January 14, 1999

project title: Behavioural Segmentation of Retail Customers
customer: Banca Popolare di Lodi
industry: Banking
function: Marketing
project contact: S.Ercoli, G. Cuzzocrea, M. Saputo

Customer situation/problem description

The Banca Popolare di Lodi is head of the banking group of the same name. The group is composed of 11 banks, 280 bank counters, 500 financial backers, about 2,500 staff and it serves about 600,000 customers.

The Banca Popolare di Lodi counts 140 bank counters, 1,400 employees and about 320,000 customers. SAS is used in the department: Divisione Pianificazione Studi, which is staffed by 10 people.

The strategy of expansion that the Banca Popolare di Lodi group has pursued in recent years has led to the acquisition of banking properties which are deeply rooted in their place of origin. Consequently, it has become aware of the need to gain in-depth knowledge of the bank's customers in each particular region, so as to offer products and services that best meet their needs. The growth of credit and the stimulus towards new forms of saving and investment cut across the knowledge of the characteristics, preferences, aspirations and specific needs of each user of the services offered by the bank.

Project Description

The objectives of the analysis are:

- To delineate behavioural segments of Banca Popolare di Lodi's customers;
- To analyse the dynamics of the behavioural segments over time.

Correct indications of this type are the basis for:

- the evaluation of the potential for personalising the financial solution offered
- the activity of cross-selling
- the identification of targets which guarantee the profitability of specific commercial activities
- planning and feedback on the bank's commercial strategy.

From an operative point of view, the goal of the project is to create the target behavioural profiles. The implementation of the segmentation allowed 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. In this way the bank tellers can optimise their assets.

Solution

The starting point of this project was to collect the wealth of data concerning the customers' behaviour and characteristics: the problem was to discover which, where and how the information was stored in the Information System of the Bank.

After that it was 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 the customers. The analysis was done using multivariate data analysis techniques made available by the SAS system.

The analysis identified six clusters:

Cluster 1 - "Traditionalists"
This is the largest group. Generally they have only one current account, make few transactions above average value, make traditional, prudent choices (only Deposit Certificates), and make little use of plastic cards for payment of goods.

Cluster 2 - "Poor but beautiful"
Those forming this segment are mainly young nuclear families characterised by high spending (mortgage payments and insurance), meagre resources and high use of credit. The commercial objective is to develop a relationship in anticipation of their future earnings. They can also be targets for Phone Banking.

Cluster 3 - "Utilitarians"
Those in this segment are frequent users of current accounts, businesses or sole traders (entrepreneurs or shopkeepers) who use leasing, E.F.T.P.O.S and insurance and have average financial means.

Cluster 4 - "Middle of the road"
Sons of the traditionalists, but with more disposable income and utilisation, they have more need for credit for their spending. They are in a transitory phase: as is shown on the fluctuation matrix, they will evolve towards the traditionalists' cluster once their debts have been reduced , or towards the "poor but beautiful" cluster and permanent indebtedness.

Cluster 5 "Property holders"
Those belonging to this cluster have large investments in shares and Certificates of Credit, ample liquidity in current accounts and scant need of services. Their accounts show high liquidity and movement of large sums, but in very small numbers.

Cluster 6 - "Elite"
Rich and progressive, very attached to the institute, whose services they use widely. They have a high average balance, but low liquidity and frequent transactions, and they hold two or three accounts. They are few but good customers and represent possible users of home banking.

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.

A general rule of classification was drafted. This should allow us to classify all the Bank's customers in the identified behavioural targets. In this way the classification can 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. The analysis was done using both, multivariate traditional tools, and neural network models made available by SAS software's Neural Network Application.

The implementation of the general rule of classification, as we have already pointed out, allows to classify each customer to the most segment, or cluster, to which he could belong.

Another important asset of the classification rule is the possibility of following the so called tracking of the customer: 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 was carried out considering a double-entry frequency table with the previous year's segments as rows and the following year's segments as columns: in each cell there is the number of customers who were, for example, Elite and now are Utilitarian.

A graphical representation of the transition matrix was created using another multivariate statistical technique, that is Correspondence Analysis. Considering the relevant information contained in the map it was possible to answer questions like these: 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. 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.

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 the development of ready-to-use operative marketing tools.

Project Summary

Work group:

- Nunatac's organisation: 2 data mining analysts
- Banca Popolare di Lodi's organisation: 2 marketing analysts, 1 business analyst, 1 programmer analyst.

Time frame:

- 3 months for the pilot release
- 8 months for the delivery of the whole project (including pilot)

Time man efforts:

- Nunatac: 5 man-months
- Banca Popolare di Lodi: 3 man-months.

Project Methodology

Define Business Problem

From a general point of view the business questions formulated were:

- who are our customers?
- what are they like?
- how many behavioural groups are there among them?

To answer to these questions it was necessary to understand customers' preferences and needs, in other words it was necessary to do a behavioural segmentation analysis. The obligatory steps were:

- to collect data
- to investigate if it was possible to create groups of customers characterised by homogeneous behaviours.

The formulation of the business questions was clear, but we have had to work a lot with the bank's analysts to choose which variables to use in the segmentation. In fact the clusters were created and explained on the basis of the variables selected.

During this phase, it became evident that we would have to create two different kinds of segmentation:
One, more simple, with less clusters to give to the tellers for their cross selling activities
One, more refined, with more clusters to give to the business and marketing analysts to design new products and to select campaign targets.

Another important issue we discussed with the analysts of the bank was the identification of the statistical unit to analyse: it was extremely important to identify a unique, company-wide definition of customers. With regards to the company-wide definition of a customer, the problem to be faced is that banks are accustomed to carrying 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.

Evaluate Environment
The environment was favourable: good knowledge of the business problems, good knowledge of the internal data.

At first, the project had the sponsorship of the department: Divisione Pianificazione Studi.
After the pilot release the project gained visibility and it was sponsored by the marketing management.

The IT environment was an IBM – MVS, but the data mining activity was done using a powerful PC.
The software: SAS BASE, SAS STAT, SAS INSIGHT, and SAS NEURAL NETWORK APPLICATION.

Make data available
The original data was in DBII tables, and wasn't organised for data mining activity. So it was necessary to:

1. define the data model for the Customer Table[1].
2. create data extraction processes from the operational environment.
3. create the processes for the implementation of the Customer Table.

The first step was done by the consultants together with the analysts of the bank. The data model was defined in about one month. In particular the transformation and classification of the operational data, together with the time lag for summarising data, were the points of major work. This step required several meetings, which were concentrated in one month.

The second step was carried out by the analyst in about fifteen working days. The data extraction processes were done in SAS.

[1] To do data mining it's necessary 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.
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 third step was accomplished by the consultants and the analyst in one and half man-months. This step also included the data screening to assess their quality. Fortunately, the preliminary activity of data cleaning was not such a burden: data was good and missing values were not a problem.

The Banca Popolare di Lodi does not have an enterprise data warehouse, but fortunately the IT analysts have a great knowledge of the operational environment and this made the data collection phase easy.

Mine in cycles

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

1. We selected a representative sample of customers and we carried out the so-called preliminary analyses whose aim is to understand the macro phenomena and to complete data cleaning. We used the bank counter as stratification variable and we selected a random sample of about 60.000 customers.
2. After that, we used Factor Analysis and Cluster Analysis 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.
3. The estimation of the general rule of classification involved different inferential engines such as Neural Networks and Discriminant Analysis, which were compared considering their ability to correctly classify customers.
4. Finally the classification of customers in the identified segments allowed 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.

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 allows 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 produced groups of customers according to their similarity in the analysed behaviours.

Once clusters' number and positioning were chosen, we gave a graphical representation of customers' clusters: the identified segments and the relative presence of financial services and products are plotted over the two main axes of the analysis.

After the previous descriptive steps of the analysis, separating the sample in training and test sub-samples, we estimated the rule of classification which should have allowed us to classify all the Bank's customers in the identified behavioural targets. To find the best classification rule in our problem, we used Neural Networks and Discriminant Analysis. 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%.

Using the classification rule it's possible to follow 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. We have already explained that measurement of the migration flows can be done considering a double-entry frequency table where in each cell we count the number of customers who changed strata: for example, Elite in the past year and now Utilitarian. A graphical representation of the transition matrix was done considering another multivariate statistical technique: Correspondence Analysis.

The most critical steps were the choice and the interpretation of the factors and, consequently, the creation of the clusters. The pilot release was very important to integrate the statistical criteria with the business criteria. Interaction between Nunatac's consultants and Lodi's analysts was of fundamental importance.

The lesson learned is that the choice of the number of factors and of the number of clusters must follow criteria that map the business requirements.
The statistical measures have to support the choice, but they cannot be considered the only parameter. The risk is to segment the population in a way that does not meet the operative business needs.

During the data mining cycles we didn't go into depth on the technical aspects of the statistical methods with the bank's analysts, but we discussed all the results which might have a business consideration. For example considering the relevant information contained in the migration flows' map: where are our customers going? Are we loosing ore gaining profitable customers? Were our marketing policies effective?

Implement in production
The whole project was 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.

Review
Currently there is no review under way.

Lessons learned

- Do not underestimate the contributions that the business analysts may give, even if they are not data mining specialists.
- The quality of input data is fundamental, even if you try to clean or to reconstruct the data used in the data mining cycles.
- Find a very good IT analysts who knows the operational environment, if this is the source of the data you have to analyse.
- Start with a pilot release to give visibility to the project in a short time.
- Create more several and compare their results.

The most important remark is that data mining activity should not remain within the Research & Development department borders: business decision makers should benefit from data mining through the development of ready-to-use operative marketing tools.

Future developments

The next step will be the monitoring of the cross-selling activity supported by the results of the segmentation.