

# Statistical Methods in CRM

## 1 Introduction

This is a brief description of the variety of statistical methods potentially applicable to CRM problems. I've tried to indicate what each method entails, give examples of where it might be applied and explain our 'implementation status' with the method.

## 2 Graphical Presentation

### 2.1 Description and Scope

This covers a wide range of tools, including, for example:

- Value-Potential matrix
- Scatter plots to illustrate/explore relationships (e.g. profit vs. one or more demographic factors)
- GIS systems
- Diagnostic and explanatory displays from other tools (e.g. from predictive modelling – see Section 3)
- Multivariate presentations of data – e.g. in analysis of customer survey data.

### 2.2 Areas of Application

One or more graphical tools will be used in almost every project, both to gain insight into the data and to explain findings to the customer. For example, in attempting to build a model to predict buying propensity, one might use the following sequence of graphical tools:

- Plots of observed propensity vs. demographic variables
- Tree diagram to illustrate and explore preliminary modelling results
- ROC (Gini) curve and partial residual plots to explore the results of fitting predictive models and to refine the models
- Display of predicted propensities against demographic variables
- Analysis of customer survey data

## 3 Predictive Modelling

### 3.1 Description and Scope

Sometimes known as regression modelling or (in the data mining literature) as supervised knowledge discovery. The intention is to predict a future outcome or measurement (e.g. 'the card holder will default') from currently observable characteristics (e.g. age, income, employment tenure...). The outcome (response variable, dependent variable, predicted variable and a load of other names) can be binary (Yes/No) categorical (e.g. 3 categories – good responder, fair and poor) or continuous (e.g. profit). In the first two cases the problem is termed 'classification' and in the last case 'regression'.

All these methods rely on having a group of customers (the 'training sample') for whom both the response and explanatory variables are known. The results of fitting the model

to this group are then extended to the remaining customers to predict the response variable.

A large number of methods fall into this category:

- Tree based methods (e.g. CART and CHAID). These partition the customers into discrete groups ('leaves') on the basis of the explanatory variables, so that customers within a leaf are 'similar' in their values of the response. These methods are computationally expensive, robust and easily explained, but generally of low accuracy. Modifications exist to classification trees to increase their accuracy substantially (e.g. boosting) but these also increase the complexity.
- Generalised Linear Models. Also known as regression modelling. The additive scorecards produced in credit scoring are a special case of regression modelling. These models are generally fairly accurate, reasonably robust and easily interpretable, but require considerable analytical skill to develop.
- Neural Network and other non-linear models (e.g. MARS). These are flexible generalisations of the linear models. Neural networks especially are 'sexy', and dramatic claims are made for their improved accuracy and ease of use. In fact, however, similar analytical skills are needed, the methods can go badly wrong and the results can be very hard to interpret.

### **3.2 Areas of Application**

Just a selection of the possible areas:

- Credit scoring
- Propensity scoring
- Fraud detection
- Prediction of lifetime value

### **3.3 Implementation Status**

## **4 Segmentation Methods**

### **4.1 Description and Scope**

Also known as clustering (in statistical literature) and unsupervised knowledge discovery (in data mining). Methods designed to group customers based on similarities in their demographics or (more usually) behaviour, in the hope that such groupings will be helpful in predicting future reactions or in selecting targets for campaigns.

A large number of clustering tools exist. They enjoy generally similar properties, and all suffer from the major defect that they generally *will* find clusters, even when no meaningful ones exist.

Related techniques such as factor analysis and multi-dimensional scaling also exist to allow us to represent a large number of related measurements in a small number of dimensions (for example, one might evaluate staff on a number of measurement scales, then attempt to find a way to represent the results in a 2 or 3 dimensional plot).

## **4.2 Areas of Application**

Again, a selection:

- (Clustering) Finding groups of customers who are 'similar' for the purposes of marketing (e.g., private client bank customers, developing target groups for campaign mailing)
- (Dimension reduction) Graphical comparison of companies using data from a multiple scale audit.

## **5 Time Series Methods**

### **5.1 Description and Scope**

These methods are appropriate when a measurement has been collected repeatedly over time (e.g. monthly sales figures) and we wish to predict likely future values of the measurement. More complex methods allow one to find the relationship between multiple time series (e.g. monthly receipts and UK economic indicators), to predict the values of one series based on projected future values for the other series.

### **5.2 Areas of Application**

Largely as described above. Currently time series methods, especially those based on the relationships between time series, are relatively little used in customer modelling because of the specialised nature of the problems to which they are appropriate.

## **6 Miscellany**

### **6.1 Description and Scope**

- Experimental design techniques are important in situations where we wish to test the results of one or more changes in customer management. For example, we may wish to discover the relative efficacy of a number of mailings in promoting sales of a new credit product. Efficient experimental designs and associated analysis methods can ensure that useful learning occurs
- Event driven marketing, i.e. tying a marketing approach to an important event in the customer's life (e.g. moving house), is frequently touted as a potentially very valuable marketing tool. Statistical techniques may help identify appropriate events and, via experimental design, quantify the value of such an approach.