No Future Without the Past? Predicting Churn in the Face of Customer Privacy

Holtrop, Niels; Wieringa, Jakob; Gijsenberg, Maarten; Verhoef, Pieter C

Published in:
International Journal of Research in Marketing

DOI:
10.1016/j.ijresmar.2016.06.001

Publication date:
2017

No Future Without the Past? Predicting Churn in the Face of Customer Privacy[a]

Niels Holtrop [b,c]

Jaap E. Wieringa

Maarten J. Gijsenberg

Peter C. Verhoef

[a] 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.

[b] Address for correspondence: Niels Holtrop, Department of Marketing, Faculty of Economics and Business, University of Groningen, PO Box 800, 9700 AV Groningen, The Netherlands, Tel. +31 503639621; N. Holtrop@rug.nl.

[c] All four authors are affiliated with the Department of Marketing, Faculty of Economics and Business, University of Groningen, PO Box 800, 9700 AV Groningen, The Netherlands

No Future Without the Past? Predicting Churn in the Face of Customer Privacy

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.

Keywords: churn prediction, database marketing, customer relationship management, data privacy, Kalman filter, mixture model

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, Lehman, and Stuart 2004). Top executives recognize these benefits, reporting that customer retention is their top priority in terms of marketing spending (*Forbes* 2011).

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 and Schleicher 2016; Lemmens and 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 et al. 2006; Risselada, Verhoef, and 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 et al. 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 heigthened 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 et al. 2005; Rust and Chung 2006; Verhoef, Kooge and 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 et al. 2005; Deighton 2005) or changing customer behavior (Lewis 2005) to negative stock market valuations (Acquisti, Friedman, and Telang 2006). The combination of governmental and public pressure has led to firms' "self-policing" (Wedel and 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, Kooge and Walk 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, and 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, Kooge and Walk 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, Friedman, and Telang 2006; Culnan and Williams 2009; Otto, Antón, and Baumer 2007), which anonymized and voluntarily stopped collecting and removed data from their systems (Culnan and 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 and Hardie 2013). We thus strike a balance between two seemingly incompatible objectives (*Marketing Science Institute*

2016; Rust and Chung 2006; Rust and 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 and 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 and 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 GMOK model, see Table 1). While we can estimate cross-sectional models on panel data by considering each panel wave as a separate cross-section, we can not estimate panel data models on cross-sectional data as the same customers need not be present in every cross-section. Consequently, when past data is unavailable, panel data models collapse due to the absence of information on the same customers in prior periods, while cross-sectional models remain feasible. We show that the GMOK model has similar performance to the HMM in the panel data world while outperforming the simpler benchmark models. Importantly, the GMOK model retains this good performance in the cross-sectional world where past data is unavailable, while the HMM cannot be estimated (see Table 1). This way we show that when past is unavailable, analytical operations need not suffer, provided adequate modeling techniques are applied.

Model	Type of Data Required	Includes Data from Prior Periods (t-1)?	Includes Heterogeneity?	Aggregate Level Data Storage?
GMOK	(Repeated) Cross-Sectional Panel	✓	✓	✓
Hidden Markov Model (HMM)		✓	✓	-
Dynamics Only Model	(Repeated) Cross-Sectional	✓	-	✓
Heterogeneity Only Model	Cross-Sectional	-	✓	-
Classification Tree	Cross-Sectional	-	-	-
Logit Model	Cross-Sectional	-	-	-

Table 1: Overview of Models Included in this Study and Their Traits

As an additional benefit of our approach compared to simpler benchmarks, in periods following the period of model estimation, the decline in predictive accuracy is negligible compared with existing methods: whereas existing methods show an average decline in predictive performance of 20% after two periods (Risselada, Verhoef, and Bijmolt 2010), this

decline is only 1-3% on average for our approach. This increased accuracy results in cost savings for firms, as time- and resource-intensive tasks such as data collection, data preparation, and model estimation can be performed less often because the same model can be used repeatedly without loss of performance.

The remainder of this article unfolds as follows: First, we motivate the conceptual development of our approach according to current practices and prior research in this field. In line with these observations, we develop our model to fulfill the criteria we identified with regard to privacy and model requirements. Next, we offer an empirical illustration of this model in the insurance industry and compare its performance with a selection of benchmark models. Here, we first compare all the models in the panel data world, and subsequently compare their performance in the cross-sectional world. In the setting of the cross-sectional world, we also validate our findings using a second data set from the telecommunications industry. We conclude with a discussion of our findings and specify directions for further research.

2. RESEARCH BACKGROUND

To understand the practical requirements for churn prediction, we first discuss how churn prediction is performed in practice—specifically, which methods practitioners use. Next, we argue that we can preserve the important methodological features recent studies have uncovered in a setting in which access to past data is limited.

2.1 Churn Prediction Practice

By predicting churn before its actual occurrence, marketers can proactively target activities toward those customers at risk of churning to convince them to stay with a firm. This approach

can reduce the costs associated with churn (Blattberg, Kim, and Neslin 2008, p. 611; Neslin et al. 2006). Targeting is achieved by attaching a churn propensity to each customer in the customer base. Subsequently, a retention program is designed to cater to a selected subgroup of customers using their churn propensity as starting point (Blattberg, Kim and Neslin 2008, p. 615; Ganesh, Arnold, and Reynolds 2000).

Several methods can determine these probabilities, all of which predict future churn on the basis of historical churn data. The most popular and best-performing methods are logistic regression and classification trees (Neslin et al. 2006). These methods can be further improved by using model averaging algorithms such as bagging and boosting (Lemmens and Croux 2006; Risselada, Verhoef, and Bijmolt 2010). A common characteristic of these prior methods is their reliance on cross-sectional data. A limitation of cross-sectional data is the unavailability of past periods. If the model is not reestimated with new data but instead is reused beyond the period of estimation (as often happens in practical situations[1]), predictive inference suffers quickly (Risselada, Verhoef, and Bijmolt 2010). The latter study also indicates that the parameters of churn models are not stable over time; they vary in size, sign, and significance over the periods investigated, which illustrates the consequences of this information loss. To remedy this problem, Ascarza and Hardie (2013) develop a dynamic hidden Markov model (HMM) using a panel of customers to capture past customer behavior and unobserved customer heterogeneity. These authors show the improved predictive ability of churn of such a model for up to five subsequent periods. Thus, a dynamic model can alleviate the problems associated with cross-sectional approaches. A downside of this approach however is the reliance on customer panel

[1] With reuse we mean that a model that was estimated in a prior time period is used without modifications in a later time period. That is, the parameter estimates that were obtained from the prior model are saved and used to generate churn propensities for a new, current dataset. In our interviews, two-thirds of the practitioners interviewed indicated they operate in this way.

data. Limiting the storage of customer data for prolonged periods of time to comply with legislative and public pressure renders constructing and maintaining such panels infeasible (e.g. Wedel and Kannan 2016).

The main limitation of panel data methods thus lies in their reliance on the past information of single customers to make inferences. Next, we argue that an alternative approach that aggregates information about individual customers can capture the same model traits investigated by Ascarza and Hardie (2013). In so doing, we attenuate the limitations of panel data models in the face of stricter compliance with regulations, i.e. data anonymization and data minimization.

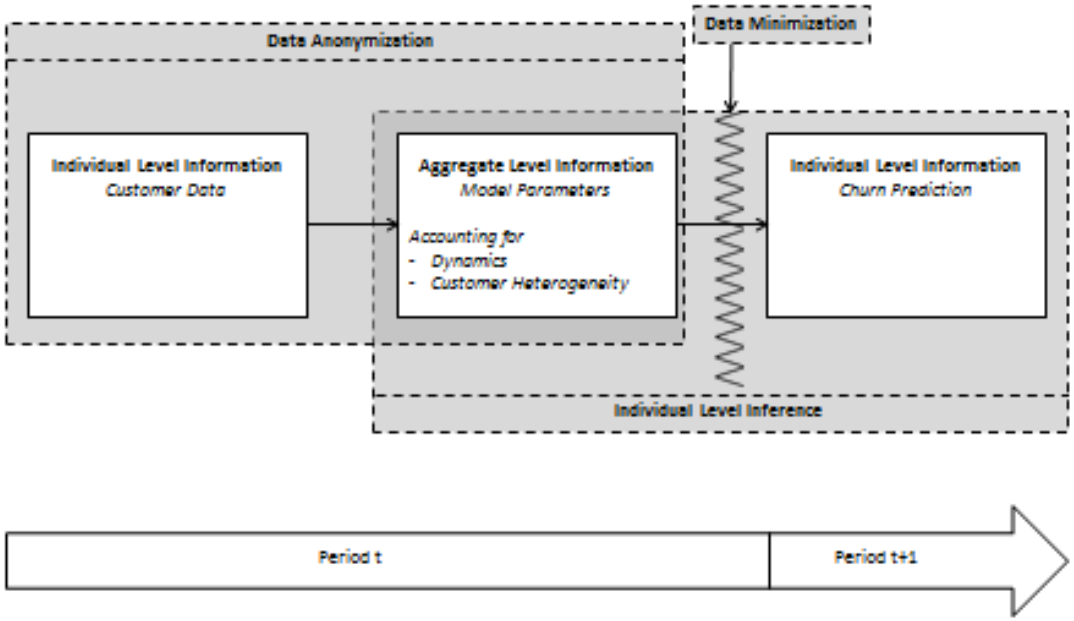
2.2 Balancing data limitations with model performance

In order to develop a method that fulfills the criteria of *data anonymization* and *data minimization* while retaining a good predictive performance, we need to consider an approach that can balance these requirements. In Figure 1 we therefore outline our approach to develop such a model. In order to retain a good predictive performance, we first analyze the model traits that have emerged from prior research and are associated with model performance. Subsequently, we accommodate these traits in a framework that accounts for data anonymization and data minimization.

The main differences between the cross-sectional methods (i.e. logistic regression and classification trees) discussed previously and the model of Ascarza and Hardie (2013) are twofold: The latter model accounts for 1) information from prior periods (dynamics), and for 2) unobserved customer heterogeneity (see also Table 1). These traits capture the relevant characteristics of the underlying data generating process. Therefore, inclusion of these traits is a

necessary and sufficient condition for an accurate model (e.g. Jerath, Fader and Hardie 2016; Wedel and Kannan 2016), while exclusion will reduce model performance (e.g. Zhao, Zhao and Song 2009).

Figure 1: Overview of Model Development[2]



However, these traits need not be captured at the individual level as in Ascarza and Hardie (2013). We propose to account for these traits at the aggregate instead of individual level similar to Jerath, Fader and Hardie (2016). However, instead of adapting the data format as these authors suggest, we adapt the model framework to accommodate these model traits. By accounting for these traits at the aggregate level we achieve the condition of data anonymization (see Figure 1). We account for these model traits at the aggregate level as follows: First, instead of capturing heterogeneity at the individual customer level, we allow for customer segments. Ascarza and Hardie (2013) support this with their finding that there exist three clusters that

[2] Our model process works as follows: At time period *t*, the individual level information serves as input for our model (Box 1). At this point, data is fed into the model (Box 2), in which it is aggregated. As such, we achieve data anonymization. Once fed into the model, data is no longer required, and can be removed (data minimization). At time $t + 1$, the model can be used on new individual data for inference at that time point (Box 3).

exhibit a different evolution of customer behavior over time. Second, we then capture the dynamics of the data generating process by allowing time-varying parameters within these segments through a state-space framework. Hence, both unobserved heterogeneity and dynamics are captured at an aggregate instead of individual level.

An important motivation to use the state-space framework is its recursive nature. This way, a state-space model is updated once new information becomes available. More importantly, the information from prior periods is retained in the model parameters. It is this feature that allows for data minimization, as data from prior periods is no longer required once it is incorporated in the model parameters. Subsequently, in period $t + 1$ inference at the individual level can be made by assigning customers to relevant segments based on their characteristics (see Figure 1).

Aggregate analysis combined with indidual level inference also alleviates the data requirements. Instead of panel data capturing the behavior of individual customers, this approach merely requires repeated cross-sectional data to capture the aggregate customer base traits of dynamics and customer heterogeneity.

3. METHODOLOGY

In this section we translate the observations of the previous section into a modeling framework that can be used to predict churn. In particular, we develop a generalized mixture of Kalman filters (GMOK) model that takes into account all sources of variation among customers and sources of variation over time using repeated cross-sectional data from the customer database. We first describe how we model the dynamics and then extend the model with unobserved heterogeneity.

3.1 Churn Dynamics Model

The starting point for our model is the standard logistic regression model, which we extend with time-varying parameters to allow for carry-over of past information (dynamics). By taking the logit model as a starting point, we follow existing literature indicating its good performance when predicting customer churn (e.g., Neslin et al. 2006). We apply a state-space approach to allow for time-varying parameters in the logit model (for prior applications, see, e.g., Cain 2005; Naik, Mantrala, and Sawyer 1998; Osinga, Leeflang, and Wieringa 2010). Our basic model specification thus becomes

$$(1)\quad \pi_{it} = P(y_{it} = 1) = \Lambda(X_{it}\beta_t) = \frac{\exp(X_{it}\beta_t)}{1 + \exp(X_{it}\beta_t)},$$

$$(2)\quad \beta_t = \beta_{t-1} + \zeta_t,\ \zeta_t \sim N(0, Q_t)\ ,$$

$$(3)\quad \beta_0 \sim N(b_0, Q_0).$$

Equation (1) is the observation equation, which relates the churn probability π_{it} for customer i in calendar period t to a vector of observed explanatory variables X_{it}. The observed variable y_{it} is a binary variable, which equals 1 if customer i churned in period t and 0 otherwise. Here, $i = 1, \ldots, n$, and $t = 1, \ldots, T$. We relate y_{it} and X_{it} through a logistic transformation denoted by $\Lambda(\cdot)$, which results in a logistic regression model with time-varying parameter vector β_t. We specify the transition equation (Equation (2)) of the state-space model as a random walk, which provides a parsimonious yet flexible way to accommodate various dynamic patterns in parameter evolution. We specify the error term ζ_t of the transition equation as a white-noise process with diagonal covariance matrix $Q_t = Q$. Equation (2) provides an aggregate measure of the dynamics that underlie the data generating process. As such, it is not required to observe individual

changes over time, but these changes are captured at this more aggregate level to achieve data anonymity. Individual deviations thereof are captured through X_{it} in Equation (1) (e.g. Lu 2002, Yan et al. 2001). Finally, Equation (3) specifies the hyperparameters required to initialize the Kalman filter that we will use to estimate this two-equation model.

The model is also recursive. In each time period, the model is fed with the most recent cross-sectional information on y_{it} and X_{it}, after which Equation (2) is updated. All that is required for the next period is new information pertaining to y_{it} and X_{it} for that period, as all the information from the past is transferred through the parameter evolution in Equation (2). In this manner, it is not necessary to store customer data from the past, but it is sufficient to store the model parameters and update them when new information becomes available. Hence, data minimization is achieved. Next, we discuss how to update the model in each period.

3.2 Kalman Filter Estimation

In general, state-space models can be estimated by the Kalman filter (Durbin and Koopman 2012). This recursive algorithm updates the model when new information becomes available. The standard linear Kalman filter assumes that the time series observations in Equation (1) are normally distributed, which is not the case here given the binary nature of y_{it}. Fahrmeir and Tutz (1994) relax the normality assumption and present Kalman filter recursions for the case in which the observation equation model is based on nonnormal time series, which includes our model specification (we provide the relevant recursions in Web Appendix A). Their approach applies to both univariate and—as is the case here—multivariate dependent variables. In the multivariate case we consider here, this algorithm updates the model over all n customers present at time t. This feature allows the algorithm to aggregate all the individual-level information on the dynamic process to update the model independent of information from the past of those same

customers. Thereby, it does not require individual-level information on those same customers anymore after inclusion in the model as in models based on panel data.

The estimation of the model proceeds in two steps: First, we determine the values of b_0, Q_0, and Q using numeric maximum likelihood estimation in the first period, given some initial values for $\beta_1^* = (\beta_0, \beta_1)$. The relevant log-likelihood to maximize is

$$(4)\ \ell(\beta_1^*) = \sum_{i=1}^{n} l_{i1}(\beta_1|y_{i1}, X_{i1}) - \frac{1}{2}(\beta_0 - b_0)'Q_0^{-1}(\beta_0 - b_0) - \frac{1}{2}(\beta_1 - \beta_0)'Q^{-1}(\beta_1 - \beta_0),$$

where l_{i1} is the logistic log-likehood contribution of customer *i* based on the data available in period 1. This is the likelihood derived in Equation B.7 of Web Appendix B, adapted for the first period. Second, given these initial values, we estimate the state parameter vector β_t^* using the Kalman filter recursions of Fahrmeir and Tutz (1994). In line with Fahrmeir and Wagenpfeil (1995), we iterate these two steps until the likelihood has converged. For subsequent periods, the likelihood in this case becomes

$$\begin{aligned}(5)\ \ell(\beta_t^*) = &\sum_{k=1}^{t-1}\sum_{i=1}^{n} l_{ik}(\beta_k^*|y_{ik}^*, X_{ik}^*) - \frac{1}{2}(\beta_0 - b_0)'Q_0^{-1}(\beta_0 - b_0) \\ &- \frac{1}{2}\sum_{k=1}^{t-1}(\beta_k - \beta_{k-1})'(Q_k)^{-1}(\beta_k - \beta_{k-1}) + l_{ik}(\beta_k|y_{ik}, X_{ik}) \\ &- \frac{1}{2}(\beta_t - \beta_{t-1})'Q_t^{-1}(\beta_t - \beta_{t-1}),\end{aligned}$$

where l_{it} is the logistic log-likehood contribution of customer *i*, $\beta_t^* = (\beta_0, \ldots, \beta_t)$, and b_0, Q_0, and Q are not updated anymore.[3] We smooth the parameter vector β_t^* using the linear Kalman

[3] Note that this procedure does require knowledge of the likelihood value l_{it}^* of previous periods if $1 < T \leq t\text{-}1$. However, this value is known when the state for period *t* has been determined. Equation (5) shows that the

smoother described by Fahrmeir and Tutz (1994), which facilitates the interpretation of the outcomes.

3.3 Adding Unobserved Heterogeneity

Next, we extend our prior model specification to allow for the other model trait identified to be important: unobserved customer heterogeneity. Following the literature on finite mixture models (e.g., Wedel and Kamakura, 1998), we assume two or more unknown groups exist in the data and allow the model parameters to differ among groups. This feature, in combination with the dynamic logistic regression model, allows parameters to vary simultaneously over time and over groups. Assume latent segments $j = 1, \ldots, J$ exist in the data, where J is fixed over time and set the first time the model is estimated. Returning to Equations (1)–(3), we now assume the data to be generated by a mixture state-space model and estimate a mixture of dynamic logistic regression models given by

$$(6)\quad \pi_{it} = P(y_{it} = 1) = \sum_{j=1}^{J} \lambda_t^j \Lambda(X_{it}\beta_t^j),$$

$$(7)\quad \beta_t^j = \beta_{t-1}^j + \zeta_t^j,\ \zeta_t^j \sim N\left(0, Q_t^j\right),$$

$$(8)\quad \beta_0^j \sim N\left(b_0^j, Q_0^j\right),$$

where λ_t^j represent the proportions that give the mixture weight for each cluster. The λ_t^j have the following constraints: $\Sigma_j\, \lambda_t^j = 1$ for each time period and $\lambda_t^j \geq 0$. Furthermore, we set $Q_t^j = Q^j$ and diagonal as before. To estimate this model, we build on the maximum likelihood approach outlined in the previous section. For applications in which the time series observations in

original data are no longer required after period t, only the likelihood values l_{it}^* summed over all customers and the corresponding parameters β_t^*. Therefore, it is sufficient to retain these values and the parameters.

Equation (6) are normally distributed, Calabrese and Paninski (2011) derive an EM algorithm to estimate a mixture of Kalman filters model. Combining this algorithm with the modified Kalman filter recursions of Fahrmeir and Tutz (1994), we arrive at a generalized mixture of Kalman filters (GMOK) model that allows the dynamic logistic regression model to be estimated for two or more groups. The model can be estimated using maximum likelihood estimation by considering cluster membership as missing data and applying the EM algorithm (Dempster, Laird, and Rubin 1977). We outline the approach here; for a full derivation of the EM algorithm for this case, see Web Appendix B.

We obtain the likelihood of the GMOK model by appending the likelihood from the model specified by Equations (1)-(3) with an unobserved cluster membership indicator variable that is treated as missing data in the EM algorithm and is replaced by its expected values. We show in Web Appendix B that this yields the expected log-likelihood function

$$
\begin{aligned}
E\left[\ell\left(\beta_t^{j*}\right)\right] = & \sum_{j=1}^{J}\sum_{k=1}^{t-1}\left(p_k^{j*}\left[\log\left(\lambda_k^j\right) - \sum_{i=1}^{n} l_{ik}\left(\beta_k^{j*}|y_{ik}^*, X_{ik}^*\right)\right]\right) - \frac{1}{2}\sum_{j=1}^{J}\left[\left(\beta_0^j - b_0^j\right)'\left(Q_0^j\right)^{-1}\left(\beta_0^j - b_0^j\right)\right] \\
& - \frac{1}{2}\sum_{j=1}^{J}\sum_{k=1}^{t-1}\left[\left(\beta_k^j - \beta_{k-1}^j\right)'\left(Q_k^j\right)^{-1}\left(\beta_k^j - \beta_{k-1}^j\right)\right] \\
& + \sum_{j=1}^{J} p_t^j\left[\log\left(\lambda_t^j\right) - \sum_{i=1}^{n} l_{it}\left(\beta_t^j|y_{it}, X_{it}\right)\right] - \frac{1}{2}\sum_{j=1}^{J}\left[\left(\beta_t^j - \beta_{t-1}^j\right)'\left(Q_t^j\right)^{-1}\left(\beta_t^j - \beta_{t-1}^j\right)\right]
\end{aligned}
$$

where $\beta_t^{j*} = (\beta_0^j, \ldots, \beta_t^j)$ for $j = 1, \ldots, J$. We subsequently maximize this likelihood. The maximization of this likelihood is similar to the mixture model case (Wedel and Kamakura, 1998) and consists of two separate parts. In the first part, we obtain the mixture weights λ_t^j which represent the relative weight of each mixture component. The second and remaining part of the likelihood is similar to the likelihood of the model given in Equations (1)–(3) and thus can be maximized by the procedure outlined in the previous section, provided we make a small

correction to the Kalman filter (for details, see Web Appendix B). The expectation and maximization steps are iterated until the likelihood value has converged.

The preceding procedure yields estimates β_t^{j*} for each of the *j* clusters, which are smoothed using the linear Kalman smoother described by Fahrmeir and Tutz (1994). These values capture the cluster-specific evolution of the regression parameters. Using the parameters thus obtained, we generate predicted churn probabilities for each customer in future periods by assigning them to the segment with the highest posterior likelihood for that period (e.g. Reimer, Rutz and Pauwels 2014; Vermunt and Magidson 2013). That is, we compute $X_{it}\beta_t^{j*}$ for customer *i* in segment *j* at time *t*, and assign customers to the segment with the highest posterior probability $P(i,j) = \frac{\lambda_t^j \exp\left(\ell\left(\beta_t^{j*}\right)\right)}{\Sigma_{k=1}^{J} \lambda_t^k \exp\left(\ell\left(\beta_t^{k*}\right)\right)}$ for that period, where ℓ is the likelihood given in Equation (B.10) in Web Appendix B.

4. DATA DESCRIPTION

Our focal data set comes from a large Dutch health care insurer with yearly data on customer churn for 2004–2012. Yearly data are appropriate in this case because the industry's contractual setting limits consumers' opportunity to churn to at most once a year (Dijksterhuis and Velders 2009). A customer churns if he or she is with the insurer at the start of the year but is no longer at the end of the year. During the observation period, the Dutch health care system was completely restructured. One of the goals of the change was to encourage customers to switch insurers (see Douven, Mot, and Pomp 2007). Whereas the churn rates for 2004 and 2005 were 8.8% and 7.4%, respectively, the policy change sharply increased the churn rate to 34.3% for 2006 in this data set. After this increase, the churn rate dropped to 3.7% in 2007 and then

steadily increased to 4.3% in 2012. The temporarily increased churn rate provides an opportunity to test our model specification in the face of changing market situations and to study how model predictions might be affected by such a change.

For our study we create two types of data sets from this data: a panel data set, and a cross-sectional data set. As noted in Section 1, while we can estimate cross-sectional models on panel data by considering each panel wave as a separate cross-section, the reverse does not hold. To allow for a fair comparison between models, we thus consider these two 'worlds' separately. In Table 2 we provide an overview of the sample sizes and churn percentages for each of these datasets. The next section provides further detail on data set construction.

Period	Panel Sample Size	Panel Churn Percentage	Cross-Sectional Sample Size	Cross-Sectional Churn Percentage	Cross-Sectional Training Sample Size	Cross-Sectional Validation Sample Size
2004	5,000	8.5	1,034,427	8.8	5,000	11,167
2005	4,573	19.1	859,063	7.4	5,000	13,583
2006	3,699	35.2	795,519	34.3	5,000	2,898
2007	2,397	4.0	606,861	3.7	5,000	27,231
2008	2,301	1.7	527,104	2.1	5,000	47,023
2009	2,261	2.1	741,059	2.1	5,000	47,641
2010	2,214	2.4	578,108	2.3	5,000	48,022
2011	2,160	2.4	812,203	3.7	5,000	27,110
2012	2,109	2.9	1,109,094	4.3	5,000	23,329

Table 2: Descriptive Information for the Panel Data and Insurance Data Sets

Both data sets contain variables that can be divided in several groups, such as sociodemographic (e.g., age, family status) or socioeconomic (e.g., income, social status) variables and information about relationship characteristics (e.g., length). All these variables are recorded for all the time periods we considered. Table 3 provides a more detailed overview of the variables in the data set.

Variables Insurance Data	
Sociodemographic	Age, Distance Insurance Shop, Family without Kids
Socioeconomic	BSR Grouping*, Education Level, Income, Social Class, Prosperity Level
Relationship characteristics	Relationship Age, # Insurance Shop Visits

***BSR Grouping is a third-party segmentation scheme used by the insurer to segment its customer base**

Table 3: Available Variables for the Insurance Data

5. MODEL COMPARISON APPROACH

We compare our models within two 'worlds': a panel data world and a cross-sectional data world, to allow for a fair comparison of models. In the panel data world, we estimate the GMOK model and a selection of benchmark models either by considering each panel wave as a cross-section, or by including all data up to the period of estimation, depending on the model type. We optimize model fit for each model specifically. This procedure implies that the variables included in each model can differ from period to period, and from model to model (see Section 6.2 for more details). Subsequently, we generate churn predictions for the same customer in future periods. We compare the predicted churn probability to the observed churn behavior using top decile lift and Gini coefficient as measures of model performance (see Section 5.4).

In the cross-sectional world, we follow prior studies (e.g. Neslin et al. 2006), and estimate the GMOK model and a selection of benchmark models on a training sample. All the models are estimated according to specific criteria so that model fit is optimized within the training sample. Next, we generate churn predictions using a holdout sample, and compare the predicted churn probabilities to the observed churn in the holdout data using the same performance measures as in the panel data world.

5.1 Creating the panel data world

The difference between the cross-sectional world and the panel data world is that in the panel data world we observe the same customer in each period until the customer churns, while in the cross-sectional world the same customer need not be included in a cross-section even though it might still be an active customer. This implies that the cross-sectional world places less restrictions on the data required, as it does not require data from previous periods on the same customer. Instead, only a sample of customers pertaining to the period of model estimation is needed. Hence, in the cross-sectional world we can compare model performance when data from the past is unavailable, while the panel data world provides us with a situation of full past data availability.

To construct the panel data world, we selected 5000 customers that were active in 2004, and tracked their behavior until 2012. While this number might seem low compared to the sample sizes of the cross-sectional world (see Table 2), the HMM benchmark model of Ascarza and Hardie (2013) requires the computationally intensive estimation of individual-level heterogeneity. We therefore limit the amount of customers included to maintain computational feasibility. In Table 2 we provide the number of active customers and churn rate in each year. Another difference between the cross-sectional world and the panel data world is that we do not split the data in training and holdout samples (see Section 5.2). Instead, we estimate the model using data relevant to the period of estimation, and generate our churn predictions for the same customer in the years after the period of model estimation

5.2 Creating the cross-sectional data world

Given the size of the full database model estimation using maximum likelihood becomes infeasible when using all observations (e.g. Reimer, Rutz and Pauwels 2014). We

therefore adopt a subsample approach (Musalem, Bradlow and Raju 2009), where we estimate the models on a subsample of the data. First, we create balanced training samples (50% churners, 50% nonchurners) by randomly sampling 5,000 observations from the period of model estimation from our data. This implies that the same customer need not appear in multiple samples, even though the customer might still be active. We use balanced training samples because prior research shows that models calibrated on such samples perform more reliably than those calibrated on proportional samples without loss of efficiency (Donkers, Franses, and Verhoef 2003; Lemmens and Croux 2006). We use an equal number of observations per period to avoid between-period biases due to sample size variation, which might influence model reliability. An additional constraint in determining the number of observations to use is the limited total number of churners available in each period due to generally low churn rates. Second, we use the remaining observations to generate holdout samples with a churn rate equal to the full database for model validation. We use non-balanced holdout samples to simulate the firm practice of using churn models to obtain churn propensities for the *entire* customer database (i.e. where the churn rate is equal to that of the full data). As we already used part of the churner and non-churner observations to create training samples, we use the remaining churner observations complemented by a random sample of the remaining non-churners to create a holdout sample with a churn rate equal to that of the full data. This way, we avoid biases due to using the same observations in both the training and holdout sample. As shown in Table 2, this leads to varying holdout sample sizes due to the varying number of churners available after training sample construction combined with a churn rate that varies from year to year.

5.3 Benchmark Models

In Table 1 we provide an overview of all the models included in this study, and the characteristics of each model included. A few things are of note. First, the HMM is the only model native to the panel data world. The other models emanate from the cross-sectional world. To estimate the cross-sectional models in the panel data world, we treat each panel wave as a seperate cross-section. Hence, these models, including the GMOK model, do not account for the specific structure of the panel data world, while the HMM does. Conversely, due to the lack of a panel structure the HMM model is not included in the cross-sectional world. Second, only three models use in some way data from prior periods: the GMOK model, the dynamics only model (a restricted GMOK model, see below), and the HMM. Third, customer heterogeneity is only accounted for in three models: the GMOK model, the heterogeneity only model (a restricted GMOK model, see below), and the HMM. Finally, the feature of data aggregation, that achieves data anonymization and data minization, is a feature that only the GMOK and dynamics only model have. In the remainder we provide more information on each model, and the reason for its inclusion.

Given its importance as a model with similar model traits, but without the characteristics of data minimization and data anonymization due to the required panel structure, we first benchmark against a version of the HMM developed by Ascarza and Hardie (2013). Because we need to observe several periods prior to the period of estimation to model transitions between states, we estimate this model only three times using data until 2009, 2010, and 2011 instead of generating predictions over the full data period as is done for the other models. Consequently, we only have churn predictions for up to three periods ahead for this model[4]. We compare these

[4] In our results section, we compare the HMM predictions to the other models where the predictions for the other models include all (hence, more) time periods for completeness and comparability to those of the cross-sectional

predictions with the observed churn behavior of these customers in later periods. The model itself is a modification of the binominal model discussed in the Web Appendix of Ascarza and Hardie (2013). We adapt this model to fit our application, as we do not have access to usage data as in the latter study (for details, see Web Appendix C). To determine the number of segments to select for this model, we use the log marginal density and deviance information criterion to select the best fitting model from models with two to four segments.

As a second benchmark, we use two models that are known to provide good predictions of customer churn: logistic regression and classification trees with a bagging procedure[5] (Lemmens and Croux 2006; Neslin et al. 2006; Risselada, Verhoef, and Bijmolt 2010). As can be seen from Table 1, these cross-sectional models neither account for dynamics nor for heterogeneity. Due to their popularity and proven effectiveness we include them as simple benchmark models in both the cross-sectional world and the panel data world. For the logistic regression model, in the panel data world we treat each year as a cross-section, while in the cross-sectional world we estimate the model separately on each training sample. In both cases, we estimate various versions of the model where we consider all possible combinations of explanatory variables, and select the best fitting model in each year according to the Bayesian information criterion (BIC; Schwarz 1978). Similarly, for the classification tree in the panel data world we consider each year as a separate cross-section, while in the cross-sectional world we

world. In a separate analysis (reported in Web Appendix D) we also compare the HMM predictions to the other models, where the other models were estimated on the same data period (i.e. 2010-2012). The results are similar in terms of ordering of the models, but show fewer significant differences between models. See also footnotes 7 and 8.

[5] For these prior studies, the authors also considered classification trees without a bagging procedure. Their findings suggest that classification trees with a bagging algorithm provide superior predictive performance compared to those without a bagging procedure. We confirmed these findings using our dataset, but do not report them to keep the exposition clear by only reporting a limited number of benchmark models. Therefore, hereinafter "classification trees" refer to the case in which a bagging algorithm is applied to these classification trees. For logistic regression, bagging algorithms did not improve predictions (as was also found in these prior studies); thus, we only consider normal logistic regression here.

estimate the model separately on each training sample. In both cases we estimate models with a variety of explanatory variables, and use a splitting rule based on the Gini index of diversity. After that we select the model with the best fit using cost-complexity pruning to avoid overfitting (Breiman et al. 1984). For the bagging procedure, we estimate the model on B bootstrap samples and average the predictions of these B models to obtain the final model prediction for that period. To determine B, we follow previous work (Lemmens and Croux 2006; Risselada, Verhoef, and Bijmolt 2010) and estimate our model for $B = 50$, $B = 100$ and $B = 150$ to determine the value for which the holdout sample top-decile lift does not change. We find that in both worlds a large improvement occurs when moving from $B = 50$ to $B = 100$, but that no improvement occurs when moving from $B = 100$ to $B = 150$. Hence, we set $B = 100$, which is the same value as reported in previous studies. In addition, when generating predictions for the cross-sectional world we apply the correction of Lemmens and Croux (2006) to correct for the balanced training samples.

In addition to these well-known models, we estimate two other benchmark models representing restricted versions of our proposed GMOK model: a model that only accounts for dynamics and one that only accounts for customer heterogeneity. By doing so, we can analyze the extent to which the introduction of dynamics or unobserved heterogeneity induces predictive performance improvements and determine whether accounting for one or the other would be sufficient. The model only accounting for dynamics is a one-segment version of the GMOK model, in which all the observations are pooled and no segments are assumed (i.e., the model described by Equations (1)–(3)). In both the panel data and cross-sectional world this model is estimated using data up to and including the year of model estimation, where for the panel data world we treat each year as a separate cross-section. By varying the number of variables included

in the model specification, and comparing these models using BIC, we select the model with the best fit. The model that only accounted for customer heterogeneity is a mixture model, estimated using the GLIMMIX algorithm (Wedel and Kamakura 1998). Similar to the logit and classification tree models, in the panel data world we treat each year as a separate cross-section, while this model is estimated separately for each cross-section in the cross-sectional world. When estimating this model, we varied the number of segments between two and six, and the variables included in the model. Based on BIC we select the model whose combination of variables and segments showed the best fit. Estimation of the GMOK model combines the above approaches: For both the panel data and cross-sectional worlds we estimate the models up to and including the data for the year of model estimation, where in the panel world each year is treated as a separate cross-section. Subsequently, we estimate the model for each time period using two to six segments per period and select the model for which the combination of variables and number of groups has the lowest BIC value.

5.4 Model Performance Measures

We use two common measures to assess model performance: top decile lift and Gini coefficient (e.g. Lemmens and Croux 2006; Neslin et al. 2006). The top decile lift is defined as the fraction of churners in the top decile divided by the fraction of churners in the whole set (Blattberg, Kim, and Neslin 2008, p. 318). We compute the Gini coefficient by dividing the area between the cumulative lift curve and the 45-degree line by the total area under the 45-degree line (Blattberg, Kim, and Neslin 2008, p. 319). We apply a bootstrap approach and estimate the model for each period on 50 bootstrap samples and then compute the top decile lift and Gini coefficient for each bootstrap sample. In so doing, we construct pointwise 95% bootstrap

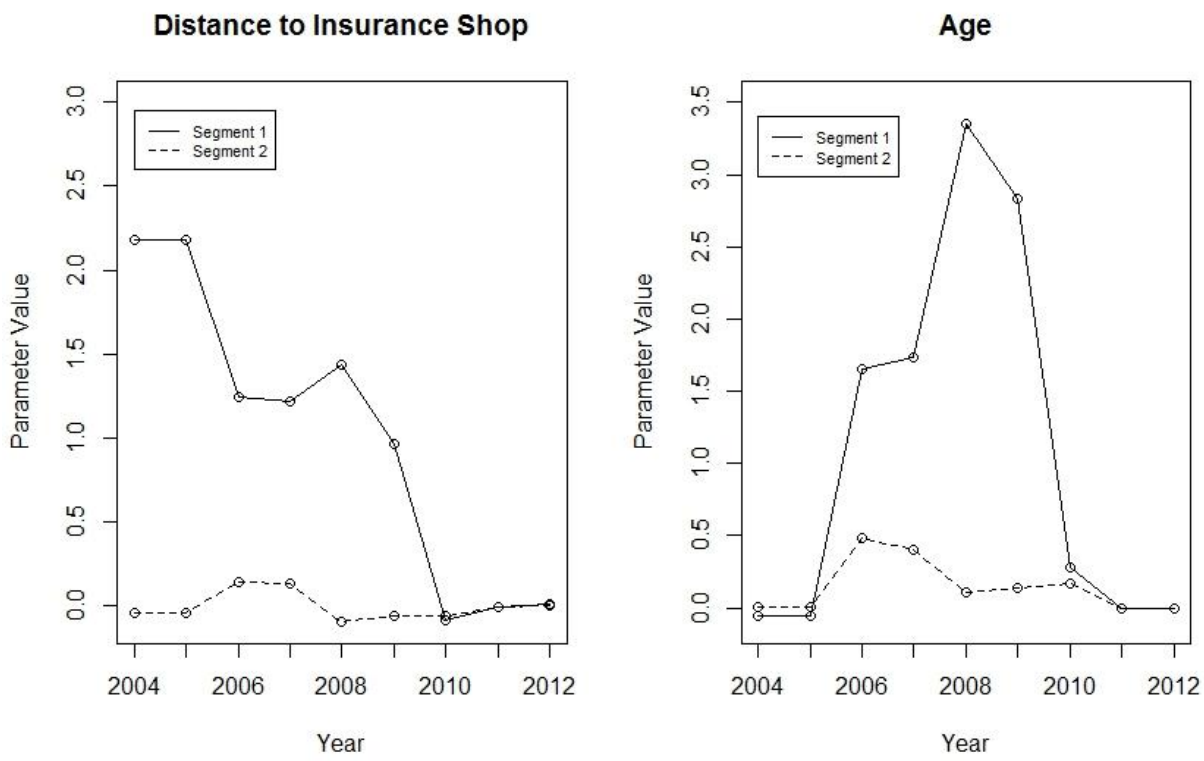
confidence intervals for the top decile lift and Gini coefficient, which we use to test for significant differences among models.

6. RESULTS

6.1 Estimation Results of the GMOK Model

Given our interest in the performance of the GMOK model compared with existing methods, we provide some additional background on the estimation results obtained from this model. According to BIC, a two-segment model was preferred to other models. Both segments contain churning customers as indicated by post-estimation comparison of segments. Thus, the results do not show a degenerate cluster solution but rather display two segments that show differences in churn behavior, as reflected through their estimated parameters, and the evolution thereof over time. To illustrate, Figure 2 presents the evolution of the (significant) parameters of two variables over time. For distance to insurance shop, the importance decreases steadily over time for segment 1 while remaining stable for segment 2. For segment 1, the variable age becomes a strong determinant of churn following the policy change in 2006, whereas this variable remains stable and unimportant for segment 2. In combination, these illustrations confirm the extent to which the GMOK model can account for differences over time and among segments.

Figure 2: Evolution of (Significant) Parameters over Time



6.2 Variable inclusion across models

Before providing the results on the performance measures, we first give some insight in the structure of the various models included. The model estimation procedures outlined in Section 5.3 in combination with the bootstrap procedure outlined in Section 5.4 have as a consequence that while all variables serve as input to all models at each point in time, different variables are included in each model depending on the model type and the data sample used. To get some insight in the importance of each variable across models, we provide an overview of which variables are included in each model in Table 4. We provide these in relative terms, as the HMM model is less frequently estimated than the other models. We find that across models, the classification tree and dynamics only model include the most variables; the heterogeneity only model includes the least. The GMOK model and HMM model include mostly similar variables,

although the HMM excludes some variables more frequently than the GMOK model. Across models, variables related to relationship characteristics (relationship age, number of insurance shop visits) seem to be the most important variables influencing churn.

Variable	GMOK (%)	HMM (%)	Logistic Regression (%)	Classification Tree (%)	Dynamics Only (%)	Heterogeneity Only (%)
Age	36	100	69	91	71	0
BSR Grouping	37	44	13	95	66	1
Distance Insurance Shop	36	92	12	100	58	0
Education Level	35	52	14	86	64	0
Family Without Kids	36	12	34	71	62	0
Income	36	48	100	79	64	0
# Insurance Shop Visits	36	72	23	98	60	88
Prosperity Level	36	36	79	99	60	0
Relationship Age	36	0	22	100	64	4
Social Class	38	8	31	89	63	0
Total # Models	450	200	450	450	450	450

Table 4 Percentage of Models that Includes Variable

6.3 Comparison of Model Predictions: Panel Data World

We provide plots illustrating the performance of the GMOK model compared to the benchmark models in the panel data world. Our findings for the top decile lift can be found in Figure 3, the findings for the Gini coefficient are given in Figure 4. For ease of interpretation, we provide graphs showing the top decile lift and Gini coefficient averaged over period of prediction. In these graphs, the time period t denotes the period of model estimation (i.e the same observations that are used for model estimation are used to generate this prediction), and $t + 1$, $t + 2$, and so on refer to predictions using a model with parameters estimated at time t (i.e. they use the observations of the same customers in future periods, provided they have not churned before).

This way, we can compare all models on usage for churn forecasts both within period of model estimation (t) as well as in future periods (t + 1 and onwards). Hereby we represent the typical usage occasions of these models in practice. The error bars in these graphs indicate 95% bootstrap confidence intervals for the top decile lift and Gini coefficient respectively.

In terms of top decile lift, we find that there is no significant difference between the GMOK model and the HMM save for period t + 3. Furthermore, both models show a better performance than the remaining benchmark models, which ranked by decreasing performance are the classification tree[6], the logit model, the model with only dynamics and the model with only heterogeneity. For the Gini coefficient, we find that the model with only heterogeneity performs best. This model is followed by the GMOK model and the HMM model, which show small significant differences in performance.[7] These models are then followed by the classification tree, logit model and the model with only dynamics

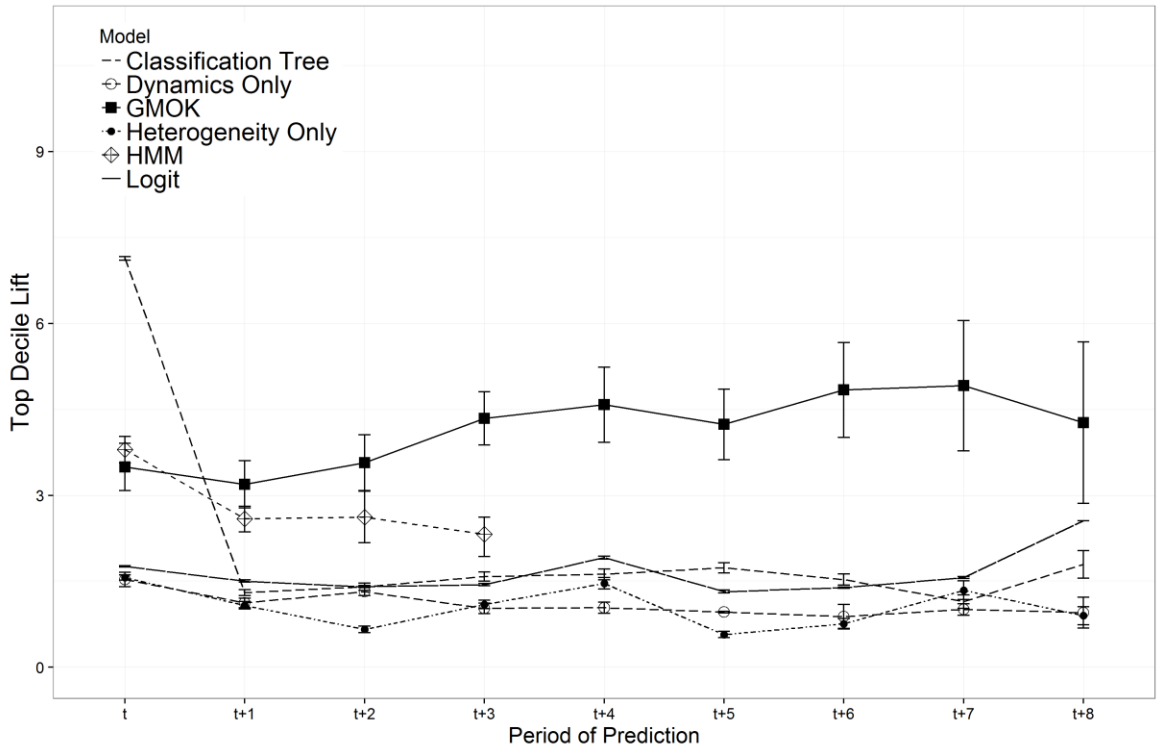
At first glance, our results show a mixed performance of models across metrics. We find that the GMOK model performs best in terms of top decile lift, but that it is second to the heterogeneity only model in terms of Gini coefficient. However, this latter finding is degenerate as the heterogeneity only model does not succeed in separating churners from non-churners, but instead assumes everyone does not churn. While this leads to good performance in terms of Gini coefficient, performance in terms of top decile lift suffers, as illustrated by Figure 3. Compared to the heterogeneity only model, the GMOK model is able to accurately distinguish churners from non-churners due to the addition of dynamics to the model, as evidenced by the good performance on both top decile lift and Gini coefficient. Hence, we conclude that on both metrics

[6] Note that the classification tree shows a good performance in period *t*, which in this case is the in-sample (same obseservations used for estimation and prediction) performance of the model. However, the predictive capabilities (period *t* +1 and onwards) of this model are worse than those of the GMOK model and the HMM.

[7] The analysis in Web Appendix D shows that when compared over the same time period, the GMOK model, HMM and heterogeneity only model show non-significant performance differences.

the GMOK model is the best performing model compared to the heterogeneity only model. The average improvement with respect to the best performing benchmark exceeds 9% in terms of the top decile lift, and 10% in terms of Gini coefficient.

Figure 3: Average Top Decile Lifts for Panel Data Models Estimated at Time *t* (95% Bootstrap Confidence Intervals, Insurance Data)



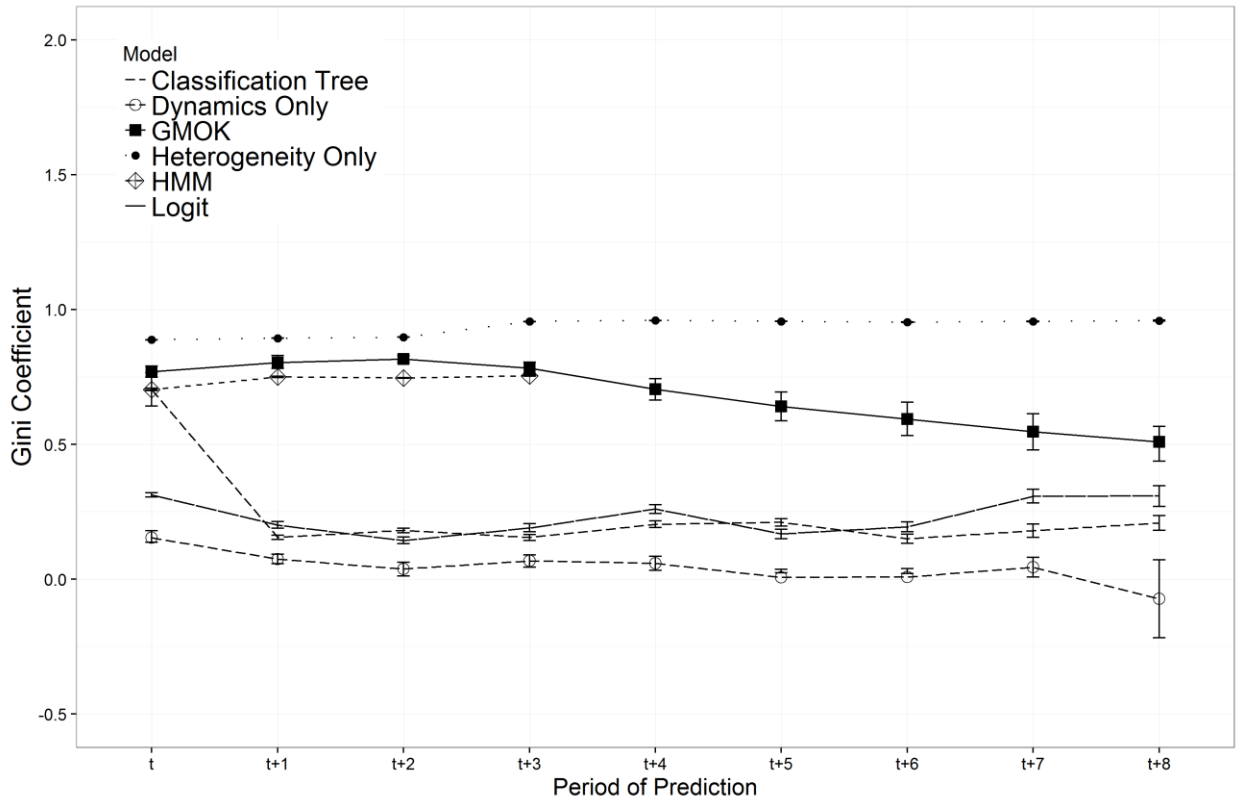
We find that there is no significant difference in performance between the GMOK model and the HMM up to period $t + 2$[8]. Hence, the GMOK model is able to achieve similar performance compared to the HMM. This implies that in the situation where a full set of past data is available, the performance of our approach equals that of the HMM. The similarities in

[8] The significant difference for period $t + 3$ is due to the single forecast we have for the HMM in this period, whereas we have multiple for the GMOK model. The analysis presented in Web Appendix D shows that when compared over the same time period, this difference becomes insignificant as well. In absolute sense, the HMM performs slightly better even. Additionally, there is no significance difference between GMOK/HMM and logistic regression in period $t + 1$.

performance can be attributed to the fact that the GMOK model captures the same underlying traits of the data generating process as the HMM model, albeit in a different fashion. Additionally, both the GMOK model and HMM outperform the simpler logit model and classification tree, illustrating the performance improvements that can be achieved by using these models over more traditional models. Given the absence of usage data for the HMM, we can interpret this result as a lower bound on its performance. The availability of such information could further improve HMM performance. However, given our interest in the case where past data is unavailable and the HMM cannot be estimated (i.e. the cross-sectional world), our interest does not lie in the question which of these two models performs best in absolute terms.

Finally, we note that to identify churners, a model should account for both dynamics and heterogeneity, as reflected in the significantly higher top decile lifts of the GMOK model and HMM compared to models that account for part or neither of these traits. Accounting for either dynamics or heterogeneity alone is not sufficient and even leads to poorer predictions than those generated by the logit model and the classification tree. In terms of overall classification, as measured by the Gini coefficient, the GMOK model is significantly more effective than all benchmark models in most periods. Thus, accounting for dynamics and unobserved heterogeneity not only significantly increases a model's capability to identify churners but also helps identify nonchurners, thereby increasing the overall model performance.

Figure 4: Average Gini Coefficients for Panel Data Models Estimated at Time *t* (95% Bootstrap Confidence Intervals, Insurance Data)



6.4 Comparison of Model Predictions: Cross-Sectional World

While the panel data world presents a scenario where full data from the past is available, the cross-sectional world presents the more interesting scenario where no past data is available. As such, the HMM model cannot be estimated, and we will compare the GMOK model to the remaining benchmarks to ascertain its performance.

To compare the performance of GMOK model with the benchmark models in the cross-sectional world, we provide plots of the out-of-sample metrics assessing predictive performance for all possible time periods. The results for the top decile lift are in Figure 5 and those for the Gini coefficient are in Figure 6. Note that in these graphs, the prediction at time *t* now constitutes

an out-of-sample fit as opposed to the panel data world, where this prediction is an in-sample fit. This emerges from the fact that in the cross-sectional world we use a separate holdout sample to generate predictions. In the panel data world, the same customers that were used for model estimation are used to generate the prediction at time *t*. For the other time periods in these graphs, interpretation remains the same as before.

In terms of top decile lift, the GMOK model performs significantly better than all other models, followed in order by the logit model, the classification tree, the model with only heterogeneity, and the model with only dynamics. In terms of the Gini coefficient, a significant difference between GMOK and the model with only heterogeneity exists until predictions for seven periods ahead, and both these models performed significantly better than the logit model, the classification tree, and the model with dynamics, in order. With respect to the best-performing benchmark model, the average improvement exceeds 11% for the top decile lift, and the relative average improvement for the Gini coefficient exceeds 36%. Hence, we find that the GMOK model continues to outperform the logit model and classification tree, in a setting where these models have been traditionally developed and applied (e.g. Neslin et al. 2006; Lemmens and Croux 2006; Risselada, Verhoef and Bijmolt 2010). This further corroborates our findings of the panel data world, illustrating the performance improvements of the GMOK model over simpler benchmarks.

We also note that the predictive performance as measured by both metrics decreases after $t + 1$ for most models, because predictions further in the future suffer from a greater amount of noise. In addition to aiming to predict further ahead, in this data set, we were also confronted with the policy change in 2006, which led to additional noise in long-term predictions. However, for the GMOK model, this decrease in predictive performance is much smaller, yielding more

stable predictions over time for both metrics. Thus, in the cross-sectional world we find evidence of the good staying of the GMOK model. In contrast, in the panel data world, we find that model staying power is lower (for top decile lift) or seems absent (for Gini coefficient). In addition, the size of top decile lift and Gini coefficient is also lower in the panel data world than in the cross-sectional world we consider here. The reduced performance and staying power in the panel data world can both be attributed to the loss of information over time as more and more customers churn, leaving only older and more loyal customers in the panel. This reduces the heterogeneity in the sample, limiting the ability of the GMOK model to adapt to the data. In the cross-sectional world this limitation does not occur, as new customers can continuously be part of the estimation sample, increasing the heterogeneity in the data. Hence, using cross-sectional data for model estimation has advantages above and beyond those related to privacy, as performance of the GMOK model is increased compared to benchmarks

Figure 5: Average Holdout Top Decile Lifts for Cross-Sectional Models Estimated at Time *t* (95% Bootstrap Confidence Intervals, Insurance Data)

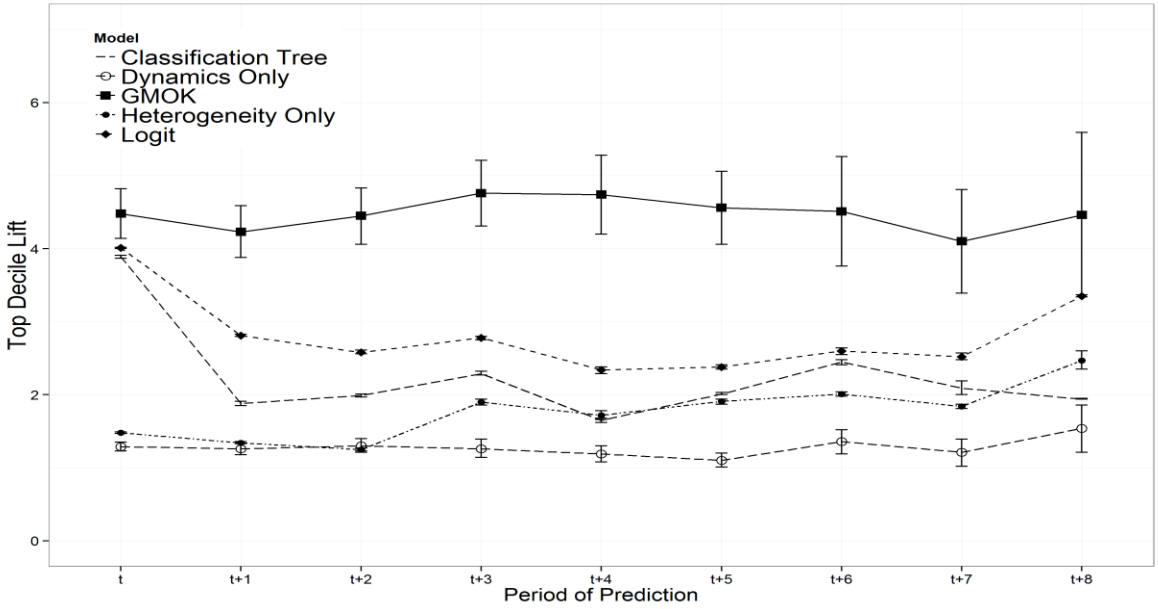
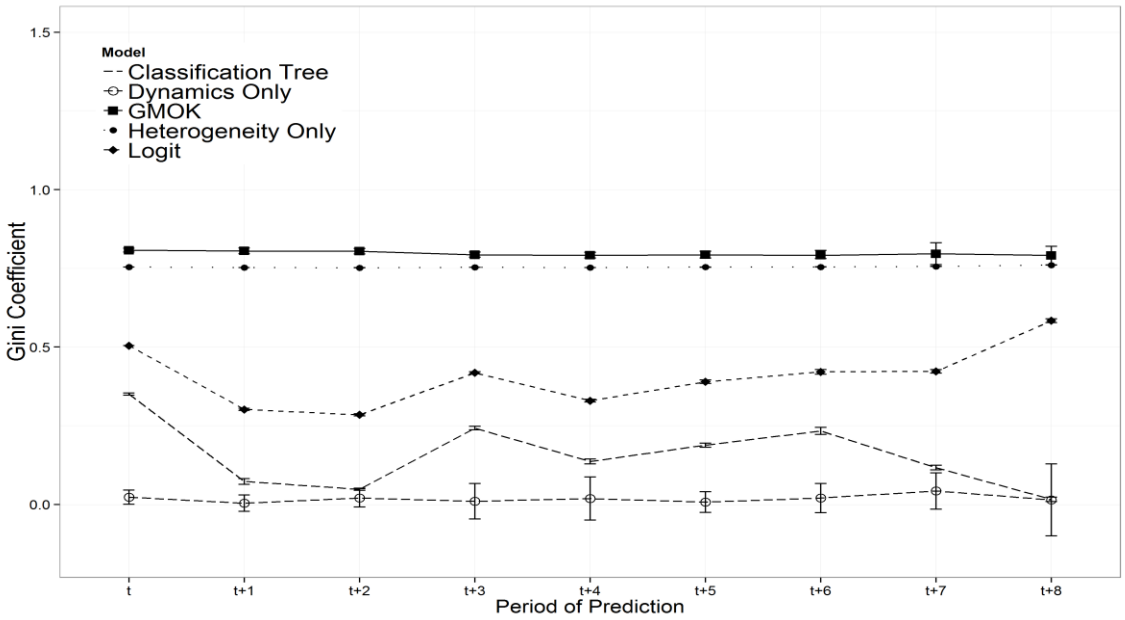


Figure 6: Average Holdout Gini Coefficients for Cross-Sectional Models Estimated at Time *t* (95% Bootstrap Confidence Intervals, Insurance Data).



To replicate our findings of the cross-sectional world in a different setting, we repeat our analysis on a second data set provided by a large European Internet service provider (ISP). These data cover three quarters from January–September 2006 and pertain to a specific Internet service. In this setting, a customer churns if he or she had a subscription to the Internet service at the beginning of the quarter but not at the end of the quarter. We use this second data set to validate our findings along two dimensions: time scale (quarterly vs. yearly) and industry (Internet vs. insurance). Our purpose is to provide more generalizable results about our model performance.

The churn rates in this data set are relatively stable over time: 1.8% for the first quarter, and 1.3% for the second and third quarters (see Table 5). Table 6 provides more details on the variables in this data set. Note that due to a lack of an individual customer identifier, we are not able to estimate the HMM for this data set as we cannot create the required panel data set. We

can estimate our GMOK model however, as this model only requires repeated cross-sectional data without the need for the same customer to be present in multiple cross-sections.

Period	Sample Size	Churn Percentage	Training Sample Size	Validation Sample Size
Q1-2006	233,780	1.8	5,000	5,501
Q2-2006	243,199	1.3	5,000	7,562
Q3-2006	246,335	1.3	5,000	7,447

Table 5: Total Sample Size, Churn Percentage, Training Sample Size and Validation Sample Size per Quarter for the ISP Data

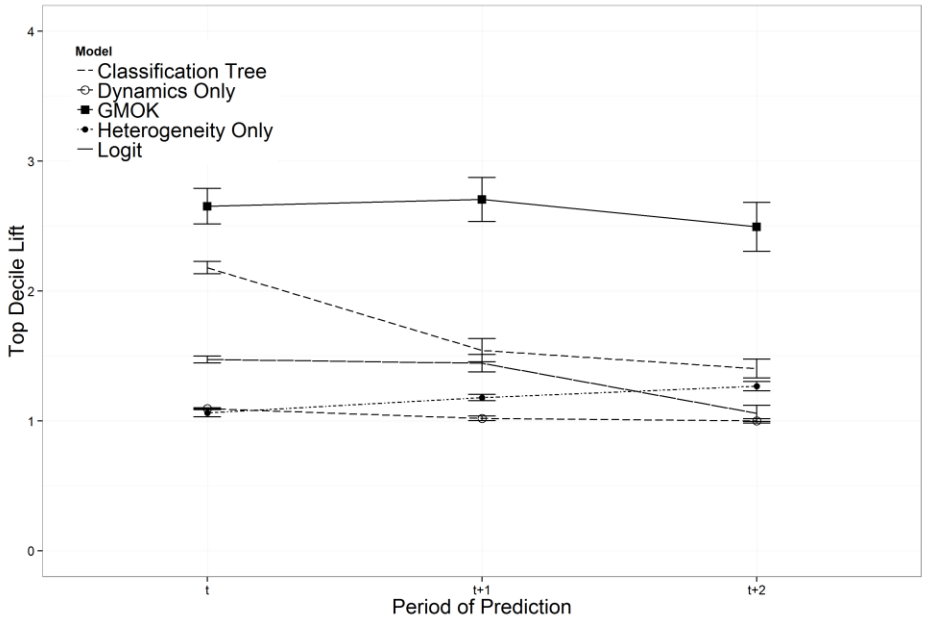
	Variables ISP Data
Sociodemographic	Age, Household size, Moved house
Socioeconomic	# Cars, Education level, Income, Employment status
Relationship characteristics	Relation ship age firm, Relationship age ISP, Revenue fixed line
Product details	Carrier preselect, Connection speed, Fixed line subscription type, value added services

Table 6: Available Variables for the ISP Data

For this data set, the results for the top decile lift are in Figure 7 and those for the Gini coefficient are in Figure 8. We confirm our main finding that the GMOK model significantly outperforms the other models in terms of both top decile lift and Gini coefficient in the cross-sectional world. The differences between the benchmark models are mostly small or insignificant. Compared with the best performing benchmark model for each metric, the improvements of the GMOK model exceed 20% for the top decile lift and 21% for the Gini coefficient.

As in the case of the insurance data set, the predictive performance for periods after *t + 1* is much lower. However, the GMOK model appears less affected by increased noise than the benchmark models: we find that the GMOK model has a greater staying power than the benchmark models, similar to our finding in the insurance data set. If we compare the results of the insurance data set with those from the ISP data set, we conclude that the GMOK model performed better in both industry settings. In addition, our findings are invariant to time scale, as we found no difference between the quarterly level ISP data and yearly level insurance data. Finally, they are independent of the patterns of churn rates.

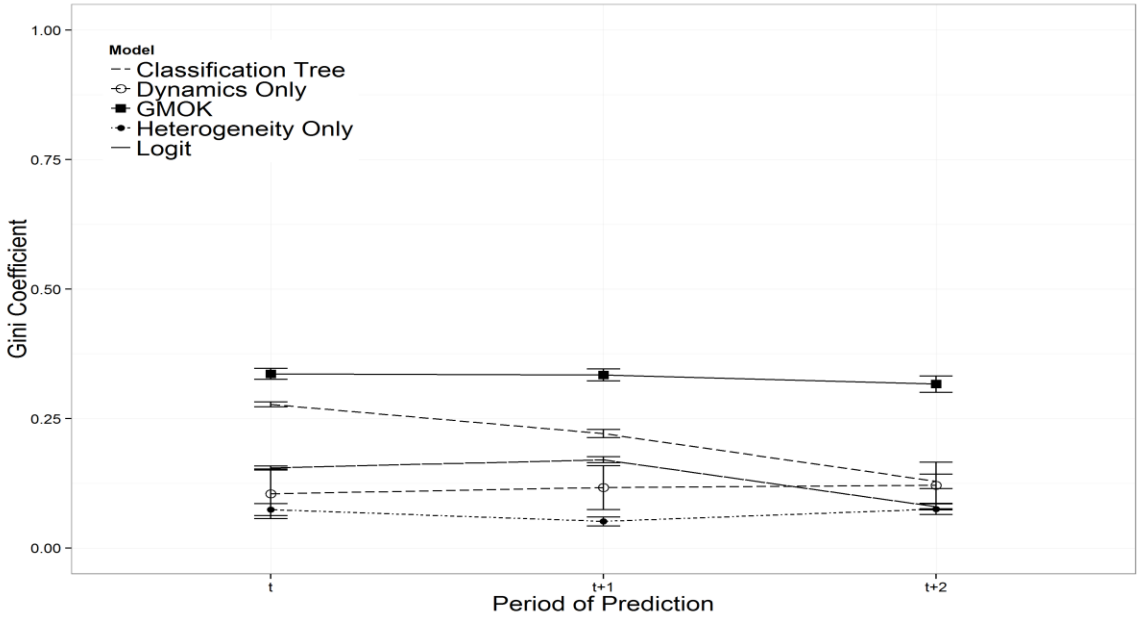
Figure 7: Average Holdout Top Decile Lifts for Models Estimated at Time *t* (95% Bootstrap Confidence Intervals, ISP Data)



Concluding, we find that in the panel data world where full past data is available the GMOK model performs equally well compared to the HMM, and both outperform the simpler benchmark models (notably the logistic regression and classification tree). This is achieved even

though the GMOK model was not specifically developed with panel data in mind and does not rely on storage of data from prior periods as the HMM does. When we consider the results of the cross-sectional world where data from the past is unavailable, we confirm the increased performance of the GMOK model compared to the simpler models in both the insurance industry and Internet service provider industry. In addition, when estimated on cross-sectional data, the GMOK model shows increased performance in top decile lift and Gini coefficient compared to the panel data world, and benefits from increased staying power. The latter finding implies that model performance deteriorates less quickly when the model is used for prolonged periods of time.

Figure 8: Average Holdout Gini Coefficients for Models Estimated at Time *t* (95% Bootstrap Confidence Intervals, ISP data)



7. DISCUSSION

Effective churn management plays an essential role in customer-centric firms. At the heart of many churn management programs lies the identification of customers with a high churn risk. Methods that use a probabilistic approach to identify such customers are widely used in practice, and their sophistication has increased significantly over the years. Most recently, Ascarza and Hardie (2013) show that using customer panel data to model the dynamic evolution of a customer base can greatly benefit churn predictions. However, in the face of legislative restrictions and public pressure to limit the storage of large amounts of customer data, firms actively limit the data they store, hampering their ability to perform advanced inference (e.g. Blatterg, Kim and Neslin 2008 p. 78; Wedel and Kannan 2016; Verhoef, Kooge and Walk 2016). In this article, we show that despite such data storage restrictions, churn prediction with the same accuracy as that of recently developed methods is still possible. In particular, we show that by applying a new method that captures the dynamics and unobserved customer heterogeneity of the underlying data generating process, improvements in churn predictions in excess of 9% are possible in comparison with existing methods (i.e. logistic regression and classification trees). This performance is similar to that of the HMM as proposed by Ascarza and Hardie (2013). Our GMOK model extends the logistic regression model with time-varying parameters to account for variation in model response parameters, and applies a mixture model approach to allow for different segments that show a different dynamic evolution of their response parameters.

The GMOK model estimated on cross-sectional data (when past data is unavailable) also has the benefit of longer staying power. Whereas commonly used methods exhibit 20% average drops in predictive performance after being in use for two periods (Risselada, Verhoef, and Bijmolt 2010), this average decline is only 1-3% for the GMOK model. Thus, the same model

can be used for a longer period of time without reestimation, resulting in costs savings in terms of data preparation and model estimation. However, if the GMOK model is estimated using panel data (i.e. when past data is available) the benefits of additional staying power are absent, and churn predictions further into the future become less accurate. In contrast, the HMM seems to be very stable in the panel data world, indicating a potential improved staying power of this model. Such an improvement could arise from the explicit incorporation of the panel structure of the data this model has. However, due to the limited amount of periods available for this model, we cannot investigate the long-term staying power of this model.

Additionally, we empirically establish that a model that accounts for either dynamics or unobserved heterogeneity, but not both (i.e. the restricted GMOK models in this study), does not perform better than the benchmark models. Hence, it is important that both effects be considered simultaneously for churn predictions to improve. This finding holds for models estimated on panel data as well as models estimated on cross-sectional data. Prior work in a different setting has empirically established similar results (e.g., Zhao, Zhao, and Song 2009). We attribute these findings to the following: As we have observed, accounting for dynamics only is not sufficient; heterogeneity is present in the data, according to a comparison of the parameter estimates of the GMOK model for different segments. These estimates differ in both size and sign for many variables (see also Figure 2). Accounting for heterogeneity only is also insufficient; prior research has shown that parameter estimates are not stable over time and that accounting for dynamics is necessary (Risselada, Verhoef, and Bijmolt 2010), as we confirm here. Combined, these results seem to confirm prior findings that accounting for both dynamics and unobserved heterogeneity is a necessary requirement for models to deliver improved performance.

Our results also show that the GMOK model provides equal predictive performance to Ascarza and Hardie's (2013) approach. We attribute these findings to the combination of dynamics and unobserved heterogeneity both models have in common. Indeed, HMMs as applied by Ascarza and Hardie (2013) are a special case of the general state-space model on which our method is based. The difference between our model and HMMs is that HMMs have a discretized rather than continuous state space (Roweis and Ghahramani, 1999). Both models show better performance than commonly used logit and classification tree models, illustrating the value of models that account for multiple sources of dynamics and heterogeneity. Concluding, we find that privacy conservation does not need to come at the cost of analytical capabilities, as indicated by the equal performance of the GMOK model and the HMM when applied under the same circumstances (i.e. the panel data world), and the improved performance of the GMOK model when past data is unavailable (i.e. the cross-sectional world).

8. LIMITATIONS AND FURTHER RESEARCH

Our research has some limitations that could not be addressed within the scope of this study. First, although we used data from two different industries in which customer churn is a common phenomenon, extensions to other industries would be helpful to confirm and generalize our findings. In addition, our second data set from the telecommunications industry did not contain as many time periods as the insurance data set. Although we established the superior performance of our approach to simpler models even with this limitation, a data set with a longer time horizon could further strengthen this evidence.

Second, careful consideration of variables to include in a model is always important, especially so for our proposed approach. In our applications, we had a full set of relevant variables for each period, but in practice, missing data and the consideration of new variables for inclusion are common. Missing data for included variables can be addressed easily in the state-space setting of our approach by not performing the correction step in the Kalman filter (see Web Appendix A) for customers with missing data. However, adding new variables would require model reestimation rather than model updating. As our model requires a consistent model structure over the periods the model is used, adding new variables would necessitate model redevelopment. This scenario places more pressure on variable selection processes, to ensure the model offers the best long-term performance. Therefore, we consider this determination a promising topic for further research.

Whenever marketing responses are modeled without knowledge of the exact firm decision process regarding marketing actions, this lack of information can lead to endogeneity issues if marketing variables are included in the model (Rossi, 2014). This type of endogeneity should not be of concern here as we use non-marketing variables as explanatory variables in our models. However, there is the notion that depending on the explanatory variables used to generate churn propensities, the importance of these variables changes for subsequent model applications due to their usage in directing retention efforts (Boulding et al. 2005; Verhoef, Kooge and Walk 2016 p. 189). In particular, a variable could become more important over time as the firm uses this variable to target customers with higher churn propensities, or be of reduced importance if the firm is succesfull in lowering the churn propensity of customers that fullfil the variable criterion. To investigate whether this is the case, we analyzed the variables included in the logistic regression model from the panel data world. In Table 7 we present the number of

models (out of 50 bootstrap iterations) in which a certain variable was included. We use the logistic regression results, because this is the model used by the insurance firm to compute churn propensities. While we do find some changing variable importance across time, as indicated by a variables` inclusion in a model, this changing importance seems to be centered around 2006, the year of the health care policy change. Hence, there seems to be no evidence of endogeneity due to firm actions, but there is evidence of an exogeneous shock that drives variable importance. As Figure 2 illustrates, the GMOK model can deal with this shock quite effectively due to the inclusion of time-varying parameters. Other cross-sectional models (i.e. logistic regression, classification tree) are also robust to this shock if re-estimated. As 2006 is included in every HMM model estimated, this model also should be robust to the shock as the information is included in the model. Hence, there appears to be no evidence of strong endogeneity issues in relation to all considered models, and our considered models appear to be robust to the exogeneous shock in 2006 as well. Even if unaccounted for endogeneity remains, given the similar (regression-based) structure of our models, it should affect all our models in a similar fashion, and our results remain valid as we compare models in relative sense. Moreover, given the predictive focus of our methods, accounting for endogeneity could negatively affect model performance on the predictive metrics we use (Ebbes, Papies and Van Heerde 2011). Concluding, while we find no strong indications of endogeneity, even in the case of unaccounted for endogeneity our results remain valid.

Finally, when developing the model it also important to correctly determine the model structure in addition to selecting the right variables. For example, the number of segments is fixed over time in our approach, making the determination of this number an important decision. This selection can easily be done by estimating the model for multiple segments, and selecting

the model with lowest BIC. When data from prior periods is available (for example because they have not been discarded yet), this information can be incorporated as well to improve model fit, provided this in line with government legislation and firm policy. As an extension of our approach, further research could investigate how to dynamically adapt the number of segments, for example by applying the the fully Bayesian reversible jump Markov chain Monte Carlo approach of Bruce, Peters and Naik (2012).

Variable	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**
Age	32	15	0	46	23	50	50	50	42
BSR Grouping	12	0	10	0	1	14	1	19	3
Distance Insurance Shop	0	2	0	13	5	6	18	5	4
Education Level	16	0	11	0	0	2	0	3	0
Family Without Kids	0	0	50	0	1	1	0	50	50
Income	0	0	1	0	0	0	0	0	0
# Insurance Shop Visits	3	0	45	0	0	32	24	1	0
Prosperity Level	0	0	47	0	0	0	0	0	0
Relationship Age	31	32	13	4	1	13	0	0	3
Social Class	3	0	40	0	0	0	0	0	0
Total # Models	50	50	50	50	50	50	50	50	50

Table 7: Number of Logistic Regression Models that Include Variable (Per Year)

References

Acquisti A, Friedman A, Telang R (2006) Is there a cost to privacy breaches? An event study. *ICIS 2006 Proceedings* paper 94,

Ascarza E, Hardie BGS (2013) A joint model of usage and churn in contractual settings. *Marketing Science.* 32(4):570–590.

Ascarza E, Iyengar R, Schleicher R (2016) The perils of proactive churn prevention using plan recommendations: Evidence from a field experiment. *Journal of Marketing Research* 53(1): 46-60.

Bart Y, Shankar V, Sultan F, Urban GL (2005) Are the drivers and role of online trust the same for all web sites and consumers? A large scale exploratory empirical study. *Journal of Marketing* 69: 133-152,

Blattberg RC, Kim B, Neslin SA (2008) *Database Marketing: Analyzing and Managing Customers* (Springer Science+Business Media, New York).

Boulding W, Staelin R, Ehret M, Johnston WJ (2005) A customer relationship management roadmap: What is known, potential pitfalls and where to go. *Journal of Marketing* 69: 155-166.

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) *Classification and Regression Trees* (Wadsworth, Belmont, CA).

Bruce, NI, Peters K, Naik PA (2012) Discovering how advertising grows sales and builds brands. *Journal of Marketing Research* 49: 793-806.

Cain PM (2005) Modeling and forecasting brand share: A dynamic system approach. *International Journal of Research in Marketing* 22: 203-220.

Calabrese A, Paninski L (2011) Kalman filter mixture model for spike sorting of non-stationary data. *Journal of Neuroscience Methods* 196:159–69.

CJEU C-131/12 (2014) Google Spain SL, Google Inc. v. Agencia Española de Protección de Datos (AEPD), Mario Costeja González, Retrieved from http://curia.europa.eu/juris/liste.jsf?num=C-131/12 [Accessed November 27 2014].

Culnan MJ, Williams CC (2009) How ethics can enhance organizational privacy: Lessons from the Choicepoint and TJX data breaches. *MIS Quarterly* 33(4): 673-687.

Deighton J (2005) Privacy and customer management. *Customer Management* (MSI Conference Summary). Cambridge, MA: Marketing Science Institute: 17-19.

Dempster A, Laird N, Rubin D (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B* 39(1):1–38.

Dijksterhuis M, Velders S (2009) Predicting switching behavior in a market with low mobility: A case study, in *Developments in Market Research,* A.E. Bronner, ed. (Spaar en Hout, Haarlem), 167–80.

DMA (2015) *Data privacy: what the consumer really thinks*.

Donkers B, Franses PH, Verhoef PH (2003) Selective sampling for binary choice models. *Journal of Marketing Research* 40(4):492–97.

Douven R, Mot E, Pomp M (2007) Health care reform in the Netherlands. *Die Volkswirtschaft* 3:31–33.

Durbin J, Koopman SJ (2012) *Time Series Analysis by State Space Methods* (2nd ed.) (Oxford University Press, Oxford).

Ebbes P, Papies D, Van Heerde HJ (2011) The sense and non-sense of holdout sample validation in the presence of endogeneity. *Marketing Science* 30(6): 1115-1122.

European Parliament (2013) Report on the proposal for a regulation of the European Parliament and of the Council on the Protection of Individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). Retrieved from http://www.europarl.europa.eu [Accessed May 2 2014].

Fahrmeir L, Tutz G (1994) *Multivariate Statistical Modelling Based on Generalized Linear Models* (Springer-Verlag, New York).

Fahrmeir L, Wagenpfeil S (1995) Penalized likelihood estimation and iterative Kalman smoothing for non-Gaussian dynamic regression models. *Sonderforschungsbereich* 386, Paper 5.

Forbes (2011) Bringing 20/20 foresight to marketing: CMOs seek a clearer picture of the customer. *Forbes Insights*, 1–13.

Ganesh J, Arnold MJ, Reynolds KE (2000) Understanding the customer base of service providers: An examination of the differences between switchers and stayers. *Journal of Marketing* 65:65–87.

Gupta S, Lehmann DR, Stuart JA (2004) Valuing customers. *Journal of Marketing Research* 41(1):7–18.

Jerath K, Fader P, Hardie B (2016) Customer-base analysis using repeated cross-sectional summary (RCSS) data. *European Journal of Operational Research* 249(1): 340-350.

Lemmens A, Croux C (2006) Bagging and boosting classification trees to predict churn. *Journal of Marketing Research* 43(2):276–86.

Lemmens A, Gupta S (2013) Managing churn to maximize profits. working paper, available at http://hbswk.hbs.edu/item/7350.html.

Lewis M (2005) Incorporating strategic consumer behavior into customer valuation. *Journal of Marketing* 69: 230-238.

Lu J (2002) Predicting customer churn in the telecommunications industry: An application of survival analysis modeling using SAS. *SAS User Group International (SUGI27) Online Proceedings*, Paper No. 114-27.

Marketing Science Institute (2016) *2016-2018 Research Priorities*. Cambridge, MA.

Musalem A, Bradlow ET, Raju JS (2009) Bayesian estimation of random-coefficients choice models using aggregate data. *Journal of Applied Econometrics* 24: 490-516.

Naik PA, Mantrala MK, Sawyer AG (1998) Planning media schedules in the presence of dynamic advertising quality. *Marketing Science* 17(3):214–235.

Neslin SA, Gupta S, Kamakura W, Junxiang L, Mason CH (2006) Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research* 43(2):204–211.

Osinga EC, Leeflang PSH, Wieringa JE (2010) Early marketing matters: A time-varying parameter approach to persistence modeling. *Journal of Marketing Research* 47(1):173–185.

Otto PN, Antón AI, Baumer DL (2007) The Choicepoint dilemma: How data brokers should handle the privacy of personal information. *IEEE Security & Privacy* 5(5): 15-23.

PCAST: President's Council of Advisors on Science and Technology (2014) Big data and privacy: A technological perspective. Executive Office of the President Report.

Podesta J, Pritzker P, Moniz EJ, Holdren J, Zients J (2014) Big data: Seizing opportunities, preserving values, Report, Executive Office of the President.

Reimer R, Rutz OJ, Pauwels K (2014) How online consumer segments differ in long-term marketing effectiveness. *Journal of Interactive Marketing* 28:271-284

Risselada H, Verhoef PC, Bijmolt THA (2010) Staying power of churn prediction models. *Journal of Interactive Marketing* 24:198–208.

Rossi PE (2014) Even the rich can make themselves poor: A critical examination of IV methods in marketing applications. *Marketing Science* 33(5): 655-672.

Roweis S, Ghahramani Z (1999) A unifying review of linear Gaussian models. *Neural Computation* 11(2):347–374.

Rust R, Kannan PK, Peng N (2002) The customer economics of Internet privacy. *Journal of the Academy of Marketing Science* 30:455-464.

Rust R., Chung TS (2006) Marketing models of service and relationships. *Marketing Science* 25(6): 560-580.

Rust R, Huang MH (2014) The service revolution and the transformation of marketing science. *Marketing Science* 33(2): 206-221.

Schwarz G (1978) Estimating the dimension of a model. *Annals of Statistics* 6(2):461–464..

Verhoef PC, Kooge E, Walk N (2016) *Creating Value with Big Data Analytics*. Routledge

Vermunt JK, Magidson J (2013) *Technical Guide for Latent GOLD 5.0: Basic, Advanced, and Syntax* (Statistica Innovations Inc., Belmont).

Wedel M, Kamakura WA (1998) *Market Segmentation: Conceptual and Methodological Foundations* (Kluwer Academic Publishers, Dordrecht).

Wedel M, Kannan PK (2016) Marketing analytics for data rich environments. *Working paper*.

Yan L, Miller DJ, Mozer MC, Wolniewicz R (2001) Improving prediction of customer behavior in nonstationary environments. *Proceedings IJCNN 01 International Joint Conference on Neural Networks* 3:2258–2260.

Zhao Y, Zhao Y, Song I (2009) Predicting new customers' risk type in the credit card market. *Journal of Marketing Research* 46(4):506–517.

\

No Future Without the Past? Predicting Customer Churn with Limited Past Data

WEB APPENDIX

Appendix A: Recursions for the Modified Kalman filter

We provide the Kalman filter recursions we used to estimate the state-space model here. Recall that the state equation for individual *i* is $\pi_{it} = P(y_{it} = 1) = \exp(X_{it}\beta_t)\,/(1 + \exp(X_{it}\beta_t))$, and the transition equation $\beta_t = F_t\beta_{t-1} + \zeta_t$, where $\zeta_t \sim N(0, Q_t)$. We implicitly assumed that $F_t = I_t$ previously, but we generalize this assumption here. Furthermore, we assume in general that $E(Y_{it}|\beta_t) = \mu_{it} = g(\eta_{it})$, where $\eta_{it} = X_{it}\beta_t$ and $g(\cdot)$ is a known link function. This function is the logistic function in the case we consider in the article. The two relaxations serve to give a general overview of the recursions. We aim to obtain estimates for β_t for *t* = 1, …, *T*, given values for Q_t for *t* = 1,…, *T*, β_0 and Q_0. To obtain these estimates, Fahrmeir and Tutz (1994) suggest the following set of recursions:

Initialization:	$\beta_{0\|0} = \beta_0, V_{0\|0} = Q_0$
Prediction step : For *t* = 1,…,*T*:	$\beta_{t\|t-1} = F_t\beta_{t-1\|t-1};\ V_{t\|t-1} = F_tV_{t-1}F_t' + Q_t$
Correction step:	$\beta_{0,t} = \beta_{t\|t-1}; V_{0,t} = V_{t\|t-1}$
For *i* = 1, …, *n*:	$\beta_{i,t} = \beta_{i-1,t} + K_{it}(Y_{it} - \mu_{it})$
	$V_{it} = (I - K_{it}D_{it}'X_{it})V_{i-1,t}$
Kalman gain:	$K_{it} = V_{i-1,t}X_{it}'D_{it}\left[D_{it}'X_{it}V_{i-1,t}X_{it}'D_{it} + \Sigma_{it}\right]^{-1}$

Here, $D_{it} = \delta g/\delta\eta_{it}$, and μ_{it}, Σ_{it}, and D_{it} are evaluated at $\beta_{i-1,t}$. The final estimates are the parameter vector $\beta_{t|t}$ and its corresponding covariance matrix $V_{t|t}$. For the case of logistic regression, these quantities read as follows: $\mu_{it} = g(\eta_{it}) = \exp(\eta_{it})\,/(1 + \exp(\eta_{it}))$, $D_{it} = \exp(\eta_{it})\,/(1 + \exp(\eta_{it}))^2$ and $\Sigma_{it} = \mu_{it}(1 - \mu_{it})$. To obtain the initial values required (b_o, Q_o, and Q_t), we used maximum likelihood estimation in an iterative algorithm, inspired by Fahrmeir and Wagenpfeil (1995) to obtain the values given the data.

Appendix B: Derivation of Likelihoods for the Regular State Space and GMOK Models

For the model without unobserved heterogeneity, we derive the full posterior log-likelihood that the Kalman filter is aimed to maximize. Next, we adapt this likelihood further when we add unobserved heterogeneity to the model through a mixture model approach (Wedel and Kamakura, 1998).

The following derivation is adapted from Fahrmeir and Tutz (1994): For customer *i*, let $y_t = (y_{1t}, \dots, y_{nt})$ and $X_t = (X_{1t}, \dots, X_{nt})$ represent all individual observations for period *t* = 1, …, *T*, and then let $y_t^* = (y_1, \dots, y_t)$ and $X_t^* = (X_1, \dots, X_t)$ denote the observations until time *t*.[9] Consider the estimation of $\beta_T^* = (\beta_0, \dots, \beta_T)$, where we assume the hyperparameters b_0 and Q_0, and the covariance matrices Q_t to be given. Repeated application of Bayes's theorem yields the following expression for the posterior distribution of the state vector (Fahrmeir and Tutz 1994):

$$(B.1)\ p(\beta_T^*|y_T^*, X_T^*) \propto$$
$$\prod_{t=1}^{T} p(y_t|\beta_t^*, y_{t-1}^*, X_t^*)\ \prod_{t=1}^{T} p(\beta_t|\beta_{t-1}^*, y_{t-1}^*, X_t^*) \prod_{t=1}^{T} p(X_t|\beta_{t-1}^*, y_{t-1}^*, X_{t-1}^*),$$

where $p(\cdot)$ denotes a (conditional) density function. Next, we make the following (weak) conditional independence assumptions:

1. Conditional on β_t, y_{t-1}^* and X_t^*, current observations y_t are independent of β_t^*; that is,
   $$(B.2)\ p(y_t|\beta_t^*, y_{t-1}^*, X_t^*) = p(y_t|\beta_t, y_{t-1}^*, X_t^*).$$
2. Conditional on y_{t-1}^* and X_{t-1}^*, X_t is independent of β_{t-1}^*; that is,
   $$(B.3)\ p(X_t|\beta_{t-1}^*, y_{t-1}^*, X_{t-1}^*) = p(X_t|y_{t-1}^*, X_{t-1}^*).$$

[9] We suppress *i* in this first part for notational convenience

3. Conditional on β_t and y^*_{t-1},with X^*_t, the individual observation y_{it} within y_t are conditionally independent; that is,

   $(B.4)\ p(y_t|\beta_t, y^*_{t-1}, X^*_t) = \prod_{i=1}^{n} p(y_{it}|\beta_t, y^*_{t-1}, X^*_t)$.

4. The parameter process is Markovian; that is,

   $(B.5)\ p(\beta_t|\beta^*_{t-1}, y^*_{t-1}, X^*_t) = p(\beta_t|\beta_{t-1})$.

With these assumptions, we can write Equation (B.1) as

$$(B.6)\ \ p(\beta^*_T|y^*_T, X^*_T) \ \propto \prod_{t=1}^{T}\prod_{i=1}^{n} p(y_{it}|\beta_t, y^*_{t-1}, X^*_t)\prod_{t=1}^{T} p(\beta_t|\beta_{t-1})p(\beta_0).$$

When we take the logarithms and write out the conditional distributions, the maximization of this conditional density is equal to maximizing the (penalized) log-likelihood:

$$(B.7)\ \ell(\beta_T) = \sum_{k=1}^{T}\sum_{i=1}^{n} l_{ik}(\beta_k|y_{ik}, X_{it}) - \frac{1}{2}(\beta_0 - b_0)'Q_0^{-1}(\beta_0 - b_0)$$
$$-\frac{1}{2}\sum_{k=1}^{T}(\beta_k - \beta_{k-1})'Q_k^{-1}(\beta_k - \beta_{k-1}),$$

where l_{ik} is the logistic log-likelihood contribution of individual *i*, and b_0 and Q_0 are the initial values for the Kalman filter. When we estimate the model at a time *t*, given information known up until that point, we evaluate the log-likelihood

$$(B.8)\ \ell(\beta^*_t) = \sum_{k=1}^{t-1}\sum_{i=1}^{n} l_{ik}(\beta^*_k|y^*_{ik}, X^*_{ik}) - \frac{1}{2}(\beta_0 - b_0)'Q_0^{-1}(\beta_0 - b_0)$$
$$-\frac{1}{2}\sum_{k=1}^{t-1}(\beta_k - \beta_{k-1})'(Q_k)^{-1}(\beta_k - \beta_{k-1}) + l_{ik}(\beta_k|y_{ik}, X_{ik})$$
$$-\frac{1}{2}(\beta_t - \beta_{t-1})'Q_t^{-1}(\beta_t - \beta_{t-1}),$$

Likelihood for the GMOK model

To formulate the likelihood for the GMOK model, we introduce the latent, unobserved indicator variable z_{it}^{j}, which indicates the membership of customer *i* to cluster *j* at time *t*. Calabrese and Paninski (2011) derive an EM algorithm to estimate a mixture of Kalman filters model with z_{it}^{j} as missing data for the case in which the y_{it} are generated by a normal distribution. To arrive at the likelihood for the GMOK model, we take the likelihood derived by Calabrese and Paninski (2011) and combine it with the likelihood from Equation (B.6). To maximize this likelihood we then derive an EM algorithm following the work of these authors. The modified likelihood of Equation (B.6) is given as[10] :

$$(B.9)\ p\left(\beta_T^*|y_T^*, X_T^*, z_{it}^j\right) \propto \prod_{k=1}^{T}\prod_{i=1}^{n} p\left(y_{ik}, z_{ik}^j|\beta_k, y_{k-1}^*, X_k^*\right) \prod_{k=1}^{T} p(\beta_k|\beta_{k-1})p(\beta_0)$$

$$\propto \prod_{k=1}^{T} p(z_{ik}^j|k) \prod_{k=1}^{T}\prod_{i=1}^{n} p(y_{ik}|\beta_k, y_{k-1}^*, X_k^*) \prod_{k=1}^{T} p(\beta_k|\beta_{k-1})p(\beta_0)$$

If we assume that $p\left(z_{it}^j|\beta_t\right) = \lambda_t^j$, where λ_t^j represents a mixture weight, such that $\sum_j \lambda_t^j = 1$ for each time period and each $\lambda_t^j \geq 0$, after taking logarithms and expectations we arrive at

$$\begin{aligned}(B.10)\ E\left[l\left(\beta_t^{j*}\right)\right] &= \sum_{j=1}^{J}\sum_{k=1}^{t-1}\left(p_k^{j*}\left[\log\left(\lambda_k^j\right) - \sum_{i=1}^{n} l_{ik}\left(\beta_k^{j*}|y_{ik}^*, X_{ik}^*\right)\right]\right) \\ &- \frac{1}{2}\sum_{j=1}^{J}\left[\left(\beta_0^j - b_0^j\right)'\left(Q_0^j\right)^{-1}\left(\beta_0^j - b_0^j\right)\right] - \frac{1}{2}\sum_{j=1}^{J}\sum_{k=1}^{t-1}\left[\left(\beta_k^j - \beta_{k-1}^j\right)'\left(Q_k^j\right)^{-1}\left(\beta_k^j - \beta_{k-1}^j\right)\right] \\ &+ \sum_{j=1}^{J} p_t^j\left[\log\left(\lambda_t^j\right) - \sum_{i=1}^{n} l_{it}\left(\beta_t^j|y_{it}, X_{it}\right)\right] - \frac{1}{2}\sum_{j=1}^{J}\left[\left(\beta_t^j - \beta_{t-1}^j\right)'\left(Q_t^j\right)^{-1}\left(\beta_t^j - \beta_{t-1}^j\right)\right]\end{aligned}$$

[10] Calabrese and Paninski (2011), and the references therein, offer a full derivation of this log-likelihood.

where

$$(B.11)\quad p_t^j = \frac{\exp\left[\log\left(\lambda_t^j\right) + \sum_{i=1}^{n} l_{it}\left(\beta_t^j\right)\right]}{\sum_{j=1}^{J} \exp\left[\log\left(\lambda_t^j\right) + \sum_{i=1}^{n} l_{it}\left(\beta_t^j\right)\right]},$$

$\beta_t^{j*} = (\beta_0^j, \ldots, \beta_t^j)$ for $j = 1, \ldots, J$ and Equation (B.11) is evaluated using current values of the parameters. The M-step of the algorithm requires the maximization of Equation (B.10), which is simplified because the cross-derivatives are zero, so we can maximize the two parts of Equation B.10 separately. Maximizing the first part of the likelihood requires maximizing

$$(B.12)\quad \sum_{j=1}^{J}\sum_{k=1}^{T} p_k^j \log\left(\lambda_k^j\right),$$

subject to the constraints on λ_t^j. This yields

$$(B.13)\quad \widehat{\lambda_t^j} = \frac{\sum_{k=1}^{T} p_k^j}{T},$$

which is similar to the mixture model case (Wedel and Kamakura 1998). The remainder of this likelihood is similar to Equation (B.7), multiplied by p_k^j. Thus, we can apply the Kalman filter of Fahrmeir and Tutz (1994) to maximize this likelihood, provided we include p_k^jas additional weight in the Kalman gain part of the filter. The Kalman gain, as given in Web Appendix A, in that case reads

$$(B.14)\quad K_{it} = V_{i-1,t} X_{it}^{'} D_{it}\left[D_{it}^{'} X_{it} V_{i-1,t} X^{'}{}_{it} D_{it} + p_t^j \Sigma_{it}\right]^{-1}$$

for the filter applied to a given segment *j*. By iterating over these E- and M-steps, we obtain the final estimates of all parameters for all groups and use these estimates for our predictions.

Appendix C: Modification of the HMM of Ascarza and Hardie (2013)

In part D of their Web Appendix, Ascarza and Hardie (2013) derive a variant of their hidden Markov model for usage and churn in which usage is modeled using a binominal model instead of a Poisson model. Inspired by this change of state-dependent usage process, we modified the model using a Bernoulli state-dependent equation to model churn. Because we do not observe usage, we can only model churn given membership of a certain segment on the basis of the binary decision of each customer that we observe each year. Thus, we chose to model the individual-specific churn probability p_{it} at time t using a Bernoulli model, given the same unobserved commitment process outlined in Ascarza and Hardie (2013). We therefore only outline the changes we made with respect to their model, following the notation used in their article. For more background information, we refer to their article.

In the HMM, observed behavior is modeled conditional on a latent, unobserved state, which is time varying. Let S_{it} denote this state process. Given a certain state *k*, a customer *i* has a probability to churn p_{it}. We let this churn probability depend on segment specific and an individual specific component. Specifically, we assume that

$$p_{it}|[S_{it} = k] = \theta_k^{\Lambda^{-1}(\beta_{0i})} .$$

This churn probability consists of two parts: a segment-specific part, indexed by *k*, and an individual-specific part, indexed by *i*. We let θ_k denote the segment-specific commitment level, with the restriction that $0 < \theta_k < 1$ for all *k* and that $0 < \theta_1 < \theta_2 < \cdots < \theta_K < 1$; that is, θ_k is increasing with the commitment level. This specification is similar to that of Ascarza and Hardie (2013), with the exception that we do not assume an individual-specific movement process through these states. Instead, we assume the same transition matrix across customers, which is

identified through the variation in churn propensity across all customers (e.g. Zucchini and MacDonald 2009). We exponentiate this commitment level with an individual-specific parameter β_{0i} following a lognormal distribution with mean βX_i and standard deviation $\sigma_{b0}$to allow for individual specific effects. The variables X_i represent the characteristics of customer *i* a the time of model estimation[11], and help with the identification of the individual specific effect (see also Datta, Foubert and Van Heerde 2015). By including this term as an exponent and by applying an inverse logistic transformation Λ^{-1}, we ensure that p_{it} remains interpretable as a probability. The likelihood for the model then becomes, in the notation of Ascarza and Hardie (2013),

$$L_i(\theta, \beta_{0i}, \beta_k | S_i = k, data) = \prod_{t=i}^{T_i} (\theta_k^{\Lambda^{-1}(\beta_{0i})})^{y_{it}} \ (1 - \theta_k^{\Lambda^{-1}(\beta_{0i})})^{1-y_{it}},$$

where y_{it} is the binary indicator indicated previously.

[11] As the variables are varying over time, we use the observations at the time of model estimation to fix them at a customer-specific value. Alternatively, the mean or mode across all time periods could be used.

Appendix D: Panel World Models for the Period 2010-2012

To compare the models on similar samples, we also present the results for all models estimated using data from 2010-2012 only. Figure D.1 and D.2 show that in terms of top decile lift, there is no significant difference between the GMOK model and the HMM in all periods but period *t*, whereas the main results show no significant differences until after period *t* + 2. This implies that the HMM has similar predictive power compared to the GMOK model. Second, the logistic regression model does not perform significantly worse in period *t* +1 as shown in the main results, but does so only after this period. Figure D.3 shows that in terms of Gini coefficient, the GMOK model, HMM and heterogeneity only model do not differ significantly in any period, indicating a worse performance of the heterogeneity only model compared to the main results.

Figure D.1: Average Top Decile Lifts for Panel Data Models Estimated at Time *t* (95% Bootstrap Confidence Intervals, Insurance Data)

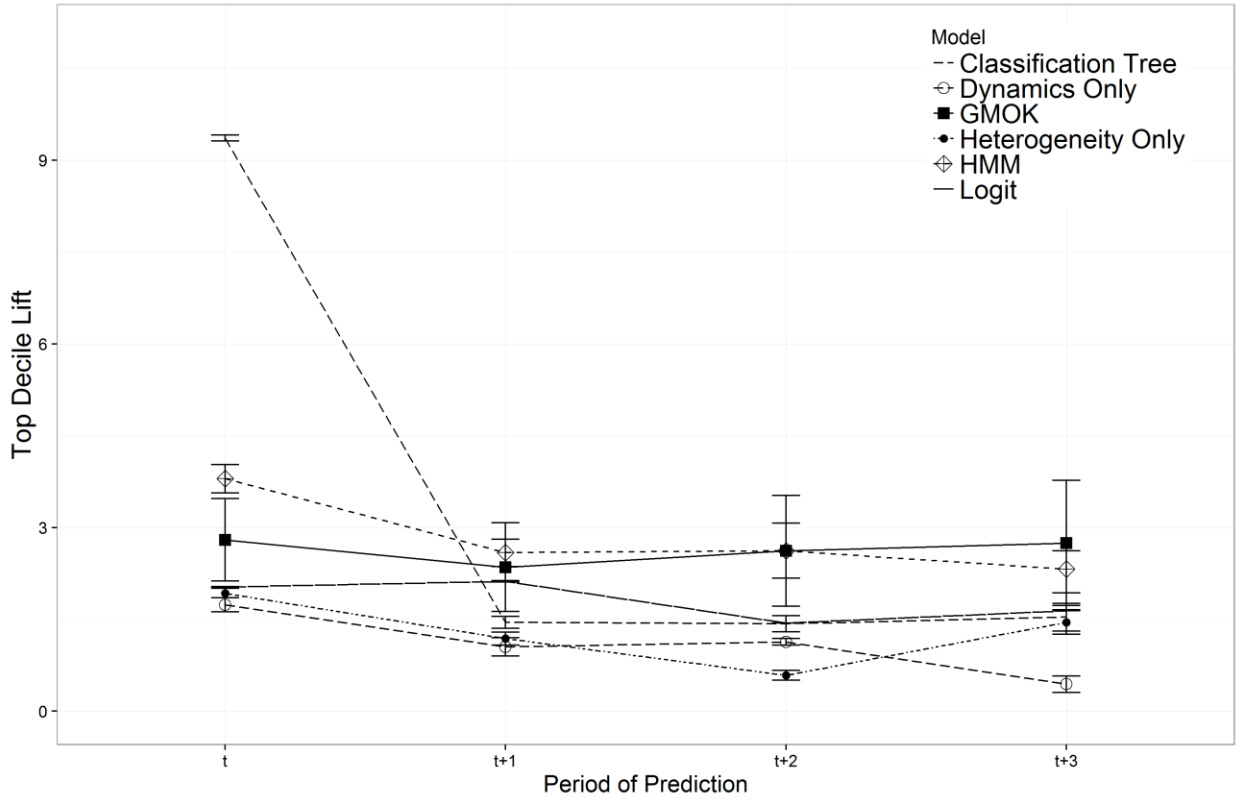


Figure D.2: Blowout of Figure D.1. for Top Decile Lifts Between 0 and 5

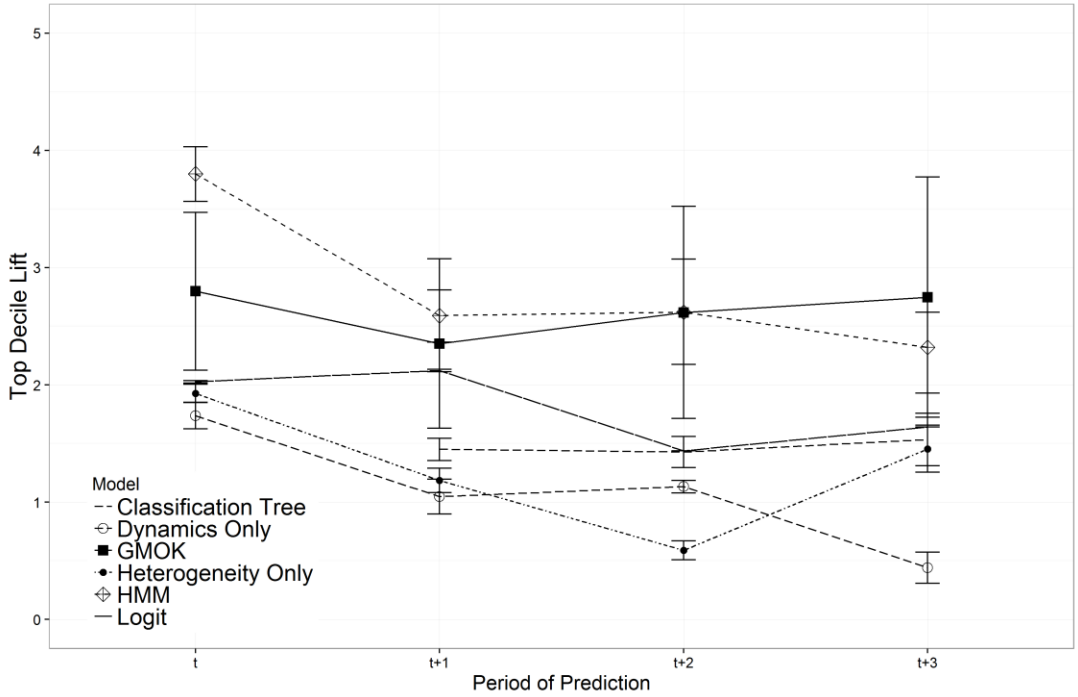
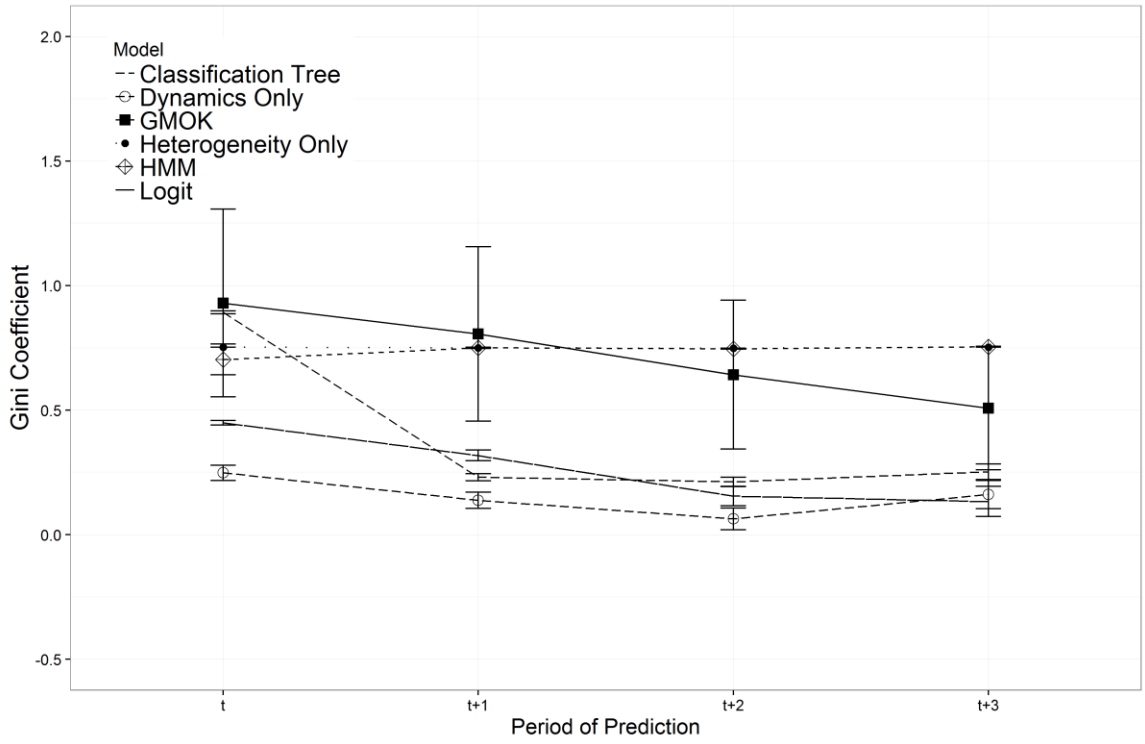


Figure D.3: Average Gini Coefficients for Panel Data Models Estimated at Time *t* (95% Bootstrap Confidence Intervals, Insurance Data)



References

Ascarza E, Hardie BGS (2013) A joint model of usage and churn in contractual settings. *Marketing Science.* 32(4):570–590.

Calabrese A, Paninski L (2011) Kalman filter mixture model for spike sorting of non-stationary data. *Journal of Neuroscience Methods* 196:159–169.

Datta H, Foubert B , Van Heerde HJ (2015) The challenge of retaining customers acquired with free trials. *Journal of Marketing Research* 52: 217-234.

Fahrmeir L, Tutz G (1994) *Multivariate Statistical Modelling Based on Generalized Linear Models* (Springer-Verlag, New York).

Fahrmeir L, Wagenpfeil S (1995) Penalized likelihood estimation and iterative Kalman smoothing for non-Gaussian dynamic regression models. *Sonderforschungsbereich* 386, Paper 5.

Zucchini W, MacDonald IL (2009) Hidden markov models for time series: An introduction using R. Chapman and Hall/CRC, Boca Raton (FL).

Wedel M, Kamakura WA(1998) *Market Segmentation: Conceptual and Methodological Foundations* (Kluwer Academic Publishers, Dordrecht).