Available online at www.sciencedirect.com

ScienceDirect

Journal of Business Research 60 (2007) 656–662

JOURNAL OF BUSINESS RESEARCH

Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression

John A. McCarty [a,*], Manoj Hastak [b,1]

[a] School of Business, The College of New Jersey, Ewing, New Jersey 08628, United States
[b] Kogod School of Business, American University, Washington, DC 20008, United States

Received 1 November 2005; received in revised form 1 April 2006; accepted 1 June 2006

Abstract

Direct marketing has become more efficient in recent years because of the use of data-mining techniques that allow marketers to better segment their customer databases. RFM (recency, frequency, and monetary value) has been available for many years as an analytical technique. In recent years, more sophisticated methods have been developed; however, RFM continues to be used because of its simplicity. This study investigates RFM, CHAID, and logistic regression as analytical methods for direct marketing segmentation, using two different datasets. It is found that CHAID tends to be superior to RFM when the response rate to a mailing is low and the mailing would be to a relatively small portion of the database, however, RFM is an acceptable procedure in other circumstances. The present article addresses the broader issue that RFM may focus too much attention on transaction information and ignore individual difference information (e.g., values, motivations, lifestyles) that may help a firm to better market to their customers.

Keywords: Database marketing; Data-mining; RFM; CHAID; Analytical procedures

1. Introduction

Segmentation in direct marketing has become more efficient in recent years because of the development of database marketing techniques. These data-mining approaches provide direct marketers with better ways to segment their current customers and develop marketing strategies tailored to particular segments and/or individuals. Over the recent years, database marketing techniques have evolved from simple RFM models (models involving recency of customer purchases, frequency of their purchases, and the amount of money they have spent with the firm) to statistical techniques such as chi-square automatic interaction detection (CHAID) and logistic regression. More recently, neural network models are employed in the database marketing arena (Yang, 2004).

In spite of recent statistical advances in data-mining, marketers continue to employ RFM models. A study by Verhoef et al. (2002) shows that RFM is the second most common method used by direct marketers, after cross tabulations, in spite of the availability of more statistically sophisticated methods. There are a couple of related reasons for the popularity of RFM. As Kahan (1998) notes, RFM is easy to use and can generally be implemented very quickly. Furthermore, it is a method that managers and decision makers can understand (Marcus, 1998). This is an important consideration in that a successful technique for a direct marketer is one that differentiates likely responders to a particular mailing from those who are unlikely to respond, yet does so in a way that is easy to explain to decision makers. However, it has been argued that the simplicity of RFM has been overemphasized, but its ability to differentiate, relative to statistical techniques, has not been considered to the extent that it should be (Yang, 2004).

Although the efficiency of RFM has been questioned, little research documents its ability relative to newer statistical techniques. This paucity of research is partly because RFM refers to a general approach to data-mining; there are a variety

[*] Corresponding author. Tel.: +1 609 771 3220.
E-mail addresses: mccarty@tcnj.edu (J.A. McCarty), mhastak@american.edu (M. Hastak).
[1] Tel.: +1 202 885 1973.

0148-2963/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbusres.2006.06.015

of ways of applying the use of recency, frequency, and monetary value. Research that has been conducted on the efficacy of RFM generally focuses on proprietary or judgmental models of RFM (e.g., Levin and Zavari, 2001; Magidson, 1988) and not on empirically based RFM models. More recently, research has moved away from RFM and has focused instead on newer, more sophisticated approaches to data-mining (c.f., Deichmann et al., 2002; Linder et al., 2004). The current study evaluates one popular, empirically based (as opposed to judgmental) approach to RFM. This RFM approach is compared to CHAID and logistic regression, in an effort to understand its capabilities as a database marketing analytical tool.

2. Analytical segmentation methods in data-mining

2.1. RFM analysis

Recency, frequency, and monetary (RFM) analysis has been used in direct marketing for a number of decades (Baier et al., 2002). This analytical technique grew out of an informal recognition by catalog marketers that three variables seem particularly related to the likelihood that customers in their house datafiles would respond to specific offers. Customers who recently purchased from a marketer (recency), those who purchase many times from a marketer (frequency), and those who spend more money with a marketer (monetary value) typically represent the best prospects for new offerings.

As noted, RFM analysis is utilized in many ways by practitioners, therefore, RFM analysis can mean different things to different people. One common approach to RFM analysis is what is known as hard coding (Drozdenko and Drake, 2002). Hard coding RFM is a matter of assigning a weight to each of the variables recency, frequency, and monetary value, then creating a weighted score for each person in the database. The assignment of weights is generally a function of the judgment of the database marketers with a particular database; for example, past experience may tell a marketer that recency should weigh twice as much as frequency and monetary value. Therefore, this application of RFM is often referred to as judgment based RFM. The weightings could also vary as a function of the particular mailing (Baier et al., 2002). The weights can, of course, be empirically derived based on offerings mailed to database members in the past, thus relying on previous data rather than judgments.

Regardless of the way that RFM is utilized, there are two common characteristics of RFM procedures. First, RFM is used to segment a house file (i.e., a company's current customers) using information related to recency, frequency, and monetary value. RFM is not applicable to the prospecting for new customers because a marketer would not have transaction information for prospects. Second, RFM analysis generally focuses on the three behavioral variables of recency, frequency, and monetary value. Although these variables are considered powerful predictors of future behavior, traditional RFM is limited to these three things.

A well known, empirically based RFM method is a procedure advocated by Arthur Hughes (2000). Hughes' approach is applicable in instances when a marketer intends to send a mailing to customers in its database and would like to find those in the database who are the most likely to respond to the specific mailing. Hughes recommends a test mailing to a sample of customers in the file; then the selection of the members of the rest of the file is made as a function of the results of the test. Thus, compared with hard coding RFM, Hughes' method is not arbitrary with respect to the weighting of recency, frequency, and monetary value. The importance of each of these is determined by the test mailing for the particular offer.

The first step in the method is for the marketer to sort the customer file according to how recently customers have purchased from the firm. The database is then divided into equal quintiles and these quintiles are assigned the numbers 5 to 1. Therefore, the 20% of the customers who most recently purchased from the company are assigned the number 5; the next 20% are assigned the number 4, and so on. The next step involves sorting the customers within each recency quintile by how frequently they purchase from the marketer. For each of these sorts, the customers are divided into equal quintiles and assigned a number of 5 to 1 for frequency. Each of these groups (25 groups) is sorted according to how much money the customers have spent with the company. These sorts are divided into quintiles and assigned numbers 5 to 1. Therefore, the database is divided into 125 roughly equal groups (cells) according to recency, frequency, and monetary value.

Hughes recommends conducting a test mailing to a randomly sampled subset of each cell (e.g., 10%). After the responses of the test mailing are received, the proportion of respondents in each cell can be calculated. The cells can then be ordered as a function of response percent. The marketer can then elect to mail to a certain portion of the remaining file (e.g., the top 20% of the cells). Alternatively, the marketer can elect to mail to the cells that are above a break even percent, given the cost of the mailing and the expected revenue for each return. For example, if a mailing costs $1.50 and the revenue received is $50.00 per order, the break even percentage would be 3%. Thus, for the 90% of the file that is left after the test mailing, the direct marketer would mail to the RFM cells that the test mailing predicted a 3% or better return.

It is important to note that Hughes' method does not assume a monotonic relationship between the dependent variable (responded/did not respond) with the variables of recency, frequency, and monetary value. Each cell is a discreet group that is considered individually in terms of its performance. Thus, if middle levels of one of the independent variables (e.g., frequency) are more related to response compared with higher or lower levels of this variable, then the procedure can accommodate the non-monotonic nature of the relationship.

2.2. CHAID

Chi Square Automatic Interaction Detector (CHAID) (see, for example, Sargeant and McKenzie, 1999) is a method of database segmentation that has been used for a number of years. Research has shown that CHAID is superior to judgment based RFM with respect to the identification of likely responders