Real-Time Prediction of the Probability of Abandonment in Call Centers

M.Sc. Research Proposal

Rony Albert GHEBALI

Advisor: Professor Avishai Mandelbaum

The William Davidson Faculty of Industrial Engineering and Management

Technion - Israel Institute of Technology

Tammuz 18, 5771-**Haifa, Israel-** July 20, 2011

Contents

Introduction

This research proposal deals with predicting the probability of abandonment of a call center customer who has a call history, that is, who already called the call center at least once. In Section 1, we first provide the reader with a brief review of a few concepts in survival analysis. Section 2 presents the SEE laboratory and describes the data we shall be using. A definition of customer patience is provided there with an introduction of the SEEStat survival kit. Then, a detailed prediction procedure is provided in Section 3.

1 A Few Basic Notions in Survival Analysis

The goal of this section is to present the theoretical framework within which we propose to carry out our research.

1.1 Survival Function and Hazard Rate

As a reminder, let T be a random variable, describing the time of occurrence of an event. Assume T has a cumulative distribution $F(t) = \int_{-\infty}^{t} f(u)du$, where f is the density function. Then its survival function is defined as

$$S(t) = 1 - F(t) = \int_{t}^{\infty} f(u)du,\ t \geq 0.$$

The above survival function clearly characterizes the variable's distribution (see the first chapter of [3]). The probability of "having an event at time t given that it did not occur up to time t", is proportional to the *hazard* rate function as follows:

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{d\ln(S(t))}{dt}.$$

Therefore, for a continuous positive random variable, one can write

$$S(t) = e^{-\int_{0}^{t} \lambda(u)du},\ t \geq 0.$$

The hazard rate function is a central object in describing failure time distributions and how the likelihood of experiencing the event changes with time. There are many possible shapes for hazard rates (see Section 2 of this proposal) and the only required constraint is that it should be nonnegative.

1.2 Censored Data

What gives its name to "survival analysis" is that a significant number of applications deal with "time-to-death" studies, mostly in medicine. These problems arise in many other fields, such as biology, economics or, as in our study, call centers. Studies may have different definitions of failure times, that is to say, different definitions for "death" but their data are usually composed of recorded times until a "failure" occurs.

In many applications, the value of the time-to-event is only partially known. In our work we consider only right censoring data, namely, the time-to-event is above the observed value but it is unknown by how much. Specifically, for the call center data, we study the customers' patience and, therefore, we look at the time until abandonment: for example, if a customer abandoned after 2 minutes of waiting, we can say that his patience (the time he is willing to wait) is exactly 2 minutes. However, if this customer was answered after 2 minutes of waiting, the only information we have on his/her patience is that it is at least of 2 minutes. Therefore, we have a left bounded interval for patience and, in our example, the data is said to be *right censored*. For the call center data we use in the present research, censoring rates can reach 99%.

Formally, let T^0 be the failure time, C the censoring time, and $T = \min(T^0, C)$ the observed time. Define $\delta = I(T^0 < C)$, which is the censoring indicator function. Then, our data consists of pairs (T, δ), each associated with a customer.

1.3 The Proportional Hazard Model

The proportional hazard regression model is the most popular regression model in survival analysis. The Cox hazard regression model is given by:

$$\lambda(t|Z) = \lambda_0(t) \cdot c(\beta^T Z)$$

where $\lambda_0(t)$ is an unspecified baseline hazard function, β is a vector of unknown regression coefficients, Z is a vector of covariates and c is a known positive function, which is usually exponential. The above semi-parametric regression model is called proportional for the following reason: Assume that we want to compare the hazards of two individuals with covariate vectors Z_1 and Z_2, respectively, under the exponential function for c, thus:

$$\frac{\lambda(t|Z_1)}{\lambda(t|Z_2)} = e^{\beta^T(Z_1 - Z_2)}$$

where $e^{\beta^T(Z_1-Z_2)}$ is called the"relative risk" and is independent of t.

1.4 The Frailty Model

The notion of frailty provides a convenient way to characterize an unobserved heterogeneity into survival models. The term frailty itself was introduced by Vaupel et al. (1979) in univariate survival models and the model was substantially promoted by its application to multivariate

survival data in a seminal paper by Clayton (1978) on chronic disease incidence in families. The frailty is a random multiplicative effect on the hazard rate, and more specifically, on the hazard function of the semi-parametric Cox regression model. The main idea of frailty models in clustered data is that the random cluster-specific variate representing the unobserved common risk is shared by cluster members.

Estimation in the frailty model has received much attention under various frailty distributions, including gamma, positive stable, inverse Gaussian, compound Poisson and log-normal. Hougaard (2000) provides a comprehensive review of the properties of the various frailty distributions. Specifically, a shared frailty model applied to the proportional hazards model would be of the form:

$$\lambda(t|Z,\omega) = \omega \cdot \lambda_0^c(t) \cdot e^{\gamma^T Z}$$

where ω is the unobservable frailty variate and its distribution is usually assumed to be known, up to some parameters, and $\lambda(t|Z,\omega)$ is the conditional hazard function. Then, the main assumption in frailty models is that given the frailty variate and the covariate vector, the cluster members' failure times are independent. As far as our data is concerned, each customer of the call center can call several times, therefore our"clusters" will be composed of each customer calling at least once.

2 (Im)patience in Call Centers

2.1 The SEE Laboratory, and its Repositories.

The SEE Laboratory (Service Enterprise Engineering) was established in 2007, within the Faculty of Industrial Engineering and Management, at the Technion. The goal of SEE is the development of engineering and scientific principles that support modeling, design and management of Service Enterprises, for example, financial services (banking, insurance), health services (hospitals, clinics), government and tele-services (telephone, internet). Presently, SEE's main activity is designing, maintaining and analyzing an accessible repository of resources and data from telephone call-centers.

One of the SEE's central projects is the DATA MOdel for Call Center Analysis (DataMOCCA). The DataMOCCA Project is an initiative of researchers from the Technion and the Wharton School - University of Pennsylvania. The goal of the project has been to collect, pre-process, organize and analyze data from Telephone Call/Contact Centers. The raw data constitute call-by-call records, over at least a one-year duration from active Call Centers. Over the past 10 years, the SEE laboratory has received several important data sets from both the US and Israeli Banks. The structure of the data is described in the different volumes of the DataMOCCA project, many of which are available in SEE's website http://ie.technion.ac.il/Labs/Serveng/. Among those are patience data consisting of customers' waiting times until service calls were either completed or abandoned. These data have been very useful in Survival Analysis of customers' (im)patience, while waiting to be served by a telephone agent (see [5]).

We shall be using SEE data in our research. The data were provided by Polina Khudyakov who used it for her Ph.D. research on the statistical analysis of Call Center Data ([6]). Part of her work was dedicated to the analysis of customer patience, and it used two files of 50,000 and 90,000 customers, respectively. These were kindly transferred to us via the SEE laboratory. The files mainly consist of listing the waiting time of each customer (several in case of repeated calls) and indicators whether the customer was served or abandoned.

Our data set is from a financial company's call center. This is the data that was used in Khudyakov's PhD. Research on customer patience analysis; her work is our starting point. A customer can get service through an IVR or through an agent. At any stage of the procedure (see [5]), the customer can either leave the system or wait in a queue until served (unless immediately served). After being served, the customer hangs up the phone and possibly calls again for further services. Each call registers the following data:

- ID number of the customer.
- Type of customer (priority level given by the system, according to some specific criteria).
- Beginning of each "call segment" (time of the beginning of the call. See [5]).
- Duration of the call.
- Type of requested service.
- Classification of the call, according to its termination: abandonment, completed service or system error.

Each customer can go through a call "series", that is a series of repeated calls. If the time that elapses between two consecutive calls is less than three days, we assume that these calls belong to the same series; otherwise, we assume that these calls belong to two different series. The idea here is that a given customer accumulates experience through repeated calls, under the condition that the calls of the same series are sufficiently close to each other. The data enables patience estimation within the framework of censored data and survival analysis. Abandonment is considered here as "failure" (event), and the observed time up to abandonment (namely, patience or impatience) is considered as censored if the customer is answered (receives service).

2.2 The Notion of (Im)Patience

The (im)patience of a customer, denoted τ is defined as the time that a customer is willing to wait for service. If the customer's patience was infinite, the time he would wait until served would be exactly the time he is required to wait in order to be served; denote this time by V. Then the time a customer effectively waits, denoted W, is the following:

$$W = min(V, \tau).$$

Our definition of patience (time willing to wait) involves both psychological and operational factors. Psychological factors are very difficult to describe and quantify since they must be shared by the population under study. In contrast, operational factors can be detected clearly when analyzing the time to failure (abandon). However, we shall discover that some data sets can also reveal psychological characteristics when studying time to abandonment.

2.3 The SEEStat Interface: A Prediction Tool

The SEEStat interface was developed by SEELab researchers. Its goal is to provide the research community with a useful tool for the analysis of call center data. It is frequently updated with new features, among those a full survival analysis kit in which survival curves, hazard rates and their smoothing curves can be created within a few clicks. SEEStat is thus a useful tool to create hazard rate curves and infer customers' abandonment probability, depending on the specific time within a day, a week, a month or even a year. Several variables can be set to be failure or censoring time. If the answer time is set to be the censor time and the time to abandonment is set to be the failure time, the plotted hazard rates correspond to the (im)patience. Figure 1 shows an example of such a smoothed hazard rate.

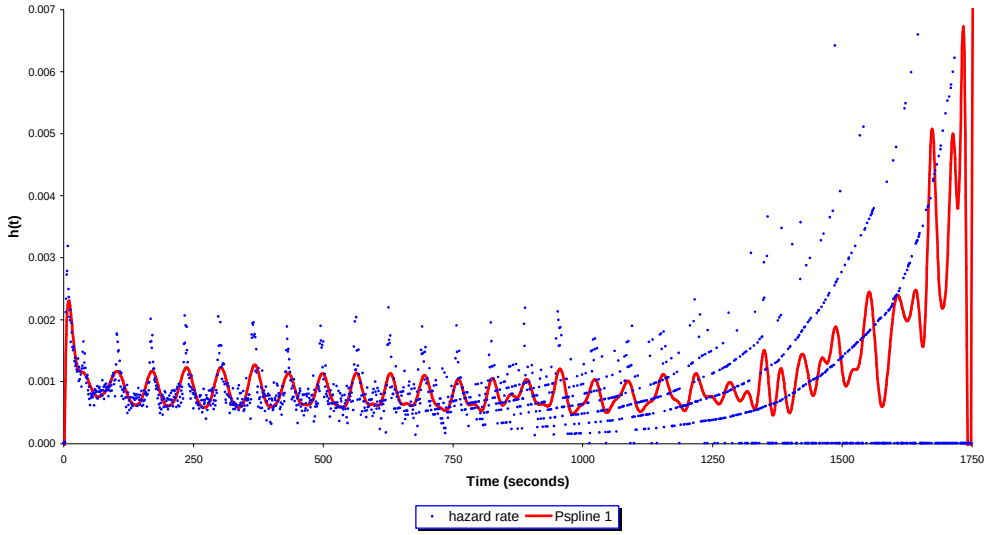

Figure 1: An example of SEEStat Graphical output in Survival Analysis, where we created the hazard rate of (im)patience (time willing to wait)

2.3.1 Relevant Outputs

Despite the fact that psychological factors are difficult to quantify, hazard rates do reveal some effects on the probability of abandonment (or the probability to be served, depending on how the failure time is defined).

In Figure 2, two smoothed curves can be seen. Each is plotted for a different failure time definition: abandonment as failure time (red curve) and service as failure time (blue). Defining the failure time as the service time allows one to visualize protocol schedules. Here, regularly-spaced peaks can be seen when observing the probability to be answered (blue curve) which reveals the protocol time schedules. The interesting observation here lies in the fact that the two curves' peaks match only every two red peaks. According to the information we have on the system protocols, it appears that announcements are made regularly which explains the red peaks: the probability to abandon increases significantly and noticeably following the

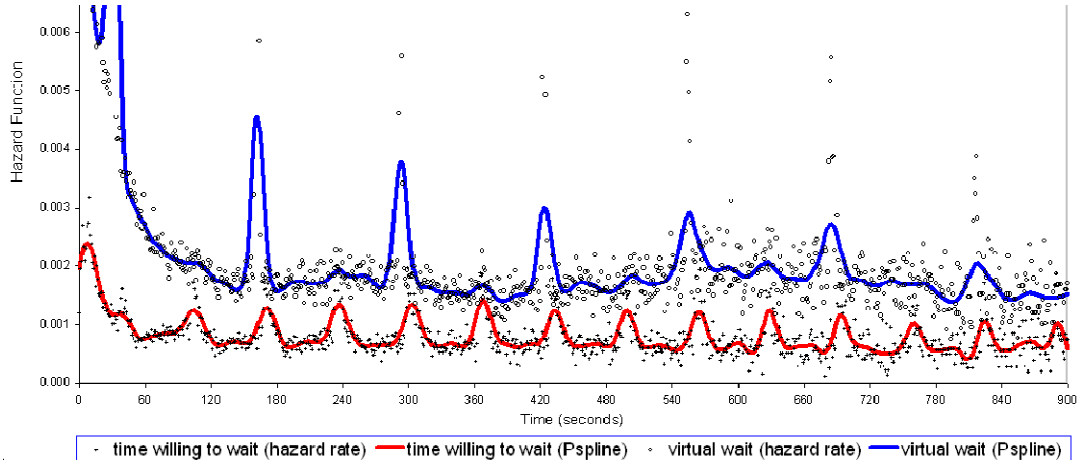

Figure 2: Customers' psychology (red) and system protocols (blue).

announcements. On the other side, the blue peaks can be explained by the answering policy: the queueing customers are answered at every blue peaks, which explains the fact that the probability to be served is very high during these periods.

2.3.2 Smoothing Challenges

The SEEStat interface features several smoothing tools covering some of the most popular ones in statistical and data mining methods. Examples include kernel smoothers, polynomial smoothers, Heft smoothers and others (Figure 3 shows the smoothing feature of the SEEStat). The interface also features parameter ranges which allow one to customize the smoothing.

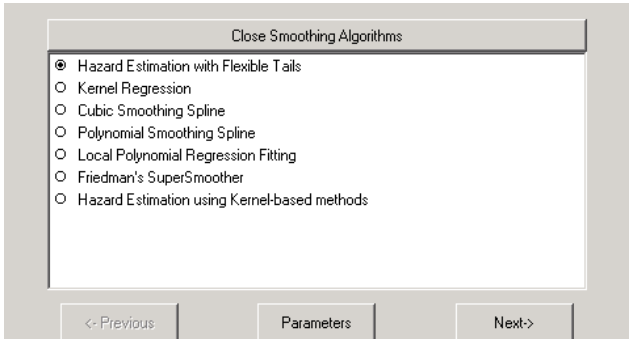

Figure 3: The SEE Stat Interface

As we showed in the previous subsection, time protocols can be revealed via smoothing. However, some smoothers appear to be more or less efficient than others, as we now demonstrate.

Figure 4 on Page 9 presents two smoothing methods for the same data set. The first is the Heft method, and the second one is a polynomial smoother with a degree exceeding 6. We notice here (several trials were carried out with different customer categories) that the many smoothers present relevant outputs for the first points of the graph but do not smooth the tail

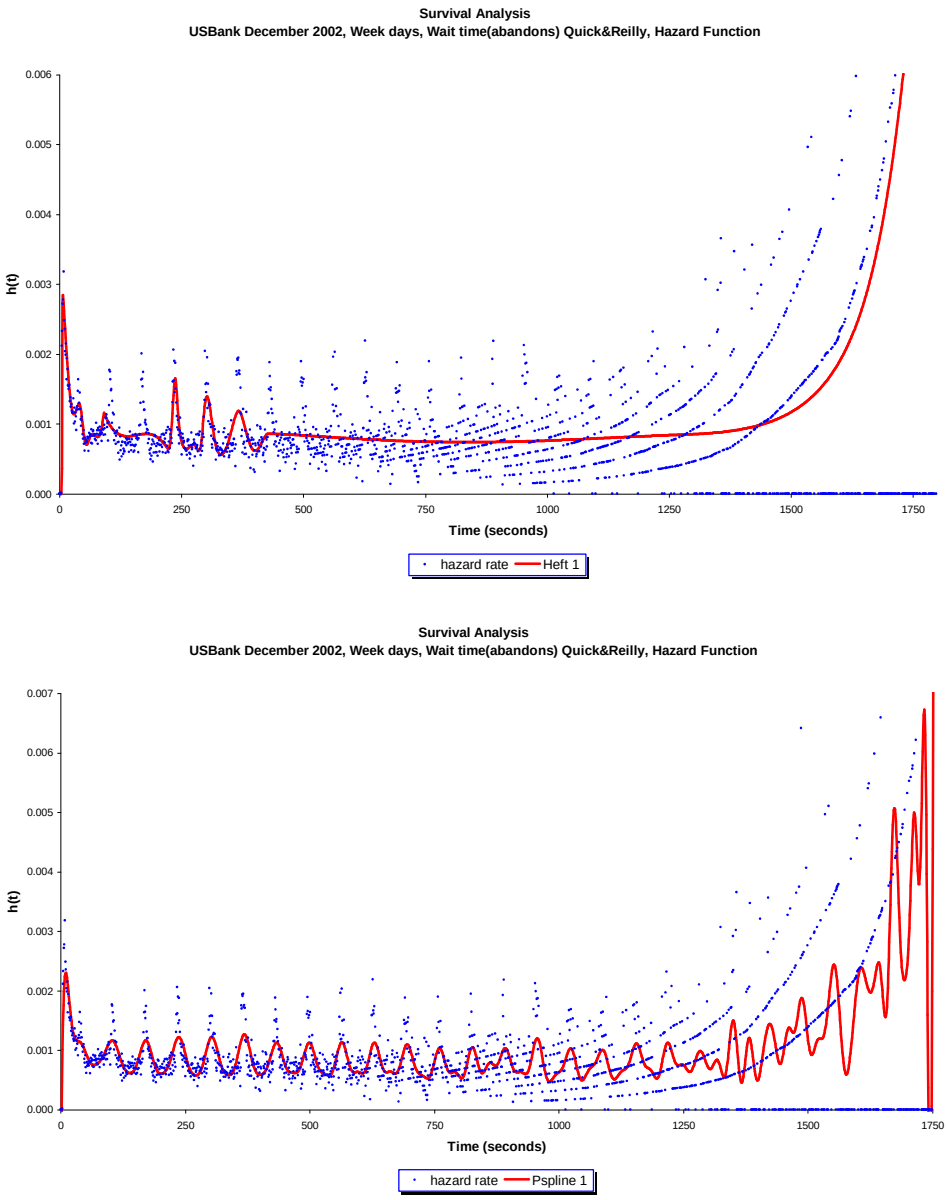

Figure 4: Two different smoothing methods

that well. To the contrary, some other smoothers seem to present good smoothing of the tail, whereas no protocol effects are detectable. One can see that this phenomenon is very common, at least in the databases to which we have access. Note here that the "goodness" of a smooth should be discussed further and it is not an easy topic. One can feel very quickly that there is a clear tradeoff between the tail smooths and the first points of a data set.

Smoothing tools are the first empirical tools to be used within the framework of SEEStat, and further trials should be carried out in order to assess the quality of the SEEStat smoother and its prediction capabilities.

3 The Proposed Research

The main goal of our research will be to predict the risk of abandonment for a repeated call of a customer, given his/her call history in the Call Center. More precisely, for a customer of a given category (among the three basic categories that are in our data base) who already called $m > 0$ times, we would like to provide a simple prediction procedure to assess the probability of abandonment during the $m + 1$ call at any time t.

Let Z_j be the covariate vector of call j, $j = 1, ..., m + 1$. Then, we assume:

$$\lambda_j(t|Z_j, \omega) = \omega \cdot \lambda_{0j}^c(t) \cdot e^{\gamma^T Z_j} \tag{3.1}$$

where $\lambda_{0j}^c(t)$ is an unspecified baseline hazard function of call j. Then, we can write the conditional survival as

$$S_j(t|Z_j, \omega) = e^{-\Lambda_{0j}^c(t)\omega e^{\gamma^T Z_j}}$$

where $\Lambda_{0j}^c(t) = \int_0^t \lambda_{0j}^c(u)du$.

If customer called m times already, we propose to estimate his unknown frailty variate ω by temporarily treating ω as an unknown parameter and adopting the approach of [2]. Let $f(\omega|\theta)$ be the frailty's density with unknown parameter vector θ; then the likelihood function based on the joint distribution of $(T_1, ..., T_m, Z_1, ..., Z_m, \delta_1, ..., \delta_m, \omega)$ is proportional to:

$$\prod_{j=1}^{m} [(\lambda_j(T_j|Z_j, \omega)^{\delta_j} \cdot S_j(T_j|Z_j, \omega))] \cdot f(\omega|\theta).$$

and the log-likelihood is proportional to

$$\ln f(\omega|\theta) + \sum_{j=1}^{m} [\delta_j \ln \omega - \Lambda_{0j}^c(T_j) e^{\gamma^T Z_j} \omega] \tag{3.2}$$

Hence, the maximum likelihood estimator of ω is the root of the following equation:

$$\sum_{j=1}^{m} \delta_j - \omega \cdot \sum_{j=1}^{m} \Lambda_{0j}^c(T_j) \cdot e^{\gamma^T Z_j} + \frac{\omega}{f(\omega|\theta)} \cdot \frac{\partial f(\omega|\theta)}{\partial \omega} = 0. \tag{3.3}$$

The unknown hazard functions $\Lambda_{0j}^c(t)$ for $j \in \{1, ..., m + 1\}$ and the unknown parameters γ and θ will be estimated based on external data sets. For estimating the regression coefficient parameters, the cumulative baseline hazard functions and the frailty distribution parameter, we use an extended estimation procedure of Zeng and Lin (2007). We have already written the R code and its performance has been assessed and confirmed under two frailty distributions—Log Normal and Gamma.

The prediction procedure is based on [4] where the method is used for genetic risk estimation. Let $T^m = (T_1, T_2, ..., T_m)$ be the observed times of the m first calls of a given customer, $Z^m = (Z_1, Z_2, ...Z_m)$ the covariates vectors of the m past calls and $\delta^m = (\delta_1, \delta_2, ..., \delta_m)$ the indicators set of the m past calls. Let $F^m = \{T^m, \delta^m, Z^m, Z_{m+1}\}$ be the observed history, and the covariates of his/her current call which is actually the one for which we want to predict the probability of abandonment. Let also $F_+^m = \{T^m, \delta^m, Z^m, Z_{m+1}, \delta_{m+1} = 0, T_{m+1}\}$ be as before, plus the information that this customer is currently waiting T_{m+1} units of time and didn't abandon, thus $\delta_{m+1} = 0$. Two prediction approaches are proposed in the following sections—the conditional approach and the marginal approach.

3.1 The Conditional Approach

Our goal is to estimate $S_{m+1}(t|F^m, \omega)$ and $S_{m+1}(t|F_+^m, \omega)$, the probability that the customer will still be waiting for service at time t at his $m+1$ call given his/her past call history and his/her frailty variate. To this end, write:

$$\begin{aligned} S_{m+1}(t|F^m, \omega) &= P(T_{m+1}^0 > t|F^m, \omega) \\ &= \frac{P(T_{m+1}^0 > t, T^m, \delta^m|Z^m, Z_{m+1}, \omega)}{f(T^m, \delta^m|Z^m, Z_{m+1}, \omega)} \\ &= \frac{P(T_{m+1}^0 > t|Z_{m+1}, \omega) f(T^m, \delta^m|Z^m, \omega)}{f(T^m, \delta^m|Z^m, \omega)} \\ &= P(T_{m+1}^0 > t|Z_m, \omega). \end{aligned} \tag{3.4}$$

Hence our estimator is defined as:

$$\hat{S}_{m+1}(t|F^m, \hat{\omega}) \tag{3.5}$$

where $\hat{\omega}$ is the solution of Eq. 3.3.

Similarly, for F_+^m we get:

$$\begin{aligned} S_{m+1}(t|F_+^m, \omega) &= P(T_{m+1}^0 > t|, T_{m+1}^0 > T_{m+1}, F_+^m, \omega) \\ &= \frac{P(T_{m+1}^0 > t, T_{m+1}^0 > T_{m+1}, T^m, \delta^m|Z^m, Z_{m+1}, \omega)}{P(T_{m+1}^0 > T_{m+1}, T^m, \delta^m|Z^m, Z_{m+1}, \omega)} \\ &= \frac{P(T_{m+1}^0 > t|Z_{m+1}, \omega) f(T^m, \delta^m|Z_{m+1}, Z^m, \omega)}{P(T_{m+1}^0 > T_{m+1}|Z_{m+1}, \omega) f(T^m, \delta^m|Z_{m+1}, Z^m, \omega)} \\ &= \frac{S_{m+1}(t|Z_{m+1}, \omega)}{S_{m+1}(T_{m+1}|Z_{m+1}, \omega)}. \end{aligned} \tag{3.6}$$

and our estimator is:

$$\hat{S}_{m+1}(t|F_+^m,\hat{\omega}) = \frac{\hat{S}_{m+1}(t|Z_{m+1},\hat{\omega})}{\hat{S}_{m+1}(T_{m+1}|Z_{m+1},\hat{\omega})}. \tag{3.7}$$

3.2 The Marginal Approach

The marginal approach does not use any maximum likelihood estimation for ω but also uses the frailty distribution $f(\omega|\theta)$:

$$\begin{aligned}
S_{m+1}(t|F^m) &= P(T_{m+1}^0 > t|F^m) \\
&= \frac{P(T_{m+1}^0 > t, T^m, \delta^m|Z^m, Z_{m+1})}{f(T^m,\delta^m|Z^m,Z_{m+1})} \\
&= \frac{\int P(T_{m+1}^0 > t, T^m, \delta^m|Z^m, Z_{m+1},\omega) f(\omega|\theta)d\omega}{\int f(T^m,\delta^m|Z^m,Z_{m+1},\omega)f(\omega|\theta)d\omega} \\
&= \frac{\int S_{m+1}(t|Z_{m+1},\omega)\prod_{j=1}^m f(T_j,\delta_j|Z_j,\omega)f(\omega|\theta)d\omega}{\int \prod_{j=1}^m f(T_j,\delta_j|Z_j,\omega)f(\omega|\theta)d\omega}
\end{aligned} \tag{3.8}$$

where $f(T_j,\delta_j|Z_j,\omega) = \lambda_j(T_j|Z_j,\omega)^{\delta_j} S_j(T_j|Z_j,\omega)$. Therefore,

$$S_{m+1}(t|F^m) = \frac{\int \omega^{\sum_{j=1}^m \delta_j} e^{-\omega(\Lambda^c_{0(m+1)}(t)e^{\gamma^T Z_{m+1}} + \sum_{j=1}^m \Lambda^c_{0j}(T_j)e^{\gamma^T Z_j})} f(\omega|\theta)d\omega}{\int \omega^{\sum_{j=1}^m \delta_j} e^{-\omega \sum_{j=1}^m \Lambda^c_{0j}(T_j)e^{\gamma^T Z_j}} f(\omega|\theta)d\omega}. \tag{3.9}$$

If it is reasonable to assume that $\Lambda^c_{0(m+1)}(t) = \Lambda^c_{0m}(t)$ for all t then our estimator will be:

$$\hat{S}_{m+1}(t|F^m) = \frac{\int \omega^{\sum_{j=1}^m \delta_j} e^{-\omega(\hat{\Lambda}^c_{0m}(t)e^{\hat{\gamma}^T Z_{m+1}} + \sum_{j=1}^m \hat{\Lambda}^c_{0j}(T_j)e^{\hat{\gamma}^T Z_j})} f(\omega|\hat{\theta})d\omega}{\int \omega^{\sum_{j=1}^m \delta_j} e^{-\omega \sum_{j=1}^m \hat{\Lambda}^c_{0j}(T_j)e^{\hat{\gamma}^T Z_j}} f(\omega|\hat{\theta})d\omega} \tag{3.10}$$

where $\hat{\gamma}$, $\hat{\Lambda}^c_{0j}(t)$ for all $j \in \{1,...,m\}$ and all t and $\hat{\theta}$ are the estimates from the extended procedure of [1].

For the other type of information, a very similar calculation yields:

$$S_{m+1}(t|F_+^m) = \frac{\int \omega^{\sum_{j=1}^{m}\delta_j} e^{-\omega(\Lambda^c_{0(m+1)}(t)e^{\gamma^T Z_{m+1}} + \sum_{j=1}^{m}\Lambda^c_{0j}(T_j)e^{\gamma^T Z_j})} f(\omega|\theta)d\omega}{\int \omega^{\sum_{j=1}^{m}\delta_j} e^{-\omega(\Lambda^c_{0(m+1)}(T_{m+1})e^{\gamma^T Z_{m+1}} + \sum_{j=1}^{m}\Lambda^c_{0j}(T_j)e^{\gamma^T Z_j})} f(\omega|\theta)d\omega} \tag{3.11}$$

and

$$\hat{S}_{m+1}(t|F_+^m) = \frac{\int \omega^{\sum_{j=1}^{m}\delta_j} e^{-\omega(\hat{\Lambda}^c_{0m}(t)e^{\hat{\gamma}^T Z_{m+1}} + \sum_{j=1}^{m}\hat{\Lambda}^c_{0j}(T_j)e^{\hat{\gamma}^T Z_j})} f(\omega|\hat{\theta})d\omega}{\int \omega^{\sum_{j=1}^{m}\delta_j} e^{-\omega(\hat{\Lambda}^c_{0m}(T_{m+1})e^{\hat{\gamma}^T Z_{m+1}} + \sum_{j=1}^{m}\hat{\Lambda}^c_{0j}(T_j)e^{\hat{\gamma}^T Z_j})} f(\omega|\hat{\theta})d\omega}. \tag{3.12}$$

If the involved integrals do not have a closed analytical solution, we will use numerical approximations such as the Gauss-Hermite approximation.

3.3 Measure of Performances

The goal of the present research is to evaluate the performance of our prediction procedure.

3.3.1 Simulation Studies

We plan to simulate the $m+1$ first calls of n customers and estimate the risk of abandonment at the $m+1$ call based on the m previous calls by using our proposed conditional and marginal approach, and compare it to the true values. The comparison can be made using the area under the curve (AUC) of the Receiver Operating Characteristic (ROC). The ROC-AUC is an established performance index for diagnostic tests and can be used for assessing the performance of various approaches in risk prediction at various times. Specifically, the ROC-AUC measures the probability that the predicted risk for a randomly selected individual who abandoned exceeds the risk prediction of a randomly selected individual who did not abandon. The ROC-AUC will be compared for different frailty distributions such as gamma and log-normal.

Another measure of performance to be used is the empirical mean-squared error for prediction,

$$\frac{1}{n} \cdot \sum_{i=1}^{n} (\hat{S}_m(t) - S_m(t))^2.$$

Often, in random effect models, it is difficult to assess from the data the true random effect distribution. For this purpose, we plan to investigate the following simulation settings: (1) the true frailty model is gamma but the analysis is based on the log-normal model; and (2) the true model is the log-skewed-normal distribution of Azalini (1985) but the analysis is based on the normal frailty model.

3.3.2 Data Analysis

Our prediction procedures can also be tested by using the call centers data to be provided to us by the SEELab. To this end we plan to:

1. Estimate the parameters of Model (3.1), namely the cumulative baseline hazard functions, the regression coefficient parameters and the frailty distribution parameter, using our proposed extended approach of Zeng and Lin (2007).

2. For each customer i that called m_i times we will predict the probability of abandonment of his/her m_i call based on the previous $m_i - 1$ calls, and compare our predictions to the true observed values based on the ROC-AUC.

3. Ideally, we would like to be able to predict the function

$$P(Abandon\ during\ (t, t+h]\ |still\ waiting\ at\ t)$$

for all $t \geq 0$, $h \geq 0$. Of special interest is the case of an infinitesimally small h, which corresponds to predicting the hazard rate function of the time a customer is willing to wait (patience, or impatience). Note that that case 2 above corresponds to $t = 0$, $h = \infty$.

The above steps will be applied using several frailty distributions. Then, our final goal is to implement the proposed prediction procedures in real-time, so that a call center's system will be able to assign priorities also based on these predictions.

References

[1] Lin DY. and Zeng D. Maximum Likelihood Estimation in Semiparametric Regression Models with Censored Data. *Royal Statistical Society*, 2007.

[2] Ha. ID., Lee Y., and Song JK. Hierarchical Likelihood Approach for Frailty Models. *Biometrika*, 2001.

[3] Klein J. and Moeschberger M. *Techniques for Censored and Truncated Data, Survival Analysis*. Statistics for Biology and Health,Second Edition, Springer. 2003.

[4] Gorfine M. and Hsu L. Genetic Risk Prediction in Complex Genetic Diseases Based on Family History. *Technical Report,Technion-IIT*, 2010.

[5] Feigin P., Mandelbaum A., Zeltyn S., Trofimov V., Ishay E., Khudyakov P., and Nadjharov E. The Call Center of a US Bank. *DataMOCCA Volumes*, 2006.

[6] Khudyakov P. Statistical Analyses of Call Center Data. *Ph.D. Research Thesis,Technion-IIT*, 2010.