Journal of the Operational Research Society (1999) 50, 1185–1190 0160-5682/99 $15.00
http://www.stockton-press.co.uk/jors

Not if but when will borrowers default

J Banasik, JN Crook and LC Thomas*
University of Edinburgh, UK

Credit scoring systems are based on Operational Research and statistical models which seek to identify who of previous borrowers did or did not default on loans. This study looks at the question when will borrowers default not if they will default. It suggests that some of the reliability modelling approaches may be useful in this context and may help identify who will default as well as when they may default.

Keywords: credit scoring; proportional hazards models; accelerated life models; competing risks

Introduction

Historically credit scoring systems were built to answer the question how likely is this applicant for credit to default by a given time in the future. The survey papers of Hand and Henley,[1] Rosenberg and Gleit,[2] and Thomas[3] outline the different modelling techniques that can be used to build such systems—discriminant analysis, logistic regression, partitioning trees, mathematical programming, neural networks, expert systems and genetic algorithms. With all these techniques the methodology is to take a sample of previous customers and classify them into 'goods' and 'bads' depending on their repayment performance over a given period. Similarly many behavioural scoring systems are based on the idea of how likely is someone with this recent performance pattern to still be performing satisfactorily for a fixed period in the future. Other behavioural scoring systems use the Markov chain idea of a borrower moving from state to state. The systems try to identify whether the borrower will end up in unsatisfactory states. This Markov chain approach emphasises that there is a dynamic element to credit status and borrowers default at different times in their credit history. The statistically-based classification approach imposes a static element over this dynamism by concentrating on an applicant's status after a fixed period of their credit history.

Once one accepts that credit status has a dynamic element then the underlying question changes. One asks not if an applicant will default, but if they default when will this be? This is a more difficult question to address both because there can be several answers not just the yes/no of the if question, but also because for many past customers the data is censored in that they cease to be borrowers (either by paying back the loan, or by death) before they default. However there are advantages in trying to answer this question. These advantages include:

(i) that such estimates of when an applicant defaults will give a better view of the likely profitability of the applicant and hence is a first step on the road to profit scoring.
(ii) that such estimates will give a forecast of the default levels as a function of time. This would be useful for firms' debt provisioning.
(iii) the estimates may guide the decision on the length of time over which the loan should be repaid.
(iv) that such an approach may make it easier to incorporate estimates of future changes in economic climate.

Such an analysis of when borrowers default may throw light on the received wisdom that 'if they go bad they go bad early'.

In this approach to credit scoring one tries to estimate how long until some event occurs even though in many cases the event will not occur. This has considerable similarities with survival analysis where one is estimating the lifetime for items such as vehicles, equipment and people, which can deteriorate and fail. The aim of this study is to show how some of the ideas of survival analysis may be applied in the credit scoring context and compare the results with more standard approaches. Narain[4] is one of the few authors who have previously investigated this approach to credit scoring. Narain applied one type of proportional hazard approach to loan data and showed that it gives a reasonable approximation to the time until default, but the paper did not make any comparison with alternative methods. In the next section we outline the survival analysis approaches which we will use. Following that we describe the application of three types of proportional hazard models and accelerated life models to loan data and compare their results with regression scorecard approaches. We then describe how the competing risks

*Correspondence: Prof LC Thomas, Management School, University of Edinburgh, Edinburgh, Scotland, UK.
E-mail: L.Thomas@ed.ac.uk

approach of survival analysis proves a useful way of dealing both with defaults and early completion of loans. The final section draws some conclusions from the analysis performed.

Survival analysis approaches

Let T be the length of time before a loan defaults. There are three standard ways of describing the randomness of T in survival analysis[5,6]

Distribution function, $F(t)$, where $F(t) = \text{Prob}\ \{T \leqslant t\}$

$$(S(t) = 1 - F(t) \text{ is the survivor function})$$

Density function, $f(t)$ where $\text{Prob}\ \{t \leqslant T \leqslant t + \delta t) = f(t)\delta t$

Hazard function $h(t)$, where $h(t) = f(t)/(1 - F(t))$ so $h(t)\delta t = \text{Prob}\{t \leqslant T \leqslant t + \delta t | T \geqslant t)$ These three formulations are mathematically equivalent in that one can obtain one from the others but they highlight different aspects of the loan lifetime. In all cases it is possible to have a positive probability mass at infinity corresponding to there being no default in a finite time.

The most common lifetime distributions used in survival analysis are the exponential and Weibull distribution. The exponential distribution with parameter λ has

$$\begin{aligned} &F(t) = 1 - e^{-\lambda t}, f(t) = \lambda e^{-\lambda t} \quad \text{and} \\ &h(t) = \lambda \text{ with mean } 1/\lambda. \end{aligned} \tag{1}$$

The Weibull family of distributions has two parameters: scale λ and shape k, where

$$\begin{aligned} &F(t) = 1 - e^{-(\lambda t)}, f(t) = k\lambda^k t^{k-1} e^{-(\lambda t)}, \\ &h(t) = k\lambda^k t^{k-1} \text{ with mean } \Gamma(1/k)/k\lambda \end{aligned} \tag{2}$$

where Γ is the Gamma function. Note that the Weibull distribution with shape 1 is an exponential distribution.

A common problem in estimating survivor and hazard functions is the number of right censored observations. These occur when an item has not failed after some time in use. Therefore for each item one records how long the item was in use and whether this ended with a failure or not. This translates in the credit environment to how long the consumer was a borrower and whether this ended with the consumer defaulting or not.

It may be that there are explanatory variables that influence the survival time of the unit under consideration. These explanatory variables allow one to cope with the heterogeneity of the population under consideration. These play the same role as the characteristics of the applicant in normal credit scoring approaches. There are two models that have found favour in connecting the explanatory variables to failure times in survival analysis—proportional hazard models and accelerated life models.[5,7] If $\mathbf{z} = (z_1, z_2, \ldots, z_N)$ is the vector of explanatory variables then the accelerated life models assume

$$S(t) = S_0(\psi(\mathbf{z})t) \quad \text{or} \quad h(t) = \psi(\mathbf{z})h_0(t\psi(\mathbf{z})) \tag{3}$$

where typically $\psi(\mathbf{z}) = \exp(b_1 z_1 + b_2 z_2 + \cdots + b_N z_N) = e^{\mathbf{b}\cdot\mathbf{z}}$ and S_0 and h_0 are the baseline survivor function and hazard rate function. So the $\mathbf{z}$ can speed up or slow down the 'ageing' of the life of the system.

The proportional hazard models assumes

$$h(t) = \psi(\mathbf{z})h_0(t) = e^{\mathbf{b}\cdot\mathbf{z}}h_0(t) \tag{4}$$

so that the explanatory variables have a multiplier effect on the base hazard rate. In both the proportional hazard model and the accelerated life model one can use a parametric approach to estimating the parameters by assuming the hazard function belongs to a particular distribution family, for example exponential or Weibull distributions. The difference between the models is that in proportional hazards the units most at risk at any one time remain the units most at risk at all other times whereas in the accelerated life models the rate at which the units 'age' is given by the explanatory variables, so that if failure risk varies non-monotonically with age then different items can be most at risk at different times. Kalbfleisch and Prentice[5] showed that the only lifetime distributions which can be both proportional hazards and accelerated life at the same time are the exponential and the Weibull families.

Cox[8] pointed out that in proportional hazards models one can estimate the parameters $\mathbf{b}$ without any knowledge of $h_0(t)$ by just using the rank of the failure times and the censored times. What is important for $\mathbf{b}$ is the ordering of the times rather than the times themselves. He used the idea of partial likelihoods and showed that if $\mathbf{z(i)}, t_i$, $i = 1, 2, \ldots, K$ are the explanatory variables and the failure or censored times of the K units under test then the conditional probability that item i fails at time t_i given that the items $R(i)$ are the items at risk, that is still operating, just before t_i is given by

$$\begin{aligned} &\exp(\mathbf{b}\cdot\mathbf{z(i)})h_0(t_i) / \sum_{k\in R(j)} \exp(\mathbf{b}\cdot\mathbf{z(k)})h_0(t_j) \\ &\quad = \exp(\mathbf{b}\cdot\mathbf{z(i)}) / \sum_{k\in R(j)} \exp(\mathbf{b}\cdot\mathbf{z(k)}) \end{aligned} \tag{5}$$

hence the likelihood function can then be obtained.

The only complication is if there are a number of ties with items failing and/or being censored at the same time. If there are d_i failures at time t_i let $\mathbf{s(i)} = \sum \mathbf{z(j)}$ summed over all the d_i failures at time t_i and let $R_d(i)$ be the set of all subsets of d_i items chosen from the items at risk just before t_i. Then the conditional probability that the actual set of failures at time t_i are the ones that actually failed given that $R(i)$ are the items at risk just before t_i is

$$\exp(\mathbf{b}\cdot\mathbf{s(i)}) / \sum_{k\in R(i)} \exp(\mathbf{b}\cdot\mathbf{s(k)}) \tag{6}$$

This is difficult to compute and approximations by Breslow[5] and Efron[9] are used. Ties are common if the time in the original problem is discrete or if the time data in a continuous problem is 'grouped' into months. There are two alternative definitions of proportional hazards models for discrete time where the baseline hazard function at times $t = 1, 2, \ldots$, is h_t. Kalbfleisch[5] suggested that it should be

$$h(t)\delta t = 1 - (1 - h_t\delta t)\exp(\mathbf{b}\cdot\mathbf{z}) \tag{7}$$

while Cox[7] suggests a linear log odds model with

$$h(t)\delta t/(1 - h(t)\delta t) = (h_t\delta t/(1 - h_t\delta t))\exp(\mathbf{b}\cdot\mathbf{z}) \tag{8}$$

(7) gives a model that agrees with that when the continuous data model has times grouped together while (8) leads to the likelihood function (6). Both lead to the continuous version (4) as δt tends to zero.

Another useful idea in survival analysis that could be useful in credit scoring is the concept of competing risks. This assumes that an item can fail because of different reasons. Suppose there are K different causes of failure labelled $i = 1, 2, \ldots, K$, and T_i is the net lifetime of the unit due to cause i. The net lifetime of the unit due to cause i is the lifetime of the unit if only that cause of failure is possible. The actual lifetime of the unit is T where

$$T = \min\{T_1, T_2, \ldots, T_K\} \tag{9}$$

However what one can measure is the crude lifetimes, Y_i, that is there is a failure at time t due to risk i. One can connect the two sets of variables by the equation

$$\text{Prob}\{Y_j \geqslant t\} = \text{Prob}\{T_j \geqslant t,\ T_j < T_i \text{ for all } i \neq j\} \tag{10}$$

To use this in practice one usually has to assume the T_i are independent which is often open to question. Competing risks can be useful in the credit scoring context if one is trying to identify why someone defaults on a loan as well as when is the default. Another application of competing risks is early repayment. If one is interested in the profit a company makes out of lending to consumers then early repayment is 'bad' in that it will diminish the profits made from that loan. Therefore one could also use early repayment and defaulting as competing risks which affect the profit on a loan.

Loan data and proportional hazards model

The survival analysis approaches to credit scoring were tried out on personal loan data from a major UK financial institution. The data consisted of application information of 50 000 loans accepted between June 1994 and March 1997 together with their monthly performance description for the period up to July 1997. The sample was split in two random groups where 70% was used to build the systems and the remaining 30% (15 018 cases) was used as a holdout sample.

The initial characteristics included information on age, marital status, employment, residency type, electoral role information as well as loan specific information such as the purpose of the loan and its term. The application variables or characteristics were split into attributes by combining answers. In some cases like if you had a phone at home, there were only two attributes so it was not necessary to combine them. In other categorical variables like purpose of loan where there were 25 categories, these were reduced to four by putting together ones with similar purposes where the default rate was not dissimilar. Therefore all the vehicle purchases whether for new or old cars, motor cycles or other vehicles were put together. For the continuous variables like age the characteristics were split first at a very fine level of attribute, age in years, and then consecutive ones of these grouped together to form larger categories if these consecutive ages had similar default rates. Therefore age became under 21, 22–24, 25–32, 33–42, 43–51 and over 51 years. This is a standard procedure in credit scoring because even for continuous variables like age and income, the default risk is not monotone. By considering such a variable as a number of binary variables each with their own coefficient (weight) in the scoring system one can allow for this lack of monotonicity.

To measure the outcome of the loan, the monthly performance indicators were used. The number of months until the loan either defaulted, paid off early or paid off on time was recorded while for about a third of the loans their times were censored since they had not reached any of these states by July 1997. Therefore for each loan one had a survival time, whether it was censored or not. In building the models for time until default all the cases of early or normal payoff as well as those which were still active in July 1997 are considered as censored data.

As was outlined in section two Cox's approach to proportional hazard models allows one to get estimates of the **b** coefficients without making any assumptions about the base hazard rate simply by using the ranks of the times. We denote this the Cox non-parametric approach (Cox).

One could also assume the base-hazard rate $h_0(t)$ is of a specific distribution and try to estimate **b** and the parameters of that distribution. The simplest assumption to make is that the baseline hazard function is exponential so that $h_0(t) = \lambda$ and this leads to proportional estimators with exponential hazard rate. In that case (3) and (4) become the same so that this is also an example of an accelerated life model with exponential baseline. We will denote it as (exp). The discrete time equivalent this would lead to the assumption that the lifetime is geometric.

One could generalise these models by allowing a larger family of possible baseline distributions, namely the Weibull distributions. In that case the baseline hazard function is

$$h_0(t) = k\lambda^k t^{k-1} \tag{11}$$

With this distribution again the hazard functions of (3) and (4) are the same and the respective **b**-vectors agree up to a constant of proportionality. So again this model is both a proportional hazard and an accelerated life one and is denoted (Weib).

In order to compare these approaches with standard credit scoring approaches, the data is also used to develop logistic regression based scorecards (LR). The exp and Weib approaches give probability functions of the avoidance of default for all durations of the loan, while Cox's proportional hazard approach gives an ordering of relative likelihood to default for each loan. In all three cases the ordering of the likelihood of default of the loans stays the same at all times, this is the proportional hazard assumption. So the same group are considered most at risk for all ages of the loans. This is not necessarily the case for the standard logistic regression approach. Thus we take two measures of how these survival approaches compare with the logistic regression approach, which concentrate on default risk in different epochs, namely

(i) how likely are the loans to default in their first 12 months
(ii) how likely are loans that survive 12 months to default in the subsequent 12 months

Therefore two separate logistic regression scorecards (LR) are built for each of these approaches. For (i) 'bads' are failures in the first 12 months, 'goods' are all others; for (ii) only loans that are still being repaid at 12 months are considered and 'bads' are ones that default before 24 months.

This is giving the logistic regression approach quite an advantage, in that two separate logistic regression-based scorecards are built and each one is tailored to be the best classifier for the criterion under which it is used. Against this the same proportional hazard model is measured under the two criteria in turn. Moreover because they are proportional hazards models with constant coefficients it is the same applicants who will be considered most likely to default whatever time period is taken. There are generalisations of proportional hazards like allowing the coefficients to be time dependent or taking accelerated life models with other distributions which would overcome this, but the aim here is to present the simplest type of survival analysis models. Note that though the same group is most at risk at all ages of loan in the proportional hazards approaches, it is not the case that the same applicants will be determined most at risk under the two criteria. In criteria (ii) those who have failed in the first 12 months have already been removed and some of the most likely to fail will thus not be in the set at risk.

The estimators and logistic regression scorecards are built on the training sample and the resulting functions applied to the holdout sample. In each case the cut-off is chosen so that the predicted number of 'bads' equals the actual number of 'bads' in the holdout sample under each of these criteria, which removes any effect of the cut-off. The following Tables 1–4 then show the scorecards predicted classification against the actual classification into the 'goods' (G) and 'bads' (B) for the 15 018 in the holdout sample. The tables also indicate which attribute had the greatest effect in each scorecard. Suppose a variable is split into M categories and the scorecard came up with a value b_i for category i and p_i was the proportion of the sample population in that category for that variable. The average value for that variable is then $\beta = \sum_i b_i p_i$. We identify which attribute has the largest effect by looking for the b_i where the difference between the b_i and the mean value β is the greatest. The results are given in the following tables for the two different time periods. The closeness of the results of the different approaches may be an example of the flat maximum effect[3] in which there are many score-cards with different coefficients whose discrimination is close to the optimal.

Competing risks approach

The competing risks approach[5,6] to loans assume that there are several reasons why repayment of the loans finish before the original intended term. We will concentrate on just two reasons; either the repayer defaults or he pays off early. If one had the data one could separate each of these in turn by looking at the reason for default or the reason for early payoff. One then can build survivor function models to estimate the distribution of

T_d = the lifetime of the loan until default
and
T_e = the lifetime of the loan until early repayment,

Table 1 Predicting default in first 12 months

	Actual numbers	*LR*	*Cox*	*Exp*	*Weib*
G-predicted G	14620	14268	14269	14266	14267
G-predicted B	0	352	351	354	353
B-predicted G	0	352	351	354	353
B-predicted B	398	46	47	44	45
Attr. largest effect		amount 8 K+(−)	purpose: refinance (−)	purpose: refinance (−)	purpose: refinance (−)
Attr. 2nd largest		purpose: refinance (−)	employ 17 + yrs (+)	employ 17 + yrs (+)	employ 17 + yrs (+)
Attr. 3rd largest		insurance £500−700 (−)	amount 8 K + (−)	amount 8 K + (−)	address 16 + (+)

Table 2 Predicting default in 12–24 months

	Actual numbers	*LR*	*Cox*	*Exp*	*Weib*
G-predicted G	8021	7850	7846	7844	7848
G-predicted B	0	171	175	177	173
B-predicted G	0	171	175	177	173
B-predicted B	191	20	16	14	18

where if T_m is the term of the loan, the number of months of repayments is

$$T = \min\{T_d,\ T_e,\ T_m\} \tag{12}$$

Models for estimating T_d were given in the previous section. Exactly the same analysis can be applied to T_e. The only difference is that in the first case loans that defaulted were considered failures and the repayment times of all other loans were considered as censored times. For T_e the loans that are paid off early are the ones that are considered as 'failures' while the repayment times of all the others are considered as censored times. The Cox, Exp and Weib methods are applied to this early repayment data. The results are presented using a comparison with a logistic regression approach under the two criteria

(i) estimating which loans will be paid off early within the first 12 months
(ii) estimating which loans which are still repaying after 12 months will pay-off early within the next 12 months

Conclusions

The analysis in the previous sections indicates that one can apply the ideas of survival analysis to credit scoring and come up with meaningful results. The results suggest that the proportional hazard models investigated here are competitive with the logistic regression approach in identifying those who default in the first year, and may be superior to that approach for looking at who will pay-off early in the first year. The proportional hazard results for the second year, where there are fewer defaulters are not so encouraging and suggest that more sophisticated models might be appropriate. The superior performance on early payment compared with default may be because the sample used has already been credit scored. Therefore there are very few bad cases under the default criterion compared with the early repayment criterion, and the survival analysis approach benefits more from a large sample of 'bads' than does the logistic regression approach.

The poor performance under the second year criterion must be partly due to the fact that the ordering of risk of default (or early repayment) does not change whatever the time period. Two obvious extensions are used in survival analysis to overcome the fixed for all time risk ordering of proportional hazards. The first is to allow the **b**-coefficients to be time dependent and the second is to use accelerated life models with other distribution families than the Weibull one. Both of these extensions would allow the risk ordering to vary over time.

Another extension of proportional hazards and accelerated life models is to allow the baseline function to be different for different groups within the population. This is a half-way house to building different scorecards for different groups in the population since the **b**-coefficients are kept the same for all groups. One can also include Bayesian analyses of the proportional hazards models or look at situations where there are two times associated with each failure each modelled by a separate survival function. In the credit scoring context this could be first time the repayments fall behind schedule as well as the time of default. The developments of multi-stage models in survival analysis to deal with these problems could also prove useful in the credit scoring context.

Table 4 Predicting early repayment 12–24 months given still repaying at 12 months

	Actual numbers	*LR*	*Cox*	*Exp*	*Weib*
G-predicted G	6375	5094	4955	4897	4953
G-predicted B	0	1281	1420	1478	1283
B-predicted G	0	1281	1420	1478	1283
B-predicted B	1837	556	417	359	415

References

1 Hand DJ and Henley WE (1997). Statistical classification methods in consumer credit. *J Roy* Stat Soc, Series A **160**: 523–541.

Table 3 Predicting early repayment in first 12 months

	Actual numbers	*LR*	*Cox*	*Exp*	*Weib*
G-predicted G	12080	9884	9907	9910	9905
G-predicted B	0	2196	2173	2170	2175
B-predicted G	0	2196	2173	2170	2175
B-predicted B	2938	742	765	768	763
Attr. largest effect		amount 8 K + (−)	term 8 + yrs (−)	term 8 + yrs (−)	term 8 + yrs (−)
Attr. 2nd largest		term 4–8 yrs (−)	term 4–8 yrs (−)	term 4–8 yrs (−)	term 4–8 yrs (−)
Attr. 3rd largest		amount 4–8 K (−)	term 1.5 yrs < (+)	purpose; home furnishing (+)	term 1.5 yrs < (+)

2 Rosenberg E and Gleit A (1994). Quantitative methods in credit management: a survey, *Opns Res* **42**: 589–613.
3 Thomas LC (1998). Methodologies for classifying applicants for credit. In: Hand D and Jacka S (eds). *Statistics in Finance*. Edward Arnold: London, pp 83–103.
4 Narain B (1992). Survival analysis and the credit granting decision. In: Thomas LC, Crook JN and Edelman (eds). *Credit Scoring and Credit Control*. OUP: Oxford, UK, pp 109–121.
5 Kalbfleisch JD and Prentice RL (1980). *The Statistical Analysis of Failure Time Data*. Wiley: New York.
6 Leemis LM (1995). *Reliability*. Prentice-Hall: Englewood Cliffs.
7 Cox DR (1972). Regression models and life-tables. *J Roy Stat Soc*, Series B, **34**: 187–202.
8 Cox DR (1975). Partial likelihood. *Biometrika*, **62**: 269–276.
9 Efron B (1977). The efficiency of Cox's likelihood function for censored data. *J Am Stat Ass* **72**: 557–565.

Received January 1998;
accepted September 1998 after two revisions