Data Preparation, Exploratory Data Analysis and Predictive Asset Management

Introduction

Data preparation and Exploratory Data Analysis (EDA) are essential prerequisites for successful data mining projects. They are early stages in the CRISP-DM (CRoss Industry Standard Process for Data Mining) methodology and it is described in *The CRISP-DM Methodology* in Analytics Modelling.

PAM carries out data preparation and EDA in the Time to Failure Transformations module (see *Time to Failure Transformations Module* in PAM Modules).

Data Preparation

Before EDA can be carried out, the data must be prepared to make sure that they are fit for purpose. The first task is to identify and correct any data errors, and the data preparation work then carried out depends on the data. The reasons for data errors are many and varied, and can include:

- inaccurately recorded data
- transcription errors (this is particularly true of manually recorded data)
- data values that are obviously incorrect
- inconsistent data
- text field values that are not defined
- fields with the same name but different value labels in different input files
- incorrectly labelled fields
- missing values.

Data preparation can include the following tasks:

- correcting inaccurately recorded data
- correcting inconsistent value labels in fields with the same name but in different input files
- identifying and then correcting or removing outliers
- correcting data values that are obviously wrong
- identifying text field values that are not defined
- removing copies of duplicate fields

- removing copies of duplicate records
- changing fields' formats (ordinal, string, date, time) to those used by **PAM**
- calculating new variables
- imputing missing data where possible (infilling), for example missing asset installation dates
- sorting the records for each asset into the correct chronological order
- classifying each intervention as one of proactive non-terminal, proactive terminal, reactive non-terminal or reactive terminal (see *Time to Transformations Module in* PAM Modules)
- deduping overlapping and nested interventions (discussed below).

Exploratory Data Analysis

The term 'exploratory data analysis' was first used by John Tukey in the 1970s and is described in his book *Exploratory Data Analysis* published in 1977. He believed that 'traditional' statistics placed too much emphasis on confirmatory data analysis using hypothesis testing and not enough emphasis on studying the data to suggest hypotheses to test.

EDA does not have a formal definition. Tukey described it in a number of ways, including 'an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as those we believe to be there'. The absence of a formal definition for EDA explains why it is often regarded as a philosophy or way of working rather than as a prescriptive procedure.

The objectives of EDA are to:

- gain an understanding of the data so that the business objectives of the project can be met
- prepare the data for the analysis and modelling, always being aware of and having a clear understanding of the business objectives of the project
- calculate some key performance indicators.

EDA is a 'walk through the data' with an open mind. Most EDA techniques are graphical rather than quantitative because graphical techniques, even simple ones, can represent data, particularly large datasets, much more clearly and efficiently than quantitative techniques. Very often, the results of the EDA suggest the most appropriate modelling methods.

The techniques used in EDA depend on the data and the application. Some of the most frequently used EDA techniques are:

- plotting a number of graphs including frequency distributions, box plots, scatter plots and Pareto charts to understand the distribution and characteristics of each field (the data for Pareto charts must be calculated)
- calculating crosstabs
- calculating summary statistics for each field, for example the minimum and maximum values, and measures of central tendency, variation and skewness
- calculating correlations coefficients to identify highly correlated fields
- establishing key relationships and dependencies
- studying the dynamic profile of the data, including autocorrelation
- looking for periods of extreme outcomes and relating them to extreme external events
- determining if the data are homogeneous or if they consist of a number of groups that have high internal homogeneity and high external heterogeneity.

EDA techniques for numeric data also depend on their type (nominal, ordinal or interval). Since nominal data are just labels, for example name, they do not have any numerical properties – all the labels are equal. EDA for nominal data is therefore very limited, for example checking label consistency, and calculating frequency distributions and crosstabs. Nominal data are written as text or numeric strings. Numeric strings are numbers usually written in single quotes, for example '1' and '2', where the numbers are labels, just like text. Label consistency can be a problem when a dataset is formed by merging at least two datasets and the source labels have different naming conventions, for example the spelling of labels with respect to lower case and upper case letters. Notwithstanding these comments, nominal data are easy to use as predictor variables in all types of regression model. Indeed, **PAM** uses them extensively in the asset survival models to describe the assets, their locations and catchments, etc. and other suitable data.

Ordinal data are ordered categories where differences between the categories are not all equal. An example of ordinal data is the scale 1 (very bad), 2 (bad), 3 (neutral), 4 (good) and 5 (very good) where differences between adjacent values do not *all* have the same meaning. Ordinal data have some numerical properties and so only some of the techniques described above can be used with ordinal data. An example of ordinal data used by **PAM** is asset criticality. The ranking of the criticalities is clear but differences between adjacent values do not *all* represent the same difference in criticality. Furthermore, as the number of values of ordinal data increases and since the values are usually based on qualitative descriptions, the meaning of each value becomes harder to define and quantify clearly, leading to increased scope for confusion and error.

Interval data have the richest numerical properties and all arithmetic operations can be carried out on them. Unlike ordinal data, differences between adjacent values of interval data have the same meaning. Interval data are the most common form of data and have the widest range of modelling methods and include all real numbers between $-\infty$ and $+\infty$.

EDA, Predictive Asset Management and PAM

EDA for predictive asset management is carried out on a range of data and data types. Example data include the asset register, asset maintenance and failure history data, other asset data, for example cost data, and external data. Since data type is the most important factor that determines which EDA techniques can be used, much of the EDA is common to all the data in the various data sources.

Time series analysis EDA techniques are particularly useful for predictive asset management because they can be used to study the dynamic nature of asset maintenance and failure. They can answer questions such as:

- do failures occur more at some times of the year than at other times, or more on some days of the week than on other days
- is the scheduled maintenance frequency adequate, particularly as the assets age
- do some events tend to occur together separated by a lag, for example extreme weather and asset failure due to severe blockages. If such relationships do exist, they must be reflected in the asset survival models in **PAM**.

A particular problem with maintenance and failure data is that some of the interventions may be nested or overlapping (see *Time to Failure Transformations Module* in PAM Modules). EDA carried out properly will detect such interventions. The solution to these very serious data problems is not straightforward and requires complex bespoke software that is an integral part of **PAM** that is not available elsewhere.

