



Customer Segmentation

Dr. Patrick Bangert, algorithmica technologies GmbH

At a major European wholesale retailer, 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 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, for instance, on their profession or on their buying behavior. In this particular case, it also allows us to trace which customers belong to which group, because we are aware of their (business) identities. This kind of 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 being offered to specific customers based on their purchasing habits.

Such an investigation has taken place for a dataset of all sold items in two stores over one calendar year. The investigation 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 simply to find any structures that might be economically interesting from the point of marketing.

We used a variety of methods such as: (1) descriptive statistics, (2) non-linear multi-dimensional regression analyses in all dimensions, (3) k-means clustering and (4) Markov Chain modeling. The aims that we pursued with these methods are described in the following box:

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 – which leads to cause and effect conclusions).
2. Nonlinear multidimensional regression [2]: To see which variables depend on which other variables. (Expressing variables in terms of each other can lead at once to understanding and also reduces the number of independent dimensions.)
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 whether consumer behavior is stable over time or merely temporary).

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 can be made numerically precise:

1. The total amount of money spent per visit is, statistically speaking, the same 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 occurs in the aftermath of the summer school holidays and the preparations for Christmas. The low seasonal business occurs 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 the following factors offer a significant potential for improving the profitability of the retail market (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 all customers can be divided into less than 10 clusters assuming that the clusters should be different from each other and that the members of each cluster should have similar behavior. . These different interest groups could now be treated differently in some ways, e.g. by sending them advertising material 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 (and raise the prices of the non-key products in the same category). 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. This is comparable to a moderately sized city. The wholesaler can focus his marketing efforts 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 (directly after school holidays end) 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 such a general high-level. The procedures of data-mining are able to output a quantitative presentation of these

results (also with uncertainty margins) that allows these conclusions to act as a firm basis for business decisions.

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 is to translate the findings into successful action items.

References

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

