Survival Data Mining

Will Potts

Data Miners Inc., Cambridge, MA 02140
wjep@data-miners.com

Abstract. Customer databases contain histories of vital events such as the acquisition and cancellation of products and services. The historical data is used to build predictive models for customer retention, cross selling, and other database marketing endeavors. The temporal nature of the target events can be accommodated with survival analysis methods. This paper outlines the application of survival analysis to predictive modeling and includes a discussion of the discrete-time logistic and piecewise exponential hazard models.

1 Customer Event History Data

Historical data, extracted from operational customer databases, can be used to build predictive models for various temporal outcomes:

- cancellation of products or services (churn)
- downgrading
- acquiring add-on products or upgrading
- product return
- loan prepayment.

The occurrence of the target event on the *i*th customer is controlled by the probability distribution of the time until the event, T_i. Customer events might be recorded at discrete increments such as months or on a continuous time scale. At the time the data was extracted for analysis, all customers usually have not experienced the event. In which case, the event time is considered (right) censored. Survival analysis is a set of statistical methods designed for censored duration data [1][2][3].

Censored event history data can be represented by an observed event time, $Y_i = \min(T_i, a - B_i)$, and an event indicator, $\delta_i = I\{T_i \le a - B_i\}$. The date of origin, B_i, can vary among customers. Typically, B_i represents the date that an account was opened. In this censoring scheme (generalized type I censoring), there is a fixed date, *a*, when the extracted data was current (Fig. 1). Another possible cause of censoring is the occurrence of an independent and mutually exclusive competing event. For example, if the event of interest is cancellation of a service, then a customer that moves out of the service area might be considered censored at the date they moved, a_i.

The data used for mining customer histories consists of retrospective samples extracted from large operational databases. In some applications, the available data consists of a cross-sectional snap-shot of customers that were active as of some fixed date *c*. Such a sample is truncated on the left. The sample is length-biased because, for a given start date, B_i, only the lengthier event times appear in the sample (Fig.1). With discrete event times the truncation date can equal the censoring date. The available data might be all accounts active at the beginning of the month, some of which experienced the event during the month.

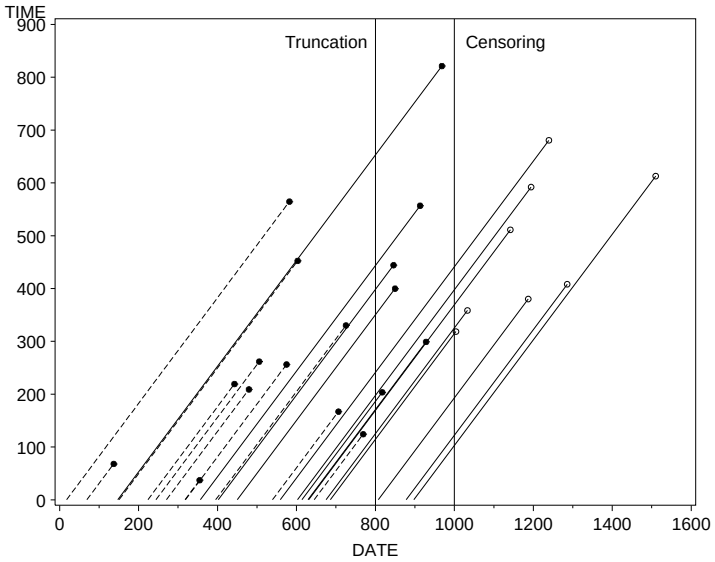

Fig 1. In this lexis diagram, a line segment represents each subject. The vertical axis is the event time. The horizontal axis is the calendar date. The beginning of each line segment corresponds to the origin $(B_i, 0)$. The end of the line segment corresponds to the event time $(B_i + T_i, T_i)$. The eight subjects with lines extending beyond the censoring date, $a = 1000$, are censored at $Y_i = 1000 - B_i$. If the sample were truncated at the date $c = 800$, the 11 dashed lines would be absent

The event time distribution is usually characterized by the survival function, $S(t) = \Pr(T_i \geq t)$, or the hazard rate. For discrete event times, the hazard rate is the conditional probability of the event given that it has yet to occur

$$h(t_j) = \Pr(T = t_j \mid T \geq t_j) = 1 - S(t_{j+1}) / S(t_j).$$

For continuous event times, the hazard rate is the limit of the conditional probability in an infinitesimally small interval

$$h(t) = \lim_{\Delta \to 0} \left(\tfrac{1}{\Delta} \Pr(t \leq T < t + \Delta \mid T \geq t) \right) = -\tfrac{d}{dt} \ln(S(t))$$

The hazard can be interpreted as an age-specific rate (events/unit time). The survival function decreases monotonically from one to zero. In contrast, the hazard rate can be any nonnegative function. The shape of the hazard rate often gives insight into the underlying system driving the occurrence of an event

Customer databases contain concomitant information that may affect the event time distribution such as demographics, account balances and payments, and the occurrence of other events such as the acquisition of new products or services. The vector of covariates for the *i*th customer, $\mathbf{x}_i(t)$, is often time-dependent. Time-dependent covariates can represent single irreversible events that occur at some point in the customer lifetime such as paying off an installment loan. Time-dependent covariates can be step functions representing the occurrences of repeatable events such as problems reported to customer service or payment delinquencies. Time-dependent covariates can be more continuously varying quantities such as the balance in an investment account.

2 Predictive Scoring

The ultimate purpose of modeling customer relationship data is usually prediction. Predictive models are used to map attributes of each customer to a score, which measures the propensity of some actionable event. The choice of an appropriate predictive score depends on how the model is to be deployed. In the most general scenario, customers would be scored at the current point in their lifetime for the propensity

of the outcome. Consequently, predictive scoring should consider the distribution of the residual event time $R = (T - t \mid T \geq t)$. The hypothetical random variable *R* is the time remaining until the event, conditional on the information available at the current time *t*. The hazard rate at *t*, $h(t)$, equals the probability density (mass) function of *R* and can be interpreted as the probability of the event in the next instant. The hazard rate is a relevant score in many applications.

Other potentially useful scores can be derived from the distribution of the residual event time. The expected value of *R* (mean residual life) is the area under survival function of *R*

$$\mu(t) = E(T - t \mid T \geq t) = \frac{1}{S(t)} \int_t^\infty S(x)dx .$$

The median of *R* (median residual life) is the half-life of the remaining time

$$m(t) = med(T - t \mid T \geq t) = S^{-1}\left(\tfrac{1}{2} S(t)\right) - t .$$

The mean and median are defined similarly for discrete event times. Smaller quantiles of the residual event time (e.g., quarter-life) can be useful with heavily censored data because the mean and median can fall far outside the range of observed events. If the model is used to score new customers at time zero, then $\mu(0)$ and $m(0)$ revert to the mean and median of *T*.

In many cases, the scores are used to forecast a future time $t + r$. The probability density function of *R* evaluated at *r* is more relevant than the hazard rate at $t + r$ because *T* is only known to be greater than *t*

$$f_R(r) = \lim_{\Delta \to 0}\left(\tfrac{1}{\Delta} \Pr(r \leq T - t < r + \Delta \mid T \geq t)\right) = \frac{f(t + r)}{S(t)} .$$

In forecasting applications, the time-dependent covariates would either need to be forecasted or lagged by *r* units in the model.

When the entire future interval $[t, t + r]$ is of interest, the survival function of *R* evaluated at *r* is pertinent

$$S_R(r) = \Pr(T \geq t + r \mid T \geq t) = S(t + r) / S(r) .$$

This quantity is monotonically related to the cumulative hazard (total risk) on the interval $[t, t + r]$. The area under the survival function of *R* on the interval $[t, t + r]$ involves all values in the interval, not just the endpoints. This area is equal to the restricted mean residual event time

$$\mu(t, r) = E(\min(R, r)) = \frac{1}{S(t)} \int_t^{t+r} S(x)dx .$$

When $t = 0$, the restricted residual event time is the restricted mean life [4].

An ideal predictive scoring model would give a sufficiently flexible estimate of the hazard rate as a function of the (possibly time-dependent) covariates. The discrete-time logistic hazard model and the piecewise exponential hazard model passably satisfy this requirement. The hazard rate uniquely characterizes the event time distribution. The survival function can be determined from the hazard rate using the identities $S(t) = \exp\left(-\int_0^t h(x)dx\right)$ and $S(t) = \prod_{t_j < t} (1 - h(t_j))$ for continuous and discrete times, respectively.

3 Discrete-Time Logistic Model

The discrete-time logistic model [5] assumes the logit of the discrete hazard is a linear combination of the logit of the baseline hazard rate (not depending on the covariates) and some suitably flexible function of the covariates (often linear).

$$\ln\left(\frac{h(t\,|\,\mathbf{x}_i(t))}{1-h(t\,|\,\mathbf{x}_i(t))}\right)=\ln\left(\frac{h_0(t)}{1-h_0(t)}\right)+\eta(\mathbf{x}_i(t),\text{ß})$$

The probability mass function of the observed data (Y_i,δ_i)

$$f_Y(y_i\,|\,\mathbf{x}_i(y_i))=h(y_i\,|\,\mathbf{x}_i(y_i))^{\delta_i}\prod_{j=0}^{i-1}(1-h(y_j\,|\,\mathbf{x}_j(y_j)))$$

equals the joint distribution of the $\delta_{ij}=\delta_i I\{i=j\}$, for $j=0,\ldots,i$; where the δ_{ij} are treated as independent Bernoulli variates with posterior probabilities equal to $h(y_i\,|\,\mathbf{x}_i(y_i))$. Consequently, the parameters can be estimated by maximum likelihood using logistic regression [2]. The Y_i need to be expanded to one pseudo-observation for each time point that is less than or equal to Y_i. Discrete time-dependent covariates may have different values for each pseudo-observation.

The model should include a sufficiently flexible parameterization of the time and covariate effects. One effective approach is to use regression spline terms [6]. Alternatively, the model can be represented as a neural network with a logistic output activation function and a Bernoulli error function [7].

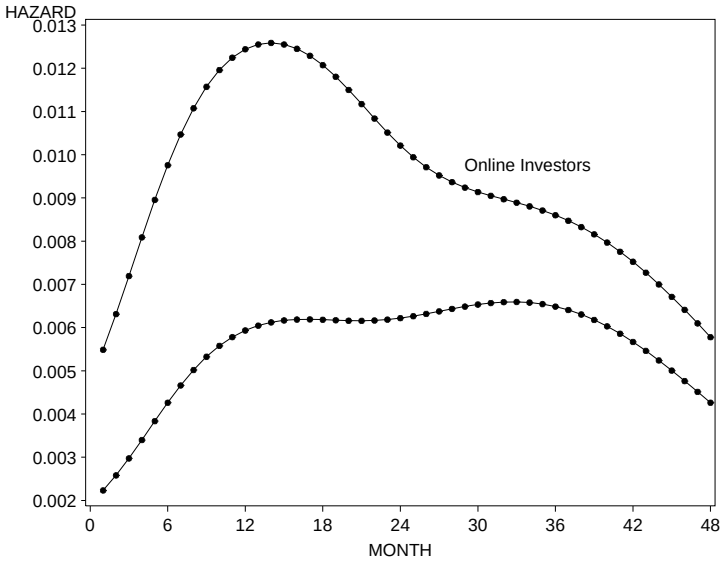

Fig 2. Discrete hazard rate for investor attrition, estimated using a discrete-time logistic neural network model. The model included the month after opening an investment account and a binary predictor indicating online investors

If the sample was left truncated at $c-B_i$, then the probability mass function of the observed data (Y_i,δ_i) becomes

$$\frac{1}{S(c-B_i)}h(y_i\,|\,\mathbf{x}_i))^{\delta_i}\prod_{j=0}^{i-1}(1-h(y_j\,|\,\mathbf{x}_j)))=h(y_i\,|\,\mathbf{x}_i))^{\delta_i}\prod_{j>c-B_i}^{i-1}(1-h(y_j\,|\,\mathbf{x}_j)))\,.$$

Thus, the parameters can be estimated using logistic regression after removing the pseudo-observations with times earlier than the truncation time $c-B_i$. In the special case where the truncation date is Y_i, each subject has one observation.

4 Weibull Model

A simple yet flexible method for modeling continuous event times is to assume they follow a Weibull distribution with shape parameter α and scale parameter $\exp(\eta(\mathbf{x}_i, \mathbf{ß}))$, where $\eta(\mathbf{x}_i, \mathbf{ß})$ is a suitable function of the covariates (often linear). The Weibull assumption leads to a proportional hazards model

$$h(t \mid \mathbf{x}_i) = \alpha t^{\alpha-1} \exp(-\alpha\eta(\mathbf{x}_i, \mathbf{ß})) .$$

The Weibull model accommodates a wide range of monotonic hazard rates. The hazard rate decreases at decreasing rate when $\alpha < 1$. The hazard is constant when $\alpha = 1$ (exponential distribution). The hazard increases at a decreasing rate when $1 < \alpha < 2$. The hazard increases linearly when $\alpha = 2$ (Rayleigh distribution). The hazard increases at an increasing rate when $\alpha > 2$.

The Weibull model can be fitted to the data (Y_i, δ_i) by maximum likelihood using methods designed for the class of accelerated failure time (AFT) models that are widely used in engineering applications. The parameter estimates can be used to estimate other characteristics of the distribution, such as the survival function

$$S(t \mid \mathbf{x}_i) = \exp\left(-t^{\alpha} \exp(-\alpha\eta(\mathbf{x}_i, \mathbf{ß}))\right),$$

the mean residual event time

$$\mu(t) = \exp(\eta(\mathbf{x}_i, \mathbf{ß})) \exp\left(t^{\alpha} \exp(-\alpha\eta(\mathbf{x}_i, \mathbf{ß}))\right) \Gamma\left(\tfrac{1}{\alpha} + 1\right) \left(1 - P(\tfrac{1}{\alpha}, t^{\alpha} \exp(-\alpha\eta(\mathbf{x}_i, \mathbf{ß})))\right)$$

(P is the incomplete gamma function), and the median residual event time

$$m(t) = \exp(\eta(\mathbf{x}_i, \mathbf{ß}))\left(t^{\alpha} \exp(-\alpha\eta(\mathbf{x}_i, \mathbf{ß})) + \ln(2)\right)^{1/\alpha} - t .$$

One shortcoming of the Weibull model (as well as other AFT models) is that standard implementations do not allow time-dependent covariates.

5 Piecewise Exponential Model

The piecewise exponential model [8] allows for a wider variety of hazard rate shapes than the Weibull model. The hazard rate is approximated by a step function, where time is partitioned into J intervals $(b_0 = 0, b_1](b_1, b_2] \cdots (b_{J-1}, b_J = \infty)$

$$h(t \mid \mathbf{x}_i(t)) = h_j \exp(\eta(\mathbf{x}_i(b_{j-1}), \mathbf{ß})) \qquad t \in (b_{j-1}, b_j) .$$

The values of the time-dependent covariates at the beginning of each interval affect the levels of the step function.

The distribution of the data (Y_i, δ_i) is proportional to the joint distribution of $\delta_{ij} = \delta_i I\{b_{j-1} < y_i \le b_j\}$, for $j = 1, \ldots, \max\{j : b_{j-1} < y_i\}$; where the δ_{ij} are treated as independent Poisson variates with means

$$\ln(E(\delta_{ij})) = \ln(h_j) + \eta(\mathbf{x}_i(b_{j-1}), \mathbf{ß}) + (\min(y_i, b_j) - b_{j-1}) .$$

Consequently, the parameters $(h_j, \mathbf{ß}')$ can be estimated by maximum likelihood using Poisson regression with a log link function and an offset [9][2]. The Y_i need to be expanded to one pseudo-observation for each interval prior to and including the current. The Poisson model can be closely approximated using logistic regression by treating the δ_{ij} as binomial variates with $(\min(y_i, b_j) - b_{j-1})$ trials.

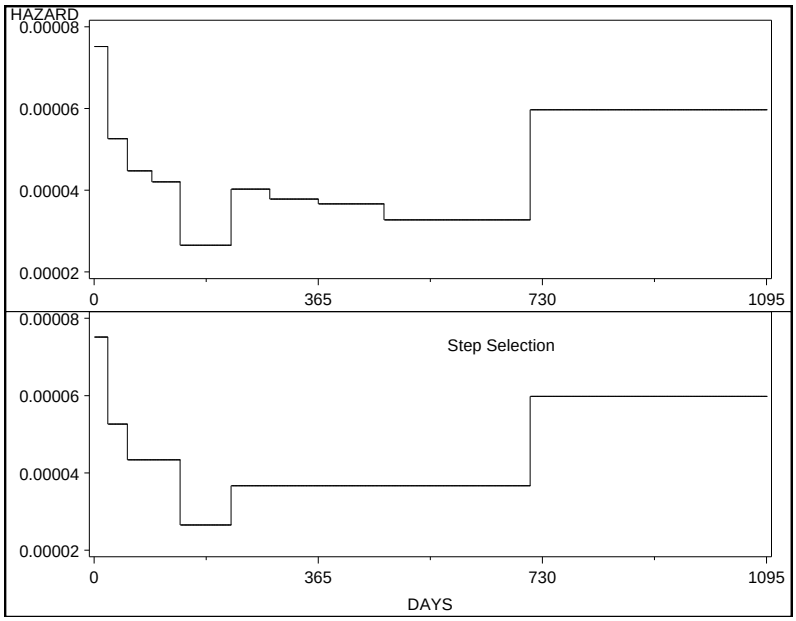

Fig. 3. Piecewise exponential hazard rate, before and after step selection. The example is from an up-selling application where the target event is upgrading to a high-value communication product. The time axis was partitioned into 10 intervals at the deciles of the event times

A more parsimonious step function can be selected without altering the interval lengths by merging consecutive steps where the incremental change $s_k = \ln(h_k) - \ln(h_{k-1})$ is small (Fig 3.). Step selection is facilitated by parameterizing the interval effect as $\ln(h_j) = \sum_{k=1}^{j} s_k$. Eliminating a term s_k from the model, for $k > 1$, merges the *k*th and (*k*-1)th step.

6 The Cox Proportional Hazards Model

In the Cox model, the functional form of the hazard rate is unspecified and unrestricted

$$h(t \mid \mathbf{x}_i(t)) = h_0(t) \exp(\eta(\mathbf{x}_i(t), \mathbf{ß})) .$$

The parameters are estimated by maximizing the partial likelihood [5], which does not involve the baseline hazard rate, $h_0(t)$. The model was designed for evaluating the effects of the covariates on the hazard without having to specify the functional form of the hazard. This seemingly attractive property is a liability in predictive modeling because the hazard rate is not estimated. Supplementary methods have been devised for nonparametrically estimating the survival function from a fitted Cox model [3]. The estimated survival function is constant between each distinct event time. This nonparametric representation does not lead to good estimates of the hazard rate and is prohibitively cumbersome as a scoring model.

The Weibull and piecewise exponential models are special cases of the general form of the Cox model. Consequently, the Cox model can be useful as an auxiliary tool for exploratory analysis, variable selection, residual diagnostics, and model validation.

References

1. Cox, D.R., Oakes, D.: Analysis of Survival Data. Chapman and Hall (1984)
2. Allison, P.D.: Survival Analysis Using the SAS System. SAS Institute Inc. (1995)

3. Klein, J.P., Moeschberger, M.L.: Survival Analysis: Techniques for Censored and Truncated Data. Springer-Verlag (1997)
4. Karrison, T.: Restricted Mean Life with Adjustment for Covariates. Journal of the American Statistical Association (1987) 1169-1176
5. Cox, D. R.: Regression Models and Life Tables (with discussion). Journal of the Royal Statistical Society, Series B (1972) 187-220
6. Efron, B.: Logistic Regression, Survival Analysis, and the Kaplan-Meier Curve. Journal of the American Statistical Association (1988) 414-425
7. Biganzoli, E., Boracchi, P., Mariani, L., Mabubini, E.: Feed Forward Neural Networks for the Analysis of Censored Survival Data: A Partial Logistic Regression Approach. Statistics in Medicine (1998) 1169-1186
8. Holford, T.R.: Life Tables with Concomitant Information. Biometrics (1976) 587-597
9. Holford, T.R.: The Analysis of Rates of Survivorship Using Log-Linear Models. Biometrics (1980) 299-305