# Modelling Customer Retention with Rough Data Models

Wojciech Kowalczyk[1] and Frank Slisser[2]

[1] Dept. of Mathematics and Computer Science, Vrije Universiteit Amsterdam, De Boelelaan 1081A, 1081 HV Amsterdam, The Netherlands

[2] Strategic Management and Marketing, University of Amsterdam, Roetersstraat 11, 1018 WB Amsterdam, The Netherlands

**Abstract.** 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. Therefore, banks are very interested in identifying some mechanisms behind such decisions and determining clients that are about to leave the given bank. One way of getting such an insight is to analyse historical data that describe customer behaviour in the past.

In this paper we present a methodology and some results of an analysis of a large data set provided by a big mutual fund investment company. Our approach, based on the concept of Rough Data Model, [7], resulted in the identification of key factors that influence customer retention. Moreover, a number of rules that characterise various groups of clients have been generated. Our results have been highly appreciated by the company and led to specific actions aimed at increasing customer retention.

## 1 Introduction

One of the objectives of a mutual fund investment company is to increase its value. It can be achieved, for example, by increasing the cash flow, or, more specifically, by increasing the cash inflow (acquisition) and reducing the cash outflow (retention). For a healthy growth acquisition and retention efforts have to be brought into balance [1]. The cash outflow can be reduced, for example, by preventing clients from quitting their relationship with the company. 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 percent points can double profits [12]. Customer retention is therefore an important issue. To be able to increase customer retention the company has to be able to predict which clients have a higher probability of defecting. It is also necessary 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.

In our research, which was carried out in cooperation with a big mutual fund investment company, we tried to discover some factors (or their combinations) which discriminate between stoppers and non-stoppers and which are early warnings for customer defection. We have focused mainly on behavioural factors, i.e., factors which reflect ways of responding to changing situation. As a starting point for our investigations we have used a fragment (about 15.000 cases, each case characterized by a few hundred values) of a database containing information about more than 500.000 clients. For our analysis we have used the TRANCE system which supports the process of building Rough Data Models (RDM's) and their evaluation, [7]. In total we have generated a few million models and selected a couple of them. Moreover, a number of significant rules which characterize various groups of clients have been found. These rules provided a lot of useful information about the phenomenon of retention.

## 2 Data description

The company registers all available data about their clients since many years. In addition to data on financial transactions the company stores data about all communications with her clients, demographic profiles and some additional administrative data. Communications can be divided in several types corresponding to different media and the contents. The demographic profile of a client contains, amongst others, the date of birth, the gender and the family size. An example of administrative data is the date of subscription. Also a lot of data about various financial indicators is stored. The most important are the prices and the returns of the company's investment funds. 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.

In the first phase of the project we had to identify some attributes that we considered important. In total we have identified 5 attributes which were 'static' (client's age, starting capital, duration of relation, duration of the investment relation[1] and stop month) and 9 'dynamic' attributes (their values were changing over time): monthly profit, risk, investment index, number of funds, number of payments and withrawals, etc. Values of these attributes were calculated for every client for every month over a period of 2 years preceding the 'stop moment'. For the 'non-stoppers' this 'stop moment' was randomly generated as a month between January 1994 and

[1] Clients can begin the relation with the company by opening a savings account and start investing later. This difference is reflected in the two duration attributes.

February 1995. Thus each dynamic attribute has been represented by 24 ordinary attributes (one for each month). However, the period spanned for every client was not the same (there were 14 different periods corresponding to the different 'stop months'). The values for the last month (the potential stop moment) were not allowed to be used in our experiments: we were supposed to predict client behaviour (will (s)he stop or not?) in month 24 on the basis of the data for preceding 23 months. The resulting data table consisted of 14394 cases, each case having 5 + 9*23 = 212 (independent) attributes and one binary decision attribute. All attributes were numerical.

## 3 Important attributes

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 [3]. 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. This coefficient, introduced in late seventies, has been originally used for measuring the quality of so-called ROC-curves (Receiver Operating Characteristic curves). Recently, it has been successfully applied in the context of Neural Networks and Genetic Programming, [13]. Information gain, [11], measures the amount of information provided by a (discrete-valued) attribute and is calculated according to the formula:

$$gain(A) = I(p,n) - \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i,n_i),$$

where $p$ and $n$ denote the number of 'positive' and 'negative' cases in the data set; $p_i$ and $n_i$ refer to subsets of the whole data set which are determined by $v$ possible values of the attribute $A$, and $I(p,n)$ is given by:

$$I(p,q) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}.$$

Some of our attributes were not discrete so we had to discretize them (into eight 'equal frequency' intervals) before applying the above formulas. It should be noticed that by discretizing continues attributes we possibly lost some information. Nevertheless, the

measure based on information gain seems to be most suitable for our problem: it assumes no linear ordering of domains nor linear dependencies between attributes.

All three measures provided similar results; Figure 1 illustrates the importance of all dynamic attributes which was measured according to the information gain. Let us note that for A-clients the amount of information which is provided by most of dynamic attributes increases when we are getting closer to the last month before the stop moment. On the other hand, in the group of B-clients this phenomenon does not occur.

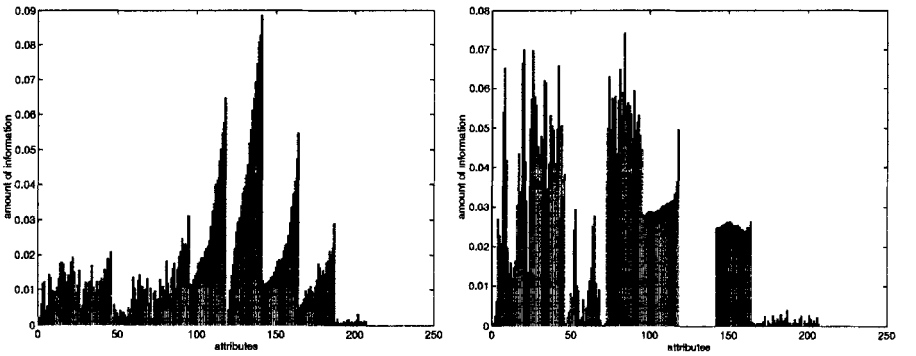


**Figure 1.** The amount of information provided by dynamic attributes for A-clients (left) and B-clients (right). Each dynamic attribute is represented by a group of 23 single valued attributes. Thus the first 23 bars correspond to the first dynamic attribute, bars 24-47 to the second one, etc.

In order to reduce data dimensionality we tried to aggregate some dynamic attributes (groups of 23 single attributes) by combining their values. For example, we tried to use weighted means, trends, coefficients of polynomials and several other ad-hoc invented combinations, [6]. Unfortunately, in all cases the amount of information provided by such 'combined' attributes was not significantly higher than the amount of information provided by the original attributes in month 23. Therefore, we decided to use in our experiments only values of dynamic attributes from this month. In this way the number of attributes was reduced to 14 (5 static and 9 dynamic taken one month before the stop moment). Finally, 6 attributes which provided least information were removed, yielding a final collection of 8 attributes.

## 4 Rough Data Models

For further analysis of the data we have used the concept of a Rough Data Model, RDM, introduced in [7]. 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, [9]. Let us consider a decision table

$T = (U, A, d)$,

where $U$ is a finite collection of objects (the universe), $A=\{a_1, ..., a_k\}$ is a set of attributes on $U$, i.e., every $a_i$ is a function from $U$ into a corresponding set of attribute values $V_i$, $a_i: U \longrightarrow V_i$, for $i = 1,..., k$, and $d$ is a decision function which takes values in a finite set of decisions $D=\{d_1, ..., d_n\}$, $d: U \longrightarrow D$. 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 $i=1, ..., k$. The relation $R$ determines a partition of $U$ into a number of (pairwise disjoint) equivalence classes $C_1, ...., C_m$, 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 $T = (U, A, d)$ as a triple:

$M = \langle C, class, \leq \rangle$, where

- $C$ is a set of clusters,

- *class*: $C \longrightarrow D$ 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, [7]. 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, [6, 7]:

(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–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.

## 5 Retention and Rough Data Models

In this section we will describe the process of building rough data models for our problem. As a starting point we had two decision tables (A-clients and B-clients) with 8 numerical attributes and one binary (stopper/non-stopper) decision attribute.

The objective of our project–identification of potential stoppers–almost immediately led to the following performance measure which should be maximized:

*model quality* = percentage of stoppers that can be found in the top 10% of cases.

In other words, clusters should be arranged according to the percentage of stoppers (the higher the better) and then cumulative percentages of stoppers should be calculated. The percentage of stoppers which are encountered in best clusters that together cover 10% of all cases is taken as the final quality measure. The choice of

'10%' was partially based on expert knowledge, partially on common sense. Namely, it was expected that among all stoppers some were 'typical' (i.e., easy to predict) others were not. Now by estimating the ratio between both types of stoppers and the fact that the model should focus on typical stoppers the figure 10% has been found to be reasonable. As a matter of fact, when evaluating model performance we were also measuring percentage of stoppers in the top 20% and 30% of all cases.

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 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.

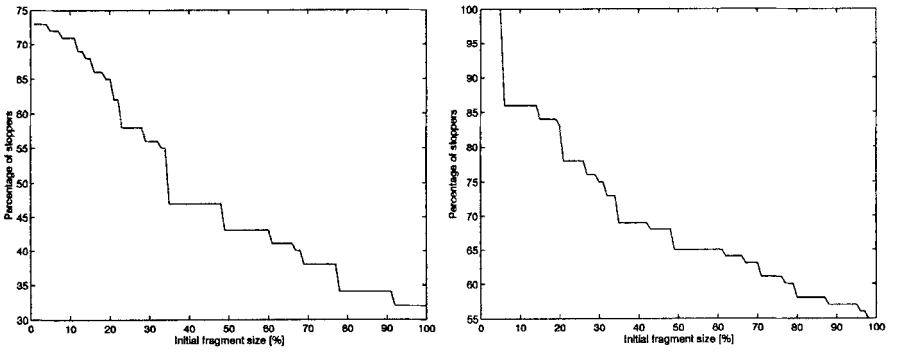


**Figure 2.** Cumulative performance of best models for A-clients (left) and B-clients (right). Note different scales on both plots.

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%).

## 6 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

**if** $(a < X_1 < A)$ **&** $(b < X_2 < B)$ **&** $(c < X_3 < C)$ **then** 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:

- attributes $X_1$, $X_2$, and $X_3$ were arbitrary combinations of attributes taken from the set of 8 most important attributes
- 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
- rules were allowed to involve only 2, 3, 4 or 5 'splitting points'.

For example, rules which involve 2 splitting points have the form:

**if** $(a < X_1 < A)$ **then** decision

or

**if** $(a < X_1)$ **&** $(b < X_2)$ **then** decision

etc.

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.

## 7. Conclusions and suggestions for further research

In this paper we traced a data mining process aimed at understanding the phenomenon of customer retention. Various phases of this process have been described: conceptual analysis of the problem, initial analysis of available data, identification of most

important attributes, construction of models and extracting rules from the data. In spite of simplicity of the presented approach (systematic search through large collections of rough data models and rules) we obtained results which helped to identify various factors which influence customer behaviour. In a comparative study, [2], the same problem has been approached by other techniques: Genetic Programming, [8], Logistic Regression, [5], and CHAID (CHI-square Automatic Interaction Detection), [4]. However, none of these techniques provided as much insight into the problem as ours.

Our approach has also some drawbacks and future research will focus on them. First of all, the presented approach can be used for building models which are based on relatively few attributes (say, 2-5). Models which are built on more attributes involve many small clusters and generalise badly. Second, the search procedures which were used in our experiments were very primitive (exhaustive search) so it is clear that by using some heuristics we could increase the size of explored search space considerably. This should result in better performance of generated models. Finally, the issue of rule extraction has been not treated (in this research) very deeply. For example, while generating rules no mechanisms for enforcing rule independence (different rules should cover different areas of the model) have been used.

The results of the reported research also led to some suggestions concerning the data. For example, the definitions of risk and profit should be adapted to reflect more the client (subjective) point of view of these factors. Also some additional factors with respect to the communication between the company and her clients should be taken into account. Moreover, recent research in the field of marketing stresses the importance of attitudinal factors in customer loyalty, [10, 12]. We believe that incorporating all these suggestions in a follow-up project will result in an even better understanding of customer retention.

## 8 Acknowledgments

We would like to thank the company which sponsored the project for an excellent cooperation, a very interesting problem and high quality data sets. Moreover, we are grateful to our colleagues: Guszti Eiben (Leiden University), Teije Euverman and Jan Wesseling (University of Amsterdam), Rob Walker and Arnold Koudijs (Cap Volmac) for many fruitful discussions.

## 9 References

1. Blattberg, R.C., Deighton, J.: Manage marketing by the customer equity test, in Harvard Business Review, July-August (1996), 136-144.
2. Eiben, A.E., Euverman, T.J., Kowalczyk, W., Peelen, E., Slisser, F., Wesseling, J.A.M.: Customer Retention: towards predicting stop moments of private investors, Internal Report of the Marketing, Intelligence and Technology Foundation (in Dutch), (1996).
3. Hair, J.F., Jr., Anderson, R. E., Anderson, T.R.L., Black, W.: Multivariate Data Analysis (fourth edition), Prentice Hall, Englewood Cliffs, New Jersey, (1995).

4. Haughton, D., Oulida, S.: Direct marketing modeling with CART and CHAID, Journal of direct marketing, volume 7, number 3, (1993), 16-26

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

6. Kowalczyk, W.: Analyzing temporal patterns with rough sets, in H.-J. Zimmermann (ed.), Proceedings of the 4th European Congress on Intelligent Technologies and Soft Computing, Verlag der Augustinus Buchhandlung, (1996), 139-143.

7. Kowalczyk, W.: TRANCE: a Tool for Rough data ANalysis, Classification, and clustEring. Proceedings of the Fourth International Workshop on Rough Sets, Fuzzy Sets, and Machine Discovery.Tokyo University, (1996), 269-275.

8. Koza, J.: Genetic Programming, MIT Press, (1992).

9. Pawlak, Z.: Rough Sets–Theoretical Aspects of Reasoning about Data, Kluwer, (1991).

10. Peelen, E., Ekelmans, C.F.W., Vijn, P.: Direct Marketing for Establishing the Relationships Between Buyers and Sellers, in Journal of Direct Marketing, Volume 3, Number 1, Winter (1989), 7-14.

11. Quinlan, R.: Induction of decision trees, Machine Learning **1** (1996), 81-106.

12. Reichheld, F.F.: Learning from Customer Defections, in Harvard Business Review, March-April (1996), 56-62.

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