the desired results

#### Data Understanding

- Identify data quality problems, and discover data insights

#### Data Preparation

- Construction of the dataset in the format to meet the requirements of the analytic approach

#### Modelling

- The statistical modelling technique identified from the requirements is applied

*Evaluation*

- Evaluate the model; is it fit for purpose?

*Deployment*

- Knowledge is presented to the customer for action or information

**Who owns data mining marts?**

The actual data mining or modelling phase is only a small part of the overall lifecycle. By far the largest is the data preparation phase. This delivers the analytic dataset (data mining mart) and is a critical part of the project lifecycle for any solution: it can influence the accuracy of the results, and is the largest effort – 70% of the overall project is not unknown.

When one considers the steps that are required to deliver the mart, it is easy to understand why it takes so much effort:

- We have to select the data that is relevant to the requirements of the project
- The data quality has to be verified and improved where necessary
- Data from multiple sources is de-duped, merged and formatted
- Derivations and/or transformations are applied to produce analytic variables

These steps are not dissimilar to the design of subsystems for a traditional ETL data warehouse architecture, yet in many organisations the DW/BI professionals job is done once the data has reached (our) data warehouse; it is left to statisticians to prepare the data for data mining.

Consider who is best placed to create the data mining mart? Statisticians are highly skilled statistical modellers but often have basic or no SQL skills whereas we, as DW/BI professionals can deliver complex data flows and transforms using the tools at our disposal.

**Table formats for Data Mining Marts**

The data mining process requires data to be organised in ways that suit the analytical method identified in the Business Understanding phase. There are two different types of table format:

- Behavioural/Demographic
  - The behavioural/demographic data layout consists of a single record per customer with multiple columns representing the behavioural and/or demographic attributes. This format is required for clustering, regression, and classification techniques
- Transactional
  - Consists of a table with three columns Transaction, Item, and Group. This format is used by association and sequence techniques

**Traditional Data Preparation**

Statisticians have traditionally extracted data from source systems and data warehouses in file formats and then moved these across networks to create their data mining marts locally on client machines, or on servers remote from the source system – sometimes in machine rooms, often under people's desks.

In my mind this approach raises a number of significant issues:

- How efficient is this? Not very from my experience!
- What metadata is available for these datasets – data lineage, column descriptions etc.?
- Who verifies that the datasets have been pieced together correctly?
- How are the datasets backed-up?
- How secure is the data?
- Why are we duplicating data?

Why do we as DW/BI professionals allow such an important business system to be managed like this? Wouldn't data mining be better supported by moving the creation of project or domain specific data mining marts within the data warehouse environment where we could leverage existing expertise, tools and techniques developed for the DW?

**Conclusion**

A key objective for a data warehouse is to provide a single data source for the integration of analytical activity. Often data mining is considered after the DW implementation – too late to ensure the appropriate architecture/infrastructure is deployed. Ensure you don't make this mistake.

Get the wizards in early in the project lifecycle; build a cross functional team. Bring the data preparation activities within the DW/BI architecture to create a centralised source where data mining marts can be refreshed on a regular basis. Get the analysts to work within your DW/BI environment so you avoid the problems and costs caused by moving and duplicating data. Familiarise yourself with and get a working knowledge of data mining.

Of course there will always be project specific requirements, but do consider a customer domain specific mart as the highest priority in any development; I can guarantee this will be used over and over again!



*Trevor Hodges is a Business Consultant at IPL, a leading UK IT services company specialising in the delivery of intelligent business solutions. He has over 10 years of experience working for a wide range of household names helping them to exploit the full potential of their information.*

*Trevor has significant information management expertise gained across a variety of industry sectors; including finance, on-line retail and automotive. He can be reached at [trevor.hodges@ipl.com](mailto:trevor.hodges@ipl.com).*