Decision Support Systems 54 (2012) 443–451

Contents lists available at SciVerse ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

Direct marketing decision support through predictive customer response modeling

David L. Olson [a], Bongsug(Kevin) Chae [b,*]

[a] Department of Management, University of Nebraska, Lincoln, NE 68588-0491, United States
[b] Department of Management, Kansas State University, Manhattan, KS 66506, United States

ARTICLE INFO

Article history:
Received 8 June 2011
Received in revised form 12 May 2012
Accepted 19 June 2012
Available online 3 July 2012

Keywords:
Customer response predictive model
Knowledge-based marketing
RFM
Neural networks
Decision tree models
Logistic regression

ABSTRACT

Decision support techniques and models for marketing decisions are critical to retail success. Among different marketing domains, customer segmentation or profiling is recognized as an important area in research and industry practice. Various data mining techniques can be useful for efficient customer segmentation and targeted marketing. One such technique is the RFM method. Recency, frequency, and monetary methods provide a simple means to categorize retail customers. We identify two sets of data involving catalog sales and donor contributions. Variants of RFM-based predictive models are constructed and compared to classical data mining techniques of logistic regression, decision trees, and neural networks. The spectrum of tradeoffs is analyzed. RFM methods are simpler, but less accurate. The effect of balancing cells, of the value function, and classical data mining algorithms (decision tree, logistic regression, neural networks) are also applied to the data. Both balancing expected cell densities and compressing RFM variables into a value function were found to provide models similar in accuracy to the basic RFM model, with slight improvement obtained by increasing the cutoff rate for classification. Classical data mining algorithms were found to yield better prediction, as expected, in terms of both prediction accuracy and cumulative gains. Relative tradeoffs among these data mining algorithms in the context of customer segmentation are presented. Finally we discuss practical implications based on the empirical results.

1. Introduction

The role of decision support techniques and models for marketing decisions has been important since the inception of decision support systems (DSSs) [25]. Diverse techniques and models (e.g., optimization, knowledge-based systems, simulation) have emerged over the last five decades. Many marketing domains, including pricing, new product development, and advertising, have benefited from these techniques and models [16]. Among these marketing domains, customer segmentation or profiling is recognized as an important area [18,19,26,43]. There are at least two reasons for this. First, the marketing paradigm is becoming customer-centric [41] and targeted marketing and service are suitable. Second, unsolicited marketing is costly and ineffective (e.g., low response rate) [15,30]. Along with these reasons, there are increasing efforts on collecting and analyzing customer data for better marketing decisions [9,26,30]. The advancement of online shopping technologies and database systems has accelerated this trend.

Data mining has been a valuable tool in this regard. Various data mining techniques, including statistical analysis and machine learning algorithms, can be useful for efficient customer segmentation and targeted marketing [4,26,38]. One such technique is RFM, standing for recency, frequency, and monetary. RFM analysis has been used for marketing decisions for a long time and is recognized as a useful data mining technique for customer segmentation and response models [3,30]. A survey [43] also shows that RFM is among the most popular segmentation and predictive modeling techniques used by marketers.

RFM relies on three customer behavioral variables (how long since the last purchase by customer, how often the customer purchases, how much the customer has bought) to find valuable customers or donors and develop future direct marketing campaigns. Having a reliable and accurate customer response model is critical for marketing success since an increase or decrease in accuracy of 1% could have a significant impact on their profits [1]. While there could be many other customer-related factors [e.g.,42], previous studies have shown that RFM alone can offer a powerful way of predicting the future customer purchase [1,3,17].

Our research builds customer response models using RFM variables and compares them in terms of customer gains and prediction accuracy. The paper aims to increase understanding of how to find knowledge hidden in customer and transactional databases using data mining techniques. This area is called knowledge-based marketing [26]. The next section briefly reviews various data mining techniques for building customer response or predictive models. Section 3 describes methodology. All the response models will be built upon the three RFM variables, while different data mining techniques are used. Then, we present a research design, including two direct marketing data sets with over 100,000 observations, a process of predictive modeling building, and methods to measure the performance of models. Section 4 includes analysis and results. There could be different methods to increase the

* Corresponding author.
E-mail addresses: Dolson3@unl.edu (D.L. Olson), kevinbschae@gmail.com (B.(K.) Chae).

0167-9236/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2012.06.005

prediction performance of an RFM-based predictive model and sophisticated data mining techniques (decision tree, logistic regression, and neural networks) appear to outperform more traditional RFM. These findings are further discussed in Section 5, comparing results with previous studies of customer response models and in the broad contexts of knowledge-based marketing. We also discuss practical implications from the findings and offer conclusions.

The contribution of this study is to demonstrate how RFM model variants can work, and supports general conclusions consistently reported by others that RFM models are inferior to traditional data mining models. This study shows that RFM variables are very useful inputs for designing various customer response models with different strengths and weaknesses and the ones relying on classical data mining (or predictive modeling) techniques can significantly improve the prediction capability in direct marketing decisions. These predictive models using RFM variables are simple and easy to use in practice than those with a complex set of variables. Besides descriptive modeling techniques popular in practice [43], thus, marketers should adopt those advanced predictive models in their direct marketing decisions.

2. Customer response models using data mining techniques

2.1. Marketing DSS and customer response models

The use of DSS in marketing goes back to the 1960s and 1970s [22,44] and has been applied in various areas, including marketing strategy, pricing, new product development, and product analysis and management [16]. There has been an increase of DSS use in customer-side marketing activities, such as customer segmentation (or profiling), direct marketing, database marketing, and targeted advertising. This reflects advances in database management and complex model building [11,16,35]. More convenient methods are available for the acquisition and storage of large amounts of customer and transactional data. In addition, knowledge-based systems or intelligent systems using data mining techniques (e.g., neural networks) [37] have emerged in the marketing domain.

This trend is broadly termed knowledge-based marketing. Knowledge-based marketing is both data-driven and model-driven: that is the use of sophisticated data mining tools and methods to find knowledge discovery from customer and transactional databases [26]. Overall, this leads to more efficient and effective communication with potential buyers and an increase in profits. An important approach to knowledge-based marketing is to understand customers and their behavioral patterns. This requires such transactional characteristics as recency of purchases, frequency of purchases, size of purchases, identifying customer groups, and predicting purchases [35]. The RFM model and other data mining-based customer response models have proven useful to marketers.

2.2. Data mining techniques for customer response models

2.2.1. RFM

R represents the period since the last purchase. F is the number of purchases made by a customer during a certain period. M is the total purchase amount by a customer over that period. It is common practice for each R, F, and M to have five groups or levels and thus there are 125 ($=5*5*5$) customer segmentation groups. Each customer is segmented into one cell or group. This model allows markets to differentiate their customers in terms of three factors and to target the customer groups that are likely to purchase products or services. This technique is known as the benchmark model in the area of database marketing [3].

Since its introduction in a major marketing journal [5], RFM has received a great deal of interest from both academic and industry communities [3,17]. Many studies [1,13,17] have recognized these three variables as important to predict the future responses by customers to potential direct marketing efforts. Certain limitations in the original RFM model have been recognized in the literature [31,45]. Some previous studies have extended the original RFM model either by considering additional variables (e.g., socio-demographics) [1] or by combining with other response techniques [6,7]. Because of the high correlation between F and M, Yang [45] offered a version of RFM model collapsing the data to a single variable "Value" = M/R. To overcome the problem of data skewed in RFM cells, Olson et al. [31] proposed an approach to balance observations in each of the 125 RFM cells.

Other variables that may be important include customer income, customer lifestyle, customer age, product variation, and so on [14]. That would make traditional data mining tools such as logistic regression more attractive. However, RFM is the basis for a continuing stream of techniques to improve customer segmentation marketing [12]. RFM has been found to work relatively well if expected response rate is high [24]. Other approaches to improve RFM results have included Bayesian networks [1,8] and association rules [46].

2.2.2. Classical data mining tools

Common data mining practice in classification is to gather a great number of variables and apply different standard algorithms. Given the set of predefined classes and a number of attributes, these classification methods can provide a model to predict the class of other unclassified data. Mathematical techniques that are often used to construct classification methods are binary decision trees, neural networks, and logistic regression. By using binary decision trees, a tree induction model with "Yes–No" format can be built to split data into different classes according to its attributes. Such a model is very easy to apply to new cases, although the algorithms often produce an excessive number of rules. Neural networks often fit nonlinear relationships very well, but are difficult to apply to new data. Logistic regression models are easy to apply to new data, although the problem of a cutoff between classes can be an issue [32].

Relative performance of data mining algorithms has long been understood to depend upon the specific data. Since data mining software is widespread, common practice in classification is to try the three basic algorithms (decision trees, neural networks, logistic regression), and use the one that works best for the given data set. Studies have compared these algorithms with RFM. Levin and Zahavi [20] compared RFM with decision trees (specifically CHAID), pointing out that decision trees are more automatic (RFM requires extensive data manipulation), but involve modeling issues such as controlling tree size and determining the best split for branches and leaves. Kim and Street [19] proposed a neural network model and applied feature selection mechanisms to reduce input variables, enabling focus upon the most important variables. Baesens et al. [1] also applied neural networks to customer response models (adding customer profile indicators to RFM), obtaining better prediction accuracy. That is a consistent finding — data mining algorithms will be expected to better predict customer response than RFM. However, RFM remains interesting because it relies upon the three fundamentally basic inputs that are readily available.

3. Methodology

3.1. Problem description and data set

This research design includes two studies (Study 1 and Study 2 hereafter) using two datasets obtained from the Direct Marketing Educational Foundation. Study 1 uses a dataset including 101,532 individual purchases from 1982 to 1992 in catalog sales. Study 2 is based on the data of 1,099,009 individual donors' contributions to a non-profit organization collected between 1991 and 2006. The purchase orders (or donations) included ordering (or donation) date and ordering amount. The last four months (Aug–Dec) of the data were used as the target period: Aug–Dec 1992 for Study 1 and