THE 6TH EDITION OF THE INTERNATIONAL CONFERENCE
EUROPEAN INTEGRATION
REALITIES AND PERSPECTIVES

Data Mining – Innovative Method for Obtaining Information in Marketingand Business Management

Mirela-Cristina Voicu[1], Alina-Mihaela Babonea[2], Andreea Paula Dumitru[3]

[1]*Nicolae Titulescu University, Faculty of Economics, voicu.cristina.m@gmail.com*

[2]*Nicolae Titulescu University, Faculty of Economics, alinababonea@univnt.ro*

[3]*Nicolae Titulescu University, Faculty of Economics, pauladumitru@univnt.ro*

Abstract: The existence of massive amounts of data raised the question of using their reorientation to a retrospective to a prospective operation. Data mining offers the promise of an important aid for discovering hidden patterns in data that can be used to predict the behavior of customers, products and processes. Data mining tools must be guided by users who understand the business, the general nature of the data and analytical methods involved. It discovers information within the data that queries and reports can't effectively reveal. It is vital to collect data and prepare properly, to face reality models. Choosing the most appropriate product data mining is to find a tool with the capabilities required, an interface that matches the skills of users and can be applied in a specific business problem. In this context, the purpose of this paper is to illustrate some of the problems of company activity problems which can be solved by using data mining techniques.

Keywords: marketing research; data mining; segmentation; cluster analysis; consumer behavior

1 Introduction

Organizations' strategic flexibility depends to a great extent on securing access to information. The fact that many companies make significant efforts to research for data supporting the hypotheses stated is well known, but, most of the time, managers don't know what to expect. To further complicate things, within many organizations, different parties following different objectives may be encountered having, surprisingly, very little information concerning their clients. It's the marketing's duty to aid in interpreting data, to present the research results and to plan actions based on the results obtained. Many companies were faced lately with a pretty serious issue: the data is available in large quantities, but the information is insufficient. This is due to the fact that only part of the data gathered is analyzed and processed, due to the fact that the data is inconsistent and divided in several databases. In average, no more than 7% of the data collected within a company is used, the rest representing a potential that should be explored in order to obtain the necessary information and being a step beyond the competition. How could we, as marketers, help our clients improve the results of their activity? What techniques can we use in order to create an advantage for our clients regarding their business?

As all the data tends to be multivariate when being interpreted, the data analysis exploratory techniques may identify patterns within the data sets (Ruxanda, 2005). For example, in case we wish to analyze the relationship between client satisfaction and company profitability, our survey shall include, evidently, questions for individual consumers, questions that may be aggregated at a high

lever (for example, all consumers preferring a certain location of the store) in conjunction with data related to the store revenues. Multivariate analysis allows marketers to mold in an objective manner the relationship between the consumers' answers, the variation in levels of satisfaction, as well as increases and decreases in profitability, apart from discovering patterns useful in forecasting future variations in company profitability and strategies for preventing decreases in profitability.

Moreover, using multidimensional data analysis in studying the data obtained may provide a certain depth in the way of looking at the organization's strong and weak points as well as unique benefits, specific to each approach.

2 Data Mining – The Process of Extracting Information

Data mining is the process of extracting information previously unknown and potentially useful from large databases. This process consists in searching for information in large databases, exploring and analyzing it by using automatic or semi-automatic methods, with a view to discover useful patterns and rules.

Data mining tools may scan databases and identify previously hidden patterns in one single step. Such an example is analyzing retail sales data in order to identify so called unconnected products that are often sold together.

The data mining tools may also automate the process of finding predictive information in large databases. Questions that normally require vast analyses may be answered to rapidly, using the available data. A typical example of a predictive issue is identifying the market segment targeted. Data mining uses data gathered from the most recent promotional post offers in order to identify the most probable targets for maximizing profits, to be used for the next offers. Other predictive problems include preventing bankruptcy or other forms of failure and identifying population segments reacting similarly to various events (Bălăiță et al., 2003).

Classification is the operation most frequently used, from the plethora of data mining tools. It is a process aiding organizations to identify certain patterns out of large and complex databases with a view to solve specific business problems.

3 Where Can We Apply Data Mining?

Data mining applies to a large category of structured databases in different ways, such as: relational databases, data warehouses, transactional databases (where each record is a transaction), advanced database systems (including multimedia and hypertext data, www. a.s.o.). Thus, the applicability of this informational technology is very wide: from business, medicine, science, engineering to games, media, virtual worlds and satellites. Specialists in the field (Decker, Focardi (1995); Brachman et. al. (1996); Simoudis (1995)) have identified two major areas of applying the data mining technology: business and science (Zaharia et al., 2005).

Data mining may be successfully integrated in economics, where we can distinguish between two main areas of applicability: those specific to industrial services and production and those restricted to certain fields. Within an enterprise, data mining may be applied to varied departments, the results being of great help for the management on all levels.

Marketing is extremely suitable for implementing this technology in the activities related to: statistics, forecasting group behavior, identifying preferences, launching a marketing campaign, advertising in the media etc.

The control activity is also necessary in a field offering a base for applying the data mining processes. Due to the great number of objects that are subject to control (products, clients, distribution channels, regions etc.) as well as to the relevant aspects resulting from this activity, methods automatically analyzing the large amount of data are needed – data mining methods. The resulting patterns are usually a combination of certain attributes and indicators. For example, an interesting result is that the rate of profitability (indicator) for a certain group of products (attributes) is constantly very large. Costs may be subject to data mining analyses as well. Here, we are especially interested in identifying the main business indicators' trends (profit, price, production cost, contribution margin etc.). The results provided to the management by data mining software concerning production may take the shape of answers to questions related to the capacity of using certain equipments arranged in a production flow, discovering unknown interruptions in the production process, identifying unknown dependencies between production quality and the production process (Shahbaz et al., 2010).

Another field is that of internal and external audit. The capacity of data mining systems to automatically identify patterns opens new horizons in allowing control bodies to identify fraud. For example, a phone company may use data mining by identifying unusual patterns in order to discover cell phone cloning frauds. Should a client develop suddenly a different phoning pattern, the company is automatically notified. Data mining provides special advantages to banks due to its financial forecast, control and credit risk evaluation abilities. Some advantages of using the system in the banking activity are presented as follows: it helps in identifying card fraud patterns, it identifies loyal customers, it predicts the behavior of card clients, it provides information related to the expenses of certain groups of clients, it discovers hidden correlations between various financial indicators etc. Last but not least, the information supplied by data mining may be used in loan extension.

The insurance industry may also benefit from using data mining systems in order to foresee which clients are to purchase new insurance policies, to identify behavior patterns of high risk clients, or for fraud-prevention purposes. Data mining may be applied to medicine as well, for instance in characterizing patient behavior or in identifying successful treatments for various diseases. The most famous type of applications of data mining solutions to business is that of a supermarket trying to find correlations between the products sold on the same cash voucher. The system's conclusion was that of a high likelihood for infant diapers to be sold in connection with beer six-packs, especially if the purchase takes place in the evening. Upon deepening the analysis, it turned out that the clients are dads send for urgent diaper purchases, taking advantage of the situation to buy a beer six-pack as well. As a result, the supermarket made sure to insert a beer stand near the diaper department, thus increasing their beer sales (upon converting several diaper buyers in beer buyers) and profitability.

4 Data Mining Techniques

Latent factor analysis - The class of latent analyses is the categorical variable analogue to the much more frequently used factor analysis. Nevertheless, latent analyses hold an advantage over the traditional factor analyses in that the variables may be continuous, categorical (nominal or ordinal), numeric, or any combination of the above.

This procedure allows for a set of mutually exclusive latent classes to be identified, accounting for the consumer distribution. An important use of latent class analysis is researching consumer typologies,

both as an empiric method of characterizing the types of latent consumers within a set of determined indicators, as well as as a method of testing in practice the adequacy of a theoretically formulated typology. If we enter, for the multiple consumer segments, categorical data over much more classical segments, such as those obtained using income and sex as criteria, we can identify the existence of latent variables (for example, the relationship with the service personnel), having a number of different clearly defined latent classes (for instance, the segment of high income women satisfied by the service personnel's behavior or the segment of low income male not satisfied with the behavior of the same personnel). We can also have multiple latent variables able to explain the relationship (the service personnel vs the telephone operating personnel vs the online-service personnel). An latent class analysis may suggest, apart from the information provided on how a process differs for various consumer segments, and the fact that this process was not surprised well enough, explicitly enough in the survey performed.

Marketing applications of latent class analysis include:

- identifying consumer types;
- identifying latent classes explaining data heterogeneity;
- identifying factors determining consumer satisfaction that are relevant for each consumer segment;
- developing complex variables out of the attitude elements within the survey.

Cluster analysis - Cluster analysis is a type of relational model not used for gaining forecast results, but solely for investigating the relationships between variables. Cluster analysis is truly a multivariate method. In researching consumer satisfaction, for instance, most frequently we ask general questions concerning major aspects of the organization's activities. Ideally, we would expect to identify the following groups: a group of customers recording high scores for all major aspects of the organization's activities, an intermediate score group and a low score group. Based on these results we cad determine that group of consumers we must focus our attention on in order to accomplish our aims and improve our activities, as well as identify the specific aspects of activity that need improving. One of the advantages in using cluster analysis is that it is not initiated starting from certain predetermined segments. On the contrary, it starts from undifferentiated segments, and the question is whether a specific group may be divided empirically in significantly differing segments. Cluster analysis is an exploratory tool used for describing, in terms of data collected, in which segment we can fit its members. Due to the fact that cluster analysis doesn't make any assumption concerning the way the population is segmented, its results may contribute to defining a formal classification scheme, much more significant than the pre-determined segments that are usually analyzed, it can suggest patterns useful for describing the populations, it can point out new rules for allocating new consumers to pre-existing segments for identification and diagnosis purposes, it can provide new measures for defining concepts that used to be very broad, or it can identify certain profiles for representing the segments.

Multilevel modeling - Due to the fact that consumer segments have inherent hierarchies or acquire certain qualities leading to a hierarchy based on the company activities' results, the process of analyzing must take into account this hierarchic structure. In other words, the existence of hierarchic data is not an accidental result or a phenomenon worthy neglecting. Segments differ, and this is obvious in the various company activity results (company profitability, for instance), as the organization's results are often a direct effect of the differences existing among the consumer segments. In other cases, the segments may be much more arbitrary, such as allotting consumers to local offices, since, in many cases, consumers having similar demographic background live in those

areas. In this case, analyzing the differences using solely the consumers' location isn't likely to provide us with much information. However, when determining the segments, they tend to differentiate themselves, this differentiation suggesting that a group is empirically different from another. Multilevel models take into account segment ordering, most of the times performed in hierarchic form. More so, multilevel models require more than linear relationship models. In researching consumer satisfaction, for instance, we are interested not only in individual consumers but in the aggregated units as well such as stores, cities, regions etc. We would have normally investigated the effects on an individual level and, in a separate analysis, on an aggregate one. However, multilevel models allow us to understand the influence an individual consumer exerts on an aggregate unit, having impact on the organization's results. In addition, the capacity of multilevel models to perform an inter-classification of segments helps us treat much better the problem of unclear boundaries of segments by recognizing the fact that, for various reasons, people can change over the course of their lives the segments they are a part of. Multilevel modeling allows us to examine the impact consumers exert over the company's activities.

CHAID method - The CHAID method (Chi-square automatic interaction detection) is a method based on the AID (automatic interaction detection) methodology, representing a family of methods used for treating regression-type data much more vigorously than using linear hypothesis method assumptions. CHAID is a useful tool for analyzing multiple variables that are not easily "untangled" one from another, due to the fact that the method is solely searching for interactions.

Although CHAID is usually used in marketing for population segmentation studies, it is at the same time one of the most powerful procedures used for fixed-response questions leading to metric or non-metric replies.

The CHAID decision tree is the main tool used in interpreting the results of CHAID analysis. The tree's "trunk" is the dependent variable under study. The following branches indicate the most adequate predictive variables for the respective dependent variable.

CHAID is an exploratory procedure helping researchers identify and analyze the complex relationship existing between higher order contingency tables by dividing the contingency tables resulting from tabulating three or more variables. This division is performed by determining the lowest number of divisions based on the group's importance, importance used as predictor. Upon performing this task, we turn to the group subcategories in order to identify the adequate predictor and thus continue partitioning, so that no further adequate predictor may be identified. A CHAID analysis may indicate, for instance, that "importance" and "occupation" are significant predictors for the organization's income, that more important clients bring the company higher incomes. However, even if the "importance" and "occupation" are significant predictors for the group of important clients, these predictors may prove to differ in case of the group of consumers bringing lower incomes to the organization. This type of interaction analysis allows for an in-depth investigation of the way in which variables interact with each other in influencing the organization activities' results and in creating multiple profiles for the organization's potential results.

Curvilinear regression - The common curvilinear regression involves a linear relationship between predictors and the result. This is represented graphically by a straight line best describing this type of relationship. We can also see cases in which non-linear relationships are present, case in which the conventional regression analysis would underestimate the predictor-result relationship. Within marketing, we often have data on the basis of which the result may not be forecasted as a straight line.

In this instance, the relationship may be curvilinear. The conventional regression analyses evidence non-linearity through a graphic representation of the residuum.

The curvilinear regression analysis is performed in a hierarchic manner. We start from the linear model to which we progressively add higher order terms in successive steps. Should a term induce a significant change in R^2 (the empiric correlation ratio) over the inferior terms, we can state that the curvilinear relationship is indeed representative for our data. Additionally, multiple trends may coexist within the data, so that we can describe the predictor-result relationship both using a linear as well as a curvilinear relationship. Tim Keiningham and Terry Vavra, authors of „The Delight Principle: Exceeding Customers' Expectations for Bottom-Line Success", have illustrated this effect. As the organization increases the level of satisfaction of extremely dissatisfied consumers up to a state of satisfaction, we can notice a significant increase in satisfaction, from the point in which profits start to stabilize up to the point where clients reach an extreme state of satisfaction, point in which profits once again start increasing.

5 Conclusion

The multivariate mining and exploratory techniques described above may prove to be extremely useful for obtaining supplementary information from the data available, and for identifying patterns within the data, within almost any research performed. The purpose of this material was not to provide an exhaustive list of methods, but to illustrate some of the company activity problems which can be solved by using the various multivariate data exploring and mining methodologies.

Although these techniques are essential tools for the activity of marketers, their adequate use and application involves sophisticated training in the field of statistics for each different technique. However, the marketer holding multiple abilities may aid even the least prepared client in understanding and incorporating the marketing research results within the marketing strategies developed within their organizations.

6 References

Bălăiță, R., Hulea, M., & Olariu, C. (2003). *Data mining*, Virtual Library for Artificial Intelligence, Retrieved from http://eureka.cs.tuiasi.ro/~fleon/bvia.htm.

Ruxanda, Gh. (2005). *Econometrics II*, course notes, Bucharest: Academy of Economic Studies.

Shahbaz, M.; Masood, S.A.; Shaheen, M., & Khan, A. (2010). Data Mining Methodology in Perspective of Manufacturing Databases. *Journal of American Science*, 2010; 6(11), pp. 999-1012.

Zaharia, M.H; Aflori, C.; Șova, I.; Amarandei, C., &Leon, F.(2005). Research Report: The prototype of an intelligent web GIS system for extracting knowledge from a database by using intelligent agents. *The Science Policy and Scientometry Review*, Special issue – 2005, Research report on grant 67/66/2004.