International Journal of Research in Marketing 34 (2017) 154–172

Contents lists available at ScienceDirect

IJRM
International Journal of Research in Marketing

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

No future without the past? Predicting churn in the face of customer privacy☆

Niels Holtrop *, Jaap E. Wieringa, Maarten J. Gijsenberg, Peter C. Verhoef

Department of Marketing, Faculty of Economics and Business, University of Groningen, PO Box 800, 9700 AV Groningen, The Netherlands

ARTICLE INFO

Article history:
First received on January 21, 2015 and was under review for 8½ months
Available online 28 June 2016

Area Editor: Werner J. Reinartz

Keywords:
Churn prediction
Database marketing
Customer relationship management
Data privacy
Kalman filter
Mixture model

ABSTRACT

For customer-centric firms, churn prediction plays a central role in churn management programs. Methodological advances have emphasized the use of customer panel data to model the dynamic evolution of a customer base to improve churn predictions. However, pressure from policy makers and the public geared to reducing the storage of customer data has led to firms' 'self-policing' by limiting data storage, rendering panel data methods infeasible. We remedy these problems by developing a method that captures the dynamic evolution of a customer base without relying on the availability past data. Instead, using a recursively updated model our approach requires only knowledge of past model parameters. This generalized mixture of Kalman filters model maintains the accuracy of churn predictions compared to existing panel data methods when data from the past is available. In the absence of past data, applications in the insurance and telecommunications industry establish superior predictive performance compared to simpler benchmarks. These improvements arise because the proposed method captures the same dynamics and unobserved heterogeneity present in customer databases as advanced methods, while achieving privacy preserving data minimization and data anonymization. We therefore conclude that privacy preservation does not have to come at the cost of analytical operations.

1. Introduction

For firms that rely on customers as their principal asset, the defection of customers, or churn, is a chief concern. This concern has exacerbated itself in the past decade as customers have become more aware of switching opportunities and switching barriers have fallen as a result of increasing market transparency and government deregulation. Annual churn rates can be as high as 63% (Blattberg, Kim, and Neslin, 2008, p. 609), which illustrates the extent to which customer churn can affect a firm's customer base. Therefore, focusing on retention is more beneficial in terms of firm value than, for example, increasing profit margins or lowering acquisition costs (Gupta, Lehmann, & Stuart, 2004). Top executives recognize these benefits, reporting that customer retention is their top priority in terms of marketing spending (Forbes, 2011).

☆ The authors thank the two anonymous companies and the Customer Insights Center at the University of Groningen for providing data for this research. Furthermore, the authors thank the Editors (Eitan Muller, former Editor, and Roland Rust, current Editor), the Area Editor and two anonymous reviewers for their valuable feedback. Participants at the 2012 EMAC Doctoral Colloquium, the 2012 Marketing Dynamics Conference, the 2014 ISMS Marketing Science Conference and the 2016 EMAC Conference, and seminar participants at the University of Groningen and Lancaster University provided additional valuable comments on earlier versions of this paper.

* Corresponding author. Tel.: +31 503639621.
E-mail address: N.Holtrop@rug.nl (N. Holtrop).

http://dx.doi.org/10.1016/j.ijresmar.2016.06.001

Firms use churn management programs to stimulate customer retention. These programs center on identifying customers at risk of churning and targeting them with a marketing program geared to increasing behavioral loyalty using retention incentives such as special offers and discounts (Ascarza, Iyengar, & Schleicher, 2016; Lemmens & Gupta, 2013). To identify at-risk customers, firms form churn propensities using statistical models calibrated on historical churn data. After ranking these churn probabilities, the customers with the highest probabilities are selected for inclusion in the retention program.

Because predicting churn plays a vital role in the design of effective churn management programs, researchers are continually exploring more accurate ways of forming these propensities. The most popular methods used in practice are logistic regression and classification trees, which use cross-sectional data and have been shown to have a good short-term predictive performance (Neslin, Gupta, Kamakura, Junxiang, & Mason, 2006; Risselada, Verhoef, & Bijmolt, 2010). More recently, Ascarza and Hardie (2013) presented an approach that takes advantage of the richness of modern databases and model the dynamic evolution of customers in a customer base while accounting for unobserved customer heterogeneity using a Hidden Markov model (HMM) and panel data on customer churn behavior. They show that such an approach provides better short- and long-term predictions of churn than a range of benchmark models.

While the richness and size of modern customer databases offer opportunities for firms, as illustrated by the previous example and numerous (big) data driven firms, this development has also raised policy maker and public awareness that increasing amounts of customer data are stored and linked at the expense of customer privacy. Policy makers in both the United States (PCAST, 2014; Podesta, Pritzker, Moniz, Holdren, & Zients, 2014) and Europe (General Data Protection Regulation, European Parliament, 2013) have, or are planning to, put forward legislation to regulate the storage of individual customer data for prolonged periods of time. At the same time, public awareness of privacy has also increased, for example due to the high profile Google Spain v. AEPD and Mario Costeja González case (CJEU C-131/12, 2014) on the right to be forgotten. Consequently, this legislative and public awareness has also heightened firm attention on the privacy topic (Marketing Science Institute, 2016), and raised the question how to balance the need for analytics with customer privacy (Boulding, Staelin, Ehret, & Johnston, 2005; Rust & Chung, 2006; Verhoef, Kooge, & Walk, 2016). This trend of heightened attention is exacerbated by potential negative consequences for firms that ignoring this topic carries, from loss of consumer trust (Bart, Shankar, Sultan, & Urban, 2005; Deighton, 2005) or changing customer behavior (Lewis, 2005) to negative stock market valuations (Acquisti, Friedman, & Telang, 2006). The combination of governmental and public pressure has led to firms' "self-policing" (Wedel & Kannan, 2016). These firms incorporate privacy preserving measures into practice at the cost of analytical operations, limiting their capability to provide detailed insights on past customer behavior (Blattberg, Kim and Neslin, 2008, p. 78; Verhoef et al., 2016). Such firm behavior is in line with the prediction of economic theory showing that there is customer demand for privacy which firms should acknowledge (Rust, Kannan, & Peng, 2002). Wedel and Kannan (2016) state two important privacy preserving measures such firms take: *data minimization* (i.e. limiting the amount of data collected, and disposing of unneeded data) and *data anonymization* (i.e. assuring that data can not be connected to specific individuals). In practice, for many firms data minimization results in limiting data storage periods, and removing customer data after this period. In addition, data is analyzed anonymously or at aggregated levels to maintan data anonymization (Verhoef et al., 2016). Customers value firms that take these steps, as data usage, data security and (length of) data storage are listed as the most important concerns when sharing personal information (DMA, 2015). One well-documented example of a firm that applied these principles is that of data broker Choicepoint (e.g. Acquisti et al., 2006; Culnan & Williams, 2009; Otto, Antón, & Baumer, 2007), which anonymized and voluntarily stopped collecting and removed data from their systems (Culnan & Williams, 2009). Similarly, the European Internet service provider who provided one of the datasets for this study stores customer data for a year only, like many others in this industry. Interviews by the authors with several firms in a variety of other industries confirmed this trend, with the majority of the interviewed firms indicating that customer privacy played a large or very large part in their decision to store customer data.

In this article, our goal is to provide a method for churn prediction that combines the principles of data anonymization and data minimization while retaining the strong predictive ability and richness of state-of-the-art churn models (e.g. Ascarza & Hardie, 2013). We thus strike a balance between two seemingly incompatible objectives (Marketing Science Institute, 2016; Rust & Chung, 2006; Rust & Huang, 2014). In doing so, we show that privacy preservation does not have to come at the cost of analytical operations (as suggested by e.g. Blattberg, Kim and Neslin 2008, p. 78; Wedel & Kannan, 2016). To this end we develop a generalized mixture of Kalman filters (GMOK) model. This dynamic state-space model accounts for unobserved heterogeneity, while its recursive nature requires only knowledge of past model parameters to generate churn predictions. Our approach achieves data anonymization by aggregating information from prior periods into the model parameters, thereby not requiring the storage of privacy-sensitive individual-level panel data on past customer behavior (as in e.g. Ascarza & Hardie, 2013). Instead, it merely requires new cross-sectional information from the current period to update the model. After inclusion in the model the data need not be stored, which achieves data minimization.

We compare our approach to several other methods besides the HMM as introduced by Ascarza and Hardie (2013). These include logistic regression and classification trees due to their extensive usage in practice and good short-term predictive performance (Neslin et al., 2006). In addition, we investigate to what extent the introduction of either dynamics (i.e. using data from prior periods) or unobserved customer heterogeneity improves model performance compared to models that include neither component (logistic regression and classification trees) or models that include both components (GMOK and HMM). In Table 1 we provide an overview of the models included in this study and their respective model traits.

We consider two "worlds" in which our models are estimated: the panel data world, and the cross-sectional world. The reason for this is that some of the models we consider were developed with full past data availability in mind using panel data (notably the HMM, see Table 1), while others were developed without reliance on past data using cross-sectional data only (notably the