The Predictive Analytics Cycle and PAM

Introduction

More and more organisations are realising the hidden value of the data they hold on all aspects of their business, including the performance of their assets. This requires them to acknowledge two things: firstly, that data are an asset to be exploited for improving the performance of the organisation; and secondly that predictive analytics is the means by which this objective can be achieved. Developing predictive analytics models requires knowledge and experience, and the entire process, including data preparation, cannot be automated. When the models have been developed with all the necessary data preparation procedures, they can be run automatically.

The analytics must always be applied in a structured and rigorous way if they are to improve the quality and validity of business decisions and so be of value to the business. The benefits of applying predictive analytics are more achievable now than ever before because of the increasing amount of data, ready availability of software and powerful hardware. However, analytics alone is not the answer to how to gain business advantage from the ever-increasing amount of data and low cost of computing power. Rather, the answer consists of a number of steps:

- understanding the business problem
- working out the data required
- identifying what data are available
- designing and building a credible and practical solution to the business problem that uses the available data and appropriate analytics; and
- implementing the solution.

These steps are a summary of CRISP-DM (see *The CRISP-DM Methodology* in Analytics Modelling).

The ready availability of easy to use software, including black-box software, does not detract at all from the importance of understanding the data, preparing them correctly and using an appropriate model that is understood and whose assumptions and limitations are known. Predictive analytics that uses incorrect data, incorrectly prepared data or unsuitable models will not address the business problem and so may lead to inappropriate and costly decisions and actions.

Predictive analytics covers a broad range of methods and includes decision analysis, forecasting (causal and time series), queueing theory, network theory, reliability analysis, simulation and optimisation. The particular methods used depend on the aim of the project, the application and the data. **PAM** uses

survival analysis and discrete event simulation to model and simulate asset failure (see the appendix in *Introduction to PAM* in PAM Introduction for a brief introduction to survival analysis).

The Predictive Analytics Cycle

There are four stages in the predictive analytics cycle: descriptive, diagnostic, predictive and prescriptive. They are expressed formally in the CRISP-DM methodology (see reference above). Each stage uses the outputs of the preceding stages. Table 1 shows the question answered by each stage.

Table 1

Analytics Type	Question Answered
Descriptive	What happened?
Diagnostic	Why did it happen?
Predictive	What is likely to happen under a range of scenarios?
Prescriptive	What should be done to address the problem?

Descriptive Analytics

Descriptive analytics is the exploratory analysis and presentation of data in summary form using tables and graphs so that their key features can be viewed and described. Since it only involves empirical analysis of the data, modelling the data is not part of descriptive analytics.

It is quite common for descriptive analytics to reveal unusual or unexpected features of the data or a small number of observations that are inconsistent or significantly different from other observations. Very often, the results of the descriptive analytics suggest the most suitable type of modelling for the diagnostic analytics.

The most common summary statistics are measures of the central tendency of the data, i.e. one figure that can be used to represent all the data (the 'centre of gravity' of the data), and the spread or variation of the data. Correlation and crosstab analysis are also part of descriptive analytics. Plotting the data using a histogram or a bar chart, as appropriate, is a quick and easy way of gaining a very good initial view of the data, including their skewness, and any unusual or unexpected observations. The larger the data set, the more important it is to plot them to get a view of their main features. If the data are a time

series, i.e. observations made at regular intervals in time (hourly, daily, weekly, annual, etc.), they should be plotted as a time series (the variable is on the Y axis and time is on the X axis) to show their dynamic variation. The plot will reveal any regular periodicity in the data, for example weekly cycles.

It is sometimes beneficial or necessary to transform some of the data to new forms, for example to changes, percentage changes between adjacent values or normalised values, i.e. values relative to a particular value in the data or to an arbitrary value (usually 100). These transformations are particularly useful when comparing variables measured in different units.

Diagnostic Analytics

Diagnostic analytics involves developing models that show the main relationships and dependencies in the data. Its aim is to determine what happened and why it happened by converting observed data into insight and understanding. When developing models, it is important to recognise that however good a model is, it is never a perfect representation of reality – models are our view of a very complex world using concepts we developed to help us explain and understand our observations.

There are many types of model and the most common type is regression, of which there are many variants, for example linear, non-linear and logit. Indeed, the model at the centre of **PAM** is a regression model (the Cox proportional hazards model). Other types of model include decision tree models, classification models and Markov models. Each type of model addresses a particular problem, and the available data and aim of the project define the required model.

Predictive Analytics

Predictive analytics uses the models developed in the diagnostic analytics stage to predict what is likely to happen under a range of different scenarios. The results are the *most likely* outcomes and cannot be assumed to be *the* outcomes because as stated above models are our (imperfect) perceptions of reality. The models may reveal unexpected results and problems that the organisation was not aware of and require further investigation, possibly immediately. If changes to the system are made after developing the model, it should be updated so that it is based on the revised system and then rerun.

Prescriptive Analytics

Prescriptive analytics uses the results of the predictive analytics to suggest the actions required to correct or overcome any problems. Although the results of the predictive analytics are very important

factors to consider when deciding the best course of action for addressing the problem, factors that were not considered in the model, for example subject matter expertise and domain knowledge, should also be considered when deciding what to do. It should always be remembered that models are decision *support* tools, not decision *making* tools.

PAM and the Analytics Cycle

Table 2 lists the analytics, the type of modelling and **PAM** module for each type of analytics.

Table 2

Analytics Type	Type of Modelling	PAM Module
Descriptive	Business intelligence	Asset Key Performance Indicators
Diagnostic	Model development	Asset Deterioration Curves, Asset Survival Models
Predictive	Forecasting, projection	Predicted Maintenance Interventions
Prescriptive	Simulation, optimisation	Asset Survival Simulations

By using predictive analytics, **PAM** can answer a range of asset management questions, including:

- which factors contribute to asset failure
- how to optimise asset management at individual asset level and at the operational, tactical and strategic levels
- how does the risk of an asset failing change as it is used and maintained
- how does proactive maintenance affect the risk of subsequent asset failure
- how does repeated asset failure affect the risk of subsequent asset failure
- how does asset criticality determine which assets to maintain
- what is the effect of standby assets on asset group reliability.

The data required for applying predictive analytics to asset management depend on the industry, organisation and asset type, and can be split into four categories: asset register data; asset maintenance and failure history data; other asset data; and external data. Table 3 shows example data for a range of asset types. It is unlikely that all the data in the table will always be available or required. Similarly, other data may be available or required.

Table 3

Data Source/Type	Examples
Asset register (asset static data)	Installation date, manufacturer, design specification *
Asset maintenance and failure history	Maintenance and failure data with dates *
Asset operating history	Time in use, number of starts *
Asset status	Asset criticalities and redundancies **
Asset costs	Asset maintenance and replacement costs **
Other costs	Costs resulting from asset failure **
Maintenance depot	Location, size of workforce *
Catchment	Type (rural/residential/urban/industrial), population, holiday location, inland/coast *
Season	Weather (rainfall, temperature), month *
Extreme events	Flooding, extreme temperatures *

* candidate predictor variables

** factors applied to models

PAM and Asset Management Optimisation

Predictive asset survival models are developed to understand the causes of asset failure. If the risk of asset failure is modelled as a dynamic phenomenon (as **PAM** does), asset performance can be optimised at individual asset level and at the operational, tactical and strategic levels. One way of modelling the risk of failure as a dynamic phenomenon is to use each asset's maintenance and failure history with dates (in addition to its static data) in the models. Static data alone cannot model the risk of failure as a dynamic phenomenon (because the data are time invariant) and neither can they identify each asset as a unique entity if the number of assets is greater than the number of asset descriptor combinations. For example, if the assets are described by three factors and the factors have 2, 3 and 4 values respectively, there are 24 value combinations. If the number of assets is smaller than 24 and the assets have different value combinations, the assets can be defined uniquely but if at least two assets have the same value combination, the assets cannot be defined uniquely.

The multi-dimensional and dynamic nature of asset failure means that it is easier to optimise individual asset performance heuristically than by an optimisation algorithm. The optimisation is carried out by

using the asset survival model to simulate the effects of a range of factors on an objective function that describes a key asset performance metric, for example asset maintenance and replacement costs, and the consequence costs of asset failure, at a predefined number of time steps. If the optimisation and therefore the simulation are for the next five years and the system is sampled at monthly intervals, the simulation is carried out for 60 time steps. The number of calculations, and the amount of intermediate data and output that have to be stored can quickly become very large. For example, in a simulation of 2,000 assets and a five year model at monthly sampling 120,000 values for *each* calculated variable are produced.

Furthermore, since the failure rate of each asset is defined by a number of parameters and depends on the failure rate at earlier times (the failure rate of an asset is strongly autocorrelated *in the absence of maintenance interventions*), it is clear that the optimisation is numerically very intensive. Continuing the above example, if 15 variables are calculated at each time step, 1.8 million values are calculated in each simulation run. A modern laptop can easily handle this scale of problem but it is instructive to note the very large number of calculations, let alone the complexity of the asset survival and simulation models.

The value of the objective function in each simulation run is calculated from its values in all the simulation periods in the run. The optimisation requires a number of simulation runs, one for each combination of factor values. The optimal asset management policy is the policy with the factor value combination that optimises the objective function.

Table 4 shows example variables and the role of each in the cost optimisation (minimisation) of assets in the water industry.

Table 4

Variable	Role in Optimisation
Total cost	Objective
No. proactive interventions per month	Factor
No. reactive interventions per month	Factor
Threshold survival probability	Factor
Maintenance capacity	Constraint
Risk tolerance	Constraint

- The objective function is the sum of the costs to be minimised.
- The factors are the variables in the simulation model that can be controlled. Each simulation run uses a different combination of factor values.
- The threshold survival probability is the survival probability below which an asset is deemed to be at very high risk of failure.
- The maintenance capacity is the maximum number of assets that can be maintained each month.
- Risk tolerance is the maximum acceptable level of repeated asset failure.

The output of the optimisation model is the set of factor values that optimises the objective function subject to the constraints.

