Building predictive models for direct mail: A framework for choosing training and test data

Received: 15th December, 1999

Tom Breur
is senior consultant at the database marketing centre of Postbank in the Netherlands. He is professionally engaged in matters involving data mining, customer segmentation and database marketing methodology.

Leonard Paas
worked on issues such as data mining, predictive modelling, controlling data quality, customer segmentation and credit scoring when he was a senior consultant at the database marketing centre of Postbank. Currently he is working on similar issues as a database marketing consultant at the Da Vinci Group. He is also working on a doctoral thesis at Tilburg University.

Abstract Predictive data mining models can be used to increase the return on investment of direct mail campaigns. In this paper the authors present a framework for choosing data sources for building and validating predictive data mining models. They propose a hierarchy which can be used to decide which behaviour is to be modelled when building and testing models. Choices made within this hierarchy depend on the cost and availability of relevant data and on campaign constraints.

Tom Breur
Adelaarshorst 47, 5042 XH
Tilburg, the Netherlands

e-mail:
tombreur@hotmail.com.

INTRODUCTION

When applying predictive data mining models for direct mail the goal is to select a segment of customers from the population that is considered the most desirable target group. The operational definition of desirability depends on the target measure used. Ideally, some measure close to bottom-line profitability is used. In general, a perfect measure of profitability is not feasible in practical business settings, but less sophisticated measures will be valuable as well.[1] An example of a simple measure is the probability that a subject will respond to a direct mail offer. A somewhat more advanced measure would be to include a prediction of expected spending.

Much has been written about different techniques for determining characteristics of the best target group.[2,3] Some possible techniques are algorithms for recursive partitioning, neural network computing, logit analysis or genetic algorithms. As the field of data mining has matured over the past decade, a wide choice of commercial software which uses the techniques mentioned above has become available. These data mining tools require a data set consisting of two types of variables: (1) a dependent variable that has the function of representing the behaviour that is to be predicted, the target variable (response behaviour for our purposes); and (2) an array of independent variables representing clients' characteristics, which are used to predict which values subjects have on the target variable (explanatory variables).

It is interesting that little attention has been given to the development of a methodology for establishing which data can be used as the target variable in the model-building process. An outstanding book on data preparation was published in 1999,[4] but it does not cover *selecting* a target variable. Moreover, some lucid discussions on explanatory variables can also be found.[5,6,7] As for information on the availability of target variables for predictive models, Paas and Kuijlen[8] give a short discussion on utilising ownership patterns of financial products for this purpose. Discussions on related topics can be found in David Sheppard Associates.[9] But often it is implicitly assumed that the only possible target variable is response to a random mailing. Below we will point out why this is not the case and not always the most desirable option. Our paper concentrates on the utilisation of different types of data as target variables for predictive modelling purposes and on how to choose between these options.

It will become apparent from this paper that this is a topic of relevance for direct marketing practice. That is to say, in practice, one typically has to find the best compromise when considering all the possible data sources which could supply the target variable. Response to a random pilot mailing usually resembles best the behaviour one is trying to model. A drawback can be that acquiring enough responses to build a stable model is costly. When, on the other hand, an easily available target variable is used (like coupon redemption requesting product information) one never knows how well this will forecast actual purchase of the product on offer. This is because coupon redemption, in this example, is clearly more remotely connected to actual responses to a direct mail offer. Thus, one needs to find a balance between sparse and expensive data with highly desirable qualities in terms of representing future response to a mail offer, and readily available data that are inexpensive. Sometimes a hybrid can be used as an alternative. In such hybrids models trained on a less than perfect target variable (in terms of predictive validity) might only be tested on relatively small samples of highly valid response behaviour. A desirable side effect of this hybrid approach is that, although the model building was done on a quite different target variable, through testing on highly valid response data (a random pilot mailing) one is able to present a quantitative prediction of the expected response on the direct mail campaign.

First we will discuss which data are potential candidates for use as a target variable in the model building process. And guidelines are given with regard to choosing which specific type of data should be utilised as the target variable. Then follows an elaboration on the hybrid approach introduced above. Finally, the paper concludes with a discussion in which the major points are summarised.

AVAILABLE OPTIONS

In general, three sorts of target variables can be extracted from databases containing information on interactions with clients. First, response to direct mailing campaigns. Secondly, data on which persons in the customer base have acquired a product in the past period, say one year, regardless of what marketing effort was made to sell the product. Thirdly, data on clients indicating an interest in the product — this might be calling in for information on a product, replying to a coupon for information on the product, or any kind of registered behaviour that might precede acquisition of the product. Data in any of these three classes can be used in order to build a predictive model for a direct mail campaign. It will be obvious that a response variable that closely resembles the actual targeted behaviour will have greater predictive power than, say, the behaviour of customers calling in for information, that is only remotely related to actually buying a product.

Utilising response to a random mailing as the target variable is not necessarily the best option, however, because this type of data is difficult to obtain and expensive. The expense is due to the low response rates associated with random mailings.[10] This means that a large random selection has to be made if the data miner is to obtain enough respondents to support model development. Our experience has shown that, as a rule of thumb, techniques like logit and decision trees generally require at least 100 respondents, to enable identification of significant differences in response probabilities between different segments of targeted clients. Models built on such a small number of respondents are generally not very robust however. A random mailing containing at least 250 respondents is usually a minimum requirement for the development of robust models. These numbers serve only as an indication of the numbers required, and come with the proviso that the data miner has domain expertise and abundant experience in model building with these tools. If this is not the case, more respondents are required. Moreover, there should be more respondents when utilising techniques such as neural networks and genetic algorithms. With response rates being as low as 0.5 per cent in some random mailings, this means that at least 20,000 clients have to be mailed at random to obtain the minimum of 100 respondents. Even more people have to be sent the offer if a robust model is to be developed or if neural networks or genetic algorithms are applied. It goes without saying that this requires substantial investment.

Data specifying which clients acquired specific products are often readily available in marketing databases, usually with reference to the date of purchase. Coupon redemption, telephone requests for information, or any dialogue one has with customers that might imply an interest in certain products should be recorded, as this information might prove invaluable in modelling who might be willing to purchase a product. Utilising such data as a target variable is relatively inexpensive, because the retrieved information is often a by-product of existing business processes. Thus, the added value of using a model based on a random mailing, instead of using other response data, has to be substantial if one is to cover the costs incurred.

WHY A *RANDOM* PILOT GROUP?

Given the potential cost involved in sending a random mailing and that many direct mail campaigns are of a more or less cyclical nature, marketing professionals sometimes wonder why pilot mailings should specifically be drawn randomly from the target population. Why not use data obtained from the previous mail campaign to build a new model? The answer has to do with representativeness. If a model is built on response that was recorded from a randomly drawn sample, one can be sure that effects found in the pilot sample will generalise to the target population it was drawn from. In the case of a targeted mailing (whichever kind of model is used), effects found in the pilot sample as to who would respond and who not, do not necessarily generalise to the entire population. What has happened is that the sample has become censored,[11] an effect that has been widely studied in operations research.

Furthermore, selecting a random group has the benefit of allowing model performance to be monitored. Assessing model stability can reveal highly relevant information, particularly when a model is reused frequently over a long period of time (as in cyclical direct mail campaigns). Such information is likely to provide invaluable input on product and model life cycles.[12]

An example is given here to illustrate the problems associated with censored samples. For the sake of simplicity we will assume that the target population for a certain product is any client in the customer base aged 18–65. It is assumed that the product manager decided to mail those customers with incomes in the top 10 per cent. A model to predict response in the 'wealthy' might be quite different from a model targeted at other types of customers. It is conceivable that the variables that discriminate between respondents and non-respondents in the high income group are quite different from the variables that discriminate between respondents and non-respondents within other groups. Moreover, if there is any kind of interaction between income and a predictive variable in the marketing database, we are unable to discover this effect since there are only wealthy customers in the training sample. So, the final conclusion will be that a model built on a censored sample only generalises to the subpopulation with the characteristics of this censored sample, and not necessarily to the *entire* target population from which it was drawn.

DATA CHOICE AMONG AVAILABLE OPTIONS

As stated above, three options exist for data selection aimed at retrieving a target variable: response to direct mail, data on product acquisition and behaviours associated with an interest in the product. Clearly, if a representative sample of responses to the proposed direct mail offer is available this will be the preferred choice of data. But how large should this sample be? To date, no seminal work on this topic has been published. Above we gave some guidelines, such as the absolute minimum of 100 respondents, based on our experience. More precisely, the necessary size of the training sample depends on multiple factors: what modelling technique is to be used, what level of prediction is sought, how heterogeneous the target population is, and the magnitude of differences between respondents and non-respondents, to name a few. If there was a single rule it would have to be as follows: the more training records, the better. So, from a model building perspective the random sample never contains enough clients. The financial feasibility of a large training sample is, however, a different issue. If response percentages in a random sample are low (compared to the targeted group), acquiring a large training sample can be very costly, as stated above. Somewhere an optimum has to be sought. Now, what if no direct mail sample for purposes of model building can be obtained at acceptable costs? What are the alternatives?

If there are no data on direct mail response available, then one could build models using data on product acquisition or interest-related behaviours. If product acquisition is used then the target variable becomes the following: an indication whether a person has acquired this product within a specified time frame preceding the planned campaign. One would typically collect data on the characteristics of customers at the beginning of this time frame who acquired the product, and compare these with data at the beginning of the time frame from customers who did not buy the product. This latter information is the array of explanatory variables in the model building process.

An example illustrates the propositions made above. Consider the following segment of clients found in a hypothetical marketing database at 31st December, 1998:

— age between 30 and 45 years of age
— monthly income of at least 4,000 Euro
— owns at least one savings account.

Furthermore, 5 per cent of these subjects acquired an investment trust in 1999, while only 1 per cent of subjects not conforming to these characteristics acquired this product in the period. In this case subjects aged 30 to 45, with a monthly income of at least 4,000 Euro, owning one or more savings accounts may be targeted in direct mail campaigns offering investment trusts. Note that in practice the above mentioned client characteristics are found using data mining techniques.

A drawback of the approach just described is that solicited and unsolicited acquisitions of products are confounded. Moreover, no prediction about the expected response rates for the direct mail campaign can be made. One could score or segment the customer base but there is no guidance in terms of profitability on where to cut off customers who should be solicited and those who should not receive the offer. Also, the ensuing workload when customers respond is impossible to predict. These can be serious limitations.

In the same way that data on product acquisition were used in the previous example, one could conceive using data on interest-related customer behaviour like calling in for information or sending in a coupon of sorts. Such information can be used in two ways:

— subjects who apply for information or send in a coupon have shown interest in a particular product. These subjects may be considered to be prospects, and can be mailed if they have not acquired the product yet
— the information can also be used to establish the characteristics of clients interested in particular products. For example, consider that the output of a data mining analysis shows that subjects between 25 years and 40 years of age with an income of at least 3,000 Euro apply for information on mortgages relatively often. This implies that this segment may be considered as the target group for a direct mail campaign in which mortgages offered by the financial services provider are introduced to its clients. Thus, applying for information may act as the target variable.

Nevertheless, calling in for information or other interest related-behaviours are even more distinct from responding to an offer than acquisition *per se*. Thus, the predictive validity of this representation of response is somewhat lower than data on acquisition. But the line is blurred. If much effort is put into acquiring customers, with enticing premiums, alongside high numbers of autonomous (unsolicited) product acquisition, data on product acquisition become a less desirable source for the target variable. If, on the other hand, people requesting a mortgage offer (a time intensive effort on the part of the client) might be targeted, they are a good bet for targeting, given short lag times (or the prospects will have acquired their mortgages elsewhere!).

In summary, response data are the most desirable, but are also the most difficult to obtain and the most expensive. If it is not feasible to obtain response data then alternative target variables have to be sought. In general, data on product acquisition are to be preferred to data on interest-related behaviours. But both suffer from the common problem: it is impossible to give an educated guess about the expected response to the proposed direct mail campaign. An answer to these problems lies in our proposed hybrid approach.

HYBRID DATA SOURCES: ONLY *TESTING* ON RANDOM RESPONSE

As stated before, in some cases there might not be a high enough response to a randomly selected direct mail offer to build a model. The answer in these cases lies in building models on less ideal data (data on product acquisition or interest-related behaviours), and subsequently testing them on response to a small random mailing. The rationale being that if there is not enough response to build (ie train) a model, there might still be enough response to validate the model. Note that less response is required when validating a model,[13] thus a smaller number of subjects have to be included in the random mailing, implying lower costs. This hybrid approach has other advantages: first, the model is validated on a data source with highly desirable validation properties, and secondly it now becomes possible to present an estimate of response to the ensuing direct mail campaign.

There is an additional advantage in the hybrid approach, which is particularly relevant if multiple models have been distilled from the available data. (This is often the case, as data mining tools will allow several models to be built on the same data set.) Based on the same data set, many models can be built with more or less equivalent performance on the training data. What is even more interesting is that it is usually possible to come up with entirely different explanatory variables in these models, especially when there is high multi-collinearity among exogenous (= explanatory, predictive) variables. Often no single model stands out as substantially better than other models.[14] The way in which predictive scores could be combined is a separate topic in its own right. The use of 'ensembles'[15] will be an interesting research topic for data miners in the future.

The application of various techniques can result in multiple models being extracted from the data. With logit models it is not uncommon to have several different sets of variables all showing comparable predictive power. With neural networks it is even possible to create an entire array of networks with differing numbers of inputs. Even with a fixed set of inputs the layout of the topology may be changed. The layout can be changed, the number of hidden layers, the number of neurones per layer, etcetera. Determining the best layout for a neural network remains an issue of debate among professionals.[16]

With recursive partitioning models one can create entirely different trees on the same data source by manipulating the choice of splitting variables at the top of the tree. Again, these different models might all show comparable discrimination power. The point is that the proposed hybrid approach can also support decisions on model choice. This approach may also prove to be particularly useful for genetic algorithms.[17] These utilise the principles of evolution to select among sets of models (generations), whereby a generation of reasonably 'fit' offspring might, at some point, be tested against

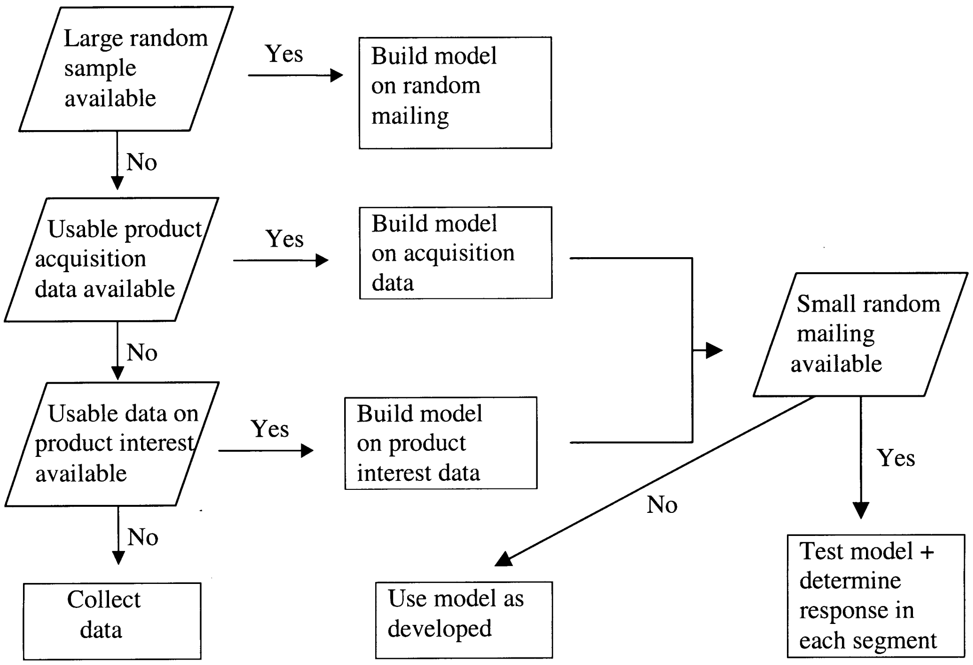
Large random sample available
Yes
Build model on random mailing
No
Usable product acquisition data available
Yes
Build model on acquisition data
No
Usable data on product interest available
Yes
Build model on product interest data
No
Collect data
Small random mailing available
No
Yes
Use model as developed
Test model + determine response in each segment

Figure 1: Response data schema

the limited random sample of direct mail responses. Last, but certainly not least, when using neural networks, the hybrid approach may also be valuable since training a network is severely impeded by having insufficient records.

DISCUSSION

The paper opened with observations regarding the lack of systematic discussion on which data sources are suitable for providing the target variable in predictive modelling for direct mailing. Note that we defined a target variable as being a variable that functions as a representation of the behaviour that is to be modelled, response behaviour for predictive model purposes. These remarks are particularly applicable for the topic of choosing between data that are to be used as a representation of the modelled behaviour. The main goal of this paper was to present a systematic discussion concerning choices that can be made with regard to the selection of this type of data.

First, we proposed that response to a pilot mailing, randomly drawn from the target population, is the most desirable target variable, for reasons of predictive validity. Practical constraints may make this impossible, usually because of high costs involved in acquiring the data or sometimes because of campaign (time) constraints.[18] All data mining techniques benefit from large training and testing samples, so at some point a trade-off has to be made. The costs involved in acquiring more training data should be justified by the resulting higher predictive power. Furthermore, if there is not enough response to a random pilot mailing, we proposed using product acquisition as the second best source for the target data, or, failing that, interest-related behaviours. If at all possible, models built on product acquisition or interest-related behaviours should be tested against a small response sample from a random pilot mailing. This testing is desirable for two reasons. First, it will allow for better validation and will thus generalise better to the target population. Secondly, because this hybrid approach enables one to forecast response, more critical validation is possible. The propositions just made can be summarised in one reasonably simple scheme, see Figure 1.

References

1 Paas, L. (1999) 'Revenue-driven direct mail campaigns', *Journal of Database Marketing*, Vol. 6, No. 4, pp 321–329.

2 Linoff, G. and Berry, M. (1997) 'Data mining techniques for marketing, sales and customer support', Wiley & Sons, New York (NY).

3 Dhar, V. and Stein, R. (1997) 'Seven methods for transforming corporate data into business intelligence', Prentice Hall, Upper Saddle River (NJ).

4 Pyle, D. (1999) 'Data preparation for data mining', Morgan Kaufman, San Francisco (CA).

5 Bult, J. R. (1993) 'Target selection for direct marketing', doctoral thesis, University of Groningen (The Netherlands).

6 Rossi, P. E., McCulloch, R. E. and Allenby, G. M. (1996) 'The value of purchase history data in targeting marketing', *Marketing Science*, Vol. 15, No. 4, pp. 321–340.

7 David Sheppard Associates (1995) 'The new direct marketing: How to implement a profit-driven database marketing strategy', Irwin Professional Publishing, New York (NY).

8 Paas, L. and Kuijlen, T. (1998) 'Analysing generic needs through Mokken Scaling and the Multi Nominal Logit model', *Journal of Targeting, Measurement and Analysis for Marketing*, Vol. 7, No. 2, pp. 145–154.

9 David Sheppard Associates (1995) *op. cit.*

10 Council on financial competition (1996) 'Present at the creation: Redefining competitive advantage through data-driven marketing and management', The advisory board company, Washington DC.

11 Heckman, J. J. (1979) 'Sample selection bias as a specification error', *Econometrica*, Vol. 47, No. 1, pp. 153–161.

12 Paas (1999) *op. cit.*

13 *idem.*

14 Wagenaar, E. (1997) 'Data mining in marketing databases', DMSA, Amsterdam (The Netherlands).

15 Merz, C. J. (1999) 'Using correspondence analysis to combine classifiers', *Machine Learning*, Vol. 36, Nos. 1–2, pp. 33–58.

16 Bigus, J. P. (1996) 'Data mining with neural networks', McGraw-Hill, New York (NY).

17 Goldberg, D. E. (1989) 'Genetic algorithms in search, optimization & machine learning', Addison-Wesley, Reading (MA).

18 Paas and Kuijlen (1998) *op. cit.*