63

Data Mining for Target Marketing

Nissan Levin[1] and Jacob Zahavi[2]

[1] Q-Ware Software Company, Israel
[2] Tel-Aviv University

Summary. Targeting is the core of marketing management. It is concerned with offering the right product/service to the customer at the right time and using the proper channel. In this chapter we discuss how Data Mining modeling and analysis can support targeting applications. We focus on three types of targeting models: continuous-choice models, discrete-choice models and in-market timing models, discussing alternative modeling for each application and decision making. We also discuss a range of pitfalls that one needs to be aware of in implementing a data mining solution for a targeting problem.

Key words: Targeting, predictive modeling, decision trees, clustering, survival analysis, in-market timing

63.1 Introduction

Targeting is at the core of marketing management. It is concerned with offering the right product to the customer at the right time and using the proper channel. Indeed, marketing has gone a long way from the mass marketing era where everybody was exposed to the same product, to today's fragmented and diversified markets. The focus has changed from the product to the customer. Instead of increasing market share the objective has shifted to increasing customer share and enhancing customers' loyalty and satisfaction. Recent developments in computer and database technologies are helping these goals by harnessing database marketing, Data Mining and more recently CRM technologies to better understand the customer thus approach her only with products and services that are keen to her. Various marketing metrics have been developed to evaluate the effectiveness of marketing programs and keep track of the profit and costs of each individual customer.

From a Data Mining point of view, we classify the targeting problems into three main categories, according to the variable that we are attempting to predict (the dependent, the choice or the response variable) – discrete choice, continuous choice and in-market timing problems. Each type of problem requires a different type of model to solve.

O. Maimon, L. Rokach (eds.), *Data Mining and Knowledge Discovery Handbook*, 2nd ed.,
DOI 10.1007/978-0-387-09823-4_63,

Discrete choice problems are targeting problems where the response variable is discrete (integer value). The simplest is the binary choice model where the dependent variable assumes two values, usually 0 and 1, e.g.: 0 – do not buy, 1- buy (a product or service). A generalization is the multiple choice model where the dependent variable assumes more than 2 nominal values, e.g., 3 values: 0 – do not buy, 1 - buy a new car, 2 - buy a used car. A special case of a discrete choice is where the dependent variable assumes several discrete values which possess some type of order, or preference. An example in the automotive industry would be: 0- no buy, 1 – buy a compact car, 2 – buy an economy car, 3- buy a midsize car, 4 – buy a luxury car, where the order here is defined in terms of the car segment in increasing order of size.

Continuous choice problems are targeting problems where the choice variable is continuous. Examples are money spent on purchasing from a catalog, donations to charity, year-to-date interest paid on a loan/mortgage, and others. What makes continuous targeting problem in marketing special is the fact that the choice variable is non-negative, i.e., either the customer responds to the solicitation and purchases from the catalog or the customer declines the offer and spends nothing.

Mixed types of problems also exist. For example, continuous choice problems which are formulated as discrete choice models (binary or ordinal), and discrete choice models which are expressed as continuous choice problems (e.g., predicting the number of purchases, where the frequency of purchase assumes many discrete values 0,1,2,... and is thus approximated by a continuous choice).

In-Market timing problems are time-related targeting problems where the objective is to predict the time of next purchase of a product or service. For example, when the customer will be in the market to purchase a new car? When s/he is up to taking the next flight or next cruise trip? Etc.

In this chapter, we discuss how Data Mining modeling and analysis can support these targeting problems, ranging from segmentation-based targeting programs to detailed ”one-to-one” programs. For each of models we also discuss the decision making process. Yet, this process is not risk free as there are many pitfalls that one needs to be aware of in building and implementing a targeting program based on Data Mining, which, if not cared for, could lead to erroneous results. So we devote a great deal of efforts to suggesting ways to identify these pitfalls and ways to fix them.

This chapter is organized as follows: In Section 63.2 we discuss the modeling process for a typical targeting application of a new product, followed by a brief review, in Section 63.3, of the common metrics used to evaluate the quality of targeting models. In sections 63.4,63.5,63.6, we discuss the three class of models to support targeting decisions - segmentation, predictive modeling and in-market timing models, respectively. In Section 63.7 we review a host of pitfalls and issues that one needs to be aware of when building and implementing a targeting application involving Data Mining. We conclude, in Section 63.8, with a short summary

63.2 Modeling Process

Figure 63.1 exhibits the decision process for a targeting application. In the case of a new product, the process is often initiated by a test mailing to a sample of customers in order to assess customers' response. Then people in the audience who ”look like” the test buyers are selected for the promotion. For a previously promoted program, the modeling process is based on the results of the previous campaign for the same product. The left hand side of Figure 63.1

corresponds to the testing phase, the right hand side to the rollout phase. The target audience, often referred to as the universe, is typically, but not necessarily, a subset of the customer list containing only customers who, based upon some previous consideration, make up potential prospects for the current product (e.g., people who have been active in the last, say, three years). The test results are used to calibrate a response model to identify the characteristics of the likely buyers. The model results are then applied against the balance of the database to select customers for the promotion. As discussed below, it is a good practice to split the test audience into two mutually exclusive data sets, a training set to build the model with and a validation (or a holdout set) to validate the model with. The validation procedure is essential to avoid over fitting and make sure that the model produces stable results that could be applied to score a set of new observations. Often there is a time gap between the time of the test and the time of the rollout campaign because of the lead time to stock up on the product. Since the customer database is highly dynamic and changes by the minute, one has to make sure that the test universe and the rollout universe are compatible and contain the same ”kind” of people. For example, if the test universe contains only people who have been active in the last three years prior to the test, the rollout universe should also include only the last three-year buyers. Otherwise we will be comparing apples to oranges thereby distorting the targeting results.

We note that the validation data set is used only to validate the model by comparing predicted to actual results. The actual decisions, however, are based only on the predicted profit/response for the training set. The decision process proceeds as follows:

- Build a model based on the training set
- Validate the model based on the validation set. Below we discuss a variety of metrics to evaluate and assess the quality of a predictive model.
- If the resulting model is not ”good enough”, build a new model by changing the parameters, adding observations, trying a different set of influential predictors, use a different type of model, new transformations, etc. Iterate, if necessary
- Once happy with the model, apply the model to predict the value of the dependent variable for each customer in the rollout universe. This process is often referred to as ”scoring” and the resulting predicted values as ”scores”. These scores may vary from model to model. For example, in logistic regression, the resulting score is the purchase probability of the customer.
- Finally, use economic criterion to select the customers for targeting from the rollout universe. These economic criteria may vary between models, and so we discuss them below in the context of each class of models. Note that the rollout universe does not have any actual values for the current promotions. Hence decisions should be based solely on predicted values, i.e., the calculated scores.

63.3 Evaluation Metrics

Several metrics are used to evaluate the results of targeting models. These are divided into goodness-of-fit measures, prediction accuracy and profitability/ROI measures.

63.3.1 Gains Charts

Prediction models are evaluated based on some goodness-of-fit measures which assess how good the model fits the data. However, unlike the scalar values used to assess overall fit (e.g.,

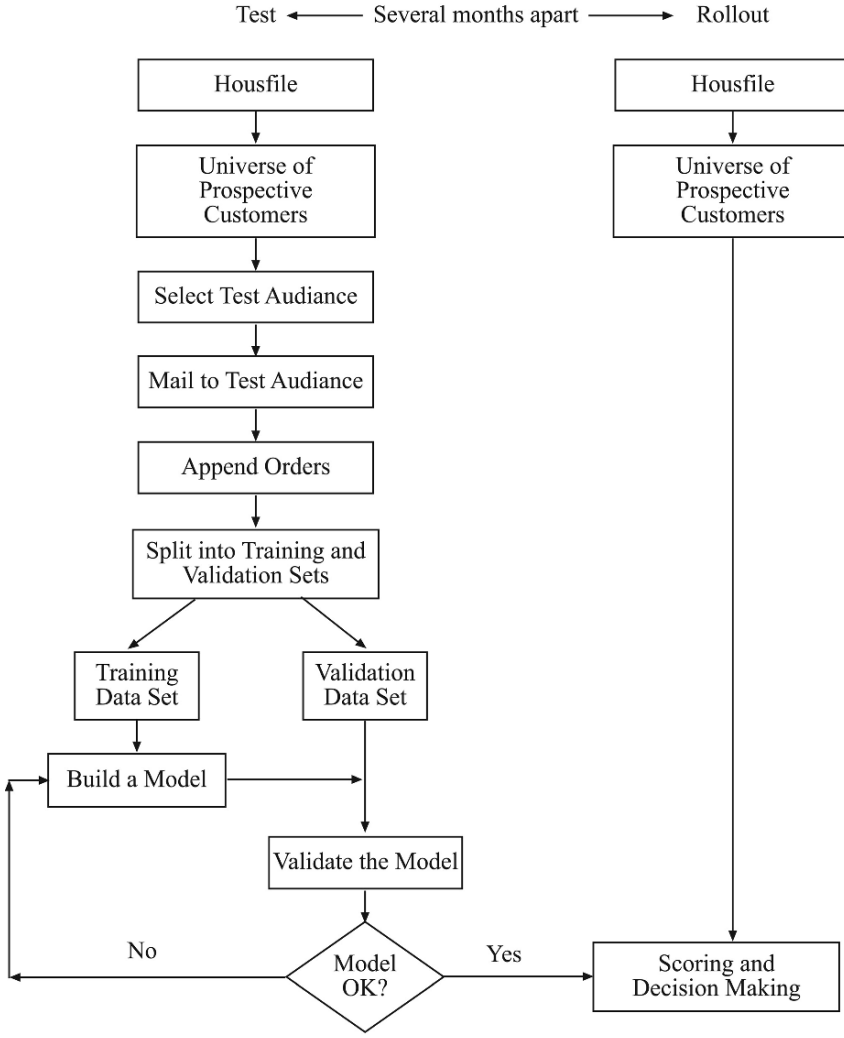

Fig. 63.1. The decision Making Process for New Promotions

the coefficient of determination in linear regression, misclassification rates in classification models, etc), in targeting applications we are interested in assessing the gains achieved by the model, or how well the model is capable of discriminating between buyers and non buyers. Thus the relevant goodness-of-fit measure is based on the distribution of the targeting results, known as gains chart.

Basically, gains-chart displays the added gains (for instance profitability or response) by using a predictive model versus a null model that assumes that all customers are the same. The X-axis represents the cumulative proportion of the population $X_i = 100 \cdot i/n$, (where n is the

size of the audience, i - the customer index). The Y-axis represents the cumulative proportion of the actual response (e.g., proportion of buyers), $Y_i = 100 \cdot \frac{\sum_{j=1}^{i} y_j}{\sum_{j=1}^{n} y_j}$ where the observations are ordered in descending order of the predicted values of the dependent variable, i.e., $\hat{y}_i \leq \hat{y}_{i+1}$. A typical gains chart is exhibited in Figure 63.2. We note that gains charts are similar to Lorenz curves in economics.

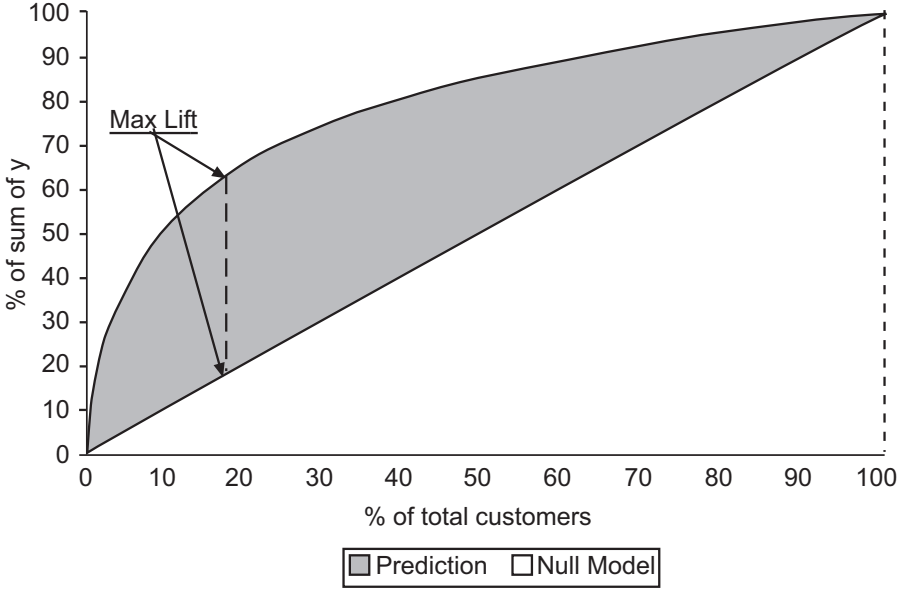

Fig. 63.2. Gains Chart

Two metrics, based on the gains chart, are typically used to assess how the model results differs from the null model:

- Maximum Lift (ML), more commonly known as the Kolmogorov Smirnov (K-S) criterion (Lambert, 1993), which is the maximum distance between the model curve and the null model. The K-S statistics has a distribution known as the D distribution (DeGroot, 1991). In most applications, a large ML indicates that the distribution of the model results is different from the null model. The D distribution can be approximated when the number of observation n is large. For large n, the null hypothesis that the two distributions are the same, is rejected with a significance level of 5% if $ML > D_{95} \approx \frac{1.36}{\sqrt{n}}$ (Gilbert, 1999, Hodges, 1957)
- The Gini Coefficient (Lambert, 1993) which is calculated as the area between the model curve and the null model (the gray area in Figure 63.2) divided by the area below the null model. In most applications, a large Gini coefficient indicates that the distribution of the model results is different from the null model.

Clearly, the closer the model curve to the upper left corner of the chart, the better the model is capable of distinguishing the buyers from the non buyers. Equivalently, the larger is the maximum lift or the larger the Gini coefficient, the better the model.

We note that the gains chart is the metrics which reflect the true prediction quality of the model. The lift and the Gini coefficients are summary measures that are often used to compare between several candidate models. Moreover, the maximum lift and the Gini coefficient may not be consistent with one other. For example, it is possible to find two alternative models, build off the same data set, where in one the Gini coefficient is larger, and in the other the *ML* is larger.

63.3.2 Prediction Accuracy

Prediction accuracy is measured by means of the difference in the predicted response versus the actual results, the closer the values the better. Again, it is convenient to view the prediction results at a percentile level, say deciles. Gains tables or bar charts are often used to exhibit the prediction results.

63.3.3 Profitability/ROI

Definitely, the ultimate merit of any model is given by profitability/ROI measures, such as the total profits/return for the target mailing audience and/or the average profits/return per mailed customer. A common ROI measure is given by the incremental profits of the targeted mailing versus using no model (null model) and mailing everybody in the list.

63.3.4 Gains Table

The tabular form of the gains chart is referred to as the gains table. Gains tables are exhibited at some percentiles level, often deciles, i.e. the predicted y-values (e.g., response probabilities) are arranged in decreasing order and the audience is divided into percentiles. The actual results are then summarized at the percentile level (see Table 63.1).

A ”good” model is a model which satisfies the following criteria:

- The actual response rate (the ratio of the number of buyers ”captured” by the model to the size of the corresponding audience) monotonically decreases as one traverses from the top to the bottom percentiles.
- A large difference in the response rate between the top and the bottom percentiles.

For example, in Table 63.1, the first decile captures 6 times as many buyers as the bottom decile (30 vs. 5) for the same audience size (708 customers). Except for minor fluctuations at the lower deciles because of few buyers, the number of buyers nicely declines as one traverse along deciles suggesting that the model is capable of distinguishing between the better and worse customers.

The economic cutoff rate for this problem (more on this below) falls in the fifth decile, at which point the profits attains a maximum value (highlighted in bold). The profits are calculated by multiplying the actual number of buyers by the profit per order and subtracting the mailing costs.

To assess the prediction accuracy of the model, one should compare the actual number of buyers to the predicted number of buyers (calculated by summing up the purchase probabilities) at the decile level, the closer the value, the better. For the interesting deciles at the top of

the list, the prediction accuracy is indeed high (e.g., 30 actual buyers for the first decile vs. 32 predicted buyers, etc.)

Note that the example above represents the results of a discrete choice model where the model performance is measured by means of the response rates. These measures should be substituted by profit values in case of continuous choice modeling. In other words, one needs to arrange the observations in decreasing order of the predicted profits, divide up the list into deciles, or some other percentiles, and create the gains chart or gains table.

Finally we emphasize that whatever model is used, whether discrete or continuous, the validation process of a model should always be based on the validation data set and not on the training data set. Being an independent data set, the validation file is a representative of the audience at large and as such is the only file to assess the performance of the model when applied in the ”real” world.

63.4 Segmentation Methods

Segmentation is a central concept in marketing. The concept was formally introduced by Smith (1956) and since then has become a core method for supporting targeting applications. Segmentation is concerned with splitting the market into groups or segments of ”like” people with similar purchase characteristics. The key to successful segmentation is identifying a measure of ”similarity” between customers with respect to purchasing pattern. The objective of segmentation is to partition the market into segments which are as homogenous as possible within the segments and as heterogeneous as possible in between segments. Then, one may offer each segment only the products/services which are of most interest to the members of the segment. Hence the decision process is conducted at the segmentation level, either the entire segment is targeted for the promotion or the entire segment is declined.

Segmentation methods which are used to address targeting decisions consist of unsupervised judgmentally-based RFM/FRAT methods, clustering methods and supervised classification methods. Supervised models are models where learning is based on some type of a dependent variable. In unsupervised learning, no dependent variable is given and the learning process is based on the attributes themselves.

63.4.1 Judgmentally-based RFM/FRAT methods

Judgmentally based or ”manual” segmentation are still commonly used to partition a customer list into ”homogenous” segments for targeting applications. Typical segmentation criteria include previous purchase behavior, demographics, geographic and psychographics. Previous purchase behavior is often considered to be the most powerful criterion in predicting likelihood of future response. This criterion is operationalized for the segmentation process by means of Recency, Frequency, Monetary (RFM) variables (Shepard, 1995). Recency corresponds to the number of weeks (or months) since the most recent purchase; frequency to the number of previous purchases or the proportion of mailings to which the customer responded; and monetary to the total amount of money spent on all purchases (or purchases within a product category), or the average amount of money per purchase. The general convention is that the more recently the customer has placed the last order, the more items s/he bought from the company in the past, and the more money s/he spent on the company’s products, the higher is his/her likelihood of purchasing the next offering and the better target s/he is. This simple rule allows one to arrange the segments in decreasing likelihood of purchase.

Table 63.1. Gains Table by Deciles

Response Prob. (%)	Cum. Aud.	Actual No. of Responders Decile	Cum.	% Buyer	Actual Profit ($)	Pred. # of Responders Decile	Cum.
2.90	708	30	30	26.55	9168	32	32
2.15	1416	20	50	44.25	14436	18	50
1.55	2124	14	64	56.64	17104	14	64
1.17	2832	10	74	65.49	18272	9	73
0.85	3540	10	84	74.34	**19440**	**7**	80
0.73	4248	5	89	78.76	18608	6	86
0.57	4956	8	97	85.84	18976	5	91
0.46	5664	7	104	92.04	18056	4	95
0.27	6372	4	108	95.58	17712	3	98
0.02	7077	5	113	100.00	16892	1	99

The more sophisticated manual methods also make use of product/attribute proximity considerations in segmenting a file. By and large, the more similar the products bought in the past are to the current product offering, or the more related are the attributes (e.g., themes), the higher the likelihood of purchase. For example, when promoting a sporting good, it is plausible that a person who bought another sporting good in the past is more likely to respond to a new sporting good offer; next in line are probably people who like camping, followed by people who like outdoor activities, etc. In cases where males and females may react differently to the product offering, gender may also be used to partition customers into groups. By and large, the list is first partitioned by product/attribute type, then by RFM and then by gender (i.e., the segmentation process is hierarchical). This segmentation scheme is also known as FRAT - Frequency, Recency, Amount (of money) and Type (of product).

RFM and FRAT methods are subject to judgmental and subjective considerations. Also, the basic assumption behind the RFM method may not always hold. For example, in durable products, such as cars or refrigerators, recency may work in a reverse way - the longer the time since last purchase, the higher the likelihood of purchase. Finally, to meet segment size constraints, it may be necessary to run the RFM/FRAT iteratively, each time combining small segments and splitting up large segments, until a satisfactory solution is obtained. This may increase computation time significantly.

63.4.2 Clustering

Clustering are methods for grouping unlabeled observations. A data item is mapped into one of several clusters as determined from the data. Modern clustering algorithms group data elements based on the proximity (or similarity) of their attributes. The objective is to partition the observations into ”homogeneous” clusters, or groups, such that all observations (e.g., customers) within a cluster are ”alike”, and those in between clusters are dissimilar. In the context

of our targeting applications, the purpose of clustering is to partition the audience into clusters of people with similar purchasing characteristics. The attributes used in the clustering process are the same as those used by the RFM/FRAT method discussed above. In fact, clustering methods take away the judgmental considerations used by the subjective RFM/FRAT methods, thereby providing more "objective" segments. Once the audience is partitioned into clusters, the targeting decision process proceeds as above.

Let $X_i = (x_{i1}, x_{i2}, ..., x_{iJ})$ denotes the attribute vector of customer i with attributes $x_{ij}, j = 1, 2, ..., J$. To find which customers cluster together, one needs to define a similarity measure between customers. The most common one is the Euclidean distance. Given that ℓ and m are two customers from a list of n customers, the Euclidean distance is defined by:

$$\text{distance } (X_\ell, X_m) = \sqrt{\sum_j \left(x_{\ell j} - x_{m j}\right)^2}$$

Clearly, the shorter the distance, the more similar are the customers. In the case of two identical customers, the Euclidean distance returns the value of zero.

An alternative distance measure, which may be more appropriate for binary or integer attributes, is the cosine distance (Herz *et al*., 1997), defined by:

$$\text{distance } (X_\ell, X_m) = \frac{\sum_{j=1}^{J} x_{\ell j}\, x_{m j}}{\sqrt{\sum_{j=1}^{J} x_{\ell j}^2 \sum_{j=1}^{J} x_{m j}^2}}$$

Note that unlike in the Euclidean distance, here if the purchase profile of customers are identical, the cosine measure returns the value of 1; if orthogonal (i.e., totally dissimilar), it returns the value of 0.

Several clustering algorithms have been devised in the literature, ranging from K-Means algorithms (Fukunaga, 1990), Expectation Maximization (Lauritzen, 1995), Linkage-based methods (Bock, 1974), Kernel Density estimation (Silverman, 1986), even Neural Network based models (Kohonen *et al*., 1991). Variations of these models that address the scalability issues are appearing recently in the literature, e.g., the BIRCH algorithm (Zhang *et al*., 1996).

The K-Means algorithm is, undoubtedly, the most popular of all clustering algorithms. It partitions the observations (customers, in our case) into K clusters, where the number of clusters K is defined in advance, based on the proximity of the customer attributes from the center of the cluster (called the centroid).

Let S_k denote the centroid of cluster k. S_k is a vector with J dimensions (or coordinates), one coordinate per attribute j. Each coordinate of the centroid of a cluster is calculated as the mean value of the corresponding coordinates of all customers which belong to the cluster (hence the name K-Means).

To find which customers belong to which cluster, the algorithm proceeds iteratively, as follows:

Step 1 – Initialization: Determine the K centroids, S_k, $k = 1, ..., K$ - (e.g., randomly).

Step 2 – Loop on all customers: For each customer, find the distance of his/her profile to each of the centroids S_k, $k = 1, ..., K$, using the specified similarity measure, and assign the customer to the cluster corresponding to the nearest centroid.

Step 3 – Loop on all clusters: For each cluster $k = 1, ..., K$, recalculate the coordinates of its centroid by averaging out the coordinates of all customers currently belonging to the centriod.

Step 4 – Termination: Stop the process when the termination criteria are met – otherwise, return to Step 2.

63.4.3 Classification Methods

Classification models are segmentation methods which use previous observations with known class labels (i.e., whether the customer responded to the offer, or not) in order to classify the audience into one of several predefined classes. Hence these models belong to the realm of supervised learning. They too take away the judgmental bias that is inherent in the subjective RFM/FRAT methods.

The leading classification models are decision trees. In the binary response case, the purpose is to segment customers into one of two classes – likely buyers and likely non buyers. Some decision trees allow splitting an audience into more than 2 classes. Several automatic tree classifiers were discussed in the literature, among them AID - Automatic Interaction Detection (Sonquist *et al*., 1971); CHAID - Chi square AID (Kass, 1983), CART - Classification and Regression Trees (Breiman *et al*., 1984), ID3 (Quinlan, 1986), C4.5 (Quinlan, 1993), and others. A comprehensive survey of automatic construction of decision trees from data can be found in Chapter 8.8 of this volume.

Basically, all automatic tree classifiers share the same structure. Starting from a ”root” node (the whole population), tree classifiers employ a systematic approach to grow a tree into ”branches” and ”leaves”. In each stage, the algorithm looks for the ”best” way to split a ”father” node into several ”children” nodes, based on some splitting criteria. Then, using a set of predefined termination rules, some nodes are declared as ”undetermined” and become the father nodes in the next stages of the tree development process, some others are declared as ”terminal” nodes. The process proceeds in this way until no more nodes are left in the tree which are worth splitting any further. The terminal nodes define the resulting segments. If each node in a tree is split into two children only, one of which is a terminal node, the tree is said to be ”hierarchical”.

Three main considerations are involved in developing automatic trees:

- Growing the tree
- Determining the best split
- Termination rules

Growing the Tree

One grows the tree by successively partitioning nodes based on the data. With so many variables involved, there is practically infinite number of ways to split a node. Several methods have been applied in practice to reduce the number of possible partitions of a node to a manageable number:

- All continuous variables are categorized prior to the tree development process into small number of ranges (”binning”). A similar procedure applies for integer variables which assume many values (such as the frequency of purchase).
- Nodes are partitioned only on one variable at a time (”univariate” algorithm).
- The number of splits per each ”father” node is often restricted to two (”binary” trees).
- Splits are based on a ”greedy” algorithm in which splitting decisions are made sequentially looking only on the impact of the split in the current stage, but never beyond (i.e., there is no ”looking ahead”).

Several algorithms exist that relax some of these restrictions. For example, CHAID is a non binary tree as it allows splitting a node into several descendants. By and large, Genetic algorithms (GAs) are non greedy methods and can also handle multiple variables to split a node.

Determining the Best Split

With so many possible partitions per node, the question is what is the best split? There is no unique answer to this question as one may use a variety of splitting criteria each may result in a different ”best” split. We can classify the splitting criteria into two ”families”: node-value based criteria and partition-value based criteria.

- Node-value based criteria: seeking the split that yields the best improvement in the node value.

- Partition-value based criteria: seeking the split that separates the node into groups which are as different from each other as possible

Termination Rules

Theoretically, one can grow a tree indefinitely, until all terminal nodes contain very few customers, as low as one customer per segment. The resulting tree in this case is unbounded and unintelligible, having the effect of ”can't see the forest because of too many trees”. It misses the whole point of tree classifiers whose purpose is to divide the population into buckets of ”like” people, where each bucket contains a meaningful number of people for statistical significance. Also, the larger the tree, the larger the risk of overfitting. Hence it is necessary to control the size of a tree by means of termination rules that determine when to stop growing the tree. These termination rules should be set to ensure statistical validity of the results and avoid overfitting.

63.4.4 Decision Making

There's quite a distinction between the decision process in unsupervised and supervised models.

In the unsupervised RFM/FRAT and clustering methods, one would normally contact the top segments in the list, i.e., the segments which are most likely to respond to the solicitation. In the RFM/FRAT approach, the position of the segments in the hierarchy of segments represents, more-or-less, its likelihood of purchase, in ordinal terms. Thus people at the segment occupying, say, the 10^{th} position in the hierarchy of segments are usually (but not necessarily) more likely to buy the current product than people belonging to the succeeding segments, but are less likely to purchase the current product than people belonging to the preceding segments. In the clustering approach, there is no such clear-cut definition of the quality of segments, and one needs to assess how good are the resulting clusters by analyzing the leading attributes of the customers in each segment based on domain knowledge.

A more accurate approach for targeting is to conduct a live test mailing involving a sample of customers from the resulting segment to predict the response rate of each segment in the list for the current product offering. Several rules of thumb exist to determine the size of the sample to use for testing from each segment. The convention among practitioners is to randomly pick a proportional sample, typically 10% from each segment, to participate in the test mailing. Then, if the predicted response rate of the segment, based on the test result, is

high enough the remainder of the segment is rolled out; otherwise, it is not. The threshold level separating the strong and the weak segments depend on economical considerations. In particular, a segment is worth promoting if the expected profit contribution for a customer exceeds the cost of contacting the customer. The expected profit per customer is obtained as the product of the customer purchase probability, estimated by the response rate of the segment that the customer belongs to, by the profit per sold item.

This decision process is subject to several inaccuracies because of large Type-I and Type-II errors, poor prediction accuracy and regression to the mean, which fall beyond the scope of this chapter. Further discussion of these issues can be found in (Levin and Zahavi, 1996).

The decision process may be simpler with supervised classification models as no test mailing is required here. The objective is to contact only the "profitable" segments whose response rate exceeds a certain cutoff response rate (CRR) based on economical considerations. As discussed in the next section, the CRR is given by the ratio of the contact cost to the profit per order, perhaps bumped up by a certain profit margin set by management.

63.5 Predictive Modeling

Predictive modeling is the work horse of targeting issues in marketing. Whether the model involved is discrete or continuous, the purpose of predictive modeling is to estimate the expected return per customer as a function of a host of explanatory variables (or predictors). Then, if the predicted response measure exceeds a given cutoff point, often calculated based on economical and financial parameters, the customer is targeted for the promotion; otherwise, the customer is rejected.

A typical predictive model has the general form:

$Y = f(x_1, x_2, ..., x_J, U)$

Where:

Y - the response (choice variable)

X - $(x_1, ..., x_J)$ - a vector of predictors "explaining" customers' choice

U - a random disturbance (error)

There are a variety of predictive models and it is beyond the scope of this chapter to discuss them all. So we will only review here the two most important regression models used for targeting decisions – linear regression and logistic regression, as well as the AI-based neural network model. More information about these and other predictive models can be found in the database marketing and econometric literature.

63.5.1 Linear Regression

The linear regression model is the most commonly used continuous choice model. The model has the general form:

$$Y_i = \beta' X_i + U_i$$

Where:

- Y_i - The continuous choice variable for observation i
- X_i - Vector of explanatory variables, or predictors, for observation i
- β - Vector of coefficients
- U_i - Random disturbance, or residual, of observation i, and there exist $E(U_i) = 0$

Denoting the coefficient estimate vector by $\hat{\beta}$, the predicted continuous choice value for each customer, given the attribute vector X_i, is given by:

$$E(Y_i \mid X_i) = \hat{\beta}' X_i$$

Since the linear regression model is not bounded from below, the predicted response may turn out negative, in contrast with the fact that actual response values in targeting applications are always non-negative (either the customer responds to the offer and incurs positive cost/revenues, or does not respond and incurs no cost/revenues). This may render the prediction results of a linear regression model somewhat inaccurate.

In addition, the linear regression model violates two of the basic assumptions underlying the linear model:

- Because the actual observed values of Y_i consists of many zeros (non responders) but only a few responders, there is a large probability mass at the origin which ordinary least squares methods are not "equipped" to deal with. Indeed, other methods have been devised to deal with this situation, the most prominent ones are the Tobit (Tobin, 1958), and the two-stage model (Heckman, 1979).

- Many of the predictors in database marketing, if not most of them, are dichotomous (i.e., 0/1 variables). This may affect the test of hypotheses process and the interpretability of the analysis results.

A variation of the linear regression model, in which the choice variable Y_i is defined as a binary variable which takes on the value of 1 if the event occurs (e.g., the customer buys the product), and the value of 0 if the event does not occur (the customer declines the product), is referred to as the linear probability model (LPM). The conditional expectation $E(Y_i/X_i)$ in this case may be interpreted as the probability that the event occurs, given the attribute vector X_i. However, because the linear regression model in unbounded, $E(Y_i/X_i)$ can lie outside the probability range (0,1).

63.5.2 Logistic Regression

Logistic regression models are at the forefront of predictive models for targeting decisions. Most common is the binary model, where the choice variable is a simple yes/no, which is coded as 0/1: 0 – for "no" (e.g., no purchase), 1 - for "yes" (purchase). The formulation of this model stems from the assumption that there is an underlying latent variable Y_i^* defined by the linear relationship:

$$Y_i^* = \beta' X_i + U_i \tag{63.1}$$

Y_i^* is often referred to as the "utility" that the customer derives by making the choice (e.g., purchasing a product). But in practice, Y_i^* is not observable. Instead, one observes the response variable Y_i, which is related to the latent variable Y_i* by:

$$Y_i = \begin{cases} 1 \;\; if\, Y_i^* > 0 \\ 0 \;\; otherwise \end{cases} \tag{63.2}$$

From (63.1) and (63.2), we obtain:

$$\begin{aligned} Prob(Y_i = 1) &= Prob(Y_i^* = \beta' X_i + U_i > 0) \\ &= Prob(U_i > -\beta' X_i) = 1 - F(-\beta' X_i) \end{aligned} \tag{63.3}$$

Which yields, for symmetrical distribution of U_i around zero:

$$Prob\,(Y_i = 1) = F\left(\beta' X_i\right)$$

$$Prob(Y_i = 0) = F(-\beta' X_i)$$

Where F(·) denotes the CDF of the disturbance U_i.

The parameters $\beta' s$ are estimated by the method of maximum likelihood. In case the distribution of U_i is logistic, we obtain the **logit** model with closed-form purchase probabilities (Ben Akiva and Lerman, 1987):

$$Prob(Y_i = 1) = \frac{1}{1+exp(-\hat{\beta}'X)}$$

$$Prob(Y_i = 0) = \frac{1}{1+exp(\hat{\beta}'X)}$$

Where $\hat{\beta}$, the MLE (Maximum likelihood estimate) of β

An alternative assumption is that U_i is normally distributed. The resulting model in this case is referred to as the **probit** model. This model is more complicated to estimate because the cumulative normal variable does not have a closed-form solution. But fortunately, the cumulative normal distribution and the logistic distribution are very close to each other. Consequently, the resulting probability estimates are similar. Thus, for all practical purposes, one can use the more convenient and more efficient logit model instead of the probit model.

Finally we mentioned two more models which belong to the family of discrete choice models - multinomial regression models and ordinal regression models (Long, 1997). In multinomial models, the choice variable may assume more than two values. Examples are a trinomial model with 3 choice values (e.g., 0 – no purchase, 1 – purchase a new car, 2 – purchase a used car), and a quadrinomial model with 4 choice values (e.g., 0 – no purchase, 1 – purchase a compact car, 2 – purchase a mid size car, 3 – purchase a full size luxury car). Higher order multinomial models are very hard to estimate and are therefore much less common.

In ordinal regression models the choice variable assumes several discrete values which possess some type of an order, or preference. The above example involving the compact, mid size and luxury car, can also be conceived as an ordinal regression model with the size of the car being the ranking measure. By and large, ordinal regression models are easier to solve than multinomial regression models.

63.5.3 Neural Networks

Neural Networks (NN) are AI-based predictive modeling method which has gained a lot of popularity recently. NN is a biologically inspired model which tries to mimic the performance of the network of neurons, or nerve cells, in the human brain. Mathematically, a NN is made up of a collection of processing units (neurons, cells), connected by means of branches, each characterized by a weight representing the strength of the connection between the neurons. These weights are determined by means of a learning process by repeatedly showing the NN with examples of past cases for which the actual output is known, thereby inducing the system to adjust the strength of the weight between neurons. On the first try, since the NN is still untrained, the input neuron will send a current of initial strength to the output neurons, as determined by the initial conditions. But as more and more cases are presented, the NN will eventually learn to weigh each signal appropriately. Then, given a set of new observations, these weights can be used to predict the resulting output.

Many types of NN have been devised in the literature. Perhaps the most common one, which forms the basis of most business applications of neural computing, is the supervised-learning, feed-forward networks, also referred to as backpropagation networks. In this model, which resulted from the seminal work of (Rumelhart and McClelland, 1986), and the PDP

Research Group (1986), the NN is represented by a weighted directed graph, with nodes representing neurons and links representing connections. A typical feedforward network contain three types of processing units: input units, output units and hidden units, organized in a hierarchy of layers, as demonstrated in Figure 63.3 for a three-layer network. The flow of information in the network is governed by the topology of the network. A unit receiving input signal from units in a previous layer aggregates those signals based on an input function I, and generates an output signal based on an output function O (sometimes called a transfer function). The output signal is then routed to other units as directed by the topology of the network. The input function I often used in practice is the linear one, and the transfer function O either the tangent hyperbolic or the sigmoid (logit) function.

The weight vector W is determined through a learning process to minimize the sum of squared deviations between the actual and the calculated output, where the sum is taken over all output nodes in the network. The backpropagation algorithm consists of two phases: feedforward propagation and backward propagation. In forward propagation, outputs are generated for each node on the basis of the current weight vector W and propagated to the output nodes to generate the total sum of squared deviations. In backward propagation, errors are propagated back, layer by layer, adjusting the weights of the connections between the nodes to minimize the total error. The forward and backward propagation are executed iteratively once for each number of iterations (called epoch) until convergence occurs.

The type and topology of the backpropagation network depends on the structure and dimension of the application problem involved, and could vary from one problem to the other.

In addition, there are other considerations in applying NN for target marketing, which are not usually encountered in other marketing applications of NN (see Levin and Zahavi, 1997b, for descriptions of these factors). Recent research also indicates that NN may not have any advantage over logistic models for supporting binary targeting applications (Levin and Zahavi, 1997a). All this suggest that one should apply NN to targeting applications with cautious.

63.5.4 Decision Making

From the marketer's point of view, it is worth mailing to a customer as long as the expected return from an order exceeds the cost invested in generating the order, i.e., the cost of promotion. The return per order depends on the economical/financial parameters of the current offering. The promotion cost usually includes the brochure and the postal costs. Denoting by:

g - the expected return from the customer (e.g., expected order size in a catalog promotion).

c - the promotion cost

M - the minimum required rate of return

Then, the rate of return per customer (mailing) is given by:

$$\frac{g-c}{c} = \frac{g}{c} - 1$$

And the customer is worth promoting to if his/her rate of return exceeds the minimal required rate of return, M, i.e.:

$$\frac{g}{c} - 1 \geq M \rightarrow g \geq c \bullet (M+1) \tag{63.4}$$

The quantity on the right-hand side of (63.4) is the cutoff point separating out between the promotable and the nonpromotable customers.

Alternatively, equation (63.4) can be expressed as:

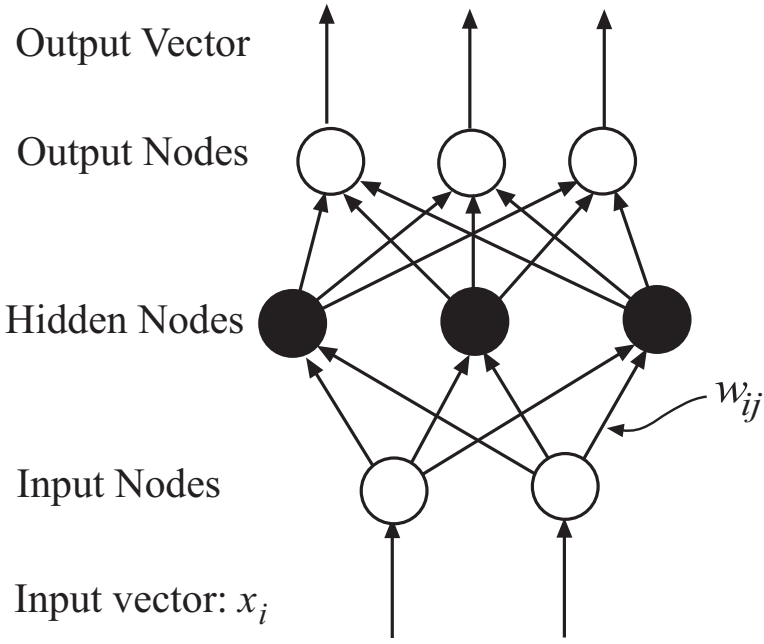

Fig. 63.3. A multi-layer Neural Network

$$g - c \bullet (M+1) \geq 0 \tag{63.5}$$

Where the quantity on the left-hand side denotes the net profit per order.

Then, if the net profit per order is non-negative, the customer is promoted; otherwise, s/he is not.

In practical applications, the quantity c is determined by the promotion cost; M is a threshold margin level set up by management. Hence the only unknown quantity is the value of g - the expected return from the customer, which is estimated by the predictive model. Two possibilities exist:

In a continuous response model, g is estimated directly by the model.

In a binary response model, the value of g is given by:

$$g = p \bullet R \tag{63.6}$$

Where:

p - The purchase probability estimated by the model, i.e. $p = Prob(Y = 1)$. Y is the purchase indicator - 1 for purchase, 0 for no purchase. R is the return/profit per responder.

In this case, it is customary to express the selection criterion by means of purchase probabilities. Plugging (63.6) in (63.4) we obtain:

$$p \geq \frac{c(M+1)}{R} \tag{63.7}$$

The right hand side of (63.7) is the cutoff response rate (CRR). If the customer's response probability exceeds CRR, s/he is promoted; otherwise, s/he is not.

Thus, the core of the decision process in targeting applications is to estimate the expected return per customer, g. Then, depending upon the model type, one may use either (63.4) or (63.7) to select customers for the campaign.

Finally we note that CRR calculation applies only to the case where the scores coming out from the model represent well-defined purchase probabilities. This is true of logistic regression, but less true for NN where the score is ordinal. But ordinal scores still allow the user to rank customers in decreasing order of their likelihood of purchase, placing the best customers at the top of the list and the worst customers at the bottom of the list. Then, in the absence of a well defined CRR, one can select customers for promotion based on "executive decision", say promote the top four deciles in the list.

63.6 In-Market Timing

For durable products such as cars or appliances, or events such as vacations, cruise trips, flights, bank loans, etc, the targeting problem boils down to the timing when the customer will be in the market "looking around" for these products/events. We refer to this problem as the in-market timing problem. The in-market timing depends on the customer's characteristics as well as the time that elapsed since last acquisition, e.g., the time since the last car purchase. Clearly, a customer that just purchased a new car is less likely to be in the market in the next, say, three months than a customer who bought his current car three years ago. Not only this, but the time until next car purchase is a random variable. We offer two approaches for addressing the in-market timing problem:

- Logistic regression – estimating the probability that the next event (car purchase, next flight, next vacation,...) takes place in the following time period, say next quarter.
- Survival analysis – estimating the probability distribution that the event will take place within the next time period t (called survival time), given that the last event took t_L units of time ago.

63.6.1 Logistic Regression

We demonstrate this process for estimating the probability that a customer will replace his/her old car in the next quarter. For this sake, we summarize the purchase information by, say, quarters, as demonstrated in Figure 63.4 below, and split the time axis into two mutually exclusive time periods – the "targeting period", to define the choice variable (e.g., 1 – if the customer bought a new car in the present quarter, 0 – if not), and the "history period" to define the independent variables (the predictors). In the example below, we define the present quarter as the target period and the previous four quarters as the history period. Then, in the modeling stage we build a logistic regression model expressing the choice probability as a function of the customer's behavior in the past quarters (the history period) and his/her demographics.

In the scoring stage, we apply the resulting model to score customers and estimate their probability of purchasing a car in the next quarter. Note the shift in the history period in the scoring process. This is because the model explains the purchase probability in terms of the customers' behavior in the previous four quarters. Consequently, and in order to be compatible with the model, one needs to shift the data for scoring by discarding the earliest quarter (the fourth quarter, in this example) and adding the present one.

We also note that the "target" period used to define the choice variable and the "history" period used to define the predictors, are not necessarily consecutive. This applies primarily

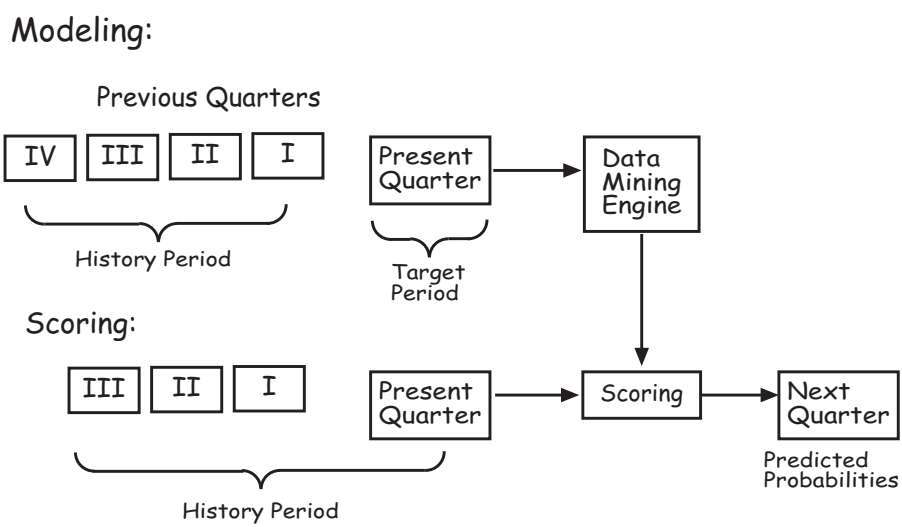

Fig. 63.4. In-Market Timing Using Logistic Regression

to purchase history and less to demographics. For example, in the automotive industry, since customers who bought a car recently are less likely to look around for a new car in the next quarter, one may discard customers who purchased a new car in the last, say, two years from the universe. So if the target period in the above example corresponds to the first quarter of 2004, the history period would correspond to the year 2001. There could also be some shift in the data because of the time lag that takes place between the actual transaction and the time the data becomes available for the analysis. Finally, we note that we used quarters in the above example just for demonstration purposes. In practice, one may use a different time period to summarize the data by, or a longer time period to express the history period. It all depends on the application. Certainly, in the automotive industry, because the purchase cycle to replace a car is rather long, the history period could extends over several years; moreover, this period should even vary from one country to the other, because the ”typical” purchase cycle time for each country is not the same. In other industries, these time periods could be much shorter. So domain knowledge should play an important role in setting up the problem. Data availability may also dictate what time units to use to summarize the data and how long the history period and the targeting period should be.

63.6.2 Survival Analysis

Survival Analysis (SA) is concerned with estimating the duration time distribution until an event occurs (called the survival time). Given the probability distribution, one can estimate various measures of the survival time, primarily the expected time or the median time until an event occurs. The roots of survival analysis are in health and life sciences (Cox and Oakes, 1984). Targeting applications include purchasing a new vehicle, applying for a loan, taking a cruise trip, a flight, a vacation. . .

The survival analysis process is demonstrated in Figure 63.5 below. The period from the starting time to the ending time (”today”) is the experimental or the analysis period. As alluded

to earlier, each application may have its own ”typical” analysis period (e.g., several years for the automotive industry). Now, because the time until an event occurs is a random variable, the observations may be left-censored or right-censored. In the former, the observation commences prior to the beginning of the analysis period (e.g., the analysis period for car purchases is three years and the customer purchased her current car more than three years ago); in the latter, the event occurs after the analysis period (e.g., the customer did not purchase a new car within the three-year analysis period). Of course, both types of censoring may occur. For example, a customer that has bought her car prior to the analysis period (left censoring) and replaced it after the end of the analysis period (right censoring).

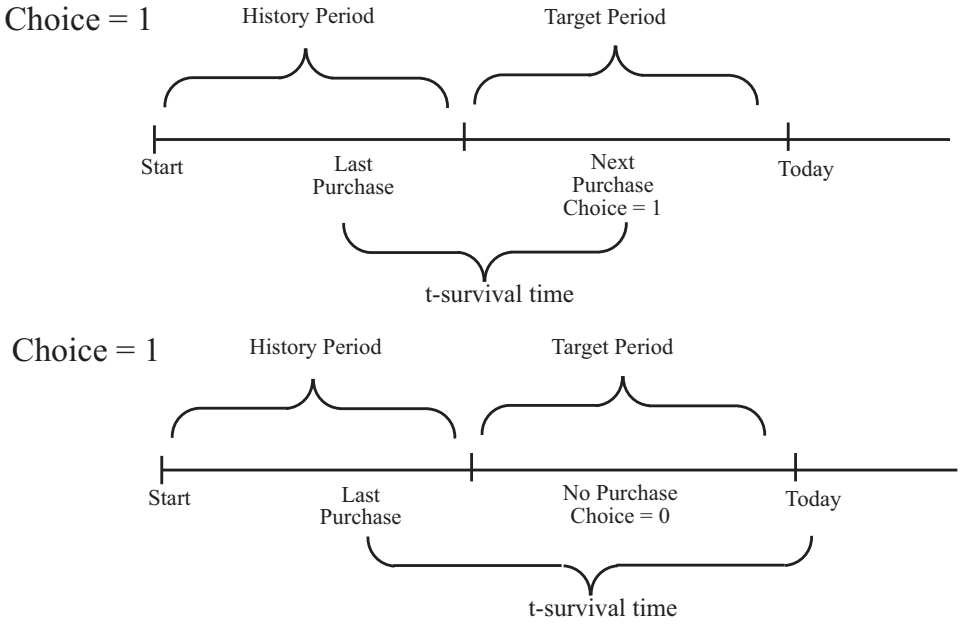

Fig. 63.5. In-Market Timing Using Survival Analysis

As in the logistic regression case, we divide the time axis into two mutually exclusive time periods – the target period, to define the choice variable, and the history period, to define the predictors. But in addition, we also define the survival time, i.e., the time between the last event in the history period and the time until the first event in the target period, as shown in Figure 63.5 (if no event took place in the history period, the survival time commences at the start of the analysis period). Clearly the survival time is a random variable expressed by means of a survival function S(t), which describes the probability that the time until the next event occurs exceed a given time t. The most commonly used distributions to express the survival process are the exponential, the Weibull, the log-logistic and the log-normal distributions. The type of the distribution to use in each occasion depends on the corresponding hazard function, which is defined as the instantaneous probability that the event occurs in an infinitesimally short period of time, given that the event has not occurred earlier. The hazard function is constant for the exponential distribution; It increases or decreases with time, for the other survival

functions, depending upon the parameters of the distribution. For example, in the insurance industry, the exponential distribution is often used to represent the survival time, because the hazard function for filing a claim is likely to be constant as the probability of being involved in an accident is independent of the time that elapses since the preceding accident. In the car industry, for the same make, the hazard function is likely to assume an inverted U-shape function. This is because, right after the customer purchases a new car, the instantaneous probability that s/he buys a new car is almost zero, but it increases with time as the car gets older. Then, if after a while the customer still did not buy a new car, the instantaneous probability goes down, most likely because s/he bought a car from a different manufacturer. Note that in the case any car is involved (not a specific brand), the hazard function is likely to rise with time as the longer one keeps her car, the larger the probability she will replace the car in the next time period. In both cases, the log-logistic distribution could be a reasonable candidate to represent the survival process, with the parameter of the log-logistic distribution determining the shape of the hazard function.

Now, in marketing applications, the survival functions are expressed in terms of a linear function of the customer's attributes (the "utility") and the scaling factor (often denoted by σ). These parameters are estimated based on observations using the method of maximum likelihood.

Given the model, one can estimate the in-market timing probabilities for any new observation for any period Q from "today", using the formula:

$$P\left(t < t_L + Q \middle| t > t_L\right) = 1 - \frac{S\left(t_L + Q\right)}{S\left(t_L\right)}$$

Where:

S(t) – The survival function estimated by the model

t – The time index

t_L – The time since last purchase

We note that the main difference between the logit and the survival analysis models is that prediction based on logit could only be made for a fixed period length (i.e., the period Q above) while in survival analysis Q could be of any length. Also, survival analysis is better "equipped" to handle censored data which is prevalent in time-related applications. This allows the marketer to target customers more accurately by going after them only at the time when their in-market timing probabilities are the highest.

Given the in-market probabilities, either using logistic regression or survival analysis, one may use a judgmentally-based cutoff rate, or a one based on economical considerations, to pick the customers to go after.

63.7 Pitfalls of Targeting

As alluded to earlier, the application of Data Mining to address targeting applications is not all that straightforward and definitely not automatic. Whether by overlooking, ignorance, carelessness, or whatever, it is very easy to abuse the results of Data Mining tools, especially predictive modeling and make wrong decisions. An example which is widely publicized is the 1998 KDD (Knowledge Discovery in Databases) CUP. The KDD-CUP is a Data Mining competition that provides a forum for comparing and evaluating the performance of Data Mining tools on a predefined business problem using real data. The competition in 1998 involved a charity application and the objective was to predict the donation amount for each customer in a

validation sample, based on a model built using an independent training sample. Competitors were evaluated based on the net donation amount obtained by summing up the actual donation amount of all people in the validation set whose expected donation amount exceeded the contact cost ($0.68 per piece). All in all, 21 groups submitted their entry. The results show quite a variation. The first two winners were able to identify a subset of the validation audience to solicit that would increase the net donation by almost 40 percent as compared to mailing to everybody. However, the net donation amount of all other participants lagged far behind the first two. In all, 12 entrants did better than mailing to the whole list, 9 did worse than mailing to the entire list and the last group even lost money on the campaign! The variation in the competition results is indeed astonishing! It tells us that Data Mining is more than just applying modeling software. It is basically a blend of art and science. The scientific part involves applying an appropriate model for the occasion, whether regression model, clustering model, classification model, or whatever. The art part has to do with evaluating of the data that goes into the model and the knowledge that comes out from the modeling process. Our guess is that the dramatic variations in the results of the 1998 KDD-CUP competition is due to the fact that many groups were "trapped" into the mines of Data Mining. So in this section we discuss some of the pitfalls to beware of in building Data Mining models for targeting applications. Some of these are not necessarily pitfalls but issues that one needs to account for in order to render strong models. We divide these pitfalls into 3 main categories – modeling, data and implementation.

63.7.1 Modeling Pitfalls

Misspecified Models

Modern databases often contain tons of information about each customer, which may be translated into hundreds, if not more, of potential predictors. Usually only a handful of which suffices to explain response. The process of selecting the most influential predictors in predictive modeling affecting response from the much larger set of potential predictors is referred to in Data Mining as the feature selection problem. Statisticians refer to this problem as the specification problem. It is a hard combinatorial optimization problem which usually requires heuristic methods to solve, the most common of which is the stepwise regression method (SWR). It is beyond the scope of this chapter to review the feature selection problem in full. So we only demonstrate below the problems that may be introduced to the feature selection problem because of sampling error. For a more comprehensive review of feature selection methods see (Miller, 2002), (George, 2000), and others.

The sheer magnitude of today's databases makes it impossible to build models based on the entire audience. A compromise is to use sampling. The benefits of sampling is that it reduces processing time significantly, but on the other hand it reduces model accuracy by introducing to the model insignificant predictors while eliminating significant ones, both result in misspecified model. We demonstrate this with respect to the linear regression model.

Recalling, in linear regression the objective is to "explain" a continuous dependent variable, Y, in terms of a host of explanatory variables $X_j,\ j = 0, 1, 2, .., J$

$$Y = \sum_{j=0}^{J} \beta_j X_j + U$$

Where:

$\beta_j,\ j = 0, 1, 2, ..., J$ – The coefficients estimated based on real observations

U – A random disturbance

Assuming all other factors are equal, one can check whether a variable, say X_k, is significant by testing the hypothesis.

$$H_0: \quad \beta_k = 0$$
$$H_1: \quad \beta_k \neq 0$$

The test statistics for testing the hypothesis is given by:

$$t = \left| \hat{\beta}_k \Big/ s(\hat{\beta}_k) \right|$$

Where:

$\hat{\beta}_k$ – the coefficient estimate of β_k

$s(\hat{\beta}_k)$ – the standard error of the coefficient estimate

In small samples, the test statistics t is distributed as the t (student) distribution with $n - J - 1$ degrees of freedom. In Data Mining applications, where the sample size is very large, often containing as many as several hundred observations, or more, the t-distribution may be approximated by the normal distribution.

Given the test statistics and its sampling distribution, one calculates the minimum probability level to reject H_0 where it is true, $P - value$:

$$P - Value = 2P(T > |\hat{\beta}_k / s(\hat{\beta}_k))$$

And if the resulting $P - value$ is smaller than, or equal, to a predefined level of significance, often denoted by α, one rejects H_0; otherwise, one does not reject H_0.

The level of significance α is the upper bound on the probability of Type-I error (rejecting H_0 when true). It is the proportion of times that we reject H_0 when true, out of all possible samples of size n drawn from the population. In fact, the P-value is just one realization of this phenomenon. It is the actual Type-I error probability for the given sample statistics.

Now, suppose that X_k is an insignificant variable having no relation whatsoever to the dependent variable Y (i.e., the correlation coefficient between X_k and Y is zero). Then, if we build the regression model based on a sample of observations, there is a probability of α that X_k will turn out significant just by pure chance, thus making it into the model and resulting in Type-I error, in contradiction to the fact that X_k *and* Y are not correlated.

Extending the analysis to the case of multiple insignificant predictors, even a small Type-I error may result in several of those variables making it into the model as significant. Taking this to the extreme case where all predictors involved are insignificant, we are almost sure to find a significant model, indicating a true relationship between the dependent variable (e.g., response) and some of the regressors, where such a relationship does not exist! This phenomenon also extends for the more realistic case which involves both significant and insignificant predictors.

The converse is also true, i.e., that there is a fairly large probability for significant predictors in a population to come out insignificant in a sample, and thus wrongly excluded from the model (Type-II error).

In either case, the resulting predictive model is misspecified. In the context of targeting decisions in database marketing, a misspecified model may result in either some profitable people being excluded from the mailing (Type-I error) and some unprofitable people being included in the mailing (Type-II error), both incur some costs: Type-I error – forgone profits due to missing out good people from the mailing as well as lost of reputation; Type-II error – real losses for contacting unprofitable people.

Clearly, one cannot avoid Type-I and Type-II errors altogether, unless the model is built off the entire database, which is not feasible. But one can reduce the error probabilities by several means - controlling the sample size, controlling the Type-I and Type-II errors using Bonferonni coefficients, False Discovery Rates (FDR) (Benjamini and Hochberg, 1995), Akaike Information Criterion (AIC) (Akaike, 1973), Bayesian Information Criterion (BIC) (Schwarz, 1978), and others.

Detecting misspecified models is an essential component of the knowledge discovery process, because applying a wrong model to target audiences for promotion may incur substantial losses. This is why it is important that one validates the model on an independent data set, so that if a model is wrongly specified, this will show up in the validation results.

Over-Fitting

Over-fitting pertains to case where the model gives good results when applied on the data used to build the model, but yields poor results when applied against a set of new observations. An extreme case of overfitting is when the model is doing too good a job in discriminating between the buyers and the non buyers (e.g., "capturing" all the buyers in the first top percentiles of the audience ("too good to be true")). In either case, the model is not valid and definitely can not be used to support targeting decisions. Over-fitting is a problem that plagues large-scale predictive model, often as a result of a misspecified model, introducing insignificant predictors to a regression model (Type-I error) or eliminating significant predictors from a model (Type-II error).

To test for over-fitting, it is necessary to validate the model using a different set of observations than those used to build the model. The simplest way is to set aside a portion of the observations for building the model (the training set) and hold out the balance to validate the model (the holdout, or the validation data). After building the model based on the training set, the model is used to predict the value of the dependent variable (e.g., purchase probabilities) in predictive model or the class label in classification models, of the validation audience. Then, if the scores obtained for the training and validation data sets are more-or-less compatible, the model appears to be OK (no over-fitting). The best way to check for the compatibility is to summarize the scores in gains table at some percentile level and then compare the actual results between the two tables at each audience level. The more sophisticated validation involves n-fold cross-validation.

Over-fitting results when there is too little information to build the model upon. For example, there are too many predictors to estimate and only relatively few responders in the test data. The cure for this problem is to reduce the number of predictors in the model (parsimonious model). Recent research focuses on combining estimators from several models to decrease variability in predictions and yield more stable results. The leading approaches are bagging (Breiman, 1996) and boosting (Friedman *et al*., 1998).

Under-Fitting

Under-fitting is the counterpart of over-fitting. Under-fitting refers to a wrong model that is not fulfilling its mission. For example, in direct marketing applications, under-fitting results when the model is not capable of distinguishing well between the likely respondents and the likely non-respondents. A fluctuation of the response rate across a gains table may be an indication of a poor fit, or too small of a difference between the top and the bottom deciles. Reasons for under-fitting could vary: wrong model, wrong transformations, missing out the influential

predictors in the feature selection process, and others. There is no clear prescription to resolve the under-fitting issue. Some possibilities are: trying different models, partitioning the audience into several key segments and building a separate model for each, enriching data, adding interaction terms, appending additional data from outside sources (e.g., demographic data, lifestyle indicators), using larger samples to build the model, introducing new transformations, and others. The process may require some creativity and ingenuity.

Non-Linearity/ Non-Monotonic Relationships

Regression-based models are linear-in-parameters models. In linear regression, the response is linearly related to the attributes; in logistic regression model, the utility is linearly related to the attributes. But more often than not, the relationship between the output variable and the attribute is not linear. In this case one needs to specify the non-linear relationships using a transformation of the attribute. A common transformation is a polynomial transformation of the form $y = x^a$ where $-2 < a < 2$. Depending upon the value of a, this function provide a variety of ways to express non-linear relationships between the input variable x and the output variable y. For example, if a<1, the transformation has the effect of moderating the impact of x on the choice variable. Conversely, if a>1, the transformation has the effect of magnifying the impact of x on the choice variable. For the special case of a=0, the transformation is defined as y=log(x). The disadvantage of the power transformation above is that it requires that the type of the non-linear relationship be defined in advance. A more preferable approach is to define the non-linear relationships based on the data. Candidate transformations of this type are the step function or the piecewise linear transformation. In step function, the attribute range is partitioned into several mutually exclusive and exhaustive intervals (say, by quartiles). Each interval is then represented by means of a categorical variable, assuming the value of 1 if the attribute value falls in the interval, 0 – otherwise. A piecewise transformation splits a variable into several non-overlapping and continuously-linked linear segments, each with a given slope. Then, the coefficient estimates of the categorical variables in the step function, and the estimate of the slopes of the linear segments in the piecewise function, actually determine the type of relationships that exist between the input and the output variables.

Variable Transformations

More often than not, the intrinsic prediction power resides not in the original variables themselves but on transformations of these variables. There are basically infinite number of ways to define transformations, and the ”sky is the limit”. We mention here only proportions and ratios, which are very powerful transformations in regression-based models. For example, the response rate, defined as the ratio of the number of responses to the number of promotions, is considered to be a more powerful predictor of response than either the number of responses or the number of promotions. Proportions are also used to scale variables. For example, instead of using the dollar amount in a given time segment as a predictor, one may use the proportion of the amount of money spent in the time segment relative to the total amount of money spent. Proportions possess the advantage of having a common reference point which makes them comparable. For example, in marketing applications it is more meaningful to compare the response rates of two people rather than their number of purchases, because the number of purchases does not make sense unless related to the number of opportunities (contacts) the customer has had to respond to the solicitation.

Space is too short to review the range of possible transformations to build a model. Suffice it to say that one needs to pay a serious consideration to defining transformations to obtain a good model. Using domain knowledge could be very helpful in defining the ”right” transformation.

Choice-Based Sampling

Targeting applications are characterized by very low response rates, often less than 1%. As a result, one may have to draw a larger proportion of buyers than their proportion in the population, in order to build a significant model. It is not uncommon in targeting applications to draw a stratified sample for building a model which includes all of the buyers in the test audience and a sample of the non buyers. These types of samples are referred to as choice-based sample (Ben Akiva and Lerman, 1987). But choice-based samples yield results which are compatible with the sample, not the population. For example, a logistic regression model based on a choice-based sample that contains higher proportion of buyers than in the population will yield inflated probabilities of purchase. Consequently, one needs to update the purchase probabilities in the final stage of the analysis to reflect the true proportion of buyers and non buyers in the population in order to make the right selection decision. For discrete choice models, this can be done rather easily by simply updating the intercept of the regression equation (Ben Akiva and Lerman, 1987). In other models, this may be more complicated.

Observations Weights

Sampling may apply not just to the dependent variable but also to the independent variable. For example, one may select for the test audience only 50% of the female and 25% of the males. However, unlike the choice-based sampling which does not affect the model, proportion-based sampling affect the modeling results (e.g., the regression coefficients). To correct for this bias, one needs to inflate the number of males by a factor of 2 and the number of females by a factor of 4 to reflect their ”true” numbers in the population. We refer to these factors as observations weights.

Of course, a combination of choice-based sampling and proportional sampling may also exist. For example, suppose we first create a universe which contains 50% of the females and 25% of the males and then pick all of the buyers and 10% of the non buyers for building the model. In this case, each female buyer represents 2 customers in the population whereas each female non-buyer represents 20 customers in the population. Likewise, each male buyer represents 4 customers in the population whereas each male non-buyer represents 40 customers in the population. Clearly, one needs to account for these proportions to yield unbiased targeting models.

63.7.2 Data Pitfalls

Data Bias

By data bias we mean that not all observations in the database have the same items of data, with certain segments of the population are having the full data whereas other segments containing only partial data. For example, new entrants usually contain only demographics information but no purchase history, automotive customers may have purchase history information only for

the so-called unrestricted states and only demographic variables for the restricted sates, survey data may be available only for buyers and not for non buyers, some outfits may introduce certain type of data, say prices, only for buyers and not for non-buyers, etc. If not taken care of, this can distort the model results. For example, using data available only for buyers but not for non buyers, say the price, may yield ”perfect” model in the sense that price is the perfect predictor of response, which is of course not true. Building one model for ”old” customers and new entrants may underestimate the effect of certain predictors on response, while over estimating the effect of others. So one need to exercise caution in these cases, perhaps build a different model for each type of data, use adjustment factors to correct for the biased data, etc.

Missing values

Missing data is very typical of large realistic data sets. But unlike in the previous case, where the missing information was confined to certain segments of the population, in this case missing value could be everywhere, with some attributes having only a few observations with missing values with others having a large proportion of observations with missing values. Unless accounted for, missing values could definitely affect the model results. There's a trade off here. Dropping attributes with missing data from the modeling process results in loss of information; but including attributes with missing data in the modeling process may distort the model results. The compromise is to discard attributes for which the proportion of observations with missing value for that attribute exceeds a pre defined threshold level.

As to the others, one can ”capture” the effect of missing value by defining an additional predictor for each attribute which will be ”flagged” for each observation with a missing value, or impute a value for missing data. The value to impute depends on the type of the attribute involved. For interval and ratio variables, candidate values to impute are the mean value, the median value, the maximum value or the minimum value of the attribute across all observations; for ordinal variables – the median of the attribute is the likely candidate; and for nominal variables - the mode of the attribute. More sophisticated approaches to dealing with missing value exist, e.g., for numerical variables, imputing a value obtained by means of a regression model.

Outliers

Outliers are the other extreme of missing value. We define an outlier as an attribute value which is several standards deviations away from the mean value of the observations. As in the case of a missing value, there's also a tradeoff here. Dropping observations with outlier attributes may result in a loss of information, while including them in the modeling process may distort the modeling results. A reasonable compromise is to trim outlier value from above by setting the value of an outlier attribute at the mean value of the attribute plus a pre-defined number of standard deviations (say 5), and trim an outlier value from below by setting the value of an outlier at the mean value minus a certain number of standard deviations.

Noisy Data

We define by noisy data binary attributes which appear with very low frequency, e.g., the proportion of observations in the database having a value of 1 for the attribute is less than a small threshold level of the audience, say 0.5%. The mirror image are attributes for which the proportion of observations having a value of 1 for the attribute exceeds a large threshold level,

say 99.5%. These types of attributes are not strong enough to be used as predictors of response and should either be eliminated from the model, or combined with related binary predictors (e.g., all the Caribbean islands may be combined into one predictor for model building, thereby mitigating the effect of noisy data).

Confounded Dependent Variables

By a confounded dependent variable we mean a dependent variable which is "contaminated" by one or more of the independent variables. This is quite a common mistake in building predictive models. For example, in a binary choice application the value of the current purchase in a test mailing is included in the predictor Money_Spent. Then, when one uses the test mailing to build a response model, the variable Money_Spent fully explains customer's choice, yielding a model which is "too good to be true". This is definitely wrong. The way to avoid this type of errors is to keep the dependent variable clean of any effect of the independent variables.

Incomplete Data

Data is never complete. Yet, one needs to make best use of the data, introducing adjustment and modification factors, as necessary, to compensate for the lack of data. Take for example the in-market timing problem in the automotive industry. Suppose we are interested in estimating the mean time or the median time until the next car replacement for any vehicle. But often, the data available for such an analysis contain, in the best case, only the purchase history for a given OEM (Original Equipment Manufacturer) which allows one to predict only the replacement time of an OEM vehicle. This time is likely to be much longer than the replacement time of any vehicle. One may therefore have to adjust the estimates to attain time estimates which are more compatible with the industry standards.

63.7.3 Implementation Pitfalls

Selection Bias

By selection bias we mean samples which are not randomly selected. In predictive modeling this type of sample is likely to render biased coefficient estimates. This situation may arise in several cases. We consider here the case of subsequent promotions with the "funnel effect", also referred to as rerolls. In this targeting application, the audience for each subsequent promotion is selected based on the results of the previous promotion in a kind of a "chain" mode. In the first time around, the chain is usually initiated by conducting a live market test to build a response model (as in Figure 63.1), involving a random sample of customers from the universe. The predictive model based on the test results is then used to select the audience for the first rollout campaign (the first-pass mailing). The reroll campaign (the second-pass mailing) is then selected using a response model which is calibrated based on the rollout campaign. But we note that the rollout audience was selected based on a response model and it is therefore not a random sample of the universe. This gives rise to a selection bias. Similarly, the second reroll (the third-pass campaign) is selected based on a response model built based upon the reroll audience, the third reroll is based on the second reroll, and so on. .

Now, consider the plausible purchase situation where once a customer purchases a product, h/se is not likely to purchase it again in the near future. Certainly, it makes no sense to

approach these customers in the next campaign and they are usually removed from the universe for the next solicitation. In this case, the rollout audience, the first campaign in the sequence of campaigns, consists only of people who were never exposed to the product before. But moving on to the next campaign, the reroll, the audience here consists of both exposed and unexposed people.

The exposed people are people who were approached in the roll campaign, declined the product, but are promoted again in the reroll because they still meet the promotability criteria (e.g., they belong to the ”right” segment)

The unexposed people are people contacted in the reroll for the first time. They consist of two types of people:

- New entrants to the database who have joined the list in the time period between the first rollout campaign and the reroll campaign.
- ”Older” people who were not eligible for the rollout campaign, but have ”graduated” since then and now meet the promotability criteria for the reroll campaign (e.g., people who have bought a product from the company in the time gap between the rollout and the reroll campaigns, and have thus been elevated into a status of ”recent buyers” which qualifies them to take part in the reroll promotion).

Hence the reroll audience is not compatible with the rollout audience, i.e., it contains ”different” type of people. The question then is how one can adjust the purchase probabilities of the exposed people in the reroll given that the model is calibrated based on the rollout audience which contains unexposed people only?

Now, going one step further, the second reroll audience is selected based on the results of the first reroll audience. But the first reroll audience consists only of unexposed and first-time exposed people, whereas the second reroll audience also contains twice-exposed people. The question, again, is how to adjust the probabilities of second-time exposures given the probabilities of the first-time exposures and the probabilities of the unexposed people? The problem extends in this way to all subsequent rerolls.

Empirical evidence show that the response rate of repeated campaigns for the same product drops down with each additional promotion. This decline in response is often referred to as the ”list dropoff” phenomenon (Buchannan and Morrison 1988). The list falloff rate is not consistent across subsequent solicitations. It is usually the largest, as high as 50% or more, when going from the first rollout to the reroll campaigns and then more-or-less stabilizes at a more moderate level, often 20%, with each additional solicitation. Clearly, with the response rate of the list going down from one solicitation to the other, there comes a point where it is not worth promoting the list, or certain segments of the list, any more, because the response rate becomes too small to yield any meaningful expected net profits. Thus, it is very important to accurately model the list falloff phenomenon to ensure that the right people are promoted in any campaign, whether the first one or a subsequent one.

Regression to the Mean (RTM) Effect

Another type of selection bias, which applies primarily to segmentation-based models, is the regression to the mean (RTM) phenomenon. Recall that in the segmentation approach, either the entire segment is rolled out or the entire segment is excluded from the campaign. The RTM effect arises because only the segments that performed well in the test campaign, i.e., the ”winners” are recommended for the roll. Now because of the random nature of the process, it is likely that several of the ”good” segments that performed well in the test happened to do

so just because of pure chance; as a result, when the remainder of the segment is promoted, its response rate drops back to the "true" response rate of the segment, which is lower than the response rate observed in the test mailing. Conversely, it is possible that some of the segments that performed poorly in the test campaign happened to do so also because of pure chance; as a result, if the remainder of the segment is rolled out, it is likely to perform above the test average response rate. These effects are commonly referred to as RTM (Shepard, 1995). When both the "good" and "bad" segments are rolled out, the over and under effects of RTM cancels out and the overall response rate in the rollout audience should be more-or-less equal to the response rate of the testing audience. But since only the "good" segments, or the "winners", are promoted, one usually witnesses a dropoff in the roll response rate as compared to the test response rate.

Since the RTM effect is not known in advance for any segment, one needs to estimate this effect based on the test results for better targeting decisions. This is a complicated problem because the RTM effect for any segment depends on the "true" response rate of the segment, which is not known in advance. Levin and Zahavi (1996) offer an approach to estimate the RTM effect for each segment which uses a prior knowledge on the "quality" of the segment (either "good", "medium" or "bad"). Practitioners use a knock down factor (often 20%-25%) to project the rollout response rate. While the latter is a crude approximation to the RTM effect, it is better than using no correction at all, as failure to account for the RTM may results in some "good" segments eliminated from the rollout campaign and some "bad" segments included in the campaign, both incur substantial costs.

As-of-Date

Because of the lead time to stock up on product, the promotion process could extend over time, with the time gap between the testing and the rollout campaign could extend over several months, sometime a year (see Figure 63.1). In case of subsequent rerolls, the time period between any two consecutive rerolls may be even longer. This introduces a time dimension into the modeling process.

Now, most predictors of response also have a time dimension. Certainly, this applies to the RFM variables which have proven to be the most important predictors of response in numerous applications. This goes without saying for recency which is a direct measure of time since last purchase. But frequency and monetary variables are also linked to time, because they often measure the number of previous purchases (frequency) and money spent (monetary) for a given time period, say a year. We note that some demographic variables such as age, number of children, etc., also change over time.

As a result, all data files for supporting targeting decisions ought to be created as of the date of the promotion. So if testing took place on January 1st, 2003 and the rollout campaign on July 30th, 2003, one needs to create a snap shot of the test audience as of January 1, 2003, for building the model and another snap shot of the universe as of July 30, 2003, for scoring the audience.

We note that if the time gap between two successive promotions (say the test and the rollout campaigns) is very long, several models may be needed to support a promotion. One model to predict the expected number of orders to be generated by the rollout campaign, based on the test audience reflecting customers' data as of the time of the test (January 1 2003, in the above example). Then, at the time of the roll, when one applies the model results for selecting customers for the rollout campaign, it might be necessary to recalibrate the model based on a snap shot of the test audience as of the rollout date (July 30th, 2003, in the above example).

63.8 Conclusions

In this chapter we have discussed the application of Data Mining models to support targeting decisions in direct marketing. We distinguished between three targeting categories – discrete choice problems, continuous choice problems and in-market timing problems, and reviewed a range of models for addressing each of these categories. We also discussed some pitfalls and issues that need to be taken care of in implementing a Data Mining solution for targeting applications.

But we note that the discussion in this chapter is somewhat simplified as it is confined mainly to targeting problem where each product/service is promoted on its own, by means of a single channel (mostly mail), independently of other products/services. But clearly, targeting problems can be much more complicated than that. We discuss below two extensions to the basic problem above – multiple offers and multiple products.

63.8.1 Multiple Offers

An ”offer” is generalized here to include any combination of the marketing mix attributes, including price point, position, package, payment terms, incentive levels... For example, in the credit card industry, the two dominant offers are the line of credit to grant to a customer and the interest rate. In the collectible industry, the leading offers are price points, positioning of the product (i.e., as a gift or for own use), packaging,...

Incentive offers are gaining increasing popularity as more and more companies recognize the need to incorporate an incentive management program into the promotion campaigns to maximize customers' value chain. Clearly it does not make sense to offer any incentive to customers who are ”captive audience” who are going to purchase the product no matter what. But it does make sense to offer an incentive to borderline customers ”on the fence” for whom the incentive can make the difference between purchasing the product/service or declining it. This is true for each offer, not just for incentives. In general, the objective is to find the best offer to each customer to maximize expected net benefits. This gives rise to a very large constrained optimization problem containing hundreds of thousands, perhaps millions, of rows (each row corresponds to a customer) and multiple columns, one for each offer combination. The optimization problem may be hard to solve analytically, if any, and a resort to heuristic methods may be required.

From a Data Mining perspective, one needs to estimate the effect of each offer combination on the purchase probabilities, which typically requires that one designs an experiment whereby customers are randomly split into groups, each exposed to one offer combination. Then, based on the response results, one may estimate the offer effect. But, because the response rates in the direct marketing industry is very low, it is often necessary to test only part of the offer combinations (partial factorial design) and then deduct from the partial experiment onto the full factorial experiment. Further complication arises when optimizing the test design to maximize the information content of the test, using feedback from previous tests.

63.8.2 Multiple Products/Services

The case of multiple products adds another dimension of complexity to the targeting problem. Not only it is required to find the best offer for a given product to each customer, but it is also necessary to optimize the promotion stream to each customer over time, controlling the timing, number and mix of promotions to expose to each individual customer at each time

window. This gives rise to even a bigger optimization problem which now contains many more columns, one column for each product/offer combination.

From a modeling standpoint, this requires that one estimate the cannibalization and saturation effects. The cannibalization effect is defined as the rate of the reduction in the purchase probability of the product as a result of over-promotion. Because of the RFM effect discussed above, it so happens that the "good" customers are often bombarded with too many mailings at any given time window. One of the well known effects of over-promotion is that it turns down customers, resulting in a decline in their likelihood of purchase of either product promoted to them. Experience shows that too many promotions may cause customers to discard the promotional material without even looking at them. The end result is often a loss in the number of active customers, not to mention the fact that over promotion results in misallocation of the promotion budget.

While the cannibalization effect is a result of over-promotion, the saturation effect is the result of over-purchase. Clearly, the more a customer buys from a given product category, the less likely s/he is to respond to a future solicitation for a product from the same product category. From a modeling perspective, the saturation effect is defined as the rate of reduction in the purchase probability of a product as a function of the number of products in the same product line that the customer has bought in the past. Since the saturation effect is not known in advance, it must be estimated based on past observations.

And these are not the only issues involved, and there are a myriad of others. Clearly, targeting applications in marketing are at the top of the analytical hierarchy, requiring a combination of tools from Data Mining, operations research, design of experiments, direct and database marketing, database technologies, and others. And we have not discussed here the organizational aspects involved in implementing a targeting system, and the integration with other operational units of the organization, such as inventory, logistics, financial, and others.

References

Akaike, H., Information Theory and an Extension of the Maximum Likelihood Principle, in 2^{nd} International Symposium on Information Theory, B.N. Petrov and F. Csaki, eds, pp. 267-281, Budapest, 1973.

Ben-Akiva, M., and S.R. Lerman, Discrete Choice Analysis, the MIT Press, Cambridge, MA, 1987.

Benjamini, Y. and Hochberg, Y., Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing, Journal Royal Statistical Society, Ser. B, 57, pp. 289-300, 1995.

Bock, H.H. Automatic Classification. Vandenhoeck and Ruprecht, Gottingen, 1974.

Breiman, L., Bagging Predictors, Machine Learning, Vol. 2, pp. 123-140, 1996.

Breiman, L., Friedman, J., Olshen, R. and Stone, C., Classification and Regression Trees, Belmont, CA., Wadsworth, 1984.

Buchanan, B. and Morrison, D.G., A Stochastic Model of List Falloff with Implications for Repeated Mailings", The Journal of Direct Marketing, Summer, 1988.

Cox, D.R. and Oakes, D., Analysis of Survival Data, Chapman and Hall, London, 1984.

DeGroot, M. H., Probability and Statistics 3^{rd} edition. *Addison-Wesley*, 1991.

Friedman, J., Hastie, T. and Tibshirani, R., Additive Logistic Regression: a Statistical View of Boosting, Technical Report, Department of Statistics, Stanford University, 1998.

Fukunaga, K., Introduction to Statistical Pattern Recognition. San Diego, CA: Academic Press, 1990.

Heckman, J., Sample Selection Bias as a Specification Error, Econometrica, Vol. 47, No. 1, pp. 153-161, 1979.

Gilbert A. and Churchill, Jr., Marketing Research. *Seventh edition The Dryden Press*, 1999.

George, E.I., The Variable Selection Problem, University of Texas, Austin, 2000.

Herz, F., Ungar, L. and Labys, P., A Collaborative Filtering System for the Analysis of Consumer Data. Univ. of Pennsylvania, Philadelphia, 1997.

Hodges, J.L. Jr., "The Significance Probability of the Smirnov Two-Sample Test," Arkiv for Matematik, 3, 469 -486, 1957.

Kass, G., An Exploratory Technique for Investigating large Quantities of Categorical Data, Applied Statistics, 29, 1983.

Kohonen, K., Makisara, K., Simula, O. and Kangas, J., Artificial Networks. Amsterdam, 1991.

Lauritzen, S.L., The EM algorithm for Graphical Association Models with Missing Data. Computational Statistics and Data Analysis, 19, 191-201, 1995.

Long, S.J., Regression Models for Categorical and Limited Dependent Variables, Sage Publications, Thousand Oaks, CA, 1997.

Lambert P.J., The Distribution and Redistribution of Income. *Manchester University Press*., 1993.

Levin, N. and Zahavi, J., Segmentation Analysis with Managerial Judgment, Journal of Direct Marketing, Vol. 10, pp. 28-47, 1996.

Levin, N. and Zahavi, J., Applying Neural Computing to Target Marketing, The Journal of Direct Marketing, Vol. 11, No. 1, pp. 5-22, 1997a.

Levin, N. and Zahavi, J., Issues and Problems in Applying Neural Computing to Target Marketing, The Journal of Direct marketing, Vol. 11, No. 4, pp. 63-75, 1997b.

Miller, A., Subset Selection in Regression, Chapman and Hall, London, 2002.

Quinlan, J.R., Induction of Decision Trees, Machine Learning, 1, pp. 81-106, 1986.

Quinlan, J.R., C4.5: Program for Machine Learning, CA., Morgan Kaufman Publishing, 1993.

Rumelhart, D.E., McClelland, J.L., and Williams, R.J., Learning Internal Representation by Error Propagation, in Parallel Distributed Processing: Exploring the Microstructure of Cognition, Rumelhart, D.E., McClelland, J.L. and the PDP Researcg Group, eds., MIT Press, Cambridge, MA, 1986.

Schwarz, G., Estimating the Dimension of a Model, Annals of Statistics, Vol. 6, pp. 486-494, 1978.

Shepard, D. (ed.), The New Direct Marketing, New York, Irwin, 1995.

Silverman, B.W., Density Estimation for Statistics and Data Analysis. Chapman and Hall, 1986.

Smith, W.R., Product Differentiation and Market Segmentation as Alternative Marketing Strategies, Journal of Marketing, 21, 3-8, 1956.

Sonquist, J., Baker, E. and Morgan, J.N., Searching for Structure, Ann Arbor, University of Michigan, Survey Research Center, 1971.

Tobin, J., Estimation of Relationships for Limited-Dependent Variables, Econometrica, Vol. 26, pp. 24-36, 1958.

Zhang, R., Ramakrishnan, R. and Livny, M., An Efficient Data Clustering Method for Very Large Databases. Proceedings ACM SIGKDD International Conference on Management of Data. 103-114, 1996.