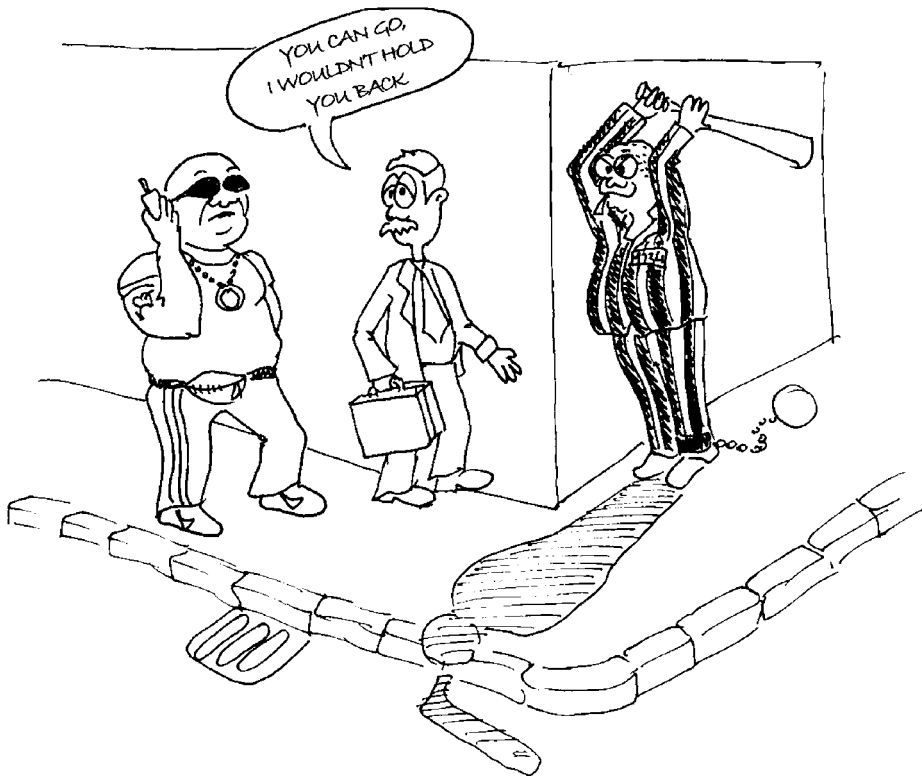


5. A Typical Data Mining Problem: Predictive Modelling

“There is no doubt that the person who has more precise knowledge of a future event than others is in a position of advantage.”



5.1. Predicting the Future

Data mining tasks – like all modelling activities – can be divided into two basic groups:

- ▶ constructive modelling (data mining); and
- ▶ predictive modelling (data mining).

|101|

The first type of modelling is used for exploring the internal operating mechanism of an existing system (economic, technical, sociological, etc.), while the second is used for *predicting* the future operations of an existing or hypothetical one. The latter is called *predictive data mining*.

According to many, the most exciting tasks for data mining are those that help us make predictions about future events. There is a simple reason for this: anything that is a puzzle will attract human curiosity. One of the many questions this curiosity raises is: “What is going to happen in the future?” It is obvious that everyone is interested in this in some way.

This question plays such a large part in our everyday lives that we seldom even consciously notice that we are speculating about the future. When we travel we monitor our environment (partly unconsciously) and we make small forecasts about the motion of the vehicle and the state of other travellers. Before leaving home we look out the window and decide whether to carry an umbrella or not. When we argue, we try to predict the other person’s reactions based on our own communications and the other’s mental abilities, habits, and state of mind. When we go bowling we analyze the course of the ball in motion and before it even hits the pin we are quite certain whether it will be a strike or not.

In a lot of cases, scanning future events is a conscious activity, particularly in competitive situations. There is no doubt that the person who has a more precise knowledge of a future event than others is in a position of advantage. The ability to know the future more accurately requires investment. This investment always takes the form of some kind of learning and analyses and it results in a ‘formula’. An infant discovering the world identifies responses and tries to create his or her own formulas which are quite simple in the beginning (“If I cry, Mom will appear and hold me.”, or “If the object is round and I push it, it will start rolling.”) A poker player uses the same method (“If he is fidgety he has a good hand.”) and so does the man who decides whether to take an umbrella with him before leaving home. How do we know if it is going to rain? We have seen clouded skies many times and we have learnt (due to an unconscious process of analysis) that the clouds and the rain are ‘cause and effect’.

Our formulas change all the time and there are two reasons for this. Firstly, these formulas improve as an effect of learning from former experiences. We use what we have lately learnt so we constantly refine our ability to predict future events. By and by, a novice driver learns how to handle a car better, and they come to know what

the effects of their actions are. A child growing up gets used to the fact that not all furry animals are as friendly to her as the ones she plays with in the garden at home. Moreover, by the time she grows up she will have created a whole system of formulae about animal behaviour, and she will know what to expect from them.

1102| On the other hand, environments can change, and if we do not adapt, old rules no longer apply. The heavy traffic of modern metropolitan areas would be strange and unpredictable to a driver who had travelled forward several decades in time and arrived in our present, since they developed their predictive abilities based on the traffic patterns of their own time.

5.1.1. Predictive Modelling

The aim of predictive modelling in data mining is to generate formulas which are based on human intelligence (but which are derived at a much faster speed than is possible for humans) which can predict the future in well-defined environments and can give a *more accurate* prognosis (compared to, for example, a human expert). The increase in speed comes from the pre-existing training samples stored in databases, while the learning algorithms can use the computer's ability to execute several million operations per second.

One should not have any illusions about the precision of any predictions, though. The limits of the accuracy of our predictions depend on the complexity of our observations. Obviously, we will be less efficient at predicting rainfall if we only take the clouds in the sky into account and omit consideration of other factors such as temperature and atmospheric pressure. For this purpose, it is essential to have a formula which can take all three aspects fully into account. This is what data mining allows: the building of the best possible predictive formula or model given the limitations imposed by the extent of our observations.

5.1.2. A Predictive Modelling Example

Let us take a concrete example! A company which loans money wants to know how reliable an applicant for a loan will be at paying back the loan. Data mining and predictive modelling can help this company in the following way: The data which is available about former loan applicants is collected (both data about those who pay instalments on time and about those who do not). A model is then built using the factors available (the debtor's age, marital status, education level, salary or the size of loan applied for, etc.). This information can 'qualify' the applicant before any credit is actually offered. The predictive model contains information which can be used to determine what weighting should be used for each factor for the pre-qualification of the debtor. The return on the invested work is manifested in the more accurate forecasts of the

company than when human experts were asked to prequalify the customers. Fewer loans will go unpaid, so the energy invested in predicting will yield abundant returns. However, one should be aware of the limitations of such modelling: often the willingness to re-pay a loan does not depend on the age, marital status, level of education, salary or the amount of credit of the debtor. Sometimes unexpected situations cause problems with payment and these ad-hoc things cannot be monitored by the company, so we cannot expect the model to predict them. One should therefore not expect to receive a perfect forecast, but one can expect better predictions than would be made without using such a model.

|103|

The credit application case described above is easily generalizable to other tasks where we have a client and a question to which we want a simple “yes” or “no” answer. Using our previous example, the question is simply this: will the client pay the money back, or not?

Since this type of problem is often the focus of business, data mining is a good tool for increasing profits. In fact, dealing with this type of task is such an important application of data mining that the authors consider the modelling methodology presented in the following section to be the primary and most representative element of data mining.

5.2. The Predictive Modelling Process

In the previous chapter about data mining methodologies, we described a standard (‘CRISP-DM’) which describes the structure of data mining processes. The process involves the following steps:

- ▶ business understanding;
- ▶ data understanding;
- ▶ data preparation;
- ▶ modelling;
- ▶ evaluation; and
- ▶ deployment.

The subsequent parts of this chapter are devoted to elaborating these elements and thereby highlighting the various steps involved in predictive modelling in detail. We show that different methods and approaches need to be used for the different phases of this process: data manipulation techniques, mathematics, creative thinking – and sometimes, intuition – all the while keeping in mind the interests of business.

Later, we refer to the fact that analysis requires many iterations. After creating a model variant, a phase of evaluation follows which reveals whether the settings used during the phases preceding modelling have yielded the expected results.