Data Quality

Introduction

Many business processes rely on the right data of high quality being available to the right people at the right time. If these conditions are not met, either incorrect decisions will be made or no decisions will be made. This may result in undesirable, unexpected and financially costly consequences or even to bad accidents and loss of life.

The imperfections of the real world mean that data observed or recorded for a defined purpose are rarely, if ever, perfect. These imperfections lead to data quality issues that must be addressed at the start of all analysis and modelling projects. The imperfections are many and varied, and include: human error; problems, inaccuracies and limitations of the measuring or recording devices; and the design of the data collection process.

Two serious data quality issues are missing data and missing objects (records). The problem of missing objects is particularly acute if the objects are a distinct group. Such skewed samples will lead to consistently poor predictions if the model is applied to a new set of data that has records that represent the distinct group. Another common problem is duplicate records for some records. The effect of duplicate records is to give undue weight in the model to these records and so bias the model towards them. Whatever the cause of the imperfections, the result is that the data are not perfect (if there is such a thing as 'perfect data') so that their quality limits the extent to which they can be used for their intended purpose.

Sorting out data quality issues can be a tedious, painstaking and unglamorous task. However, its importance cannot be over-estimated to ensure that the data are fit for purpose, and so the time spent at the start of projects understanding and preparing the data is always time well spent and of great benefit later in the projects as all the analysis and modelling are contingent on the quality of the input data. Indeed, improving the quality of the data is an essential first step in all analytics projects (see *The CRISP-DM Methodology* in Analytics Modelling).

Definition

Data quality can be defined in a number of ways. One definition is that data are of high quality if 'they are fit for their intended uses in operations, decision making and planning' (Tom Redman, (2008), 'Data driven: Profiting from your most important business asset', Boston, Mass., *Harvard Business Press*). An

alternative definition is that data are of high quality ‘if they correctly represent the real world construct to which they refer’.

The Dimensions of Data Quality

A data quality dimension is an attribute of data that can be measured or assessed against defined standards to determine the quality of the dimension. There are six data quality dimensions: completeness, uniqueness, timeliness, validity, accuracy and consistency.

Completeness: The amount of data actually available as a proportion of the maximum amount of data available.

Uniqueness: The presence of one record for each object based on how it is identified and its state at that time.

Timeliness: The availability of the data as specified when it is required.

Validity: The correct properties of the data for their intended use. Example data properties include: format (numeric, string, date); real or integer for numeric data; type (scale, ordinal, nominal); range (infinite [-∞ to + ∞] or semi-infinite [0 to +∞]).

Accuracy: The degree to which the data correctly describe the object being monitored or recorded.

Consistency: The perfect similarity between at least two representations of an object under the same conditions and at the same time.

Aspects of Data Quality

Data quality has many aspects. This document discusses four of them: outliers; missing data; inconsistent data; and duplicate data. They relate to the dimensions of data quality described above.

Outliers and Noise

There isn’t a formal definition of an outlier. One definition is that it is a value that differs significantly from its expected value and is inconsistent with respect to adjacent data. Outliers can sometimes be explained but even if they cannot be explained initially, they may still be valid observations that require

further work to understand. An example of when this can occur is if the aim is to detect infrequent or unusual observations as in fraud detection and network failures.

Noise is the random part of measurement errors. By definition, noise does not have structure and cannot be explained. It is a nuisance that should be as small as possible so that the recorded value of an attribute is as close as possible to its true value.

Missing Data

Missing data are described in detail in the document ‘PAM Analytics Projects: Database Marketing’.

Inconsistent Data

Inconsistent data in a record are data that cannot occur simultaneously. An example of inconsistent data in a record is a person of age 10 who has two children.

It is impossible to give rules for correcting inconsistent data. Sometimes the reasons for the inconsistencies are clear, for example if the start date and end date of an event have been transposed. A common cause of inconsistent data is incorrect manual entry of data. These reasons are usually due to poor company procedures, and therefore require new data quality practices and procedures to be adopted. In other cases, the reasons are less clear and so require considerable work to establish.

Duplicate Data

Databases with identical records or records that differ very slightly but refer to the same object can occur quite easily, particularly the latter instance when databases are merged. A common cause of duplicate records that refer to the same object is when name fields in at least two records have two very similar values. One way in which the duplicate records can be isolated is to select records using increasingly tight filtering criteria and then investigate which records have been deleted how the number of records changes. Continuing the name example, the presence of other fields, for example date of birth and address, can help identify duplicate records.

