Application of survival analysis to cash flow modelling for mortgage products.

Abstract

In this article we describe the construction and implementation of a pricing model for a leading UK mortgage lender. The crisis in mortgage lending has highlighted the importance of incorporating default risk into such pricing decisions b y mortgage lenders. In this case the underlying default model is based on survival analysis, which allows the estimation of month-to-month default probabilities at a customer level. The Cox proportional hazards estimation approach adopted is able to incorporate both endogenous variables (customer specific attributes) and time-covariates relating to the macro-economy. This allows the lender to construct a hypothetical mortgage portfolio, specify one or more economic scenarios, and forecast discounted monthly cashflow for the lifetime of the loans. Monte Carlo simulation is used to compute different realisations of default and attrition rates for the portfolio over a future time horizon and thereby estimate a distribution of likely profit. This differs from a traditional scorecard approach in that it is possible to forecast default rates continually over a time period rather than within a fixed horizon, which allows the simulation of cashflow, and differs from the company's existing pricing model in incorporating the possibilities of both default and early closure.

Keywords

Survival analysis, Cox Proportional Hazards, default risk

Introduction

The mortgage crisis that has shaken the financial stability of many developed countries in 2007 and 2008 has highlighted how important it is to accurately asses the

risk in mortgage lending, in order to price these risks correctly. There are two critical issues which have to be addressed in such pricing models and which it can be argued were partly the cause of the sub prime mortgage crisis. The first is the impact that changes in the economy, particularly in house prices, have on the default and attrition risks involved in mortgage lending. The second is that these risks vary over the duration of the loan, and so one needs to develop a dynamic model which reflects the particular structure of the loan and reflects the economic changes that may occur while it is being paid back.

This case study describes a pricing model that was built for a leading UK mortgage lender. It combines survival analysis (Allison 1999) and Monte Carlo simulation, and allows the lender to experiment with different portfolios, pricing structures and economic scenarios. The outputs is a monthly cashflow forecast which incorporates the possibility that loans will terminate before running their full term either because the borrowers default or because they choose to repay or refinance (early closure). The frequency of these events and their likely impact varies by customer quality, loan type, and changes in economic conditions over the lifetime of the loan. The choice of explanatory variables and modelling assumptions attempts to account for as many of these influences as possible. At the same time, limitations in the amount of available data and the lack of significant shocks to the UK economy in the time period (2000-2006) over which the data was collected meant there is significant scope for the model to be updated and refined over time. The model was built in such a way that this will be easy to do.

The UK mortgage lending market has traditionally had a very high proportion of two stage mortgages (akin to the US 2/28 and 3/27 mortgages) which have an initial period of two, three or five years at a fixed rate or a rate tied to a Central Bank set rate (a tracker mortgage) and which then move to variable-rates thereafter. This traditionally led to a rapid turnover among customers particularly after the initial stage during which there are high penalties for changing to another mortgage. Mortgage lenders aim to price loans strategically, taking into account a number of factors including market position, customer retention and profitability, liquidity risk, competition, shareholder value and the likely performance of the economy. The most critical aspect of the price is the interest charged both in the initial stage and in subsequent stages, but arrangement and early redemption fees also can be considered part of the pricing package. At the time of writing (late 2008), a significant slowdown in the interbank lending markets and a simultaneous desire among banks and other mortgage lenders to shore up their capital reserves has led to a sharp decline in mortgage lending, which may in turn bring into question assumptions regarding the relationship between base rates and actual lending rates. For the purposes of modelling, however, it is convenient to assume that a lender charges interest at the base rate plus a margin intended to cover the 'risk' of the investment, which still seems to be the case even though this margin is now considerably increased. Future cashflows from the loan can be discounted at the Bank of England rate, which is considered the risk-free rate.

The model developed incorporates time covariates and monthly probabilities of default, and so differs markedly from the typical default models that are developed for application scorecards. These generally assume that the default behaviour of future

customers will be broadly similar to that of past applicants, regardless of the broader economic climate. A model (nearly always logistic regression) is fitted to the application characteristics of past customers. For each new customer, this model outputs a probability of defaulting within a fixed time horizon (say, six or twelve months), and the lender can impose a threshold on this default risk above which he or she is unwilling to lend. Such a model cannot, however, be used to estimate the value of a loan, since profit or loss on a mortgage loan is strongly dependent on the exact time that the default event occurs, the capital outstanding, and the interest that has been paid up to that point.

Previously the lender had used a traditional default scorecard (Thomas et al 2002) to assess the default risk of its borrowers and an economic pricing model that was able to simulate returns for mortgage product under given interest scenarios. However the latter model was used only to assess interest rate risk in pricing new products; it took no account of default events or their consequent losses and how these were affected by the economic climate. The new model allowed these two aspects of a loan . together with its other features to be combined and to give the results at a portfolio level as well as at an individual loan level.

Modelling Approach

The central feature of the modelling approach in this project is a default model based on survival analysis (Thomas et al 1999, Stepanova and Thomas 2001, 2002). Survival analysis has its origins in medical and actuarial sciences, where it is the standard model for predicting the lifetime of individuals contingent on particular risk factors. These factors may be endogenous, individual specific variables (e.g. smoker /

non-smoker), or exogenous time-covariates affecting all individuals under consideration (e.g. economic or social trends).

In the model presented here one needs to predict the time to default of mortgage borrowers in our portfolio. The endogenous variables are the application characteristics of the consumer, and the exogenous variables are macro-economic time series including the base interest rate (see Tang et al (2007) for a similar approach to product purchase) . One of the advantages of survival analysis is the ability to incorporate 'censored' data, or individuals for which the default event has not yet been observed. This means that all of the available customer data can be used to build a model, even where loans were still active at the time of the most recent observation. This concept is illustrated in **Figure 1** below.

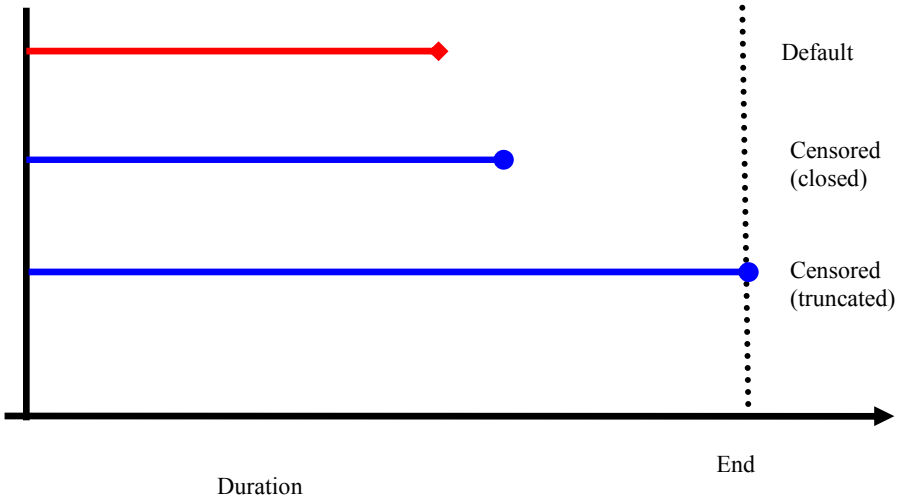

Figure 1: The data on all loans can be used though some may be censored as default does not occur

The lender offered a range of mortgages, including products tailored to first time buyers and buy to let investors. As the application characteristics and default

behaviour of customers in different groups were known to differ widely, separate default models were built for each product type.

Two important concepts in survival analysis are the *survivor function* and the *hazard rate*. The survivor function is a continuous function representing the probability that the 'failure time' *T* of an individual is greater than time *t*.

$$S(t) = \Pr(T > t) \qquad (1)$$

The hazard function *h(t)* represents the point in time default 'intensity' at time *t* conditional upon survival up to time *t*.

$$h(t) = \lim_{\delta t \to 0} \left(\frac{\Pr(t \le T \le t + \delta t \mid T > t)}{\delta t} \right) \qquad (2)$$

The survivor function and hazard rate are linked via the cumulative hazard rate Λ(t), defined as

$$\Lambda(t) = \int_0^t h(u)du = -\log S(t) \qquad (3)$$

Since most mortgage lenders record repayment and default data on a monthly basis, the model built was a discrete time one. The survivor function is the chance the borrower will have not defaulted in the first t months of the mortgage while the hazard function may be thought of as the probability that a given borrower, having 'survived' to month *t*, will default in the next month.. One can therefore produce comparative global survivor function and hazard rate estimates for different mortgage products simply by plotting the month by month survival rates and default rates. The model uses the standard definition of default as being three months in arrears with repayments (note that a default event does not therefore necessarily correspond to repossession). Examples of these curves, for the first 32 months of the loan term, are

shown in **Figure 2** below. Note that the names of the product types and the vertical scale are not shown for commercial reasons. However, it is evident that some specialised product types are considerably more 'risky' than others, and that some hazard rates appear to be increasing as a loan advances. These plots proved very informative to the company, and were in agreement with their intuition regarding the riskiness of particular products.

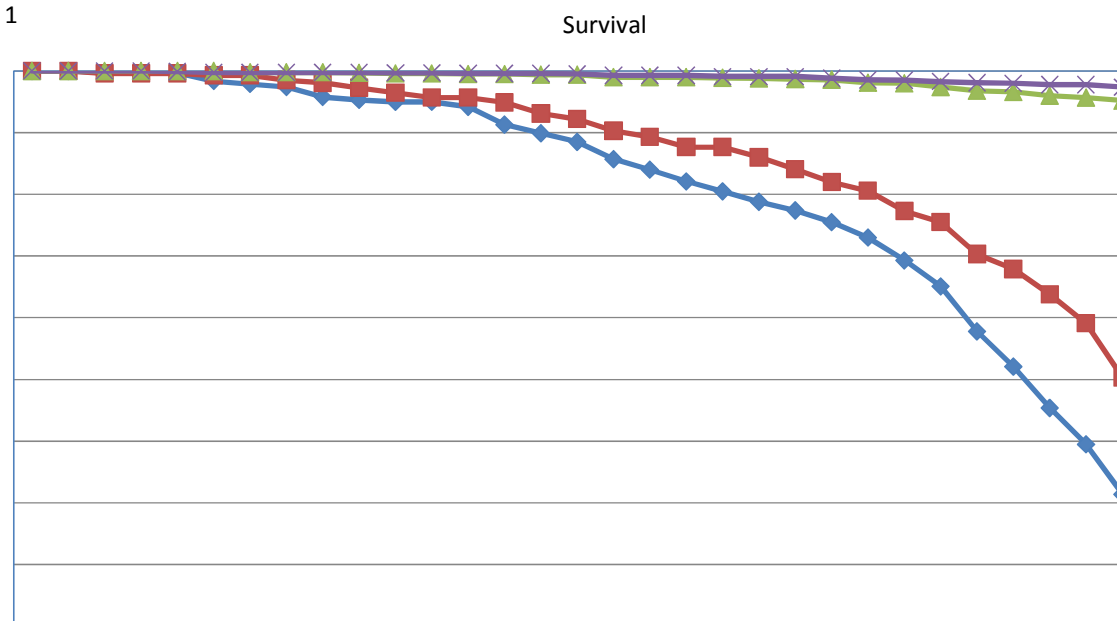

Figure 2: The chance that lenders with different mortgage products have not yet defaulted as a function of how long they have had the loan

Cox Proportional Hazards (Cox 1972, Therneau 2000) approach to survival analysis allows the building of default models for specific combinations of individual characteristics. This approach assumes that there exists a 'baseline hazard' $h_0(t)$ which is common to every individual. This baseline hazard is multiplied by a term which depends on both the application characteristics **x** of the applicant and on time covariates **y** (t)**,** which are the time dependent economic conditions. So the hazard rate *t* months into the loan is modelled as:

$$h(t; x, y(t)) = h_o(t)e^{\boldsymbol{\beta}_1^T \cdot \mathbf{x} + \boldsymbol{\beta}_2^T \cdot \mathbf{y}(t)} \qquad (4)$$

where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are vectors of coefficients.

Our application variables consisted of a number of application characteristics including the application score under the company's existing default scorecard (which may itself be viewed as a summary of application characteristics). These were categorised in such a way that the vector **x** was a list of binary indicators according to which categories a customer fell under. **y(**t) were the values of the macro-economic variables t months into the loan. These were obtained from publicly available sources and included the log of the Bank of England base rate and an index of house prices . **Figure 3** below shows a plot of two macro-economic factors, the base rate and the Halifax Seasonally adjusted house price index over a period from January 2000 to September 2007.

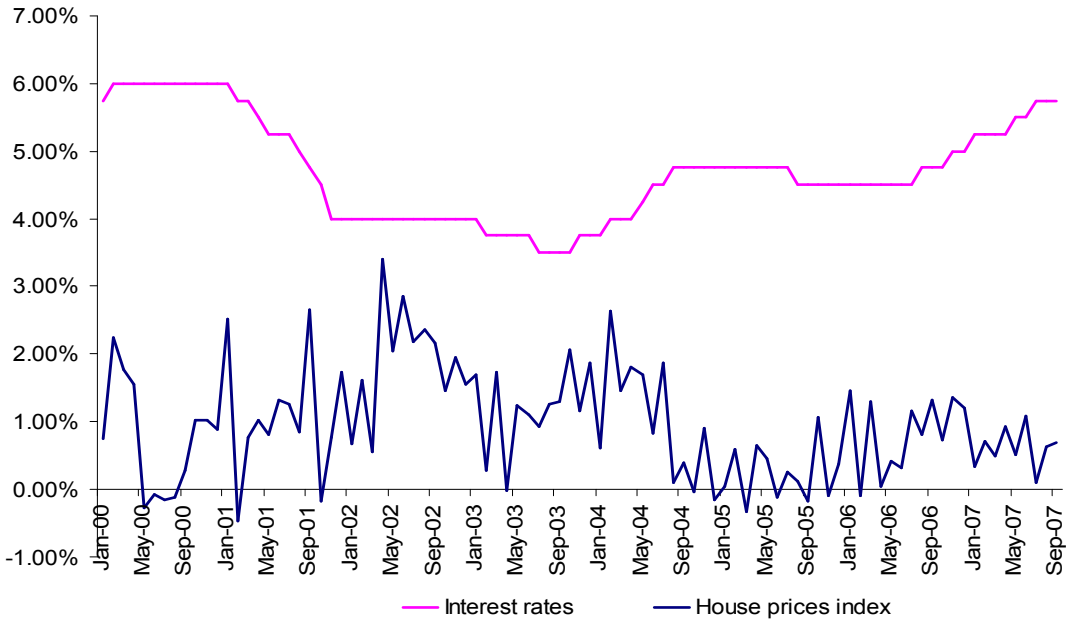

Figure 3: The values of the two economic variables used in the model for the period 2000 to 2006 on which the model was built

Estimation of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ was performed in SAS using the *phreg* proc for Cox regression (with the Efron method used for breaking ties). Given these estimates, the baselines $h_0(t)$ were computed via the Nelson-Aalen (Anderson et al 1993) formula:

$$h_0(t) = \frac{d_t}{\sum_{l \in R_t} \exp\left[\boldsymbol{\beta}_1^T \cdot \mathbf{x} + \boldsymbol{\beta}_2^T \cdot \mathbf{y}(t)\right]} \qquad (5)$$

Here d_t is the number of defaulters for the current product in month t and R_t is the 'risk set' at time t; ie. all accounts that were open at the beginning of month t. Estimates of the unsmoothed baselines computed for individual products are shown in **Figure 4** below (note that the vertical scales are not directly comparable as the baseline has no fixed scale).

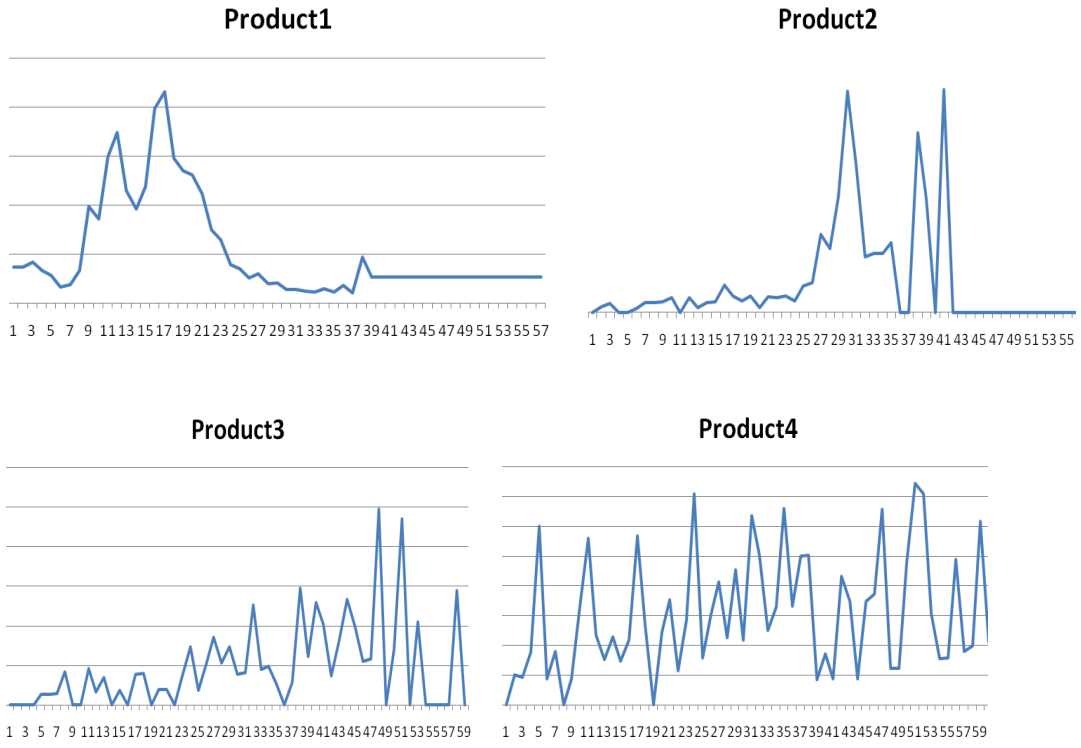

Figure 4: The baseline hazard rates for four different products shows how the risk of default varies over the duration of the loan

The baseline shapes represent the characteristic risk profile for each product independent of the covariates. Most products show a distinctive 'spike' in the hazard

curves after a certain time period. This reflects the fact that customers are most likely to default at the end of their introductory fixed rate period, when they transfer onto a less favourable rate of interest. Because the data available did not cover the full period of a mortgage loan (up to forty years), it was necessary to smooth and extend these baselines. In doing so, it was assumed that the inherent risk would diminish over time (in keeping with the received wisdom that most loans which fail because of fraud or unaffordability do so near the beginning of the term). Standard smoothing procedures were used. As all the estimates of the model coefficients and the baselines can be updated periodically by the company, the model fit should improve over time and the validity of our smoothing assumptions can be tested.

The combination of model coefficients and baseline estimates allowed calculation, via equation (4), of the monthly estimates of the default rate of any consumer for a given combination of application characteristics and given trajectories of the macro-economic variables.

Treatment of Early Closures

One of the most important characteristics of a mortgage loan portfolio is the high frequency (at least in economic conditions that favour a competitive marketplace) with which borrowers repay or refinance loans. Repayments tend to be low during the discounted or fixed rate period due to the penalties incurred, very high at the point where this period ends (up to 70% in some portfolios), and relatively low from this point onward. For simplicity, this model assumed there are three repayment rates – one during the time when the fixed or discounted rate is in operation, a one-off repayment probability at the end of the fixed or discounted rate period and a

repayment rate for the remainder of the loan. Currently these are subjective estimates input by the lender. An obvious means of extending the model, given sufficient data, would be to build a competing-risks type (Stepanova and Thomas 2002) model for the probabilities of both early closure and default, which might also capture early repayment behaviour under changing economic circumstances.

Stucture of the Model

The application delivered to the company was coded in VBA for Excel. The structure of the full model is illustrated in **Figure 5** below.

The inputs of the model fall into three broad categories. The loan parameters are generic inputs common to all loans in the hypothetical portfolio to be constructed. They include factors such as the average loan size, the term, the repayment pattern (eg. amortisation, interest only etc.), the probability of repossession given a default event and the haircut given repossession (ie. the proportion of a property's nominal value that is not recovered due to a forced sale), fees and early closure penalties. Some of these parameters can be given different values under different economic scenarios. The user is also able to specify the margin charged over the base rate, which impacts on the profitability of the loan. It is assumed that if there is a fixed rate introductory period of a loan, funds are hedged via financial instruments in such a way that the equivalent variable rate is recovered. (This is the way the lender hedges his loan book in practice).

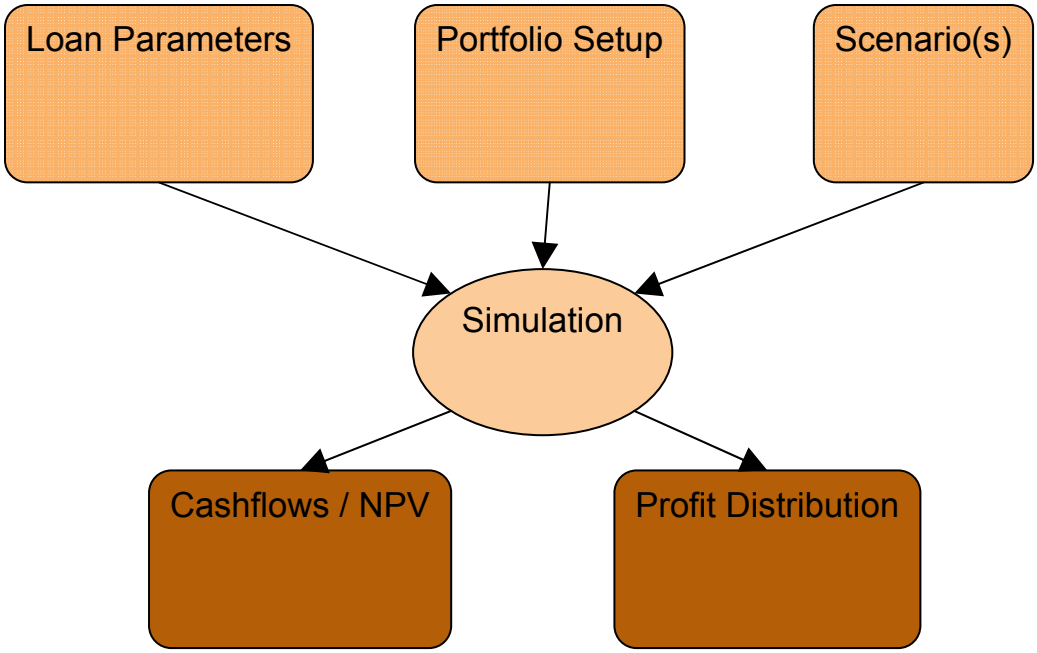

Figure 5: The model was structured with Excel front and back end so that the three types of data could be introduced via the sections on loan parameters, portfolio set–up and scenarios, while the outputs were given as profit distributions and cash flows forecasts.

The portfolio setup parameters allow the user to construct a hypothetical loan portfolio comprising one or more of the different product types. The user is able to specify the frequencies with which different combinations of application characteristics occur for each product type.

The user may input a number of different future economic scenarios, along with the probability of their occurring. An economic scenario consists of a set of monthly time series for the macro-economic variables. The most important of these is the base rate, which serves as the discounting rate for future cashflows as well as determining the interest repayments for the life of a loan. The user also inputs monthly changes in

house price index, which are used to track the value of a property in the event of repossession.

Once these parameters have been entered, the model calculates cashflow patterns for all possible combinations of application characteristics in the portfolio, and for each scenario. These include the monthly capital and interest payments, the early repayment penalties should the loan be repaid in any given month, and the recovery amount should repossession occur in any given month (up to a maximum of the amount outstanding at the time of sale). The model also computes the monthly discount factors for all future cashflows based on the base interest rate.

The model also allows for the possibility that defaults can be "cured" so the property is not necessarily repossessed. In that case, the mortgage company always assumes that the future repayments will eventually equate to those if no default had occurred, and this is the assumption used in this model.

Model Output

Once the model is run to determine the parameters for the default hazard functions, and the parameters given by the user describing the other aspects of the loan portfolio, the application performs Monte-Carlo simulation (Ross 2006), to generate a distribution of cashflow forecasts. The simulation is run for a set number of iterations which the user may vary according to the size of the hypothetical portfolio and the computational resources available. At each iteration an economic scenario is chosen at random from the scenarios determined by the user. Note the different economic scenarios to be considered and the likelihood with which they are chose is set by the

user. The algorithm then cycles through each customer in the portfolio, and through each month in the loan term, randomly determining whether a default or an early closure occurs, and recording the total repayments and losses in that period.

Having run the model for *N* iterations over the possibly different economic scenarios one is left with *N* potential cashflow forecasts for the hypothetical loan portfolio. From these cashflows, one can calculate the distribution of net present values for the portfolio (using the risk free discount rate), including the expected net value and the maximum and minimum values. One can also use the cashflows to estimate the distribution of internal rates of return and an effective interest rate. The outputs of the simulation are also expressed graphically. For example **Figures 6a and 6b** are one output which show the cashflows for the best, worst and median net present values for a hypothetical loan portfolio and scenario, their cumulative values, and the capital outlay after 100 simulation runs. Note that the spike at month 25 occurs because of the large number of repayments occurring at the end of the fixed rate period. **Figure 7,** is another graphical output , which shows the distribution of profit, at the current value of money, across all runs.

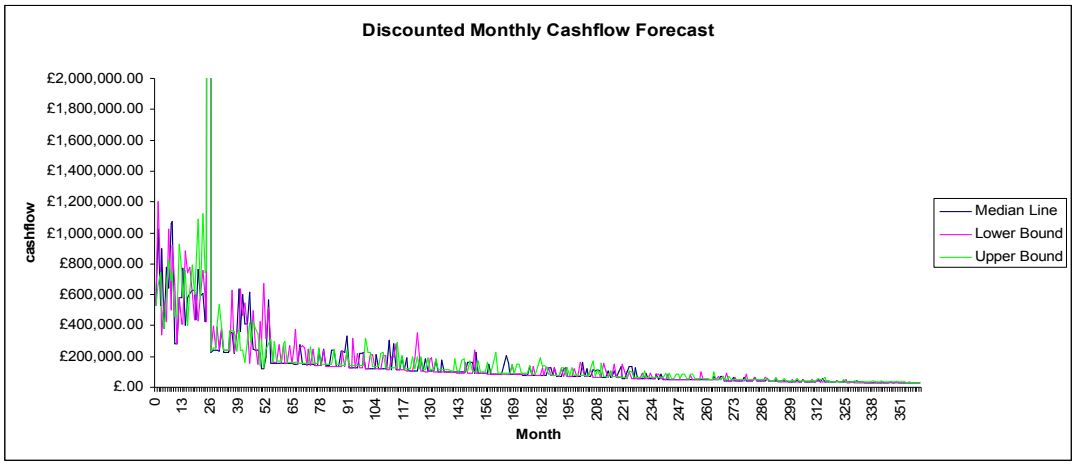

Figure 6a: The median, maximum and minimum cash flow from the N potential cash flows for the whole portfolio obtained from the Monte Carlo simulation.

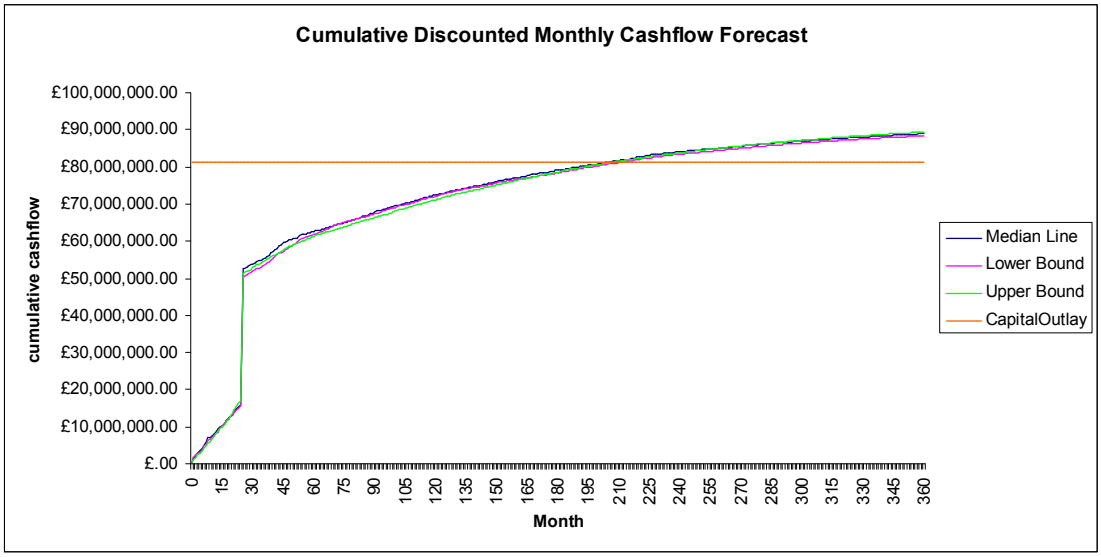

Figure 6b: The equivalent cumulative cash flows, using the data from Figure 6a, compared with the original capital outlay

.

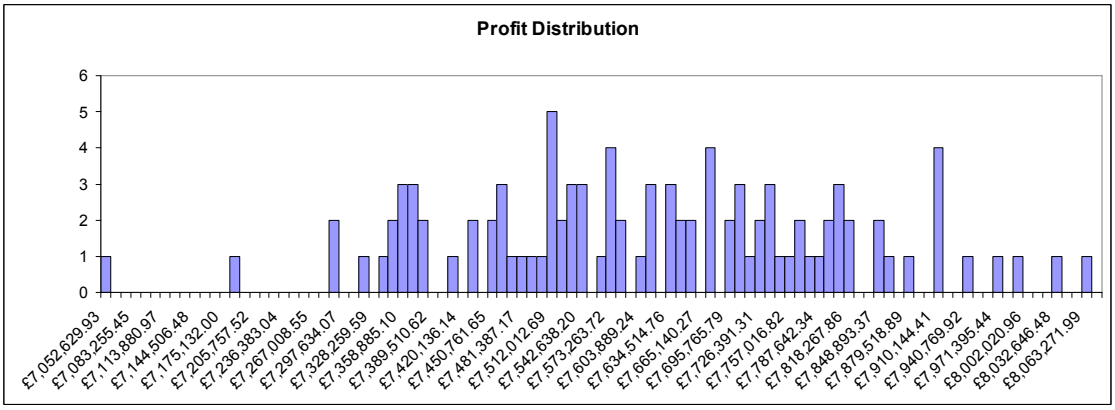

Figure 7: The histogram of profits obtained from the portfolio of loans using the N iterations of the Monte Carlo simulation

Model Usage

The main use of the model by the lender is to "price" new portfolios of loans. This lender traditionally obtains its loan capital from its depositors and standard money markets, rather than from the currently illiquid securitization method. Recently the

UK, along with other governments made extra capital available through government backed bonds to stimulate the mortgage market. When a tranche of money is made available for lending, the organisation must determine whom to target the loan and what price to charge for the loans. Most loans have a two stage structure, so the "price" involves setting the interest rate in both the initial "discount" stage and the second stage. It is also involves setting the early redemption penalties during the different stage and any up front arrangement fees. The targeting of the loans also involves determining which loan type – prime, subprime, buy to let, self-certification (Alt-A)-, what Loan to Value limits and what borrower default risks, through application score cut-offs, are acceptable.

These decisions involve input from both the marketing and risk groups within the organisation. These were often in conflict as the marketing group wanted attractive prices and low application score cut-offs to ensure the loans were taken up by borrowers while the risk group was concerned about good margins to cover any potential losses and higher application score cut-offs. Since there was no model that incorporated both risk and profitability it was difficult to resolve this argument. The model described above is now used to look at the profit distribution under different possible price structures and so allows both groups to understand the likely long term consequences of these different pricing structures. As well as changing the pricing structure , the model allows the lender to change the distribution of the borrower types who are likely to apply for and be accepted for such loans.

As well as being used for new lending, the model is also used to give cash flow forecasts for existing loan portfolios. This is proving useful both for debt

provisioning and also for Treasury functions. The cash flow projections also suggest the level at which new loans can be offered without having to seek new funding .

Benefits and Refinements

There are a number of ways in which this model might be refined and improved in future. At present, no statistical model has been built for early repayment events, with the user simply inputting the rates at which early repayment is expected to occur. There has been no attempt to model correlations between default, early repayment and the macro-economy. These effects are likely to be significant since borrowers are more likely to refinance their loans when conditions favour a competitive market, and an inability to refinance may lead to a default. However, modelling such correlations would likely to require more data than is currently available.

Another issue encountered in building the survival model was the long unbroken period of relative economic stability in the UK over the historic timeframe in which the model was fitted. As mentioned above though , this has changed dramatically in 2008-9 . Prior to this, the last major economic shock in the UK that seriously impacted on the mortgage sector occurred in the early 1990s under an arguably very different market structure. The practical implication of this is that the model performance is likely to weaken under economic scenarios that are very different from those in the training period. For this reason, the model was deliberately designed so that all the coefficients of the survival analysis can be updated on an ongoing basis as new data is obtained.

At present the model makes broad assumptions regarding the probability of repossession given default, haircuts (the drop in expected sale price) if a repossessed house is sold, and the time between repossession and sale (which is likely to be longer in an adverse climate). One particular issue in modelling such factors is that they are heavily influenced by actions taken by the mortgage lender itself and their procedures for handling bad debt. Many lenders will offer to refinance a loan in some circumstances in preference to undertaking repossession, though again this will depend on economic conditions. These decisions need to be reflected in the model estimates for the losses if there is a default.

The cashflow model is only one of a set of tools that the mortgage lender uses to decide how to price a loan. Given suitable input, it can incorporate changing macro-economic conditions and produce monthly cashflow forecasts. This is very useful as it can use the same scenarios that the lender is using for other portfolios. Unlike the existing pricing model, the model described here includes default events and calculations relating to repossession and recovery rate (though these are based on broad assumptions). This approach provides useful information to the lender and has helped in making decisions regarding the pricing of mortgage products. Some of the assumptions in the model will be replaced over time as more data becomes available, but in general we believe that the survival analysis approach conveys many benefits over comparable methodologies. Given the changes that have been occurring in the housing market a tool that allows for a number of different future economic scenarios and does model the impact of these economic changes on the profitability of the mortgage portfolio has already proved very useful.

References

Allison, P. D. 1995. Survival Analysis Using the SAS System: A Practical Guide. Cary NC: SAS Institute.

Andersen P.K., Borgan O, Gill R.D., Keiding N., (1993), Statistical models based on counting processes, Springer Verlag, New York.

Cox D R (1972). Regression models and life-tables (with discussion). *J Royal Statist Society, Series B* **74:** 187-220.

Ross S.M., (2006), Simulation, Elsevier Academic Press, Burlington, MA.

Stepanova M. and Thomas, L.C., (2001). PHAB scores: Proportional hazards analysis behavioural scores. J. Operational Research Society, 52, 1007-16.

Stepanova M., Thomas, L.C (2002). Survival analysis methods for personal loan data. Operations Research, 50, 277-289.

Tang L, Thomas LC, Thomas S, Bozzetto J-F (2007). It's the economy stupid: modelling financial product purchases. *International Journal of Bank Marketing*. Vol.25, issue 1, pp.22-38.

Therneau T. M., Grambsch P. M., (2000): Modeling survival data: extending the Cox model; Springer.

Thomas LC, Banasik J, Crook JN (1999) Not if but when will borrowers default, Journal of Operational Research Society 50: 1185-1190;

Thomas, L.C., Crook, J.N. and D.B. Edelman (2002). Credit Scoring and its Applications. Philadelphia: SIAM,