

Data mining and statistical models in marketing campaigns of BT Retail

*Francesco Vivarelli and Martyn Johnson
Database Exploitation, Segmentation and Targeting group
BT Retail
Pp501 Holborn centre
120 Holborn
London EC1N 2TE*

In this paper we present some applications we develop to support marketing campaigns of BT Retail Consumer Division by using data mining techniques and statistical modelling to segment and build propensity models for our 19.5M base of customers.

The base of customer has been segmented by K-means clustering algorithm, where location and width of the K Gaussian were optimized with the Expectation-Maximization algorithm. The 19.5M customers have been clustered on the basis of transactional summaries, demographic and lifestyle variables; balance of these features guarantees that segments are logical across each type of variable.

We also build propensity models to optimise the selection of suitable customers who will be more likely to positively respond to marketing campaigns. We show how decision trees, logistic regression and neural networks score our base of customers by using traffic and billing data as well as demographic and lifestyle features.

All our applications have been developed using the SAS System (release 8.2) for Microsoft Windows 98 (2nd edition) and release 4.1 of the Enterprise Miner Software.

1. Introduction

The success of a marketing campaign can be determined by the knowledge we have about the lifestyle and behaviour of our customers. In particular, market segmentation and targeting the right customer with the right product play key roles in order to build up a complete picture of BT's customers. In this paper we present data mining and statistical techniques we use to segment our customer base and build propensity models to support our marketing campaigns.

Like most major retailers, BT has segmented its consumer market for many years. Over time, segmentation structures have evolved from simple revenue based schemes to classification based on demographic factors such as life-stage, presence of children etc.

These schemes were successful, in that they enabled us to address our segmented markets more effectively but, as in all our marketing activity, we try to develop even better and more effective methodologies. With our Customer segmentation, in particular, we were keen to develop a scheme which allowed us to obtain a holistic picture of our customers, based on how and when they use our services and on many demographic attributes. With this improved understanding of our customers - you might call it "what makes them tick" - we would be able to develop products and services, approaches and campaigns that would be truly tailored to their needs and lifestyles. We have termed this our "data logical" approach because we have allowed the data, using SAS programs, to create the segments, rather than using any element of preconception to achieve this.

BT is one of the world's leading communications companies, and we have a large share of the UK Consumer market. This is positive, but it does give us the problem that, to understand our customers properly, we need to create and maintain very extensive knowledge systems, capable of dealing with vast amounts of data generated by the activities of many millions of people. A proper, robust segmentation exercise presents quite a challenge in this context!

It is also important to develop methodologies which help marketing campaigns to target the right people with the right product - i.e. answering the question: How can we discover who will positively respond to a contact strategy? In this way customers can be contacted only with the message which is relevant to them, improving significantly customer satisfaction.

We model the behaviour of each customer as a binary variable, i.e. the customer either responds or does not respond to a marketing campaign, the customer either buys or does not buy a certain product;

thus the target associated to each customer's profile can be either 0 or 1. A list of *hot contacts* for our telemarketing advisors is generated using some form of predictive models, such as decision trees, logistic regression or feed-forward neural networks. For each customer the statistical model generates a single number which represents the propensity of that customer to do something, i.e. the probability of behaving in a certain way.

The paper is organised as follows. In Section 2 we describe the data we use in our activity and the features which describe the customer base. Section 3 presents the procedures we follow in order to pre-process data for a data mining project. The methods used to segment the customers into subgroups and to target them for our marketing campaigns are reported in Sections 4 and 5, respectively. Conclusion of the paper and future work are reported in Section 6.

In this paper we will not go into any technical detail of the techniques used. For a technical introduction, we suggest the references [1], [2], [3] and [4] listed at the end of the paper.

2. The database

As you might expect, BT holds a considerable amount of data about how our customers use the telephone, and we can aggregate this in many ways to build up the holistic picture of customer behaviour referred to previously. Not all the information can be employed in the data mining process though, since the use of data that BT processes to manage the flow of traffic across our network and to bill customers for the telecommunications services is subject to rigorous and complex regulation. Data collected about each customer can be divided in two subsets, the traffic and billing data (TB) and data describing demographic and lifestyle features (DL).

TB data are generated from information obtained from traffic and billing data; however for practical, regulatory and competitive reasons, it is very rare that we employ TB data in data mining models for marketing campaigns.

DL data describes demographic and lifestyle features of customers. They are partly provided by a third party supplier and have been obtained through "shoppers survey" questionnaires and product registration response forms and services. Attributes available include demographic attributes (e.g. primary and partner age band, marital status, number and age band of children, occupation type), financial information (e.g. household income, credit cards, stocks and shares) and lifestyle information (e.g. hobbies and interests, newspapers read, car ownership, home ownership status).

These are just a few examples - there are literally hundreds of fields of data available. You should note that, while actual responses are used to complete the fields for a great many observations, the remainder are modelled.

TB and DL data are collected in one input vector which describes features of BT's customers. Thus we have as a potential basis for our data mining set a large number of variables, comprising aggregations from billing data records and the hundreds of lifestyle variables. Add to this the fact that BT has a very large number of customers and you will see that we end up with a very large data mining set which, even after some careful pruning, will still need a very powerful tool to analyse effectively. We feel that SAS Enterprise Miner, operated on a client/server basis provides us with the necessary power to deliver appropriate analyses against such large datasets. Incidentally, we maintain a version of this huge dataset in our SAS environment, and we use it as a basis for many of our data mining and analytical activities - we call it the "SAS Mother" because it took the mother of all queries to set it up!

3. Data pre-processing

Data pre-processing plays a crucial role in a data mining project, since the final results we will obtain depend on the quality of data used in our models. The procedure is illustrated in Figure 1, which shows the SAS Enterprise Miner desktop, upon which a simple data mining project has been set up.

Data are loaded into a project via the INPUT DATA SOURCE node. In order to reduce the number of data to a manageable size, the data loaded are sampled by using the SAMPLING node. This node allows to choose the sampling methods, the sample size and the random seed. The five sampling methods offered are simple random sampling (default), sampling every n th observation, stratified sampling, sampling the first n observations and the cluster sampling. Ideally we would like to sample the database randomly. However it is often the case that the actual proportion of the target event level is tiny (sometimes less than 5% of the total number of observations in the predecessor input data source). Since the number of observations for the targets is very small, we usually decide to stratify the sample obtaining two subsets of equal sizes from both classes 0 and 1.

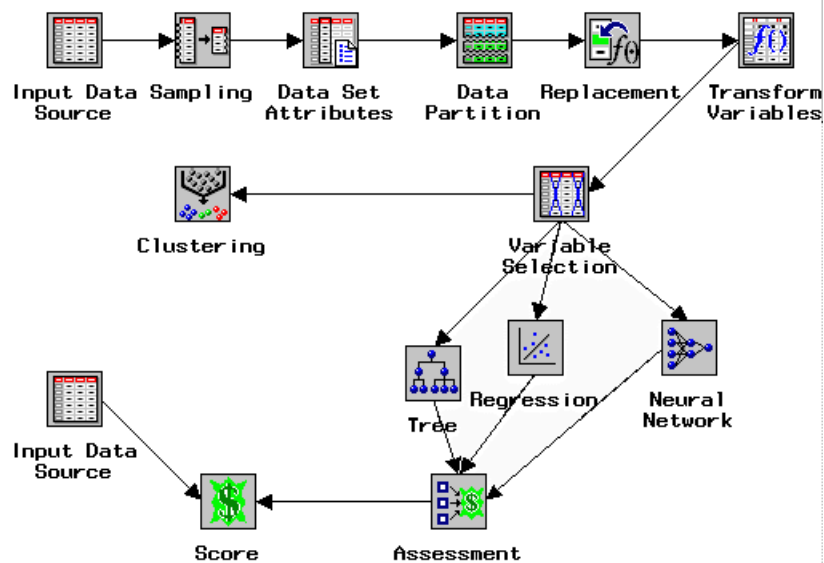


Figure 1 The SAS Enterprise Miner desktop upon which a simple data mining project is set up. The use of each single node has been explained in the text.

The advantage of the stratified sample is that it provides us a better chance of finding useful patterns for the rare target event. Unfortunately the sample is biased with respect to the original proportions of the target levels in the input data source; in order to develop valid, meaningful models which can be applied on real world data, we have to take into account the effect of the biased sample later on in the analysis. This is achieved by editing the *prior vector* of the target profile in a DATA SET ATTRIBUTES node. This option adjusts the probability values for each target level back to those in the original data. By default, the prior probability values are proportional to those in the data; however we can specify our own prior probability values, by typing the values of the true prior probabilities for each occurrence of the targets. These values (which must be between 0 and 1 and add up to 1) should reflect our prior knowledge of the problem we are dealing with.

This node can also alter the attributes of input data - for instance many of the lifestyle variables have a 1 to N coded value which might be interpreted by SAS as an interval value - clearly we would need to change this to an ordinal value for the downstream nodes to work properly.

In order to train the models and to assess their generalisation capabilities, the data available are randomly split in three subsets by using the DATA PARTITION node; the training (containing 40% of the total), validation and test sets (containing 30% of the data each). Each set is used for a different purpose during the data mining project; the training set is used to estimate the parameters of the model, the validation set to select the best structure for the project (e.g. the number of hidden nodes in a neural network) and the test set is used to estimate the generalisation capabilities of the model built. Unless otherwise specified, in the following we always report results obtained on the test set.

Some models (such as the neural networks) omit entire observations from training if any of the input variables are missing. Hence we need to replace missing data with imputed values. The DATA REPLACEMENT node enables us to replace interval missing values with the input's mean, median, or midrange. Missing values of a categorical input can be replaced with the mode.

The input vector describing each customer is composed by features whose values may differ for several order of magnitude; thus we need to transform each component, obtaining linearly scaled values. We transform variables by using the TRANSFORM VARIABLES as follows:

- an interval variable is linearly transformed so that it has mean 0 and variance 1;
- a binary variable is replaced by a variable which contains values 0 or 1;
- nominal variable with n categories is expanded in n dummy variables set to 0; only the one corresponding to the level we want to code is set to 1.

We can also use this node to quickly add new variables - e.g. we might want to call charges higher than £100 per quarter as "high spend".

We can investigate and select the variables that will go forward into the final modelling node by using the VARIABLE SELECTION node. To do this in a scientific way, the node allows us to test to correlation values between the variables and exclude those which have low values and which would not contribute to the decisions made by the final nodes. We note that usually a small amount of input variables proves to be useful in predicting the target variable.

In the remaining sections of the paper we illustrate the two most relevant data mining applications we have developed, namely Market Segmentation and Database Targeting Marketing. These lie at the heart of our marketing activities and they are therefore commercially sensitive. We will try to be as specific as possible in describing how we have used SAS to drive these activities, but we will not go into detail about the segments themselves. The charts and diagrams which will be presented are intended to describe and illustrate the methodologies, but to preserve commercial confidentiality we have based them on dummy data.

4. Market Segmentation

Market Segmentation is carried out by using the CLUSTERING node. In this case we have used the K-Means clustering algorithm which exists as one of several clustering or classification algorithms in SAS. Viewed in two dimensional terms, this partitions the observations that have the closest values, as shown in Figure 2.

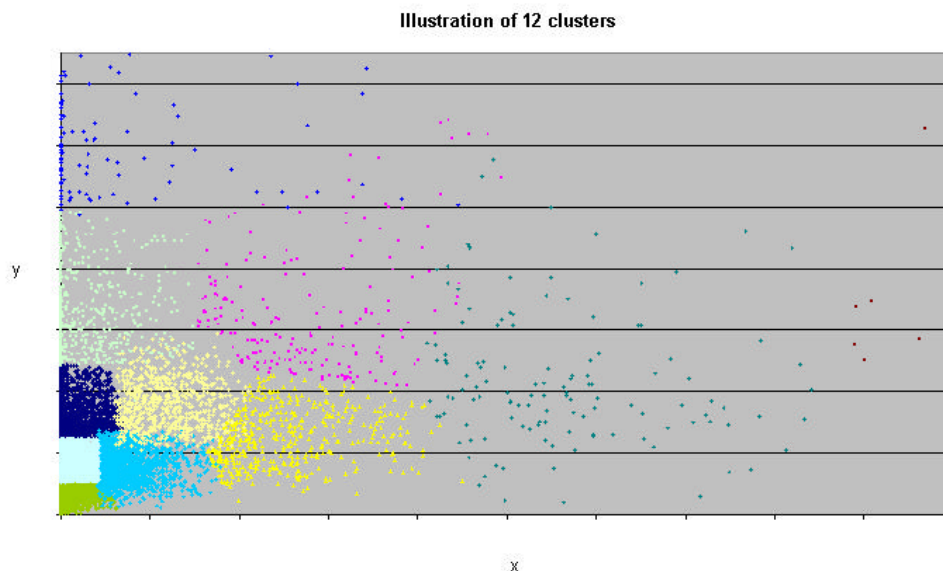


Figure 2 Clusters illustrated in two dimensions. Clusters defined by the model have been denoted by different colours. Note how some clusters are tightly packed while others are sparsely distributed (illustration based on dummy data).

....but of course this is really done in a multidimensional space, not just two dimensions. The diagram illustrates how some of the resulting clusters are very closely packed - in other words the members of the cluster strongly share the attributes, and others are widely or sparsely distributed - meaning that the members have a lesser level of similarity to their fellows. For example, the key attribute y might be spend per quarter; hence the closely packed observations could represent those customers with a bill size close to modal, whereas the sparse observations could represent those with a high bill. Note that there's an argument for considering the sparse observations to be closer to one another than the observations with lower spend are to the modal value.

The CLUSTERING node enables us to choose from a number of different classification algorithms - the default is a least squares type of model - and there are various options available by which we can refine the model to optimise the end result.

This node, once run, allows us to examine the statistics of the resulting clusters and we can use these to evaluate the model, make refinements along the PROCESS FOR DIAGRAM as necessary, and reiterate. We can view the selection of clusters in a decision tree format, which is excellent for sharing the information with our Marketing colleagues. This example has been created using dummy data, but

you can see that the most important variables for deriving and describing the clusters can be readily seen.

I can tell you that, for the final version of our segmentation scheme, we identified over 20 clusters (see Figure 3), some of which were small and were subsequently aggregated.

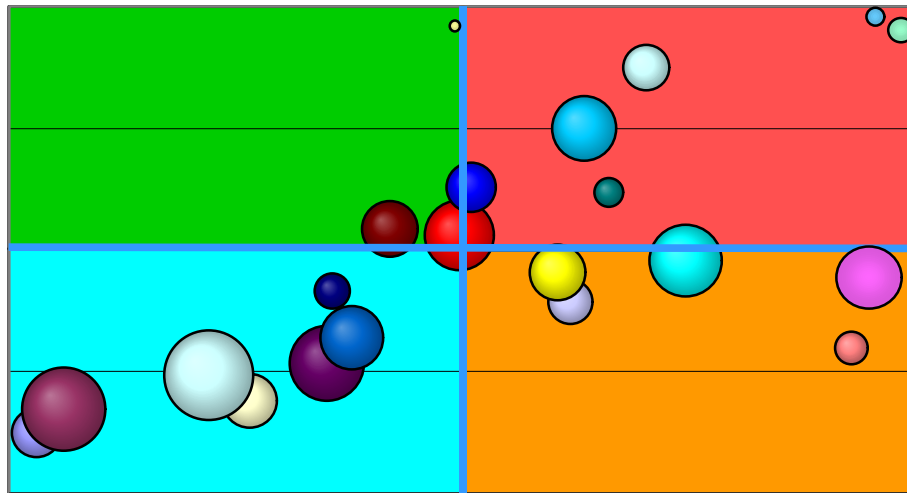


Figure 3 Illustration of 2D view of clusters.

In introducing the segmentation scheme to our Marketing colleagues, we found it useful to illustrate the segments in a two dimensional grid with the axes being the key dimensions as above. Here is an illustration of how the segments look on this grid, with the size of the bubbles representing the frequency of observations per cluster. These axes turned out to be somewhat interdependent since the x dimension will generally contribute to a higher y value, hence when viewing the clusters against these axes we see a clear bottom left to top right trend.

5. Database Targeting marketing

A Data mining project for database targeting marketing should provide a model which is able to estimate customer's propensity to behave in a certain way (propensity model).

We model the behaviour of each customer with a binary variable and the target associated to each customer's profile can be 0 or 1, i.e. the customer either responds or does not respond to a marketing campaign, the customer either buys or does not buy a certain product. For each campaign we test several models and in the following we illustrate the use of some of them.

Decision tree

DT represents a segmentation of the data that is created by applying a series of simple rules. Each rule assigns observations to a segment based on the value of one input. One rule is applied after another and results in splitting each segment in sub-segments. The hierarchy is called a tree, and each segment is called a node. The criterion for evaluating a splitting rule may be based on either a statistical significance test (an F test or a χ^2 test) or on the reduction in variance, entropy, or Gini impurity measure. An advantage with respect other models is that a DT produces a set of interpretable rules (see Figure 4).

Unfortunately, sometimes the simplicity of the rules can not fully explain the complexity of the data at hand and more powerful models should be applied. Lack of granularity is a particular problem for us in using DT - a tree with even as many as 40 leaf nodes would mean that we have large groups - hundreds of thousands - of customers all receiving the same score. However, as may be seen from the tree diagram itself, DT is an excellent way of describing to our Marketing colleagues the key variables that drive the decisions.

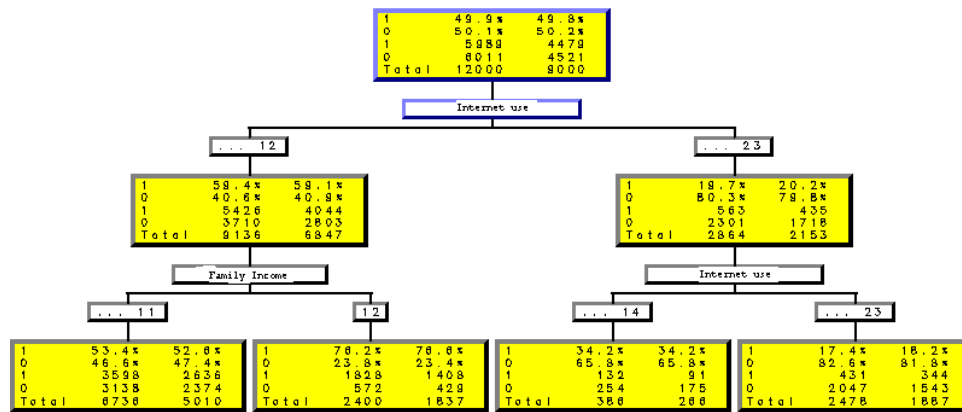


Figure 4 Graphical representation of the first few nodes of a decision tree. The initial database has been segmented in two subsets (on the basis of the attribute called Internet use) and then in two further sub-segments on the basis of the Family income and Internet use attributes.

Logistic regression and Feed forward neural networks

A Logistic regression (LR) is a linear model which attempts to predict the probability that a customer will behave in a certain way on the basis of one or more independent inputs. It is important to stress the fact that the model is linear, i.e. it can discover only linear relation between customer's features and customer's behaviour. It can be implemented in the SAS desktop with a REGRESSION node (see Figure 1).

Non linear mapping between customer's features and customer's response can be modelled by feed-forward neural networks (NNs), by using a layer of hidden processing units. In a NN the input feature of a customer is processed by a layer of hidden units, producing as output the probability that a customer behaves in a certain way. We note that a NN without hidden nodes corresponds to a LR and it is usually known as generalised linear model (GLM).

A good performance of NN can be achieved by setting the right number of units in the hidden layer (a high number of hidden units increases the risk of overfitting the data, building a model which is not able to generalise on unseen data). In order to avoid overfitting, we choose the optimal number of hidden units on the basis of the error reported on the validation set (see Figure 5).

The structure (i.e. the number of hidden units) of a neural network depends on the problem at hand and thus it is not possible to suggest a standard setting of the a NEURAL NETWORK node (Figure 1) for a data mining project. However there are some standard options we tend to choose.

Variables in the input layer should be normalised (as we suggested also in Section 3) since this can avoid overfitting of the training data.

The activation function of the units in the hidden layer is the hyperbolic tangent. The activation function of the output unit suitable for a binary classification is the sigmoid function; this enables us to interpret the output value of the NN as the probability that a customer behaves in a certain way.

An optimisation algorithm which performs well on the kind of problems we deal with is the Conjugate gradient; this algorithm is suitable for large problems (memory requirements are only linear in the number of parameters) and it is also fast in converging to a local minimum (since it makes use of information about the curvature of the objective function).

In our applications, the CG optimises the logarithm of the likelihood of the data, which is the Bernoulli error function in case of binary classification targets.

Results

The ASSESSMENT node evaluates and compares the performance of classification models; among the several methods available, the one which better suits our needs is the lift chart. In the lift charts (for binary target) the test set is sorted level in descending order according to the posterior probabilities of the event and the observations are grouped into percentiles (reported on the x-axis). The y-axis reports either the percent response or the cumulative percent captured response obtained within each percentiles. An example is shown in Figure 6.

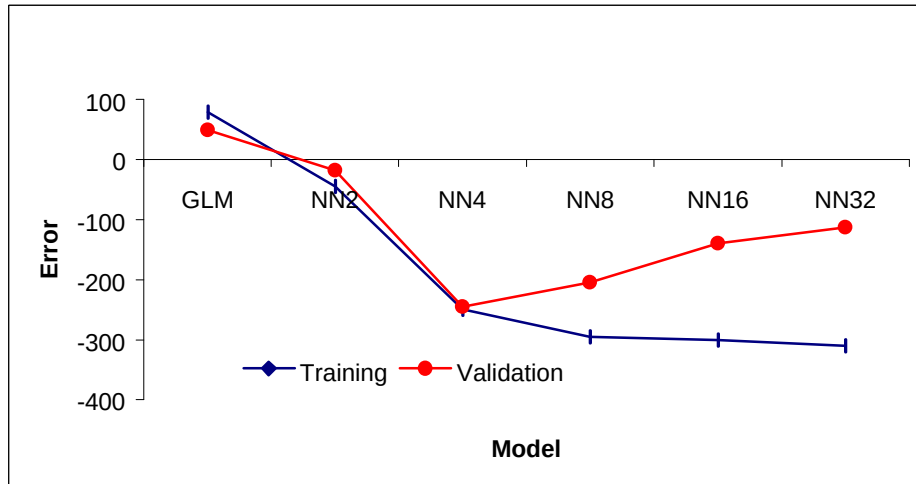


Figure 5 Training and validation errors (reported on the y-axis) as functions of model complexity (on the x-axis). NNx indicates a neural network with x hidden units. The graph suggests that the optimal number of hidden units for the problem at hand is 4.

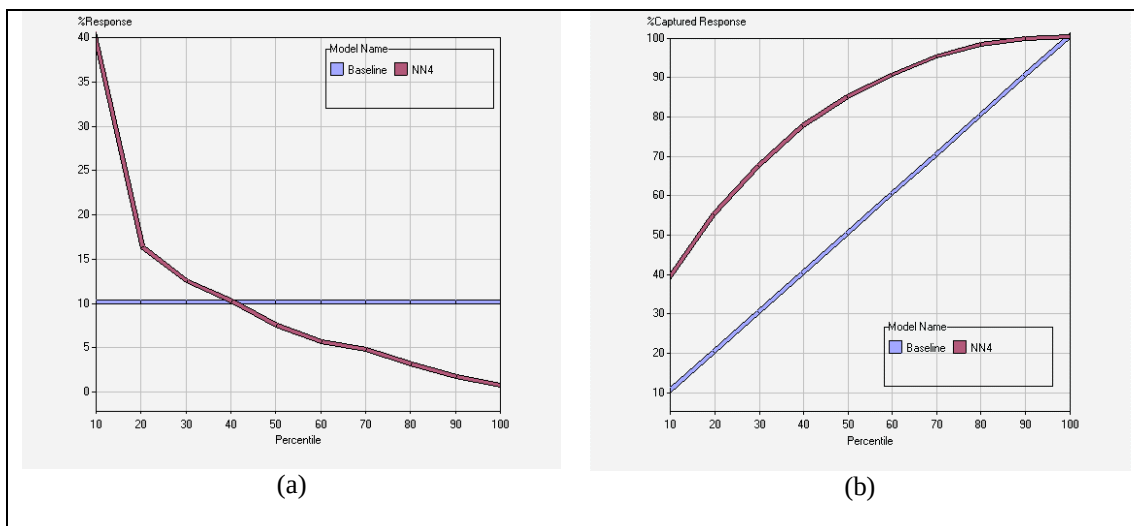


Figure 6 In the above lift charts, the test set is sorted according to the posterior probabilities of the event level in descending order and the observations are grouped into percentiles (shown on the x-axis). The y-axis reports the percentage of customers who are correctly targeted by the model (a) and the cumulative percentage of the captured response (b). The baseline represents the performance of the random classifier.

Figure 6 (a) shows the percent response we obtain from a statistical model we prepared to support a marketing campaign. We can notice how the percent response decrease significantly as the model targets lower percentiles (i.e. customers less likely to positively respond) and after the percentile 40 the model performs worse than the random classifier. This means that, in order to optimise costs, the telemarketing department has to contact only people scored within the top 40 percentiles. Note also that the percentage of response in the top 10 percentiles is 40%, i.e. four times better than the random classifier.

Figure 6 (b) shows the cumulative percent captured response for the same campaign. The graph shows that, contacting customers of the first 40% of our customer base, the model is able to identify about 80% of the total number of customers who will positively respond to the campaign; this is twice better than the performance of a random classifier.

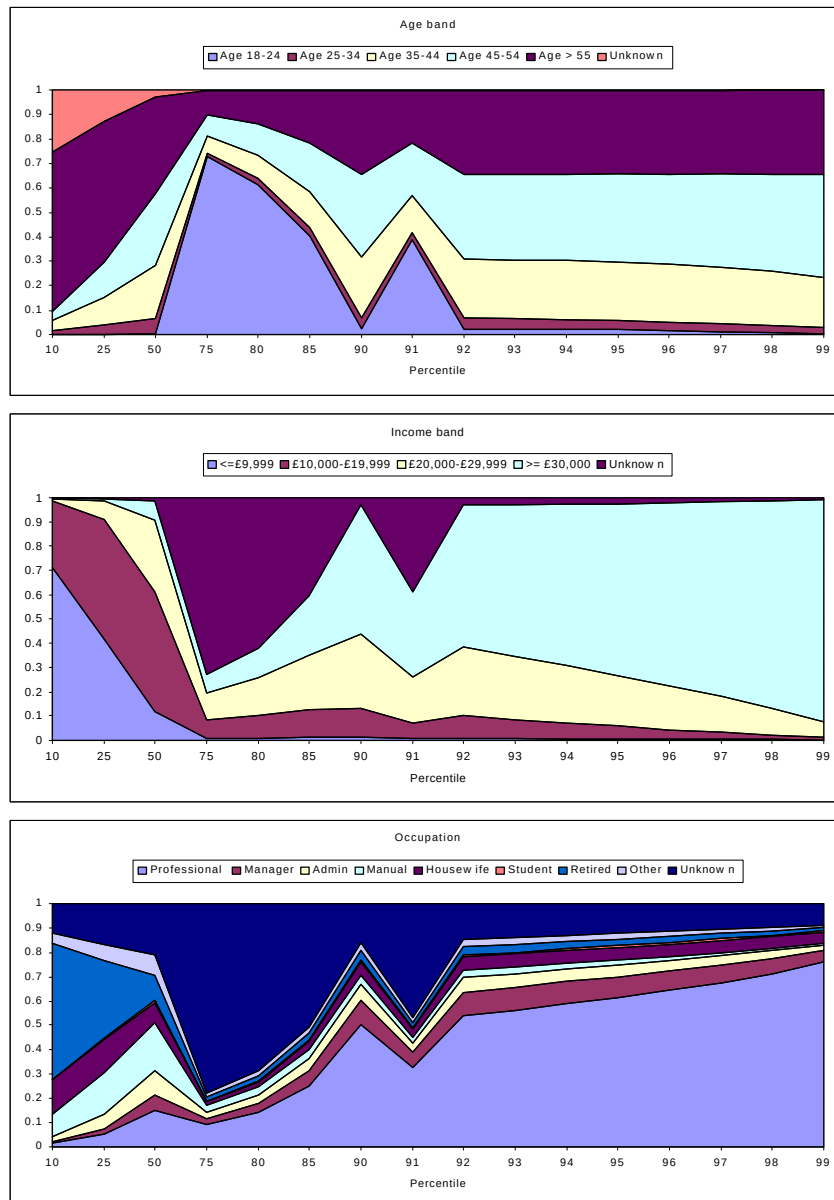


Figure 7 The charts show a graphical representation of the distribution of Age, Occupation and Income bands (on the y-axes) reported in each percentile. Note that customers in top percentiles high percentiles share an homogenous demographic profile. Homogeneity is lost in lower percentiles, when predictions of the model become less accurate.

Another way to presents the results of a propensity model is to describe the customer belonging to each percentile. This can be done by looking at the distribution of their demographic characteristics and lifestyle as a function of the percentile. An example is reported in Figure 7, where for each percentile (on the x-axis), we report on the y-axis the distribution of age, income and occupation. Note that top percentiles are characterised by customers with highly regular profiles; on the contrary the less accurate prediction of the model lose regularity in customers' profiles.

The list of *hot contacts* can be produced by scoring the whole customer base with the SCORE node.

6. Conclusion and future work

In this paper we presented data mining techniques and statistical modelling which we use to segment and target the 19.5 million customers base for marketing campaigns of BT Retail.

Gaussian mixture models achieve a satisfactory segmentation of the customers, whereas decision trees, linear and non-linear regression are implemented for our database targeting marketing.

So far we have used TB and DL data to understand the nature of the individuals within the segments. Whilst valuable, this is only part of the story - to develop a more thorough understanding of attitudes and behaviour we need to examine, in detail, attributes on such diverse subjects as media consumption, transport, money, TV & radio and attitudinal statements from which we can gain some insight into their personalities - as we said in the introduction, learning 'what makes them tick'. Similarly, it will be our aim to conduct primary research on the segments, developing segmentation and targeting models localised in each single segment.

References

- [1] Bishop, C.M. (1995), **Neural Networks for Pattern Recognition**, Oxford, Oxford University Press.
- [2] Duda, R.O. and Hart, P.E. (2000), **Pattern Classification**, New York, John Wiley & Sons.
- [3] Berry, M.J.A. and Linoff, G.S. (1997), **Data Mining Techniques for Marketing, Sales and Customer Support**, New York, John Wiley & Sons.
- [4] SAS Institute Inc.(2000), **Getting Started with Enterprise Miner Software, Release 4.1**, Cary, NC, SAS Institute Inc.