Expert Systems with Applications 36 (2009) 10475–10484

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Benefits of quantile regression for the analysis of customer lifetime value in a contractual setting: An application in financial services

Dries F. Benoit, Dirk Van den Poel *

Ghent University, Faculty of Economics and Business Administration, Department of Marketing, Tweekerkenstraat 2, B-9000 Ghent, Belgium

ARTICLE INFO

Keywords:
Customer relationship management (CRM)
Database marketing
Customer segmentation
Quantile regression
Prediction interval
Customer lifetime value

ABSTRACT

The move towards a customer-centred approach to marketing, coupled with the increasing availability of customer transaction data, has led to an interest in understanding and estimating customer lifetime value (CLV). Several authors point out that, when evaluating customer profitability, profitable customers are rare compared to the unprofitable ones. In spite of this, most authors fail to recognize the implications of these skewed distributions on the performance of models they use. In this study, we propose analyzing CLV by means of quantile regression. In a financial services application, we show that this technique provides management more in-depth insights into the effects of the covariates that are missed with linear regression. Moreover, we show that in the common situation where interest is in a top-customer segment, quantile regression outperforms linear regression. The method also has the ability of constructing prediction intervals. Combining the CLV point estimate with the prediction intervals leads to a new segmentation scheme that is the first to account for uncertainty in the predictions. This segmentation is ideally suited for managing the portfolio of customers.

1. Introduction

Over the past decade, customer relationship management (CRM) has become a leading strategy in highly competitive business environments. Companies increasingly derive revenue from the creation and enhancement of long-term relationships with their customers (Coussement & Van den Poel, 2008). This move towards a customer-centric approach to marketing, coupled with the increasing availability of customer-transaction data, has led to an interest in estimating and understanding customer lifetime value (CLV). CLV is viewed as the present value of the future cash flows associated with a customer (Pfeifer, Haskins, & Conroy, 2005). Knowing the CLV of individual customers enables the decision maker to improve the customer segmentation and marketing resource allocation efforts (Kumar, Lemon, & Parasuraman, 2006; Kim & Lee, 2007) and this in turn will lead to higher retention rates and profits for the firm (Hawkes, 2000).

Donkers, Verhoef, and de Jong (2007) give a detailed overview and comparison of the wide range of different approaches that have been used for CLV modeling. From their outline it is clear that regression-type models are often used in this context (e.g. linear regression model (Glady, Baesens, & Croux, 2009; Malthouse & Blattberg, 2005; Hansotia & Rukstales, 2002; Verhoef & Donkers, 2001); Probit model (Bolton, Kannan, & Bramlett, 2000); multivariate Probit model (Donkers et al., 2007); multivariate Logit model (Prinzie & Van den Poel, 2007). Several authors (Kim, Jung, Suh, & Hwang, 2006; Duboff, 1992; Gloy, Akridge, & Preckel, 1997) point out that, when evaluating customer profitability, profitable customers are rare compared to the unprofitable ones. Gupta et al. (2006) remark that the regression-type models easily break down when applied to settings where the behavior of interest is rare. In spite of this, most authors fail to recognize the implications of this skewed distribution for the models they use. Note that in this study, the focus is on the situation where the buyer-seller relationship is governed by a contract. The Pareto/NBD model, which is state-of-the-art in a non-contractual setting (Glady et al., 2009; Wübben & Wangenheim, forthcoming), is inappropriate in this contractual context (Schmittlein, Morrison, & Colombo, 1987) and is therefore excluded from the current analysis.

In this paper, we propose to analyze CLV by means of quantile regression. Quantile regression (Koenker & Basset, 1978; Koenker, 2005) is a method for fitting a regression line through the conditional quantiles of a distribution. Therefore, the method is less influenced by long-tailed, skewed distributions that are typical in CLV modeling. Moreover, the manager's interest is often not in the large group of unprofitable customers, but in the smaller group of more lucrative customers. In this case, mean regression gives only very limited information and it is worthwhile considering the more extreme quantiles of the CLV distribution. Consequently, explicit investigation of the effects of the covariates via quantile regression can provide a more nuanced view of the stochastic

* 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).

0957-4174/$ - see front matter © 2009 Published by Elsevier Ltd.
doi:10.1016/j.eswa.2009.01.031

relationship between the covariates and the dependent variable (Koenker & Hallock, 2001). Therefore quantile regression results in a more informative empirical analysis. Furthermore, quantile regression produces prediction intervals that give insight into the uncertainty about a CLV prediction. This information is important for the decision maker when quantifying the risk associated with any given customer (Gupta et al., 2006).

This study contributes to the existing customer lifetime value literature by investigating the usefulness of the quantile regression approach in the financial services industry. First, we demonstrate that by using quantile regression, we provide management with a more detailed insight into the complex relationship of the CLV drivers than mean regression does. Secondly, we show that quantile regression outperforms the often used mean regression as predictive tool in the context of CLV modeling. Thirdly, we provide our CLV forecasts with prediction intervals or even with the entire predictive distribution. This makes the task of customer risk assessment more straightforward for the manager. A new segmentation scheme is proposed, based on CLV forecast and the related prediction interval. This method of customer segmentation distinguishes itself from other CLV schemes by explicitly taking uncertainty into account.

The structure of this paper is as follows. In Section 2, we review the previous studies related to customer value. It illustrates the limitations of existing studies and sets the stage for this paper. In Section 3, we describe our methodology for the analysis and prediction of customer lifetime value. In Section 4, we present an observational application of this methodology in the financial services industry. We also discuss the market segmentation and managerial implications for this application. Finally, Section 5 concludes with remarks on the limitations of this study and future research directions.

2. Related works

Customer lifetime value has been studied under the name of LTV, Customer Value, Customer Equity and Customer Profitability. The concept is defined as the sum of the revenues gained from company's customers over the lifetime of transactions after deduction of the total cost of attracting, selling and servicing customers, taking into account the time value of money (Hwang, Jung, & Suh, 2004). The basic formula for calculating CLV for customer i at time t for a finite time horizon T (Berger & Nasr, 1998) is:

$$\mathrm{CLV}_{i,t} = \sum_{\tau=0}^{T} \frac{\mathrm{Profit}_{i,t+\tau}}{(1+d)^{\tau}}, \tag{1}$$

where d is a pre-determined discount rate. In multi-service industries, $\mathrm{Profit}_{i,t}$ is defined as:

$$\mathrm{Profit}_{i,t} = \sum_{j=1}^{J} \mathrm{Serv}_{ij,t} * \mathrm{Usage}_{ij,t} * \mathrm{Margin}_{ij,t} \tag{2}$$

here J is the number of different services sold, $\mathrm{Serv}_{ij,t}$ is a dummy indicating whether customer i purchases service j at time t, $\mathrm{Usage}_{ij,t}$ is the amount of that service purchased and $\mathrm{Margin}_{ij,t}$ is the average profit margin for service j.

Theoretically, CLV models should estimate the value of a customer over the entire customer's lifetime. However, in practice most researchers use a finite time horizon of three or four years (e.g. Donkers et al., 2007; Rust, Zeithaml, & Lemon, 2000). Three to four years is a good estimate for the horizon over which the current business environment would not substantially change and even then, there is significant uncertainty in predicting customer behavior (Venkatesan, Kumar, & Bohling, 2007). Moreover, some research considers an even shorter time horizon (Hwang et al., 2004).

CLV has been analyzed in a substantial number of different domains, varying from econometric models to computer science techniques. However, the key questions are usually very similar: *"What are the drivers of CLV?"*, *"Which customers are the future most valuable ones?"*, *"How to address the top customers?"*, etc. Several authors give an overview of the variety of modeling procedures that were used in search for answers to the key questions (Ngai, Xiu, & Chau, 2009; Gupta et al., 2006; Donkers et al., 2007; Berger & Nasr, 1998; Venkatesan & Kumar, 2004). In general, one can distinguish two broad classes of models in the current contractual setting. First, a large group of models focuses on the choices customers face when buying an additional service or product. A customer's lifetime value is then seen as the sum of the distinct contributions per service or product. This approach is appealing because of the natural way in which the CLV prediction is built up. In a first stage, an estimation is made on the probability of a customer buying a given product or service. The second stage is then to combine these probabilities with the margins associated with the product or service into an aggregate prediction of a customer's lifetime value. This approach also has the advantage of providing more insight into the factors that drive customer value. The main drawbacks are the amount of modeling required and the often poorer predictions. Examples of this approach are found in Venkatesan and Kumar (2004) and Hwang et al. (2004). The second large group of models does not follow the two stage method, but focuses directly on relationship length and total profits. Since the individual-level choice modeling is left aside, the process of producing CLV estimates is much more straightforward and prediction accuracy is higher (Verhoef & Donkers, 2001). As such, this approach turns the disadvantages of the first approach into benefits. However, due to aggregation, insight into the factors that drive consumer profitability is limited compared to the choice-based approach. Examples of CLV research following this direct approach are found in Malthouse and Blattberg (2005) and in Hansotia and Rukstales (2002).

Given that one of the key issues when decision makers use the CLV metric is whether the firm can provide an adequate prediction of the CLV of each customer in the database (Malthouse & Blattberg, 2005; Venkatesan & Kumar, 2004), it is clear that the predictive accuracy of the CLV is of primordial importance. Furthermore, these predictions are often used as guidelines for investments in segments of customers (Zeithaml, Rust, & Lemon, 2001). However, the previously used regression techniques are often not ideally suited for the purpose of modeling customer lifetime value. When evaluating customer profitability, marketers are often reminded of the 80/20 rule (80% of the profits are produced by top 20% of profitable customers and 80% of the costs are produced by top 20% of unprofitable customers) (Duboff, 1992; Gloy et al., 1997). This finding has important implications for both the two-stage approach as well as for the approach that models CLV directly. For researchers using the two-step CLV approach, the problem arises when modeling the choice problem. Since the largest group of customers buys no or only a very limited amount of products or services and only a small group of customers buys many products or services, the researcher should be aware of the fact that he or she is modeling rare events. In this rare-event situation, it is known that parametric choice models easily break down (Gupta et al., 2006). The other approach, where the researcher focuses directly on the relationship length and total profits, leaves aside the individual-level choice modeling step. However, the problem of rare events can not be totally avoided. This is because the underlying process (the 80/20 rule) results in a lifetime value variable that tends to have a strong non-normal distribution and the usual assumption of homoscedasticity is hard to maintain (Fader, Hardie, & Lee, 2005; Malthouse & Blattberg, 2005). In contrast, the proposed quantile regression technique does not suffer from these