Customer-Centric Forecasting Using Survival Analysis

White Paper

Customer-Centric Forecasting Using Survival Analysis White Paper was developed by Michael J. A. Berry and Gordon S. Linoff. Editing and production support was provided by the Curriculum Development and Support Department.

Customer-Centric Forecasting Using Survival Analysis White Paper

Course code BDMWP, prepared date 24Feb2009. BDMWP_001

Table of Contents

To learn more…

For information on other courses in the curriculum, contact the SAS Education Division at 1-800-333-7660, or send e-mail to training@sas.com. You can also find this information on the Web at support.sas.com/training/ as well as in the Training Course Catalog.

For a list of other SAS books that relate to the topics covered in this Course Notes, USA customers can contact our SAS Publishing Department at 1-800-727-3228 or send e-mail to sasbook@sas.com. Customers outside the USA, please contact your local SAS office.

Also, see the Publications Catalog on the Web at support.sas.com/pubs for a complete list of books and a convenient order form.

Chapter 1 Customer-Centric Forecasting Using Survival Analysis

1.1 Overview

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 white paper points out shortcomings of the traditional approach and 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 discusses forecasting from three perspectives. First, we introduce the various components of a forecast. Next, we provide an overview of the usual approach to forecasting customer levels using a technique called ARIMA. This technique, which was originally developed to understand macro-economic trends, is quite powerful, but because it is usually applied at the aggregate level, it does not provide good insight at the level of particular customer segments. Lastly, we introduce survival analysis as a method for building customer-centric forecasts. Survival analysis, an underutilized tool in the data miner's toolkit, is very well suited to this purpose. It provides a simple way to incorporate customer-specific characteristics into the forecast. This paper uses one of the simplest methods of developing survival models, the empirical hazard model. In practice, more sophisticated methods are often used.

The data used for the examples in this paper is available for download at www.data-miners.com.

1.2 Forecasting

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

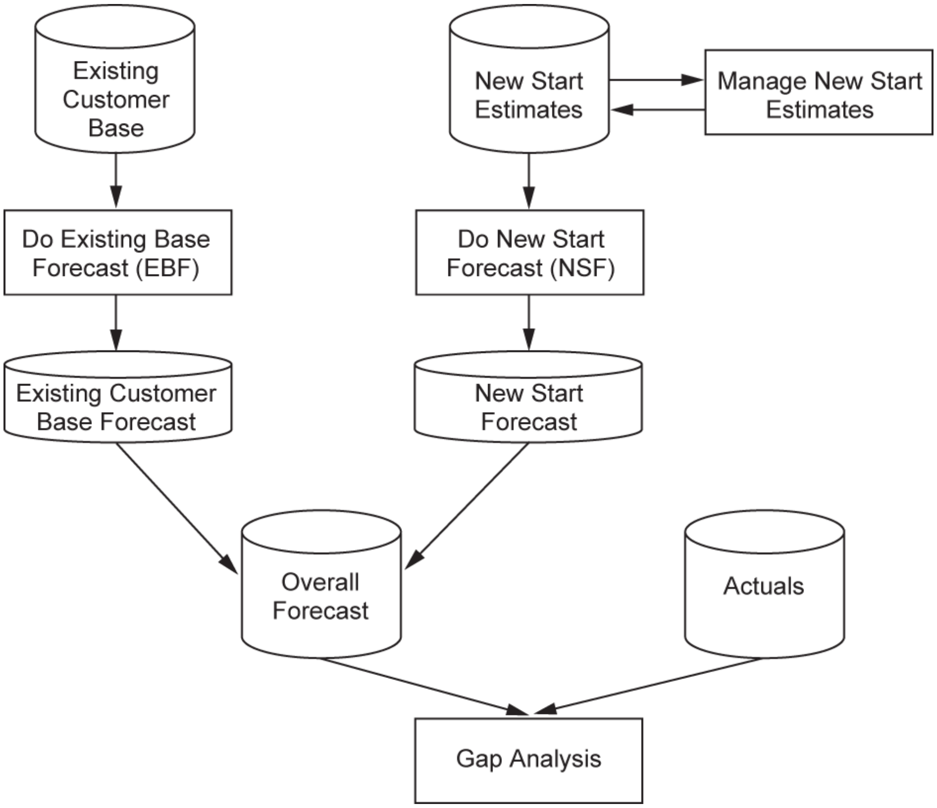
First, it needs to estimate the number of active customers at different times in the future. Typical time frames might be six months, one year, or two years. The actual time frames depend on business needs. Often, this forecast needs to be organized in different ways for different purposes – by geography, 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 following figure shows different components of a customer-centric forecasting system:


Existing Customer Base
New Start Estimates
Manage New Start Estimates
Do Existing Base Forecast (EBF)
Do New Start Forecast (NSF)
Existing Customer Base Forecast
New Start Forecast
Overall Forecast
Actuals
Gap Analysis

The overall 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.

1.3 The Usual Approach to Forecasting

The standard approach to forecasting time series is called ARIMA, which stands for **a**uto-**r**egressive, **i**ntegrated, **m**oving **a**verage. This is actually a family of related techniques used for forecasting time series of all kinds. This section introduces the ARIMA approach in the context of the existing base forecast. In practice, though, the ARIMA methods would typically be done on the overall forecast. It is not our intention to show all the many complex variations on the core idea that a skilled forecaster would make use of in practice. On the contrary, our goal is to walk through the simplest imaginable ARIMA model and the simplest imaginable survival model to illustrate the basic differences between the two approaches.

As a reminder, the existing base forecast tells us how many of the currently active customers will still be around at various points in the future. There are three steps involved in building the simple ARIMA model for the existing base forecast:

1. First, the time series is made stationary by removing the overall trend. (This is the **integrated moving average** part of ARIMA).

2. Next, any regular, cyclical patterns are removed by expressing the value at time *t* in terms of the value *t-n* for various values on *n*. (This is the **auto-regressive** part of ARIMA).

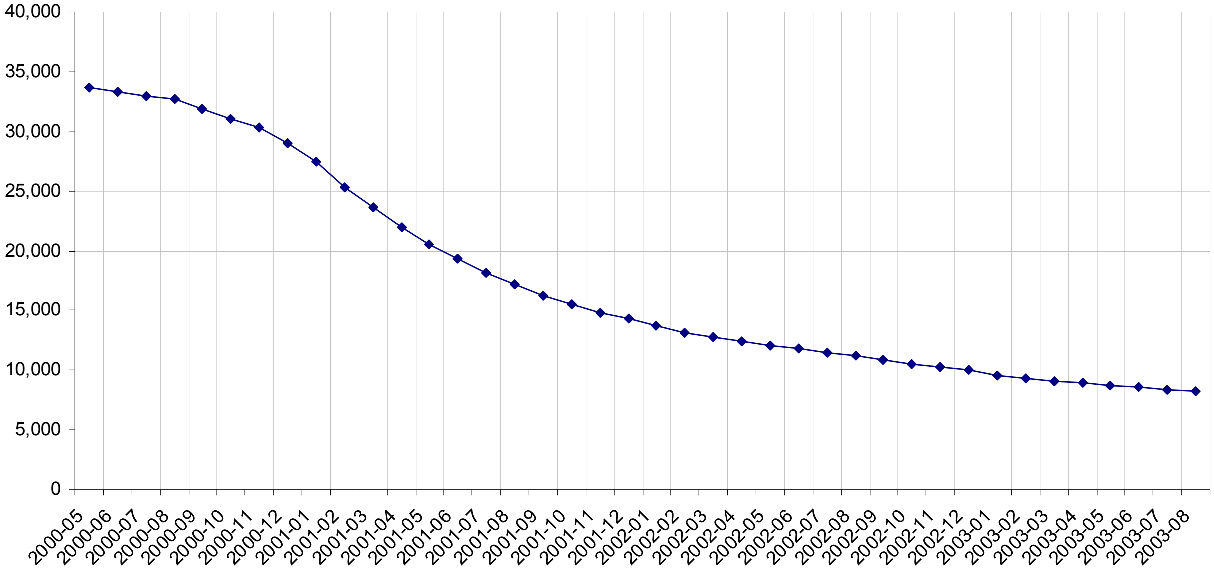
3. Finally, the remaining variation is modeled in terms of whatever explanatory variables are found to be significant.

In this simple example, we eliminate step 3 because, as is often the case, trend and seasonality alone are sufficient to describe the time series fairly well.

The rest of this section walks through an example.

The Original Time Series

The following chart shows what happens to the 33,624 customers who were active as of May Day 2000. Although the company records new starts and stops on a daily basis, here the data has been summarized at the monthly level.

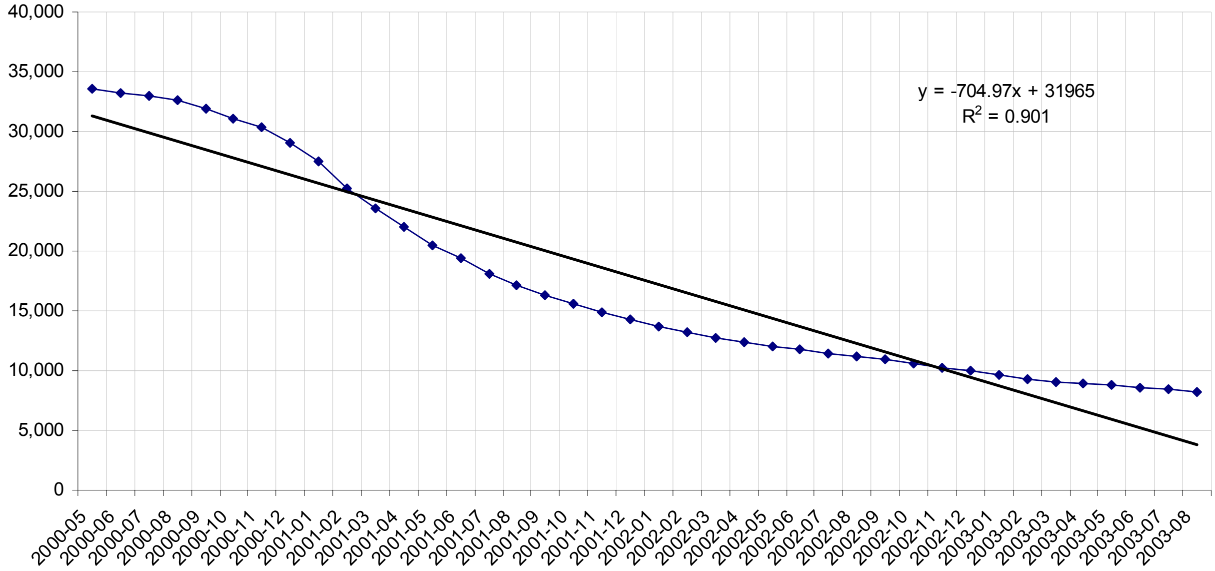


Customers remaining from active base of May 2000

Summarizing at the monthly level is very common, but not without problems: some months are longer than others, some have more holidays, and so on. In this case, however, it achieves the desired effect, which is to make the overall pattern in the data very clear.

Calculating the Trend

The first step in the standard forecasting process is to remove the trend. The trend is simply the best-fit line to the data.

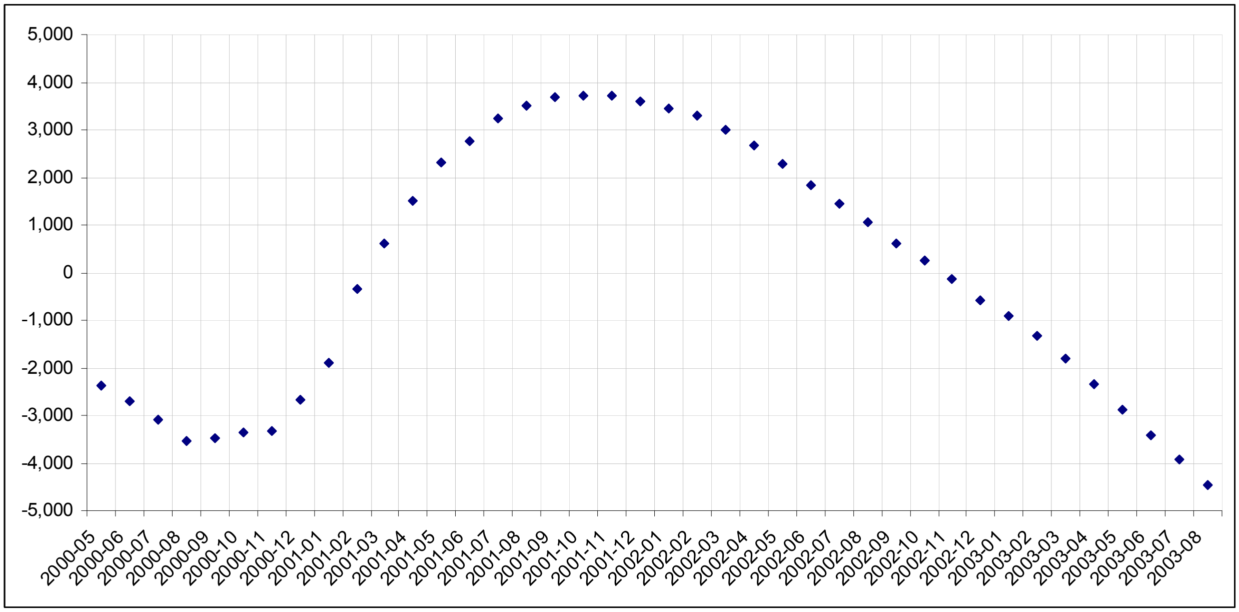


The linear trend captures 90% of the variance in the base subscriber population

The high R^2 value shows that the best fit line explains most of the variance in the number of base subscribers over the period. For some purposes, this very simple model may be good enough. Its slope is the rate that customers leave over time. We have found the first component of our model: the trend. The next step is to remove the trend so that we can see what else is happening.

An Autoregressive Model on the Residuals

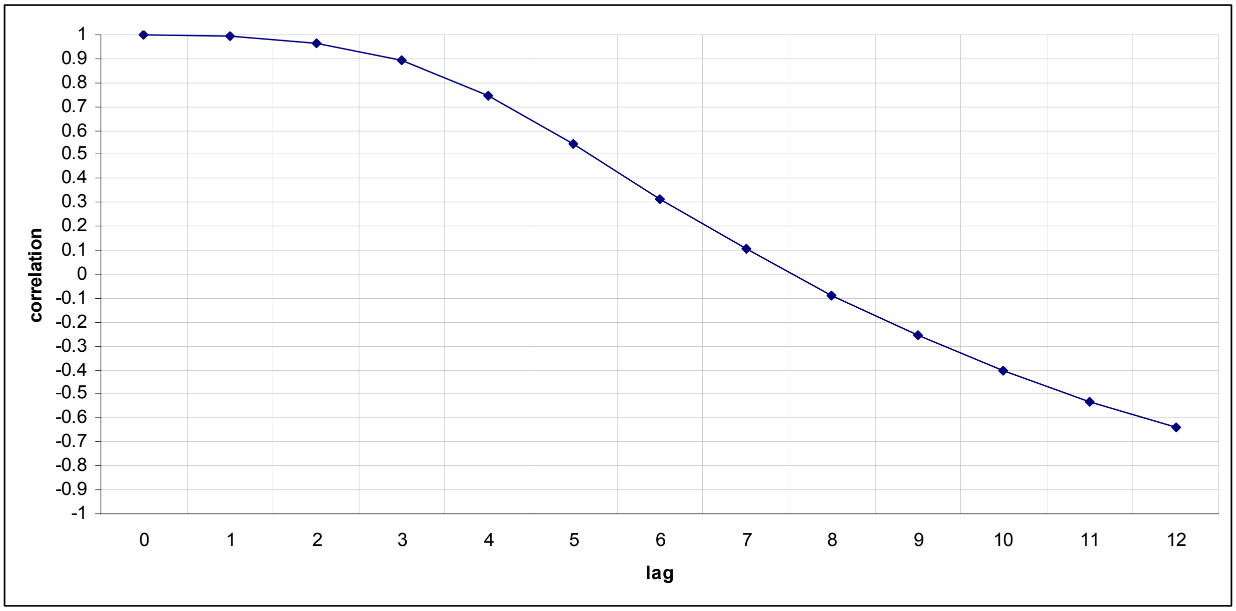
After removing the trend component, the residuals look like this:



Residuals after subtracting trend

The residuals start low, then go high, and then low again. This is certainly not a seasonal pattern linked to the calendar such as “we always miss high in winter and low in summer,” but it could conceivably be part of a cyclical pattern with some other period. One way to explore that is to look at the correlation of observations at different lags. This is the **auto-regressive** part of ARIMA.

To investigate further, we measure the correlation of the residual values at various lags. Such a chart is called a *correlogram*.



Correlogram for residuals after trend has been removed from base subscriber numbers

The correlogram starts with a correlation of 1 for lag 0 because, of course, any observation is perfectly correlated with itself. The correlation at lag 1 is also strong. This is typical of many time series. If yesterday was hotter than average, today is likely to be as well. In fact, weather forecasters have to work quite hard to beat the simple forecast that tomorrow will be just like today. By lag 7, the correlation is close to 0. At lag 12, it has become fairly strongly negative.

Which of the lags show strong enough correlation to warrant inclusion in the model? The test for significance of an observed correlation *r* based on N observations is taken from the *t* distribution.

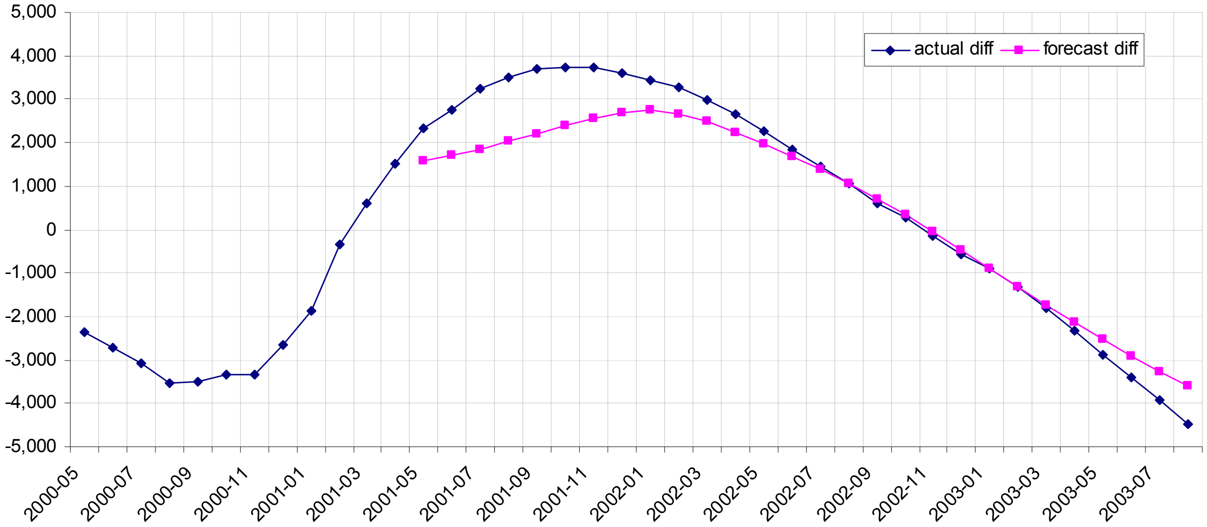
$$t = \frac{r}{\sqrt{(1-r^2)/(N-2)}}$$

Our observed correlation is calculated based on the 27 months of data for which lags 1 through 12 have been calculated. With N=27, any observed correlation that differs from 0 by 0.38 is significant at the 0.05 level. The correlations at lags 1 and 12 are strong enough to be considered significant. This suggests an auto-recursive model of the form $y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-12}$.

Fitting this model to the observed, de-trended data yields $y_t = -169.56 + 1.02 y_{t-1} - 0.09 y_{t-12}$. Note that this formula describes the residuals, after the trend has been removed. To get the full model, we simply subtract this predicted residual from the original linear forecast.

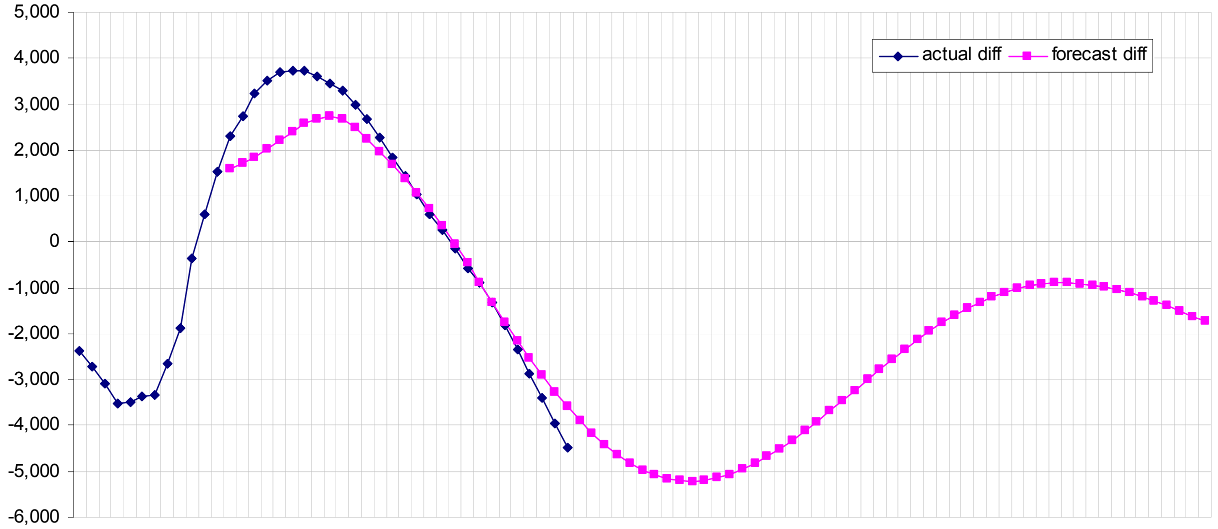
Forecasting the Past

It is tempting to apply the model to the training data to see how well it captures the pattern of stops we have seen in the past. After all, if we cannot predict the very data used to create the model, how can we expect to predict the future? In any case, assessing a forecast of the future requires patiently waiting for the future to arrive so the forecast can be compared with what actually happened. While waiting, we can try forecasting the past, which ought to be considerably easier. Because our model depends on lags 1 and 12, we use the actual values for May 2000 and April 2001 to forecast a value for May 2001. To forecast June 2001, we use the forecast value for May 2001 as the lag 1 value, not the actual value even though it is available.



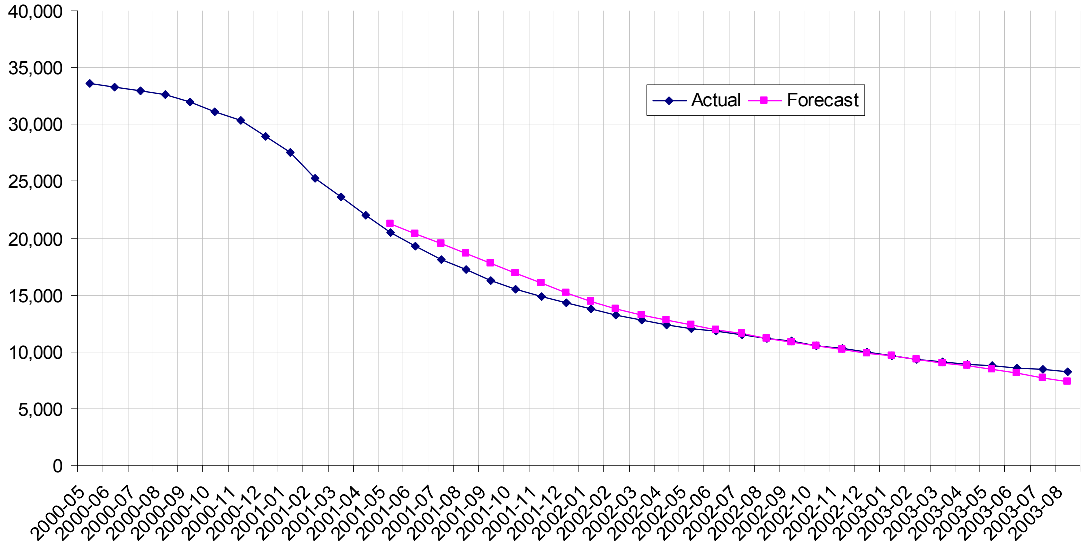
Auto-regressive portion of the forecast compared with actual values

The auto-regressive residuals model does appear to fit the data fairly well, at least for the period for which we have actual observations. It is a bit unsettling to realize that if we extend the forecast a bit further the predicted residual curve will head back up again. There is no obvious reason to think that the actual residuals will do the same. The model **describes** the training data without in any way **explaining** the training data. Nevertheless, we press on.



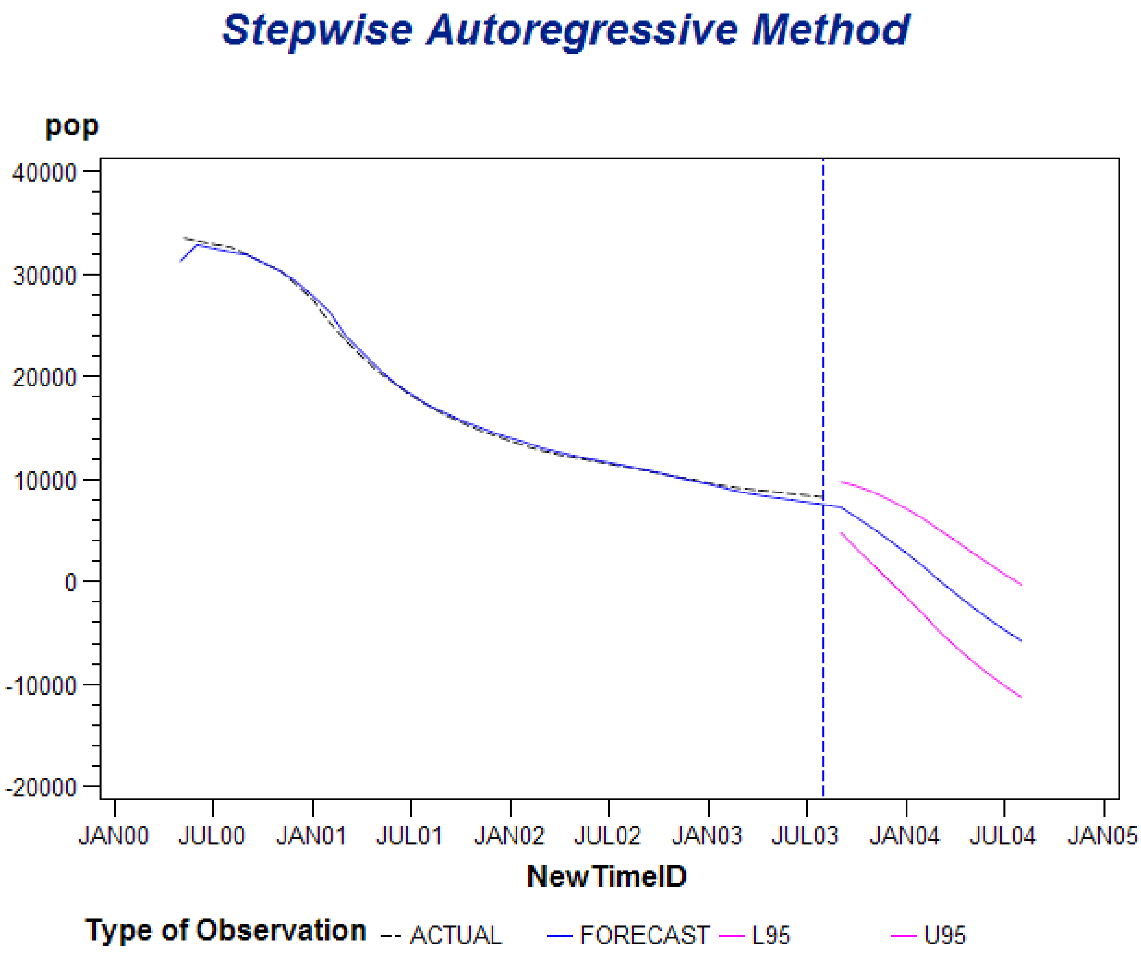
Will the actual residuals continue to follow the predicted pattern?

We achieve the final model by subtracting the forecast residuals from the linear forecast.



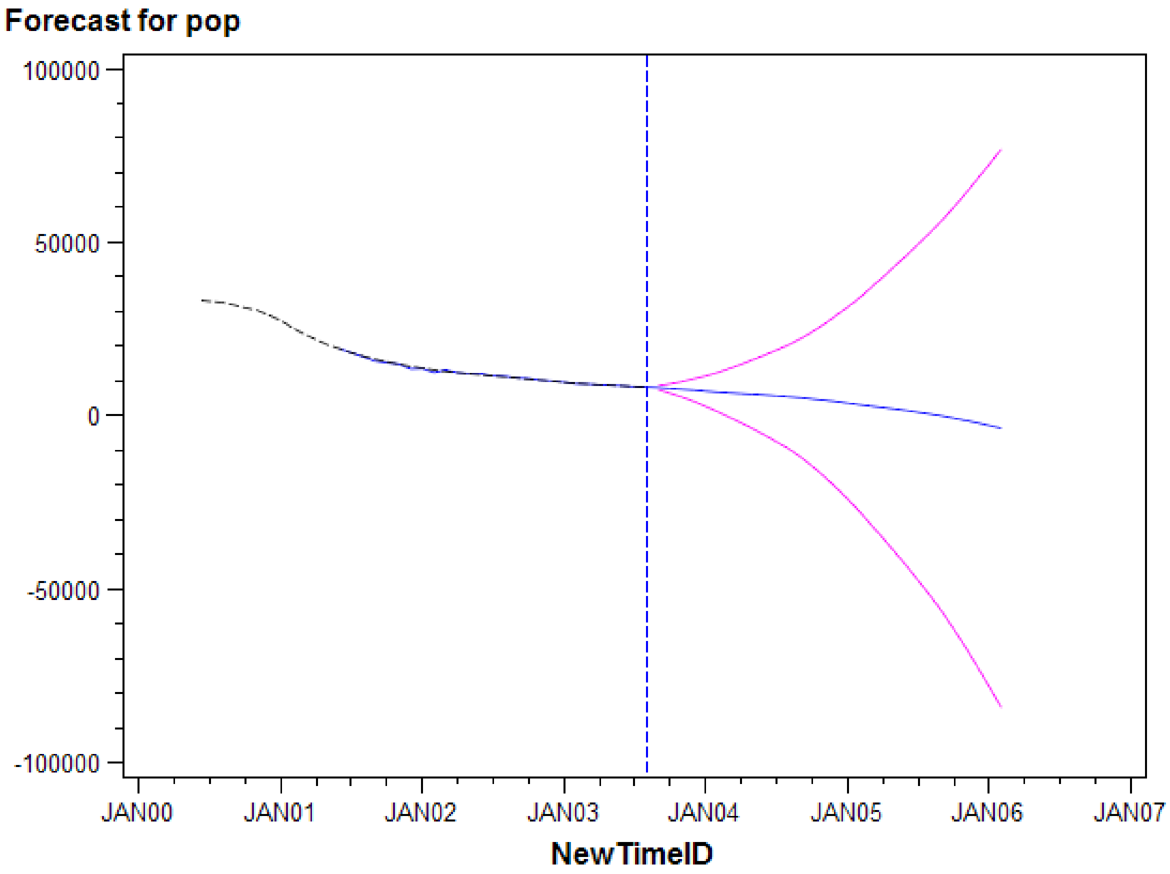
Final forecast including trend and auto-regressive component compared to actual values

Over the period for which we have actual data, the final forecast looks pretty good, although because of the lag it has nothing to say about the first 12 months of the forecast period. The forecast does have the problem that if we extend the forecast period farther into the future, extrapolating the downward trend in the size of the base population will cause it to merrily predict negative populations. The following chart was produced by the SAS Enterprise Guide Basic Forecasting node, which calls the FORECAST procedure. Given the same data as was used above, this procedure created a model with a linear trend component and autoregressive components for lags 1 and 7.



PROC FORECAST predicts negative population size

The more sophisticated ARIMA node calls the ARIMA procedure. This model also predicts future negative population sizes, albeit farther in the future.



PROC ARIMA forecasts negative population size

1.4 Problems with the Time-Series Approach

The time-series approach to forecasting easily captures major, aggregate-level patterns such as trend and seasonality, but not the causes of deviations from these major patterns. We have seen many cases where a forecast fails to predict fluctuations that, upon investigation, should have been foreseeable. For example, one client was surprised by a sudden spike in deactivations. Meetings were hastily called and managers demanded to know what was going on. Was there a sudden decrease in service quality? Had a competitor announced a particularly attractive offer? Actually, no. One year earlier they had run a very successful acquisition campaign that added many new subscribers on one-year contracts. Those contracts were now expiring and, **as expected**, many subscribers failed to renew. We emphasize **as expected** because the fact that many people cancel their subscriptions at the end of the contract period was well known. The problem was the forecast did not take that into account. The customer-centric approach to forecasting advocated in this paper is specifically designed to address two weaknesses of traditional forecasts:

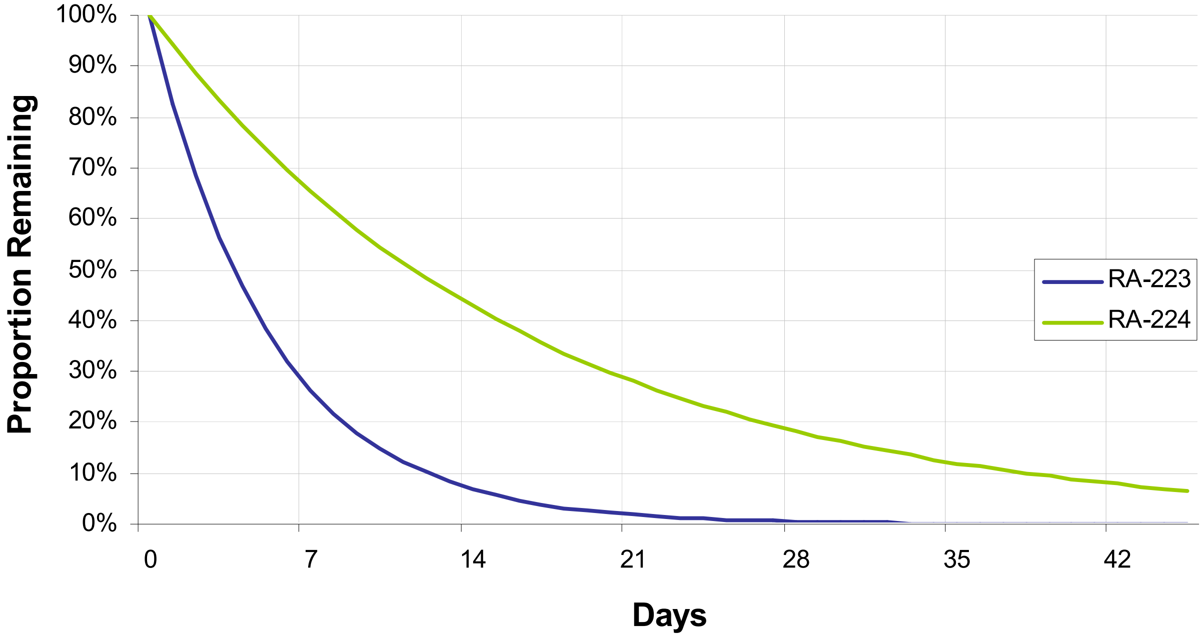
- lack of sensitivity to changes in the subscriber mix
- inability to drill down into the forecast to diagnose errors and discrepancies

The next two sections expand on these in turn.

Lack of Sensitivity to Changes in the Subscriber Mix

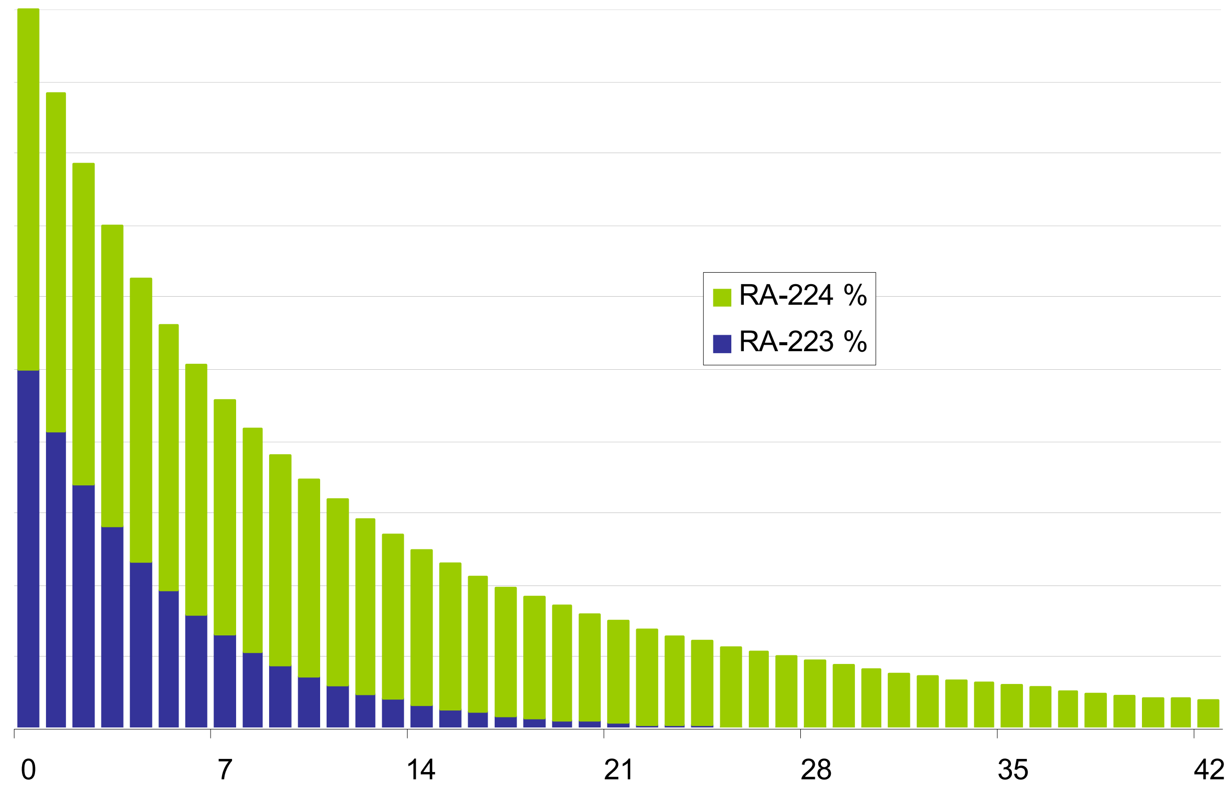
Past patterns in aggregate behavior are good predictors of future behavior as long as future subscribers resemble past subscribers. As companies grow, they typically expand into new markets, encounter new competitors, introduce new products and services, and do countless things that change the subscriber mix. A population once dominated by well-off, style-conscious, urban, early-adopters may come to be dominated by middle-income, suburban, value-conscious subscribers. Such changes are bound to affect the attrition rate, but how?

Subscriber populations are complicated, so we often use an example from the simpler world of physics and radioactivity to illustrate the point. Consider two isotopes of radium. Chemically, they are the same, but they decay at very different rates. One, radium 223, has a half-life of 3.6 days; another, radium 224, has a half-life of 11.4 days.



Survival curves for two isotopes of radium

If we start with a half-and-half mixture of the two isotopes, the mixture will decay at a varying rate over time. The proportion of the two isotopes changes because one of them decays faster than the other.



Changing proportions of two isotopes with different rates of decay

The mixture, which started out half-and-half, quickly comes to be dominated by the slower-decaying radium 224. The analogy with customer populations is clear. Think of the two isotopes as customer segments defined by variables such as credit class, rate plan, or geographic market. The overall population is an ever-shifting mix of these. And to further complicate matters, in contrast to isotopes of radium that decay at a constant rates, each customer segment has a deactivation rate that is a function of subscriber tenure.

The customer mix changes due to changes in acquisition strategies, changes in the competitive landscape, and maturing of the customer base. Customer-centric, survival-based forecasting takes all this into account.

Inability to Drill Down to Diagnose Errors

Was it really Yogi Berra who said, “It’s hard to make predictions—especially about the future”? Or was it Mark Twain? Or perhaps Confucius? In any case, the saying’s enduring popularity is due to its evident truth: Forecasts are often wrong. When a forecast misses the mark, it is natural to want to know why. If the forecast is based on aggregate numbers, there is no satisfactory answer. Answers such as “The slope of the trend component seems to have changed” or “Seasonal effects were not as pronounced as in the past” do not shed much light.

In customer-centric forecasting, the aggregate forecast is the summation of independent forecasts for each customer segment. Each of these mini-forecasts can be assessed independently. This makes it possible to answer the question “What went wrong?” with something like “The hazard probability for the outbound telemarketing channel was way up. Did we switch call centers?” or “Something has changed drastically in Denver. Is there a new competitor there?”

The ability to track forecast performance along multiple dimensions means that localized problems are flagged even when the overall forecast is highly accurate.

1.5 Survival-Based Forecasting

Customer-centric forecasting need not involve survival analysis. Any forecasting method can be employed for customer-centric forecasting by using it to provide separate forecasts for each segment based on customer characteristics and then aggregating those segment-level forecasts to get an overall forecast. 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 the 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

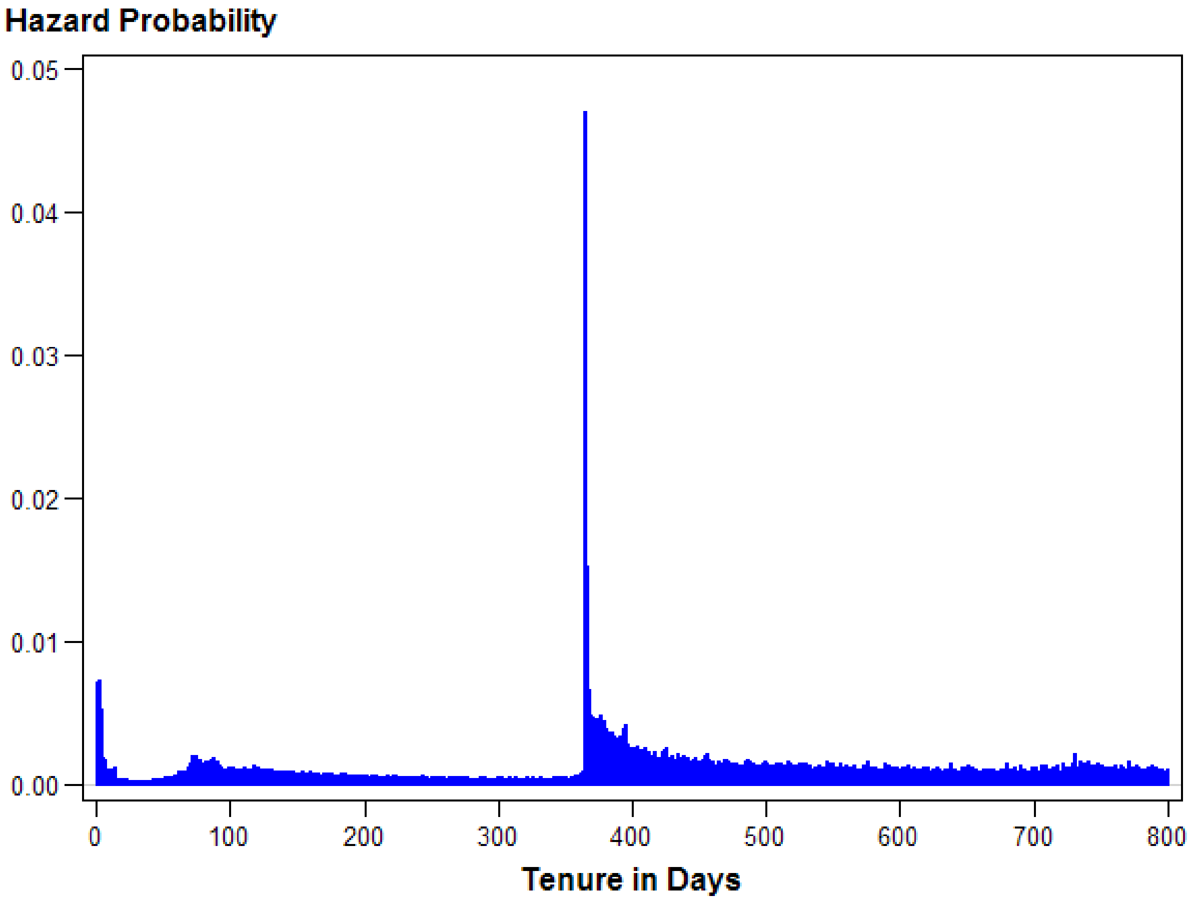
The descriptions and definitions we give here differ from ones you might have seen elsewhere. Survival analysis is most often introduced in the context of clinical trials. Without getting into too much technical detail, these differences mostly stem from the very large quantities of data typically available 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:



Deactivation hazard (any reason) for first 800 days

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

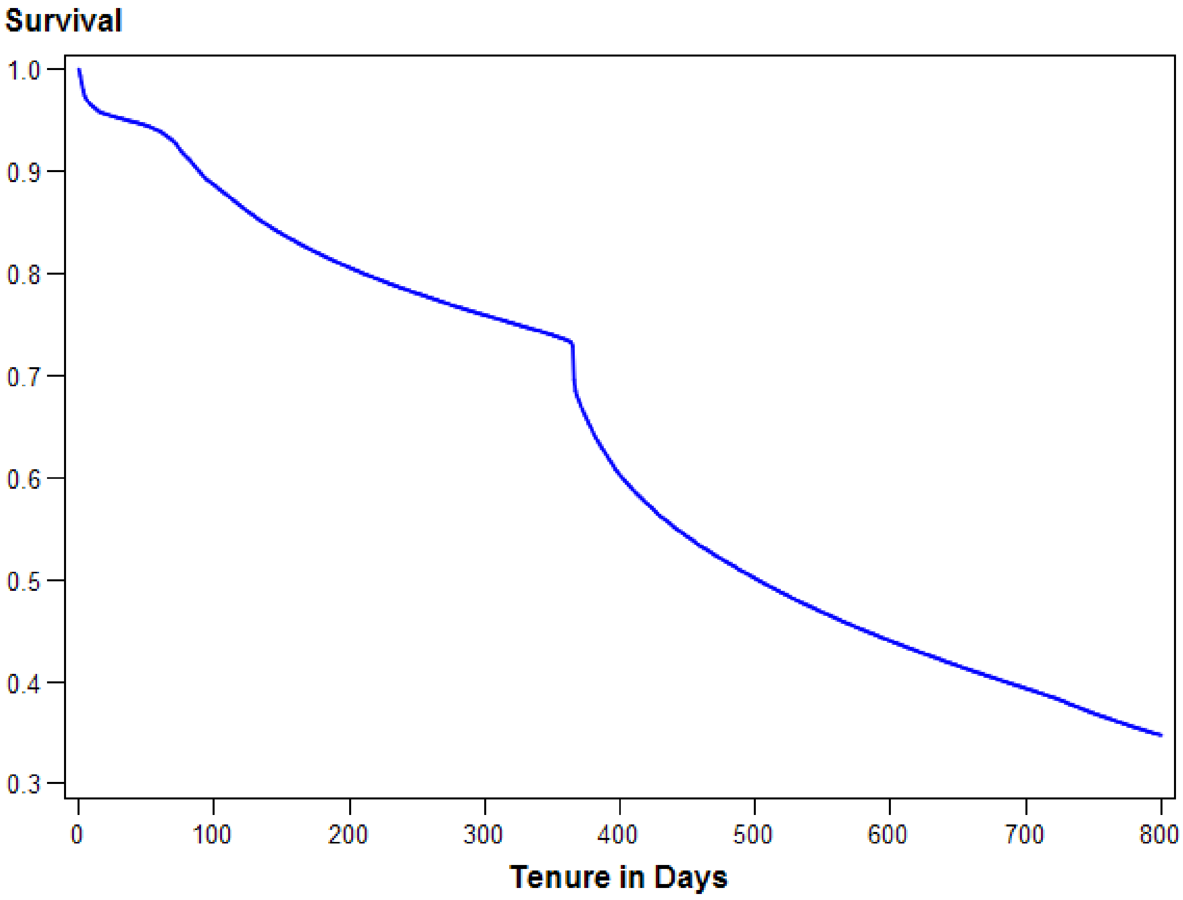
Although the hazard plot is interesting in and of itself, the primary reason for calculating hazard probabilities is to compute the survival curve. The survival curve for a customer segment shows what proportion of the segment survives to each tenure. Because deactivated customers do not come back (or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

$$S(T) = \prod_{t=1}^{T} (1 - h(t-1))$$

$$S(0) = 1$$

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 corresponding to hazard plot in previous illustration

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:

Day	Hazard
0	0.006170
1	0.007202
2	0.007447
3	0.005407
4	0.002587
5	0.002031
6	0.001651
7	0.001804
8	0.001267
9	0.001088

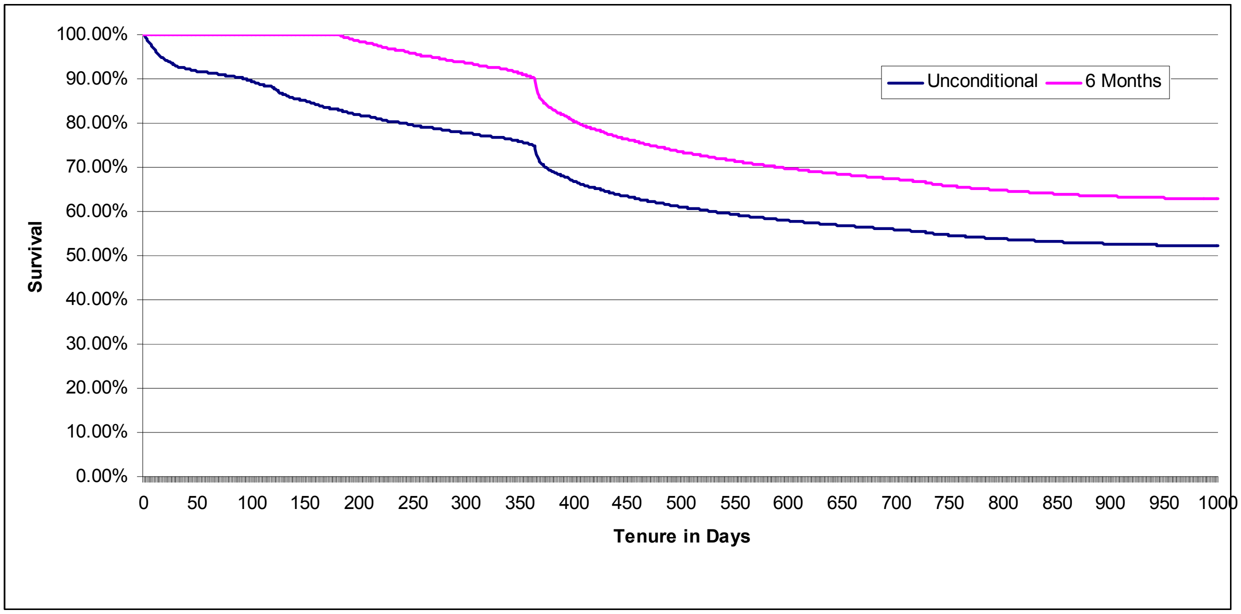
The Day 0 hazard of 0.006170 means that of the 1,000 D1 subscribers that started today, we expect 1,000*0.006170 = 6.17 to fail to make it until tomorrow. That means we expect 993.83 of these subscribers to be around tomorrow to be exposed to the slightly higher Day 1 hazard. Of these, 993.83*0.007202=7.16 will fail to see the following day, their Day 2.

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

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>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.



Conditional survival for subscribers lasting at least six months compared 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:

	Cohort 1	Cohort 1	Cohort 1	Cohort 2	Cohort 2	Cohort 2	Cohort 2	Total
Day	Hazard	Remaining	Lost	Survival	Hazard	Remaining	Lost	Lost
0	0.006170	1000	6.17	1	0.002031	800.00	1.62	7.79
1	0.007202	993.83	7.16	0.997969	0.001651	798.38	1.32	8.48
2	0.007447	986.67	7.35	0.996322	0.001804	797.06	1.44	8.79
3	0.005407	979.32	5.30	0.994525	0.001267	795.62	1.01	6.30

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. Note that, in contrast to the ARIMA forecast, predicted losses in the survival-based forecast are always a fraction of the remaining population, so this forecast method will never predict a population size smaller than 0.

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 in "Forecasting Future Stops for New Subscribers" 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.

Where Does the Add Plan Come from?

Ideally, the acquisition plan has been developed jointly by the marketing and finance departments and has been organized along planning dimensions such as market, channel, and product that are also suitable for defining customer segments for forecasting. If the acquisition plan has not been broken down by suitable dimensions, we can use historic data to allocate planned starts to segments in the proportionally. If the forecast is to be made at the daily level, it will almost certainly be necessary to allocate daily starts from a monthly acquisition plan. As a last resort, we can forecast starts using a simple trend plus seasonal effects model as described in Section 1.3, "The Usual Approach to Forecasting." In general, it is not advisable to attempt to model things that are or should be largely in the company's control.

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?”

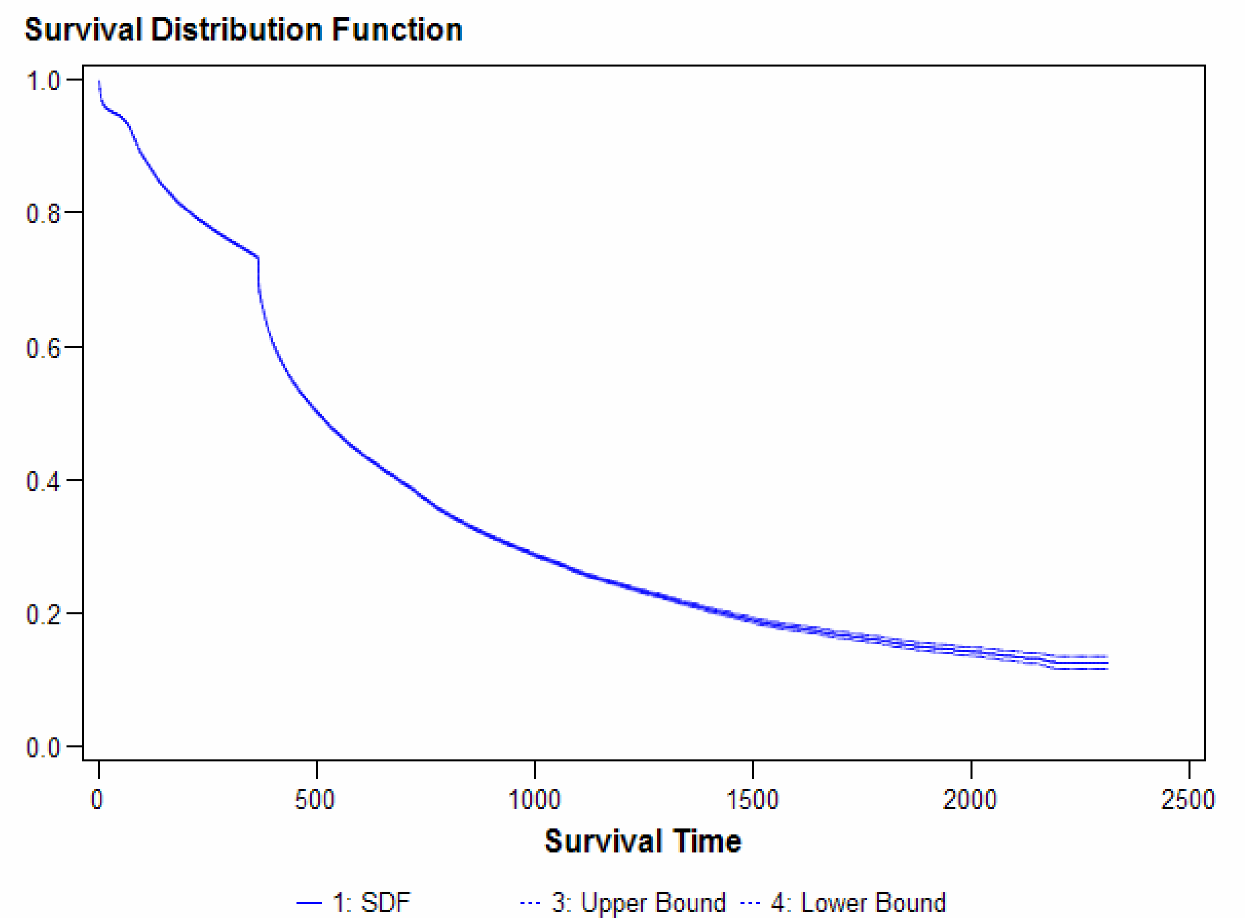
The initial evaluation of the forecast is made at the aggregate level, but with a customer-centric forecast it is possible and desirable to also evaluate each cohort-level forecast. Because they are based on smaller populations, these will naturally show greater variance than the aggregate forecast. 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. In the following section, we take this approach to compare a very simple survival-based forecast with the equally simple time-series-based-forecast presented earlier.

1.6 Applying the Survival-Based Method

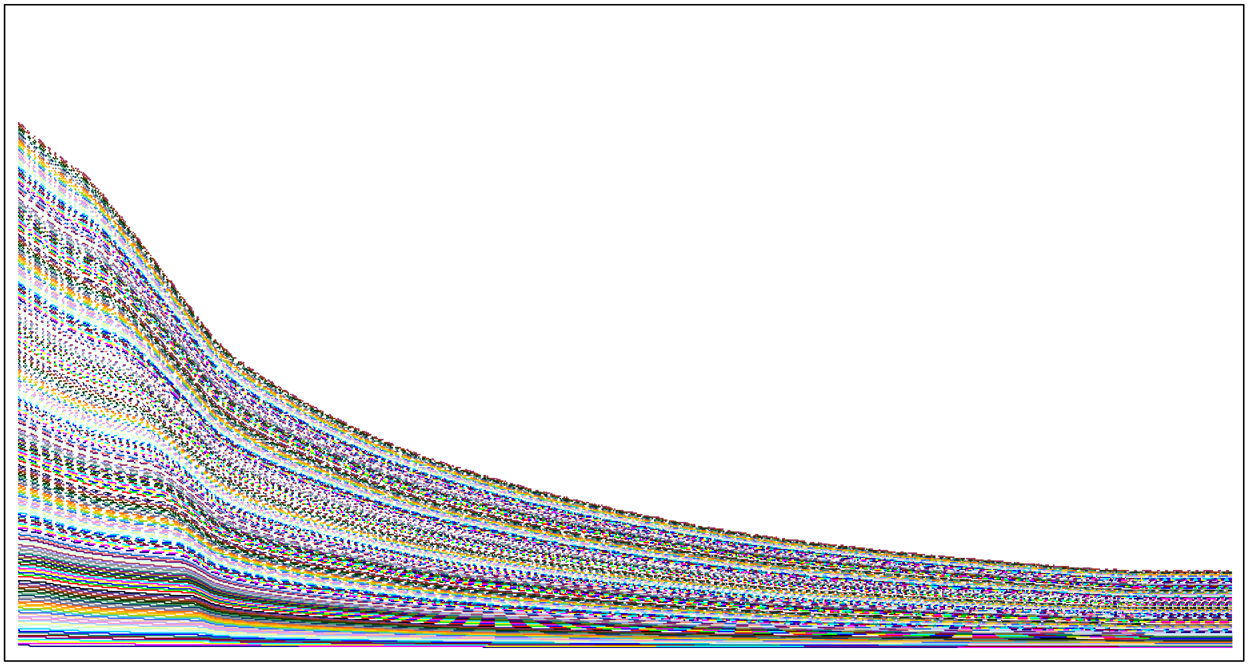
We now show how to create the existing base forecast using the survival-based approach. Normally, the first step is to define customer segments because each segment should have its own family of survival curves. Several good variables for segmentation are available including market, acquisition channel, 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 in "The Hazard Probability" in Section 1.5. 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. Charts showing the hazard probability and survival curve for this data are in "The Hazard Probability." 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.



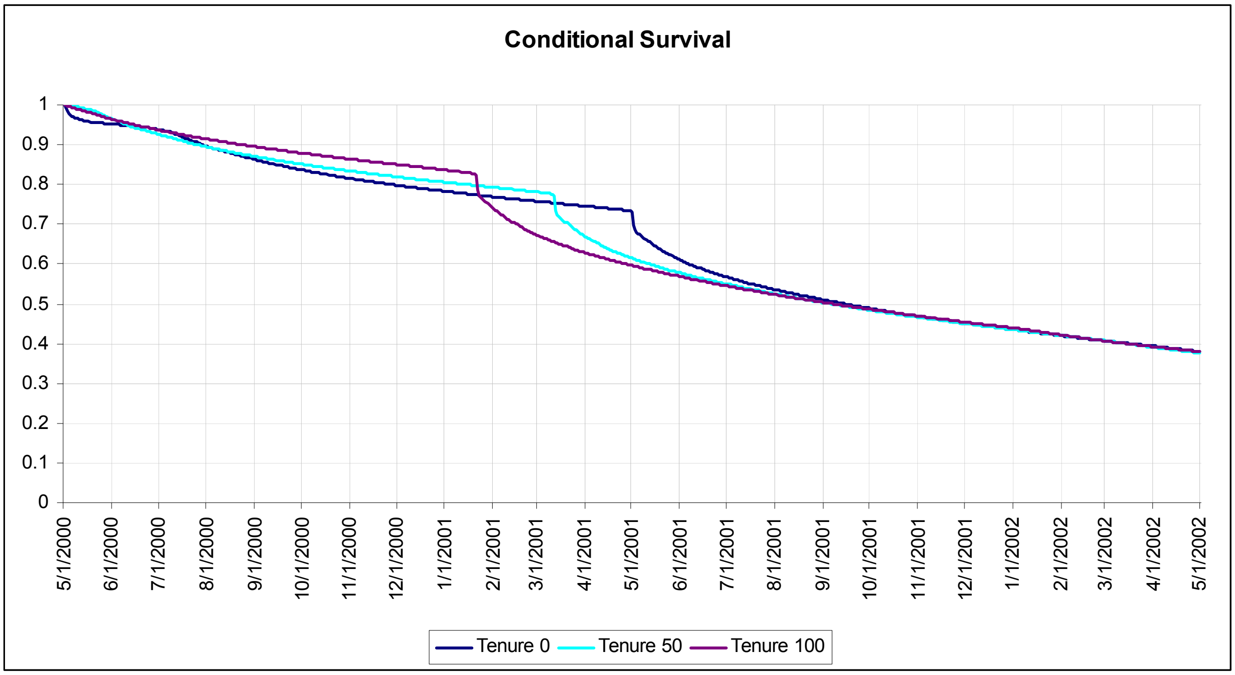
Survival curve with 95% confidence bounds

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.



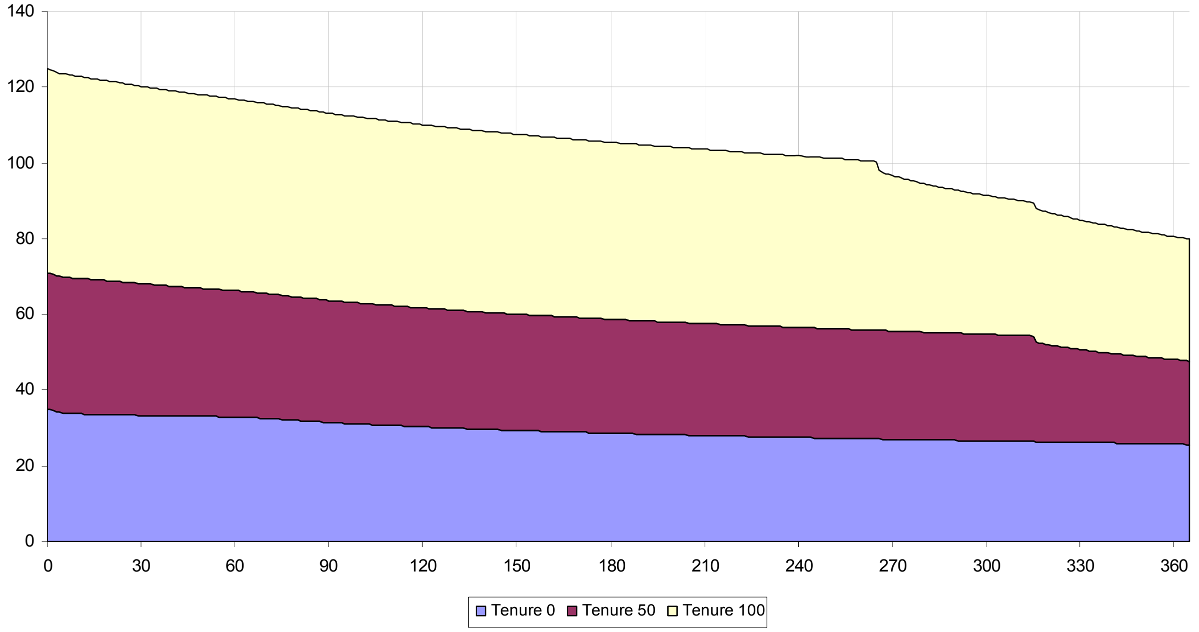
Overall survival is the sum of the survival of many cohorts

This chart shows the contribution towards overall survival of a few hundred of the more than a thousand cohorts used in this example. The following chart shows just three cohorts, so it gives a better idea of how things work.



Conditional survival curves for selected tenures at start of forecast period

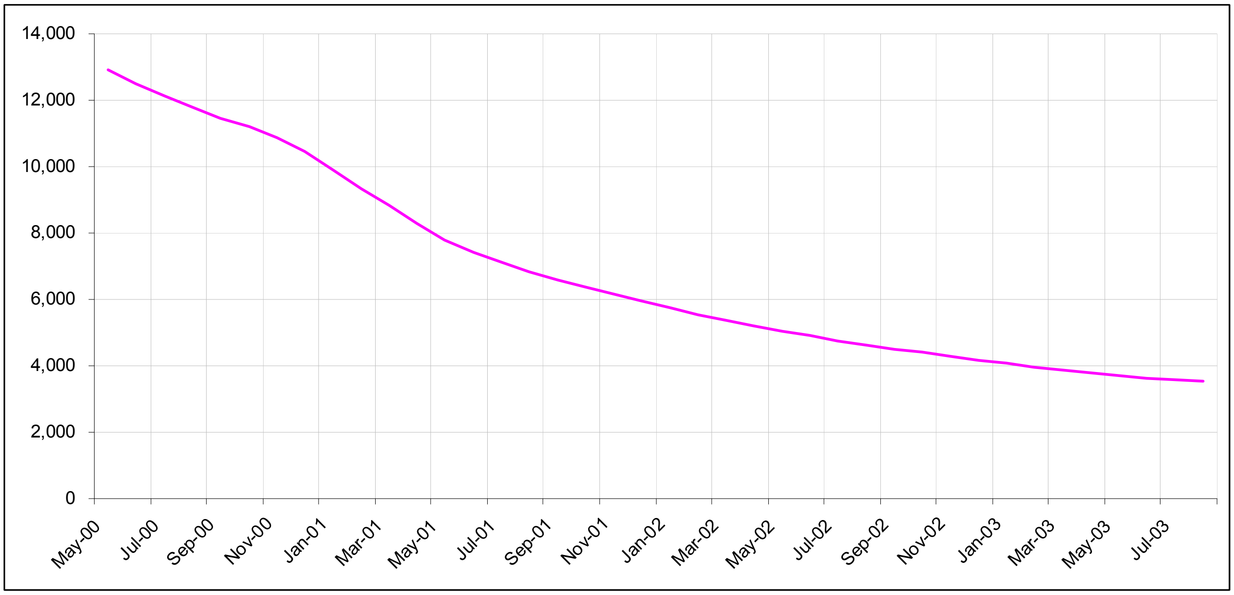
The survival curves for three cohorts are shown here: the cohort that started on 01 May 2000, the cohort that had tenure 50 on that day, and the cohort that had tenure 100 on that day. By multiplying the conditional survival curves for each cohort by the size of the cohort on the first day, this chart 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.



Contribution of three cohorts to overall stops in first year of forecast

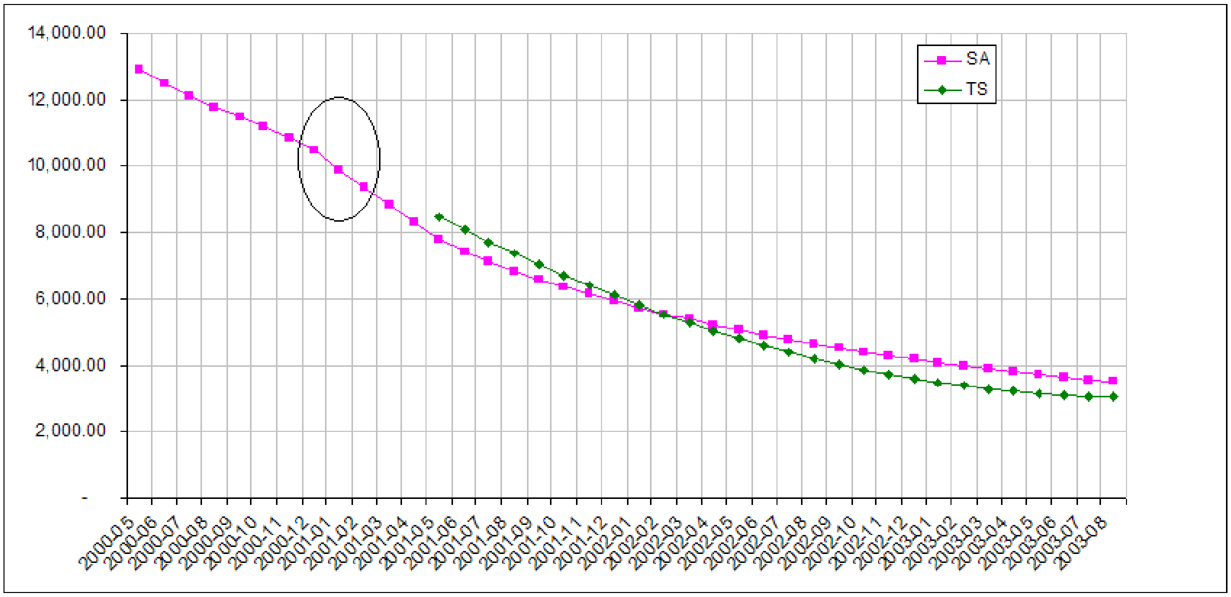
To get the final forecast, we simply sum the expected surviving population from each cohort. We lump all subscribers with tenure over 1,100 into a final cohort for 1,100 and up. Note that when calculating the expected attrition for days far in the future for cohorts that already had high tenure at the start of the forecast period, we reach some tenures for which no estimate of the hazard probability has been calculated. Beyond about tenure 2,000, there is not enough data to make good estimates of the hazard. In this simple model, we handle this by assuming a small, constant hazard of 0.0005 based on the average hazard estimate for tenures over 1,500 days. A constant hazard for high tenures is not as unreasonable as it might first sound. When customers reach high tenures, the effects of variables such as credit class and acquisition channel have worn off. These subscribers have demonstrated loyalty and willingness to pay. Much of the attrition from this group is of the unavoidable kind; subscribers die or move out of the service area.

To get the base forecast, we take the starting population for each cohort and multiply it by the survival probability for each tenure. Summing these produces a daily forecast for the whole subscriber base.



Daily forecast for the existing base on the test data

In order to compare this forecast with the time-series forecast, we aggregate it to the monthly level.

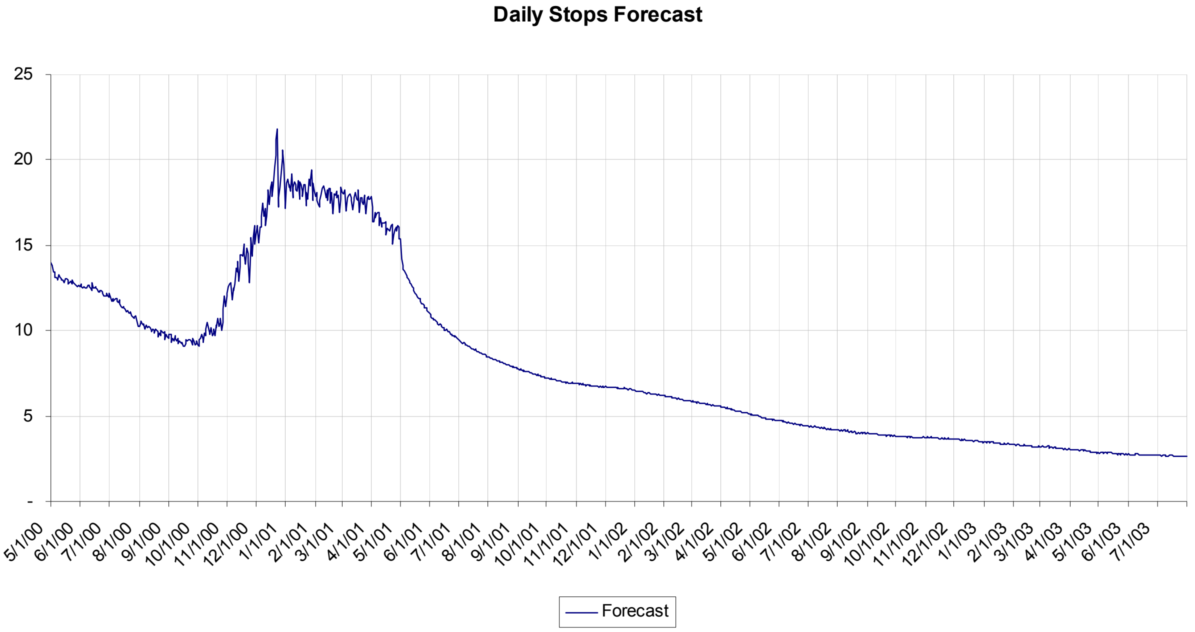


Comparison of two existing base survival forecasts on the test data

For the period where both forecasts are able to produce estimates, the two forecasts don't look much different. Is there really a reason to prefer one to the other? As explained in the following section, the answer is yes. The survival-based forecast is more useful because it allows us to drill down into the forecast to see the contribution made by each cohort and explain anomalies such as the steeper drop in population predicted for the highlighted period around January 2001.

Why the Survival-Based Forecast Is Preferable

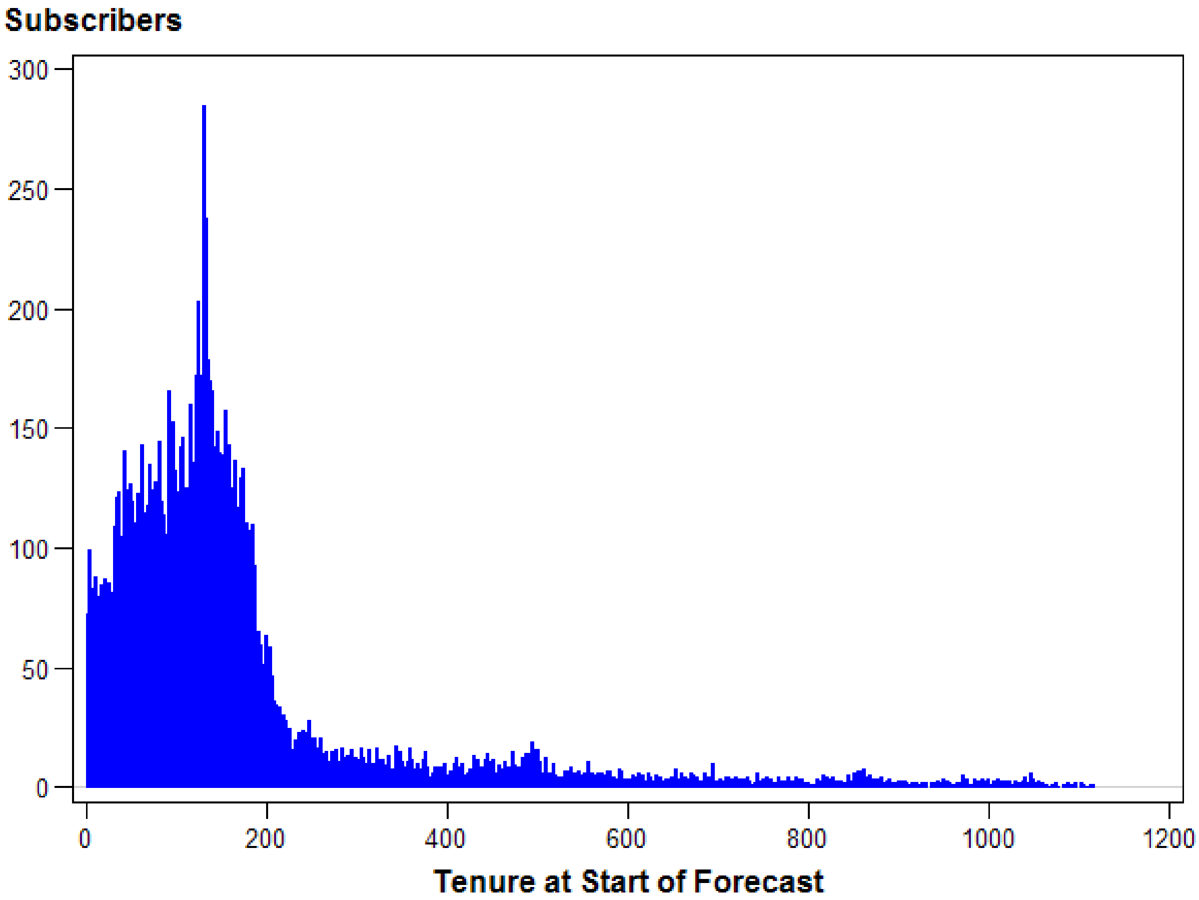
The added richness of the survival forecast is best seen by looking at forecast stops rather than forecast survival. The survival forecast is easily converted into a stops forecast by subtracting the forecast number of survivors at each time *t* from the number at time *t*-1. The following chart is a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



Daily forecast for stops in test data from among subscribers active on 01 May 2000

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. Fewer 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?

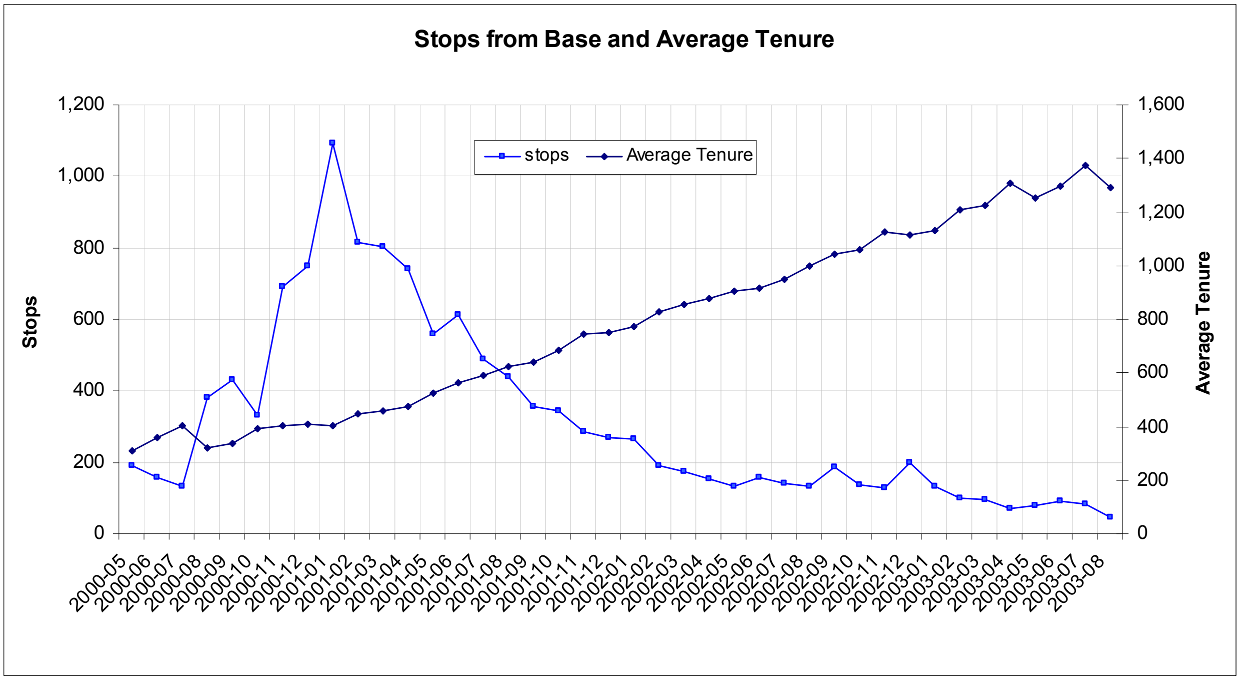
The answer lies in the distribution of tenure cohorts.



Distribution of tenures of active customers as of 01 May 2000in the training data

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.



Actual stops and change in average tenure of the base over the forecast period

Notice that the average age at deactivation is around one year for most of the first year of the forecast period. Then, when all the people in the base have reached tenures over one year, the average tenure at deactivation increases steadily as expected and the number of deactivations per month begins to decline as the base shrinks.

The actual data shows an even greater surge in deactivations around January of 2001 than predicted by the model. Something else must be going on that this very simple model fails to capture. We emphasize that the models compared here are the simplest imaginable for each class of model. Neither includes any covariates such as market, handset type, rate plan, or credit rating, all of which are known to be important. Neither distinguishes between voluntary and involuntary deactivations. Neither takes calendar time effects into account such as the tendency of people to give phones as Christmas presents which means that many one-year contracts end when the year ends. Adding these explanatory variables to the time-series model can be quite difficult. Adding them to the survival-based model is quite simple. In this example, the hazard estimate for each tenure is a simple ratio of the historical count of subscribers who have stopped with that tenure to the total number who have been at risk for that tenure. To incorporate additional effect, we make the hazard estimate a function of the new explanatory variables. There are many approaches to modeling hazard probabilities, the details of which are beyond the scope of this report. The simplest is stratification: hazards are calculated separately for each market, credit class, start month, and so on.

The survival analysis framework also makes it easier to account for the effects of explanatory variables. Clearly, factors such a 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.

In conclusion, because it is customer-centric, even the simplest survival-based model captures effects of a changing subscriber mix that are very hard to capture with a traditional, time-series-based approach. Furthermore, the simple survival model is easier to extend and build upon than the simple time-series model to which we have compared it.