UNIVERSITEIT GENT

FACULTEIT ECONOMIE
EN BEDRIJFSKUNDE

TWEEKERKENSTRAAT 2
B-9000 GENT
Tel. : 32 - (0)9 – 264.34.61
Fax. : 32 - (0)9 – 264.35.92

WORKING PAPER

Data Augmentation by Predicting Spending Pleasure Using Commercially Available External Data

Philippe Baecke[1]

Dirk Van den Poel[2]

June 2009

2009/596

[1] PhD Candidate, Ghent University

[2] Corresponding author: Prof. Dr. Dirk Van den Poel, Professor of Marketing Modeling/analytical Customer Relationship Management, Faculty of Economics and Business Administration, dirk.vandenpoel@ugent.be; more papers about customer relationship management can be obtained from the website: www.crm.UGent.be

D/2009/7012/48

Data Augmentation by Predicting Spending Pleasure

Using Commercially Available External Data

Philippe Baecke and Dirk Van den Poel*

*Ghent University, Faculty of Economics and Business Administration, Department of Marketing, Tweekerkenstraat 2, B-9000 Ghent, Belgium.

Abstract

Since customer relationship management (CRM) plays an increasingly important role in a company's marketing strategy, the database of the company can be considered as a valuable asset to compete with others. Consequently, companies constantly try to augment their database through data collection themselves, as well as through the acquisition of commercially available external data. Until now, little research has been done on the usefulness of these commercially available external databases for CRM. This study will present a methodology for such external data vendors based on random forests predictive modeling techniques to create commercial variables that solve the shortcomings of a classic transactional database. Eventually, we predicted spending pleasure variables, a composite measure of purchase behavior and attitude, in 26 product categories for more than 3 million respondents. Enhancing a company's transactional database with these variables can significantly improve the predictive performance of existing CRM models. This has been demonstrated in a case study with a magazine publisher for which prospects needed to be identified for new customer acquisition.

Keywords: customer relationship management (CRM), data augmentation, commercially available external data, new customer acquisition, random forests, purchase behavior, attitude, spending pleasure.

Corresponding author: Dirk Van den Poel (Dirk.VandenPoel@UGent.be); Tel. +32 9 264 89 80; Fax. + 32 9 264 42 79; website about CRM teaching: http://www.mma.UGent.be website about CRM research: http://www.crm.UGent.be .

1. Introduction

Among business practitioners and marketing scientists today, there has been a shift in focus from the traditional mass marketing to customer relationship management (CRM) (Kannan and Rao, 2001). This is reflected by the expanding number of articles on CRM that have recently been published in the literature (Kamakura et. al., 2005). Earlier, one-to-one marketing was laborious, time-consuming and costly, but in recent years the rise of new media such as the internet enabled companies and their customers to communicate in a more direct manner and exchange information valuable to each other (Van den Poel and Buckinx, 2005). Moreover, the significant drop in costs of data warehousing and the exponential increase in computational power contributed to the fact that plenty of organizations started to acquire transactional data of their clients (Bult and Wansbeek, 1995; Petrison et al., 1993). Consequently, customer databases of huge magnitude are created and processed in order to get more insights into their consumers' buying behavior which should help to improve the marketing strategies.

This process of collecting and analyzing a firm's information regarding customer interaction in order to enhance the customers' value to the firm has been studied extensively in the marketing literature (Kamakura et. al., 2005). Analytical CRM can be used in a variety of stages of the customer lifecycle. Most research has been done on customer churn, which is focused on detecting those customers who have a high probability of leaving the company. This should enable the company to make the correct interventions in order to increase loyalty and prolong the lifetime of a customer. Customer retention has received a lot of attention in the domain ever since it has proven that even a small reduction in customer defection can have a great impact on a firm's profitability (Reichheld and Sasser, 1990; Van den Poel and Larivière, 2004; Gupta et al., 2004). The value of a customer can also be enhanced through customer development activities such as cross-selling and up-selling. Cross-selling is involved with encouraging customers to buy across categories, while up-selling is focused on increasing the demand of customers in existing categories (Prinzie and Van den Poel, 2006 and 2008; Ansari et al., 2000). Besides the immediate profit, both techniques deepen the customer relationship by increasing the share of products that is purchased at the company, thereby increasing the switching costs associated with purchasing from a competitor. Before a company is able to enhance their customer relationship, they first need to attract these customers. Customer acquisition is another stage of the customer life cycle where CRM can contribute useful insights. The objective in this domain is to attract more and profitable customers.

In recent years, academic work in the field of direct marketing has focused on the development, improvement and comparison of new statistical techniques. These techniques are mainly used for segmentation or response modeling. Market segmentation involves dividing the total market into different clusters that are internally homogenous and mutually heterogeneous (Hung and Tsai, 2008). The desires of each cluster should be responded to with separate marketing actions. Response modeling refers to the use of costumer information in order to predict whether a customer will reply to a certain marketing action. Marketers will send mails or catalogs only to those consumers who have a high response probability and spend a large amount of money (Suh et al., 1999). A well-targeted mail increases profit while an irrelevant mail not only increases marketing cost but can also affect the customer-company- relationship in a negative way (Kim et al., 2008). Over the recent years, database marketing techniques have evolved from RFM models (based on the recency, frequency and monetary value of customer purchases) to statistical techniques such as chi-square automatic interaction detection (CHAID) and logistic regression (Bult and Wansbeek, 1995; McCarty and Hastak, 2007). Recently, more advanced machine learning techniques were introduced like support vector machines, neural networks and random forests (Shin and Cho, 2006; Zahavi and Levin, 1997).

Besides the data mining technique that has been used, also the precision and depth of the database will have an important influence on the performance of such a response model and the potential of data analyses to increase profitability. The customer database can be seen as the foundation of CRM which will be used as input for the data mining techniques. The omission of relevant variables can lead to incorrect interpretations and poor predictions. In other words, if the quality of the data is inferior, even the best data mining techniques will still result in mediocre performance (Petrison et al., 1993; Verhoef et al., 2003). As a result, companies constantly try to augment their database through data collection as well as trough the acquisition of commercially available external data.

The remainder of this paper is organized as follows: In Section 2 the limitations of a classic transactional databases are discussed, and commercially available external data is presented as a solution. Section 3 presents the purchase behavior and attitude matrix on which the creation of spending pleasure variables is based. The complete methodology to create these variables is elaborated on in Section 4. Section 5 demonstrates the value of these variables in a case study with a magazine publisher. Finally, conclusions and directions for further research are given in Section 6.

2. External databases as a solution for database limitations

Although companies try to improve their database quality by collecting data themselves, these transactional databases will still suffer from a couple of limitations. First of all, these databases are typically single source in nature. The data collection is limited to the information a company retrieves from their own customers which often results in an inward-looking view of the customer, as competitive information is mostly impossible to obtain. These databases do not capture the purchasing behavior of its customers in the total product category. Hence, the company has no indication about the total potential of each customer (i.e., the total needs of the customer for products in a certain product category) (Buckinx et al., 2007). However, this information can be extremely valuable in several applications. For example, when a company would like to target existing customers in a cross- or up-selling case, this information can help direct marketers to focus their marketing action on clients with a rather low purchase behavior at the company in proportion to their full potential in the product category. The issue of single source data is even a bigger problem when a company wants to attract new customers. Because these persons have never had any contact with the company, no information is available to efficiently target prospects. Secondly, a great part of the data collected by most companies focuses on the past behavior of the individual. Although some authors recognize the predictive value of transaction information summarized in variables such as recency, frequency and monetary value, others remark that relational information should not be ignored and provide important additional insights to the company (McCarty and Hastak, 2007). Focusing only on transactional information is a very sales-oriented approach without understanding the underlying attitudes and motivation of the customer. Such an emphasis may increase sales in the short run, but does not improve the long-term relationship with the customer (Zahay et al., 2004). For example, when a company wants to acquire new customers, it would be easier to attract the customers who are not committed to the product category. This would increase sales at short notice. However, it will be very difficult to build a long relationship with such customers. Hence, in the long run, it would be more profitable to target customers who have a positive attitude towards the product category.

In order to solve these two limitations companies can enhance their databases with commercially available databases sold by external data vendors (Lix et al, 1995). Such databases differ from traditional company databases in two ways. First of all, they contain a large amount of demographic, socio-economic and life style variables for an extensive population. Such non-behavioral information can only be derived by directly questioning the respondent. Due to financial reasons and because it is impossible to reach every respondent in the database at all time, data about every variable is only available for a limited number of respondents in the total database of the external data vendor. A second characteristic of commercially available external databases is the fact that they are not related directly to a specific brand and often not to a specific product category (Lix et al, 1995). In other words, when a particular company wants to enhance his database with commercially available external data, he often has to deal with a lot of irrelevant variables and a large amount of missing values that are difficult to interpret. This is probably an important reason why the majority of the companies do not buy external data for their current customers (Verhoef et al., 2003).

To the best of our knowledge, little research has been done on the usefulness of commercially available external data for customer relationship management. Only Lix et al. (1995) described these databases and explored the linking of them to limited individual based survey data. But in this study they did not focus on the usefulness of the linked variables themselves. Moreover, they did not test the extra value of these variables when used for database enhancement of a company. In this study we will present a methodology that should help external data vendors to create variables that are a solution to the limitations stated earlier. These variables, called spending pleasure variables, have the following advantages. First of all, the variables are created based on a combination of behavioral information and attitudinal information surveyed for a limited number of respondents. The addition of relationship information over behavioral information will help the company to identify customers who are accessible for a long term relationship (Zahay et al., 2004). Secondly, due to financial reasons it is infeasible to obtain this information for each individual respondent in the external database. Consequently, in this study, we will show that it is sufficient to collect this information for only a limited number of respondents and use a predictive data mining technique (i.e. random forests) to extrapolate this spending pleasure information to all respondents in the commercially available external database in a similar way as in the study of Buckinx et al. (2007) and Lix et. al. (1995). As a result predictions for a large amount of respondents were created, which will make the variables more useful in a CRM context. Enhancing a company's single source database with this information will improve the performance of existing CRM models which will have a positive effect on the profitability. Moreover, because this data is also available for customers who are not a client of the company yet, it can also be used to identify prospects for acquisition. Thirdly, the respondents were questioned directly about specific product categories. This variable will be predicted based on a large amount of variables who are not relevant on their own for a direct marketer, but combined they can help to predict the spending pleasure variable, which is much more interpretable by managers.

3. The purchasing behavior and attitude matrix

Previous studies have indicated that transactional information, like RFM variables for example, are very valuable in predicting response and can increase profits in short term, but it would be advisable to also take relational information into account. Although this data have less predictive ability, they may be enormously useful in understanding the underlying tendencies and identifying those customers who are approachable for a long-term relationship (McCarty and Hastak, 2007; Zahay et al., 2004). As a result, we will predict spending pleasure variables based on a two-dimensional matrix including purchasing behavior and attitude in a product category. This matrix, displayed in Figure 1, is constructed in a similar way as in the study of Bandyopadhyay and Martell (2007), where loyalty was split into a behavioral and an attitudinal dimension.

INSERT FIGURE 1 OVER HERE

Based on this matrix we can position every surveyed customer in one of the following four quadrants. There is little that can be done with a person who is a non-user and is characterized with a weak attitude towards a certain product category. This person has no interest at all in the category and little resources should be wasted in an attempt to convince these individuals to buy products from this category. A respondent who spends a lot of money in a product category but doesn’t have any affection with these products will be classified in the functional expenses quadrant. On the other hand, a respondent with a low purchase behavior and a high attitude could potentially become a person with a lot of spending pleasure in the product category, but due to financial reasons his spending power is limited (Rossiter, 1995). This study aims to identify the respondents in the spending pleasure quadrant. These are the most valuable customers in the product category because they are big spenders and committed to the product category.

4. Methodology

4.1. Data description

This study will be based on data of one of the largest external data vendors in Belgium. This database includes about 10000 socio-demographic, economic and life-style variables of more than 3 million respondents. A typical example of the information in the database can be found in Table 1. Such a database is compiled by linking the data of many large, mostly online, surveys. Of course, not every respondent in the database has responded on all surveys. As a result the database suffers from a high number of missing values.

INSERT TABLE 1 OVER HERE

4.2. Data collection

Because no data about the purchase behavior and attitude towards specific product categories was available in the commercially available external database, it was necessary to survey this information from a limited number of respondents. The creation of the purchase behavior construct is similar to the construct used by Dahl et al. (2001). Table 2 represents an overview of the four items measured on a seven point semantic differential scale. This scale consists out of the three RFM items and one item measuring the familiarity with the product category. The attitude construct on the other hand is constructed based on five items displayed in Table 3, also measured on a seven-point semantic differential scale. The selection of these items was based on other studies which involved the measurement of attitude towards a certain product category (Voss et al, 2003; Martin et al., 2001).

INSERT TABLE 2 OVER HERE

INSERT TABLE 3 OVER HERE

4.3. Survey response

In this study spending pleasure variables will be created for 26 product categories. An overview of these product categories can be found in Table 4. Because responsiveness to lengthy questionnaires has decreased and it would be too repetitive to question one respondent for all 26 product categories, the questionnaires were compiled so that the number of surveyed categories was limited to a maximum of five or six per respondent, similar to the one discussed in the study of Kamakura and Wedel (2003).

INSERT TABLE 4 OVER HERE

150000 respondents or about 5% of the total external database were randomly addressed with an online questionnaire about their purchase behavior and attitude in several product categories. 22083 persons responded on questions about at least one product category which results in a response rate of 14.72%. After elimination of bad and inconsistent respondents we maintained on average 3178 respondents per product category.We tested construct reliability per product category by means of Cronbach's coefficient alpha. All coefficients, presented in Table 5, clearly exceed the 0.7 level recommended by Nunnally and Bernstein (1994), which proves we use reliable constructs, especially given the fact that reversed coding was used to measure certain items.

INSERT TABLE 5 OVER HERE

4.4. Identification of spending pleasure respondents

Based on the construct scores we can position every surveyed respondent in the purchase behavior and attitude matrix. By means of a cluster analysis these respondents can be divided into four segments that correspond with the four quadrants discussed earlier in this study. Respondents who are member of the segment with a high purchase behavior and a high attitude towards a certain product category will be classified as having spending pleasure for that product category. This dummy variable will be the dependent variable in the predictive models used to make spending pleasure predictions for all members of the commercially available external database.

4.5. Prediction of spending pleasure using random forests

For the extrapolation of the spending pleasure variables over the total database we opted for the use of random forests, a machine learning technique introduced by Breiman in 2001 based on the principle of a decision tree (Breiman, 2001). A decision tree can be described as a flow of decision rules and their outcomes. Each node corresponds to a variable and a leaf represents a possible value of the target variable. It classifies an example by starting at the root of the tree and moving through the decision nodes until a leaf is reached, which provides the classification of the instance. This is a very popular technique because of its ease and interpretability, which is especially useful in a business context (Duda et al., 2001). Moreover, the technique is flexible in terms of input features and able to handle covariates at different measurement levels. A major drawback of this technique is its instability or lack of robustness (Hastie, Tibshirani and Friedman, 2001). Small variations in data structure or feature space often result in very different series of spits, tree structures and predictions. Random forests solve this problem by combining a large number of trees based on bootstrap sampling techniques with random feature selection techniques. Eventually, the forest chooses the classification having the most votes over all the trees in the forest. Random forests have a number of advantages that are particularly attractive in this study, using a commercially available external database. First of all, this algorithm often outperforms classic predictive techniques like logistic regression (Coussement and Van den Poel, 2008). Secondly, as stated earlier external databases typically contain a lot of variables that have no or little predictive value. This technique has proven that the outcomes of the classifier are very robust when the data contains a lot of noise (Breiman, 2001). Thirdly, in this study we will build a model based on a limited number of respondents including a large number of variables. This can easily lead to over-fitting of the model, but because random forests is based on a large number of subset trees this problem is avoided. Finally, random forests are easy to implement because it includes a good method for estimating missing data and maintains accuracy when a large proportion of the data are missing. Further, there are only two free parameters to set. Based on the suggestions of Breiman (2001) the number of randomly chosen predictors was set equal to the square root of the total number of variables included in the model and 1000 trees were grown per random forests model.

Taking into account that a separate model for each of the 26 product categories has to be created and scored for more than 3 million respondents in the commercially available external database, it is computationally too expensive to include all 10000 variables into the random forests model. Although random forests models have a randomly feature selection embedded in the algorithm and can deal with large feature sets, this database is still too big to work efficiently on. Therefore, preceding the random forests model, a simple maximum-relevance variable selection based on the correlation between the variables and the classification variable is performed to reduce the total number of input variables to 300 per product category, which is computationally more feasible to work with.

4.6. Predictive performance of the resulting models

In order to be able to evaluate the predictive performance of each of the 26 models, each surveyed sample was split into two parts. Firstly, the predictive model is estimated on a training set, containing 70% of the surveyed sample. Afterwards, this model is validated on the remaining 30% of the surveyed sample. It is essential to evaluate the performance of the classifiers on a holdout validation sample in order to ensure that the training model is able to extrapolate the spending pleasure variables well. The area under the receiver operating characteristic curve (AUC) is used as evaluation metric of the classifiers (Hanley and McNeil, 1982). The receiver operating characteristic (ROC) curve is a graphical plot of the sensitivity (i.e. the number of true positives versus the total number of events) and 1-specificity (i.e. the number of true negatives versus the total number of non-events) for all possible cut-off values used. The AUC measures the area under this curve and can range from 0.5, if the predictions are as good as random, to 1, if the model's predictions are perfect. The advantage of an AUC in comparison with other evaluation metrics, like the percent correctly classified (PCC), is the fact that PCC is highly dependent on the chosen threshold. The PCC gives only an indication of the performance at one cut-off, while AUC is a performance metric including all cut-off levels.

Table 6 ranks for each product category the AUC values of the models on the validation sample. The predictive performance varies from 0.6378 to 0.8293 with an average AUC of 0.7279. Apparently the commercially available external database includes more valuable data to identify spending pleasure respondents in product categories as non-profit, active sports, risk investments, and newspapers than to predicting spending pleasure in product categories as grocery, food and drinks, extra insurance and personal hygiene.

INSERT TABLE 6 OVER HERE

Based on these models all respondents in the commercially available database were scored 26 times, once per product category. This results in the creation of 26 variables for 3,218,759 respondents indicating their probability of having spending pleasure in a certain product category.

5. Application

These spending pleasure variables can be very attractive for other companies to enhance their database because they solve a couple of shortcomings. These are easy interpretable variables containing information about the purchase behavior and attitude as well in the total product category, whereas the classic database of a firm is mostly limited to socio-demographic and single-source transactional information between the customer and company. Moreover, the spending pleasures variables are known for a large amount of respondents, also non-customers of the company, and do not include missing values. Enhancing a company's database with such variables will result in a better predictive performance of existing CRM models and increase the profitability. Especially in the case of new customer acquisition, for which buying external data is most popular because companies do not possess data about prospects on their own, these variables can be very valuable (Verhoef et al., 2002). In this section we will use the spending pleasure variables to enhance the database of a monthly issued magazine in order to improve the selection of prospects for new customer acquisition.

5.1. Research question of the company

An application of the spending pleasure variables has been implemented in cooperation with a magazine publisher. This monthly issued magazine is positioned in the market as a magazine specially designed for elderly people and contains information about topics such as law and finance, healthiness, people and opinions, leisure time, lifestyle and multi media. The main target group of this magazine is persons older than fifty. Consequently and not surprisingly, age is an important variable in identifying prospects.

In this study the commercially available external data is used in order to identify prospects with the same profile as the existing magazine subscribers. A logistic regression model will be built in order to predict the magazine subscribers. Based on this model all respondents in the external database will be scored and the company can target a top section of the respondents who are still not a client with the highest probability of being a subscriber. A comparison will be made between the predictive performance of the model based on data excluding and including the spending pleasure variables.

5.2. Methodology

Logistic regression with a stepwise feature selection was chosen in order to solve this binary classification problem because this is a statistical technique that is frequently used in the commercial world and has a better performance than other popular techniques as chi-square automatic interaction detection (CHAID) for example (Verhoef et al., 2002).

The analyses are performed on a database containing 125,434 respondents, consisting of 62,717 existing subscribers and the same amount of randomly chosen non-subscribers. The dependent variable will be the binary variable subscription (i.e. one for the subscribers and zero for the non subscribers). First, a model is built based on 125 socio-demographic independent variables. Subsequently, we enhance this database with 26 spending pleasure variables and compare the predictive performance. Both models will be built on a training sample of 70% of the total database and evaluated on a validation sample, containing the remaining 30% of respondents. AUC will be used to evaluate the predictive performance of both models.

5.3. Results

A model based on only socio-demographic data performs already very well with an AUC of 0.8045 on the validation sample. This was more or less expected since the magazine is positioned as a magazine for elderly people and the socio-demographic data contains several well discriminating age group variables. Despite the fact that this model performs already very well, which makes it more difficult to improve it, enhancing the data with only 26 spending pleasure variables lifts the AUC to a value of 0.8385 on the validation sample. This significant improvement in predictive performance of 0.0340 will result in better predictions of potential subscribers which will increase the success ratio of the acquisition campaign and improve the profitability.

All variables that have a significant influence (alpha = 0.05) on predicting magazine subscribers are presented in Appendix 1, ranked by the absolute values of their standardized betas. Looking at the top of this table it is clear that besides socio-demographic variables like age and gender, the spending pleasure variables contribute extra value to the database in order to improve the model's prediction. This table demonstrates that there is a negative relationship between all the age groups lower than fifty years old and the dependent subscription variable. This confirms the fact that this magazine is positioned as a magazine for elderly people. Also the spending pleasure variables are easily interpretable. Obviously, respondents with a high purchase behavior and attitude toward magazines are more likely to subscribe to this magazine. But also variables as spending pleasure for vacation and omnium insurance have a positive relation with the magazine subscriptions. This is probably due to the topics leisure time as well as law and finance in the magazine.

6. Conclusion and directions for further research.

The emergence of customer relationship management in marketing resulted in the fact that the company's database becomes more and more important to improve customer relationships and attract new customers.

Although companies are able to collect a large amount of data from their own customers, these transactional databases will still suffer from several limitations. Firstly, such databases contain only single source data coming from the company's own customers. They contain no information about non-customers but also not about the purchase behavior of existing customers in the total product category. Secondly, these databases typically contain transactional information about the purchase behavior of the customer, like recency, frequency and monetary value. Including attitudinal information could help to identify the customers who are committed to the product category and more approachable to build up a long term relationship with. These limitations can be solved by enhancing the company's database with commercially available external data. But among business practitioners, these external databases are not always very popular because they suffer also from a couple of drawbacks. Typically, these databases include a lot of missing values and most variables are not related directly to a specific brand or product category. Consequently, these variables are difficult to interpret and not attractive to enhance a company's database with. This study describes a methodology for an external data vendor to create variables that solve all of these limitations. The spending pleasure variables are composed of purchasing behavior and attitude dimension in specific product categories, predicted for a large amount of respondents (customers and non-customers) without missing values.

Such spending pleasure variables were created by questioning a limited number of respondents about their purchase behavior and attitude in a specific product category. By combing these two constructs in a two dimensional matrix respondents in the spending pleasure segment can be identified. This dummy variable is predicted for all respondents in the commercially available database by means of the random forests predictive modeling technique. This results in the creation of 26 spending pleasure variables for more than 3 million respondents. These easily interpretable variables can be very valuable to a company and improve the predictive performance of existing CRM models. This has been demonstrated in a new customer acquisition case for a magazine publisher. Enhancing a predictive model based on socio-demographic variables with spending pleasure variables resulted in a significant increase of the AUC performance.

While we strongly believe that this research paper fills a large gap in today’s literature, there are still some directions for future research. Firstly, this study demonstrates the usefulness of spending pleasure variables in a new customer acquisition case. It would also be interesting to investigate how the spending pleasure variables perform in other contexts of the CRM field, like in cross-selling, up-selling or churn models. Secondly, in this study we predicted only the respondents who are positioned in the spending pleasure segment because these are valuable and interesting respondents for most companies, but in particular cases it could also be useful to identify the respondents who see the product category as a functional expense or have a lack of spending power. It is advisable to target these respondents with different communication strategies than the spending pleasure respondents. For example, customers with a lack of financial resources to spend in the product category but a high attitude towards the product category can still be convinced by offering easy credit facilities for the product.

Acknowledgement

We would like to thank the anonymous external data provider for supporting this research and form making the data available. We are also grateful to Leo Breiman for his inspiring research and the public availability of the Random Forests software.

Appendix 1: Significant socio-demographic and spending pleasure variables in the acquisition model

		Standard	Wald Chi		Standardized
Variable	**Beta**	**Error**	**Square**	**P-value**	**Beta**
age group 36-40	-4.1614	0.1161	1283.8251	0.0000	-0.5146
age group 41-45	-3.4665	0.0862	1618.6206	0.0000	-0.4740
age group 31-35	-4.4344	0.1432	959.1675	0.0000	-0.4594
age group 26-30	-4.6407	0.1966	556.9500	0.0000	-0.3922
age group 22-25	-5.2401	0.4508	135.1342	0.0000	-0.3076
age group 36-50	-1.9967	0.0629	1008.7171	0.0000	-0.2930
age group 18-21	-5.6122	0.9958	31.7604	0.0000	-0.2572
gender: female	0.7544	0.0342	487.4052	0.0000	0.2030
SP for magazines	2.0310	0.0969	439.2208	0.0000	0.1477
SP for vacation	1.1228	0.0584	369.1060	0.0000	0.1086
language: French	-0.3550	0.0340	108.8498	0.0000	-0.0943
number of household members	-0.1245	0.0130	91.1683	0.0000	-0.0840
SP for omnium insurance	1.2624	0.1380	83.6999	0.0000	0.0705
SP for cell phones	-1.5492	0.1353	131.1670	0.0000	-0.0673
SP for clothes	-0.7908	0.0895	78.1094	0.0000	-0.0637
% of unemployed people in the neighborhood	-0.0184	0.0021	76.0148	0.0000	-0.0603
% of higher educated people in the neighborhood	-0.0139	0.0022	39.2756	0.0000	-0.0510
SP for non-profit organizations	0.6718	0.0746	81.1742	0.0000	0.0500
SP for faster internet	0.6155	0.1042	34.9082	0.0000	0.0479
SP for phoning	-0.9283	0.1426	42.3819	0.0000	-0.0473
age group 51-55	-0.2401	0.0462	26.9907	0.0000	-0.0466
number of children between 12 and 15 in the household	-0.2528	0.0472	28.6546	0.0000	-0.0464
SP for consumer credit	1.1327	0.1632	48.2024	0.0000	0.0435
high status factor variable	0.0005	0.0001	25.5749	0.0000	0.0434
age group 71-75	0.2511	0.0324	60.1730	0.0000	0.0387
SP for no risk investments	0.6925	0.1104	39.3654	0.0000	0.0379
number of women between 51 and 55 in the household	0.1891	0.0393	23.1101	0.0000	0.0372
SP for risk investments	1.2698	0.2379	28.4932	0.0000	0.0364
SP for grocery	0.5012	0.1032	23.5987	0.0000	0.0362
number of children between 16 and 17 in the household	-0.2699	0.0535	25.4243	0.0000	-0.0348
age group 66-70	0.2044	0.0326	39.2156	0.0000	0.0343
SP for newspapers	0.4397	0.0764	33.1197	0.0000	0.0330
SP for active sports	-0.4646	0.0810	32.9254	0.0000	-0.0326
director of a private limited company	-0.3520	0.0604	33.9222	0.0000	-0.0313
number of women between 46 and 50 in the household	0.1962	0.0490	16.0042	0.0001	0.0313
presence of a phone	0.1060	0.0195	29.5658	0.0000	0.0292
number of men between 41 and 45 in the household	-0.1936	0.0620	9.7456	0.0018	-0.0268
age group 60-65	0.1399	0.0389	12.9077	0.0003	0.0266
SP for food and drinks	-0.4396	0.0948	21.4944	0.0000	-0.0265
number of men between 46 and 50 in the household	-0.1744	0.0489	12.7215	0.0004	-0.0255
number of children between 6 and 11 in the household	-0.1102	0.0510	4.6582	0.0309	-0.0244
number of men between 76 and 80 in the household	0.1947	0.0381	26.0608	0.0000	0.0235
number of men between 51 and 55 in the household	-0.1328	0.0353	14.1237	0.0002	-0.0233
number of women older than 80 in the household	-0.2053	0.0421	23.8074	0.0000	-0.0216
SP for personal hygiene	0.3253	0.1117	8.4746	0.0036	0.0204
number of men between 18 and 21 in the household	-0.1383	0.0425	10.5884	0.0011	-0.0193
neighborhood with Mediterranean people	-0.1776	0.0547	10.5652	0.0012	-0.0187
head of the household	0.0684	0.0272	6.3214	0.0119	0.0187
SP for extra insurance	0.2939	0.1024	8.2357	0.0041	0.0185

number of women between 61 and 65 in the household	0.0974	0.0362	7.2217	0.0072	0.0182
high carrier people	-0.2971	0.0822	13.0665	0.0003	-0.0182
SP for multimedia	-0.2848	0.0924	9.4949	0.0021	-0.0171
life stage: middle age	0.0002	0.0001	8.5214	0.0035	0.0157
older couples	0.1535	0.0468	10.7533	0.0010	0.0156
number of women between 18 and 21 in the household	-0.1134	0.0444	6.5363	0.0106	-0.0155
Italian roots	-0.1702	0.0685	6.1840	0.0129	-0.0146
urban residential neighborhood	-0.1018	0.0356	8.1651	0.0043	-0.0138
SP for pay TV	0.4235	0.1906	4.9368	0.0263	0.0132
revenue class 1	0.0720	0.0300	5.7672	0.0163	0.0130
semi-urban residential neighborhood	0.0715	0.0280	6.5298	0.0106	0.0123
number of men older than 80 in the household	0.1245	0.0471	6.9806	0.0082	0.0120
revenue class 3	-0.0552	0.0234	5.5467	0.0185	-0.0116
household with teenagers	-0.1068	0.0518	4.2472	0.0393	-0.0105
rural zone	0.0775	0.0382	4.1167	0.0425	0.0101

References

A. Ansari, S. Essegaier and R. Kohli, Internet Recommendation Systems, Journal of Marketing Research 37 (3) (2000), p. 363-375.

S. Bandyopadhyay and M. Martell, Does attitudinal loyalty influence behavioral loyalty? A theoretical and empirical study, Journal of Retailing and Consumer Services 14 (1) (2007), p. 35-44.

L. Breiman, Random Forests, Machine learning 45 (1) (2001), p. 5-32.

G.C. Bruner, P.J. Hensel and K.E. James, Marketing Scales Handbook 4: A Compilation of Multi-Item Measures for Consumer Behavior & Advertising, Ohio: Thomson/South Western (2005).

W. Buckinx, G. Verstraeten and D. Van den Poel, Predicting customer loyalty using the internal transactional database, Expert Systems with Applications 32 (1) (2007), p 125-134.

J.R. Bult, T. Wansbeek, Optimal selection for direct mail, Marketing Science 14 (4) (1995) 378-394.

K.Coussement and D. Van den Poel, Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques, Expert systems with applications 34 (1) (2008), p. 313 -327.

D.W. Dahl, R.V. Manchanda and J..J. Argo, Embarrassment in Customer Purchase: The roles of Social Presence and Purchase Familiarity, Journal of Consumer Research 28 (3) (2001), p.473-481.

R.O. Duda, P.E. Hart and D.G. Stork, Pattern classification, New York: Wiley (2001).

S. Gupta, D.R. Lehmann and J.A. Stuart, Valuing Customers, Journal of Marketing 41 (1) (2004), p. 7-19.

J.A. Hanley and B.J. McNeil, The meaning and use of area under a receiver operating characteristic (ROC) curve, Radiology 143 (1) (1982), p.29-36.

T. Hastie, R. Tibshirani and J. Friedman, The elements of statistical learning: Data mining, inference and prediction, New York: Springer-Verlag (2001)

C. Hung and C. Tsai, Market segmentation based on hierarchical self-organizing map for markets of multimedia on demand, Expert Systems with Applications 34 (1) (2008), p. 780-787.

W. Kamakura, C.F. Mela, A Ansari, A. Bodapati, P. Fader, R. Iyengar, P. Naik, S. Neslin, B. Sun, P.C. Verhoef, M. Wedel and R. Wilcox, Choice Models and Customer Relationship Management, Marketing Letters 16 (3) (2005) p. 279-291.

W.A. Kamakura and M. Wedel, List augmentation with model based multiple imputation: a case study using a mixed-outcome factor model, Statistica Neerlandica 57 (1) (2003), p.46-57.

P.K. Kannan, H.R. Rao, Introduction to the special issue: decision support issues in customer relationship management, Decision Support Systems 32 (2) (2001), P. 83-84.

D. Kim, H. Lee and S. Cho, Response modeling with support vector regression, Expert Systems with Applications 34 (2) (2008), p. 1102-1108.

T.S. Lix, P.D. Berger, T.L. Magliozzi, New Customer Acquisition: Prospecting Models and the Use of Commercially Available External Data 9 (4) (1995), p.8-19.

I.M. Martin and D.W. Stewart, The Differential Impact of Goal Congruency on Attitudes, Intensions, and the Transfer of Brand Equity, Journal of Marketing Research 38 (4) (2001), p471-484.

J.A. McCarty, M. Hastak, Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression, Journal of business research 60 (6) (2007) p. 656-662.

J.C. Nunnally and I.H. Bernstein, Psychometric theory, New York: McGraw-Holl (1994).

L.A. Petrison R.C. Blatteberg and P. Wang, Database marketing past, present and future, Journal of direct marketing, 7 (3) (1993) p. 27-43A. Prinzie and D. Van den Poel, Exploiting Randomness for Feature Selection in Multinomial Logit: A CRM Cross-Sell Application, Lecture Notes in Artificial Intelligence 4065 (2006), p.310-323.

A. Prinzie and D. Van den Poel, Random Forests for Multiclass classification: Random Multinomial Logit, *Expert Systems with Applications*, 34 (3) (2008)

F. F. Reichheld, and W.E. Jr. Sasser, Zero defections: quality comes to services. Harvard Business Review, 68(5) (1990), p105-112.

J.R. Rossiter, “Spending Power” and the Subjective Discretionary Income (SDI) Scale, Advances in consumer research 22 (1) (1995), p. 236-241.

H. Shin and S. Cho, Response modeling with support vector machines, Expert Systems with Applications 30 (4) (2006), p. 746-760.

E.H. Suh, K.C. Noh and C.K. Suh, Customer list segmentation using the combined response model, Expert Systems with Applications 17 (2) (1999), p. 89-97.

D. Van den Poel and W. Buckinx, Predicting online purchasing behaviour, European Journal of Operational Reasearch 166 (2) (2005), p. 557-575.

D. Van den Poel and B. Larivière, Customer attrition analysis for financial services using proportional hazard models, European Journal of Operational Reasearch 157 (1) (2004), p. 196-217.

P. C. Verhoef, P.N. Spring, J. C. Hoekstra and P.S.H. Leeflang, The commercial use of segmentation and predictive modelling techniques for database marketing in the Netherlands, Decision Support Systems 34 (4) (2003), p. 471-481.

K.E. Voss, E.R. Spangenberg and B. Grohmann, Measuring the Hedonic and Utilitarian Dimensions of Consumer Attitude, Journal of Marketing Research 40 (3) (2003), p.310-320.

J. Zahavi and N. Levin, Applying Neural Computing to Target Marketing, Journal of direct marketing 11 (1) (1997), p.5-23.

D. Zahay, J. Peltier, D.E. Schulz and A. Griffen, The Role of Transactional versus Relational Data in IMC Programs: Bringeng Customer Data Together, Journal of advertising research 44 (1) (2004), p.3-18.

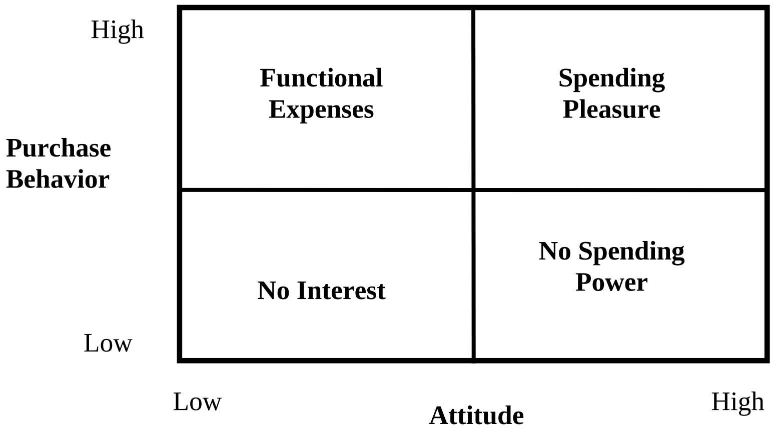

Figure 1: The purchase behavior and attitude matrix.

Socio- demographic	Economic	Lifestyle
Age group	Vehicle information	Leisure activities
Number of household members	Newspaper and magazine subscriptions	Favourite radio and television station
Gender	Average telecom payments	Sports
Life stage	Bank accounts	Favourite products
Social class	House information	Cultural intrests

Table 1: Example of commercially available external data.

In comparison with your friends, how often do you purchase (product category)?	
	never or very rare - very often
How familiar are you with the purchase of (product category)?	
	not experienced - very experienced
In comparison with your friends, when was the last time you purchased (product category)?	
	very recent - never or very long ago (reversed)
In comparison with your friends, how much money do you spend at (product category)?	
	no or little money - a lot of money

Table 2: The purchase behavior items.

I consider (product category) to be:	
	unpleasant – pleasant
	exciting – dull (reversed)
	awful - delightful
	enjoyable – unenjoyable (reversed)
	boring – fun

Table 3: The attitude items.

- active sports
- cars
- cell phone
- cleaning products
- clothes
- consumer credit
- culture
- decoration
- extra insurance
- faster internet
- food and drinks
- grocery
- magazines
- multimedia
- multimedia equipment
- newspapers
- non-profit
- no-risk investments
- omnium insurance
- passive sports
- pay-tv
- personal hygiene
- phoning
- risk investments
- vacation
- wellness

Table 4: Product categories overview.

	Cronbach's Alpha	
Product category	**Purchase behavior**	**Attitude**
active sports	0.9146	0.9368
cars	0.8202	0.8681
cell phone	0.8192	0.8914
cleaning products	0.8591	0.9374
clothes	0.8086	0.8988
consumer credit	0.8664	0.8993
culture	0.9296	0.9619
decoration	0.8530	0.9424
extra insurance	0.8814	0.9284
faster internet	0.8534	0.9079
food and drinks	0.7497	0.8976
grocery	0.7367	0.9003
magazines	0.9003	0.9151
multimedia	0.8999	0.9236
multimedia equipment	0.8758	0.9257
newspapers	0.9386	0.9447
non-profit	0.9362	0.9463
no-risk investments	0.9395	0.9491
omnium insurance	0.9212	0.9109
passive sports	0.9082	0.9612
pay-tv	0.9282	0.9497
personal hygiene	0.8831	0.9288
phoning	0.8270	0.9420
risk investments	0.9351	0.9303
vacation	0.8393	0.8693
wellness	0.9535	0.9679

Table 5: Cronbach’s Alphas per product category.

Product category	AUC values
non-profit	0.8193
active sports	0.8177
risk investments	0.8149
newspapers	0.8121
passive sports	0.7933
culture	0.7784
wellness	0.7559
pay-tv	0.7455
phoning	0.7437
multimedia equipment	0.7327
clothes	0.7284
consumer credit	0.7283
vacation	0.7230
omnium insurance	0.7194
cars	0.7095
faster internet	0.7073
multimedia	0.7029
no-risk investments	0.6968
cleaning products	0.6958
magazines	0.6956
cell phone	0.6951
decoration	0.6837
personal hygiene	0.6751
extra insurance	0.6669
food and drinks	0.6458
grocery	0.6378
Average AUC	**0.7279**

Table 5: Predictive performance (in terms of AUC) per product category.