Expert Systems with Applications 41 (2014) 7889–7903

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Dynamic churn prediction framework with more effective use of rare event data: The case of private banking

Özden Gür Ali [a,*], Umut Arıtürk [b]

[a] College of Administrative Sciences and Economics, Business Administration, Koç University, Sarıyer, Istanbul, Turkey
[b] Department of Industrial Engineering, Koç University, Sarıyer, Istanbul, Turkey

ARTICLE INFO

Article history:
Available online 19 June 2014

Keywords:
Dynamic churn prediction
Data mining
Customer retention
Private banking
Customer relationship management
Rare event
Sampling
Training data generation

ABSTRACT

Customer churn prediction literature has been limited to modeling churn in the next (feasible) time period. On the other hand, lead time specific churn predictions can help businesses to allocate retention efforts across time, as well as customers, and identify early triggers and indicators of customer churn. We propose a dynamic churn prediction framework for generating training data from customer records, and leverage it for predicting customer churn within multiple horizons using standard classifiers. Further, we empirically evaluate the proposed approach in a case study about private banking customers in a European bank.

The proposed framework includes customer observations from different time periods, and thus addresses the absolute rarity issue that is relevant for the most valuable customer segment of many companies. It also increases the sampling density in the training data and allows the models to generalize across behaviors in different time periods while incorporating the impact of the environmental drivers.

As a result, this framework significantly increases the prediction accuracy across prediction horizons compared to the standard approach of one observation per customer; even when the standard approach is modified with oversampling to balance the data, or lags of customer behavior features are added as additional predictors.

The proposed approach to dynamic churn prediction involves a set of independently trained horizon-specific binary classifiers that use the proposed dataset generation framework. In the absence of predictive dynamic churn models, we had to benchmark survival analysis which is used predominantly as a descriptive tool. The proposed method outperforms survival analysis in terms of predictive accuracy for all lead times, with a much lower variability. Further, unlike Cox regression, it provides horizon specific ranking of customers in terms of churn probability which allows allocation of retention efforts across customers and time periods.

1. Introduction

Customer retention is an important element of the customer relationship management (CRM) literature e.g. (Bijmolt et al., 2010). Several studies show the economic value associated with customer retention: The costs arising from acquiring a new customer surpass the expenditure to retain an existing one (Dawes & Swailes, 1999); long-term customers buy more and bring in new customers (Reichheld, 1996); for a bank, the longer the customer relationship, the higher the customer's worth (Reichheld & Kenny, 1990).

Accurate churn prediction helps the company to target the retention effort on the right customer at the right time and thereby convince valuable customers to stay. Further, churn probability is an important input to the calculation of the customer lifetime value (CLV) metric.

While customer specific churn rate prediction is well accepted in the literature, the typical approach does not consider its variation over time, i.e., lacks the dynamic aspect. On the other hand, dynamic churn prediction would allow the business to allocate retention and service improvement resources across customers and time (Venkatesan & Kumar, 2004). From the customer relationship management point of view, managers need insights into the reasons and timing of churn to create retention strategies and diagnose how much of the attrition can be controlled (Braun & Schweidel, 2011). Further, earlier detection of would-be-churners is valuable as it provides more time to the company to convince those customers to stay before they make their final decision (Verbeke, Dejaeger, Martens, Hur, & Baesens, 2012). Finally,

* Corresponding author. Tel.: +90 212 3381450; fax: +90 212 3381653.
E-mail address: oali@ku.edu.tr (Ö. Gür Ali).

http://dx.doi.org/10.1016/j.eswa.2014.06.018
0957-4174/

dynamic churn predictions are important for evaluating the aggregate value of the firm's customer base and hence its financial health (Fader & Hardie, 2010).

Recently, researchers have started investigating the impact of customer data preprocessing and representation of longitudinal customer data on the predictive accuracy and identified the need for dynamic churn prediction; i.e., churn probability in time (Crone, Lessmann, & Stahlbock, 2006; Lee, Wei, Cheng, & Yang, 2012). Typically, the churn training data consists of the churn response in one particular time period, while customers who have churned earlier are discarded; and the time series nature of the customer behavior is aggregated away in favor of simplicity and interpretability (Cao, 2010; Lee et al., 2012). The recent studies which focus on how to incorporate the-longitudinal data about the customer's historical behavior as predictors still use one outcome observation per customer (Crone et al., 2006; Lee et al., 2012; Orsenigo & Vercellis, 2010; Prinzie & Van den Poel, 2006). Such training data (a) throws away valuable data regarding customers who churned earlier, and (b) does not incorporate variation in the economic and competitive environment (c) does not facilitate dynamic customer churn prediction, which would allow the company to target retention programs better by selecting the customers as well as the time period.

In this paper, we propose a framework for generating training data that facilitates dynamic churn prediction, by generating multiple observations per customer from different time periods and hence including customers who churned earlier. By including the previously churned customers in the training data, this framework also alleviates the rare event problem associated with churn prediction in many organizations (Burez & Van den Poel, 2009), which poses difficulties for learning classification models (Weiss, 2004). Further, it incorporates the impact of the environmental drivers affecting customer churn; for example, customers may reevaluate the value proposition of their service provider in changing economic conditions and switch to competing institutions whose offerings are more suitable to their changing needs. Training datasets that contain churn response of only one time period do not contain variation in the environmental variables, and cannot provide insights about their impact on churn. Thus, their predictive power is expected to be limited in dynamic business environments. We apply and evaluate the proposed training data generation scheme and dynamic churn prediction approach against current standard approaches on the private banking data from a European bank. Private banking customers are high worth, high status individuals (Lassar, Manolis, & Winsor, 2000). As such, they constitute a very valuable customer segment that is sought after by competing banks and financial institutions where retention of even single individual can have significant impact on the company profitability. We empirically show that regardless of the classification method the proposed framework significantly improves the predictive accuracy of the resulting churn models, enables identification of the impact of the environmental factors, and provides additional insights about customer churn drivers compared with the standard approach even when the data is balanced by oversampling, and even when additional lags of customer behavior features are included. Further, the proposed dynamic churn prediction method - independent binary classifiers – outperforms the current standard approach, survival analysis (Cox regression) for all prediction horizons, and provides lead-time specific ranking of customers in terms of churn probability.

To summarize, our contributions are as follows: First, we introduce a dynamic churn prediction for generating training data from customer records and leverage it for predicting customer churn within a specified horizon with standard classifiers. To our knowledge, this is the first paper to perform dynamic churn prediction. Second, we provide a comparison of the accuracy impact of using single versus multiple observations per customer for churn prediction, both by identifying the theoretical drivers, as well as by empirically evaluating in the private banking customer churn context.

The remainder of this paper is organized as follows: In Section 2 we review the relevant literature for predictive churn modeling, representation of longitudinal customer data and environmental variables, and the rare event problem in churn prediction and associated remedial sampling approaches. The following section describes the proposed dynamic churn prediction framework and the proposed independently trained binary classifiers approach, and provides theoretical justifications. Section 4 describes the private banking application and the experimental design and presents the empirical evaluation results. We conclude with a summary of the results, managerial implications and limitations of the research, and offer future research directions.

2. Relevant literature

There is substantial body of work on churn prediction models. The initial focus of the literature was comparison of classification methods in terms of prediction accuracy. Table 1 provides a sample of papers since 2004. Logistic regression, decision trees, neural networks, support vector machines and survival analysis are the most popular methods e.g. (Buckinx & Van den Poel, 2005; Coussement & Van den Poel, 2008; Karahoca & Karahoca, 2011; Neslin, Gupta, Kamakura, Junxiang, & Mason, 2006). While SVM and decision trees have been used for predictive purposes, survival analysis is used descriptively (Burez & Van den Poel, 2008). Many of the later applications use hybrid models or ensembles that integrate multiple classifiers and/or develop variants of the existing algorithms (Buckinx & Van den Poel, 2005; Chu, Tsai, & Ho, 2007; De Bock & Poel, 2011) which improve prediction accuracy while decreasing interpretability (Verbeke, Martens, Mues, & Baesens, 2011). In their large comparative study Verbeke et al., find that accuracy performance of many classifiers are comparable since different classifiers yield better performances in different contexts and datasets (Verbeke et al., 2012).

Recently, researchers have started investigating the impact of data, preprocessing, sampling and representation of customer data on predictive accuracy independent of the classification algorithm. For example, Crone et al., have shown that attribute scaling, sampling, coding of categorical and continuous attributes have a significant impact on predictive accuracy on the classifier performance of decision trees, neural networks and support vector (Crone et al., 2006). Ballings and Van den Poel (2012) investigated how long an event history should be used for creating customer behavior summaries for churn prediction, while Tsai and Chen explored association rules for feature selection (Tsai & Chen, 2010). Next, we review the work related to the representation of longitudinal customer data and environmental variables; and the rare event problem in churn prediction in detail.

2.1. Representation of longitudinal customer data and environmental variables

Predictive models in customer relationship management rely on both static (non-time varying) characteristics of the customer, such as demographics; as well as dynamic characteristics consisting of customer behavior time series including customer transactions and interactions with the company. Each customer is represented with one observation. Customer demographics are obtained from a data warehouse, while the behavior information comes from transaction (Cao, 2010) databases. Traditionally, the transaction and interaction information in different time windows

Table 1
Sample of churn prediction papers since 2004.

Reference	Context				Predictor variables					Observations		Modeling technique
	Financial	Retail	Telecommunications	Other	Customer behavior	Customer perceptions	Customer demographics	Macro-environment	Customer–company interaction	Non-time-varying	Time-varying	
Larivière and Van den Poel (2004)	x				x				x		x	Survival analysis (hazard modeling)
Van den Poel and Lariviere (2004)	x				x		x	x			x	Survival analysis
Buckinx and Van den Poel (2005)				x	x		x			x		Logistic regression, neural network and random forests
Jamal and Bucklin (2006)			x		x		x		x	x	x	Survival analysis (Weibull hazard modeling)
Lemmens and Croux (2006)			x		x		x		x	x	x	Bagging, stochastic gradient boosting, binary logit model
Neslin et al (2006)			x		x		x					Logistic regression, neural network, decision tree, discriminant analysis, ...
Chu et al. (2007)			x		x		x		x	x		Decision tree, self organizing maps
Burez and Van den Poel (2008)				x	x		x		x	x		Random forests and survival analysis
Coussement and Van den Poel (2008)				x	x					x	x	Support vector machine, logistic regression, random forests
Mavri and Ioannou (2008)	x					x	x	x		x	x	Survival analysis
Kumar and Ravi (2008)	x				x		x			x		Multi perceptron, logistic reg, decision tree, random forests, RBF, SVM
Burez and Van den Poel (2009)		x	x							x		Random forests, logistic regression
Xie et al. (2009)	x				x		x			x		Improved balanced random forests, ANN, CWC-SVM, decision tree
Tsai and Chen (2010)			x		x		x		x	x		Neural network and decision tree
Huang et al. (2010)			x		x		x			x		Decision tree, multilayer perceptron neural network, support vector machine
Owczarczuk (2010)			x		x					x		Logistic regression, linear regression, discriminant analysis, decision trees
Huang et al. (2010)			x		x		x			x		Multiobjective feature selection approach using optimization
Karahoca and Karahoca (2011)			x		x		x			x		Adaptive neuro fuzzy inference, fuzzy c-means + ANFIS, decision tree
Verbeke et al (2011)			x		x		x			x		AntMiner+ and ALBA, C4.5, RIPPER, logistic regression
Xiao et al. (2012)	x		x		x					x		Dynamic classifier ensemble method for imbalanced data
Lee et al. (2012)		x			x						x	K-NN classification of time series

is summarized in static variables (Cao, 2010; Huang et al., 2010; Lee et al., 2012; Prinzie & Van den Poel, 2006). The recency, frequency and monetary value (RFM) variables which identify the time period since last transaction, the frequency of transactions and the money spent in a specified time window, respectively are very popular in practice. Prinzie and van den Poel, cluster financial services customers based on the changes in their account balance over time. They show that using the sequential cluster information in addition to the aggregate summaries of customer behavior improves churn prediction accuracy significantly (Prinzie & Van den Poel, 2006). Lee et al., take an interesting approach to using univariate time series data in telecommunications industry, and use the *k*-nearest neighbor classifier on the call volume time series to identify churners (Lee et al., 2012).

Chen et al. call the creation of summaries or features of longitudinal customer data the "standard framework for customer churn prediction" (Lee et al., 2012), which they contrast with "feature construction techniques" where each time point of the customer's each behavior is used as an attribute. They provide the use of multiple time windows and the proportional hazard (Cox) model with longitudinal behavioral variables (Van den Poel & Lariviere, 2004) as examples. Finally, they propose to use the customer longitudinal data with the hierarchical multi kernel support vector machine that simultaneously selects the static and dynamic features and the time subsequence for churn prediction (Lee et al., 2012).

While Chen et al. extend the "standard framework for customer churn prediction" by allowing the method to select the feature and its time period for longitudinal inputs; the framework still uses one observation per customer in terms of the churn output. Therefore, it cannot account for the impact of the economic and competitive environment that varies across time but not across customers.

The lack of environmental considerations and the non-time-varying nature of observations in the literature can be observed from Table 1, as the only papers using customer observations from multiple time periods and using environmental variables involve survival analysis.

Hazard modeling and survival analysis provide an avenue to incorporate environmental variables, such as the GDP index (Van den Poel & Lariviere, 2004). Interestingly, the proportional hazard (Cox) model with longitudinal variables by Van den Poel and Lariviere in financial services indicates that major drivers of attrition are demographic characteristics and environmental changes, whereas customer behavior predictors have only limited impact.

Jamal and Bucklin split the customer duration times into customer-month observations and organized the training data such that each observation corresponds to a unique customer time period, it (i.e., observation t for customer (i) to perform hazard modeling (Jamal & Bucklin, 2006). This type of data enables the researchers to model the impact of the time-varying covariates on the churn behavior.

2.2. The rare event problem in churn prediction and sampling approaches

Churn is a rare event in most organizations, since churners typically represent a small fraction of the active customers at a given time interval (Burez & Van den Poel, 2009; Kamakura et al., 2005; Lemmens & Croux, 2006; Xie, Li, Ngai, & Ying, 2009). In some settings, the number of customers is large enough to make up for this relative rarity and provides a sufficient number of positive examples for statistical learning. However, highly valuable customer segments consist of few customers who are particularly relevant for churn prediction and prevention. In fact, the Pareto principle suggest that 80% of firm profits are obtained from 20% of customers (Stahl, Matzler, & Hinterhuber, 2003). The industry practice is to develop a separate model for this top segment of customers since their behavior, and hence parameter values can be substantially different than the overall customer body (Storbacka, 1997). In those settings, the relative rarity is combined with the modest number of customers implying absolute rarity, and making it difficult for the classification algorithms to detect patterns of minority classes (Weiss, 2004). Particularly the greedy search algorithms, such as decision trees, have difficulty "finding the needle in the haystack" when the number of dimensions (variables) that describe the minority class is large, and hence the data may be fragmented due to greedy splits in the top layers of the tree. Indeed, research has shown that logistic regression provides better accuracy for smaller training sets while tree induction is better for larger data sets (Perlich, Provost, & Simonoff, 2003). Logistic regression predictive performance is sensitive to absolute rarity (King & Zeng, 2001).

While methods for addressing rarity in the literature include using expert knowledge to construct and select better features, use of boosting algorithms, or learning only the rare class; in this paper we restrict our attention to the classification framework with a predetermined set of features and standard classification algorithms. Within this context sampling approaches are frequently used to alleviate the problems associated with rarity: Undersampling refers to randomly removing the majority class and is useful under relative rarity. Burez and Van den Poel report that in experiments with datasets containing minimum of about three thousand churners, random undersampling improved churn prediction accuracy, while boosting did not help (Burez & Van den Poel, 2009). Oversampling randomly resamples the minority class which may result in overfitting. Based on their experiment with eleven datasets and 21 classification methods, Verbeke et al., report that oversampling did not result in significant improvement in accuracy and suggest the use of more sophisticated sampling methods, such as SMOTE (Verbeke et al., 2012). Synthetic minority over-sampling technique (SMOTE) creates new minority observations as a weighted average of the k nearest neighbor minority observations (Chawla, Bowyer, & Hall, 2002). A study on the statistical properties of SMOTE, which is a very popular method, finds that it decreases the data variability and it introduces correlation between samples of the rare class, which impacts the independence assumption of some classifiers, such as logistic regression, and the variable selection routines such as *t*-test (Lusa, 2013). Another approach to handle class imbalance combines ensemble learning with cost-sensitive learning, and for each test sample selects an ensemble strategy that performs best for similar observations (Xiao, Xie, He, & Jiang, 2012).

3. Proposed dynamic churn prediction framework

In this section we introduce the dynamic churn prediction framework that utilizes the longitudinal nature of the customer data. The customer data has the following components: y_{it} denotes whether the *customer i* is active or churned at time *t*.

$$y_{it} = \begin{cases} 0 & \textit{if customer is active at } t \\ 1 & \textit{if churned at } t \end{cases}$$

x_{ijt} stands for the time varying behavior attribute j at *time t* for customer i, and s_{il} is the non-time varying attribute l of customer i, where $i = 1,\ldots,N$; $j = 1,\ldots,m$; $t = 1,\ldots,T$, and $l = 1,\ldots,p$.

3.1. Standard approach – SPTD, proposed MPTD and alternative SPTD+lags

3.1.1. Sptd

The "standard framework for customer churn prediction" as defined by Lee et al. (2012) entails one observation per customer with features that summarize the historical customer behavior

time series. Recency, Frequency, Monetary Value (RFM) are popular examples of such features. The models are trained to predict the next feasible period churn behavior of the customer, where operational feasibility due to data processing can dictate an additional idle period. We define *SPTD* (Single Period Training Data) as a training dataset that consists of one observation per customer who is active as of the prediction time, even though the predictor variables may summarize customer history.

In this approach the training data is $SPTD_{t_0} = [\,Y_{t_0} \quad S \quad D_{t_0-\delta}\,]$, where $t_0 - \delta$ is the most recent operationally feasible time period to use for predicting churn at t_0, $N_{t_0-\delta}$ is the number of active customers at $t_0 - \delta$, and Y_{t_0} is an $N_{t_0-\delta}$ x 1 column vector containing the y_{it_0}. S is an $N_{t_0-\delta}$ x p matrix containing the s_{il}. $D_{t_0-\delta}$ is the $N_{t_0-\delta}$ x q matrix, consisting of the constructed features $f(X_i, t_0 - \delta, w)$, where the feature generation function $f()$ maps X_i (the T x m matrix consisting of the x_{ijt}) to a row vector with q features. w is the time window used for feature generation. Typically, $t_0 = T$, to ensure that the model captures the most recent relationships.

3.1.2. Mptd

We define *MPTD* (Multiple Period Training Data) as follows

$$MPTD = \begin{bmatrix} SPTD_T \\ SPTD_{T-1} \\ \vdots \\ SPTD_{t_1+1} \\ SPTD_{t_1} \end{bmatrix} E = \begin{bmatrix} Y_T & S & D_{T-\delta} & E_{T-\delta} \\ Y_{T-1} & S & D_{T-\delta-1} & E_{T-\delta-1} \\ & \vdots & & \vdots \\ Y_{T+1} & S & D_{T_1-\delta+1} & E_{T_1-\delta+1} \\ Y_{T_1} & S & D_{T_1-\delta} & E_{T_1-\delta} \end{bmatrix}$$

Here, $t_1 \geqslant w + 1$, and E is an $T - t_1$ x r matrix with r environmental variables that vary with time but not across customers.

MPTD contains an observation for every time period within the analysis time frame when the customer was active. The customer behavior predictor variables are calculated for each customer and time period t, using the same period length of historical customer data between t and $t - w$. Therefore, the customer behavior data that is used for calculating D_t (the constructed features at time t), through D_{t-w+1} overlap partially, where the highest overlap occurs between consecutive observations D_{t-1} and D_t.

The environmental variables take the same value for all observations in a given time period, but vary across time. Time invariant customer characteristics, such as gender vary across customers, but not across time. The customer behavior features vary across time and customers.

3.1.3. SPTD+lags

SPTD+lags is defined as a training data set that explicitly creates k lag variables for each customer behavior variable in addition to those in *SPTD*. For example, it may contain the yearly return on customer's investment portfolio calculated not just in the last period but in the last k periods.

$$\begin{aligned} SPTD + lags_{t_0} &= [\,SPTD_{t_0} \quad D_{t_0-\delta-1} \quad \ldots \quad D_{t_0-\delta-k+1}\,] \\ &= [\,Y_{t_0} \quad S \quad D_{t_0-\delta} \ldots \quad D_{t_0-\delta-k+1}\,] \end{aligned}$$

Notice that $SPTD+lags_T$ contains lags of customer behavior summary $D_{T-\delta}$ to $D_{T-\delta-k+1}$ as predictors for the most current churn behavior, Y_T, while *MPTD* contains $D_{T-\delta}$ to $D_{t_1-\delta}$ as predictors for the churn behavior δ periods after the behavior summary, Y_T to Y_{t_1}.

The decision as to how many lags to include is a design variable similar to how many features of the customer behavior time series to create and include in the training dataset. While a higher k provides more information about the behavior path that brings the customer to the present, it also decreases the ratio of number of observations per parameter.

3.2. Theoretical justification and comparison

In this subsection we provide theoretical justifications as to the pertinence of the proposed MPTD framework, and we compare it to SPTD and SPTD+lags in terms of the characteristics that affect the out of sample accuracy. Table 2 provides an overview of the theoretical comparison.

Clearly, other things being equal, any learning algorithm will perform better when it is provided with more examples. A large scale learning curve analysis that compares the generalization performance (out of sample accuracy) of decision trees and logistic regression as a function of the training-set size shows that as sample size increases the out of sample accuracy of the model increases at a decreasing rate (Perlich et al., 2003). Using the simplifying assumption that the number of customers, N, is the same in each time period, the SPTD and SPTD+lags training datasets contain N observations, while the MPTD dataset has $N\ (T - t_1)$ observations. Hence MPTD provides an improvement in predictive accuracy compared to SPTD, while the improvement is decreases with the number of customers, N, and increases with the number of time periods included in the MPTD $(T - t_1)$.

A related argument is that the *MPTD* contains $(T - t_1)$ times more churner, i.e., rare event, observations than *SPTD* or *SPTD+lags* which addresses the absolute rarity problem. "The lack of data makes it difficult to detect regularities within the rare classes/cases" (Weiss, 2004). In fact, MPTD is expected to result in better generalization performance than SPTD with oversampling or SMOTE – which are the popular remedies for addressing the rarity problem. Oversampling simply repeats existing churn examples in the training dataset, and hence may lead to memorization of irrelevant characteristics of churners. Similarly, SMOTE approach which constructs "new" rare cases as a linear combination of the existing rare cases (Chawla et al., 2002), which may give unrealistic examples when different churner profiles are present in the data. In comparison, MPTD reflects the actual variation in churn profiles.

Beyond the training dataset sample size, its relationship to the complexity of the dataset – the number of variables – is an important driver of the resulting generalization performance. The term *curse of dimensionality* refers to the phenomenon that the sampling density is proportional to $N^{\frac{1}{p}}$, where p is the number of dimensions (variables) (Bellman, 1960; Hastie, Tibshirani, & Friedman, 2009), with unfavorable consequences for variable selection and identification of patterns. Particularly when the churn rate is low, models may overfit the training data and fail to generalize properly. Evaluated based on the sampling density, *SPTD+lags* performs the worst among the three frameworks, as it involves additional variables and low sample size, whereas MPTD does best due to its higher sample size.

Another advantage of *MPTD* is that it allows variation in the environmental variables. In the *SPTD* and *SPTD+lags* all observations pertain to the same time period; therefore, the resulting models cannot estimate the impact of the environment on churn probability. For example, we cannot account for the competitor's temporary campaign or a decrease in interest rates in *SPTD* or *SPTD+lags* frameworks. Insights about the impact of the environmental drivers, such as the GDP growth, are very valuable for correctly positioning the company/ product in dynamic environments (Van den Poel & Lariviere, 2004); and they can only be generated with *MPTD*.

On the other hand, it can be argued that in a dynamic environment, where the triggers for customer churn evolve rapidly over time, *SPTD* and *SPTD+lags* facilitate learning the most relevant churn drivers based on the most recent churn examples. While the main effects of the environmental changes (E) on churn (Y) will be captured in the MPTD framework, its interaction effects with the customer information (S and D) may be more difficult to cap-

Table 2
Theoretical comparison of SPTD, SPTD+lags and MPTD in terms of characteristics that affect predictive accuracy. Most desirable level of each characteristic is printed in bold.

Characteristics	SPTD	SPTD+lags	MPTD
Training data sample size	Low	Low	**High**
Absolute rarity problem	High	High	**Low**
Sampling density	Medium	Low	**High**
Environmental variation	None	None	**Present**
Recency of the training data (under rapidly changing churn behavior patterns and absence of rarity)	**High**	**High**	Medium

ture. In the machine learning literature it is recognized that "a changing context can induce changes in the target concepts", a notion known as *concept drift* (Widmer & Kubat, 1996). "Keeping only the recent window of trusted examples and hypotheses", as in SPTD and SPTD+lags, is one of the approaches to deal with concept drift; however the more complex and/or the more rare the concept the larger the window size that results in better performance (Widmer & Kubat, 1996). Therefore, under rapidly changing churn behavior patterns and absence of rarity, frequently retraining the churn model with recent *SPTD* or may capture the most recent and representative behavior patterns.

In the econometrics literature *MPTD* is referred to as panel data. The observations in panel data may be subject to disturbances that are not independently and identically distributed (iid), which can potentially lead to biased estimates of the parameter values and underestimation of their variability in methods that assume iid disturbances, such as logistic regression (Greene, 2012). This is an important concern when the main purpose of the study is to measure the impact of particular drivers. Particularly, all observations of some customers may be more prone to churn due to an unobserved characteristic, which can be addressed by specifying fixed or random effects models. The fixed effects approach is equivalent to creating a dummy variable for each customer, which makes the model specific to the customers in the sample, and hence not useful for out of sample prediction, and also prevents the model from learning the impact of time invariant customer characteristics. The random effects approach models the customer-specific constant terms as randomly distributed across individuals, and does not suffer from the issues mentioned for fixed effects; however the estimation of the random effects models for large datasets is difficult or impossible due to memory limitations and computational complexity (Gebregziabher et al., 2012; Pennell & Dunson, 2007). We acknowledge that the potential lack of independence introduced by multiple observations of the same customer may bias the parameter estimates and/or artificially increase the significance of the parameters, on the other hand, our primary objective is predictive accuracy and the increased sample size in *MPTD* should decrease the variance of the parameter estimates.

3.3. Training data for dynamic churn prediction

A customer who appears to be active this period may be planning to leave two months later when the bonds in her portfolio reach maturity. Incorporating the same customer under different circumstances along with whether they churned in the next δ time periods, makes the naturally occurring experiment data available to the model to learn the impact of the customer behavior and company relationship variables while keeping demographics constant.

We define *Periods-to-churn* as a discrete ordinal variable as follows. If the customer does not churn within a specified maximum horizon n, it takes the value $n+$, otherwise it is equal to the number of periods to churn, *Periods-to-churn* = 1,2,…,n, $n+$. Further, we create n dummy variables for churn within δ periods ($W\delta C$), $\delta = 1,\ldots,n$.

$$W\delta C = \begin{cases} 1 & \text{if } \textit{Periods-to-churn} \leqslant \delta \\ 0 & \text{otherwise} \end{cases}$$

Period	Level of Activity	Periods to churn	W1C label	W2C label	W3C label	..	Wn-1C label	WnC label
1	active	n+	0	0	0	0	0	0
:	active	n+	0	0	0	0	0	0
t_0-n+1	active	n	0	0	0	0	0	1
:	active	n-1	0	0	0	:	1	1
:	active	:	0	0	:	:	:	:
t_0-2	active	3	0	0	1	1	1	1
t_0-1	active	2	0	1	1	1	1	1
t_0	active	1	1	1	1	1	1	1
t_0+1	inactive		.	.	.	.	.	.
:	inactive		.	.	.	.	.	.

Fig. 1. Churn label generation for churners.

Fig. 1 exhibits how the churn labels are generated for a customer that churns in some period t_0, $t_0 < T$. As shown in the figure, the customer is active up to t_0. Consistent with the objective of detecting churners before they do so, the *W1C* label (the "usual" churn label) is assigned 1 at period t_0 and 0 beforehand. Also notice that once the customer churns, no observation is generated in the subsequent time periods. With respect to *W2C*, the customer is recorded as a churner in periods $t_0 - 1$ and t_0, and a non-churner beforehand. This is repeated with the same logic until all the labels *W3C* through *WnC* are obtained. Missing values are generated for periods following the churn event.

The non-churner behavior is displayed in Fig. 2. Accordingly, the customer is labeled as a non-churner in the *W1C* label in all periods up to and including t_1. Even though we may know that the customer is active in period $t_1 + 1$, since we have set it aside as the test time period, we treat the *W2C* churn label at time period t_1 as "not observable" and set to missing. Similarly, *W3C* label is set to missing from period $t_1 - 1$ onwards, etc. Not observing this convention may lead to unrealistically optimistic evaluation results of the test dataset. The last columns of Fig. 1 exhibits how the churn labels are generated for a customer that churns in some period t_0, $t_0 < T$. As shown in the figure, the customer is active up to t_0. Consistent with

Period	Level of Activity	Periods to churn	W1C label	W2C label	W3C label	..	..	WnC label
1	active	n+	0	0	0	:	:	0
2	active	n+	0	0	0	:	:	0
3	active	n+	0	0	0	:	:	0
4	active	n+	0	0	0	:	:	0
5	active	n+	0	0	0	:	:	0
:	active	n+	0	0	0	:	:	0
T-n	active	n+	0	0	0	:	:	0
T-n+1	active	n+	0	0	0	:	:	0
:	active	n+	0	0	0	:	:	.
:	active	n+	0	0	0	:	:	.
:	active	n+	0	0	0	:	:	.
T-1	active	n+	0	0	.	:	:	.
T	active	n+	0	.	.	:	:	.
T+ 1	active		.	.	.	.	.	.
:	:		:	:	:	:	:	:
:	:		:	:	:	:	:	:
T + n	active		.	.	.	.	.	.

Fig. 2. Churn label generation for non-churners.

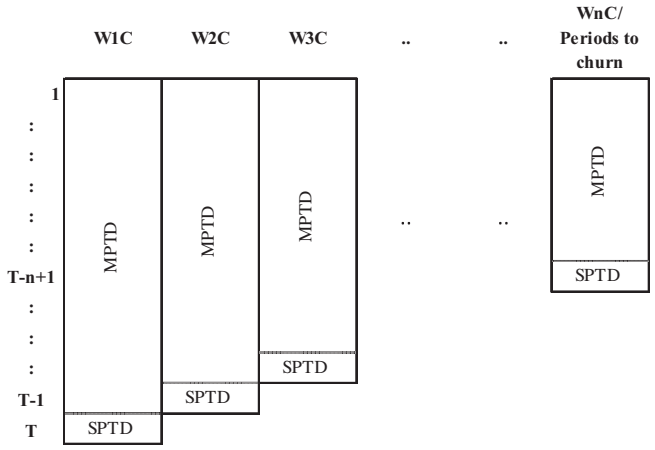

Fig. 3. Illustration of the dynamic churn prediction training dataset size for *WδC* and *Periods-to-churn* prediction, with *MPTD* and *SPTD* methods.

the objective of detecting churners before they do so, the *W1C* label (the "usual" churn label) is assigned 1 at period t_0 and 0 beforehand. Also notice that once the customer churns, no observation is generated in the subsequent time periods. With respect to *W2C*, the customer is recorded as a churner in periods $t_0 - 1$ and t_0, and a non-churner beforehand. This is repeated with the same logic until all the labels *W3C* through *WnC* are obtained. Missing values are generated for periods following the churn event.

Figs. 1 and 2 show the alternative label: *Periods-to-churn*, which indicates the number of periods to the churn event, if the customer will be churning within n periods or less, *n+* if the customer is observed for the next *n* periods and does not churn. If the customer does not churn within the observed time period, the last *n Periods-to-churn* labels will be set to missing, by the same reasoning as the *WδC* labels.

Fig. 3 illustrates dynamic churn prediction training datasets with respect to time. Customers who churn at time *t* result in positive instances in periods $t - 1$ through $t - \delta$, Assuming a constant churn probability, p_c across time, $E[W\delta C] = 1 - (1 - p_c)^{\delta}$. Therefore, as δ increases, the relative rarity decreases, regardless of whether *MPTD* or *SPTD* approach is used.

On the other hand, as the prediction lead time, δ, increases the training data is moving farther away from the test period, as shown in Fig. 3. The most recent *WδC* label that can be determined by using the available information is $T - \delta + 1$. Since assignment of the *Periods-to-churn* value requires knowledge of customer behavior in the next *n* time periods, the most recent training data for predicting *Periods-to-churn* is for $T - n + 1$.

If the *MPTD* approach is used, the largest dataset can have up to *T W1C*, and up to $T - n + 1$ *WnC* or *Periods-to-churn* observations per customer. As the prediction horizon, δ, increases the sample size in the *MPTD* for the *WδC* label increases.

In summary, training data for *Periods-to-churn* has fewer observations and does not contain the most recent observations for most of the prediction horizons, compared to training data for *WδC* labels.

3.4. Dynamic churn prediction

3.4.1. Common method – survival analysis

Survival analysis is a class of statistical methods modeling the occurrence and timing of events – customer churn in our context. It measures the probability that an individual will survive given that it lived until the measuring time. In churn datasets the duration (time since the customer has become a customer of the bank until churn) is right censored, since for the active customers the complete duration of the relationship cannot be observed. In addition, we can observe the customer only for a specific period of time, i.e., the data is subject to left truncation. An advantage of survival analysis is that censored and truncated data can be accounted for.

In survival analysis the time that elapsed since the starting point until the failure (churn) event, *t*, is a continuous random variable with cumulative density function $F(t)$, and pdf $f(t)$. The survival function $S(t)$ describes the unconditional probability of survival up to the elapsed time *t*, whereas the hazard function $\lambda(t)$ denotes the instantaneous risk that the event will occur at time *t* given that the individual survived up to that point.

$$S(t) = 1 - F(t) = P(T > t)$$

$$\lambda(t) = \lim_{\Delta t \to \infty} \frac{P(t \leqslant T < t + \Delta t | T \geqslant t)}{\Delta t} = \frac{f(t)}{S(t)} = -\frac{dS(t)}{S(t)dt}$$

The Cox regression is a proportional hazard model (Kleinbaum & Klein, 2005) that dominates the field of the dynamic survival models. Here, the basic assumption is that the covariates affect the hazard rate multiplicatively as follows, where $\lambda_0(t)$ is the base hazard rate, and the Z_j and β_j are the value for covariate *j* at time *t*, and the coefficient for covariate *j*, respectively.

$$\lambda(t|Z) = \lambda_0(t) exp\left[\sum_j \beta_j Z_j(t)\right]$$

Denoting $S_i(t)$ as the estimated survival probability of *i*th customer at time *t*, the probability of within-δ-periods churn for the corresponding customer, given that s/he has survived up to t_0, is computed as follows:

$$p_\delta = \frac{S_i(t_0) - S_i(t_0 + \delta)}{S_i(t_0)}$$

While survival models, in particular the Cox regression model has been used for modeling customer churn behavior over time, they are primarily descriptive models providing insights. For example, Mavri and Ioannou use survival analysis to estimate the contribution of each factor to the switching behavior of Greek banking customers in different periods of time (Mavri & Ioannou, 2008). Lariviere and van den Poel study customer churn in financial services and build a hazard model to detect the most convenient product categories to cross-sell in order to reduce the churn likelihood (Larivière & Van den Poel, 2004). In a related study they use the Cox proportional hazard model which they explain as an appropriate choice since customer tenure is a major driver of churn in financial services, with higher hazard rates in the early years and in twenty plus years. They find that demographics and environmental changes have great impact on retention (Van den Poel & Lariviere, 2004).

3.4.2. Proposed approach with standard classifiers

We propose that dynamic churn prediction can be accomplished more accurately using standard classifiers, such as logistic regression, decision trees or ensembles; using the advanced churn label framework we developed in the previous section. While the survival analysis considers the time since birth (start of customer relationship), classification methods focus on the discrete *Periods-to-churn*. We identify three approaches to estimate the probability that the customer will churn within δ time periods, p_δ, given that he/she is an active customer at the time of prediction.

(a) Ordinal classifier assuming that indicators of churn and their relative impact are the same regardless of the lead-time δ. In other words, this approach assumes that the ranking of customers in terms of their probability to churn is the same independent of the churn horizon.

(b) Multi-nominal classifier assuming that, for example, churn in the next month and in four months may have different triggers and/ or indicators. Hence, this approach allows lead-time specific ranking of customers in terms of churn probability. The multinomial classifier predicts the probability that the customer will churn δ periods later, rather than within δ periods, which is then converted to the p_δ.

(c) n Independently trained binary classifiers predicting $W\delta C$, $\delta = 1 \ldots n$. Here each classifier models the probability that the instance belongs to the class or higher order classes (Frank & Hall, 2001). An important distinction of this approach versus the other two is the following: By training each $W\delta c$ model independently we can take advantage of the additional and more recent training data that is available for $W\delta C$, $\delta < n$. Further, unlike the ordinal classifier it allows horizon specific predictors and churn probability ranking.

Table 3 summarizes the characteristics of the four approaches in terms of two characteristics: whether they allow prediction horizon specific customer ranking and predictors, and whether they consider the most recently available data. Binary classifiers predicting $W\delta C$s would be equivalent to the multinomial classifier, if trained on the same dataset. Multinomial classifier would be equivalent to the ordinal classifier, if predictors and their coefficients were restricted to be the same across horizons.

4. Application in private banking and experimental results

We applied the proposed and benchmark training data generation and dynamic churn prediction approaches to predict churn of the private banking customers at a European bank. Private banking customers have a very large portfolio that they invest with the bank, hence they are among the most valuable customers, and their retention is of utmost importance. The bank operates in an emerging market with dynamic economic conditions, as well as numerous existing and new competitors. Based on the bank's definition, a customer is declared churned if her portfolio size falls below a specific threshold value and stays that way for six consecutive months.

In this environment we developed and tested churn prediction models for the next time period and for several time periods ahead. Our goals in this experiment are to answer the following questions

(1) Does the *MPTD* framework significantly increase next period churn predictive accuracy vs. the "standard approach" *SPTD*?

- Does *MPTD* improve predictive accuracy over and above balancing the dataset with respect to rare event, i.e., the churn observations?
- Does adding the lags of customer behavior features as additional columns to the *SPTD*, rather than as additional rows corresponding to lagged churn behavior, as in *MPTD*, provide a significant accuracy improvement? Does *SPTD+lags* improve predictive accuracy significantly vs. *SPTD*?
- Are the answers to the above questions classifier dependent?
- Does *MPTD* capture the impact of the environment on customer churn behavior?

(2) Do independently trained binary classifiers with $W\delta C$ labels offer significantly better dynamic churn prediction than the popular Cox regression?

- Does the *MPTD* framework significantly improve dynamic churn prediction accuracy vs. the *SPTD* and *SPTD+lags*?
- Is there improvement due to the use of most recent available data for each horizon?
- Is there improvement due to horizon-specific predictors and hence horizon-specific customer ranking?
- How does the prediction lead time affect the answers?

4.1. Data and generated variables

In the analysis we focused on 7204 private banking customers who were active as of April 2009, and have been active customers for at least ten months. We used the 12 months from April 2009 to March 2010 for constructing the training datasets, and set aside the subsequent five months as a common holdout test dataset to evaluate the predictive accuracy of the churn prediction models. Since the churn definition involves tracking customer portfolio size for six consecutive months, the last six months of data had to be used to determine whether the customer had churned or not. Fig. 4 shows an overview of the dataset time periods used for developing and testing *W1C* models.

The *MPTD* comprises the customer observations of periods April 2009 through March 2010, while the *SPTD* includes only the most recent customer observations, i.e., March 2010. As a result, there are 84,180 observations in the *MPTD* and 6821 observations in the *SPTD* and *SPTD+lags*. The churn rate in both training datasets is similar. Note that each customer observation (in *SPTD* and *MPTD*) utilizes customer behavior data from the previous 12 months to create the predictor variables summarizing past customer behavior, w=12. *SPTD+lags* further takes 6 lags of each such summary, $k = 6$.

Advance churn labels *W1C* through *W6C*, and the *Periods-to-churn* variable are constructed as explained in Section 3.2. The *W1C* model uses customer data up to 201003, while the *W6C* and *Periods-to-churn* models use training data up to 200910. Fig. 5 illustrates the implications terms of sample size and number of positive cases versus the prediction horizon. We observe that as the prediction horizon increases the number of positive cases increases while in the *MPTD* the sample size decreases, alleviating the absolute and relative rarity problem. The *SPTD* sample size is much smaller than *MPTD*.

The preprocessing steps consisted of normalization of the variables to mean 0 and standard deviation 1; outlier elimination, where we capped the numeric variables which are above the 99th percentile or below the 1st percentile of the population; and creation of categorical dummies.

We developed and used a total of 169 variables which can be classified under four main categories, with the number of variables given in parentheses: customer behavior (139), customer demographics (10), customer-company interactions (8), and economic indicators (12) which are only present in the *MPTD* dataset, as explained earlier. The customer behavior category includes variables that are related to the current and historical portfolio value, return and product usage information. Some variables in this

Table 3
Theoretical comparison of dynamic churn prediction approaches.

Characteristics	Ordinal	Multinomial	Independently trained binary	Cox regression
Horizon specific customer ranking and predictors	No	Yes	Yes	No
Most recently available data considered	No	No	Yes	Yes

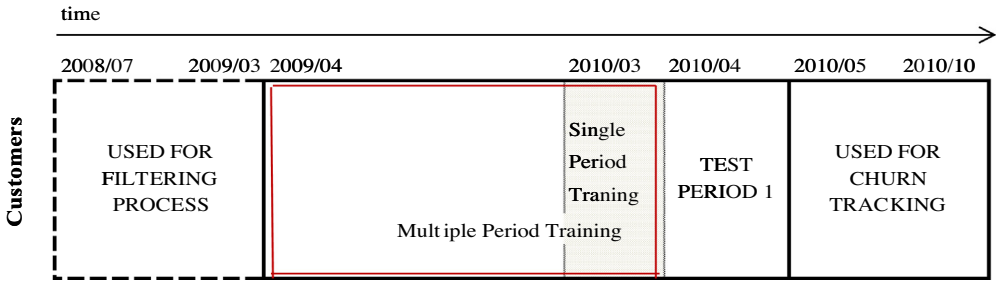

Fig. 4. Training and test dataset time periods for next period churn models (*W1C*).

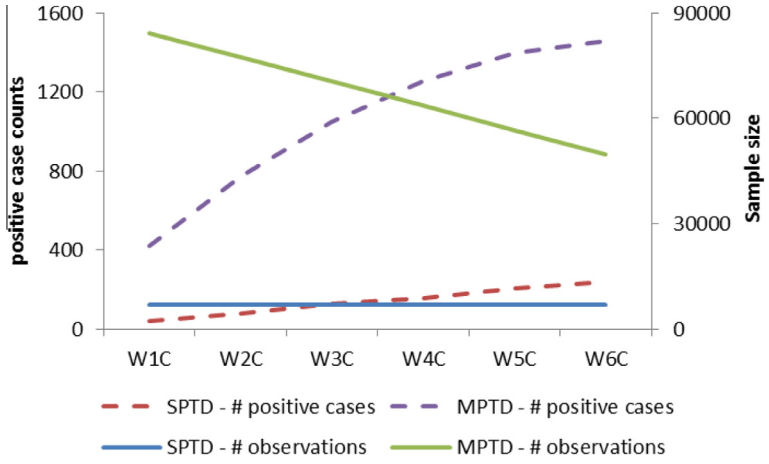

Fig. 5. Sample size and absolute rarity by prediction horizon for *MPTD* and *SPTD*.

category are current portfolio value and its change, the weight of various financial instruments in the portfolio, whether certain type of instruments are currently in the portfolio and have ever been used by the customer. We have also created six lag variables for each of the customer behavior variables to be used in the *SPTD+lags* dataset. The customer demographics category consists of demographic variables such as age, gender, education level, and nationality. The customer-company interactions category of variables reflects the interaction between the customer and the bank. Yavas, Benkenstein, and Stuhldreier (2004) find in their survey based descriptive study of German private banking customers that service quality and timeliness of services are linked to switching; i.e. customer churn. We created variables that reflect the service level of the customer's branch and representative, in addition to customer tenure. The last category incorporates economic indicators which reflect the environmental conditions. These variables include consumer confidence index, consumer price index, local stock market index, USD exchange rate and the yield curve slope which are commonly used as economic indicators (Stock & Watson, 1989). In summary, the SPTD has 157 variables, the SPTD+lags has 991 variables and the MPTD has 169 variables.

As expected, churn is a rare event in the considered organization, with a monthly churn rate of less than 1% throughout the training and test time periods.

4.2. Next period churn prediction experiments

To investigate first group of questions raised at the beginning of this section we designed the following full factorial experiment with three factors and ten replications, yielding a total of 120 models. To observe the variability of the accuracy results we have created ten bootstrap samples for each of the training datasets while keeping the test dataset the same. Notice that a v-fold cross validation approach would result in situations where the training data is from a later time period than the test data – which is not realistic in a predictive setting with time series data.

4.2.1. Factors varied in the experiment

The first factor is the training dataset creation method with three levels (*SPTD*, *SPTD+lags*, *MPTD*).

The second factor is the classifier with two levels, logistic regression and decision tree, which are widely used methods in many churn studies as robust benchmarks. Further, they differ in how they react to the rare event problem, as discussed in Section 2.2. The stepwise logistic regression was implemented with SAS 9.2 with $\alpha = 0.1$, whereas the classification tree (J48) was trained with WEKA 3.6., using the default settings.

The last factor is whether the data is balanced or not. We use the "Synthetic Minority Over-sampling Technique" (SMOTE) (Chawla et al., 2002), which has been successfully used in churn prediction problems in the literature, e.g. (Kumar & Ravi, 2008), to balance the classification categories. We use WEKA 3.6., with five nearest neighbors and with a target ratio of 1:1 between non-churn and churn events.

4.2.2. Evaluation criteria

The primary evaluation criterion for a predictive model is consistent out-of-sample predictive accuracy. Classification accuracy is a function of the probability threshold that is used to declare a customer a churner. Higher threshold values require more evidence before declaring that the customer is a churner, thus decreasing false positives while increasing false negatives.

The first measure of accuracy we use, the AUC (the area under the curve), considers the classifier performance for all possible threshold values. It is the area under the receiver operating characteristics (ROC) curve which plots the true positive rate versus the

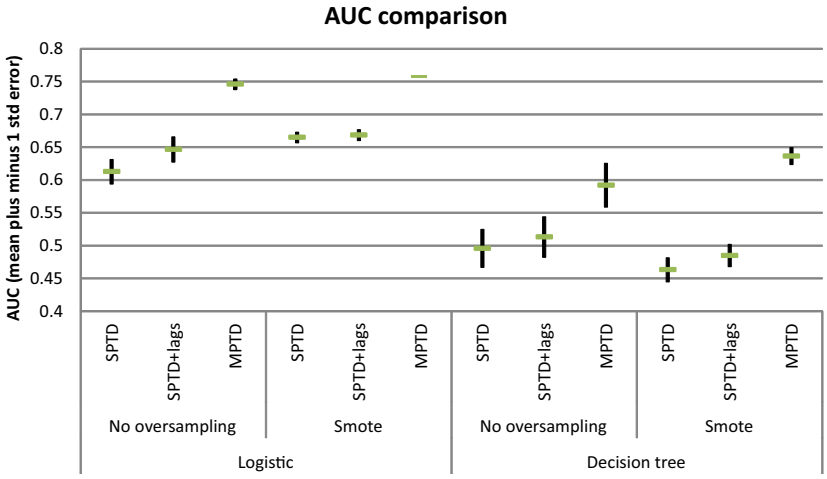

Fig. 6. AUC comparison of the *SPTD* and *MPTD* methods for *W1C* in 2010/04.

false positive rate for all possible threshold values. AUC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one, hence its value can theoretically range from 0 to 1, while 0.5 corresponds to random guessing. An AUC value less than 0.5 may occur, for example, when we reverse the predictions of a good classifier (Fawcett, 2006); in general it indicates that the classifier is misleading. For rare event classification problems, AUC is the preferred accuracy performance measure to use (Weiss, 2004).

The second measure of accuracy we use, the top-decile lift (TDL), focuses specifically to the prediction of the customers most likely to churn. It measures the lift in predictive accuracy for the top ten percent of customers predicted to be most likely to churn versus a random selection. The top-decile lift is particularly relevant from a managerial point of view as companies contact those customers most likely to churn, and therefore used in many churn prediction studies, e.g. (Verbeke et al., 2011; Xie et al., 2009).

Thus the AUC and TDL measures provide complementary views on the accuracy performance. We evaluate accuracy across ten bootstrap samples for each experimental setting. We also report the complexity of the resulting models, as a diagnostic measure.

4.2.3. Results

Figs. 6 and 7 compare the mean AUC and TDL values for the three training set generation methods (*MPTD*, *SPTD*, and *SPTD+lags*) for each combination of the classification methods and use of oversampling. The error bars represent plus minus one standard error around the mean of the results obtained with ten bootstrap samples. The results are for predicting churn in the first test period, namely 2010/04 with a sample size of 6783 customers. When training did not result in a model (a bootstrap sample on both the *SPTD* and *SPTD+lags* with the logistic model without oversampling, and another bootstrap sample on *SPTD* with the decision tree model without oversampling), we set the AUC to 0.5 and TDL to 1; i.e., random model results. Table 4. Comparison of MPTD vs. SPTD and SPTD+lags for next period churn prediction using the AUC measure. Tables 4 and 5 provide the mean accuracy performances, the improvement that MPTD provides, as well as the p-values for testing the hypothesis that there is no difference between the mean accuracy of MPTD versus of SPTD and SPTD+lags, using AUC and TDl as the accuracy measure respectively.

Finally, we report the resulting model complexity. Fig. 8 shows the average number of parameters in the logistic regression and the average number of leaves in the decision tree models, with plus minus one standard deviation.

Considering Fig. 6, the left most block indicates that when the logistic regression model with no oversampling is trained with *MPTD* the mean AUC is much higher than when it is trained with *SPTD*, which is not significantly different from *SPTD+lags*. The same pattern is repeated in all other blocks (logistic regression with oversampling, decision tree with and without oversampling). Also considering Fig. 7, Tables 4 and 5, we observe that the mean AUC

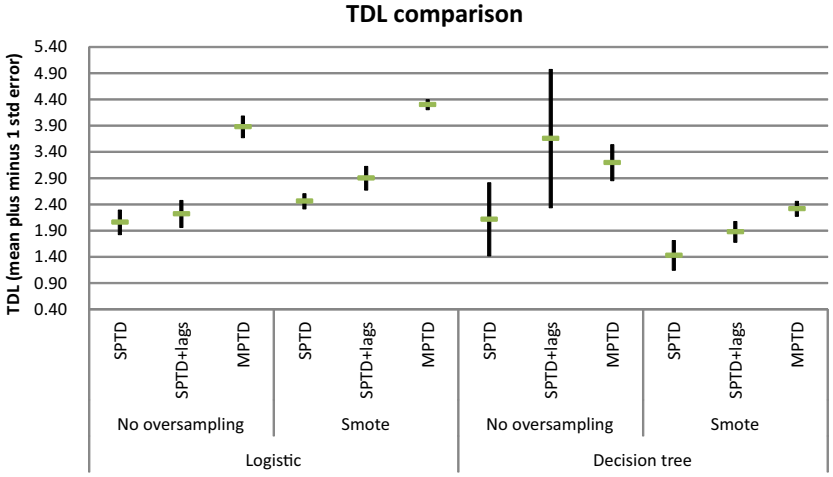

Fig. 7. TDL comparison of the *SPTD* and *MPTD* methods for *W1C* in 2010/04.

Table 4
Comparison of MPTD vs. SPTD and SPTD+lags for next period churn prediction using the AUC measure.

Classifier	Sampling	Mean AUC			MPTD improvement		p Value H_0 no difference		
		SPTD	SPTD+lags	MPTD	vs. SPTD (%)	vs. SPTD+lags (%)	MPTD vs. SPTD	MPTD vs. SPTD+lags	SPTD vs. SPTD+lags
Logistic reg	No oversampling	0.61	0.65	0.75	22	15	0.00	0.00	0.91
	Smote	0.66	0.67	0.76	14	13	0.00	0.00	0.63
Decision tree	No oversampling	0.50	0.51	0.59	19	15	0.01	0.03	0.67
	Smote	0.46	0.48	0.64	37	31	0.00	0.00	0.82

Table 5
Comparison of MPTD vs. SPTD and SPTD+lags for next period churn prediction using the TDL measure.

Classifier	Sampling	Mean TDL			MPTD improvement		p Value H_0 no difference		
		SPTD	SPTD+lags	MPTD	vs. SPTD (%)	vs. SPTD+lags (%)	MPTD vs. SPTD	MPTD vs. SPTD+lags	SPTD vs. SPTD+lags
Logistic reg	No oversampling	2.06	2.22	3.88	88	75	0.00	0.00	0.68
	Smote	2.46	2.90	4.30	75	48	0.00	0.00	0.95
Decision tree	No oversampling	2.11	3.65	3.20	51	−13	0.08	0.63	0.85
	Smote	1.43	1.87	2.32	62	24	0.00	0.04	0.90

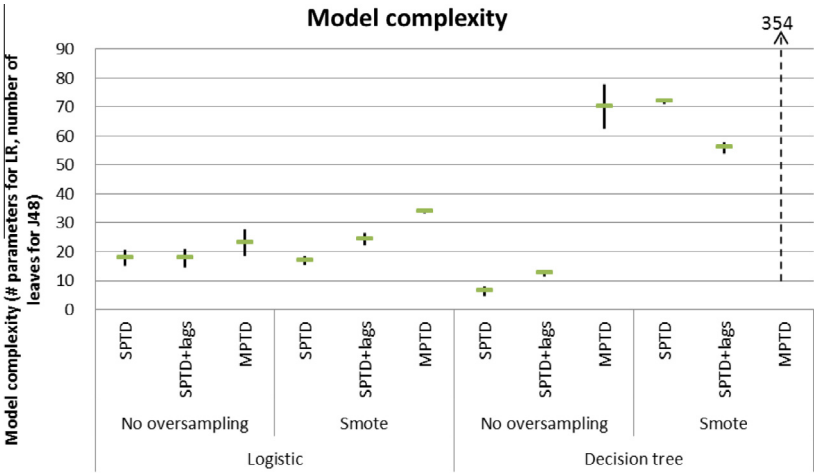

Fig. 8. Model complexity comparison for *W1C* models.

and TDL values for *MPTD* are significantly higher than *SPTD* at the 0.05 level or better – with the exception of the TDL measure of the decision tree on non-oversampled data, where there is no significant difference. Here, the TDL results have very large variability, which is in line with the finding in the literature where decision trees were found to be unstable in high percentiles of the lift curve (Owczarczuk, 2010). As expected based on the theory, decision trees generally perform worse than logistic regression due to the rare event problem; interestingly, synthetic oversampling further degrades the performance of SPTD or SPTD+lags, producing a misleading classifier that performs worse than random classification in terms of AUC. We observe that the models trained on *MPTD* outperform the models trained on *SPTD*, irrespective of the classification algorithm and whether oversampling is applied or not. On average, *MPTD* provides 69% improvement in TDL, and 23% improvement in AUC versus *SPTD* across classifiers and oversampling. Both accuracy measures point to significant accuracy improvement due to MPTD, while the TDL measure signals an even larger improvement for the detection of the top ten percent customers who are most likely to churn compared with the detection of those least likely to churn or those in between – which is measured by the AUC.

Considering Fig. 8 we observe that the logistic regression models trained on *MPTD* have slightly more parameters while the decision tree models are substantially more complex than models trained with *SPTD*. In light of the significant accuracy improvement that the MPTD has over SPTD, we conclude that MPTD allowed the classifiers to identify more details of the churner profiles than the SPTD, and that this complexity is not a manifestation of overfitting.

Models trained on *MPTD* (without oversampling) have significantly higher accuracy compared with models trained on balanced *SPTD*, on average 20% higher AUC and 90% TDL across classifiers. Therefore, we conclude that *MPTD* improves predictive accuracy over and above balancing. Further, using *MPTD* on balanced data does not significantly improve accuracy beyond *MPTD* on original data while increasing the model complexity.

The *MPTD* contains additional rows with lags of historical customer behavior features and customer churn behavior lags. The *SPTD+lags* also contains the lagged customer behavior information as additional columns to the *SPTD*. While adding the additional lag variables to the *SPTD* (i.e., *SPTD+lags*) improves the average predictive performance, the improvement is not statistically significantly (see Tables 4 and 5). One potential explanation is that while *MPTD* increases the observations per variable, *SPTD+lags* substantially

decreases in the number of observations per variable due to the additional 834 variables and increases the chances of overfitting.

Unlike the traditional *SPTD* training data with all observations referring to the same time period, the *MPTD* training data with customer-month observations allows the environmental conditions to be included in the model. For example, with the *MPTD* logistic regression model we find that increasing monthly deposit interest rates result in lower churn tendency, but as the yield curve slope increases, then customers are more likely to terminate their relationship with the bank. Considering that the majority of the customer portfolios in private banking in the country are invested in fixed deposits, it is not surprising that the significant environmental variables are related to interest rates. The customers may be shifting their investment style and channel as the environment changes resulting in churn for the bank.

4.3. Dynamic churn prediction experiments

Next, we focus on predicting churn several periods ahead. Survival analysis is commonly employed to assess how much longer the customers will continue their relationship with the company. We compare the accuracy of the Cox regression model versus the n independently trained binary classifiers. We use respective versions of logistic regression which proved to be more accurate than decision trees in the next-period churn prediction for this application. We compare the independently trained classifiers with the multinomial logistic regression to infer the impact of the horizon specific and thus more recent and larger training dataset. We compare multinomial logistic regression with ordinal logistic regression to infer the impact of the horizon specific predictors on accuracy.

4.3.1. Factors varied in the experiment

Hence, the computational experiment to investigate the second group of questions raised at the beginning of Section 4 involves an experimental design with three factors and ten replicates. The factors are the classifier (ordinal logistic regression, multinomial logistic regression, n independently trained binary logistic regressions and the Cox regression), the training data (*SPTD*, *SPTD+lags*, *MPTD*) and the prediction lead time (1 to 5 months).

Multinomial logistic regression models the log of the odds ratio as a linear function of the independent variables, as indicated below. The odds ratio refers to the ratio of probability of class i (p_i) to the probability of the reference class k (p_k), where $K = 2$ refers to the binary logistic (*or just logistic*) regression (Hastie et al., 2009). Parameters are estimated using iterative procedures for maximum likelihood, and can become time consuming especially for multiple classes.

$$log\frac{P_i}{P_K} = \beta_{i0} + \beta_i^T x, \quad \text{for } \forall i = 1, \ldots, K-1$$

The ordinal logistic regression is a special case of the multinomial logistic regression, where the outcome classes are ordered, and the log odds of the probability of belonging to class i or lower, i.e. the cumulative probability, versus higher is modeled. Further, the variable coefficients β_i are restricted to be the same for all i, implying that the impact of the independent variables is assumed to be the same for all classes (Hastie et al., 2009). While this is a drawback, the relatively quick parameter estimation for Ordinal Logistic Regression is an advantage when variable selection procedures are used.

We construct the Cox regression model for the tenure variable which denotes the length of the relationship between the customer and the service provider. To incorporate the impact of the time-varying covariates on the hazard function we generate customer-month observations.

The ordinal logistic regression predicts the *Periods-to-churn* which takes values 1–6, *n+* and directly provides the p_δ values. The multinomial logistic regression also predicts *Periods-to-churn* and provides P (*Periods-to-churn* = δ), which we convert to $p_\delta = \sum_{d=1}^{\delta} P(periods - to - chrun = d)$. Independently trained binary logistic regression models predict advance churn labels $W\delta C$, which provides the p_δ values. The approaches with *Periods-to-churn* as the dependent variable, i.e., ordinal and multinomial regression can use the training data up to 200910, while each binary logistic regression $W\delta C$ uses as recent as possible with the particular prediction horizon. In the case of *MPTD* this results in advantages for the independently trained binary logistic regressions approach due to sample size as well as recency.

We implement the binary and ordinal logistic regression models with SAS 9.2 the Logistic procedure and the Cox regression with SAS 9.2 the PHREG procedure. For The logistic and Cox regression models, we select variables with the stepwise method with a significance level of 0.1. In order to make sure that the bootstrap samples contain complete customer data, we create ten bootstrap samples of customers stratified by churn within the training period, and pull the observations that belong to these customers to create the bootstrap datasets. The customers are scored using the latest information available as of March 2010 for churn in the next five months. Due to data limitations, we test predictions for churn with a lead time of up to five time periods starting with April 2010.

4.3.2. Evaluation criteria

Expectations from a dynamic churn predictor are high predictive accuracy that is robust within and across prediction lead times. A certain degradation in the predictive accuracy with the lead time is expected, due to additional uncertainties that may arise during the longer lead time. However, monotonicity of the predictive accuracy with the prediction lead time and lack of a sudden decrease is valuable for the operationalization of the dynamic churn predictors. Evaluation of the accuracy time series along with the costs and benefits of the ensuing retention actions would help decide the maximum prediction lead time.

Therefore, in this section we extend our evaluation criteria to the mean and standard deviation of the AUC and TDL measures for each lead time, and provide a graph to evaluate the degradation of the mean accuracy over time.

4.3.3. Results

We have trained models with the *MPTD*, *SPTD* and *SPTD+lags*, and observed that *MPTD* outperforms *SPTD* for all methods and prediction horizons, in terms of average AUC and TDL, with lower variability. The *SPTD+lags* method performed on worse than the *SPTD* method in average AUC and TDL. Hence, for the comparison of the classifier performance we only consider the results with the *MPTD*. Survival analysis dataset also contains all customers present in the *MPTD* dataset. Tables 6 and 7 contain the average and standard error for the AUC and the TDL measures across the bootstrap samples by classifier and prediction horizon. They also contain the average accuracy across lead times for an overall comparison, as well as the pooled within lead standard deviation.

Figs. 9 and 10 compare the classifiers dynamic churn prediction accuracy measured in AUC and TDL, respectively. The x axis refers to the prediction horizon, and the performance of the independently trained binary logistic regression, ordinal logistic regression, multinomial logistic regression and survival models are represented with blue dashed, long green dash, solid blue and purple lines respectively, while the error lines indicate plus minus one standard error.

As expected, the accuracy of all methods degrades as the prediction lead time increases, both in terms of AUC and TDL. We note that all classifiers provide much better accuracy than a random

Table 6
The average and standard error of the AUC values for dynamic churn prediction. The font of the mean AUC figures indicate significance level of the performance difference between the classifier and benchmark survival analysis as follows: **bold** 0.01, ***bold italic*** 0.05 *italic* 0.1.

Classifier	Mean AUC						Standard error of the mean AUC					
	W1C	W2C	W3C	W4C	W5C	All leads	W1C	W2C	W3C	W4C	W5C	Pooled
Ordinal logistic regression	***0.706***	***0.701***	**0.698**	**0.697**	***0.689***	**0.698**	0.007	0.006	0.004	0.003	0.003	0.002
Multinomial logistic regression	**0.731**	***0.698***	***0.687***	***0.688***	0.681	**0.697**	0.007	0.005	0.006	0.006	0.005	0.003
Independently trained binary logistic regressions	**0.746**	**0.724**	**0.726**	**0.724**	**0.699**	**0.724**	0.006	0.007	0.006	0.004	0.003	0.002
Survival analysis	0.668	0.673	0.657	0.655	0.651	0.661	0.020	0.013	0.015	0.017	0.017	0.007

Table 7
Average and standard error TDL for dynamic churn prediction. The font of the mean TDL figures indicate significance level of the performance difference between the classifier and benchmark survival analysis as follows: **bold** 0.01, ***bold italic*** 0.05 *italic* 0.1.

Classifier	Mean TDL						Standard error of the mean TDL					
	W1C	W2C	W3C	W4C	W5C	All leads	W1C	W2C	W3C	W4C	W5C	Pooled
Ordinal logistic regression	2.969	2.725	2.699	2.628	2.561	2.716	0.182	0.173	0.118	0.107	0.100	0.061
Multinomial logistic regression	3.349	3.136	3.087	2.962	2.822	3.071	0.114	0.085	0.083	0.083	0.078	0.040
Independently trained binary logistic regressions	***3.879***	3.181	***3.321***	*3.126*	3.042	**3.310**	0.206	0.153	0.073	0.106	0.097	0.057
Survival analysis	3.213	2.954	2.811	2.734	2.752	2.893	0.341	0.254	0.250	0.250	0.233	0.119

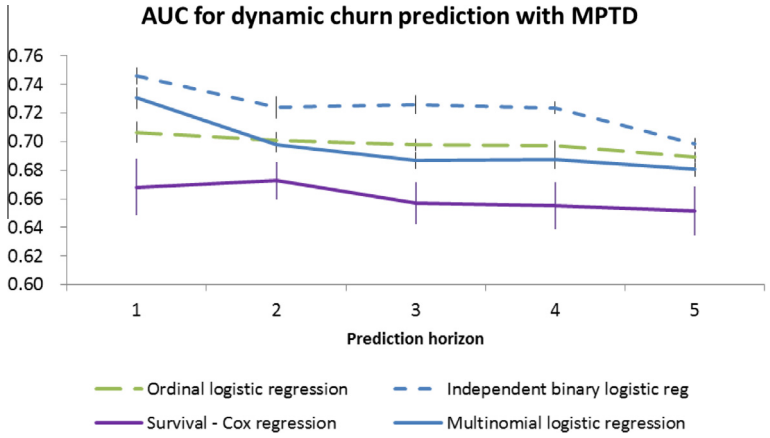

Fig. 9. AUC performance of the alternative classification models and the survival analysis for advance churn prediction.

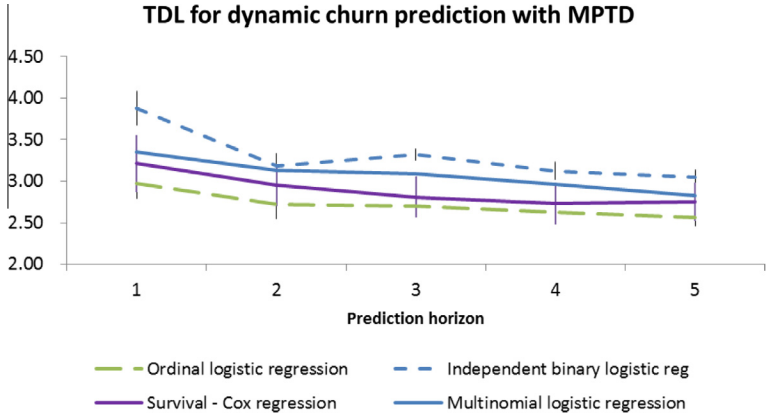

Fig. 10. TDL performance of the alternative classification models and the survival analysis for advance churn prediction.

assignment (AUC > 0.5 and TDL > 1) even at a lead time of five months. However, there are performance differences between classifiers that persist across time periods – we see that the lines in Figs. 9 and 10 do not cross, with one exception. Independently trained binary and multinomial logistic regressions with $W\delta C$ labels perform better than the Survival (Cox regression) model in terms of both the AUC and TDL measures for all lead times. The average differences across lead times are significant for both the AUC and TDL, additionally the lead time specific differences are also significant for the AUC measure.

The variability of both the AUC and TDL measures is much higher for the survival analysis than the classifiers with $W\delta C$ labels. The pooled standard errors of the AUC and TDL measures for survival analysis are 2 to 3.5 times larger than those for the classifiers with $W\delta C$ labels. This indicates that the proposed dynamic churn prediction methods are more consistent in their lead time specific prediction accuracy than the benchmark survival analysis.

Particularly the independent binary logistic regressions with $W\delta C$ labels have superior performance compared to all other methods, as their AUC and TDL performance is consistently better than other methods for any horizon. The independent binary logistic regressions method has significantly better performance than the multinomial logistic regression (except for the $W2C$ TDL). Since the difference between the two methods is only due to the training dataset, we conclude that ability to use the most recently available data and thus increasing the recency and the size of the training data significantly contributes to better prediction accuracy across horizons.

Comparing the ordinal and multinomial logistic regression approaches that differ in terms of the use of horizon-specific predictors, we observe that there is no clear winner. Multinomial logistic regression performs better in terms of TDL, while ordinal logistic regression is better in terms of AUC (except for next period churn).

5. Conclusions, implications, limitations and future research

In this paper we have introduced a dynamic churn prediction framework for generating training data and the independently trained binary classifiers approach for dynamic churn prediction. We empirically evaluated them in terms accuracy of next time period and multi horizon churn predictions using the case of private banking customers of a European bank. The benchmarks were the standard framework that includes only one observation per customer and survival analysis, respectively.

We showed that using multiple training observations per customer from different time periods (*MPTD*) increases the predictive accuracy of churn models, compared with the traditional approach of using the most recent observation per customer, regardless of the classifier or prediction horizon. Top decile lift, which is a very relevant measure for managing the retention actions for the top ten percent of customers who are most likely to leave, increased by 50% for next period churn prediction. The resulting accuracy is significantly higher than the traditional single observation approach even when the data is oversampled with the sophisticated SMOTE method to balance the churn and non-churn class observation counts.

The accuracy improvement observed in churn prediction has the following theoretical justifications: *MPTD* increases the training data sample size and decreases the absolute rarity, i.e., the number of the rare (churn) events compared with the standard approach, further it contains environmental variation. Thus, it allows the model to generalize across different time periods and identify insights regarding impact of the environmental conditions which may include economic conditions, competition, or companywide initiatives on churn. For example, in the private banking study the model indicated that lower monthly deposit interest rates increase the churn, which made sense given that a major portion of the customer assets are in fixed deposits.

Another contributor to the improvement in accuracy is the increase in the sampling density. We have shown that the alternative approach of using lags of customer behavior features as predictors while keeping one observation per customer decreases the sample size per parameter and results in the problem known as the curse of dimensionality. As a result, this approach does not result in an improvement over the standard framework: and in fact performs worse multi-period horizon models with more parameters.

For dynamic churn prediction, we have shown that the proposed independently trained binary classifiers with advanced churn labels approach has significantly superior prediction accuracy for all horizons compared to survival analysis (Cox regression) that is commonly used for this purpose.

We have shown that the proposed dynamic churn prediction approach performs better (a) because it is able to use the most recent data – unlike multinomial or ordinal classifiers that need to consider all classes together, and (b) it allows horizon specific predictors – unlike ordinal classifiers or Cox regression, in addition to using the MPTD framework with the above specified advantages. Further, while the above result has been obtained with logistic regression, the approach can be used with any standard classifier.

An important managerial advantage of the lead-time specific ranking of customers in terms of churn probability is that it facilitates targeting retention actions across time and customers. Further, horizon specific predictors allow discovery of triggers or indicators of churn within a specified horizon, which allows the company to modify the products, services or processes to alleviate the impact. These advantages are not available with the Cox regression approach which assumes that the hazard ratio of one customer versus another remains constant over time.

A limitation of the proposed framework is the potential lack of independence introduced by multiple observations of the same customer may bias the parameter estimates and/or artificially increase the significance of the parameters with methods that assume independent and identically distributed error terms, such as logistic regression. On the other hand, oversampling methods, which are used frequently with churn studies, similarly endanger the independence assumption.

Another limitation of this research is that the empirical results are based on one particular case study.

Future research can explore the improvement due to MPTD as a function of the relative and absolute rarity in the data, and compare it with other methods of addressing the rare event problem. This will help establish guidelines for the circumstances under which MPTD offers a significant improvement in accuracy. The number of observations per customer to include in the MPTD and whether they should be sequential or selected based on a particular criterion are other design questions about the implementation of the MPTD that may provide further improvement in predictive performance and/or decrease the computational complexity.

Another research dimension involves use of the dynamic churn predictions for optimal allocation of retention resources over time and across customers. Here, the multi-period-ahead churn prediction models can include the retention actions as predictors. Business relevance of the dynamic churn prediction will depend on the accuracy profile with respect to lead time. Therefore, characterization of the factors that drive the degradation in prediction accuracy with lead time would also be a useful research dimension.

Acknowledgments

The authors would like to thank Hamdi Özçelik of YKB for defining the business need and making the industry – university collaboration possible. We would also like to thank former MS student Kübra Yaman for data processing. Further, we acknowledge that this research was supported by the Scientific and Technological Research Council of Turkey, Project Number TEYDEB 1501-3100085, and that Tübitak also provided the scholarship for Umut Arıtürk's MS studies. Finally, the authors would like to thank the anonymous reviewers for their valuable comments and suggestions which greatly improved the paper.

References

Ballings, M., & Van den Poel, D. (2012). Customer event history for churn prediction: How long is long enough? *Expert Systems with Applications, 39*(18), 13517–13522.

Bellman, R. (1960). Directions of mathematical research in nonlinear circuit theory. *IRE Transactions on Circuit Theory, 7*(4), 542–553.

Bijmolt, T. H. A., Leeflang, P. S. H., Block, F., Eisenbeiss, M., Hardie, B. G. S., Lemmens, A., et al. (2010). Analytics for customer engagement. *Journal of Service Research, 13*(3), 341–356. http://dx.doi.org/10.1177/1094670510375603.

Braun, M., & Schweidel, D. A. (2011). Modeling customer lifetimes with multiple causes of churn. *Marketing Science, 30*(5), 881–902.

Buckinx, W., & Van den Poel, D. (2005). Customer base analysis: Partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. *European Journal of Operational Research, 164*(1), 252–268. http://dx.doi.org/10.1016/j.ejor.2003.12.010.

Burez, J., & Van den Poel, D. (2008). Separating financial from commercial customer churn: A modeling step towards resolving the conflict between the sales and credit department. *Expert Systems with Applications, 35*(1–2), 497–514. http://dx.doi.org/10.1016/j.eswa.2007.07.036.

Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications, 36*(3, Part 1), 4626–4636. http://dx.doi.org/10.1016/j.eswa.2008.05.027.

Cao, L. (2010). In-depth behavior understanding and use: The behavior informatics approach. *Information Sciences, 180*(17), 3067–3085. http://dx.doi.org/10.1016/j.ins.2010.03.025.

Chawla, N. V., Bowyer, K. W., & Hall, L. O. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357.

Chu, B. H., Tsai, M. S., & Ho, C. S. (2007). Toward a hybrid data mining model for customer retention. *Knowledge-Based Systems, 20*(8), 703–718. http://dx.doi.org/10.1016/j.knosys.2006.10.003.

Coussement, K., & Van den Poel, D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert Systems with Applications, 34*(1), 313–327. http://dx.doi.org/10.1016/j.eswa.2006.09.038.

Crone, S. F., Lessmann, S., & Stahlbock, R. (2006). The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing. *European Journal of Operational Research, 173*(3), 781–800. http://dx.doi.org/10.1016/j.ejor.2005.07.023.

Dawes, J., & Swailes, S. (1999). Retention sans frontiers: Issues for financial service retailers. *International Journal of Bank Marketing, 17*(1), 36–43.

De Bock, K. W., & Poel, D. V. d. (2011). An empirical evaluation of rotation-based ensemble classifiers for customer churn prediction. *Expert Systems with Applications, 38*(10), 12293–12301.

Fader, P. S., & Hardie, B. G. (2010). Customer-base valuation in a contractual setting: The perils of ignoring heterogeneity. *Marketing Science, 29*(1), 85–93.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874.

Frank, E., & Hall, M. (2001). A simple approach to ordinal classification. In *Paper presented at the proceedings of the 12th European conference on machine learning*.

Gebregziabher, M., Egede, L., Gilbert, G. E., Hunt, K., Nietert, P. J., & Mauldin, P. (2012). Fitting parametric random effects models in very large data sets with application to VHA national data. *BMC Medical Research Methodology, 12*(1), 163.

Greene, W. H. (2012). *Econometric analysis* (7th ed.). Boston: Prentice Hall.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York, NY: Springer.

Huang, B. Q., Kechadi, T.-M., Buckley, B., Kiernan, G., Keogh, E., & Rashid, T. (2010). A new feature set with new window techniques for customer churn prediction in land-line telecommunications. *Expert Systems with Applications, 37*(5), 3657–3665.

Jamal, Z., & Bucklin, R. E. (2006). Improving the diagnosis and prediction of customer churn: A heterogeneous hazard modeling approach. *Journal of Interactive Marketing, 20*(3–4), 16–29. http://dx.doi.org/10.1002/dir.20064.

Kamakura, W., Mela, C. F., Ansari, A., Bodapati, A., Fader, P., Iyengar, R., et al. (2005). Choice models and customer relationship management. *Marketing Letters, 16*(3–4), 279–291.

Karahoca, A., & Karahoca, D. (2011). GSM churn management by using fuzzy c-means clustering and adaptive neuro fuzzy inference system. *Expert Systems with Applications, 38*(3), 1814–1822.

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis, 9*(2), 137–163.

Kleinbaum, D. G., & Klein, M. (2005). *Survival analysis: A self-learning text* (2nd ed.). New York, NY: Springer.

Kumar, A. D., & Ravi, V. (2008). Predicting credit card customer churn in banks using data mining. *International Journal of Data Analysis Techniques and Strategies, 1*(1), 4–28 [10.1504/IJDATS.2008.02002].

Larivière, B., & Van den Poel, D. (2004). Investigating the role of product features in preventing customer churn, by using survival analysis and choice modeling: The case of financial services. *Expert Systems with Applications, 27*(2), 277–285. http://dx.doi.org/10.1016/j.eswa.2004.02.002.

Lassar, W. M., Manolis, C., & Winsor, R. D. (2000). Service quality perspectives and satisfaction in private banking. *Journal of Services Marketing, 14*(3), 244–271.

Lee, Y.-H., Wei, C.-P., Cheng, T.-H., & Yang, C.-T. (2012). Nearest-neighbor-based approach to time-series classification. *Decision Support Systems, 53*(1), 207–217.

Lemmens, A., & Croux, C. (2006). Bagging and boosting classification trees to predict churn. *Journal of Marketing Research, 43*(2), 276–286. http://dx.doi.org/10.1509/jmkr.43.2.276.

Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics, 14*(1), 106.

Mavri, M., & Ioannou, G. (2008). Customer switching behaviour in Greek banking services using survival analysis. *Managerial Finance, 34*(3), 186–197.

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

Orsenigo, C., & Vercellis, C. (2010). Combining discrete SVM and fixed cardinality warping distances for multivariate time series classification. *Pattern Recognition, 43*(11), 3787–3794. http://dx.doi.org/10.1016/j.patcog.2010.06.005.

Owczarczuk, M. (2010). Churn models for prepaid customers in the cellular telecommunication industry using large data marts. *Expert Systems with Applications, 37*(6), 4710–4712.

Pennell, M. L., & Dunson, D. B. (2007). Fitting semiparametric random effects models to large data sets. *Biostatistics, 8*(4), 821–834.

Perlich, C., Provost, F., & Simonoff, J. S. (2003). Tree induction vs. logistic regression: A learning-curve analysis. *The Journal of Machine Learning Research, 4*, 211–255.

Prinzie, A., & Van den Poel, D. (2006). Incorporating sequential information into traditional classification models by using an element/position-sensitive SAM. *Decision Support Systems, 42*(2), 508–526. http://dx.doi.org/10.1016/j.dss.2005.02.004.

Reichheld, F. F. (1996). Learning from customer defections. *Harvard Business Review, 74*, 56–69.

Reichheld, F. F., & Kenny, D. W. (1990). The hidden advantages of customer retention. *Journal of Retail Banking, 12*(4), 19–23.

Stahl, H. K., Matzler, K., & Hinterhuber, H. H. (2003). Linking customer lifetime value with shareholder value. *Industrial Marketing Management, 32*(4), 267–279.

Stock, J. H., & Watson, M. W. (1989). New indices of coincident and leading economic indicators. In *Paper presented at the NBER macroeconomics annual*.

Storbacka, K. (1997). Segmentation based on customer profitability—retrospective analysis of retail bank customer bases. *Journal of Marketing Management, 13*(5), 479–492.

Tsai, C.-F., & Chen, M.-Y. (2010). Variable selection by association rules for customer churn prediction of multimedia on demand. *Expert Systems with Applications, 37*(3), 2006–2015.

Van den Poel, D., & Lariviere, B. (2004). Customer attrition analysis for financial services using proportional hazard models. *European Journal of Operational Research, 157*(1), 196–217. http://dx.doi.org/10.1016/S0377-2217(03)00069-9.

Venkatesan, R., & Kumar, V. (2004). A customer lifetime value framework for customer selection and resource allocation strategy. *Journal of Marketing, 68*(4), 106–125.

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research, 218*(1), 211–229. http://dx.doi.org/10.1016/j.ejor.2011.09.031.

Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems with Applications, 38*(3), 2354–2364. http://dx.doi.org/10.1016/j.eswa.2010.08.023.

Weiss, G. M. (2004). Mining with rarity: A unifying framework. *SIGKDD Exploration Newsletter, 6*(1), 7–19. http://dx.doi.org/10.1145/1007730.1007734.

Widmer, G., & Kubat, M. (1996). Learning in the presence of concept drift and hidden contexts. *Machine Learning, 23*(1), 69–101. http://dx.doi.org/10.1023/a:1018046501280.

Xiao, J., Xie, L., He, C., & Jiang, X. (2012). Dynamic classifier ensemble model for customer classification with imbalanced class distribution. *Expert Systems with Applications, 39*(3), 3668–3675.

Xie, Y., Li, X., Ngai, E. W. T., & Ying, W. (2009). Customer churn prediction using improved balanced random forests. *Expert Systems with Applications, 36*(3, Part 1), 5445–5449. http://dx.doi.org/10.1016/j.eswa.2008.06.121.

Yavas, U., Benkenstein, M., & Stuhldreier, U. (2004). Relationships between service quality and behavioral outcomes: A study of private bank customers in Germany. *International Journal of Bank Marketing, 22*(2), 144–157.