MANAGEMENT SCIENCE
Vol. 51, No. 2, February 2005, pp. 264–276
ISSN 0025-1909 | EISSN 1526-5501 | 05 | 5102 | 0264

DOI 10.1287/mnsc.1040.0296

Customer Targeting: A Neural Network Approach Guided by Genetic Algorithms

YongSeog Kim
Business Information Systems Department, Utah State University, Logan, Utah 84322, yong.kim@usu.edu

W. Nick Street
Management Sciences Department, University of Iowa, Iowa City, Iowa 52242, nick-street@uiowa.edu

Gary J. Russell
Marketing Department, University of Iowa, Iowa City, Iowa 52242, gary-j-russell@uiowa.edu

Filippo Menczer
School of Informatics, Indiana University, Bloomington, Indiana 47408, fil@indiana.edu

One of the key problems in database marketing is the identification and profiling of households that are most likely to be interested in a particular product or service. Principal component analysis (PCA) of customer background information followed by logistic regression analysis of response behavior is commonly used by database marketers. In this paper, we propose a new approach that uses artificial neural networks (ANNs) guided by genetic algorithms (GAs) to target households. We show that the resulting selection rule is more accurate and more parsimonious than the PCA/logit rule when the manager has a clear decision criterion. Under vague decision criteria, the new procedure loses its advantage in interpretability, but is still more accurate than PCA/logit in targeting households.

Key words: database marketing; neural networks; genetic algorithms; customer relationship management
History: Accepted by Jagmohan S. Raju, marketing; received March 26, 2003. This paper was with the authors $10\frac{1}{2}$ months for 1 revision.

1. Introduction

Due to the growing interest in micromarketing, many firms devote considerable resources to identifying households that may be open to targeted marketing messages. The availability of data warehouses combining demographic, psychographic, and behavioral information further encourages marketing managers to use database-based approaches to develop and implement marketing programs.

Database marketers use different tools, depending on what is known about particular households (Winer 2001). Routine mailings to existing customers are typically based on a statistical analysis of the household purchase history (DeSarbo and Ramaswamy 1994, Schmittlein and Petersen 1994, Bult and Wansbeek 1995, Rao and Steckel 1995, Berger and Nasr 1998, Gönül and Shi 1998, Reinartz and Kumar 2000). Marketing consultants often implement the so-called RFM (recency, frequency, monetary) approach, which targets households using summary measures computed from the customer's purchase history (Schmid and Weber 1998, David Sheppard Associates, Inc. 1999).

Mailings to households with no prior relationship with the firm are based on the analysis of the relationship between demographics and the response to a test mailing of a representative household sample (David Sheppard Associates, Inc. 1999). Given the large number of potential demographics available, data dimension reduction is an important factor in building a predictive model that is easy to interpret, cost effective, and generalizes well to unseen cases. Commonly, principal component analysis (PCA) of demographic information (Johnson and Wichern 1992) is used to prepare new variables for this type of analysis. These new variables are then used as predictors in a logistic regression on the test mailing responses.

However, PCA has some drawbacks, in terms of both predictive modeling and dimensionality reduction. PCA does not take into account the relationship between the independent and dependent variables in the process of data reduction. Further, the resulting principal components from PCA can be difficult to interpret when the space of input variables is huge. Finally, constructing principal components still requires collection of all the original predictive variables.

In this study, we propose a new approach to building predictive models for identifying prospective households. The new methodology combines genetic algorithms (GAs) to choose predictive demographic

variables with artificial neural networks (ANNs) to develop a model of consumer response. ANNs (Riedmiller 1994, Sarle 1994) and GAs (Goldberg 1989, Yang and Honavar 1998, Krishna and Murty 1999) have been widely used in machine learning, pattern recognition, image analysis, and data mining. In particular, ANNs have been recognized as a relatively new approach in finance and marketing applications such as stock market prediction (Saad et al. 1998, Pan et al. 1997), bankruptcy prediction (Wilson and Sharda 1994), customer clustering (Gath and Geva 1988, Ahalt et al. 1990), and market segmentation (Hruschka and Natter 1999, Balakrishnan et al. 1996). In this work, we exploit the desirable characteristics of GAs and ANNs to achieve two principal goals of household targeting: model interpretability and predictive accuracy. Our approach is different from previous studies on direct marketing because of our consideration of multiple objectives (Ling and Li 1998) and data reduction (Bhattacharyya 2000).

Data reduction of demographic information is performed via feature selection in our approach. Feature selection is defined as the process of choosing a subset of the original predictive variables by eliminating features that are either redundant or possess little predictive information. If we extract as much information as possible from a given data set while using the smallest number of features, not only can we save a great amount of computing time and cost, but we can also build a model that generalizes better to households not included in the test mailing. Reducing the dimensionality of the input space tends to reduce overfitting in the predictive model, especially for highly flexible models like ANNs, thereby improving generalization. Feature selection can also significantly improve the comprehensibility of the resulting classifier models. In database marketing applications, it is important for managers to understand the key drivers of consumer response. Even a complicated model—such as a neural network—can be more easily understood if constructed from only a few variables.

In our work, a specifically designed GA, the Evolutionary Local Selection Algorithm (ELSA), is used to search through the possible combinations of features (Menczer et al. 2000a, Kim et al. 2000). Two quality measurements—hit rate (which is maximized) and complexity (which is minimized)—are used to evaluate the quality of each feature subset. ELSA performs a local search in the space of feature subsets by evaluating genetic individuals based on both their quality measurements and on the number of similar individuals in the neighborhood in objective space. The bias of ELSA toward diversity makes it ideal for multiobjective optimization, giving the decision maker a clear picture of Pareto-optimal solutions from which to choose.[1]

The approach to feature selection considered in our study is somewhat different from previous research based upon the ELSA algorithm. For example, Menczer et al. (2000a) applied ELSA to select the feature subset that returns the highest classification accuracy over all records. However, in our study, we evaluate individuals based on their ability to rank records based on the estimated probability of belonging to a target class, and to select the feature subset that maximizes classification accuracy over a predetermined number of records (say, the top 20% of records with highest probability of membership in the target class). Feature selection with ELSA in the current study is also different from Kim et al. (2000), in which the main goal is to find the feature subset for constructing optimized clusters, not classification accuracy.

In our approach, the input features selected by ELSA are used to train an artificial neural network that predicts "buy" or "not buy." Using information from households with an observed response, the ANN is able to learn the typical buying patterns of customers in the data set. The trained ANN is tested on an evaluation set, and a proposed model is evaluated both on the hit rate and the complexity (number of features) of the solution. This process is repeated many times as the algorithm searches for a desirable balance between predictive accuracy and model complexity. The result is a highly accurate predictive model that uses only a subset of the original features, thus simplifying the model and reducing the risk of overfitting. Because the algorithm identifies variables with no predictive value, it also provides useful information on reducing future data collection costs.

This paper is organized as follows. In §2, we explain ELSA in detail. In §3, we describe the structure of the ELSA/ANN model, and review the feature subset selection procedure. In §4, we present experimental results of both the ELSA/ANN and PCA/logit model algorithms. Using test-mailing responses on insurance policies, we show that there is a trade-off between model interpretability and predictive accuracy. In particular, we obtain both high model interpretability and high predictive accuracy only when the firm is specific about the way model forecasts will be used to select households in future mailings. In contrast, interpretability must be sacrificed to preserve predictive accuracy when the firm is vague about its selection rule. In §5, we discuss three important issues: interpretability, time complexity, and scalability of the ELSA/ANN approach. Section 6

[1] Pareto-optimal solutions are a group of solutions that are superior to others in at least one objective quality. We define this more formally in §2.3.

concludes the paper and provides suggestions about future research directions.

2. Evolutionary Local Selection Algorithm (ELSA)

2.1. Local Selection

ELSA is a variation of evolutionary (or genetic) algorithms, motivated by artificial life models of adaptive agents in ecological environments (Menczer and Belew 1996). The point of departure between ELSA and traditional GAs is the observation that reproduction and selection, like other processes of biological organisms, are locally mediated by the environment in which the agents are situated.

In a standard evolutionary algorithm, an individual (that is, a candidate solution) is selected for reproduction based on how its fitness compares to that of other individuals. In ELSA, an individual agent may die, reproduce, or neither, based on an endogenous energy level that fluctuates via interactions with the environment. The energy level is compared against constant thresholds for reproduction and survival. An individual's energy is increased based on its fitness, and decreased based on the number of agents with similar fitness. This *local* selection scheme naturally enforces the diversity of the population. It also makes ELSA appropriate for multiobjective optimization problems, as fitness can be measured—and energy allocated—separately along each objective.

The following subsection briefly describes the ELSA implementation for the feature selection problem. A more extensive discussion of the algorithm and its application to Pareto optimization problems can be found elsewhere (Menczer et al. 2000a, b).

2.2. ELSA Algorithm Details

Figure 1 outlines the ELSA algorithm at a high level of abstraction. The representation of an agent consists of D bits, with each of the bits indicating whether the corresponding feature is selected or not (1 if a feature is selected, 0 otherwise). Each agent in the population is first initialized with some random solution and an initial reservoir of energy.

Mutation is the main operator used to explore the search space. At each iteration an agent produces one mutated clone to be evaluated. The clone is identical to its parent except for one mutated bit (that is, one feature either added or removed). The agent competes for energy based on its multidimensional fitness and the proximity of other agents in the solution space.

In each iteration of the algorithm, an agent explores a candidate solution (the mutated clone). The agent collects ΔE from the environment and is taxed with

Figure 1 ELSA Pseudocode

```
initialize population of agents, each with energy θ/2
while there are alive agents and for T iterations
  for each energy source c
    for each v (0 .. 1)
      E^c_envt(v) ← 2vE^c_tot
    endfor
  endfor
  for each agent a
    a' ← mutate(clone(a))
    for each energy source c
      v ← Fitness(a', c)
      ΔE ← min(v, E^c_envt(v))
      E^c_envt(v) ← E^c_envt(v) − ΔE
      E_a ← E_a + ΔE
    endfor
    E_a ← E_a − E_cost
    if (E_a > θ)
      insert a' into population
      E_a' ← E_a/2
      E_a ← E_a − E_a'
    else if (E_a < 0)
      remove a from population
    endif
  endfor
endwhile
```

Note. In each iteration, the environment is replenished and then each living agent executes the main loop. In sequential implementations, the main loop calls agents in random order to prevent sampling effects. We stop the algorithm after T iterations.

a constant cost E_{cost} ($E_{\text{cost}} < \theta$) for this "action." The net energy intake of an agent is determined by its fitness. This is a function of how well the candidate solution performs with respect to the criteria being optimized. However, the energy also depends on the state of the environment. The environment corresponds to the set of possible values for each of the criteria being optimized.[2] We imagine an energy source for each criterion, divided into bins corresponding to its values. So, for criterion fitness F_c and bin value v, the environment keeps track of the energy $E^c_{\text{envt}}(v)$ corresponding to the value $F_c = v$. Further, the environment keeps a count of the number of agents $P_c(v)$ having $F_c = v$. The energy corresponding to an action (alternative solution) a for criterion F_c is given by

$$Fitness(a, c) = \frac{F_c(a)}{P_c(F_c(a))}. \tag{1}$$

Candidate solutions receive energy only inasmuch as the environment has sufficient resources; if these are depleted, no benefits are available until the environmental resources are replenished. Thus, an agent is rewarded with energy for its high fitness values, but also has an interest in finding unpopulated niches in objective space, where more energy is available.

[2] Continuous objective functions are discretized.

The result is a natural bias toward diverse solutions in the population.

In the selection part of the algorithm, each agent compares its current energy level with a fixed threshold θ. If its energy is higher than θ, the agent reproduces: The mutated clone that was just evaluated becomes part of the population, with half of its parent's energy. When an agent runs out of energy, it is killed.

Instead of being constant, the population size is maintained dynamically over the iterations and is determined by the carrying capacity of the environment, depending on the costs incurred by any action, and on the replenishment of resources, both described below (Menczer et al. 2000b). The population size is also independent of the reproduction threshold, θ, which only affects the energy stored by the population at steady state.

When the environment is replenished with energy, each criterion c is allocated an equal share of energy:

$$E_{\text{tot}}^{c} = \frac{p_{\max} E_{\text{cost}}}{C}, \tag{2}$$

where C is the number of criteria considered. This energy is apportioned in linear proportion to the values of each fitness criterion, so as to bias the population toward more promising areas in objective space. Note that the total replenishment energy that enters the system at each iteration is $p_{\max} \cdot E_{\text{cost}}$, which is independent of the population size p, but proportional to the parameter $p_{\max}$. This way, we can maintain p below $p_{\max}$ on average, because in each iteration the total energy that leaves the system, $p \cdot E_{\text{cost}}$, cannot be larger than the replenishment energy.

2.3. Feature Selection with ELSA vs. Standard GA

Feature selection with standard GAs has been widely applied for various applications and has shown some success. However, a standard GA can handle only a single evaluation criterion. This is a considerable disadvantage when decision makers need to take into account multiple objectives simultaneously. For example, in our study, we consider two principal quality measures, model interpretability and predictive accuracy, in evaluating each feature subset. Note that often these measures can be in conflict; no single criterion for feature selection is best for every application (Dy and Brodley 2000).

The most common approach in a standard GA framework for considering multiple objectives is to linearly combine them into one evaluation criterion in a subjective manner (Ishibuchi and Nakashima 2000, Opitz 1999, Yang and Honavar 1998). However, this approach cannot capture nonlinear trade-offs among criteria. More importantly, this approach may not be useful for the decision maker because only she can determine the relative weights of criteria for her application.

To provide a clear picture of the trade-offs among the various objectives, feature selection has been formulated as a *multiobjective* or *Pareto* optimization problem. A number of multiobjective extensions of evolutionary algorithms have been proposed in recent years (Deb and Horn 2000). Most of them, such as the Niched Pareto Genetic Algorithm (Horn 1997), employ computationally expensive selection mechanisms like fitness sharing (Goldberg and Richardson 1987) and Pareto tournaments to favor dominating solutions and to maintain diversity. Instead, ELSA maintains diversity over multiple objectives by employing a more efficient local selection scheme.

In ELSA, we evaluate each feature subset in terms of multiple objectives. Formally, each solution s_i is associated with an evaluation vector $F(s_i) = (F_1(s_i), \ldots, F_C(s_i))$, where C is the number of quality criteria. One solution s_1 is said to *dominate* another solution s_2 if $\forall c$: $F_c(s_1) \geq F_c(s_2)$ and $\exists c$: $F_c(s_1) > F_c(s_2)$, where F_c is the cth criterion, $c \in \{1 \ldots C\}$. Neither solution dominates the other if $\exists c_1, c_2$: $F_{c_1}(s_1) > F_{c_1}(s_2)$, $F_{c_2}(s_2) > F_{c_2}(s_1)$. We define the *Pareto front* as the set of nondominated solutions. In feature selection as a Pareto optimization problem, the goal is to approximate as well as possible the Pareto front, presenting the decision maker with a set of high-quality solutions from which to choose. Non-Pareto solutions will not be considered because they are inferior to those in the Pareto front by definition.

Local selection naturally enforces the diversity of the population by evaluating agents based on both their quality measurements and the number of similar individuals in the neighborhood in the objective space. Therefore, ELSA can gradually search for new feature subsets that are more promising, but have not yet been explored. Because identifying predictive feature subsets requires an extensive search for new and better solutions, the maintenance of diversity within the population is more important than a speedy convergence to the optimum. This bias toward exploration also results in a more complete picture of the Pareto front, giving the decision maker more information with which to judge the trade-offs among the various objectives.

Another advantage of ELSA is its minimal centralized control over agents. By relying on local selection, ELSA minimizes the communication among agents (e.g., the ranking of agent fitness relative to the other agents), which makes the algorithm efficient in terms of computational time and scalability (Menczer et al. 2000a). In this application, however, the computation time is dominated by the training of the individuals rather than the evolutionary mechanism.

3. ELSA/ANN Model for Customer Targeting

Our predictive model of household buying behavior is a hybrid of the ELSA and ANN procedures. In this approach, ELSA identifies relevant consumer descriptors that are used by the ANN to forecast consumer choice. We focus here on the structure of the approach and the criteria used to select an appropriate predictive model.

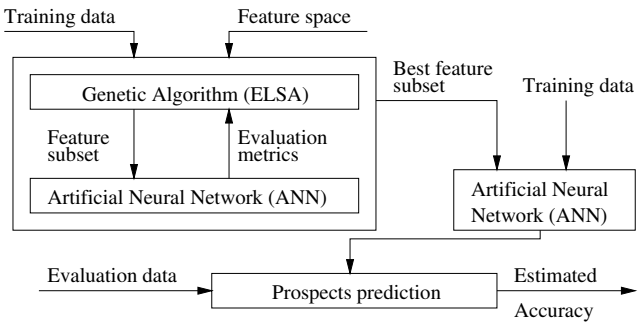
3.1. Structure of the ELSA/ANN Model

The model setup is shown in Figure 2. ELSA searches for a set of feature subsets and passes them to an ANN. The ANN extracts predictive information from each subset and learns the patterns using a randomly selected 2/3 of the training data. Once an ANN learns the data patterns, the trained ANN is evaluated on the remaining 1/3 of the training data, and returns two evaluation metrics, F_{accuracy} and $F_{\text{complexity}}$, to ELSA. It is important to note that in both the learning and evaluation procedures, the ANN uses only the selected features.

Based on the returned metric values, ELSA biases its search direction to maximize the two objectives. This routine continues until the maximum number of iterations is attained. All evaluated solutions over the generations are saved into an offline solution set without comparison to previous solutions. In this way, high-quality solutions are maintained without affecting the evolutionary process.

Among all the evaluated subsets, we choose for further evaluation the set of candidates that satisfy a minimum hit-rate threshold. With these chosen candidates, we start a more rigorous selection procedure, 10-fold cross-validation. In this procedure, the training data are divided into 10 nonoverlapping groups. We train an ANN using the first nine groups of training data and test the trained ANN on the remaining group. We repeat this procedure until each of the 10 groups is used as a test set once. We then take the average of the accuracy measurements over the 10 folds and call it an *intermediate* accuracy. We repeat the 10-fold cross-validation procedure five times and average the five intermediate accuracy estimates. We call this the *estimated* accuracy through the following sections.

Figure 2 The Structure of the ELSA/ANN Model



Note. ELSA searches for a good subset of features and passes them to an ANN. The ANN calculates the "goodness" of each subset and returns two evaluation metrics to ELSA.

Note that evaluating candidates through cross-validation is computationally expensive. However, this is necessary to have an accurate estimate of actual hit rate of candidate solutions. If we select a solution without cross-validation or other rigorous testing, our chosen solution may or may not perform well on unseen data. Therefore, trade-offs between computational cost and accurate estimation should be considered in advance as a part of experimental design. This notion raises an open question about the effectiveness of cross-validation on the performance of a chosen solution.[3] However, our main focus in this study is to introduce a new methodology for customer targeting and to compare the proposed algorithm to others. Further, we use the same cross-validation routine for all the algorithms compared. Therefore, we leave the sensitivity analysis of cross-validation to future research.

For evaluation purposes, we select a single best solution in terms of both estimated accuracy and complexity. We subjectively decided to pick a solution with the minimal number of features at the marginal accuracy level.[4] Once we decide on the best solution, we train the ANN using all the training data with the selected features only. The trained model is then used to rank the potential customers (the records in the evaluation set) in descending order of purchase probability, as predicted by the ANN. We finally select the top x% of the prospects and calculate the *actual* accuracy of our model using the actual choices of the evaluation-set households.

3.2. Evaluation Metrics

We define two heuristic evaluation criteria, F_{accuracy} and $F_{\text{complexity}}$, to evaluate selected feature subsets. Each objective, after being normalized into 25 intervals to allocate energy, is maximized by ELSA.

F_{accuracy}. The purpose of this objective is to favor feature sets with a higher hit rate. Each ANN takes a selected set of features to learn data patterns and predicts which potential customers will actually purchase the product. In our application, we define two different measures, F^1_{accuracy} and F^2_{accuracy} for two different experiments. Experiment 1 assumes that the managers can specify in advance the rule to be used in

[3] The authors thank an anonymous referee for pointing out this issue.

[4] If other objective values are equal, we prefer to choose a solution with small variance.

selecting households for mailings. We select the top 20% of potential customers in descending order of the probability of purchasing the product and compute the ratio of the number of actual customers, AC, out of the chosen prospects, TC. We calculate F^1_{accuracy} as follows:

$$F^1_{\text{accuracy}} = \frac{1}{Z^1_{\text{accuracy}}} \frac{AC}{TC}, \tag{3}$$

where Z^1_{accuracy} is an empirically derived constant to normalize F^1_{accuracy}.

In Experiment 2, we consider a generalization of Experiment 1. We first divide the range of customer selection percentages into 50 intervals with equal width (2%) and measure accuracy at the first m intervals only.[5] At each interval $i \le m$, we select the top $(2 \cdot i)$% of potential customers in descending order of the probability of purchasing the product and compute the ratio of the number of actual customers, AC_i, out of the total number of actual customers in the evaluation data, Tot. We multiply by the width of the interval and sum those values to get the area under the lift curve over m intervals. Finally, we divide by m to get our final metric, F^2_{accuracy}. We formulate it as follows:

$$F^2_{\text{accuracy}} = \frac{1}{Z^2_{\text{accuracy}}} \frac{1}{m} \sum_{i=1}^{m} \frac{AC_i}{Tot} \cdot 2, \tag{4}$$

where $Tot = 238$, $m = 25$, and Z^2_{accuracy} is an empirically derived constant to normalize F^2_{accuracy}.

$F_{\text{complexity}}$**.** This objective is aimed at finding parsimonious solutions by minimizing the number of selected features as follows:

$$F_{\text{complexity}} = 1 - \frac{d-1}{D-1}, \tag{5}$$

where d and D represent the dimensionality of the selected feature set and of the full feature set, respectively. Note that at least one feature must be used. Other things being equal, we expect that lower complexity will lead to easier interpretability of solutions, as well as better generalization.

4. Application

The proposed ELSA/ANN methodology is applied to the prediction of households interested in purchasing an insurance policy for recreational vehicles (RVs). To benchmark the new procedure, we contrast the performance of the ELSA/ANN methodology to an industry-standard logit approach that summarizes household background information using principal components analysis. We evaluate the ELSA/ANN approach using two experiments. In Experiment 1, we inform the algorithm of the way in which the predictive model will be used by managers to select households for a direct mail solicitation. In Experiment 2, we leave this information vague. We show that the new approach provides improvements in forecasting accuracy, but that model complexity is contingent on the amount of information about the managerial decision rule.

4.1. Data Description

The data are taken from a solicitation of 9,822 European households to buy insurance for an RV. These data, taken from the CoIL 2000 forecasting competition (Kim and Street 2000), provide an opportunity to assess the properties of the ELSA/ANN procedure in a customer-prospecting application.[6] In our analysis, we use two separate data sets: a training set with 5,822 households and an evaluation set with 4,000 households. The training data are used to calibrate the model and to estimate the hit rate expected in the evaluation set. Of the 5,822 prospects in the training data set, 348 purchased RV insurance, resulting in a hit rate of $348/5822 = 5.97\%$. From the manager's perspective, this is the hit rate that would be obtained if solicitations were sent out randomly to consumers in the firm's database.

The evaluation data are used to validate the predictive models. Our predictive model is designed to return the top x% of customers in the evaluation data set judged to be most likely to buy RV insurance. The model's predictive accuracy is examined by computing the observed hit rate among the selected households. It is important to understand that only information in the training data set is used in developing the model. Data in the evaluation data set are used exclusively for forecasting.

In addition to the observed RV insurance policy choices, each household's record contains 93 additional variables, containing information on both sociodemographic characteristics (Variables 1–51) and ownership of various types of insurance policies (Variables 52–93). Details are provided in Table 1. The sociodemographic data are based upon postal code information. That is, all customers living in areas with the same postal code have the same

[5] This could be justified in terms of costs to handle the chosen prospects and the expected accuracy gain. As we select more prospects, the expected accuracy gain will go down. If the marginal revenue from an additional prospect is much greater than the marginal cost, however, we could sacrifice the expected accuracy gain. Information on mailing cost and customer value was not available in this study.

[6] We use a data set on consumer responses to a solicitation for "caravan" insurance policies. A "caravan" is similar to an RV in the United States. For more information about the CoIL competition and the CoIL data sets, refer to http://www.dcs.napier.ac.uk/coil/challenge/.

Table 1 Household Background Characteristics

Feature ID	Feature description
1	Number of houses owned by residents
2	Average size of households
3	Average age of residents
4–13	Psychographic segment: successful hedonists, driven growers, average family, career loners, living well, cruising seniors, retired and religious, family with grownups, conservative families, or farmers
14–17	Proportion of residents with Catholic, Protestant, other, and no religion
18–21	Proportion of residents of married, living together, other relation, and singles
22–23	Proportion of households without children and with children
24–26	Proportion of residents with high, medium, and lower education level
27	Proportion of residents in high status
28–32	Proportion of residents who are entrepreneur, farmer, middle management, skilled laborers, and unskilled laborers
33–37	Proportion of residents in social class A, B1, B2, C, and D
38–39	Proportion of residents who rented home and owned home
40–42	Proportion of residents who have 1, 2, and no car
43–44	Proportion of residents with national and private health service
45–50	Proportion of residents whose income level is <$30,000; $30,000–$45,000; $45,000–$75,000; $75,000–$123,000; >$123,000; and average
51	Proportion of residents in purchasing-power class
52–72	Scaled contribution to various types of insurance policies such as private third party, third-party firms, third-party agriculture, car, van, motorcycle/scooter, truck, trailer, tractor, agricultural M/C, moped, life, private accident, family accidents, disability, fire, surfboard, boat, bicycle, property, social security
73–93	Scaled number of households holding insurance policies for the same categories as in scaled contribution attributes

sociodemographic attributes. The insurance firm in this study scales most sociodemographic variables on a 10-point ordinal scale (indicating the relative likelihood that the sociodemographic trait is found in a particular postal code area). This 10-point ordinal scaling includes variables denoted as "proportions" in Table 1. For the purposes of this study, all these variables were regarded as continuous. The psychographic segment assignments (Attributes 4–13), however, are household specific and are coded into 10 binary variables.

In our subsequent discussion, the word feature refers to one of the 93 variables listed in Table 1. For example, the binary variable that determines whether or not a household falls into the "successful hedonist" segment is a single feature. Accordingly, in the feature selection step of the ELSA/ANN model, the algorithm can choose to use any possible subset of the 93 variables in developing the predictive model.

4.2. Experiment 1

In Experiment 1, we maximize the hit rate when choosing the top 20% potential customers as in Kim and Street (2000). We select the top 20% of customers in the evaluation data set using the model created by the ELSA/ANN procedure. The actual choices of these households provide a measure of the hit rate. For comparison purposes, we implemented a principal component analysis (PCA) of the household background characteristics followed by a logistic regression of the insurance policy choice data. PCA is analogous to our feature selection procedure to reduce data dimension. The logistic regression is, in fact, an example of a very simple ANN. The PCA/logit approach is commonly used by industry consultants in developing household selection rules.

We also implemented an intermediate model, ELSA/logit, for comparison purposes. The ELSA/logit model is different from ELSA/ANN in the sense that it uses only one hidden node.[7] We use the same criterion to select the final solution of ELSA/logit as is done in ELSA/ANN. The motivation behind the ELSA/logit model is the decomposition of the accuracy gain of ELSA/ANN into two sources: feature selection and response function approximation. The difference in results between PCA/logit and ELSA/logit can be attributed to characteristics of feature selection, while the difference in results between ELSA/logit and ELSA/ANN can be attributed to the greater flexibility of ANN in approximating the response model.

Before discussing results, we first briefly summarize our implementation of the PCA/logit benchmark model in Figure 3. We selected 22 principal components. This is the minimum required to explain more than 90% of the variance in the training set. To get the estimated hit rate, we implement 10-fold cross-validation on the training set, as shown in Figure 4. In the cross-validation procedure, the scores of the PCs are estimated using different portions of the data each time to get the estimated hit rate.

We set the values for ELSA parameters in the ELSA/ANN and ELSA/logit models as follows: Pr(mutation) $= 1.0$, $p_{\max} = 1{,}000$, $E_{\text{cost}} = 0.2$, $\theta = 0.3$, and $T = 2{,}000$. In both models, we select the single solution that has the highest expected hit rate among those solutions that use less than 10% of the available featues. (This criterion was adopted to restrict attention to models that are relatively parsimonious.) We evaluated each model on the evaluation set. Table 2 first shows the estimates of the actual hit rates of three different models based on five replications of 10-fold cross-validation.

In terms of the actual hit rate, all three models work very well. Even the model with lowest actual hit rate (PCA/logit) is 2.3 times better than the hit rate expected by mailing to these households at random

[7] ELSA/ANN models use $\sqrt{node_{\text{in}}}$ hidden nodes, where $node_{\text{in}}$ represents the number of input nodes.

Figure 3 The Implementation Procedure of the PCA/Logit Model

Apply PCA on training data D_{train}
Determine appropriate number of PCs, n
Reduce the dimensionality of D_{train} using n PCs, creating D'_{train}
Perform logistic regression on D'_{train} and save $\hat{\beta}_i$ and $\hat{\alpha}$ where $i = 1, \ldots, n$.
Reduce the dimensionality of evaluation data D_{eval} using n PCs, creating D'_{eval}
Calculate p(not buy) for each record in D'_{eval} using

$$p = \frac{\exp(\hat{\alpha} + \sum_{i=1}^{n} \hat{\beta}_i \cdot PC_i)}{1 + \exp(\hat{\alpha} + \sum_{i=1}^{n} \hat{\beta}_i \cdot PC_i)}$$

Select 20% of records, R, with lowest p
for each selected record r
  if r is an actual customer
    $counter = counter + 1$
  endif
endfor
$Hit_{rate} = counter/R$

(5.97%). The model generated by the ELSA/ANN procedure returns the highest actual hit rate. (The difference of the estimated hit rates between the PCA/logit and the ELSA/ANN models is statistically significant at $\alpha = 0.05$.) As noted earlier, the difference in actual hit rate between PCA/logit and ELSA/logit provides an estimate of the accuracy gain that comes from the ELSA feature selection procedure. The difference in actual hit rate between ELSA/logit and ELSA/ANN provides an estimate of the accuracy gain that comes from the additional flexibility that ANN provides in approximating the true response function. In this application, both aspects of the ELSA/ANN procedure contribute equally to the improved accuracy of the model.

Figure 4 The Implementation Procedure of Cross-Validation for the PCA/Logit Model

Divide training data D_{train} into 10 equal-sized subsets
for each subsets $D_{train}[i], i = 1, \cdots, 10$
  Define $D_{train}[i]^c$ = D_{train} - $D_{train}[i]$
  Apply PCA on $D_{train}[i]^c$, and select n PCs
  Reduce the dimensionality of $D_{train}[i]^c$ using n PCs
  Do logistic regression on reduced $D_{train}[i]^c$
  Reduce the dimensionality of $D_{train}[i]$ using n PC scores
  Calculate p(not buy) using the formula in Figure 3
  Pick 20% of records, $R[i]$, with lowest p
  for each selected record r
    if r is an actual customer
      $counter[i] = counter[i] + 1$
    endif
  endfor
endfor
$Hit_{rate} = \sum_{i=1}^{10} counter[i] / \sum_{i=1}^{10} R[i]$

Note. We used the same number of PCs, $n = 22$, as we did in Figure 3.

Table 2 Results of Experiment 1

Model (# features)	Training set Hit rate (%) ± s.d. (%)	Evaluation set # Correct	Evaluation set Hit rate (%)
PCA/logit (22)	12.83 ± 0.498	109	13.63
ELSA/logit (6)	15.73 ± 0.203	115	14.38
ELSA/ANN (7)	15.92 ± 0.146	120	15.00

Note. The hit rates from the three different models are shown as percentages with standard deviation. The column marked "# Correct" shows the number of actual customers who are included in the chosen top 20%. The number in parentheses represents the number of selected features except for the PCA/logit model, where it represents the number of PCs selected.

We also compare ELSA/ANN to an ANN without feature selection to check if feature selection simplifies the model at the expense of sacrificing predictive accuracy. This is interesting because it has been strongly assumed that more information leads to a better predictive model. However, our experiments indicate otherwise. The average accuracy of a single ANN from 10 independent runs on the evaluation set is 13.62% ± 0.314%, which is significantly lower than that of the ELSA/ANN procedure. We attribute this to the fact that the ELSA/ANN model can eliminate noisy features that are negatively correlated with other predictive features and thus deteriorate the overall performance of a final model. Therefore, it is not the quantity, but the quality, of information that makes a better predictive model.

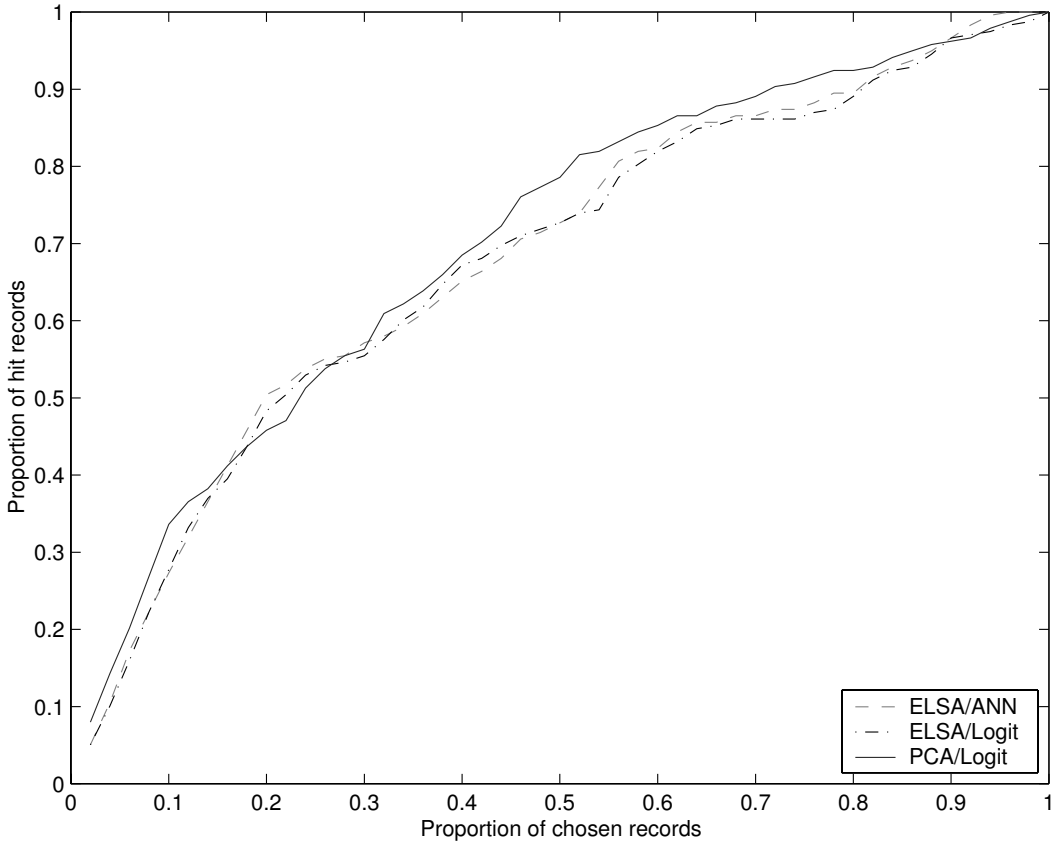
Judging the interpretability of a model is necessarily subjective. An advantage of the ELSA/ANN approach is that predictive features are clearly highlighted. In contrast, the PCA/logit model uses all of the features in constructing the principal component scores. We show the seven features that the ELSA/ANN procedure selected in Table 3.

With the exception of the "Average Family" psychographic segment, all other features are reports of the insurance-buying behavior of the household's postal code area. The feature reporting car insurance makes considerable sense, given the fact that the firm is soliciting households to buy insurance for RVs. Further evaluation shows that prospects with at least two insured autos are the most likely RV purchasers. Moped policy ownership is justified by the fact that many people carry their mopeds or bicycles on the back of RVs. Those two features are selected again by

Table 3 Selected Features by ELSA/ANN in Experiment 1

Feature type	Selected features
Demographic features	"Average Family" psychographic segment
Behavioral features	Amount of contribution to third-party policy, car policy, moped policy, and fire policy, and number of households holding third-party policies and social security policies

Figure 5 Lift Curves of Three Models That Maximize the Hit Rate when Targeting the Top 20% of Prospects



the ELSA/logit model.[8] Using this type of information, we are able to build a potentially valuable profile of likely customers (Kim and Street 2000).

In general, the results are in line with marketing science work on customer segmentation, which shows that information about current purchase behavior is most predictive of future choices (Rossi et al. 1996). The fact that the ELSA/ANN model used only seven features for customer prediction also implies that the firm could reduce data collection and storage costs considerably. This is possible through reduced storage requirements ($86/93 \approx 92.5\%$) and the reduced labor and data transmission costs.

We also compare the three models in terms of lift curves.[9] Figure 5 shows the cumulative hit rate over the top $2 \leq x \leq 100\%$ prospects. Clearly, our ELSA/ANN model is the best when the firm selects the top 20% of prospects for a direct mail solicitation. However, the performance of ELSA/ANN and ELSA/logit over all targeting percentages was worse than that of PCA/logit. This occurs because our solution is specifically designed to optimize the hit rate when managers select the top 20% of prospects. In contrast, the PCA/logit model is estimated without any knowledge of how model forecasts will be used in decision making. This observation motivated a second experiment in which we attempt to improve the performance of the ELSA/ANN model over a greater range of decision rules.

4.3. Experiment 2

In this experiment, we search for the solution that best maximizes the accuracy defined in a more global sense. The algorithm is designed to maximize the area under the lift curve, up to the top 50% of potential customers. Logically, the best solution from Experiment 1 is not necessarily the best solution in the more generalized environment of Experiment 2. In fact, our results are consistent with this observation. We also implemented the PCA/logit and the ELSA/logit model again for comparison purposes. We first show the generalized procedure of PCA/logit to get the estimated accuracy in Figure 6.

[8] The other four features selected by the ELSA/logit model are contribution to bicycle policy and fire policy, and number of trailer policies and lorry policies.

[9] Lift is defined as the percentage of all buyers in the database who are in the group selected for a direct mail solicitation. Under random sampling, the lift curve is a 45-degree line starting at the origin of the graph.

Figure 6 The Generalized Implementation of the PCA/Logit Model

```
Apply PCA on training data D_train
Determine appropriate number of PCs, n
Reduce the dimensionality of D_train using n PCs,
    creating D'_train
Perform logistic regression on D'_train and save β̂_i
    and α̂ where i = 1, ..., n.
Reduce the dimensionality of evaluation data D_eval
    using n PCs, creating D'_eval
Calculate p(not buy) for each record in D'_eval using
            p = exp(α̂ + Σ_{i=1}^{n} β̂_i · PC_i) / (1 + exp(α̂ + Σ_{i=1}^{n} β̂_i · PC_i))
for each  i = 1 to int_num
   x = int_width · i
   Select x% records with lowest p
   for each  selected record r
      if  r is an actual customer
         counter = counter + 1
      endif
   endfor
   Hit_rate = counter/Tot
   Accuracy = Accuracy + Hit_rate * int_width
endfor
Accuracy = Accuracy/int_num
```

Note. We use $n = 22$ (as in Experiment 1), $\text{int}_{\text{num}} = 25$, $\text{int}_{\text{width}} = 2$, and $Tot = 238$.

The ELSA/ANN and ELSA/logit models are adjusted to maximize the overall area under the lift curve over the same intervals as in PCA/logit. Because this new experiment is computationally much more expensive, we take a slightly different approach to choose the final solutions of ELSA/ANN and ELSA/logit. We used twofold cross-validation estimates of all solutions and set the values of the ELSA parameters identically with the previous experiment, except $p_{\max} = 200$ and $T = 500$. Following the approach used in Experiment 1, we initially examined only those predictive models having less than 10% of the total set of predictive variables. Interestingly, this criterion resulted in an ELSA/ANN model (not shown) with worse predictive performance than PC/logit. We then recomputed the predictive models, choosing the best solutions having less than half of the original features. Our intuition was that a less parsimonious ELSA/ANN model might dominate PC/logit. Indeed, we found this to be the case. Results of Experiment 2 are summarized in Table 4 and in Figure 7.

Table 4 shows that the ELSA/ANN model has higher hit rates than PCA/logit over the solicitation range between 15% and 50% of total households. In particular, ELSA/ANN is best when choosing 15%, 20%, 25%, and 50% of the targeting points, and tied for the best at 30%, 35%, and 45%. The overall performance of ELSA/logit is better than that of PCA/logit. We attribute this to the fact that both models benefit from the ELSA feature selection methodology.

The lift curves in Figure 7 show that ELSA/ANN has much-improved global characteristics relative to Experiment 1. However, we note that there are costs associated with this improved performance. First, the hit rate of ELSA/ANN at the 20% solicitation rate is now lower than in Experiment 1 (14.42% versus 15.00%). Second, the well-established parsimony and interpretability of the models selected by ELSA/ANN in Experiment 1 is largely lost in Experiment 2. We attribute this partially to the fact that different selection points may have related but different optimal subsets of features. Correlation among features seems to contribute to the loss of parsimony. For instance, a particular variable related to insurance policy ownership that is part of the optimal subset at a 20% selection rate could easily be replaced by a different, correlated feature at 30%. It should be noted that the ELSA/ANN model is superior to the PCA/logit model in the sense that ELSA/ANN works with feature subsets, while PCA/logit always requires the whole feature set to construct PCs.

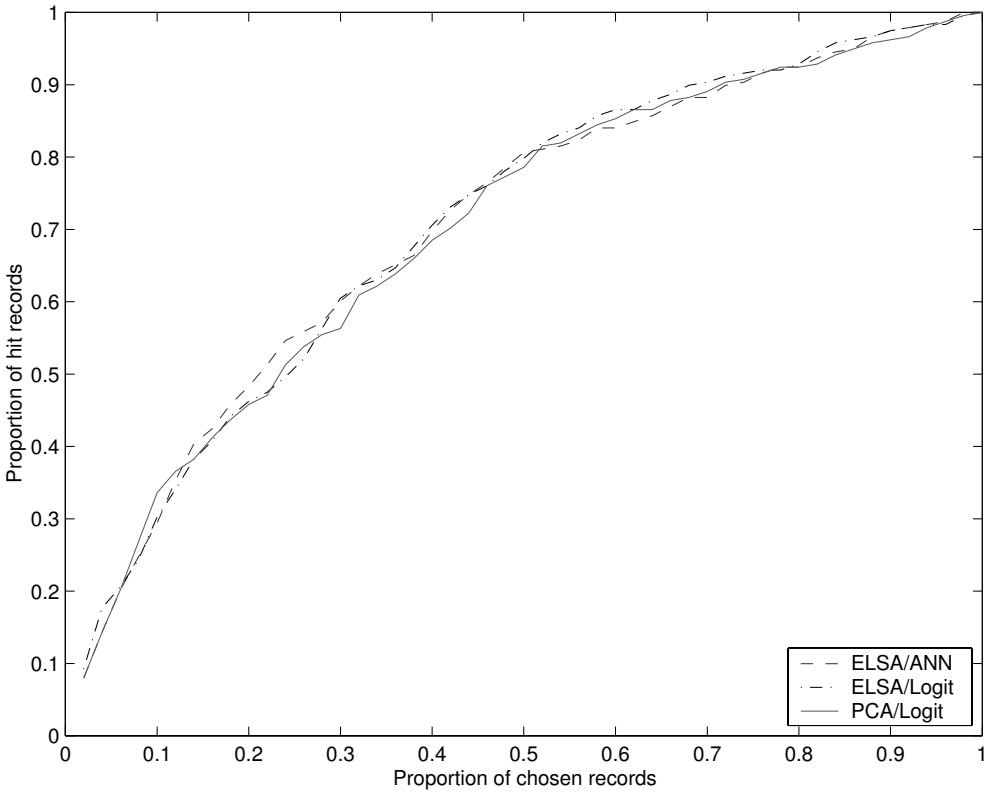
These aspects of the solution provide strong evidence that there exists a key trade-off in building a predictive model. By focusing on a specific decision scenario (as in Experiment 1), we are able to construct a procedure that is parsimonious and has superior predictive performance. When the decision scenario is more ambiguous (as in Experiment 2), we can improve predictive performance over a broad range, but sacrifice model interpretability.

Table 4 Summary of Experiment 2

	% Selected									
Model (# features)	5	10	15	20	25	30	35	40	45	50
PCA/logit (22)	20.06	20.06	16.04	13.63	12.44	11.20	10.81	10.22	9.87	9.38
ELSA/logit (46)	23.04	18.09	15.56	13.79	12.13	12.04	10.97	10.54	10.03	9.53
ELSA/ANN (44)	19.58	17.55	16.40	14.42	13.13	11.96	10.97	10.40	9.98	9.64

Note. The hit rates of three different models are shown over the top 50% of prospects.

Figure 7 Lift Curves of Three Models That Maximize the Area Under Lift Curve when Targeting Up to Top 50% of Prospects



Note. In practice, we optimize over the first 25 intervals which have the same width, 2%, to approximate the area under the lift curve.

5. Interpretability, Time Complexity, and Scalability

We have shown that our ELSA/ANN model is a promising approach to database-based marketing programs. In this section, we address three important issues that should be considered before applying the ELSA/ANN approach to real-world marketing programs: interpretability, time complexity, and scalability.

The ELSA/ANN model improves the interpretability of the resulting classifier by constructing a parsimonious set of input variables that drive the response. We noted in our empirical work that the variables selected to predict insurance-buying behavior are largely consistent with the view that past-purchase information is effective in forecasting future buying behavior. By itself, the ELSA/ANN model has limited capability to explain why the chosen features predict customers' response to a specific marketing campaign. This is due to the fact that the underlying classifier (an ANN) is essentially a black-box algorithm.

In discussing model interpretability, it is useful to keep in mind a key trade-off in developing a predictive model. From the standpoint of firm profitability, the marketing manager should always use the model with the highest predictive accuracy, even if it is based upon an ANN structure. The superior performance of ELSA/ANN in our empirical work suggests that the proposed model will be a useful predictive tool. However, in explaining the model to a manager, it may be necessary to undertake additional work (say, using regression or decision trees) to understand more fully how the procedure works. The advantage of the ELSA/ANN algorithm is that the number of features that must be considered in this additional work is severely constrained. The reduction of the feature space aids managers by focusing attention on a small number of key inputs, and aids the researcher by allowing a relatively simple post hoc analysis. The parsimony of the ELSA/ANN optimal feature set is a major step forward in model interpretability.

Another important aspect related to the interpretability is whether or not independent runs of the ELSA/ANN model result in the same feature subsets. Like other GAs, ELSA is a stochastic algorithm. Therefore, it is very likely that we will observe different feature subsets from each run of ELSA. This is due to the fact that strong correlations typically exist among the input variables that are available to the researcher. For example, "contribution to car policies" is highly correlated with "number of car policies," so that one of these features could easily substitute for the other in a given model. However, even if different feature

subsets were selected by ELSA, we would expect that performance in terms of the estimated hit rate would be stable. This claim is supported by the small standard deviation (0.146%) of the hit rate, as estimated by cross-validating the training set.

In terms of time complexity, the ELSA/ANN model is expensive. This is mainly because the ELSA/ANN model is an example of the wrapper approach to feature selection that evaluates many different models and finally selects the best model from evaluated models. Typically, the time complexity of ELSA/ANN is k times more expensive than that of models without feature selection (e.g., a single ANN with the complete set of features), where k represents the number of models evaluated in ELSA/ANN. The fact that the ELSA/ANN model is computationally expensive raises the question of whether the ELSA/ANN approach can be applied to much larger data sets.

To study the scalability of our approach, we applied the ELSA/ANN algorithm to a large database containing information on donations to a national veterans organization.[10] These data are large in terms of record size (95,412 and 96,367 records for learning and evaluation purposes, respectively) and in terms of the dimensionality of the feature space (481 fields of numerical and categorical features). We first preprocess the data to eliminate all multicategorical fields, resulting in 406 features. To scale the method to a data set of this size, we use a small sample of available records for feature selection because the most computationally expensive procedure in the wrapper approach is feature selection. We therefore divide the training data set into two parts: one with a randomly chosen 5,412 records with which to select relevant features through our ELSA/ANN approach, and the other with the 90,000 remaining records to train the ANN with the chosen feature subset. Finally, we apply the trained ANN to an evaluation data set (96,367 records) to compute the hit rate after selecting the top 20% of customers based on their probability of donating money.

For the feature selection process, we set the values of the ELSA parameters identically with Experiment 2 except $T = 2{,}000$. Our experiment took about 40 hours on a Window XP machine with an Intel Pentium 4 processor at 2.2 GHz and 512 MB system memory. Based on accuracy estimates, we chose the solution with less than 10% of the original features that has the highest estimated accuracy. The chosen solution has 40 features and returns an average hit rate of 8.19% based on 10 independent runs on the evaluation data set, increasing by 19.3% the hit rates of single ANN with the complete set of features (average hit rate 6.87%). Because previous published studies on this data set attempted to maximize the net revenue rather than the hit rate of the fund-raising campaign, direct comparison of results is not possible. Nevertheless, this example demonstrates that the ELSA/ANN approach developed here can be adapted for the analysis of large customer databases.

[10] The data are publicly available at http://kdd.ics.uci.edu/databases/kddcup98/kddcup98.html.

6. Conclusion

In this paper, we presented a novel approach for customer targeting in database marketing. We used an evolutionary algorithm, ELSA, to search for possible combinations of features and an artificial neural network (ANN) to score customers. When the decision rule was precise, the overall performance of ELSA/ANN was superior to the industry standard PCA/logit model in terms of both accuracy and interpretability. However, this superiority in interpretability is confined to specific decision conditions defined during model development and calibration. Under a more general decision scenario, ELSA/ANN yielded a more accurate model over a broad selection percentage range at the cost of increasing the number of predictive features in the specification.

One of the clear strengths of the ELSA/ANN approach is its ability to construct predictive models that reflect the direct marketer's decision process. Unlike a standard statistical approach like PC/logit, the ELSA/ANN procedure can be easily modified to take into account different objectives. With information of campaign costs and profit per additional actual customer, a direct marketer could use ELSA/ANN to choose the best selection point where expected total revenue is maximized. In this way, it would be possible to determine the type of decision rule that the marketer should adopt, both in terms of solicitation percentage as well as predictive rule. Because all mailing lists do not have the same potential for the marketer, this approach would allow a predictive model and solicitation-mailing rule to be customized as the firm's database changes.

Our work also provides additional evidence that there exist strong dependencies between model specification and managerial decision making. When managers are clear about how a model will be used, the analyst can construct a highly specialized model that does better than general approaches (such as PC/logit). When managers are vague, a less parsimonious model can be constructed that does better under some region of the decision space. The ELSA/ANN approach provides a new tool in which these tradeoffs can be understood in the context of direct mail marketing applications.

Acknowledgments

The authors thank Peter van der Putten and Maarten van Someren for making the CoIL data available for this paper.

This work was partially supported by NSF grant IIS-99-96044.

References

Ahalt, S. C., A. K. Krishnamurthy, P. Chen, D. E. Melton. 1990. Competitive learning algorithms for vector quantization. *Neural Networks* **3** 277–290.

Balakrishnan, P. V. S., M. C. Cooper, V. S. Jacob, P. A. Lewis. 1996. Comparative performance of the FSCL neural net and *K*-means algorithm for market segmentation. *Eur. J. Oper. Res.* **93**(10) 346–357.

Berger, P. D., N. Nasr. 1998. Customer lifetime value: Marketing models and applications. *J. Interactive Marketing* **12** 17–30.

Bhattacharyya, S. 2000. Evolutionary algorithms in data mining: Multi-objective performance modeling for direct marketing. *Proc. 6th Internat. Conf. Knowledge Discovery Data Mining* (*KDD*-00), ACM, New York, 465–473.

Bult, J. R., T. Wansbeek. 1995. Optimal selection for direct mail. *Marketing Sci.* **14**(4) 378–394.

David Sheppard Associates, Inc. 1999. *The New Direct Marketing: How to Implement a Profit-Driven Database Marketing Strategy*. McGraw-Hill, Boston, MA.

Deb, K., J. Horn. 2000. Special issue on multi-criterion optimization. *Evolutionary Comput. J.* **8**(2).

DeSarbo, W. S., V. Ramaswamy. 1994. CRISP: Customer response based iterative segmentation procedures for response modeling in direct marketing. *J. Direct Marketing* **8**(3) 7–20.

Dy, J. G., C. E. Brodley. 2000. Visualization and interactive feature selection for unsupervised data. *Proc. 6th Internat. Conf. Knowledge Discovery Data Mining* (*KDD*-00), ACM, New York, 360–364.

Gath, I., A. B. Geva. 1988. Unsupervised optimal fuzzy clustering. *IEEE Trans. Pattern Anal. Mach. Intelligence* **11**(7) 773–781.

Goldberg, D. E. 1989. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, New York.

Goldberg, D. E., J. Richardson. 1987. Genetic algorithms with sharing for multimodal function optimization. *Proc. 2nd Internat. Conf. Genetic Algorithms*, Lawrence Erlbaum, Hillsdale, NJ, 41–49.

Gönül, F., M. Z. Shi. 1998. Optimal mailing of catalogs: A new methodology using estimable structural dynamic programming models. *Management Sci.* **44**(9) 1249–1262.

Horn, J. 1997. Multicriteria decision making and evolutionary computation. *Handbook of Evolutionary Computation*. Institute of Physics Publishing, London, U.K.

Hruschka, H., M. Natter. 1999. Comparing performance of feedforward neural nets and *K*-means for market segmentation. *Eur. J. Oper. Res.* **114**(2) 346–353.

Ishibuchi, H., T. Nakashima. 2000. Multi-objective pattern and feature selection by a genetic algorithm. D. Whitley, D. Goldberg, E. Cant-Paz, L. Spector, I. Parmee, H.-G. Beyer, eds. *Proc. Genetic Evolutionary Comput. Conf.* (*GECCO'*2000), Morgan Kaufmann, San Francisco, CA, 1069–1076.

Johnson, R. A., D. W. Wichern. 1992. *Applied Multivariate Statistical Analysis*, 3rd ed. Prentice Hall, Englewood Cliffs, NJ.

Kim, Y., W. N. Street. 2000. CoIL challenge 2000: Choosing and explaining likely caravan insurance customers. Technical report 2000–09, Leiden Institute of Advanced Computer Science, Sentient Machine Research, Amsterdam, The Netherlands.

Kim, Y., W. N. Street, F. Menczer. 2000. Feature selection in unsupervised learning via evolutionary search. *Proc. 6th Internat. Conf. Knowledge Discovery Data Mining* (*KDD*-00), ACM, New York, 365–369.

Krishna, K., M. N. Murty. 1999. Genetic *K*-means algorithm. *IEEE Trans. Systems, Man, Cybernetics—Part B: Cybernetics* **29**(3) 433–439.

Ling, C. X., C. Li. 1998. Data mining for direct marketing: Problems and solutions. *Proc. 4th Internat. Conf. Knowledge Discovery Data Mining* (*KDD*-98), ACM, New York, 73–79.

Menczer, F., R. K. Belew. 1996. Latent energy environments. R. K. Belew, M. Mitchell, eds. *Adaptive Individuals in Evolving Populations: Models and Algorithms*. Addison Wesley, Reading, MA.

Menczer, F., M. Degeratu, W. N. Street. 2000a. Efficient and scalable Pareto optimization by evolutionary local selection algorithms. *Evolutionary Comput.* **8**(2) 223–247.

Menczer, F., W. N. Street, M. Degeratu. 2000b. Evolving heterogeneous neural agents by local selection. V. Honavar, M. Patel, K. Balakrishnan, eds. *Advances in the Evolutionary Synthesis of Intelligent Agents*. MIT Press, Cambridge, MA.

Opitz, D. 1999. Feature selection for ensembles. *Proc. 16th National Conf. Artificial Intelligence* (*AAAI*), AAAI Press/MIT Press, Cambridge, MA, 379–384.

Pan, Z., X. Liu, O. Mejabi. 1997. A neural-fuzzy system for forecasting. *J. Comput. Intelligence Finance* **5**(1) 7–15.

Rao, V. R., J. H. Steckel. 1995. Selecting, evaluating and updating prospects in direct mail marketing. *J. Direct Marketing* **9**(2) 20–31.

Reinartz, W., V. Kumar. 2000. On the profitability of long-life customers in a non-contractual setting: An empirical application and implications for marketing. *J. Marketing* **64**(3) 17–35.

Riedmiller, M. 1994. Advanced supervised learning in multi-layer perceptrons—From backpropagation to adaptive learning algorithms. *Internat. J. Comput. Standards Interfaces* **16**(5) 265–278.

Rossi, P. E., R. McCulloch, G. Allenby. 1996. The value of household information in target marketing. *Marketing Sci.* **15**(3) 321–340.

Saad, E., D. Prokhorov, D. Wunsch. 1998. Comparative study of stock trend prediction using time delay, recurrent and probabilistic neural networks. *IEEE Trans. Neural Networks* **9**(6) 1456–1470.

Sarle, W. S. 1994. Neural networks and statistical models. *Proc. 19th Annual SAS Users Group Internat. Conf.*, SAS Institute, Cary, NC, 1538–1550.

Schmid, J., A. Weber. 1998. *Desktop Database Marketing*. NTC Business Books, Lincolnwood, IL.

Schmittlein, D. C., R. A. Petersen. 1994. Customer base analysis: An industrial purchase process application. *Marketing Sci.* **13**(1) 41–67.

Wilson, R. L., R. Sharda. 1994. Bankruptcy prediction using neural networks. *Decision Support Systems* **11**(3) 545–557.

Winer, R. S. 2001. A framework for customer relationship management. *California Management Rev.* **43**(4) 89–105.

Yang, J., V. Honavar. 1998. Feature subset selection using a genetic algorithm. *IEEE Intelligent Systems Appl.* **13**(2) 44–49.