Direct Estimation of the Survival Function for Credit Default Using Nested Time Intervals*

Jeff Dugger, PhD and Michael McBurnett, PhD
Equifax Data Science Labs
Atlanta, GA USA

This paper describes a patent pending method for constructing nested survival models refining prediction of timing to default for credit risk. Traditional credit risk models focus on predicting the probability that an account will default in a defined performance window, typically 24 months. This paper focuses on predicting the probability of when an account will default in the performance window. Knowing when a borrower is likely to default allows lenders to price term loans more accurately as well as help them predict time between defaults for a given customer so they can better manage portfolio risk. Survival analysis predicts the probability of when an event will occur by using three interrelated functions: the survival function, the hazard function, and the probability function. The survival function predicts the probability an account will remain good up to a given time, the hazard function provides the default rate over time, and the probability function shows the distribution of default times. We explore nested survival models predicting auto loan defaults. The nested survival model is equivalent to a set of standard credit risk models on nested time intervals. Time intervals can be optimally selected in all cases. In regimes where adverse action codes are required, regulatory compliance can be guaranteed for these models. We demonstrate the performance of nested survival models.

Introduction

Traditional credit risk models focus on predicting the probability that an account will default in a given performance window, typically the next 24 months. However, much more useful information can be gained from predicting the probability of when an account will default within the performance window. Knowing when a borrower is likely to default will allow lenders to more accurately price term loans as well as help lenders better manage their loan portfolio risk.

Extensive research has been published discussing applications of survival analysis to credit risk modeling. A good overview of this literature as well as a collected outline of various modeling approaches appears in [1]. While we originally explored applying survival analysis to credit risk modeling for the purpose of helping auto lenders better price auto loans, another intriguing application for survival models in credit risk regards incorporating macroeconomic data along with borrower behavioral data into models for stress-testing bank retail credit portfolios [2] [3].

We propose a very simple way to model survival functions directly as a set of standard credit risk models built on nested intervals obtained by subdividing the performance window into discrete, overlapping sets which cumulatively indicate occurrences of default over time. This approach has the benefits of being easily integrated into standard credit modeling practice, being easily explainable to bank risk managers, and complying with regulations in jurisdictions requiring model explainability.

Overview of Survival Analysis

Survival analysis evaluates the probability of surviving up to an instant of time \( t \) [4] [5]. In credit risk, this is the probability of remaining “good” on an account until time \( t \), i.e. not being in default on an account. Right censoring, a key concept in survival analysis, occurs when the

---

* (c) Equifax, Inc., 2019, all rights reserved
event of interest has not happened during the period of interest. Left censoring occurs when we cannot observe the time the event occurred. In cases of loan defaults, this does not apply. In the most common case, right-censoring, the event occurs beyond our performance window, if at all [4] [5].

Define \( T \) as a non-negative random variable representing time until default occurs [5]. The distribution of default times is defined by a probability function, \( f(t) \). The function typically estimated in survival analysis is the hazard function, \( h(t) \), which defines the bad rate over time. The primary function of interest is the survival function, \( S(t) \), as discussed above. These three functions are interrelated as defined in Equations (1)-(3) below [5]. These equations apply to the discrete-time case, which is appropriate for credit modeling since data is reported on a daily basis.

Equation (1) gives the mathematical definition of the survival function as

\[
S(t_j) = P(T > t_j)
\]  \hspace{1cm} (1)

Once the survival function is known, the probability density function and hazard function can respectively be derived as

\[
f(t_j) = S(t_{j-1}) - S(t_j)
\]  \hspace{1cm} (2)

\[
h(t_j) = \frac{f(t_j)}{S(t_{j-1})}
\]  \hspace{1cm} (3)

Survival analysis comprises three broad modeling approaches—parametric, non-parametric, and semi-parametric [4] [5].

The parametric approach assumes a specific functional form for the hazard function. Commonly used probability density functions from which parametric hazard functions are
derived are the exponential, Weibull, log-normal, log-logistic, and gamma distributions. In all cases, the parameters can be fit from the data using maximum likelihood [4] [5].

The Cox Proportional Hazards (CPH) model is the most widely used non-parametric model in survival analysis. This approach assumes all cases have a common hazard function of the same functional form. A regression model provides scale factors that translate this “baseline” hazard function into survival functions for the various predicted cases. This method requires selecting a particular set of coefficients as a “baseline case” for which the common hazard function can be estimated [4] [5].

Semi-parametric methods subdivide the time axis into intervals and assume a constant hazard rate on each interval, leading to the Piecewise Exponential Hazards model. This model approximates the hazard function using a step-wise function. Intervals can be identically sized or optimized to provide the best fit with the fewest models. When time is discrete, a logistic regression model can be used on each interval [5]. One particular advantage of this approach over parametric and non-parametric modelling techniques is that it is more flexible because it doesn’t require the assumption of a fixed parametric form across all time.

**Direct Estimation of the Survival Function Using Nested Time Intervals**

Standard credit risk models typically define a performance window of 24 months. In our data, consumers who go 90 days past due (dpd) or worse are defined as default whereas those
who do not are defined as not-default. In the context of survival analysis, a standard credit risk model could be considered as representing the probability of a consumer remaining “good” up to 24 months. Those consumers who are “good” in a 24 month window may or may not go “bad” later, and would be considered right-censored.

Inspired by the semi-parametric approach to survival analysis discussed earlier, we subdivided the time axis into nested intervals of overlapping performance windows and built logistic regression models on each to directly estimate the survival function as shown in Figure 1a. This is equivalent to building several credit risk models on performance windows of increasing duration and leads to direct estimation of the survival function, as will be shown. Figure 1b illustrates this concept and shows how one could linearly interpolate between model predictions for a better estimate of the underlying survival function.

While logistic regression is by far the most common approach to building standard credit risk models, any binary classification scheme, such as neural networks, decision trees, etc. can be used. Because standard credit risk models are used, this approach complies with regulations in jurisdictions that require explanatory models.

**Verification with Simulated Data**

Simulated data provides a known test-case against which to determine the feasibility of the nested-interval survival function modeling approach. The simulated data set consists of a sample containing 200,000 observations from a set of five log-normal distributions as shown in Figure 2a. We chose the log-normal distribution as it is one of the simpler standard parametric models used in survival analysis.
In the simulation, we built 21 logistic regression models on nested intervals from month 3 to month 24 with single-month increments. Figure 2b compares model results for each distribution with Kaplan-Meier estimates direct from the data. The nested interval logistic modelling approach compares very well with the actual survival function.

Figure 2c illustrates probability density functions derived from these survival function models using (2) above; taking differences tends to amplify disparities between the modeled and actual survival function as seen in the figure. However, the results are still quite good.

**Results on Auto Loan Data**

We applied the set of nested-interval survival models to real auto data from the Equifax credit database. We initially considered data on accounts opened over a twelve-month period covering June 2012 through May 2013 to observe time-to-default over a longer performance window than two years. Initial histograms of time-to-bad looked nearly uniform until we examined default-time distributions segmented by loan term-length as shown in Figure 3. As the figure shows, most loans are for a 60-month or 72-month term, and in these cases the distributions are close to uniform (after an initial “ramp up”) and even trends slightly upwards in the 72-month case. In both the 60 and 72 month cases, the median default rate is right censored. These loans disguise the structure seen in the 36 and 48-month term cases which do show time-to-default distributions that look similar to the log-normal simulated data we examined above. In what follows paper, we focus on modeling results from the 36-month term case.

We constructed our final modeling data set based on auto loan accounts with a 36-month term opened over a period of twelve months covering June 2014 through May 2015 and observed the time-to-bad performance of each account over a two-year window.

Think about how a consumer defaults on a loan from the timing perspective. It takes at least three months for the loan to be in default (in general, default for retail bank products is 90 days past due or worse). Therefore, unless one “straight rolls” by not making any payments on the loan once originated, one should expect few defaults in three months. That is exactly what we see in Figure 3 above, top left. In fact there are three consumer defaults in month three. This is insufficient to build a reliable model. While our simulation worked reasonably well following the
log-normal distribution because we could “invent” any number of cases defaulting in any month; in the real data, we have far too few observations in early months to build a stable model. This led us to the idea of nesting the time intervals.

Figure 4 illustrates modeling results for real auto account data. We constructed each survival function in the plot by averaging model predictions for every individual who defaulted within a particular time interval, e.g. 6 months. In actuality, there will be a distribution of survival functions for individuals going bad at a particular time. The various colored line segments in Figure 4b distinguish consumers who defaulted in month 3 to month 24 in increments of 3 months. Also shown is the survival function for consumers who did not default in the 24-month performance window labeled as \( t = 36 \), because we assume these consumers survived to complete their loan given no other information—i.e. they were right-censored. We initially built a set of 20 logistic regression models on nested-intervals covering month 3 through month 23 as shown in Figure 4a. Modeling over performance windows of 6, 12, 18, and 24 months provides more observations for building stable models as mentioned above, but yields a coarser approximation as shown in Figure 4b. This is equivalent to subsampling the original set of 20 models (and in this case, adding an extra model for month 24). Building monthly models and selecting subsets allows us to calibrate between model reliability and refinement of survival function approximation.

In an ideal case, the survival functions would look like a set of sharp sigmoid functions surrounding the respective time-to-default for each group, and the distribution of individuals around each average would be narrow. Our particular data set does not show this ideal case; in fact our data revealed a very wide dispersion in predicted survival times for consumers in each on the intervals we tested. We may be able to remedy this by extending our model to cover time-varying covariates or adding more informative features to our model.

**Conclusion**

Knowing the likelihood when a borrower as well as if a borrower will default gives a lender additional information useful for pricing loans and managing their credit portfolio. Survival analysis is a well-known statistical approach aimed at modeling how predictor variables influence time-to-event. Our research demonstrates a very simple approach to directly estimate the survival function by building a set of standard credit risk models on a set of nested intervals indicating the likelihood of when a borrower will default.

**References**


