Expert Systems with Applications 40 (2013) 2007–2012

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Predicting customer profitability during acquisition: Finding the optimal combination of data source and data mining technique

Jeroen D'Haen [a], Dirk Van den Poel [a,*], Dirk Thorleuchter [b]

[a] Ghent University, Faculty of Economics and Business Administration, Department of Marketing, Tweekerkenstraat 2, 9000 Gent, Belgium
[b] Fraunhofer INT, D-53879 Euskirchen, Appelsgarten 2, Germany

ARTICLE INFO

Keywords:
Marketing analytics
Predictive analytics
Data source
B2B
Web mining
Web crawling
Bagging
Profitability
Customer acquisition
External commercial data

ABSTRACT

The customer acquisition process is generally a stressful undertaking for sales representatives. Luckily there are models that assist them in selecting the 'right' leads to pursue. Two factors play a role in this process: the probability of converting into a customer and the profitability once the lead is in fact a customer. This paper focuses on the latter. It makes two main contributions to the existing literature. Firstly, it investigates the predictive performance of two types of data: web data and commercially available data. The aim is to find out which of these two have the highest accuracy as input predictor for profitability and to research if they improve accuracy even more when combined. Secondly, the predictive performance of different data mining techniques is investigated. Results show that bagged decision trees are consistently higher in accuracy. Web data is better in predicting profitability than commercial data, but combining both is even better. The added value of commercial data is, although statistically significant, fairly limited.

1. Introduction

The acquisition of new customers is considered a multi-stage process in which only certain leads become real customers (Cooper & Budd, 2007; Patterson, 2007; Yu & Cai, 2007). This process is generally a stressful undertaking for sales representatives. Fortunately, these sales reps are assisted by models that assist them in selecting the 'right' leads to pursue. Two factors are important in selecting the 'right' lead: the probability the lead will convert into a customer and the profitability of that lead once he/she is a customer. This paper focuses on the latter. The goal is to design a model that is able to predict a dichotomous version of profitability (i.e., yes a customer is profitable or no a customer is not profitable). Profitability models exist, however, the main bottleneck they have is a lack of quality data. A new data source is introduced to solve this problem and it is compared in its performance to a more traditional data source. Furthermore, we investigate the impact of the data mining technique utilized on the estimated models of both data sources and examine which combination provides the highest accuracy.

This paper investigates the impact of three techniques: logistic regression, decision trees and bagged decision trees. While logistic regression is a more basic data mining technique that is often used in research, (bagged) decision trees are more advanced and less popular. The reason to consider different data mining techniques is twofold. First, according to Neslin, Gupta, Kamakura, Lu, and Mason (2006), which data mining technique is used has an impact on the predictive performance of the created models. So, employing different techniques is a way to increase the overall predictive performance by finding the optimal technique. Second, the data mining techniques are used as a proxy of data complexity and noisiness. Basic techniques are only capable of estimating simple, linear relations, while more advanced techniques are able to fit more complex, noisy data. If (bagged) decision trees are not able to perform better than logistic regression for a specific data source, we can conclude that this data source is most likely linear and noise-free in nature.

A quality model to predict profitability can only be constructed if quality data is available. Most models rely on commercial data purchased from specialized vendors (Rygielski, Wang, & Yen, 2002; Wilson, 2006). A relatively new and underinvestigated source of input for customer profitability models is textual information extracted from websites. Web mining and text mining can be used to gather this information from existing and potential customers' websites (Thorleuchter, Van den Poel, & Prinzie, 2012). However, textual information is seldom used as input for analyses in companies (Coussement & Van den Poel, 2009). The reason for this is that web data contains unstructured data that is hard to analyze. Nevertheless, latent indexing techniques can be used to make the data more structured and available as input for acquisition models (Thorleuchter et al., 2012).

* Corresponding author. Tel.: +32 9 264 89 80; fax: +32 9 264 42 79.
E-mail address: dirk.vandenpoel@ugent.be (D. Van den Poel).
URL: http://www.crm.UGent.be (D. Van den Poel).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.10.023

This paper makes two main contributions to the existing literature. Firstly, it investigates the predictive performance of two sources of data: web data and commercially available data. The aim is to find out which of these two has the highest accuracy as input predictor for profitability and to research if they improve accuracy even more when combined. Secondly, the predictive performance of different data mining techniques is investigated. So the overall research question can be formulated as follows: which technique is most accurate in combination with which data source? These two main contributions also show in what way this paper is different from the one presented by Thorleuchter et al. (2012). It investigates and compares different data sources and data mining techniques instead of simply focusing on only web data using a logistic regression. In this way there is a clear benchmark (i.e., commercial data) to which web data can be compared. As a result, this paper can be seen as the first real test of using textual data extracted using web mining as input for profitability models. Furthermore, the results obtained in this paper are discussed in more detail.

The remainder of the paper is structured as follows. First, the web vs. the commercially available data are discussed. Next we go deeper into the different data mining techniques. Third, a short description of the used data is given. Then, the results are presented. Finally, we end with a conclusion and discussion and we discuss the limitations of this paper and give suggestions for further research.

2. Web data vs. commercially available data

Today, most companies construct huge databases containing a wealth of information on their customers and their buying behaviors (Shaw, Subramaniam, Tan, & Welge, 2001). In order to extract the knowledge hidden in these databases, data mining can be applied to them (Mitra, Pal, & Mitra, 2002). Nevertheless, this source of data is not applicable to identify new profitable customers (Arndt & Gersten, 2001). The databases constructed by companies represent company-internal information, which means that they only contain information on their own customers.

Most companies purchase lists of information on potential (i.e., new) customers from specialized external vendors (Wilson, 2006). These lists tend to be of poor quality. Superior quality lists exist, though at a much higher expense (Buttle, 2009; Shankaranarayanan & Cai, 2005). Inferior data will render inferior results: this is the so-called garbage in, garbage out rule (Baesens, Mues, Martens, & Vanthienen, 2009). The main quality problem of purchased data is the high amount of missing values.

An alternative to the commercially available data is the use of web mining to extract customer information data (Shaw et al., 2001). The challenge of web data is twofold (Stumme, Hotho, & Berendt, 2006). On the one hand, the data is so unstructured that only humans are capable of understanding it. On the other hand, the amount of data is too huge for humans to handle and it can therefore only be processed by computers. This challenge can be solved by combining web- with text- and data-mining. Web mining can extract different types of data: content, structure, usage and user profile data (Srivastava, Cooley, Deshpande, & Tan, 2000). Content data is utilized in this paper as input to the proposed models. This type of data refers to the textual content that is seen when visiting a site. The textual information of customers' websites is consequently converted into term vectors in a term-space model (Thorleuchter et al., 2012). Latent semantic indexing is used to group related terms. Subsequently, singular value decomposition is applied to generate semantic generalizations. These generalizations are linked to the appearance of terms in similar web pages. Each generalization is a concept that refers to the hidden (latent semantic) patterns in the textual information. Companies get a score on each concept and these scores reflect how well a website loads on a specific concept (see Thorleuchter et al. (2012) for a more in-depth overview of this approach).

3. Data mining techniques

Data mining techniques are a way of extracting hidden knowledge in large databases (Ngai, Xiu, & Chau, 2009). Their importance is increasing in CRM analyses as the size of databases keeps growing (Ngai et al., 2009; Rygielski et al., 2002). Moreover, data mining is being used in the decision making process of companies (Baesens et al., 2009). The next part elaborates on the data mining techniques employed in this paper.

3.1. Logistic regression

Logistic regression is a regression analysis for categorical dependent variables and is based on the logit transformation of a proportion (Everitt & Skrondal, 2010; Field, 2009; Miguéis, Van den Poel, Camanho, & Falcao e Cunha, 2012). It is a standard parametric technique (Bellotti & Crook, 2008). The formula of a logistic regression is:

$$F(z) = \frac{1}{1+e^{-z}} \quad \text{where} \quad z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \qquad (1)$$

Blattberg, Kim, Kim, and Neslin (2008a), Hansotia and Wang (1997), Pampel (2000), Thomas (2010), Van den Poel and Buckinx (2005) As logistic regression is an often used and well-known data mining technique we will not expatiate on this subject.

3.2. Decision trees

A decision tree divides a dataset in subsets, using the values of the independent variables as selection criteria, in order to predict the dependent variable (Blattberg, Kim, Kim, & Neslin, 2008b). The top of a decision tree is called the root node (Berk, 2008b). This root node contains the full dataset. The outcome of a decision at each node is called a split (Duda, Hart, & Stork, 2001). Splits after the root node are termed branches and the final splits are the terminal nodes. All splits after the initial split imply interaction effects, unless they use the same predictor (Berk, 2008b). After the full tree is built, it needs to be pruned. Pruning is used to find the right size of the tree to avoid overfitting (Blattberg et al., 2008b; Duda et al., 2001). The bigger a tree is, the less cases there are in the terminal nodes and the more chance there is of having an overfitted tree. Pruning a tree starts at the terminal nodes and works its way up to the top (Berk, 2008b). It eliminates nodes that do not reduce heterogeneity enough compared to the complexity they add to the tree. This is a version of Occam's razor that prescribes that researchers should prefer the simplest model that explains the data (Baesens et al., 2009; Duda et al., 2001). Decision trees have several specific advantages (Tirenni, Kaiser, & Herrmann, 2007). They are a non-parametric method, invariant to monotonic predictor transformations (i.e., no variable transformations are required). Parametric methods yield poor results when the dimensionality of data is high (as is in our case) (Petersen, Molinaro, Sinisi, & van der Laan, 2007). Furthermore, decision trees are robust to the effects of outliers. Fig. 1 shows a graphical representation of a simple

3.3. Bagging

A problem with a decision tree is that it has been shown to be unstable (Breiman, 1996b). This means that small changes in the