Modelling Clustered Survival Data with Cured Fraction

Angela D. Nalica[1], Iris Ivy M. Gauran[2], Erniel B. Barrios[3]
[1,2,3] University of the Philippines, Diliman
Corresponding author: Angela D. Nalica, E-mail: angela_r_delapaz@yahoo.com

In modelling lifetime data, standard parametric theory assumes that all observations will eventually experience the event of interest if they are monitored for a very long period. While every unit starts as susceptible to the event of interest, a fraction of observations may switch into a non-susceptible group. A mixture cured fraction model with covariates is modified to incorporate random clustering effect to characterize the switch mechanism. Simulation studies and telecommunications data show that cured fraction models with random clustering effect perform better than their parametric counterpart in terms of predictive ability.

Key Words: Mixture Cured Fraction Models, Random Clustering Effect, Right-censored Lifetime Data

1. Introduction

Customer care units of service companies aim to cultivate satisfied customers who will eventually develop loyalty and subsequent retention as a customer. Companies are becoming more and more aware of the crucial importance to fully exploit their existing consumer database to serve as inputs for marketing planning and strategies. A lot of information can be derived from these data such as predicting customer behavior, customer loyalty and customer satisfaction among others.

Customer relationship management (CRM) plays a critical role in establishing customer-company relationship that is mutually beneficial. Chalmeta (2005) defines CRM as a set of business, marketing and communication strategies and technological infrastructures designed with the aim of building lasting relationship with the customers, which involves identifying, understanding and meeting their needs. Coltman (2007) illustrates that a superior CRM capability can create positional advantage and subsequent improved performance. Further, it is shown that to be most successful, CRM programs should focus on latent or unarticulated customer needs that emphasize a proactive rather than a reactive market orientation.

There is now a shift in focus from building a large client base to keep the existing ones since markets are becoming saturated and competitive pressure is becoming intense. Extant literature shows that it is much cheaper to retain old clients than to gain new ones. Thus, attention has been directed on detecting customers that are becoming less loyal, also called churners.

The churn model presented in this study is based on the theory of survival analysis. It is assumed that if complete follow-up were possible for all elements of the study population, then each element would eventually experience the event of interest. However, the focus is shifted towards modelling survival data wherein a substantial proportion of the individuals does not experience the event at the end of the observation period. These individuals are classified as long term survivors and are viewed as "cured" in the sense that even after an extended follow-up period, the event of interest may not be

observed. Failing to account for such cured subjects would lead to incorrect inferences (Corbiere and Joly, (2007)).

When right-censoring is a possibility, the survival function is usually modelled by a mixture model (Boag, (1949) and Farewell, (1982)). This approach allows the simultaneous estimation of the occurrence of the event of interest (incidence) and time of occurrence, given that it can occur (latency). Let $U = 1$ indicate that an individual is susceptible while $U = 0$ denote that the individual is non-susceptible to the event of interest. Define T to be a non-negative random variable denoting the failure time of interest, defined only when $U = 1$. The mixture model is given by

$$S(t \mid x, z) = \pi(z) S(t \mid U = 1, x) + 1 - \pi(z) \qquad \text{Equation (1)}$$

where $S(t \mid x, z)$ is the unconditional survival function of T for the entire population, $S(t \mid U = 1, x) = P(T > t \mid U = 1, x)$ is the survival (latency) function for susceptible individuals given a covariate vector $\boldsymbol{x}$ and $\pi(z) = P(U = 1 \mid z)$ is the probability of being susceptible given a covariate vector $\mathbf{z}$, which may include the same covariates as $\boldsymbol{x}$. The survival function of cured individuals can be set to one for all finite values of t because they will never experience the event of interest. It should be noted that $S(t \mid x, z) \to 1 - \pi(z)$ as $t \to \infty$. If all the individuals are susceptible to the event of interest, $\pi(z_i) = 1$ for all $\mathbf{z_i}$. This means that when there is no cured fraction then the mixture cure model reduces to the standard survival model.

The conditional latency distribution $S(t \mid U = 1)$ can take the form of parametric distributions. Among the parametric models, Weibull distribution is commonly used to model survival data. After reparametrization (Gamel, et. al., 2000), this distribution can be expressed as:

$$S(t \mid U = 1) = \exp\left\{ -\exp\left(\frac{\log(t - \mu)}{\sigma} \right) \right\} \qquad \text{Equation (2)}$$

In proportional hazards models, $S(t \mid U = 1, x) = S_0(t \mid U = 1)^{\exp(\beta' x)}$ where $S_0(t \mid U = 1)$ is the baseline hazard function. If $S_0(t \mid U = 1)$ is left arbitrary, the model is defined as the Cox's Proportional Hazards (PH) Mixture Cure Model (Cox, 1972).

The data utilized in this study are represented as clustered survival information to take into account the homogeneity of the individuals belonging in the same cluster. However, in most of the survival data models described in the literature, heterogeneities between individuals have been taken into account only in the form of the observable covariates. Thus, to be able to model clustered survival data, the Cox PH Model was modified to include a cluster-specific random component.

To establish a direct comparison, two different predictive churn models are considered in this study: the modified Cox PH model and the Weibull model. The predictive abilities of these models are compared using simulation and a telecommunications data. Comparison is done through the relative difference in median absolute percentage error. In the context of the telecommunications data set, the definition of prepaid churn is based on a number of successive months with zero top-up. Due to the constraints in data available, definition of churn is restricted to having zero top-up in 6 consecutive months.

2. Results and Discussion

Simulation Study

To compare the parametric model and the proposed model, a simulation study was performed. The simulation boundaries include (1) the percentage of observations classified as right-censored and (2) presence of misspecification in the model. We considered 5%, 10%, and 20% censoring in assessing the predictive ability of the model. We also assessed the effect of misspecification error into the model.

Table 1. Relative Difference in Mean of MAPE based on the Simulation Scenarios

Simulation Scenarios	% Censoring	Relative Difference in Mean of MAPE
With Misspecification	**5%**	4.3411
	10%	4.9330
	20%	4.8184
Without Misspecification	**5%**	46.1603
	10%	46.6819
	20%	47.1401

In Table 1, the cured model performs better than the parametric model especially if misspecification is present in the model. It can also be noted that as the percentage of the right-censored observations increases, the corresponding relative difference in the mean MAPE also increases. This indicates that if a sizable amount of observations have been subject to an intervening variable or a curing situation, then the predictive ability of the model will be affected.

Moreover, in the ideal scenario where contamination is not present, the model produces relatively lower MAPE compared to the scenario wherein misspecification error is introduced into the model. This explains why data sets with limited information on covariates and with some degree of contamination would greatly affect the predictive ability of the two models.

Application to Telecommunications Data

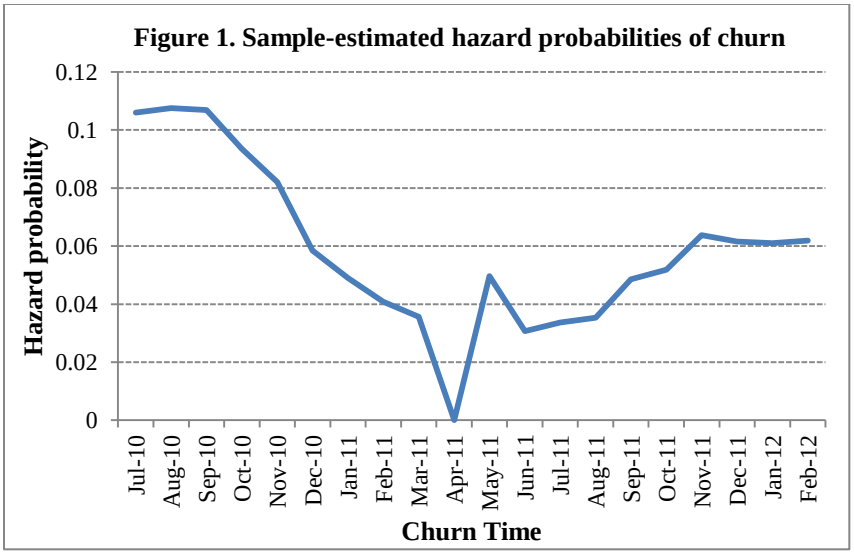
Based on the activation date of the subscribers, there are two telecommunications data sets utilized in this paper. First, we have monthly transactional details of prepaid customers of a telecommunications company observed in (a) January 2010 to February 2012 and (b) January 2011 to February 2012. The data is limited only on billing information (amount of and frequency of top-up). There are 38,000 subscribers whose activation date started on January 2010 and around 42,000 subscribers with activation date starting on January 2011. The churn models are applied on these two data sets to characterize the effect of length of history on the amount of contamination or misspecification.

According to the amount of top-up, ten clusters were identified. This number is based on the result of the study of Arceneaux and Nickerson (2009) which highlights that an increase in the number of clusters leads to an increase in efficiency. The first cluster includes all subscribers whose total amount of top-up is less than twenty pesos per month. This implies that the subscribers belonging in this cluster are the ones that bought the Subscriber Identity Module (SIM) card, used all the freebies and opted to recharge on a short duration only. Almost all subscribers belonging in this cluster were the earliest churners. Meanwhile, the tenth cluster is characterized by subscribers whose total amount of top-up is more than PhP 250 per month. Upon consideration of the wide range of promos and freebies that this network offers, this amount is classified high already.

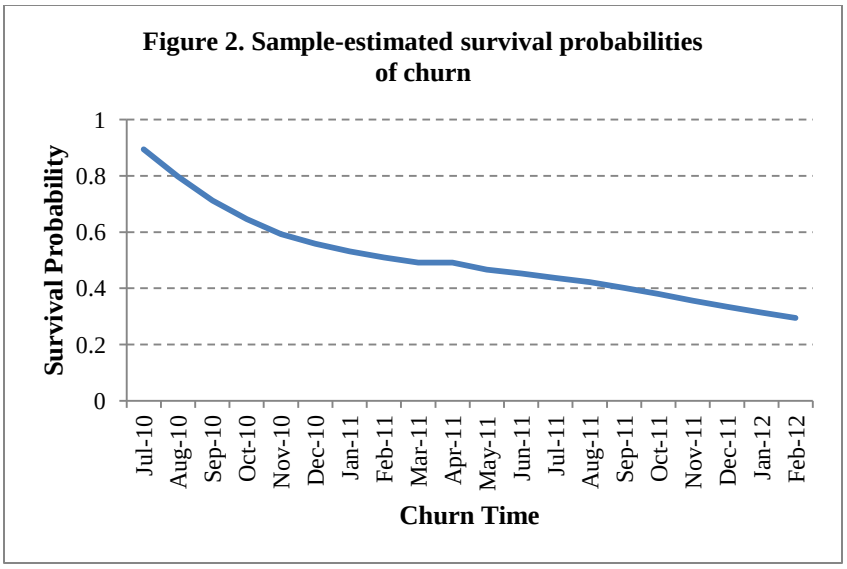
Estimation of Hazard Probabilities

The sample estimates of hazard probabilities, also known as the marginal hazard probability estimates, are calculated from the event history data. The estimated hazard probability for time *j* is the number of events that are observed to occur in time period *j* divided by the total number of subjects at risk in time period *j*. "Subjects at risk" in this context refers to all those observations in period *j* that are not censored during period *j*.

The 20 sample hazard probabilities are plotted by month as shown in Figure 1, suggesting that the marginal hazard function is decreasing from July 2010 to April 2011 but there is a slightly increasing pattern from April 2011 to February 2012 wherein the fluctuations range from 0.03 to 0.06.



The proportions of subscribers surviving through each month (i.e., the survival probabilities) can be estimated directly from the estimated hazard probabilities. Figure 2 displays the plot of the estimated survival probabilities by month. There is an increase in the proportion of the total subscribers churned over time with almost 70% churned by the end of the 20th month.



Comparison of the Parametric Model and the Cured Fraction Model

The parametric method with covariate and the cure mixture model with clustering variable are applied on two data sets. Customers are grouped according to the mean amount of top-up incurred in the study period. Although the models include only one covariate, four different indicators were used to determine if there will be differences on the performance of the models.

Table 2. Relative Difference in MAPE of the Parametric and Cured Fraction Model

Covariate	Relative Difference in MAPE	
	14-month data	26-month data
Average frequency of recharge	9.0872	13.4647
Median frequency of recharge	4.0689	18.2262
Maximum frequency of recharge	23.2084	15.6208
Total frequency of recharge	9.0872	13.4647

Regardless of the length of usage history of a subscriber, the proposed method has smaller MAPE compared to the parametric method. Except for the covariate maximum frequency of recharge, 26-month data shows larger difference in MAPE between the two models in favor of the proposed model. However, the resulting MAPEs for longer history are larger which can be explained by the lack of access to other possible covariates or the existence of some random shocks on certain period. Data on number of inbound/outbound calls as well as the minutes of use among other variables may help better explain the churning behavior.

3. Conclusions

The simulation study confirms that clustered survival data can be better characterized by the proposed model. In a perfect scenario (i.e. there is no misspecification error), the modified cured model is superior than its parametric counterpart, relatively robust to varying censoring percentages. With misspecification error, while predictive ability declines, the proposed model still outperforms the parametric model.

The cured model with random clustering effect also performed better (predictive ability) than the parametric counterpart in the application to telecommunications data. Regardless of the covariate used, the proposed model still exhibits lower MAPE.

References

Arceneaux, K. and Nickerson, D. (2009), "Modeling Certainty with Clustered Data: A comparison of Methods," *Political Analysis*, **17**, 177-190.

Boag, J. W. (1949), "Maximum Likelihood Estimates of the Proportion of patients cured by cancer therapy," *Journal of the Royal Statistical Society* **11**, 15-44.

Caroni C. and Economou P. (2012), "A Hidden Competing Risk Model for Censored Observations," *Brazilian Journal of Probability and Statistics*.

Chalmeta, R., (2006), "Methodology for Customer Relationship Management," *The Journal of Systems and Software*, 79:1015-1024.

Coltman (2007), "Why build a customer relationship management capability?," *Journal of Strategic Information Systems*, **16**, 301-320.

Corbiere, F. and Joly, P. (2007), "A SAS Macro for Parametric and Semiparametric Mixture Cure Models," *Computer Methods and Programs in Biomedicine*, **85(2)**: 173-180.

Cox, D. R. (1972), "Regression models and life–tables (with discussion)," *Journal of the Royal Statistical Society: Series B* **34**, 187–220.

Farewell, V. T. (1982), "Mixture Models in Survival Analysis: Are they worth the risk?," *Canadian Journal of Statistics*, **14**, 257-262.

Gamel, J. W., Weller, E. A., Wesley, M. N., Feuer, E. J. (2000), "Parametric Cure Models of Relative and Cause-Specific Survival for Group Survival Times," *Computer Methods and Programs in Biomedicine*, **61**, 99-110.

Glady, N., Baesens, B., Croux, C., (2009), "Modeling Churn Using Customer Lifetime Value," *European Journal of Operational Research*, **197**: 402-411.