Modelling Customer Retention with Statistical Techniques, Rough Data Models, and Genetic Programming

A. E. Eiben[1] *T. J. Euverman*[2] *W. Kowalczyk*[3⋆] *F. Slisser*[4]

[1] Dept. of Comp. Sci., Leiden University
e-mail: gusz@wi.leidenuniv.nl

[2] Faculty of Economics and Econometry, University of Amsterdam
e-mail: teije@fee.uva.n l

[3] Dept. of Math. and Comp. Sci., Vrije Universiteit Amsterdam
De Boelelaan 1081A, 1081 HV Amsterdam, The Netherlands
e-mail: wojtek@cs.vu.nl

[4] Strategic Management and Marketing, University of Amsterdam
e-mail: slisserf@fee.uva.nl

Abstract

This paper contains results of a research project aiming at modelling the phenomenon of customer retention. Historical data from a database of a big mutual fund investment company have been analyzed with three techniques: logistic regression, rough data models, and genetic programming. Models created by these techniques were used to gain insights into factors influencing customer behaviour and to make predictions on ending the relationship with the company in question. Because the techniques were applied independently of each other, it was possible to make a comparison of their basic features in the context of data mining.

1 Introduction

Banks, as many other companies, try to develop a long-term relationship with their clients. When a client decides to move to another bank it usually implies some financial loses. A climbing defection rate is namely a sure predictor of a diminishing flow of cash from customers to the company–even if the company replaces the lost customers–because older customers tend to produce greater cash flow and profits. They are less sensitive to price, they bring along new customers, and they do not require any acquisition or start-up costs. In some industries, reducing customer defections by as little as five percents can double profits, Reichheld (1996). Customer retention is therefore an important issue.

⋆ The corresponding author

To be able to increase customer retention the company has to be able to predict which clients have a higher probability of defecting. Moreover, it is important to know what distinguishes a stopper from a non-stopper, especially with respect to characteristics which can be influenced by the company. Given this knowledge the company may focus their actions on the clients which are the most likely to defect, for example, by providing them extra advice and assistance. One way of obtaining such knowledge is analysis of historical data that describe customer behaviour in the past.

In our research, which was carried out in a cooperation with a big mutual fund investment company [1] we have analyzed a fragment of a database containing information about more than 500.000 clients. In our analysis we have used three different techniques: logistic regression, e.g., Hair et al. (1995), rough data models, Kowalczyk (1996a), and genetic programming, Koza (1992).

Logistic regression is a well-known, "classical" method of analyzing data and requires no further explanations.

Rough data models have been introduced recently by Kowalczyk (1996a, 1996b). They consist of a simple partitioning of the whole data set, an ordering of elements of this partition and some cumulative performance measures. In a sense rough data models can be viewed as an extension of the concept of rough classifiers, Lenarcik and Piasta (1994).

Genetic Programming, introduced by Koza (1992), is also a relatively new technique which is based on evolutionary principles. It aims at finding complex expressions which describe a given data set as good as possible (with respect to a predefined objective criterion).

All the techniques were applied to the same data set independently, providing us, in addition to the main objective of the project (analysis of retention), a unique opportunity of comparing their various features (accuracy, comprehensibility of results, speed, etc.).

Our research was carried out in different phases, similarly to earlier projects, Eiben et al. 1996:

1. defining the problem and designing conceptual models with particular attention to relevant variables;
2. acquiring and arranging data;
3. exploratory data analysis;
4. building models by three techniques;
5. analysis and interpretation of the obtained models.

The organization of this paper reflects the order of these steps. In the next section we describe the problem and the available data. Sections 3-5 describe results obtained by the three techniques. In section 6 we discuss all the results and compare the techniques.

[1] For confidentiality reasons we are not allowed to disclose the name of this company. Also other details like the exact meaning of some variables and their actual values are not given.

2 Problem and Data Description

The company collects various data about their clients since many years. On the basis of these historical data we were supposed to investigate the following issue:

> *What are the distinguishing (behavioural) variables between investors that ended their relationship with the company (stoppers) from investors that continued (non-stoppers) and how well can different techniques/models predict that investors will stop the relation within the next month.*

The company offers at this moment about 60 different investment forms which attract customers with different profiles. Due to this diversity of clients and investment forms we had to restrict our research to a homogeneous group of clients that invest money in a specific form. In particular, we have focused on clients which were "real investors" (i.e., clients which had only a simple savings account or a mortgage were not considered). Further, we restricted our attention to clients that stopped their relation between January 1994 and February 1995 (14 possible "stop months). These restrictions led to a data set with about 7.000 cases (all stoppers). As we were interested in discriminating stoppers from non-stoppers, the data set has been extended by about 8000 "non-stopper" cases. Each record in the dataset contained the history of a single client over a period of 24 months before the moment of stopping (for non-stoppers dummy "stop-moments" were generated at random). By a "history" we mean here sequences (of length 24) of values which represent various measurements like the number of transactions, monthly profit, degree of risk, etc. Additionally, some "static" variables were stored, e.g., client's age, starting capital, etc. The dataset we finally extracted from databases consisted of 15.000 records, each record having 213 fields. Some of the most relevant variables are listed in Table 1.

3 Statistical analysis of data

When statistical methods are applied on very large data sets, the emphasis is on the explanatory significance rather than statistical significance. For example, a correlation of .001 can be statistically significant without having any explanatory significance in a sufficiently large data set. Statistical estimation becomes computation of meaningful statistics. In this section we shall describe the sequence of actions which we undertook in order to arrive at an intelligible model.

3.1 Data reduction

We separately analysed each set of dynamic attributes trying to reduce the number of variables in each set. Using growth curve analysis, see, e.g., Timm (1975), we tried to discover different average polynomial trends for the two groups of stoppers and non-stoppers in order to retain only those that were discriminating between groups. Necessary for the existence of such salient components is a sufficiently large multivariate difference between the group averages. A meaningful

statistic is Wilks' $1-\Lambda$, which ranges from zero to one. It may be conceived as a multivariate generalisation of $1-R^2$. A value near zero means that the difference between groups is negligible compared to the differences within groups. In that case there is no gain in information when two groups are distinguished instead of envisaging just one group of clients with stoppers and non stoppers mixed together. The values of $1-\Lambda$ of the sets of dynamic variables ranged from .003 to .06. These meaningless magnitudes led us to consider only aggregate values over time of each set.

3.2 Univariate exploration

Inspection of the outcomes of standard data exploration techniques has led us to categorise some of the aggregated variables in order to enhance the interpretabily. For this categorisation we took into consideration the distributional characteristics as well as the domain of content. For the next five variables we explain our categorisation.

investments
One category was formed for about 35% of the clients for which this variable did not change over time. The remaining five categories were based on quintiles.

risk
One category was formed for about 25% of the clients for which the variable did not change over time. Quintiles were used for the remaining clients.

number of transactions A
One category was formed for about 40% of the clients for which there were no transactions of this type over time. Almost all of the remaining clients had a mean number of such transactions within the interval $(0, 1]$. A further subdivision was therefore not considered as meaningful. So a two-category variable was constructed.

number of transactions B
A two-category variable similar to the preceding one.

funds
For almost 60% of the clients for which the number of funds did not change over time five categories were formed corresponding to 0, 1, 2, 3, 4. The sixth category was formed of the clients that varied their number of funds over time. The quintiles in this remaining group formed no meaningful separation (resp. .522, .870, 1.435, 2.391).

We verified the meaningfulness of the recodings in two ways. Firstly, by a verification of non-uniformness of each bar chart with each bar representing the relative frequency of stoppers. Secondly, by a two-dimensional correspondence analysis (Krzanowski 1993) in order to study the placing of the categories in one-dimensional space in relation to stoppers and non-stoppers categories.

3.3 Simple logistic regressions

We carried out ten logistic regressions, one for each variable. We used three indices for judging the results:

1. $R^2_{logistic\ regression}$, which can be interpreted as the proportional reduction of the lack-of-fit by incorporating the variable of interest above a model based only on the intercept parameter (Agresti, 1990). It is defined as:

$$R^2_{logistic\ regression} = 1 - \frac{log(likelihood_{intercept+variable}))}{log(likelihood_{intercept})}.$$

2. $\lambda_{logistic\ regression}$, which can be interpreted as the proportional reduction of errors in classification by incorporating the variable of interest above a model with only an intercept. A model with only an intercept classifies all clients in the group with the largest observed frequency, being the non-stoppers in the present case. This means that all stoppers are misclassified and regarded as errors (see, e.g., Menard, 1995). So it is defined as:

$$\lambda_{logistic\ regression} = \frac{\#errors_{intercept} - \#errors_{intercept+variable}}{\#errors_{intercept}}.$$

3. $\gamma_{logistic\ regression}$, which can be interpreted as a measure of ordinal association between the predicted probabilities of being a stopper and actually being a stopper. The measure is widely used for cross-tabulations. It was proposed by Goodman and Kruskal (1954). It measures a weak monotonicity and ignores ties. It is easily interpretable as it ranges from -1 to +1. It is based on the number of concordant pairs C and discordant pairs D. A pair (non-stopper, stopper) is concordant when the predicted probability for a non-stopper is lower than for a stopper and disconcordant in the reverse case (see , e.g., Coxon (1982) for a discussion on measures for association). It is defined as:

$$\gamma_{logistic\ regression} = \frac{C - D}{C + D}.$$

An overview of the results is given in Table 1.

Except for the categorised aggregated dynamic variables the Table displays bad results for R^2 and λ. These figures gave us reasons to try to improve the results. We decided to do a quintile-categorisation for all variables for which R^2 and λ were zero in two decimals. A quintile-categorisation for duration of investment relation did not produce interpretative results in the sense that the differences between the values were too small to justify an interpretation in a scale running from “extremely short” to “extremely long” for example. The results of the categorisations are displayed in the next Table.

As can be seen in Table 2, slight improvements were achieved.

Table 1. An overview of results.

	R^2	λ	γ
investments	.07	.10	.42
risk	.05	.11	.35
transactions A	.02	.00	.32
transactions B	.08	.18	.59
funds	.04	.05	.37
profit A	.00	.00	.18
profit B	.00	.00	.11
emotion index A	.00	.00	.31
emotion index B	.00	.00	.34
duration of relation A	.02	.00	.15
duration of relation B	.00	.00	.06
starting capital	.00	.00	.31

Table 2. Results of quintile categorization.

	R^2	λ	γ
profit A	.02	.00	.24
profit B	.01	.00	.14
emotion index A	.05	.14	.31
emotion index B	.06	.17	.37
duration of relation A	.01	.00	.13
starting capital	.04	.07	.31

3.4 Multiple logistic regression

In the final model we used only one profit and one emotional variable in order to prevent redundancy in the model and for reasons of interpretation. We used *profit B* and *emotion index A* because they had the best performance in the univariate regressions. Consequently, ten variables were put into the logistic regression model. The use of categories for most of the variables means that choices for interesting contrasts were made possible. We do not go into details here, because this article is mainly meant to compare different methods of modelling on the same data set. It is possible to improve (slightly) the predictive power of a logistic regression, but the main advantage of classical techniques in the present context is to build an interpretative model with sufficient predictive power. Note that individual statistical significance of each variable or their categories is of less importance, although one can state that non-significant results for such large data sets may cast doubts on the importance about the inclusion of the variable. The following values of the indices were obtained for this training set: $R^2 = .15, \lambda = .25, \gamma = .52$. Another useful statistic is also the Spearman's correlation coefficient ρ between the predicted probabilities and observed relative frequencies. We categorised the predicted probabilities using percentiles. Within each category the observed relative frequency was computed

and between these 100 values the value of n was calculated. In the training set $\rho = .974$. In the validation set we obtained the following values for the various statistics: $\lambda = .24, \gamma = .53, \rho = .923$.

4 Analysis of data with Rough Data Models

This section contains a brief presentation of the concept of Rough Data Models, Kowalczyk (1996a, 1996b), and results our experiments. A more detailed description of these experiments can be found in Kowalczyk and Slisser (1997).

4.1 Rough Data Models

Informally, a Rough Data Model consists of a collection of clusters that form a partition of the data set, some statistics calculated for every cluster (e.g., cluster size, number of elements of specific type), and a linear ordering on clusters. This ordering is supposed to reflect cluster importance and is used for calculating various cumulative performance measures. To define the concept of RDM more formally we need some notation and terminology used in the theory of rough sets, Pawlak (1991). Let us consider a decision table

$$\mathbf{T} = (\mathbf{U}, \mathbf{A}, \mathbf{d}),$$

where U is a finite collection of objects (the universe), $A = \{a_1, \ldots, a_k\}$ is a set of attributes on U, i.e., every a_i is a function from U into a corresponding set of attribute values $V_i, a_i : U \to V_i$, for $i = 1, \ldots, k$, and d is a decision function which takes values in a finite set of decisions $D = \{d_1, \ldots, d_n\}, d : U \to D$. Elements of U are often called *patterns* and associated decision values types, thus if $d(u) = d_1$ then u is called a pattern of type d_1. Let R denote the indiscernibility relation which is defined by the set of attributes A, i.e., for any $u_1, u_2 \in U, R(u_1, u_2)$ iff $a_i(u_1) = a_i(u_2)$, for $i = 1, \ldots, k$. The relation R determines a partition of U into a number of (pairwise disjoint) equivalence classes $C_1, \ldots, C_m$, which will further be called *clusters*. Every cluster may contain elements of different types. However, elements that belong to the same cluster are, by definition, not distinguishable, so they will be classified (by any classifier) as elements of the same type. Therefore, any classifier is determined by assigning to every cluster C its type, $class(C)$, which is an element of D. Given a partitioning of the universe and a classification function $class$, a number of useful parameters which characterise clusters can be introduced:

- cluster size, $size(C_i)$, which is just the number of elements of C_i,
- number of elements of a given type, $size(C_i, d_j)$, which is the number of elements of type d_j that are members of C_i,
- number of correctly classified elements, $corr(C_i)$, which is the number of elements of C_i which are of type $class(C_i)$,
- cluster accuracy, $accuracy(C_i)$ which is defined as the ratio $corr(C_i)/size(C_i)$.

These parameters can be used for ranking clusters according to some, user specified, criteria. For example, clusters might be ordered according to their size (the bigger the better), according to their accuracy or according to the percentage of elements of specific type.

Now we can formally define a *rough data model* of a decision table $\mathbf{T} = (U, A, d)$ as a triple:

$$\mathbf{T} =< C, class, \leq >,$$

where

- C is a set of clusters,
- $class : C \rightarrow D$ is a function that assigns to every cluster its type,
- $\leq$ is a linear ordering on C.

Performance of rough data models can be measured in many different ways, Kowalczyk (1996a). In addition to some problem independent measures like cumulative accuracy, gain curves, response curves, etc., one can introduce problem specific measures, for example, the percentage of elements of specific type in "best" (in sense of the $\leq$ relation) clusters which cover 10% of all cases.

There are two important features of RDMs:

1. there are almost no restrictions on the form of performance measure which is used for evaluating model quality; this measure is defined by the user and is problem dependent,
2. computational complexity of generating RDMs is very low (linear in the size of the data set); this feature allows for exploring huge number of alternative RDMs and focusing on these models that optimise the given performance criterion.

In practice, the process of generating high quality models consists of three major steps:

1. formulation of a performance measure that should be optimised (e.g., classification rate, percentage of correctly classified cases of the given type in specific fragment of the model, total misclassification cost, etc.).
2. determination of a search space, i.e., a collection of models which should be searched to find an optimal one (for example, a collection of models which are based on k attributes which are taken from a set of n attributes, or a collection of models determined by various discretization procedures, etc.)
3. determination of a search procedure (for example, exhaustive search, local search, branch & bound, etc.)

Usually rough data models are used as an efficient tool which helps to get an insight into data sets. The user first specifies some objective function, then proposes a number of data transformations, formulates some restrictions on model complexity (e.g., "the model should be based on at most four attributes") and then models which satisfy all these criteria are automatically generated and evaluated. In spite of its simplicity, this approach often provides models which have relatively high accuracy.

4.2 Retention and Rough Data Models

To get some idea about the importance and relationships between various attributes a number of standard tests were carried out. First of all, we have generated numerous plots which are routinely used in statistical data analysis: frequency histograms, means, density estimates, etc., see Hair et al. (1995). Visual inspection of these plots led to the discovery of a large group of clients (4809) which behaved differently from the rest. Therefore, we decided to split the whole data set into two subsets and analyse them independently. We will refer to both groups as to A-clients and B-clients. In order to identify most important attributes we have calculated, for every attribute, values of three importance measures: correlation coefficients, coefficients of concordance and information gain. Correlation coefficients measure linear dependency between attributes, are widely used and require no further explanations. Coefficient of concordance (sometimes called the CoC index or just the c index) measures the degree of similarity of an ordering (of all cases) which is induced by values of the measured attribute and the ordering induced by the decision attribute. Information gain measures the amount of information provided by a (discrete-valued) attribute and is explained in Quinlan (1986). As a result of this analysis we have identified 8 attributes which were used as the basis for construction of RDMs. To guide the search process we had to specify some performance measure that should be optimized. After some discussions with bank experts we took the percentage of stoppers that can be found in the top 10% of cases as our objective function. We have restricted our attention to models that were based on all combinations of 2, 3 or 4 attributes taken from the set of 8 important attributes mentioned in section 3. Each attribute has been discretized into 5 intervals, according to the 'equal frequency' principle. Unfortunately, a model which is based on 4 variables which are discretized into 5 intervals may have 5*5*5*5=625 clusters–too many to expect good generalisation. Therefore, we allowed each attribute to be split into 3 intervals only; ends of these intervals were taken from 6 points determined by the discretization into 5 intervals. Thus every attribute could be partitioned into 15 ways, which leads to 15*15*15*15=50.625 various models which are based on 4 variables. Moreover, there are 70 ways of selecting 4 attributes out of 8, so the total number of models based on 4 variables is about 3.5 million; adding models which are based on 2 or 3 variables does not increase this figure too much. Due to computational simplicity of RDMs we could systematically generate all these models, evaluate them and select the best one. It turned out that the performance of best models which were based on 4 attributes was almost the same as of models based on 3 attributes. Moreover, in both groups of models there were several models which were very close to the optimal ones. All these models have been carefully analysed on basis of their performance curves and the structure of clusters. Figure 6.2 contains plots of response curves which are based on best models. Clusters, together with their definitions (formulated in terms of values of attributes which determine them) can be used for formulating some rules about the data. For example, the best cluster from the model of B-clients captured clients who were investors for a long time, invested money in

funds with very small risk, and got small profits-all of them have stopped their relation with the company. Clearly, a detailed analysis of all clusters provided a good insight into customer behaviour.

Additionally, the models have been tested on an independent validation set in order to evaluate their generalisation capabilities. Not surprisingly (models based on 3 attributes had only 27 clusters), they generalised very well (performance dropped less than 1%).

4.3 Rule extraction

As mentioned above, clusters which are determined by best models can be directly translated into decision rules. However, such rules do not cover large fragments of the model. In order to identify some general rules we have run a systematic search algorithm which generated rules in the form

$$\textbf{if} \quad (a < X_1 < A)\&(b < X_2 < B)\&(c < X_3 < C) \quad \textbf{then} \quad \textit{decision}$$

(where X_1, X_2 and X_3 are attribute names and a, A, ..., c, C are some numbers), and tested them in terms of the number of covered cases and accuracy. The search process was restricted to rules such that:

1. attributes X_1, X_2 and X_3 were arbitrary combinations of attributes taken from the set of 8 most important attributes,
2. splitting points a, A, ..., c, C were determined by an 'equal frequency' discretization of the corresponding attributes into 7 intervals: they could be chosen from the set of ends of these intervals,
3. rules were allowed to involve only 2, 3, 4 or 5 'splitting points'.

Out of several million rules generated in this way (only for the group of B-clients) we have focused on rules which were 'interesting' in the following sense: they had to cover at least 10% of all cases and had accuracy at least 80% (i.e., at least 80% of all cases which were covered by the rule had to be 'stopper'-cases). The resulting collections of rules were relatively small (1, 24, 98 and 132 rules which involved 2, 3, 4 and 5 splitting points, resp.). A similar collection of rules has been found for A-clients. All rules have been carefully analysed by experts and their analysis led to the discovery of some interesting patterns in customer behaviour.

5 Genetic Programming

Genetic Programming is a new search paradigm which is based on evolutionary principles, Koza (1992). Potential solutions (individuals) are represented by (usually complex) expressions which are interpreted as definitions of functions (models). The quality of individuals (fitness) is measured by evaluating performance of the corresponding models (e.g., classification rate). The search process mimics the evolution: a collection of individuals (population) "evolves" over time

and is subjected to various genetic operators. In our research we used the system OMEGA which is developed by Cap Volmac, Holland.

OMEGA is a genetic programming system that builds prediction models in two phases. In the first phase a data analysis is performed, during which several statistical methods are carried out to find the variables with the highest individual predictive power. In the second phase the genetic modelling engine initializes with a targeted first generation of models making use of the information obtained in the first phase. After initialization, further optimization then takes place with a fitness definition stated in terms of practical objectives.

In modelling applications the most frequently applied measure for evaluating the quality of models is 'accuracy', which is the percentage of cases where the model correctly fits. When building models for binary classification the so-called CoC measure is a better option for measuring model quality. The CoC (Coefficient of Concordance) actually measures the distinctive power of the model, i.e. its ability to separate the two classes of cases, see 1994 for the definition. Using the CoC prevents failures caused by accepting opportunistic models. For instance, if 90 % of the cases to be classified belongs to class A and 10 % to class B, a model simply classifying each case as A would score 90 % accuracy. When the CoC is used, this cannot occur. Therefore, we used the the CoC value of the models as fitness function to create good models. Let us note that selection of the best model happened by measuring accuracy (on the test set), and that the final comparison between the four different techniques was also based on the accuracy of the models. Yet, we decided to use the CoC in order to prevent overfitting and some control experiments (not documented here) confirmed that evolving models with a fitness based on accuracy results in inferior performance.

As stated previously, OMEGA builds prediction models in two phases. The first phase, a data analysis, selects the variables that are of interest out of all supplied variables. Of these selected variables a summarisation of their performance measured in CoC is shown in Table 3.

Table 3. Values of CoC for 6 variables

relation A	60.4
relation B	62.0
start capital	64.9
funds	62.7
investments	71.1
risk	70.7

By the data analysis carried out in the first phase OMEGA is able to create a good initial population by biasing the chances of variables to be included in a tree. The initial models were generated into 2 populations of 20 expressions each. During this initialization, the 40 models varied in performance from 72.3 till 79.9 measured in CoC. Computing the accuracy of the best model on the training

set gave the value of 73%. The relatively good CoC values after initialization show an increased joint performance compared to the best individual variable performance (which was 71.1 for *investments*). This is due to the use of special operators that act on the variables and the optimised interactions between the selected variables in the models. Till sofar, no genetic optimization has taken place and several satisfactory models have been found.

During the genetic search we were using 0.5 crossover rate, 0.9 mutation rate and 0 migration rate between the two sub-populations, that is no migration took place. The two populations were only used to maintain a higher level of diversity. The maximum number of generations was set to 2000. After the genetic search process the best performing tree had a CoC value of 80.8 and a corresponding accuracy of 75% on the training set. For this particular optimization problem the genetic optimization phase does not show a dramatic improvement in performance as the initialization did.

6 Analysis of the results

In this section we evaluate the outcomes from two perspectives. Firstly, we will concentrate on the original problem of modelling customer behavior. Secondly, we compare the predictive power of our genetically created model to the best models the other techniques obtained. In this analysis other models are competitors of the genetic model.

6.1 Interpretation of the models

When evaluating the results it is important to keep in mind that the company providing the data is not only interested in good predictive models, but also in conclusions about the most influential variables. Namely, these variables belong to customer features that have the most impact on customer behaviour. In case of models built with logistic regression the interpretation of results was straightforward: we simply had to look into coefficients involved in these models. Rough data models were also easy to interpret: the meaning of cluster characteristics and extracted rules was obvious. The situation was a bit more complicated with genetic programming. The resulting models (complex expressions) were very difficult to interpret. Therefore we performed a sensitivity analysis on the variables involved in the generated models by fixing all but one variables of the best model and varying the value of the free variable through its domain. The changes in the performance of the model are big for very influential variables, and small for variables with less impact. The results of the sensitivity analysis are given in Table 4, where a high value indicates a high influence, while lower values show a lower impact.

Comparing the interpretations of the different techniques, i.e. (dis)agreement on the importance of variables we observed a high level of agreement. Based on this observation it was possible to bring out a well-founded advice for the company that specific risk values in the portfolio substantially raises the chance

Table 4. Results of sensitivity analysis.

relation B	4.32%
relation A	14.37%
start capital	13.50%
funds	2.88%
investments	18.56%
risk	82.35%

of ending the business relationship. This conclusion allows the company to adapt its policy and perform directed actions in the form of advising customers to change the risk value of their portfolio.

6.2 Comparison of applied techniques

The three techniques can be compared to each other with respect to various criteria. In our study we have focussed on the following aspects: accuracy of the generated models, their interpretability, time needed for their construction and expertise required by each technique.

Accuracy. Model accuracy was measured by creating cumulative response rate tables for each technique. The results are presented in Figure 6.2. It can be noticed that the model provided by genetic programming was slightly better than the best rough data model. The best model produced with statistical techniques was much worse that the previous two. It should be noted that the best rough data model was based on 3 variables only, in contrast to 9 variables used in the model generated by logistic regression and 24 variables used by the "genetic model.

Interpretability. In all cases it was possible to provide an interpretation of the generated model. Logistic regression provided a list of most influential variables (or their combinations) together with their weights. The OMEGA system generated a complex formula that was too difficult to interpret. Instead, a sensitivity analysis provided a list of most significant variables. Rough Data Models provided the best insight into the analyzed data. They provided a lot of information about meaningful combinations of variables, lists of significant clusters and explicit rules.

Time. The computer time needed for generating models was different for different techniques. Calculations necessary for logistic regression took about 30 min., but a lot of conceptual work (about one week) was needed first. The time required by the other two techniques was significantly longer (a few days). Precise comparison is not possible because both systems (OMEGA and TRANCE) were running on different computers (PC and UltraSparc).

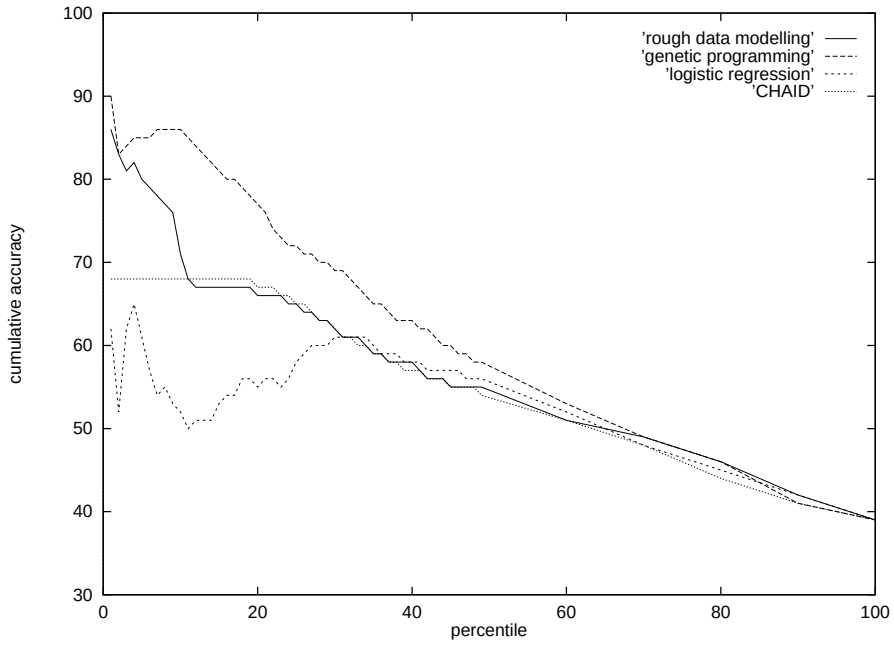

Fig. 1. Predictive power of the best models created by the four techniques

Necessary expertise. In case of logistic regression an extensive statistical knowledge and some acquaintance with a statistical package (in our case: SPSS) was indispensable. Almost no knowledge was necessary for building rough data models. However, due to the current status of the TRANCE system (a prototype implemented in MATLAB), a lot of programming work (scripting) and knowledge of the MATLAB system were necessary. In contrast, the use of the OMEGA system is very simple. This commercial tool has a friendly user interface and experiments can be run by a user with no extensive knowledge of genetic programming.

7 Concluding Remarks

In this paper we described an application oriented research project on applying different modelling techniques in the field of marketing. Our conclusions and recommendations can be summarized as follows.

- Cross-validation on the most influential variables based on models developed with other techniques raises the level of confidence. In our project we observed good agreement between conclusions of the three approaches.
- Non-linear techniques such as genetic programming and rough data modelling proved to perform better than linear ones with respect to the predictive

power of their models on this problem. This observation is in full agreement with the outcomes of an earlier comparative research on a different problem, see Eiben et al. 1996.
- It is advisable to use CoC as fitness measure in the GP. Control runs with accuracy as fitness led to models that had a worse accuracy than those evolved with CoC.
- Simple statistical data analysis can distinguish powerful variables. Using this information during the initialization of the GP works as an accelerator, by creating a relatively good initial population.
- It is somewhat surprising that even a long run of the GP could only raise the performance of the initial population by approximately 2 % in terms of accuracy on the training set. Note, however, that in financial applications one percent gain in predictive performance can mean millions of guilders in cost reduction or profit increase.

The global evaluation of the project is very positive. We have gain a good insight into the phenomenon of retention and constructed models with satisfactory accuracy. The company in question highly appreciated obtained results and decided on further research which should lead to an implementation of a system that could be used in the main business process.

Ongoing and future research concerns adding new types of variables, validation of the models for other time periods as well as developing models for a longer time horizon.

References

Agresti, A., 1990, *Categorical Data Analysis*, John Wiley: New York.

Coxon, A.P.M., 1982, *The User's Guide to Multidimensional Scaling*, Heinemann: London.

Goodman, L.A. and Kruskal, W.H., 1954, *Measures of Association for cross-classifications, Journal of the American Statistical Association*, 49, 732-764 and 54, 123- 163 and 58, 310-364.

Eiben, A.E., T.J. Euverman, W. Kowalczyk, E. Peelen, F. Slisser and J.A.M. Wesseling. Comparing Adaptive and Traditional Techniques for Direct Marketing, in H.-J. Zimmermann (ed.), In *Proceedings of the 4th European Congress on Intelligent Techniques and Soft Computing*, Verlag Mainz, Aachen, pp. 434-437, 1996.

Haughton, D. and S. Oulida, Direct marketing modeling with CART and CHAID, In *Journal of direct marketing*, volume 7, number 3, 1993.

Hair, J.F., Jr., Anderson, R. E., Anderson, T.R.L., Black, W. (1995). *Multivariate Data Analysis* (fourth edition), Prentice Hall, Englewood Cliffs, New Jersey.

Hosmer, D.W., and L. Lemeshow. (1989). *Applied logistic regression*, New York, Wiley.

Kowalczyk, W. and Slisser, F., (1997). Analyzing customer retention with rough data models. In *Proceedings of the 1st European Symposium on Principles of Data Mining and Knowledge Discovery*, PKDD'97, Trondheim, Norway, Lecture Notes in AI 1263, Springer, pp. 4–13.

Kowalczyk, W. (1996). TRANCE: A tool for rough data analysis, classification and clustering. In S. Tsumoto, S. Kobayashi, T. Yokomori, H. Tanaka and A. Nakamura

(eds.), *Proceedings of the Fourth International Workshop on Rough Sets, Fuzzy Sets, and Machine Discovery, RSFD'96*, Tokyo University, pp. 269–275.

Kowalczyk, W. (1996). Analyzing temporal patterns with rough sets. In *Proceedings of the Fourth European Congress on Intelligent Techniques and Soft Computing, EUFIT'96*, Volume I, Aachen, Germany, pp. 139–143.

Koza, J. (1992). *Genetic Programming*, MIT Press.

Krzanowski, W.J. (1993). *Principles of Multivariate Analysis: A User's Perspective*, Clarendon Press: Oxford.

Lenarcik, A. and Piasta, Z. (1994). Rough classifiers. In: Ziarko, W. (ed.), *Rough Sets, Fuzzy Sets and Knowledge Discovery*, Springer-Verlag, London, pp. 298–316.

Magidson, M. (1988). Improved statistical techniques for response modeling. In *Journal of direct marketing*, volume 2, number 4.

Menard, S. (1995), *Applied Logistic Regression Analysis*, Sage: London.

Pawlak, Z. (1991). *Rough Sets: Theoretical Aspects of Reasoning About Data*, Kluwer Academic Publishers, Dordrecht.

Quinlan, R. (1986). Induction of decision trees, *Machine Learning* 1, 81-106.

Reichheld, F.F. (1996). Learning from Customer Defections, in *Harvard Business Review*, march-april.

Timm, N.H. , (1975), *Multivariate Analysis with Applications in education and psychology*, Brooks/Cole:Monterey.

Walker, R. D. Barrow, M. Gerrets and E. Haasdijk. (1994). Genetic Algorithms in Business. In J. Stender, E. Hillebrand and J. Kingdon (eds.), *Genetic Algorithms in Optimisation, Simulation and Modelling*, IOS Press.

This article was processed using the LaTeX macro package with LMAMULT style