Client Segmentation in Wholesale Markets using Data Mining

Bangert, Patrick[1) 2) 3)]

[1)] algorithmica technologies GmbH, Bremen, Germany
[2)] Advanced International Research Institute on Industrial Optimization GmbH, Bremen, Germany
[3)] Department of Mathematics, University College London, London, United Kingdom

Abstract

A wholesaler wants to have information about the customers that visit his stores as well as those that may potentially visit the stores in order to be able to devise successful, targeted marketing initiatives. Data warehouse systems store data about a client's purchase history but the challenge is to translate these raw data into knowledge that can be used for marketing purposes. Data mining is a branch of mathematics and computer science that aims to provide solutions to this challenge. Hereafter, we demonstrate how to convert transactional data from *a leading international player in self-service wholesale* into a few statements that are understandable, enlightening and useful for marketing purposes. The methods used are mathematical methods that can be run on any similar dataset to automatically and without human intervention produce specific, precise and accurate results.

Introduction

At the retail stores, clients such as hoteliers, restaurants, caterers, canteens, small- and medium size retailers as well as service companies and businesses of all kinds, find everything they need to run their daily business. Every customer has a dedicated membership number and card. Due to this, it is possible to attribute every item sold to a particular customer.

Customer segmentation in general is the problem of grouping a set of customers into meaningful groups based e.g. on their profession or based on their buying behavior. In this particular case, it also allows us to trace which customers belong to these groups because we are aware of their (business) identities. This trace possibility is attempted by many other retailers via loyalty programs in which clients also allow the retailer to attach their identity to the products purchased.

Globally speaking it is interesting to find out buying patterns that can be detected in a certain group of clients. Based on a more detailed description of these groups and investigations on cause and effect for the actions of these groups, it is then possible to adjust the business model to react to such features, for example with targeted advertising such as specific products offerings to specific customers based on their purchasing habits.

Problem Statement and Approach

Consumer behavior investigation has taken place for a dataset of all sold items in two stores over one calendar year. This included over 31 million transactions. *The investigation included no particular questions to be answered and no a priori hypotheses to be confirmed or denied. The goal was to find any structures that might be economically interesting from the point of marketing to these clients.*

The methods used to treat the data were diverse in nature. We used descriptive statistics, non-linear multidimensional regression analyses in all dimensions, k-means clustering and Markov Chain modeling. The aims were as follows:

1 **Descriptive Statistics [1]:** To get an overall feel for the dataset and its various sections as discovered by the other algorithms. This includes correlation analyses. In supporting the Markov chain methodology, this also includes Bayesian prior and posterior distribution analysis, which is able to tell, for example, in which order in time events happen (leads to cause and effect conclusions).

2 **Nonlinear multidimensional regression [2]:** To get a dependency model of the variable among each other. Expressing variables in terms of each other can lead at once to understanding and also dimensionality reduction.

3 **k-means clustering [3]:** To find out which purchases/clients belong into the same phenomenological group and thus determine the actual segmentation that the other methods describe.

4 **Markov Chain modeling [3]:** To model the time-dependent dynamics of the system and thus to find out what stable states exist.

Results

Several descriptive conclusions are available to help with the understanding of the dataset. We present them here in a descriptive format as this is all that is required for understanding the final result. In the actual case study, these conclusions are different and also numerically precise:

1. The total amount of money spent per visit is, statistically speaking, the same per visit for any particular client. Thus, in order to increase total revenues, the key is to increase customer traffic - either by getting a client to come more often or by attracting new customers.

2. Customers will generally go to the store closest to their own locations (in this study their place of business since we are dealing with a wholesaler). The probability of visiting another store decreases exponentially with distance.

3. The high seasonal business is focused in the aftermath of the summer school holidays and the preparations for Christmas. The low seasonal business is focused in the summer school holidays and during the early year post Christmas.

4. The total amount of money spent per year and per visit as well as the number of articles purchased depends highly on the type of client and the geographical region. This has a significant effect upon storage and logistics planning.

5. The majority of clients very rarely shop in the store. There is a core group of clients that shop quite regularly.

6. The products and product groups sold depend strongly on regional effects and on the visit frequency of a customer.

7. Certain products are generally bought in combination with certain other products. Thus, we may speak of a “bag of goods” that is generally bought as a whole. The contents of this bag depend upon the customer group and geography.

8. Via Bayesian analysis and Markov Chain modeling it is possible to deduce that the purchase of a certain product causally leads to the purchase of another product as an effect of the initial purchase. An

example is that a purchase of fresh meat directly leads to the purchase of vegetables, cheese, and other milk products.

To summarize these conclusions, we may say that the customer behavior depends upon geography, product availability, time of the year and certain key products. It was determined that certain factors offer a significant potential in order to improve the profitability of the retail market. Below we present some of those factors (most significant first):

1. **Individual marketing [4]:** Customers tend to be interested in a narrow range of products. It is educational to cluster the customers into interest groups. We find that there are less than 10 clusters that hold a significant number of customers and that are sufficiently heterogeneous in terms of the products they offer to really divide the customers into different groups. These different interest groups could now be treated differently in some ways, e.g. by sending them advertising materials specifically targeted towards their interest group.

2. **Price arbitrage:** In each important product group there is a particular product that is the causal product in the group. This means that *if* the customer buys this product, *then* the customer will also buy a variety of related products in this category. This cause-effect relationship may be used to make this key product more attractive in order to boost sales in the entire product group. One way to do this is to lower the price of the key product. It can be shown that the causal relationship is independent of price changes. However, the identity of the key product is not a universal in that there are regional differences.

3. **Geography:** Most sales are made to customers whose place of business is 20 to 40 minutes away from the store. At an average travel speed of 30 km/h, this is an area of approximately 940 km^2, which is comparable to the size of a moderate sized city. The wholesaler can focus his efforts, e.g. when establishing one-to-one contacts with his customers, in this area. Promotional activities in this area like billboard advertising on major roads may also be effective.

4. **Time of the year:** The main purchase times are March, August and pre-Christmas. The low times are January, February and summer holidays. The rest of the year corresponds to the average purchase activity. The advertising should reflect this trend, focusing on and exploiting the seasonal peaks.

Due to non-disclosure, we have presented the conclusions at a high-level and somewhat modified. The procedures of data-mining are able to output a quantitative presentation of these results (also with uncertainty corridors) that allows these conclusions to act as a firm basis for business decisions.

We note that these conclusions were the result of blind analysis. That is, data of 31 million transactions were given to the mining algorithms without specifying either questions or hypotheses. These algorithms output data that could be interpreted by an experienced human analyst to the above conclusions in just a few hours. Based on these results we may now ask a number of specific questions to make the results clearer, especially when decisions have to be taken to implement changes based on these findings. We will not go into such an interactive question-answer process.

Despite the wish to know more, these conclusions are quite telling and provide valuable material for high level decision making. This illustrates very well the power of data-mining. We have converted a vast collection of data into small number of understandable actionable conclusions that can be presented to corporate management. Moreover, we have been able to do so quickly. This procedure may well be automatically reproduced monthly to track changes to customer behavior. One caveat remains however: The challenge for any data-mining approach in a “bricks & mortar” business is to translate the findings into successful operational business concepts.

Fulfilling the basic reason for being of data-mining: Data and information has quickly become useful and actionable knowledge.

References

[1] Mann, P.S.: Introductory Statistics, 2nd edition. Weinheim: Wiley 1995
[2] Bates, D.M., Watts, D.G.: Nonlinear Regression Analysis and Its Applications. Weinheim: Wiley 1988
[3] Bishop, C.M.: Pattern Recognition and Machine Learning. Heidelberg: Springer 2006
[4] Beyering, L.: Individual Marketing. Landsberg: Verlag Moderne Industrie 1987

Client Segmentation in Wholesale Markets using Data Mining

Dr. Patrick Bangert, CEO
algorithmica technologies GmbH
Außer der Schleifmühle 67, 28203 Bremen, Germany
Tel: +49 (0) 421 337-4646
Email: p.bangert@algorithmica-technologies.com
www.algorithmica-technologies.com