Database Marketing Applications in the Italian Banking Sector

Alberto Saccardi, NUNATAC
Guido Cuzzocrea, NUNATAC

Paper presentato al convegno SEUGI'98, Praga 9-12 Giugno 1998

Abstract
Many database marketing projects in Italian banks have been started before the development of a Data Warehouse and without a specific Data Mining expertise within the company: these projects have the goal of obtaining operative tools for marketing activities, such as cross-selling.
In our experience, these projects are the best way to test costs and benefits derivable from larger investments in Data Warehousing and Data Mining: dealing with the selection of a work group, the design of a Marketing Data Model, the implementation of data analysis results and, finally, with the evaluation of the return on investment, the company has the opportunity to validate and justify the strategic evolution towards Data Warehousing and Data Mining.
To gain a competitive advantage, banks have the opportunity to exploit the enormous amount of data at their disposal. Data Mining should supply the marketing decision makers with operative tools for the evaluation of client potential, the planning of an integrated offer and the selection of campaign targets.
Two case histories will be presented:
- *the behavioural segmentation of retail customers in a mid-sized bank,*
- *the development of a scoring system to select the best campaign target in a large-sized bank.*

Introduction.

Nunatac is a consulting firm which is made up of specialists in statistics, marketing and information technology. Its current business solutions fall into the following categories: data warehousing, data mining, business reporting and campaign management. In particular we are experts in database marketing activities, comprising data warehousing and data mining to do business.
Nunatac focuses on supplying tailor made business solutions for its clients, using the SAS SYSTEM and is certified as a SAS Quality Partner. The most recent projects are in the publishing, manufacturing and banking sectors. Specifically, in the banking sector we have developed projects with Banca Popolare di Lodi, whose client base is roughly two hundred thousand and with CARIPLO, whose two million clients make it one of Italy's largest private banks.

In the first part of this paper, 1. Database Marketing Activity, we'll introduce the business problem and what's necessary for a database marketing activity. In the second part, 2. Case Histories, we will explain two case studies in the banking sector. At the end some concluding remarks.

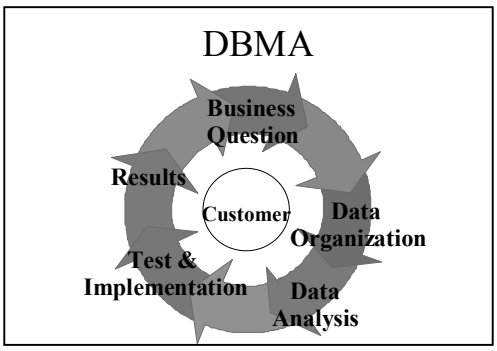
1. Database Marketing Activity.

Some of the most important business questions a marketing manager must answer are:
- who are our customers?
- what are they like?
- how many behavioural groups are there among them?
- who are the most profitable?
- how can we maximise customer lifetime value?

When trying to answer, the starting point is: the interaction between customers and the company produces an information flow. If this information is collected, stored and analysed appropriately, one can expect to profile customers, identifying their preferences and understanding who is likely to be more profitable.

1.1. The approach.

In answering the questions above, Nunatac has adopted an approach to carry out Database Marketing Activity (DBMA).



The DBMA puts the customer at the centre of interest and the approach may be summarised in the following steps.

Business question.

This step is not always easy, for example we may be interested in defining the best customers. Who are the *best* customers? The ones that generate the highest revenue or the ones that are the most loyal? Furthermore, will today's good customer be as good tomorrow?

Data Organisation.

Once the business question has been defined, the second step is to collect and to arrange the data needed to solve the problem.

In this phase it's necessary:

- to define which are the statistical units: e.g. individuals, families,…;
- to census the data presents in the Information System;
- to transform and aggregate data for preparing the data marts for the analysis.

Data Analysis.

The third step is the analysis of the data. There are many techniques that may be used. To select the appropriate techniques there are two main criteria:

- the type of problem,
- the type of data.

About the type of problem, we might be interested in:

- finding similar groups of customers considering their behaviour not related to a specific phenomenon;
- identifying the most important attributes which explain a specific phenomenon: e.g. a policy subscription.

From a statistical point of view the first question is a problem of interdependence analysis while the second is a problem of dependence analysis. In the first we want to investigate the intra-correlation structure between all the variables, in the second there one is phenomenon, represented by a variable, which is explained by a set of explanatory variables.

Having fixed the statistical nature of the problem, the type of data will determine the most appropriate technique. In fact, if we have qualitative data we will have to use techniques which deal with frequencies, while if we have quantitative data, we will have to use techniques which deal with measures.

Test & Implementation.

The fourth step is to test the models, which have been estimated, using control samples of data. Model fitness and model robustness are the criteria which must be considered in the choice of the final model. When the analysis is considered satisfactory it's necessary to implement the rule, defined by the model, on the customer file.

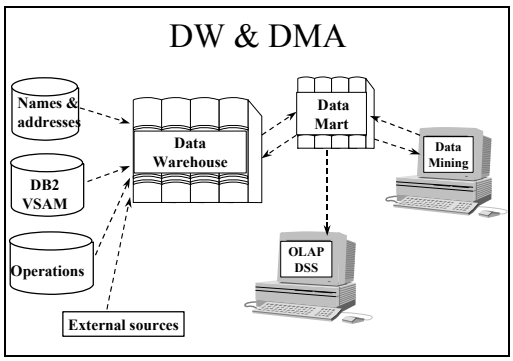
Results.

The fifth and last step is to measure the results which are obtained using this rule. When possible, we should compare these results with those we might have obtained using different models, used to estimate the same phenomenon. The comparison should be done, at least, with the results obtained from a control sample selected using traditional criteria.

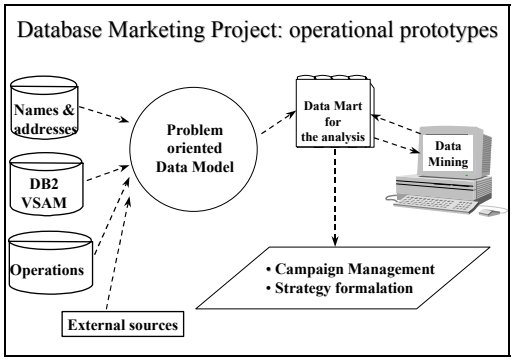
What's necessary to implement this approach?

- First, an organised work group with diverse expertise: Marketing, Statistics, Information Technology.
- Second, data access and management technology.
- Third, a data model by which it is possible to organise and to prepare data for data analysis activities: reporting, OLAP, data mining.

The ideal environment for DBMA is characterised by a data warehouse, where there is an integration of the information sources and offering fast and easy to access data. Data mining software or EIS applications allow marketing analysts to make more informed decisions.



Unfortunately we know that the reality is not so easy and that data warehousing is a complex and long process.
What can we do if the data warehouse is not ready, and how can database marketing activity help in the planning of the data warehouse?
One possible way is the construction of an operational prototype. The starting point is the census of the external an internal sources and the screening of the operational processes. Afterwards the project team have to define the problem oriented data model, to prepare the input data set for the analysis, to do the analysis, to implement the results obtained managing a campaign or formulating a business strategy.



1.2. Data model for Database Marketing Activity.

Usually the sources of the data are the operational tables. Exogenous information may also be considered, for example external mailing lists or census data. The internal and external sources supply the environment prepared for the data analysis activity. This environment is composed of fact tables (detailed and certified data) and customer tables (summarised data related to the customers).
The data extraction processes populate the fact tables whose data model is problem oriented. So it's not necessary to consider all the operational tables or all the possible relationships present in the operational environment in its design. The main characteristic of the fact tables is that they contain certified data, and so it might be necessary to implement data cleaning processes. Furthermore the data have to be at the maximum detail: one record per variation of the fact considered. In this phase data are not summarised and they are most of all used for reporting activities: standard tabulate, OLAP and so on.

The data at this level are not organised for data mining activity, where the customer is the logical dimension of interest. To do data mining, first of all, it's necessary to define which are the statistical units to analyse. In other words, it is necessary to define exactly who are the customers, whose behaviour we want to investigate and to design a data model where one customer corresponds to one record.
The data sets, which contain such data, are called customer tables. From a logical point of view, the customer tables are thematic tables, organised by customer identification code. They contain summarised data regarding: the analysis variables, in the data warehousing terminology the facts, e.g. account transactions; the classification variables, the dimensions, e.g. transaction type, the interaction between facts and dimensions; the fixed time lag for summarising the data.

The following picture is an example of a customer table. The first column contains the customer identification code, the analysis variables are the number of orders and the revenue (label ORD and label INC). These analysis variables are crossed with a classification variable, e.g. the Business Unit, whose categories are A, B, C, and the data is summarised, for example by quarter.

Customer Table

SALES REVENUE						
Cust_id	Ord_A	Ord_B	Ord_C	Inc_A	Inc_B	Inc_C
247893	3	1	0	300	10	0
248790	2	0	0	200	0	0
249420	0	0	1	0	0	1000
250393	1	0	0	100	0	0
256793	2	5	0	200	50	0

A critical step in designing the customer tables is deciding the level of the interaction between facts and dimensions to control the level of granularity. A very simple example: we consider one fact, e.g. the purchase of a financial product and we suppose that the investment may be described by three dimensions. First the time to maturity: short, middle, long; the risk level: stock only, stock balanced, bond balanced, bond only, cash; the country: Home Country, Europe, U.S.A., Pacific Area, International, Emerging Markets. If we design the customer table considering all

the interactions between the dimension, we will create a table with 3*5*6=150 variables! So it's very important to design a data model where the trade-off between information intactness and data granularity is mediated.

The customer tables are thematic tables and in a marketing data model we might have: a customer table for the revenue, one for the promotions, one for the customer assistance and one for the personal data.
Once the tables have been designed and the data is ready, it's possible to start the data mining activity. The results obtained by a behavioural segmentation or by a scoring model will be assigned to each customer and used for business purposes.

2. Case Histories.

In the following, we will present two data mining experiences that we have brought about in the financial sector in Italy: in both cases we will discuss the approach to database marketing activity of two Italian Banks.
As you will have the opportunity to notice, the logical flow chart we followed in these projects resembles the SAS Institute's SEMMA methodology. This methodology is the clear summary of several years of experience gained in the application of Data Analysis to support business decisions.

Speaking about data mining applications there is often some confusion in distinguishing Segmentation from Scoring: in the first case presented we will deal with behavioural segmentation as a problem of general classification of customers, while in the second case we will see how to score different targets of customers depending on their probability to purchase or not purchase a specific product. More in general, scoring models attempt to estimate the probability of any fixed binary phenomenon, such us purchase or not, accident or not and so on, depending on a set of explanatory variables such as sex, age, income and so on.

From a general point of view the goal of the marketing departments of Italian banks is, today more than ever before, to increase the average profitability level per customer. To obtain a competitive advantage banks as well as insurance companies have the opportunity to exploit the large amount of data at their disposal to know their customers better: To understand their preferences and needs in order to increase their loyalty and motivation.
From a concrete point of view data mining activity should supply the marketing decision makers with operative tools for the evaluation of client potential, the planning of an integrated offer, for the selection of campaign targets and for cross-selling.

2.1. The behavioural segmentation of retail customers in a mid-sized bank.

The starting point of this project is the already mentioned large amount of data concerning the customers' behaviour and characteristics: if a marketing data warehouse is not available yet, the problem is to discover which, where and how the information is stored in the Information System of the Bank.
After that it is necessary to extract a representative sample of customers, integrating and cleaning all the pieces of information coming from different company sources. On the sample, the data analysis step will generate the descriptive profile of the different groups of behaviours that characterise our customers. The description of the groups will be discussed and validated with business users, in order to obtain an efficient classification both from the statistical and the marketing point of view. Then, separating the sample in training and test sub-samples, it is necessary to estimate a general rule of classification which should allow us to classify all the Bank's customers in the identified behavioural targets. The classification can, therefore, be performed at the present time and in the future, without repeating the analysis and building a time-stable and company-sharable business information system.

From a more analytical point of view the notable steps of the project are as follows:

1. it is extremely important to identify a unique, company-wide definition of a customer;
2. then it is necessary to analyse the complete set of data at our disposal, organising and cleaning the subset of data we decide to use in the project, select a representative sample of customers and carry out the so called preliminary analyses;
3. after that, Factor Analysis and Cluster Analysis are the multivariate statistical techniques which allow us to simplify the structure of the data, identifying the main dimensions of behaviour and then producing groups of customers

characterised by similar values of those dimensions;

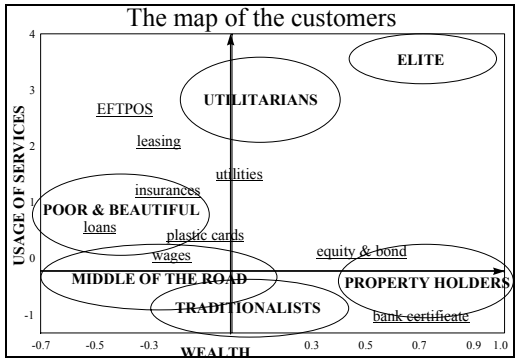
4. the estimation of a general rule of classification is a crucial step in which different inferential engines such as Neural Networks and Discriminant Analysis are compared, considering their ability to correctly classify customers;
5. finally the classification of customers in the identified segments allows us to build operative marketing tools such as the customer and target profile to monitor the variation in customer potential and the migration flow matrix for customer tracking.

With regards to the company-wide definition of a customer, the problem to be faced is that banks are used to carry out business thinking about current accounts and not about individuals. Moreover branches have their own customers and they do not necessarily share information and business with each other.

Concentrating on the data mining techniques used in this application, after having chosen the set of input variables necessary to completely describe the customer behaviour, we considered Factor Analysis as a powerful tool which allow us to simplify the complex structure of interrelated data, identifying the main dimensions of the phenomenon under analysis.
Considering all the transactions of the current accounts in a given year, we found that the general behaviour of the Bank's customers could be summarised by a so called wealth factor and an orthogonal factor measuring the propensity towards the usage of financial products and services, either for private purposes or for professional ones.

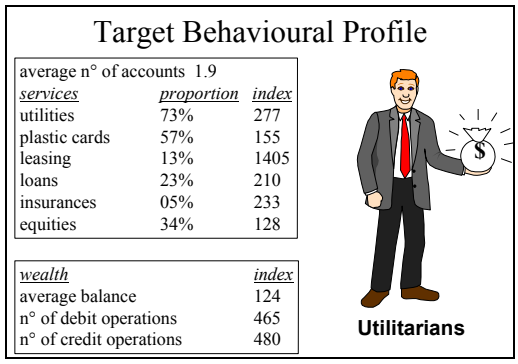
After the reduction of the data matrix dimensions, the application of Cluster Analysis produces groups of customers according to their similarity in the analysed behaviours.

Once clusters' number and positioning are chosen, it is possible to graphically represent the map of customers: the identified segments and the relative presence of financial services and products are plotted over the two main axes of the analysis.



The Poor and Beautiful segment is the segment of young couples who apply for loans and have a positive propensity towards the usage of financial products and services.

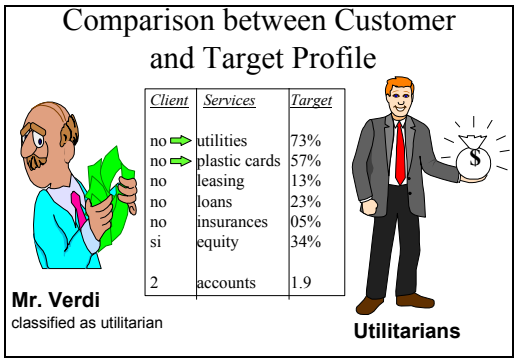
The quantitative summary of the previous graphical representation for each segment is the Target behavioural Profile, which is a clear and comprehensive description of each segment's average characteristics.
The Utilitarians, for example, are slightly wealthier than the general average (balance index equal to 124 and equities index 128) and are heavy users of financial services such as utilities, leasing and insurance policies: most of them are retailers or professionals who use financial services for their business.



After the descriptive steps of the analysis, separating the sample in training and test sub-samples, it is necessary to estimate a general rule of classification which should allow us to classify all the Bank's customers in the identified behavioural targets.
Neural Networks are a powerful tool of analysis that one can consider for solving problems of pattern recognition. In classification problems Neural Networks can learn from significant examples the patterns of data that allow us to correctly classify the statistic units, that is, in our application, customers.

To find the best classification rule in our problem, we considered either the original accounts variable for each customer, or the transformed main Factors. The best result measured on the test set was obtained by an MLPs Neural Network having the pre-determined Factors as input layers. In that case the misclassification error was just over 5%.
In order to have the possibility of classifying new customers, we also trained a Neural Network with personal data like sex, age, number of components of the family and so on, as input variables. Of course the error level increased a lot, reaching about 25%.

The implementation of the general rule of classification allowed us to associate each customer to the most probable segment to which he could belong. Once segmentation is applied to all clients, a simple and ready-to-use marketing tool is the comparison between customer and target profile: one can easily compare what a client has in his portfolio with the average presence of products/services in the segment he belongs to.



Another important asset of the classification rule is the possibility of following the so called tracking of the customer.
While data mining analyses have to be repeated year after year to verify that models are still valid (and you can do that over a test sample!), the implementation of the same rule in subsequent periods allows the evaluation of migration flows.
From the statistical point of view, the quantitative measurement of the migration flows can be done considering a double-entry frequency table with the previous year's segments as rows and the next year's segments as columns: in each cell we count the number of customers who were, for example, Elite and now are Utilitarian.
A graphical representation of the transition matrix can be done considering another multivariate statistical technique, that is Correspondence Analysis.
We are not as interested in deepening the technical aspects of Correspondence Analysis, as in considering the relevant information contained in the map: where are our customers going? Are we loosing ore gaining profitable customers? Were our marketing policies effective?

The whole project has been developed using SAS System products, considering the standard modules that were already available in the company.
The SAS System proved to be a well integrated software tool across the different steps of the project, from data extraction to model assessment, up to business reporting.

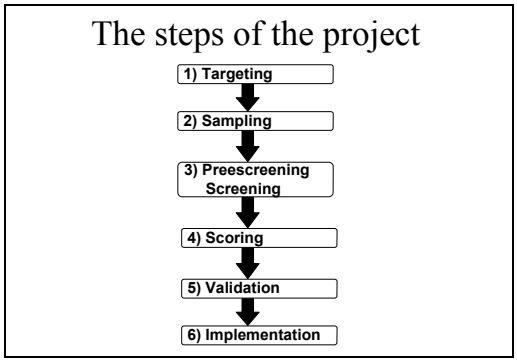
Conclusions: customer relationship management!
The important remark is that data mining activity should not remain within the R&D department borders: business decision makers should benefit from data mining through out the development of ready-to-use operative marketing tools.

2.2 The development of a scoring system to select the best campaign target in a large-sized bank.

In this case the main problem is to select the subset of customers with the highest probability of purchasing.
In particular the business problem is the identification of the best target for the direct offer of a life insurance policy to the customers of the bank.
From a statistical point view this problem is one problem of dependence analysis, where we want to investigate a phenomenon, the policy's agreement, using behavioural and personal data as explanatory variables.
In more detail the goal of the project was, in our case, to identify the best 100.000 potential customers from the customer file, composed of 1.500.000 names.
There are many criteria for solving this problem: experience, chance. In the data mining approach it may be solved constructing a Scoring System.
We can think of the Scoring System as a process whose goal is to assign a score to each customer. This score is proportional to the probability of assuming a fixed status: in this case the status is accepting the commercial proposal, while the probability is estimated by

a statistical model applied to a historical customer file.
In this way we are able to identify customer subsets characterised by different scores, and of course select the subsets with highest score.

The Scoring System may be summarised by the following flow-chart:



We will now briefly analyse each step.

The goal of the targeting step is to identify the a priori universe for the direct offer.
In this case, the potential recxipients had to have the following characteristics:

- only retail customers of the bank, so the companies were excluded,
- age between 25 and 50 years,
- prospects who had already been mailed were also excluded.

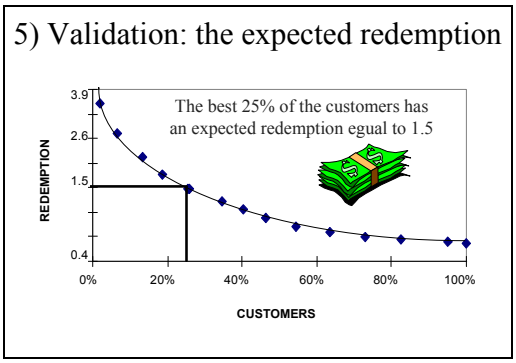
The so identified universe has been divided into two subsets: the non clients group is that of non policy holders, while the clients are those who had signed the life insurance policy during the last year.

Two samples have been extracted from the identified universe, composed of 2.418 clients and more or less 775.000 non clients.
All the most important behavioural and personal data was joined to every sample unit.
So, the probability to became a client or not is the phenomenon to be investigated or, in other words, the dependence variable. The behavioural and personal data represent the explanatory variables of that phenomenon.

In the pre-screening and screening phase it is necessary to reduce the complexity of the data matrix, considering the correlation structure between the roughly 100 explanatory variables and the dependent variable.
This will allow us to obtain more robust statistical models and more general results.
In this case the explanatory variables were first grouped into three logical categories: personal-data, investment data and current account operations.
After that, the most correlated variables were grouped into macro orthogonal variables which were then used as input variables in the final scoring model.

The modelling step considered both Neural Networks and Logit models. In this case Logit models worked better and the possible technical explanation for that is the categorical nature of the constructed macro-variables, corresponding to too many dummy variables in the Neural Network solution.

One way to select the best model and validate the discriminating power of the scoring system is to consider the expected redemption graph: on the horizontal axis we measure the progressive percentage of potential clients ordered by score, while on the vertical axis we measure the corresponding redemption level predicted by the model.



If we selected the best 25% of our customers, we would expect a 1.5% redemption level offering them the life insurance policy.

One of the most important requirements in database marketing activity is the opportunity of measuring the obtained results of each marketing action. In this case, three months after the campaign the observed redemption level over the contacted customers was measured both for the scoring target of 100,000 and for a control sample of 10,000.
Results have to be transformed to figures, for a correct costs/benefits analysis.

6) Implementation: benefits evaluation

Benefits Analysis after the campaign				
TARGET	# MAIL	REDEMP.	POLICIES	REVENUE
MODEL	100.000	2.3%	2.300	2.300 x £ire
CTRL SAMPLE	100.000	1.5%	1.500	1.500 x £ire

Success is often determined by the technology at our disposal.
This paper has discussed some database marketing experiences we have carried out using traditional SAS products: SAS BASE, SAS STAT and so on.
Now we are working with the new solutions presented by SAS INSTITUTE: SAS WAREHOUSE ADMINISTRATOR for data warehousing, SAS ENTERPRISE MINER for data mining and SAS ENTERPRISE REPORTER for the business reporting.

The importance of data availability and data analysis can only increase.

Data warehouse and *data mining* are new terms applied to old ideas. What is new, in our opinion, is making these ideas reality.

References

Agresti, A. (1990). Categorical Data Analysis, John Wiley & Sons, New York

Bouroche, J.M. e Saporta, G. (1980). L'analyse des données, C.L.U. Editrice, Napoli.

Jobson, J.D. (1992). Applied Multivariate Data Analysis, Springer-Verlag New York.

Sarle, W.S. (1994). Neural Networks and Statistial Models, SAS Institute Inc., Cary, NC, USA.

Sarle, W.S. (1995). Neural Networks Implementation in SAS Software, SAS Institute Inc., Cary, NC, USA.

SAS Institute Inc., *SAS/STAT User's Guide, Version 6, Fourth Edition,* Cary, NC: SAS Institute Inc., 1989.

Wasserman, P.D. (1993), Advanced Methods in Neural Computing, New York: Van Nostrand Reinhold.

Alberto Saccardi - Guido Cuzzocrea
NUNATAC S.a.s.
via S. Martino 11C - 20122 Milano (Italy)
Tel. **+39 2 58327005**
Fax. **+39 2 58327122**
E-mail: mailto:alberto.saccardi@nunatac.it
mailto:guido.cuzzocrea@nunatac.it