ELSEVIER

Expert Systems with Applications 29 (2005) 630–640

Expert Systems with Applications

www.elsevier.com/locate/eswa

Constrained optimization of data-mining problems to improve model performance: A direct-marketing application

Anita Prinzie[1,*], Dirk Van den Poel

Department of Marketing, Faculty of Economics and Business Administration, Ghent University, Hoveniersberg 24, 9000 Ghent, Belgium

Abstract

Although most data-mining (DM) models are complex and general in nature, the implementation of such models in specific environments is often subject to practical constraints (e.g. budget constraints) or thresholds (e.g. only mail to customers with an expected profit higher than the investment cost). Typically, the DM model is calibrated neglecting those constraints/thresholds. If the implementation constraints/thresholds are known in advance, this indirect approach delivers a sub-optimal model performance. Adopting a direct approach, i.e. estimating a DM model in knowledge of the constraints/thresholds, improves model performance as the model is optimized for the given implementation environment. We illustrate the relevance of this constrained optimization of DM models on a direct-marketing case, i.e. in the field of customer relationship management. We optimize an individual-level response model for specific mailing depths (i.e. the percentage of customers of the house list that actually receives a mail given the mailing budget constraint) and compare its predictive performance with that of a traditional response model, neglecting the mailing depth during estimation. The results are in favor of the constrained-optimization approach.

Keywords: Constrained optimization; Data mining; Direct marketing; Customer relationship management; Targeting

1. Introduction

Firms typically make marketing decisions in a constrained business as well as operating environment. For instance, a financial-services provider planning a cross-sell action knows in advance that it is not optimal to propose more than 1, 2,…, or k services to the customers being targeted. Similarly, an online retailer might want to recommend a limited set of k products on the visitor's customized welcome page. Today, most companies in the pharmaceutical industry are facing the problem of drugs for which the patent will expire in the near future. In an attempt to compete with the generic substitutes for these medicines, the pharmaceutical company might opt to send its sales team to k physicians currently prescribing their patented drug. Given this high cost/impact marketing strategy, it may not be profitable to visit *all* doctors currently prescribing the drug. For instance, the sales team could visit only doctors whose net expected LTV is higher than a certain *threshold*. A hypermarket wants to create store-traffic by mailing a reduction coupon for a certain product. Given the stock limits, the hypermarket will have to make a selection of k customers to mail the coupon in order to meet demand (Buckinx, Moons, Van den Poel, & Wets, 2004). Similarly, management of a supermarket loyalty program may want to limit defection of their k% top customers (Buckinx & Van den Poel, 2005). Finally, a car manufacturer wants to promote a new sports car by inviting prospects to join a professional racer in making laps on a circuit on a specified day. Given the time constraint (i.e. only one day) and the availability of a single free seat, the manufacturer can only invite a limited number of k prospects. These examples clearly indicate that a lot of marketing problems heed practical constraints resulting in implementation limitations. Mostly, these constraints are facts known beforehand. Nevertheless, in practice, data-mining models used to solve these problems are commonly applied ignoring the constraints/thresholds. When it comes to implementation, the solution is restricted to fit within the practical boundaries. In this paper, we advocate a direct approach estimating a data-mining (DM) model in awareness of

* Corresponding author. Tel.: +32 9 264 89 80; fax: +32 9 264 42 79.
E-mail addresses: anita.prinzie@ugent.be (A. Prinzie), dirk.vandenpoel@ugent.be (D. Van den Poel).
[1] Tel.: +32 9 264 35 20; fax: +32 9 264 42 79.

0957-4174/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2005.04.022

the constraints/thresholds. The constrained-optimization perspective on data mining is (usually) superior to the indirect approach (i.e. estimate general DM model and restrict solution afterwards to obey the implementation limitations) as it results in higher model performance. The latter is due to directly tailoring (i.e. optimizing) the DM model to fit the implementation environment at hand. We illustrate this constrained optimization of DM models on a target-selection problem faced by a direct mailer heeding budget constraint.

The remaining of this paper is structured as follows. In Section 2, we outline how the constrained optimization applies to direct-marketing problems. We note that the focus of the illustration of constrained optimization of DM models will be on target-selection DM problems given budget constraints. In Section 3, we introduce a constrained optimization of a binary response DM model. Employing weighted maximum likelihood to estimate a binary logistic regression model, we directly optimize the response model for a specific mailing depth stemming from a budget constraint. Section 4 describes the application details and specifies how the new constrained optimization of the response DM model is applied on our real-life direct-marketing case. We compare the model accuracy of the indirect response model to that of our new innovative response model optimized for a given mailing depth (i.e. implementation constraint). We conclude in Section 5 by discussing the main findings and suggesting some avenues for further research.

2. Constrained optimization of DM models for direct marketing

The importance of direct marketing is both reflected in the economic value of this industry (e.g. advertising budgets for direct marketing represent more than a quarter of total advertising expenditures; http://www.the-dma.org) as in the amount of research devoted to this subject (see Table 1). Within direct marketing, response modeling is very important as it involves several decisions crucial to the development of an optimal direct-marketing policy: (1) who to target, (2) with what offer, (3) using what action (i.e. how), (4) when (Nash, 1992) and (5) how often? The first issue is the most important (Roberts & Berger, 1989) as the success of a direct-marketing campaign is highly dependent on who is being targeted. The importance of target selection is also apparent from the fact that almost all previous studies are solely tackling this problem. Only recently the timing and/or the frequency of the direct-marketing policy is simultaneously modeled with the target-selection problem (Bitran & Mondschein, 1996; Elsner, Krafft, & Huchzermeier, 2003; Gönül & Shi, 1998; Jonker, Piersma, & Van den Poel, 2004; Piersma & Jonker, 2004). Besides the decision variable(s) investigated, the type of the objective function modeled, can further categorize response modeling in direct marketing. Basically, response models link some measure of customer response (i.e. dependent) to independent variables by an objective function. The independent variables usually include target characteristics. Among the most frequently used targets variables are recency, frequency and monetary value (RFM; Cullinan, 1977). Besides these RFM variables other behavioral variables might be incorporated. One could also include socio-demographical or psychographical target features in the objective function. However, several authors in the direct marketing literature remark that behavioral variables are far better predictors than non-behavioral variables (Heilman, Kaefer, & Ramenofsky, 2003; Rossi, McCulloch, & Allenby, 1996). In addition to target characteristics also mailing features can be treated as predictor variables (Berger & Magliozzi, 1992; Bult, van der Scheer, & Wansbeek, 1997; Morwitz & Schmittlein, 1998; Piersma & Jonker, 2004; Smith & Berger, 1996; Van den Poel & Leunis, 1998). The objective function intends to link these types of independent variables to a dependent variable. Many researches define the dependent as binary response thereby trying to maximize the number of responders (cf. purchase-incidence modeling). In other studies the objective function is to predict the slope of expenditures (Baesens et al., 2004) or to maximize the generated profit. The latter is often operationalized by the lifetime value (LTV) of the customers or by a product of the number of responses and the amount purchased for. Finally, depending on the time interval in which the dependent is measured, we can distinguish static from dynamic response models. Static models measure the response (binary or continuous) for a single direct-marketing action. These models are referred to as single-period models. In contrast, dynamic models aim at maximizing the response/profit of the direct-marketing company over a(n) (in)finite time horizon (e.g. several direct-marketing campaigns). These multi-period models might target a customer not profitable for a single campaign but who is expected to be in the near future (Bitran & Mondschein, 1996; Elsner et al., 2003; Gönül & Shi, 1998; Gönül & Srinivasan, 1996; Jonker et al., 2004).

No matter how the response model is defined, the implementation of such a model is mostly subject to environmental constraints like the available budget, inventory limits, size of the house list, production capacity, operational capacity, etc. to name just a few. Almost none of the previous research studies modeling response include these constraints, neither directly nor indirectly. Although Gönül and Shi (1998) acknowledge that firms make their mailing decisions in a constrained operating environment, they refrain from modeling these constraints in order to keep the model structure manageable. Elsner et al. (2003) estimate the optimal mailing frequency and subsequently test the feasibility of this optimal solution given the operational constraints. Luckily, they find that the optimal mailing frequency of 25 mailings a year is the volume the medium-sized company can handle efficiently, given