Quantitative approaches for profit maximization in direct marketing
van der Scheer, H.R.

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
1998

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):
Chapter 2

Target selection techniques

2.1 Introduction

Target selection, also called list segmentation, can be seen as the process of dividing the market, i.e. the mailing list, into two distinct groups, viz. a group that should receive a mailing and a group that should not receive a mailing. It is obvious that target selection is a crucial component for the development of a direct mailing campaign since a campaign can only be effective if the mailing reaches the proper targets. Therefore, direct marketers have expended considerable effort towards list segmentation techniques. By eliminating the least promising targets of the mailing list, the direct marketing organization can increase its profits.

Generally speaking, in order to achieve an effective segmentation for a promotional campaign, the distinct segments or groups must satisfy at least the following criteria (cf. Leeffang 1994, p. 41, and Wedel 1990, p. 17): first, the size and composition of the selected group must be known. Secondly, it must be possible to reach the selected group effectively in a communicative manner. Thirdly, the selected group should be economically large enough for the development of a promotional campaign. Usually, these conditions are met in direct marketing when a mailing list is available. The size is simply measured by counting the number of targets and the composition is determined by the variables available on the list. Since a mailing list contains at least names and addresses, each individual can be reached by mail. The last criterion is met when the number of individuals in the group that should receive a mailing is sufficiently large. This number implicitly depends on the revenues to a positive reply, the costs of a single mailing piece and the fixed costs. Hence, a mailing
list offers an organization a perfect possibility for effective segmentation, which coincides with target selection here.

The objective of this chapter is to review the selection techniques that have either been proposed in the literature or have commonly been employed by direct marketers. There are at least two aspects that have a large impact on the target selection process, viz. the objective of the direct marketing campaign and the (availability of) selection variables. Since it is important to keep these two aspects in mind, we start in section 2.2 and 2.3, respectively, with a brief discussion on these aspects. Then, in section 2.4, we present the selection techniques. In section 2.5 we discuss the various criteria on which the models can be compared and briefly discuss the literature that compares some of these techniques. Section 2.6 states our conclusion.

2.2 Objective of the direct marketing campaign

Each direct marketing campaign should, of course, have a specific objective. We distinguish three classes of objectives for the consumer market (e.g. Roberts and Berger 1989, p. 9):

1. The accomplishment of the sale of a product or service. This is the most frequently used objective.
2. The generation of leads, denoting the request of an individual for additional information about the subject. In the next stage, an organization may try to convert these leads into sales. Expensive or complex products, like insurance policies, are often sold by this two-stage process. For a successful campaign the organization should not only focus on generating leads but also in the conversions of these leads.
3. The maintenance of customer relationships. The mailing under consideration does not have the purpose to sell the product or to generate leads, but to strengthen the relation (e.g. a newsletter of an insurance company which is sent to all the policy holders).

When the target selection strategy is directed to profit maximization the first two objectives, roughly speaking, focus on short-run profit maximization, whereas the latter focuses on long-term profit maximization. This could imply that an organization aims at maximizing the lifetime value of the individuals, which can be described as “the net present value of future contributions to profit expected from the individual”. Different direct marketing objectives, in particular the difference between short-run and long-run profit maximization,
lead to different selections. However, the critical issue in each case is to select those individuals whose expected revenues are sufficiently high to offset the costs. The difference in selection is caused by the way the costs and expected revenues are determined. Note that, when a profit maximization approach is used, the number of individuals that will receive a mailing is determined by a cost-benefit analysis and not by a fixed budget figure. Sometimes, however, direct marketing campaigns are based on a fixed budget. In that case the same selection techniques can be used but the number of individuals that will receive a mailing is specified a priori by the size of the budget.

2.3 Selection variables

Selection variables, often called segmentation variables in the direct marketing literature, are of course a crucial element for target selection. If the available variables do not have any explanatory or discriminatory power to distinguish respondents from nonrespondents, segmentation is not very useful. Four basic categories of selection variables can be distinguished (Roberts and Berger 1989, p. 106): (1) geographic variables, (2) demographic variables, (3) psychographic or lifestyle variables and (4) behavioral variables. Within direct marketing the behavioral variables, in particular variables related to the individuals’ purchase history, are considered the most important. They include the recency of the last purchase, the frequency of purchases and the monetary amount spent on the purchases (e.g. Roberts and Berger 1989, p. 106); the so-called RFM-variables. Recency includes the number of consecutive mailings without response and the time period since the last order. Frequency measures include the number of purchases made in a certain time period. Monetary value measures include the amount of money spent during a certain period. It is clear that this kind of information solely relates to individuals previously mailed. A database with this type of customer information is called the internal list (e.g. Bickert 1992). This list often also includes individuals who have never bought a product but only made an inquiry or to whom a mailing has been directed.

Other information on the individuals is often available on external lists. These lists contain geographical variables (region, degree of urbanization), demographic variables (age, sex, family size) and lifestyle characteristics (habits, leisure interests). External lists may be put together for direct marketing support or for other reasons (e.g. telephone directories). It is important to realize that external lists can be used as additional information to the internal list but
may also be the only source of information. The latter is often the case when the mailing is directed to acquire new customers. External lists can be rented from commercial companies. In the United States these lists are offered by, among others, R.L. Polk, National Demographics and Lifestyles and Database America. In the Netherlands lists can be rented from Geo-Markt profiel, Mosaic and recently from Calyx (in cooperation with Centrum voor Consumenten Informatie). A comprehensive discussion on the use of commercial databases can be found in Lix et al. (1995).

Often information can be rented either at an individual level or at a postal code level. The latter means that the average of a group of individuals is known. Consequently, the selection takes place at the aggregate level rather than at the individual level. As a result, the maximally attainable profit decreases. In the literature it is shown (e.g. Lix et al. 1995, Roberts and Berger 1989, pp. 99-103) that an organization should make a trade-off between the cost of renting additional information (e.g. an external list or information at the individual level instead of at the postal code level) and a more effective selection. However, to the best of our knowledge, there is not a single paper that discusses the analytical and statistical aspects involved. In chapter 5 we introduce a comprehensive framework for analyzing the value of information.

Before we turn to the discussion of the selection techniques, we would like to stress that an important aspect of modeling, whichever technique is used, consists of processing, auditing and understanding the data. This includes the selection of variables from a database and transformation of variables. An example of the former is whether a step-wise procedure should be used to select the explanatory variables or a so-called forced entry, i.e. including all the variables (chosen a priori). Transformation of variables is employed, for instance, in the following two situations: (1) recode a categorical variable that has (too) many categories; (2) incorporate a continuously measured variable in a flexible, i.e. nonlinear, way. A straightforward way to do so is by dividing this variable into a number of categories and using a dummy variable for each category; this is often called binning (Lix and Berger 1995).

### 2.4 Selection techniques

Generally, selection takes place on the basis of some measure of response. For direct mail, three kinds of responses can be distinguished, depending on the offer submitted in the mailing. The first kind concerns mailings with fixed
Selection techniques

revenues (given a positive reply), such as subscriber mailings of a magazine, membership mailings and single-shot mailings offering just one product, for example a book. A second kind concerns mailings where the response is the number of units ordered, e.g. the number of compact discs ordered by direct mail selling or the subscription time (a quarter of a year, half a year, a full year) when a magazine is offered through direct mail. Third, there are mailings with a response that can take on all positive values. This may involve total revenues in purchases from a catalog retailer, or the monetary amount donated to a charitable foundation raising funds by mail.

Nearly all the selection techniques that have been proposed deal with the case of fixed revenues to a positive reply. Although it is recognized in most of the papers published that many direct mail campaigns generate a response of which the revenues vary across individuals, it is hard to find publications that take this aspect into account. So, in the case that the revenues vary across individuals one should use the average of the individuals who responded as the fixed revenues. In chapter 3 we present several methods to incorporate the quantity of response into the selection process.

Before we turn to the discussion of the selection techniques, we want to describe an important characteristic of a technique, namely its parameterization. Since we will regularly refer to this characteristic, we introduce the parameterization issue here. The intention is to give the basic idea of the possible parameterizations rather than a thorough discussion. Three approaches can be distinguished: parametric, semiparametric and nonparametric.

In a parametric approach, it is typically assumed that the dependent variable functionally depends on the independent variables and an unobservable error, in accordance with a fixed relation. That is, the functional form between the dependent variable and the independent variables is known as well as the functional form of the distribution of the error term. The error term is introduced to account for, among other things, the lack of a perfect fit of this relation, omitted variables, or badly measured variables (i.e. measurement error). An example of a parametric specification is the classical linear regression model. It assumes that the relation between the dependent variable and independent variables is linear and that the errors are independent and identically normally distributed. The main advantage of the parametric approach is its ease of interpretation and implementation.

A nonparametric approach hardly imposes any restrictions on the functional form of the relations. Only some very weak conditions have to be set on the model. For instance, the aim of the nonparametric approach in a regression
model is to determine a reasonable approximation for the curve that describes the relation between the dependent variable and independent variables. The problem is to determine a smooth curve in such a way that it shows the structure of the data without showing all the details. In other words, one has to find a balance between fitting and smoothing (Härdle 1990). The advantage of the nonparametric approach is that there is little room for misspecification. However, the disadvantage is that even with a large data set the precision of the estimates is often poor (Powell 1994). This is caused by the slow rate of convergence; a problem which is exacerbated by many explanatory variables, the so-called curse of dimensionality.

A semiparametric approach is a hybrid of parametric and nonparametric approaches. Broadly speaking, a semiparametric approach maintains the structure of a model that is most useful for interpreting the results, but hardly imposes restrictions on the other aspects of the model. A linear regression model with an unspecified functional form of the distribution of the error term is an example of a semiparametric model. Its advantage is that it is more flexible than a parametric approach but avoids the curse of dimensionality. Estimation of semiparametric and nonparametric models can be quite awkward since the function that has to be optimized does not behave neatly. Furthermore, the results are often sensitive to the degree of smoothing.

The parametric-nonparametric distinction loosely characterizes the difference of empirical analysis by statisticians and econometricians over the last few decades. Statistical analysis of data is in essence a descriptive activity. That is, the aim is to describe the data patterns; the value of the results rests on the notion that new data sets will exhibit the similar patterns. The nonparametric approach emphasizes this view. In the words of Lehmann (1975): “Why assume anything if you don’t have to?” In contrast, econometric modeling makes use of economic theory and reasoning in two ways. First, economic reasoning suggests certain kinds of variables that have to be included in the empirical analysis. Second, and more fundamental, is that econometric modeling regards the process generating the data as a behavioral regularity. That is, an econometric model specifies an economic behavioral process, as well as the connection between that process and the observed data. The aim of the analysis is the interpretation of this process, and the value of the results depends on the question whether the specified model adequately represents new data sets. Although it may seem that this discussion is somewhat beyond the scope of this thesis, it should be realized that it indicates a fundamental difference
underlying the various models presented in the thesis. We revert to this issue in section 2.5.

We start the overview with the RFM-model, which is the pioneering method of list segmentation. Subsequently, we discuss two techniques, the contingency table and CHAID, which are sometimes classified as clustering-like methods (Cryer et al. 1989), since the analysis generates clusters as output. Then we present several specifications that focus on the binary choice, like probit and logit models, discriminant analysis, the beta-logistic model and latent class analysis. Finally, we present the neural network approach, which is an alternative to conventional statistical methods. We assume that a test mailing is sent to a relatively small sample of individuals of the mailing list. Then, one of the techniques is employed to link the (non)response to the segmentation variables. Next, the technique is used to select the most promising targets on the mailing list.

RFM-model Traditionally, the so-called RFM-model is most frequently used for target selection. As its name suggests, it is solely based on the RFM-variables. Each RFM-variable is split into a number of categories, which are chosen by the researcher. To each of those categories a score is assigned, which is the difference between the average response rate and the observed response rate in that category. For an individual we obtain a score by adding his category scores. This enables us to rank the individuals on the mailing list. The advantage of this method is its simplicity. The drawback is that the model does not account for multicollinearity unless the RFM-variables are mutually independent. Furthermore, it is not clear which individual should receive a mailing. Bauer (1988) extended this method by using the RFM-model in a functional relationship in which the unknown parameters are determined a priori.

Contingency table The simplest method of modeling response is by a contingency table. It is a two- or more dimensional matrix of which the dimensions are determined by the number of variables involved. Each of these variables is divided into a number of categories. Each specific combination of categories, i.e. one category for each variable, defines a cell. For target selection contingency tables can be used as follows: the researcher chooses the variables to be taken into account. If a variable is continuous, it has to be categorized. On the basis of these categories the individuals are scanned and assigned to a cell.
In each cell the response rate is computed, which is used to decide whether the individuals belonging to that cell should be selected. It is a nonparametric method which can be classified as non-criterion based (Magidson 1994), denoting that the clusters, i.e. cells, are not explicitly derived on the basis of a dependent variable. This implies that the continuous variables, or variables with too many categories that have to be recategorized, are categorized in a rather ad-hoc manner. The major drawback of this method is that it becomes very disorderly in the case of many variables, with the risk of having not enough observations per cell, which is the curse of dimensionality. It can, however, be very useful for a first data analysis in order to choose selection variables for further analysis.

**Chi-square automatic interaction detection (CHAID)** Automatic interaction detection (AID), the predecessor of CHAID, is based on a stepwise analysis-of-variance procedure (Sonquist 1970). The dependent variable is interval-scaled and the independent variables are categorical. AID consists of the following steps: (a) for every independent variable the optimal binary split is determined in such a way that the within-group variance of the dependent variable is minimal. That is, AID merges the categories of an independent variable which are, more or less, homogeneous with respect to the dependent variable. For each independent variable this results in an optimal binary split; (b) the variable that generates the lowest within-group variance is used to subdivide the dependent variable in accordance with its optimal binary split; (c) given this binary split, two groups are constructed. Each group is analyzed in the same manner as defined by (a) and (b). This process continues until some stopping criterion is satisfied. This criterion can be either a number of segments or number of individuals in a segment. The result is a treelike group of segments; it is a so-called tree-building method. The individuals that belong to a segment with a sufficiently high response rate should be selected.

The CHAID, developed by Kass (1980), improves the AID in several respects. CHAID merges categories of an independent variable that are homogeneous with respect to the dependent variable, but does not merge the other categories. The result of this merging process, unlike AID, is not necessarily a dichotomy. Hence, steps (a) and (b) may result in more than two subgroups. On the other hand, also in contrast with AID, only those independent variables are eligible to be split into subgroups for which the subgroups differ significantly from each other; the significance is based on a $\chi^2$-statistic. This explicitly defines a stopping rule. Note that it is a nonparametric technique since it does
Selection techniques

not impose a functional form for the relation between the dependent and independent variables. In contrast with cross-tabulation, it may be classified as criterion-based since the splits are derived on the basis of a dependent variable. The main advantage of (CH)AID is that the output is easy to understand and therefore easy to communicate to the management. In practice it is often seen as a useful exploratory data analysis technique, in particular to discover interaction effects among variables, of which output may be used as input for other techniques with more predictive power (Shepard 1995). Magidson (1994) provides an extensive examination on CHAID.

A closely related technique is the so-called classification and regression trees (CART), which was developed by Breiman et al. (1984). The result of this technique is also a treelike group of segments. In contrast with CHAID, the independent variables may be categorical and continuous. The splits are based on some homogeneity measure between the segments formed. Like CHAID, it is a useful method for explorative data analysis and also easy to communicate. A comprehensive discussion on CART can be found in Haughton and Oulabi (1993) and Thrasher (1991).

Linear probability model (LPM) 

Regression analysis with a dichotomous dependent variable (response/nonresponse) is called a linear probability model. The predicted value can be interpreted as the response probability. This probability is assumed to depend on a number of characteristics of the individuals receiving the mailing. That is

\[ y_i = x'_i \beta + u_i, \]

where \( y_i \) is a random variable that takes the value one if the \( i \)th individual responds and zero otherwise, \( x_i \) is a vector of characteristics of \( i \), i.e. RFM-variables, geographic and demographic variables etc.; \( u_i \) is a random disturbance, and \( \beta \) is an unknown parameter. The unknown parameters can be estimated by ordinary least squares (OLS). On the basis of the estimated response probability the potential targets can be ranked in order to make a selection. This model has several readily apparent problems (Judge et al. 1985, p. 757). For example, the use of OLS will lead to inefficient estimates and imprecise predictions. The relation is very sensitive to the values taken by the explanatory variables, and the estimated response probability can lie outside the \([0, 1]\)-interval.
Target selection techniques

Probit and logit model. An alternative way to handle discrete response data is by the probit or logit model. These models overcome the disadvantages of the linear probability model. Probit and logit models assume that there is a latent variable, $y_i^*$, which measures the inclination to respond, which depends on a vector of individual characteristics. That is,

$$y_i^* = x_i' \beta + u_i.$$  \hspace{1cm} (2.1)

As we cannot observe the inclination but only whether or not $i$ responded, we define the dependent variable which we observe, as

$$y_i = \begin{cases} 1 & \text{if } y_i^* \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$  \hspace{1cm} (2.2)

The probit model assumes that the distribution of $u_i$ is normal and the logit model assumes that $u_i$ is logistic. With these fully parametric specifications, the model can easily be estimated by maximum likelihood (ML). The individuals for which the predicted response probabilities is sufficiently high are selected for the direct mailing campaign. In chapters 3, 4 and 5 we elaborate on the probit model in more detail.

Bult and Wansbeek (1995) argue that the selection procedure rests heavily upon the functional form of the disturbance, which could cause inconsistent estimates and would consequently lead to a suboptimal selection rule. Therefore, they use the semiparametric method developed by Cosslett (1983) to estimate the parameters. This method not only maximizes the likelihood function over the parameter space, $\beta$, but also over the space which contains all the distribution functions. Of course, other semiparametric estimators of the binary choice model, e.g. Klein and Spady (1993), could be used as well.

Beta-logistic model. Rao and Steckel (1995) consider a beta-logistic model for the binary choice to allow for heterogeneity. Let $p_i$ be the probability of obtaining a response from $i$. The heterogeneity in $p_i$ is assumed to be captured by a beta distribution with parameters $a_i$ and $b_i$. That is,

$$f(p_i) = \frac{1}{B(a_i, b_i)} p_i^{a_i-1} (1-p_i)^{b_i-1}, \quad 0 \leq p_i \leq 1,$$

where $B(a_i, b_i)$ is the beta function defined by

$$B(a_i, b_i) = \int_0^1 t^{a_i-1} (1-t)^{b_i-1} \, dt.$$
Rao and Steckel favor this distribution mainly because it is flexible so that it can take on a variety of shapes (flat, unimodel, \( J \)-shaped), depending on its parameters. The mean, \( a_i/(a_i + b_i) \) could serve as the response probability for individual \( i \). The variation in the response probability among similar individuals is thus described by the beta distribution. The parameters \( a_i \) and \( b_i \) are modeled as exponential functions of the selection variables (Heckman and Willis 1977), i.e. \( a_i = \exp(\beta_1 x_i) \) and \( b_i = \exp(\beta_2 x_i) \). Loosely speaking, the \( a_i \)s and \( b_i \)s can be used to define segments. The differences in response probabilities within a segment due to variables not included in the analysis are captured by the beta distribution. Although the heterogeneity can be modeled in this way, it cannot be used for making specific predictions. Therefore the mean of the distribution is used to determine the response probability of an individual, i.e. \( p_i \)

\[
p_i \equiv \frac{\exp(\beta_1 x_i)}{\exp(\beta_1 x_i) + \exp(\beta_2 x_i)}.
\]

Note that this expression is equivalent to a logit model with parameters \( \beta_1 - \beta_2 \), hence the name beta-logistic. These parameters can be estimated by ML.

To sum up, in the beta-logistic model a beta distribution is used to allow for heterogeneity in the response probabilities. However, there is still one parameter vector that describes the relation between the selection variables and the response probability; thus we have a logit model with parameter vector \( \beta_1 - \beta_2 \). Rao and Steckel (1995) propose this model since it enables them to determine the response probability for an individual that received \( T \) mailings (for the same offer) but did not respond to any of these. This so-called posterior response probability, \( p_{Ti} \), is given by

\[
p_{Ti} = \frac{\exp(\beta_1 x_i)}{\exp(\beta_1 x_i) + \exp(\beta_2 x_i) + T}.
\]

Note that this probability is decreasing in \( T \). The parameters can be estimated by ML if the organization followed the strategy for at least a sample of the list, of sending individuals follow-up mailings (up to a certain maximum) until they responded. The estimated model can then be used to decide whether or not an individual should receive a follow-up mailing.

**Discriminant analysis** In discriminant or classification analysis we try to find a function that provides the best discrimination between the respondents
and nonrespondents (see Maddala 1983, p. 16). That is, the following loss function is minimized:

\[ L = \sum_{i=1}^{n} | \hat{y}_i - y_i |, \]

where \( y_i \) is one if \( i \) responded and zero otherwise, \( \hat{y}_i = 1 \) if \( i \) is classified as a respondent and \( \hat{y}_i = 0 \) otherwise. On the basis of the estimated discriminant function, individuals can be classified in one of the two groups. Most widely used is the linear discriminant function, \( x' \beta \), which is based on a normality assumption of the explanatory variables. Hence it is a fully parametric specification. Individual \( i \) is classified in the group of nonrespondents if

\[ |x'_i \hat{\beta} - \bar{x}_0| > |x'_i \hat{\beta} - \bar{x}_1|, \]

where \( \bar{x}_0 \) and \( \bar{x}_1 \) are the averages of the explanatory variables of the nonrespondents and respondents, respectively; \( \hat{\beta} = S^{-1}(\bar{x}_0 - \bar{x}_1) \), with \( S \) the overall variance of \( x \) (see Maddala 1983, pp. 15-18). Shepard (1995) shows that if the response is 10% or less, the analysis is very sensitive to the violations of the normality assumptions. This problem can be overcome by using a logit or probit model as the discriminant function (Sapra 1991). Then \( \hat{y}_i = 1 \) if \( P(y_i = 1) \geq 0.5 \) and \( \hat{y}_i = 0 \) if \( P(y_i = 1) < 0.5 \), with \( y_i \) defined in (2.2). It is clear that the misclassification of an individual that responded is penalized as much as the misclassification of an individual that did not respond. In practice, however, the economic costs of these two kinds of misclassification are very different. Hence, the asymmetry in costs of misclassification should be taken into account. This can be accomplished by employing the following classification rule: \( \hat{y}_i = 1 \) if \( P(y_i = 1) \geq \tau \), where \( \tau \) is the ratio between the costs of a mailing and the returns to a positive reply. Note that, although this classification rule captures the asymmetry in the costs of misclassification, the estimated parameters are still based on symmetric costs of misclassification.

Manski (1975) proposes a semiparametric approach to estimate a classification model, the so-called maximum score method. This method yields consistent estimates under weak distributional assumptions. It enables one to incorporate, in a very straightforward manner, the asymmetry of misclassification (Manski 1985). Bult (1993) employs this method in a direct marketing application. The function Bult considers is

\[ L(\beta) = \sum_{i=1}^{n} \{ \rho \mid y_i - \text{sgn}(x'_i \beta) \mid I(y_i = 1) \\
+ (1 - \rho) \mid y_i - \text{sgn}(x'_i \beta) \mid I(y_i = -1) \}, \]

(2.3)
Selection techniques

where \( y_i \) is the observed response, \( y_i = 1 \) if \( i \) responded and \( y_i = -1 \) otherwise, \( \rho = (a - c)/a \), with \( a \) the revenues to a positive reply and \( c \) the cost of a mailing; \( I(\cdot) \) is an indicator function which is one if the argument is true and zero otherwise, and \( \text{sgn}(\cdot) \) is an indicator function which is one if the sign of the argument is positive and minus one otherwise. If \( \rho \) is close to one, it is economically more important to correctly classify the outcome \( y_i = 1 \) than the outcome \( y_i = -1 \). The vector \( \beta \) is estimated by minimizing \( L(\beta) \). Note that, in contrast with the discriminant function described above, the estimated parameters are based on the asymmetry of misclassification.

Model (2.3) can be estimated in LIMDEP (Greene 1992) in a straightforward manner. It can, however, be quite hard to find the optimum since the function generally only reaches a local optimum. Problems in the optimization may also arise by the fact that at least one explanatory variable must have an infinite support (Manski 1985), which is often not available. Moreover, many observations are needed since the estimator has a slow rate of convergence (Manski and Thompson 1986). This means that even 2000 observations, which is the maximum number of observations LIMDEP can handle, might not be sufficient.

Bult and Wittink (1996) apply this method to mailings with response that can take all positive values. On the basis of past purchase behavior individuals are a priori classified in segments. For each segment a classification model is estimated, which is used for the selection of individuals. The disadvantage of the approach is the fact that it is generally difficult to specify the segments a priori in a satisfactory way.

**Latent class analysis** A crucial assumption in the (binary choice) models considered so far is that individuals respond in a homogeneous way. This means that a single model, i.e. one vector of parameters, describes the relation between the selection variables and the dependent variable. This can be misleading if there is substantial heterogeneity in the sample with respect to the magnitude and direction of the response parameters across the individuals (DeSarbo and Ramaswamy 1994). We illustrate this with a very simple example. Assume that there are two regressors, \( x_1 \) and \( x_2 \) to predict \( y^* \), as defined in (2.1). Furthermore, assume that there are two segments of approximately equal size where the models are given by

\[
\begin{align*}
y_i^* &= \beta_1 x_{1i} - \beta_2 x_{2i} + u_i \quad \text{for individuals in segment 1;} \\
y_i^* &= -\beta_1 x_{1i} + \beta_2 x_{2i} + u_i \quad \text{for individuals in segment 2.}
\end{align*}
\]
Thus the two regressors have an opposite effect on the inclination to respond in the two segments. The use of a single response model would yield regression coefficients $\beta_1 \approx \beta_2 \approx 0$. This shows that a single response model is of little benefit for selection if the given underlying segments drive the response behavior.

Therefore DeSarbo and Ramaswamy (1994) propose a (parametric) method that accounts for heterogeneity. Their method simultaneously derives market segments and estimates response models in each of these segments. Thus, instead of specifying a single model, this approach allows for various models. The method provides a (posterior) probability of membership of a certain segment, so-called probabilistic segmentation. Consequently, each individual can be viewed as having a separate response model whose parameters are a combination, based on the posterior probabilities, of the model parameters in the various segments.

In the validation sample it is somewhat tricky to compute these posterior probabilities. Ideally, this should be based on information that is not used to estimate the response models, but useful information to do so is not always available. Another drawback of this approach is that it often has several locally optimal solutions. Furthermore, a lot of effort is needed to specify and estimate the whole model. Since latent class models are developed for descriptive rather than predictive purposes, it would be worthwhile to compare the predictive power with other methods; this is not provided by DeSarbo and Ramaswamy (1994).

A similar approach is given by Wedel et al. (1993). Instead of binary response data they focus on count data, i.e. the number of units ordered in a certain time period. In their application this number ranges from 0 to 75 with an average of 5.2. They use a latent class poisson regression model. This is a parametric specification which allows for heterogeneity across individuals in two ways. First, the mean number of ordered units has a discrete mixture distribution, i.e. it varies over a finite number of latent classes. Second, the mean number of ordered units varies within each segment, depending upon the explanatory variables. Like DeSarbo and Ramaswamy (1994), the model provides a posterior probability that an individual belongs to a certain segment.

Neural Networks (NNs)  Neural Networks have been developed rapidly since the mid-eighties and are now widely used in many research areas. Also in the field of business administration, there has been a tremendous growth in recent years; see Sharda (1994) for a bibliography of applications in this field. Interest
Selection techniques

among researchers in applying neural networks to marketing applications did not start until a few years ago (Kumar et al. 1995). With respect to target selection for direct marketing campaigns, NNs have been mentioned in several theoretical and applied journals as a promising and effective alternative to conventional statistical methods (e.g. Zahavi and Levin 1995, 1997). What makes a NN particularly attractive for target selection is its ability of pattern recognition for automatically specifying and estimating a relation between independent variables and a dependent variable. This property is especially useful when the relationship is complex.

Neural networks can be thought of as a system connecting a set of inputs to a set of outputs in a possible nonlinear way. The links between inputs, the selection variables, and outputs, the response, are typically made via one or two hidden layers. The number associated with a link from node $i$ to node $j$, $w_{ij}$, is called a weight. Figure 2.1 shows a neural network with $k$ input nodes, one hidden layer, which consists of three nodes, and one output node. The input to a node in the hidden layer or output layer is some function of the weighted (using $w_{ij}$) combinations of the inputs. There are several methods to determine the optimal weights for a given NN; the most commonly used method is the back-propagation algorithm.

In a NN, the individual’s likelihood of purchase is expressed by means of a NN score which, although it cannot be interpreted as response probabilities, can be used to rank individuals. Thus, the higher the score, the higher the likelihood of response (Zahavi and Levin 1995). Since we may not interpret the NN score as response probability, it is difficult to determine an optimal selection rule, i.e. a critical score. The individuals with a score above this critical score should receive a mailing and the other individuals should not. Zahavi and Levin (1995, 1997) suggest to use the score that maximizes the profits for the estimation sample as the critical score.

Although NNs are positioned as a method differing from existing statistical methods, which is emphasized by the completely different terminology, there is a close link between these two. Cheng and Titterington (1994) point out the similarities between statistical methods and NNs. Ripley (1994) shows the analogy between nonlinear discrimination analysis and NNs. Thus, many special cases of the neural networks are similar to statistical models. Consequently, in these cases the NN score can be interpreted in the statistical sense, i.e. as (estimated) response probabilities. Hence, an optimal selection rule can be determined. Similarly, a frequently used criticism on NNs is that it is more or less a ‘black box’, and that as a consequence the resulting specification is
2.5 Model comparison

From the various possible selection techniques, we want to select the most appropriate. However, there are many - sometimes conflicting - criteria to evaluate and compare the techniques. Four broad classes of criteria are:

1. The theoretical foundations of the model. This basically refers to the parametric-nonparametric distinction. In the nonparametric approach the aim is to describe the data adequately whereas in the parametric approach an (economic) theory specifies the behavioral process and hence the model (see page 14). This includes the statistical assumptions and the choice of variables.

2. Goodness-of-fit measures. For example, the (adjusted) $R^2$ can be used to evaluate and compare regression based techniques.
3. Predictive validity statistics. These include measures such as the mean squared error and median relative absolute error. These statistics are based on the so-called split-sample analysis. That is, part of the data set is used for estimation, the estimation sample; the other part is used for validation, the validation set. It is a classic cross-validation technique which is commonly employed for models that are built for prediction. Among other things, the analysis guards against overfitting the data.

4. Practical aspects of the model. These include the ease of use, interpretability of the output, intuitive appeal, computer time and so on.

Although these criteria may determine the choice of a technique and may be useful for specifying a model, the bottom-line is to compare the models on the basis of their list segmentation performance. That is, which model generates the highest net profits? Hypothetically, if a method exists about which nothing is known but which guarantees to provide the superior selection at any time, it “would be a sufficient argument to use that method” (Lix and Berger 1995). Note that this argument coincides with that of Lehmann (1975) with respect to the choice between a nonparametric and a parametric approach (see page 14). The argument of Lix and Berger, however, does not mean that theory or ‘understanding’ is unimportant. On the contrary, the best predictive models are often those that are generated or guided by theoretical concerns (e.g. Leahy 1992). In addition, the reliability of such models is also enhanced when the processes by which the observed data are generated are understood. Similarly, practical aspects, goodness-of-fit measures and predictive measures may play an important role in model choice and variable selection.

A method that is often used to analyze the performance with respect to net profits is the gains charts or Pareto curves (e.g. Banslaben 1992, Magliozzi and Berger 1993). This method consists of the following steps: first, the individuals are ranked on the basis of the estimated model; second, groups of equal size are defined. Usually, ten groups are considered. The first group contains the first 10% of the rank-ordered individuals, the second group the following 10% of the rank-ordered individuals, etc. Finally, the average actual and cumulative response probabilities in each of the groups are determined. The gains chart or Pareto curve gives particular insights into the performance of the model. A gains chart should be downward sloping, ideally quite steep. The gains chart of the estimation sample provides useful information to judge the performance of the model on the validation sample. The gains chart of the validation sample could be used to compare various models. It can be rather difficult, however, to compare various models on the basis of this method, since the differences are
generally small and the curves may intersect. That is, one model is better able to identify the top-segments whereas the other model better identifies the mid-range. It is clear that the former is preferred when the organization wants to select the top-segment but it is not necessarily the preferred model when many more individuals are selected. Hence, the gains chart is helpful in comparing models but still does not answer the bottom-line of which method generates the highest net profits. This question can only be answered by specifying a selection rule a priori. Therefore, Bult and Wansbeek (1995) compare models on the basis of an optimal cutoff point, which is defined as the point where the marginal costs equal the marginal revenues; they call this the profit maximizing approach. We take this approach as the validation criterion for our models.

To summarize, there are various ways to compare selection techniques. Our view, which lies at the basis of the models developed in this dissertation, entails that the models should be compared on the basis of a selection rule, which is obtained by equating marginal cost and marginal returns. It does not mean, however, that we believe that the other ways to compare the techniques are useless. Given the diverse criteria for model selection it is, unfortunately, impossible to come up with a model that is universally preferred. In the remainder of this section we will review the scarce literature on model comparisons.

Magidson (1988) compares the CHAID with log-linear modeling. Although the paper reports a successful application of the CHAID, there is no comparison with different techniques using real data. Bult and Wansbeek (1995), on the basis of the profit maximizing approach, show that CHAID performs better than the LPM, parametric, and semiparametric approach of the binary choice model (i.e. logit model and Cosslett’s approach) on the estimation sample. However, as expected, CHAID performs considerably worse than the three other methods on the validation sample. Surprisingly, the parametric approach of the binary choice model outperforms the semiparametric approach. In contrast, Bult (1993) shows that the semiparametric approach of the discriminant model performs better than the parametric approach (logit model).

Within the field of direct marketing, neural networks are not more successful than traditional methods. For instance, Kumar et al. (1995) are inconclusive as to the relative performance of NN in comparison with log-linear regression. Lix and Berger (1995) show that a particular NN performed better than regression based methods, whereas another network performed worse. In a rather comprehensive study, Zahavi and Levin (1997) demonstrate that logistic regression models achieve approximately the same results as NN. They conclude
that “these results are not encouraging for the neural network approach since the process of configuring and setting up a NN is not straightforward”.

With respect to the question whether it is preferred to force variables to enter in the regression equation versus stepwise regression, two studies have been performed. Both compare the models on the basis of a gains chart. Maglioaggi and Berger (1993) show that stepwise regression performs better whereas Lix and Berger (1995) show the opposite.

Altogether, three general conclusions may be made. First, data-exploration methods such as CHAID are less useful for selection than regression type models (e.g. probit or discriminant model). Second, regression models perform reasonably well and, with step-wise entry, have the advantage of clear interpretation. Third, neural networks did not turn out to be the revolution in target selection since the results are comparable to the conventional statistical techniques.

2.6 Discussion and conclusion

We have discussed various techniques that can be employed to select individuals from a mailing list. Unfortunately, it is not possible to draw general conclusions as to the best model for target selection. The main reason for this is that many - sometimes conflicting - criteria play a role in qualifying a technique. Furthermore, even from a theoretical point of view we cannot define a ranking of the models. Moreover, there is no literature that compares all the proposed methods on all these criteria.

As to future research, it would be very interesting to obtain a general conclusion with respect to the generated net profits, given a certain selection rule. It should be realized, however, that there is a number of aspects which could affect the performance of the models. For instance, the kind of information that is available, number of individuals in the estimation sample and the response rate. It is likely that these aspects will make it harder to obtain such a conclusion.

All the techniques discussed in this overview deal with the case of fixed revenues to a positive reply. That is, the approaches focus on binary choice modeling. However, most direct marketing campaigns do not generate simple binary response but rather a response of which the revenue varies between the individuals. Hence, a striking lack in the literature is that hardly any attention has been given to this kind of response. This is not only theoretically a matter
of concern but it can be of major practical importance. In chapter 3 we fill this
gap by incorporating the quantity of response in the selection process.

In the selection process considered so far, the organization specifies a
model and formulates an (optimal) selection rule in order to select the targets.
This selection rule is implicitly based on the assumption that the parameters
are known. Of course, the parameters are unknown and have to be estimated.
These estimates are then plugged into the selection rule, assuming that they
are the true values. It is, however, well known from the statistical literature on
decision making that separation of estimation and decision making, as is the
case here, generally yields lower profits. The reason for this is that estimation
uncertainty in neglected. In chapter 4 we show how these two steps can be
integrated. This leads to a better selection rule and hence to higher expected
profits.