The Application of Survival Analysis to Customer-Centric Forecasting
Michael J. A. Berry, Data Miners, Inc., Cambridge, MA

ABSTRACT
Survival analysis, also called time-to-event analysis, is an underutilized tool outside the clinical and manufacturing worlds where it is well established. The particular application addressed in this paper is creating a long-range daily forecast for a subscriber population that is a constantly changing mix of customers segments, each with its own survival function. A bottom-up forecast of this sort is more useful as a planning tool than the typical aggregate forecast because the survival-based forecast is very sensitive to changes in assumptions about the characteristics of new customers. The work described here is based on several successful forecasts developed for newspapers, telephone companies, and other subscription-based businesses.

INTRODUCTION
In many industries, forecasting future customer levels is a critical business function that relies on expertise from both the marketing and financial sides of the business. The traditional approach to forecasting uses standard time series analysis techniques to extrapolate historical summary data. This paper introduces a different approach based on survival analysis.

Forecasting is one aspect of the larger task of understanding customers. Media companies sell advertising based on subscription numbers. Telephone companies plan everything from handset inventory to call center staffing levels based on expected subscribers. Budgeting decisions depend on the expected impact of the alternatives under consideration so the forecasting process should include the ability to do what-if analysis related to budget levels: What will happen if we shift resources from one acquisition channel to another? What impact can we expect from a loyalty program? How many existing customers will we lose if we raise our prices by 10%?

Similar problems exist in many industries, particularly where there is a subscription or account relationship with customers. For instance, a wireless phone company not only wants to forecast overall customer numbers, they also want to forecast churn – the proportion of customers who leave each month. In addition, they want to forecast customer numbers for particular services and particular handsets in particular markets. Similar examples appear in financial services, insurance, cable television, pharmaceuticals, and a wide variety of other industries. This paper uses the example of a wireless telephone service provider, but the techniques discussed have broad applicability.

This paper uses one of the simplest methods of developing survival models, the empirical hazard model. In practice, more sophisticated methods are often used.

FORECASTING
At the highest level, the forecast needs to accomplish multiple goals.

First, it needs to estimate the number of active customers at different times in the future. Typical time frames might be six months, one year, or two years. The actual time frames depend on business needs. Often, this forecast needs to be organized in different ways for different purposes – by geography, by product, and by customer segment, for instance. One of the challenges in creating a complete forecasting system is incorporating all the relevant business processes into the forecast.

Second, the forecasting system needs to be able to measure actual results and to report on differences between the forecast estimates and the actual results. Such gap analysis is often as important as the forecast itself. The forecast is a “best attempt” based on what was known at the time of the forecast. The gap analysis can provide insight into how business conditions have changed causing the forecast to be too high or too low.

The forecast itself consists of two different components. The existing base forecast is for customers who are currently customers at the time of the forecast. Much is already known about these customers, because they have already started.
The new start forecast is for customers who are going to begin in the future. Most businesses who acquire customers already have business processes for managing new customers. In fact, there are usually several different new start forecasts: the budget forecast that was estimated before the fiscal year began, actual counts of customers who have already started during the year, and revised estimates for the remainder of the year. Managing these different forecasts of new starts is an important part of a forecasting system.

The forecasting system has three data sources. The first is the existing customer base at the time of the forecast. The number of base customers can only decline over time since no new customers are added to the base.

The second data source is the add plan, an estimate of future starts. Usually, the add plan is produced by the managers responsible for the starts. However, in some cases, it is also possible to estimate new starts analytically. In either case, estimates must be provided for the entire period of the forecast.

The third data source is the actual values for customers starting and stopping as these become available. These actuals are needed for gap analysis. Typically, the forecast will be redone on a regular basis to take into account newly available actuals. Both the original forecast and the adjusted forecast must be available for analysis.

**SURVIVAL-BASED FORECASTING**

For the subscription-based companies we work with—newspapers, wireless telephone carriers, internet service providers, software-as-a-service businesses, and the like—survival analysis is a natural choice. For these industries, the major determinate of future subscription levels is how long current and future subscribers are retained. Because survival analysis was developed specifically to study how long things last, it has become our method of choice for creating subscriber forecasts.

The basic idea is simple. New customers are acquired according to an add plan developed by management. The characteristics of new customers—their ages, incomes, credit ratings, choices of initial products, county of residence, and so forth—determine to which customer segment they are assigned. Each segment has an associated survival curve. On any given day, members of the same segment who started on different days will be on different parts of the curve. The overall subscriber population is always the sum of the segments.

We will work through an example of survival-based forecasting using the same telephone subscriber data used in section 1.3. First, we must explain the two fundamental ideas that underlie survival analysis:

- the hazard probability
- the survival function

The descriptions and definitions we give here differ from ones you might have seen elsewhere. Survival analysis is most often introduced in the context of clinical trials. Without getting into too much technical detail, these differences mostly stem from the very large quantities of data typically available in the business world (millions of customers versus dozens of patients) and to the fact that in the business world we tend to treat time as discrete rather than continuous. That is, everyone who cancels a subscription on the same day is considered to have stopped at the same time.

**THE HAZARD PROBABILITY**

The hazard probability at time t is the probability that a subscriber will stop at exactly time t given that they have not stopped at any earlier time. The hazard probability is estimated by dividing the number of subscribers who have ever stopped at time t by the number of subscribers who were ever at risk at time t. In the simple case, everyone who achieved a tenure of t or greater could have stopped at tenure t and so was in the population at risk for that tenure.

One important feature of survival data is that many of the observations are censored. That just means that for anyone who is still active on the observation date, we cannot say what their eventual tenure will be. These censored observations are counted in the at-risk pool until the tenure at which they are censored and not included in the risk pool for larger tenures.

The following chart shows daily hazard probabilities for deactivation of wireless telephone subscribers:
This hazard plot has some interesting features. The hazard is high for the first couple of days and then drops off sharply. This is called infant mortality, and it is a feature of many hazard plots including, as the name implies, those of human births. In this case, it probably represents “buyer’s remorse” and people who discovered that the coverage offered was insufficient for their needs. The next peak represents people who never paid their first bill. After these have been removed, the hazard drifts lower as the people who are going to go bad by failing to pay are weeded out. In the first year, while subscribers are under contract, deactivations are dominated by “involuntary” stops. The huge spike at day 365 shows that voluntary attrition is a big concern, however. All through the year, people place calls to the call center requesting to stop. When reminded of the one-year contract, many of them request to be deactivated on the first possible day. After the anniversary is passed, the hazard settles down again at a somewhat higher level than during the contract period. Some monthly seasonality is visible as each monthly bill causes some customers to decide they want to cancel and starts the clock on other customers being considered delinquent.

FROM HAZARD PROBABILITIES TO SURVIVAL CURVES

Although the hazard plot is interesting in and of itself, the primary reason for calculating hazard probabilities is to compute the survival curve. The survival curve for a customer segment shows what proportion of the segment survives to each tenure. Because deactivated customers do not come back (or rather, come back but are treated as new), the curve is monotonically decreasing. There is a simple relationship between the hazard probability and survival:

\[
S(T) = \prod_{t=1}^{T} (1 - h(t \cdot 1))
\]

\[
S(0) = 1
\]

By definition, there is 100% survival at tenure 0. At tenure 1, everyone survives except those few who succumb to the time 0 hazard. It is the same at every tenure: to get the number of subscribers surviving to time \( t \), take the
number that survived to time t-1 and multiply by the chance of making it one more day, which is 1-h(t-1).
Here is the survival curve produced by the hazards calculated from the example data:

![Survival Curve](image)

Notice the steep drop off in survival at the anniversary, and how this corresponds to the very high hazard at day 365.

**FORECASTING FUTURE STOPS FOR NEW SUBSCRIBERS**

When new subscribers arrive, we immediately assign them to customer segments so we know which survival curve they are riding. This means that the customer segments should be defined in terms of things that are known at time 0. Typical dimensions include geography, initial product, acquisition channel, initial rate plan, credit class, and age at application time.

Suppose that 1,000 subscribers start today and are assigned to segment d1. Suppose further that the hazard table for segment d1 for days 0-9 is the following:

<table>
<thead>
<tr>
<th>Day</th>
<th>Hazard</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.006170</td>
</tr>
<tr>
<td>1</td>
<td>0.007202</td>
</tr>
<tr>
<td>2</td>
<td>0.007447</td>
</tr>
<tr>
<td>3</td>
<td>0.005407</td>
</tr>
<tr>
<td>4</td>
<td>0.002587</td>
</tr>
<tr>
<td>5</td>
<td>0.002031</td>
</tr>
<tr>
<td>6</td>
<td>0.001651</td>
</tr>
<tr>
<td>7</td>
<td>0.001804</td>
</tr>
<tr>
<td>8</td>
<td>0.001267</td>
</tr>
<tr>
<td>9</td>
<td>0.001088</td>
</tr>
</tbody>
</table>

The day 0 hazard of 0.006170 means that of the 1,000 d1 subscribers that started today, we expect 1,000*0.006170 = 6.17 to fail to make it until tomorrow. That means we expect 993.83 of these subscribers to be
around tomorrow to be exposed to the slightly higher day 1 hazard. Of these, $993.83 \times 0.007202 = 7.16$ will fail to see the following day, their day 2.

<table>
<thead>
<tr>
<th>Day</th>
<th>Hazard</th>
<th>Survival</th>
<th>Lost</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.006170</td>
<td>1000.00</td>
<td>6.17</td>
</tr>
<tr>
<td>1</td>
<td>0.007202</td>
<td>993.83</td>
<td>7.16</td>
</tr>
<tr>
<td>2</td>
<td>0.007447</td>
<td>986.67</td>
<td>7.35</td>
</tr>
<tr>
<td>3</td>
<td>0.005407</td>
<td>979.32</td>
<td>5.30</td>
</tr>
<tr>
<td>4</td>
<td>0.002587</td>
<td>974.03</td>
<td>2.52</td>
</tr>
<tr>
<td>5</td>
<td>0.002031</td>
<td>971.51</td>
<td>1.97</td>
</tr>
<tr>
<td>6</td>
<td>0.001651</td>
<td>969.54</td>
<td>1.60</td>
</tr>
<tr>
<td>7</td>
<td>0.001804</td>
<td>967.94</td>
<td>1.75</td>
</tr>
<tr>
<td>8</td>
<td>0.001267</td>
<td>966.19</td>
<td>1.22</td>
</tr>
<tr>
<td>9</td>
<td>0.001088</td>
<td>964.97</td>
<td>1.05</td>
</tr>
</tbody>
</table>

Here, the calculation has been extended for a few more days. The lost column contains the expected contribution to overall attrition from this particular cohort of 1,000 subscribers for 10 days starting with today.

**FORECASTING FUTURE STOPS FOR EXISTING SUBSCRIBERS**

On the day we do the forecast, many subscribers already exist. Because they have already achieved some tenure $\geqslant 0$, they should not be placed on the survival curve originally calculated for their segment. Instead, they follow the conditional survival curve for their segment given that they have already survived to their current tenure. The conditional survival is simply the survival as originally calculated divided by the survival value for the subscriber’s current tenure. The chart below compares conditional survival for subscribers who survive for at least six months to overall survival.

Note that although customers who survive at least 6 months have the same steep drop-off at one year, the overall survival of this group remains higher than for the population as a whole as far out as we can calculate. This is because most involuntary attrition due to nonpayment of bills happens early in the tenure. After the first few months, most attrition is voluntary and the most common time to quit voluntarily is at the end of the one-year contract period. A real forecast would treat involuntary and voluntary stops as separate competing risks, but that topic is beyond the scope of this paper.

To return to our small example, let us say that in segment D1, on the day when the cohort of 1,000 in section 0 started, there were 800 subscribers in their 5th day. What is the hazard probability of not surviving to day 6 for these customers? By definition, the conditional survival at day 5 is 100%. The conditional survival at day 6 is the unconditional survival at day 6 divided by the unconditional survival at day 5. For this data,
969.54/971.51 = 0.99797. And 1 minus that number is the day 5 hazard probability, 0.002. Multiplying that hazard by 800 leads us to expect to lose 1.6 subscribers from the second cohort on the first day of the forecast period.

More generally, today's survival reflects yesterday's hazard, so $h(t-1) = 1 - \frac{S(t)}{S(t-1)}$.

If these were the only two cohorts, total forecast losses for the first few days would be as follows:

<table>
<thead>
<tr>
<th>Day</th>
<th>Cohort 1 Hazard</th>
<th>Cohort 1 Remaining</th>
<th>Cohort 1 Lost Survival</th>
<th>Cohort 2 Hazard</th>
<th>Cohort 2 Remaining</th>
<th>Cohort 2 Lost</th>
<th>Total Lost</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.006170</td>
<td>1000</td>
<td>6.17</td>
<td>1</td>
<td>0.002031</td>
<td>800.00</td>
<td>1.62</td>
</tr>
<tr>
<td>1</td>
<td>0.007202</td>
<td>993.83</td>
<td>7.16</td>
<td>0.997969</td>
<td>0.001651</td>
<td>798.38</td>
<td>1.32</td>
</tr>
<tr>
<td>2</td>
<td>0.007447</td>
<td>986.67</td>
<td>7.35</td>
<td>0.996322</td>
<td>0.001804</td>
<td>797.06</td>
<td>1.44</td>
</tr>
<tr>
<td>3</td>
<td>0.005407</td>
<td>979.32</td>
<td>5.30</td>
<td>0.994525</td>
<td>0.001267</td>
<td>795.62</td>
<td>1.01</td>
</tr>
</tbody>
</table>

In reality, there are many cohorts—one for each tenure as of the forecast day for each defined customer segment. Each cohort has its own forecast based on its own conditional survival curve. The overall deactivation forecast for the existing base is the sum of these individual cohort forecasts.

**FORECASTING FUTURE STOPS FOR FUTURE SUBSCRIBERS**

The future subscriber population will include current customers who don't stop plus new customers who have yet to be acquired. As soon as these new customers arrive, they will be at risk for stopping. Customers who start in the future are handled in the same way as customers who started today. They are assigned to a customer segment based on the values of explanatory variables such as geography, initial product, acquisition channel, initial rate plan, credit class, and age at sign-up. On the day they are added, they are exposed to the time 0 hazard for their segment, and things proceed as described earlier in this section.

The only difference (admittedly, an important one!) is that we don't actually know how many subscribers will be added each day in each segment. These numbers come from an acquisition plan developed outside the forecasting process. The company actively manages to the add plan, but it may or may not hit its targets. The more acquisitions deviate from the plan, the less accurate the forecast will be.

**WHAT-IF ANALYSIS**

The add plan represents one possible scenario for the coming year. One very valuable use of the forecasting system is to create multiple scenarios to see what effect they have on the forecast. By changing the add plan, we can simulate shifting resources from one acquisition channel or market to another or changing the mix of credit scores.

**EVALUATING A FORECAST**

The only way to evaluate a forecast is to compare it to the actual numbers as they become available. This process tends to lead to a proliferation of forecasts as the original forecast is adjusted in light of what actually happens. For any given month, a natural way to evaluate the forecast is to calculate the percentage difference between the forecast and actual values. To evaluate the forecast over a longer period, a measure such as the root mean squared error (RMSE) is useful. Either way, only the business context can provide an answer to the question “How good is good enough?”

The initial evaluation of the forecast is made at the aggregate level, but with a customer-centric forecast it is possible and desirable to also evaluate each cohort-level forecast. Because they are based on smaller populations, these will naturally show greater variance than the aggregate forecast. Any cohort that is off by more than some threshold amount should be investigated. It might be an early warning sign of something that will go wrong in other cohorts at a later date.

It is important to separate out the error that is due to deviation from the acquisition plan from error that is due to other causes, such as incorrect hazard estimates. For many business purposes, it doesn’t matter why the forecast...
is wrong, but for technical evaluation of the forecasting system it is important to know how good the forecast would have been if actual acquisitions had exactly followed the add plan. One way to do this is to rerun the forecast substituting the actual acquisitions for the original plan. A shortcut is to simply evaluate the existing base forecast. This is not quite the same because it will not detect whether the hazard probability for new customers is drifting, but it is easier because the existing base forecast already exists as one of the components of the overall forecast.

**APPLYING THE SURVIVAL-BASED METHOD**

We now show how to create the existing base forecast using the survival-based approach. Normally, the first step is to define customer segments because each segment should have its own family of survival curves. Several good variables for segmentation are available including market, acquisition channel, and credit class. Here our goal is to compare the simplest form of the survival-based model with the simplest form of the time-series-based model, so we treat all subscribers as one big segment.

Following standard data mining procedure, we split the data into training and test sets, even though with a data set this size, we can be quite confident that the two sets will have nearly identical distributions for all variables. Next, using the training data, we calculate the hazard for each tenure as shown previously. This calculation can be done fairly simply in DATA step code or using SQL queries, but SAS® provides the LIFETEST procedure, which makes things even easier. In the following chart, the survival curve has been augmented with pointwise 95% confidence bounds. For most tenures, these confidence bounds are so narrow that they are not visible on the chart, but for tenures greater than 1,500 days they start to become visible. This is because there are relatively few examples of subscribers with higher tenures so the hazard probability estimates and the survival calculated from them are less confident.

![Survival Distribution Function](image)

Next we count up the number of active customers having each tenure as of the start date for the forecast, 01 May 2000. We divide the survival curve by the survival at each tenure to get the conditional survival curves for each tenure. The final forecast is the sum of thousands of conditional survival curves.
The survival curves for three cohorts are shown here: the cohort that started on 01 May 2000, the cohort that had tenure 50 on that day, and the cohort that had tenure 100 on that day. By multiplying the conditional survival curves for each cohort by the size of the cohort on the first day, this chart becomes an estimate of the number of subscribers remaining from each cohort on any given day. In the training data, there are 90 people who started on 01 May 2000, 69 who started 50 days earlier on 12 March 2000, and 120 that started 100 days earlier on 22 January 2000. To get the final forecast, we simply sum the expected surviving population from each cohort.

APPEALING FEATURES OF THE SURVIVAL-BASED FORECAST

The explanatory power of the survival forecast is best seen by looking at forecast stops rather than forecast survival. The survival forecast is easily converted into a stops forecast by subtracting the forecast number of survivors at each time t from the number at time t-1. The following chart is a daily forecast for stops on the test data:

At first glance, this is quite surprising. The forecast warns that there will be a surge in the number of stops around January of 2001. Because this is the base forecast, the number of subscribers can only decrease over time. Few-
er subscribers should mean fewer stops, and that is indeed the predicted long-term trend. What accounts for the predicted increase for the months beginning in October of 2000?
The answer lies in the distribution of tenure cohorts.

Notice the large number of customers with tenures less than 200 days. All of these will experience their first anniversary within the first year of the forecast period. Most of them will come off contract at that point and experience a large anniversary spike in the hazard probability. Different cohorts will hit the anniversary drop-off on different dates. On days when a large cohort comes off contract, there will be an up-tick in total deactivations. The largest cohorts have tenures 129 and 130 at the start of the forecast period. These subscribers will go off contract 236 and 235 days later and sure enough, that is when the model predicts a surge.

When a business is growing rapidly, it is natural that on any given day, a large proportion of existing customers has low tenure. As we forecast further and further into the future, the tenure distribution of the base changes. This change is folded into the forecast automatically as different hazards are applied for each tenure. The survival analysis framework also makes it easy to account for the effects of explanatory variables. Clearly, factors such as price, customer experience, and credit class must have an impact on deactivations, so one way to improve any forecast is to add additional explanatory variables. Unfortunately, this is not as easy as might be expected when these must be forecast as well. For example, suppose we hypothesize that the deactivation rate is partly a function of gas (petrol) prices and consumer interest rates. It may be possible to build a very good model of deactivations at a given time $t$ based on these variables, but the resulting forecast for time $t+n$ now depends on the forecast gas price and interest rate at time $t+n$. If these forecasts are not reliable, the deactivation rate forecast will be unreliable as well. This suggests trying to find explanatory variables for which either the value does not change over time or for which it is possible to predict the effect of an initial value on later behavior.

**CONCLUSION**

In conclusion, because it is customer-centric, even the simplest survival-based model captures effects of a changing subscriber mix that might be hard to capture with a traditional, time-series-based approach. Furthermore, the simple survival model is easy to extend by adding covariates and by using more sophisticated functions to estimate the hazard probabilities.
ACKNOWLEDGEMENTS
SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.
Other brand and product names are registered trademarks or trademarks of their respective companies.

CONTACT INFORMATION

Michael J. A. Berry  
Data Miners, Inc.

mjab@data-miners.com  
www.data-miners.com