Journal of Business Research xxx (2012) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Journal of Business Research

Data accuracy's impact on segmentation performance: Benchmarking RFM analysis, logistic regression, and decision trees

Kristof Coussement [a,*], Filip A.M. Van den Bossche [b,c], Koen W. De Bock [a]

[a] IESEG School of Management, Université Catholique de Lille (LEM, UMR CNRS 8179), Expertise Center for Database Marketing (ECDM), Department of Marketing, 3 Rue de la Digue, F-59000, Lille, France
[b] Hogeschool-Universiteit Brussel, Faculty of Economics and Management, Warmoesberg 26, B-1000 Brussels, Belgium
[c] Katholieke Universiteit Leuven, Faculty of Business and Economics, Naamsestraat 69, B-3000 Leuven, Belgium

ARTICLE INFO

Article history:
Received 1 February 2012
Received in revised form 1 July 2012
Accepted 1 September 2012
Available online xxxx

Keywords:
Customer segmentation
Direct marketing
Data quality
Data accuracy
RFM
Decision trees

ABSTRACT

Companies greatly benefit from knowing how problems with data quality influence the performance of segmentation techniques and which techniques are more robust to these problems than others. This study investigates the influence of problems with data accuracy – an important dimension of data quality – on three prominent segmentation techniques for direct marketing: RFM (recency, frequency, and monetary value) analysis, logistic regression, and decision trees. For two real-life direct marketing data sets analyzed, the results demonstrate that (1) under optimal data accuracy, decision trees are preferred over RFM analysis and logistic regression; (2) the introduction of data accuracy problems deteriorates the performance of all three segmentation techniques; and (3) as data becomes less accurate, decision trees retain superior to logistic regression and RFM analysis. Overall, this study recommends the use of decision trees in the context of customer segmentation for direct marketing, even under the suspicion of data accuracy problems.

1. Introduction

Nowadays, increased digitization of transactions results in a boost of customer information stored in large transactional databases (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). This evolution has led to the emergence of the database marketing domain as a popular discipline in academic research and business practice (Ko, Kim, Kim, & Woo, 2008). A prominent database marketing application is customer segmentation for direct marketing, where the analyst tries to find homogeneous groups of customers with respect to their response behavior by means of so called data-mining tools (Akaah, Korgaonkar, & Lund, 1995; Cortinas, Chocarro, & Villanueva, 2010; McCarty & Hastak, 2007; Merrilees & Miller, 2010; Morganosky & Fernie, 1999). The usage of data-mining tools in direct marketing is subject to the knowledge discovery in databases (KDD) process, of which the growing importance is reflected by the large number of publications and applications in both academia and business (e.g. Bose & Mahapatra, 2001).

KDD prescribes a multi-level process to derive valuable top-level strategic insights from low-level raw data (Fayyad et al., 1996). A typical KDD process consists of the following five consecutive steps: (1) problem identification, in which the application domain is defined and objectives are formulated; (2) data preparation, or selecting, preprocessing, reducing, and transforming the data; (3) data mining, or choosing and applying an appropriate analysis technique; (4) the analysis, evaluation, and interpretation of results; and (5) presentation, assimilation, and use of knowledge (Han & Kamber, 2006; Martínez-López & Casillas, 2009).

Although the success of implementing a KDD process depends on the value of each of its five constituent steps (Crone, Lessmann, & Stahlbock, 2006; Fayyad et al., 1996), a significant proportion of recent research in direct marketing has unilaterally focused on the data-mining phase and its tools for segmenting customers (e.g. RFM (recency, frequency, and monetary value) analysis McCarty & Hastak, 2007, logistic regression McCarty & Hastak, 2007, decision trees Haughton & Oulabi, 1993 and more advanced techniques, such as artificial neural networks Zahavi & Levin, 1995, 1997, support vector machines Viaene et al., 2001, and genetic fuzzy systems Martínez-López & Casillas, 2009).

In addition to the choice of the best segmentation tool, data quality (DQ) is an equally important concept in customer analytics (Feelders, Daniels, & Holsheimer, 2000; Ko et al., 2008). Prior research shows that bad data yield bad analytical results, often referring to this process as the "garbage in, garbage out" principle (Baesens, Mues, Martens, & Vanthienen, 2009). Consider a marketing department of a direct marketing company that wants to profile its segmented customers according to their monetary value, i.e. their past total money spent at the company. If parts of the monetary value figures are not

* Corresponding author. Tel.: +33 320545892.
E-mail addresses: K.Coussement@ieseg.fr (K. Coussement), Filip.VandenBossche@hubrussel.be (F.A.M. Van den Bossche), K.DeBock@ieseg.fr (K.W. De Bock).

0148-2963/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbusres.2012.09.024

correct, the uncertainty of calculating the correct average monetary value per segment increases, and consequently the information quality and the segmentation performance decrease. DQ is often considered as a multi-dimensional construct having four subcategories (Wang and Strong, 1996); (1) *intrinsic DQ* denoting that data have quality in their own right, (2) *contextual DQ* referring to the fact that DQ should be considered within the context of the task at hand, (3) *representational DQ* and (4) *accessibility DQ* both linked to the importance of the information system(s).

Although many DQ attributes in each of these four subcategories have been introduced in the literature, this study focuses on how segmentation performance is impacted by the intrinsic DQ attribute *accuracy* which is defined as conformity with the real world (Wand & Wang, 1996). Three arguments are given to motivate the need for investigating the impact of data accuracy on segmentation performance. First, data accuracy is one of the well-documented attributes in the DQ literature. Second, data inaccuracy can be simulated and its impact is measureable in an objective manner, something which is impossible for other more subjective dimensions of data quality. Third, no research is available that investigates the impact of data accuracy problems upon segmentation performance in a direct marketing setting.

In the KDD process, DQ results from choices made in the data-preparation phase (Fayyad et al., 1996). The data-preparation phase consists of the following sub-steps: (i) data selection, aimed at the selection of relevant information while minimizing noise; (ii) data preprocessing; and (iii) data reduction and transformation. Previous research mainly focuses on strategies to improve DQ within the preprocessing and transformation phases and discusses topics such as feature selection (Kim, Street, Russell, & Menczer, 2005), re-sampling (Chawla, Bowyer, Hall, & Kegelmeyer, 2002), outlier detection (Van Gestel et al., 2005), the discretization of continuous attributes (Berka & Bruha, 1998), and the mapping and scaling of categorical variables (e.g., Zhang, Zhang, & Yang, 2003). However, the data selection phase has a significant impact on DQ, and thus on its impact on segmentation performance. Problems that may arise here are missing values, outdated data values and inaccurate data (Even, Shankaranarayanan, & Berger, 2010). Several authors investigate the merits of missing value imputation in the KDD process (e.g., Batista & Monard, 2003; Brown & Kros, 2007), but research on the direct impact of inaccurate and outdated data on the performance of segmentation models for direct marketing is not available. In summary, the objectives of this study are to assess the impact of data accuracy problems on the quality of customer segmentation approaches and to uncover whether some segmentation techniques are more resistant to these problems than others.

The paper has the following organization. Section 2 describes the segmentation approaches and the evaluation metric. Section 3 describes the experimental setup of this study, and Section 4 describes the impact of data accuracy problems on the segmentation performance for real-life direct marketing data. Section 5 revises the managerial implications of the impact of poor data accuracy, and finally Section 6 summarizes the results and offers suggestions for further research.

2. Methodology

2.1. Segmentation approaches

This paragraph details the segmentation techniques employed throughout this study. The impact of data accuracy issues on the performance of RFM analysis, decision trees, and logistic regression is evaluated. These techniques are chosen given their popularity in business and their extensive use in the direct marketing literature (e.g. McCarty & Hastak, 2007).

2.1.1. RFM analysis

RFM analysis originates from the practice of direct marketing in catalog sales companies in the 1960s (Blattberg, Kim, & Neslin, 2008). Such analysis prescribes a segmentation of customers in the company's database based on past behavior (Bitran & Mondschein, 1996; Hughes, 2000). Three variables represent this past behavior: (1) the period elapsed since the customer's last purchase (i.e., *recency*; R), (2) the number of purchases in an arbitrary period in the past (i.e., *frequency*; F), and (3) the total monetary value of past purchases (i.e., *monetary value*; M). Customers who purchased recently, frequently, and spent large amounts of money are more likely to respond to mailings and therefore represent more attractive prospects for future marketing campaigns. The objective of RFM analysis is to identify a segment of customers who have a high probability of responding to a marketing campaign. By focusing on these customers and avoiding spending resources on customers who would not have responded anyway, a company makes its marketing actions more targeted.

Although several RFM model variations exist in literature, this research relies on the RFM procedure proposed by Hughes (1996, 2000). In this approach, the RFM variables transform into discrete codes that take values in the set {1, 2, 3}. Thus, every customer receives one code for every RFM variable. In detail, the attributed codes come from the following procedure. The first step summarizes customers' purchase histories in RFM variables. This historical purchase information is easily derivable from a company's transactional database in which all past customer purchases before a direct mail campaign are recorded. The second step sorts customers on recency, divides them into three equal customer groups, and assigns them one of the three discrete codes. For example, the 33.33% of customers who purchased most recently receive code 3. The third step, within each recency group, sorts customers on their purchase frequency, attributing the codes in a similar way. Finally, the fourth step sorts each frequency group on monetary value and again attributes the codes to the subsets. From this procedure, every customer receives three codes that indicate membership to one of 27 ($3\times3\times3$) groups of equal size. The analyst then concatenates different codes for recency, frequency, and monetary value of customers and uses them to rank the customers. This ranking allows marketing decision makers to formulate a set of rules that helps to identify customers who should be targeted in a direct mail campaign. In real-life, several criteria can guide the selection of the correct number of customers to target. For example, the analyst could take an arbitrary proportion of the customer file or determine an optimal number of target customers to maximize the company's profits (Blattberg et al., 2008).

On the one hand, RFM analysis is a popular approach in database marketing because of its simplicity and reasonable performance. The relationship between the response and the RFM variables is not assumed to be monotonic or known in advance (McCarty & Hastak, 2007). On the other hand, several important disadvantages exist. First, the discretization procedure introduces a loss of explanatory information. Second, the customer coding procedure is arbitrary. Depending on the case, more or fewer categories might be more appropriate. For instance, depending on the budget, finer or cruder RFM coding schemes could be employed (Blattberg et al., 2008; Hughes, 1996). Finally, the technique is not suited to add other features that might relate to a customer's future response behavior (Blattberg et al., 2008).

2.1.2. Decision trees

Because of their combination of simplicity, transparency, and strong performance, decision trees are a popular modeling technique in business (Duda, Hart, & Stork, 2001). In the context of customer segmentation, the analyst constructs a decision tree by subsequently splitting the entire group of heterogeneous customers into smaller and more homogeneous subsets of customers. The top of the decision tree, or the node in which all customers enter the model, is the “root node”. Splits are made into two or more child nodes according to the values of one or more independent variables. In particular, the algorithm identifies a

Please cite this article as: Coussement, K., et al., Data accuracy's impact on segmentation performance: Benchmarking RFM analysis, logistic regression, and decision trees..., *Journal of Business Research* (2012), http://dx.doi.org/10.1016/j.jbusres.2012.09.024