Available online at www.sciencedirect.com
SCIENCE DIRECT®
Decision Support Systems 41 (2006) 542–553
Decision Support Systems
www.elsevier.com/locate/dsw

Toward a successful CRM: variable selection, sampling, and ensemble

YongSeog Kim*

Business Information Systems, Utah State University, UMC 3515 Old Main Hill, Logan, UT 84322, USA

Received 3 June 2003; received in revised form 1 September 2004; accepted 2 September 2004
Available online 2 November 2004

Abstract

This paper studies the effects of variable selection and class distribution on the performance of specific logit regression (i.e., a primitive classier system) and artificial neural network (ANN; a relatively more sophisticated classifier system) implementations in a customer relationship management (CRM) setting. Finally, ensemble models are constructed by combining the predictions of multiple classiers. This paper shows that ANN ensembles with variable selection show the most stable performance over various class distributions.

Keywords: CRM; Variable selection; Sampling; Ensemble; Neural network

1. Introduction

Advances in data warehouses and the vast amount of customers' demographic, psychographic, and behavioral information provide marketing managers a new marketing channel—database micromarketing. Traditional mass marketing channels (e.g., advertisements in TV and newspapers) have been very successful and are still important. However, customer relationship management (CRM) programs for macromarketing slowly give way to new CRM programs for micromarketing. This is because, in micromarketing programs, the firms can develop a marketing message directed toward a specific group of households that are most likely to open to the customized message.

Both marketing [7,13] and data mining researchers [3,10] have presented various database marketing approaches for successful CRM programs. The simplest example is the RFM (recency, frequency, monetary) approach that targets households by using knowledge of the customer's purchase history [21]. When targeting new households with no prior relationship, the analysis of the relationship between demographics and the response to a test mailing of a representative household sample can be utilized. In Piatetsky-Shapiro and Masand [18], the protability condition of a campaign was explicitly formulated as a function of the lift of the model, uniform campaign cost per mailing, and marginal revenue per identified

* Tel.: +1 435 797 2271; fax: +1 435 797 2351.
E-mail address: YonSeon.Kim@Bussiness.usu.edu.

0167-9236/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2004.09.008

positive record. Chou et al. [9] devised an effective model for identifying prospective insurance buyers when buyer versus nonbuyer information is not available. Gersten et al. [12] presented a model to select prospects in the automotive industry where the buying decision takes a long time. A good summary of related studies on CRM programs from marketing and data mining communities can be found in Kim and Street [15].

Traditionally, the optimal selection of customer targets has been considered one of the most important factors for a successful CRM program. Thus, many models have been proposed to identify as many customers as possible who will respond to a specific solicitation campaign letter, or who will end further relationships with the firm. In particular, with exceptionally high annual churn rates (20–40%), firms in mobile telecommunications industry try to develop predictive models that accurately identify which customers are most likely to churn. In addition to the predictive accuracy, the comprehensibility of a model becomes another important issue in developing CRM programs. A rule-based system that consists of too many *if–then* statements makes it difficult for marketing to understand key drivers of consumer behaviors. The poor comprehensibility can greatly reduce managers' trust in the system itself, and prevent decision makers from developing long-term CRM programs.

The ultimate goal of this study is to provide practitioners useful guidelines to building predictive models for effective CRM programs. In particular, this paper investigates the importance of variable selection and class distribution, and how they can affect the performance of predictive models. Specific implementations of two different learning algorithms are used: logit regression (i.e., a primitive model) and artificial neural networks (ANNs; a relatively more sophisticated classifier model). These algorithms have been widely used in developing predictive models for CRM applications. In particular, ANNs have been used in many other marketing applications such as customer clustering [1] and market segmentation [14]. In this study, ANNs and logit linear regression are used to estimate each individual's likelihood of ending current relationships after learning linear or possibly nonlinear relationships between given input variables and the churner indicator.

Variable selection is used to enhance the comprehensibility of models in this paper. Variable selection is the process of choosing a subset of the original predictive variables by eliminating variables that are either redundant or possess little predictive information. By identifying key determinants of churning behaviors of customers, variable selection can not only enhance the comprehensibility of predictive models but also save a great amount of computational time and cost. However, eliminating many input variables may have different effects on the predictive accuracy of models depending on their representational powers and structural complexities. Therefore, this study aims to analyze the relationship between variable selection and the predictive accuracy of predictive models over various class distributions.

Various sampling techniques are also used to shorten computational time and enhance the accuracy of a predictive model by removing noise records that have the same values for input variables except class indicator. Sampling in this paper is also used to vary class distributions to study the relationships between class distribution and the predictive accuracy of classifiers. When class distributions in the training and test data are significantly different, the calibrated model on the training data may not perform well on the test data. However, the same class distribution of the training and test data does not necessarily make the calibrated model perform best on the test set [23]. Therefore, it is necessary to vary the class distribution of the training data to the optimal class distribution for calibrating predictive models. In particular, this paper intends to identify optimal class distributions for different types of classifiers.

This paper also presents an ensemble approach that combines the predictions of multiple models. In order to build an ensemble, the estimated probabilities of being a churner from multiple models are combined with the equal weight. Furthermore, the effects of variable selection, structural complexity of models, and class distribution on the ensemble models are empirically estimated to provide decision makers and data analysts useful guidelines on how to develop an accurate model for CRM programs.

This paper is organized as follows. Section 2 briefly reviews variable subset selection, sampling, and ensemble decision making. Section 3 introduces the original data set and presents a reduced variable