Expert Systems with Applications 37 (2010) 2132–2143

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Improved marketing decision making in a customer churn prediction context using generalized additive models

Kristof Coussement [a,b,*], Dries F. Benoit [b], Dirk Van den Poel [b]

[a] University College HUBrussel, Faculty of Economics and Management, Stormstraat 2, B-1000 Brussels, Belgium
[b] Ghent University, Faculty of Economics and Business Administration, Department of Marketing, Tweekerkenstraat 2, B-9000 Ghent, Belgium

A R T I C L E I N F O

Keywords:
Customer Relationship Management (CRM)
Churn modeling
Marketing decision making
Generalized Additive Models (GAM)

A B S T R A C T

Nowadays, companies are investing in a well-considered CRM strategy. One of the cornerstones in CRM is customer churn prediction, where one tries to predict whether or not a customer will leave the company. This study focuses on how to better support marketing decision makers in identifying risky customers by using Generalized Additive Models (GAM). Compared to Logistic Regression, GAM relaxes the linearity constraint which allows for complex non-linear fits to the data. The contributions to the literature are three-fold: (i) it is shown that GAM is able to improve marketing decision making by better identifying risky customers; (ii) it is shown that GAM increases the interpretability of the churn model by visualizing the non-linear relationships with customer churn identifying a quasi-exponential, a U, an inverted U or a complex trend and (iii) marketing managers are able to significantly increase business value by applying GAM in this churn prediction context.

1. Introduction

Today, the business environment is characterized by fierce competition and saturated markets. In this context, companies increasingly derive revenue from the creation and enhancement of long-term relationships with their customers. This move towards a customer-centric approach to marketing, coupled with the increasing availability of customer-transactional data has made Customer Relationship Management (CRM) the leading strategy for marketing decision makers and this is reflected in firms' significant investments in CRM (Reinartz & Kumar, 2002; Teo, Devadoss, & Pan, 2006). Companies realize that their existing customer database is their most valuable asset (Athanassopoulos, 2000; Jones, Mothersbaugh, & Beatty, 2000; Thomas, 2001). It has been shown that it is more profitable to keep and satisfy existing customers than to constantly attract new customers who are characterized by a high attrition rate (Reinartz & Kumar, 2003). It is even suggested that it costs 12 times more to gain a new customer than to retain an existing one (Torkzadeh, Chang, & Hansen, 2006). Moreover, retained customers produce higher revenues and margin than new customers (Reichheld & Sasser, 1990).

It is clear that customer retention rates are important metrics in CRM (Hoekstra, Leeflang, & Wittink, 1999), resulting in an increasing number of customer churn related papers (e.g. Burez & Van den Poel, 2009; Coussement & Van den Poel, 2009). Predicting churn enables the elaboration of targeted retention strategies to limit the losses and to improve marketing decisions (Shaffer & Zhang, 2002). For example, specific incentives may be offered to the most risky customer segments, i.e. the most inclined to leave the company, with the hope that they remain loyal (Burez & Van den Poel, 2007). Two possible outcomes are observed: the customer churned or he/she stayed with the company. So, technically spoken, customer churn modeling is a binary classification problem. Moreover, for the prediction of customer churn using the information stored in the data warehouse one needs tools that can handle large amounts of data, i.e. data-mining tools (Shaw, Subramaniam, Tan, & Welge, 2001).

A broad range of data mining techniques have been used in the past, but one of the most popular models in the churn context remains the (binary) Logit model (Lemmens & Croux, 2006). This model has been used extensively in marketing decision making to model churn problems (e.g Buckinx & Van den Poel, 2005; Hwang & Euiiho Suh, 2004; Kim & Yoon, 2004). The popularity of Logistic Regression is not surprising since the model has some outspoken advantages. First of all, the method is able to combine relative simplicity and good performance. The parameter estimates of the Logit model are interpretable in terms of odds ratios which facilitates full understanding of the results. Moreover, the technique is relatively robust (Buckinx & Van den Poel, 2005), while

* Corresponding author. Address: University College HUBrussel, Faculty of Economics and Management, Stormstraat 2, B-1000 Brussels, Belgium. Tel.: +32 (0)2 608 14 19.
E-mail addresses: Kristof.Coussement@HUBrussel.be (K. Coussement), Dries.Benoit@UGent.be (D.F. Benoit), Dirk.VandenPoel@UGent.be (D. Van den Poel).

0957-4174/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2009.07.029

its popularity is reflected in the availability in almost all statistical software packages. However, as often, one of the advantages can also be considered as a disadvantage. The simplicity of the model is achieved by assuming that the functional form of the dependence on the explanatory variables is known. However, the functional form is seldom known in practice. If the functional form is misspecified then the estimates of the coefficients and the inferences based on them are misleading (Horowitz & Savin, 2001). In addition, with Logistic Regression, a linear relationship is modeled on the data, which is an oversimplification of the real relationship in the data (Allison, 1999). It is possible to relax these restrictive assumptions by applying the Generalized Additive Models (GAM) approach.

The GAM approach has at least two distinct advantages compared to Logistic Regression. First, by relaxing the usual assumptions it becomes possible to uncover (non-linear) structures in the data that otherwise might be missed. These structures often give substantial new insights into the effects of the covariates. Furthermore, since the GAM approach allows for more complex relationships between the dependent variable and the covariates, more accurate predictions are likely (Hastie & Tibshirani, 1986). The GAM approach has shown its relevance in a broad range of different research domains ranging from biology (e.g. Jowett, Parkyn, & Richardson, 2008) over cancer research (e.g. Lee et al., 2007) and supply-chain management (e.g. Cakir, 2009). However, only very few marketing-related studies have used a related non- or semi-parametric methodology (e.g. Kalyanam, 1998; Kumar, Scheer, & Steenkamp, 1998; Shively, 2000). Moreover, to the best of our knowledge this is the first study in a customer churn context that evaluates and uses GAM models.

This study contributes to the existing literature in several ways. Three distinct perspectives on GAM models in the current churn context are elaborated. First, a comparison in terms of predictive performance is made between the popular Logistic Regression model and the GAM approach. Second, it is shown that relaxing the linearity assumption results in a much richer insight into the effects of the covariates which helps decision makers in a thorough understanding of the churn problem of their company. Finally, it is shown how GAM models increase business value of the company. The added value of the GAM approach is therefore shown in monetary terms based on a real-life business setting.

The remainder of this paper is organized as follows. Section 2 focuses on the methodology behind both the logistic model and the GAM approach. The evaluation criteria used in the empirical part of this study are explained in the same section. Section 3 applies the proposed methodology to a real-world churn setting, i.e. the newspaper case. Section 4 treats the managerial implication of the proposed methodology and finally, Section 5 gives conclusions and directions for further research.

2. Methodology

2.1. Base learner

The methodological link between the Logistic Regression and the GAM approach is explained starting from the Generalized Linear Model (GLM) framework. The most common representation of the GLM is as follows (Tabachnick & Fidell, 1996):

$$E(Y) = \mu = g(X'\beta) \tag{1}$$

with $E(Y)$ representing the expected value of Y, $X'\beta$ is a linear combination of the data with unknown parameters β and g is the link function. As Eq. (1) shows, each outcome of the dependent variable Y is assumed to be generated from a particular distribution function. Different choices of distribution function and link function result in different statistical techniques. For example, the normal distribution in combination with the identity link function results in the normal multiple regression model. When the responses Y are binary, the distribution function is the binomial distribution and the interpretation of μ_i is then the probability, p_i, of Y_i taking on the value one. A Logistic Regression model is then obtained by applying the logit link function by

$$\text{logit}\{P(X)\} \equiv \log\left\{\frac{P(X)}{1-P(X)}\right\} = \alpha + \sum_{j=1}^{p}\beta_j' X_j \tag{2}$$

with $P(X) = Pr(Y = 1|X)$. The parameter vector β is estimated from observations (y, x) and this can be done using maximum likelihood among others. More technical details on Logistic Regression are found in Anderson (1983).

2.2. Generalized additive models

An attractive alternative to Logistic Regression is Generalized Additive Models (GAM) (Hastie & Tibshirani, 1986; Hastie & Tibshirani, 1987; Hastie & Tibshirani, 1990). Additive Logistic Regression relaxes the linearity constraint and suggests a nonparametric fit to the data. In other words, the regression function is modeled in a nonparametric way and the data itself decides on the functional form. However, Hastie and Tibshirani (1990) argue that nonparametric methods perform worse when the number of explicative variables increases, because the sparseness of the data inflates the variance of the estimates. This is often referred to as the curse of dimensionality (Bellman, 1961). It was Stone (1985) who proposed additive models to approximate the multivariate regression function. Consequently, the curse of dimensionality is avoided since each individual additive term is estimated using a univariate smoother. Hastie and Tibshirani (1990) pick up the additive model principle by extending the GLM framework.

This study gives a general overview on the GAM principle. For more details, we kindly refer to Hastie and Tibshirani (1986), Hastie and Tibshirani (1987, Hastie and Tibshirani (1990) or Hastie, Tibshirani, and Friedman (2001). Suppose that Y is the response variable with binary target labels and $X_1, X_2, \ldots, X_p$ is a set of independent variables, GAM generalize the Logistic Regression principle by replacing the linear predictor $\sum_{j=1}^{p}\beta_j X_j$ in Eq. (2) with an other additive component where

$$\text{logit}P(X) \equiv \log\left\{\frac{P(X)}{1-P(X)}\right\} = \alpha + \sum_{j=1}^{p} s_j(X_j) \tag{3}$$

with $s_1(.), s_2(.), \ldots, s_p(.)$ as smooth functions. This study uses smoothing splines for $s_1(.), s_2(.), \ldots, s_p(.)$ to estimate the nonparametric trend for the dependence of the logit on $X_1, X_2, \ldots, X_p$. A smoothing spline solves the following optimization problem: among all functions $\eta(x)$ with continuous second order derivatives, find the function that minimizes the penalized residual sum of squares

$$\sum_{i=1}^{n}(y_i - \eta(x_i))^2 + \lambda\int_a^b(\eta''(t))^2\,dt \tag{4}$$

where λ is a fixed constant and $a \leqslant x_1 \leqslant x_2 \leqslant \cdots \leqslant x_n \leqslant b$. The goodness-of-fit is measured by the first part of Eq. (4), while the second term is a penalty term that penalizes curvature in the function by the smoothing parameter λ. The complexity of $\eta(x)$ is measured by λ and it is inversely related to the degrees of freedom (df). If λ is small (i.e. the df are large), $\eta(x)$ is any function that approaches an interpolation to the data, when λ is large (i.e. the df are small), $\eta(x)$ is closely related to a simple least squares fit. It is shown that an explicit and unique minimizer for Eq. (4) exists, i.e. a natural