Credit Scoring Modelling for Retail Banking Sector.

Elena Bartolozzi, Matthew Cornford,
Leticia García-Ergüín, Cristina Pascual Deocón,
Oscar Iván Vasquez & Fransico Javier Plaza.

II Modelling Week, Universidad Complutense de Madrid,
16th - 24th June 2008.

Contents

Acknowledgements

We would like to thank Ignacio Villanueva for his guidance in this project and Estela Luna of Accenture for raising the problems.

1 Credit Scoring

1.1 Introduction

Our problem is concerned with how and who a bank should loan its money to. When a client applies for a loan the bank would like to be sure that the client will pay back the full amount of the loan. In the past the decision was made solely on the bank's experience in lending money. This method is highly subjective and the banks would like a more systematic approach.

One way of doing this, which we present here, is to fit a Generalized Linear Model to past data and use this to produce a probability that the borrower will repay the loan. This probability, along with the lenders experience is then used to decide if the bank should lend to a particular client. This method can also be used for the problem of issuing general insurance, for example car insurance.

This problem is a routine one that is already in use by many banks and insurance firms. Many tools exist in `MATLAB` and `SAS` which make the implementation of each step trivial.

1.2 Methodology and Data

There are five clear steps which we took to complete this problem. These are

- Univariate Analysis,
- Multivariate Analysis,
- Model Creation,
- Model Validation,
- Model Calibration.

Our data was provided by Accenture and include details of completed loan agreements. The variables included are age, income, wealth, marital status, length as a client, amount of loan and length of loan.

1.3 Solution

1.3.1 Univariate Analysis

This involves calculating statistics for each variable like mean, median, standard deviation, skewness etc... The purpose of this task is to provide a general feeling for the data. This information can be used as a first check before applying the model to a particular client. For example, if the average age in the data is 65 and the client is 18, it is likely that the model will not be relevant.

1.3.2 Multivariate Analysis

We need to decide which variables from the data to include within our model. Firstly we compute the correlation matrix. With this information, if any two variables are highly correlated we may discard one of them as it does not provide any new information. The second test is the χ^2-test. This tests the dependance between each variable and the response variable, in our case whether the client defaulted. If there is significant statistical evidence that a variable is independent of default, it does not make sense to include it in the model. With our data the χ^2-test indicated we should drop the variables length of loan and amount of loan.

1.3.3 Model Creation

We are looking to compute the default probability using the regression model:

$$\text{Default} = f(x_1, \ldots, x_n) + \epsilon,$$

where $x_1, \ldots, x_n$ are the explainatory variables (age, income,), ϵ is the residual, and f is our function for regression. As our response variable is binomial we use the logistic (logit) model where

$$P(\text{default} : X) = \frac{1}{1 + \exp(-X\boldsymbol{\beta})}$$

where $\boldsymbol{\beta}$ is a vector of parameters to be determined from the data.

$\boldsymbol{\beta}$ can be easily calculated using the `MATLAB` function `glmfit` or using the `SAS` procedure `proc logistic`.

Calculated β_i.

β_1	Intercept	-1.85136
β_2	Age	-0.02678
β_3	Income	0.10025
β_4	Wealth	-0.01761
β_5	Marital Status	0.79651
β_6	Maturity	0.00892

1.3.4 Model Validation

We have used Powerstat method, as it is implicit in the proc logistic at SAS.

We have been able to obtain a wrong rate of 23.6%. We establish an estimation like default if it is greater than the frontier: 0.31

Obtained Results:

Defaults Estimated:	137
Wrong Estimations:	173
Observed Defaults:	138

1.3.5 Model Calibration

The expected loss is defined as:

$$EL = PD \times EAD \times LGD,$$

where PD is the is the percentage of defaults, EAD is the exposition to default and LGD are losses given default.

PD is defined as default probability calibrated for a year

Scoring probabilities allows us to sort people against default. However, when we try to understand these probabilities we must realize that these probabilities do not take into account when default happens.

This is the reason for calibrate the Scoring results, and obtain the yerly average probability.

To obtain the calibration, we have to get a sample of people who have been observed in periods of years. The model is applied to each person to obtain the scoring and then grouped by score.

We count how many people are in default, obtaining a default observed rate. The aim of the calibration is to model with a formula that default rate:

$$A(C + score)^B.$$

A, B, C must be estimated by minimizing the least squares error. We compute this in `MATLAB` using the function `fminsearch`. The values obtained are:

$$A = 0.0004, \quad B = 3.7410, \quad C = 2.7870.$$

2 Capital Allocation

2.1 The problem

In this problem a lender has a fixed amount of money to lend, EAD, between n blocks of similiar customers. The lender would like to know, for a given level of risk, how he should distribute his money between the blocks to maximise his profit.

Each block has associated with it an interest rate ρ_i, an a priori probabilty of default PD_i, the loss given default LGD_i and the number of customers N_i.

If each customer in each block is independent of the rest then we can easily compute the probability of k defaults using the binomial distribution. However the customers are correlated via the economy. Using Gaussian-Copula we can introduce a loss distrbution for each customer

$$Z_i = \sum_{j=1}^{m} a_j^i Y_j + r_i w_i.$$

The Y_j are indices for different parts of the economy, maybe unemployment rate, stock market index etc... We assume that the state of the economy is fixed. The a_j^i represent the weighting given to each part of the economy for the customers in block i. w_i is a standard normal variable for the systemic risk associated with each customer and r_i the standard deviation. We now have that

$$\Phi(Z_i) < PD_i \iff \text{Default}.$$

We can use the above to show that the probability of default given a particular state of the economy is

$$p_i = \Phi\left(\frac{\Phi^{-1}(PD_i) - \sum_{j=1}^{m} a_j^i Y_j}{r_i}\right).$$

When presented this problem starting point was to simulate the w_i for each customer in every block and repeat thousands of times to produce a simulated loss distribution. This approach take a long time to compute and it is not clear how to then optimise given the simulated loss distribution. We would like to seek an analytical expression for the loss distribution.

The probability p_i represents the independent probabilty of default for each customer in block i. We are now able to use the binomial distribution.

$$P(\text{k defaults}) = \binom{N_i}{k} p_i^k (1-p_i)^{N_i-k}.$$

As the N_i are in the order of 10^3 we can use the Central Limit Theorem to approximate the binomial distribution with a normal random variable, D_i, with mean $N_i p_i$ and variance $N_i p_i(1-p_i)$.

$$D_i \sim N(N_i p_i, N_i p_i(1-p_i)).$$

We use $\alpha_1 \dots \alpha_n$ to denote the fraction of EAD allocated to each of the n blocks. We assume that each person in a block borrows an equal fraction of the moeny, namely $\frac{\alpha_i EAD}{N_i}$. The loss distribution can now be constructed as

$$L = \sum_{i=1}^{n} \frac{\alpha_i EAD}{N_i}(LGD_i D_i - (N_i - D_i)\rho_i).$$

As the D_i are normal random variables we see the L is also a normal random variable.

$$L \sim N(\mu_L, \sigma_L^2)$$

where

$$\begin{aligned} \mu_L &= \sum_{i=1}^{n} \alpha_i EAD(LGD_i\ p_i - (1-p_i)\rho_i) \\ \sigma_L^2 &= \sum_{i=1}^{n} \frac{\alpha_i^2 EAD^2 (LGD_i + \rho_i)^2}{N_i} p_i(1-p_i). \end{aligned}$$

The problem is to minimise expected loss, μ_L, such that the α_i's sum to one and the level of risk is fixed. To measure risk we use Value at Risk (VaR) with a 99% confidence level. Formally:

Minimise

$$f(\boldsymbol{\alpha}) = \mu_L$$

subject to the constraints

$$\begin{aligned} \sum_{i=1}^{n} \alpha_i &= 1 \\ -2.3262 \times \sigma_L + \mu_L &= VaR99, \end{aligned}$$

where $VaR99$ is the fixed level of risk the lender is willing to take.

2.2 Implementation

We used `MATLAB` to implement the above optimisation problem.

To start with we took 3 blocks of customers with 3 economic indices. Initially we ran through a large number of possible $\boldsymbol{\alpha} = (\alpha_1 \dots \alpha_3)$. For each $\boldsymbol{\alpha}$ we were able to compute the $VaR99$ and μ_L. This was then plotted to produce an efficient border. This did not take long to compute for 3 blocks, however for larger number of blocks this process is computationally expensive.

We then used the `MATLAB` optimisation function `fmincon` to find the optimal $\boldsymbol{\alpha}$ for a given $VaR99$. Figure 1 shows the two methods compared. The block parameters suck as $\rho_i, N_i, ...$ were chosen at random.

We can see that the optimisation has worked well in choosing the $\boldsymbol{\alpha}$ which gives the lowest loss for a given level of risk. As a check we also implemented the simulation of w_i for a given $\boldsymbol{\alpha}$ to compare with our analytic distribution. We got very good agreement between the two, on the order of 10^{-4}.

The next step was to increase the number of blocks to 5. The brute force approach worked but took considerably longer than with 3 blocks. The optimisation however did not perform. The reason for this is not known for sure. It could be down to unrealistic parameters, a bad starting point or a bad choice of solver. Figure 2 shows the brute force approach applied to five blocks.

2.3 Conclusions

We found that the analytical method outperformed the simulation of w_i as expected. To optimise for more than 3 blocks the choice of optimiser needs to be investigated

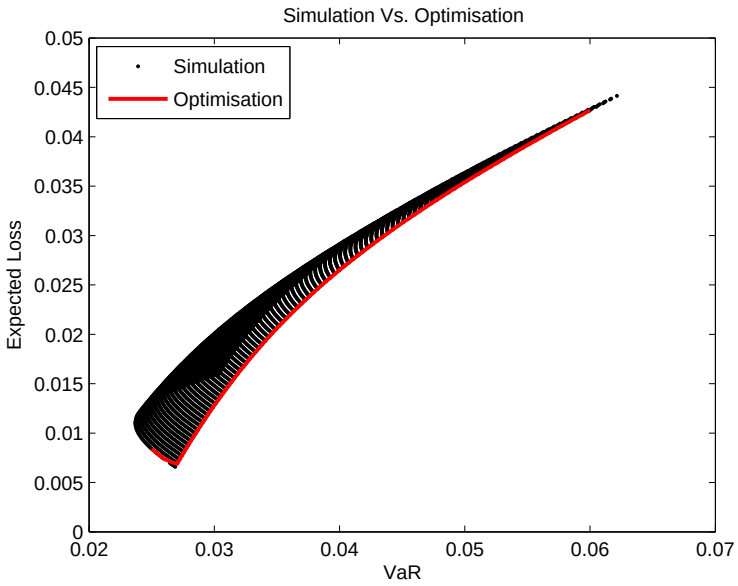

Figure 1: Brute Force vs. Optimisation for 3 blocks

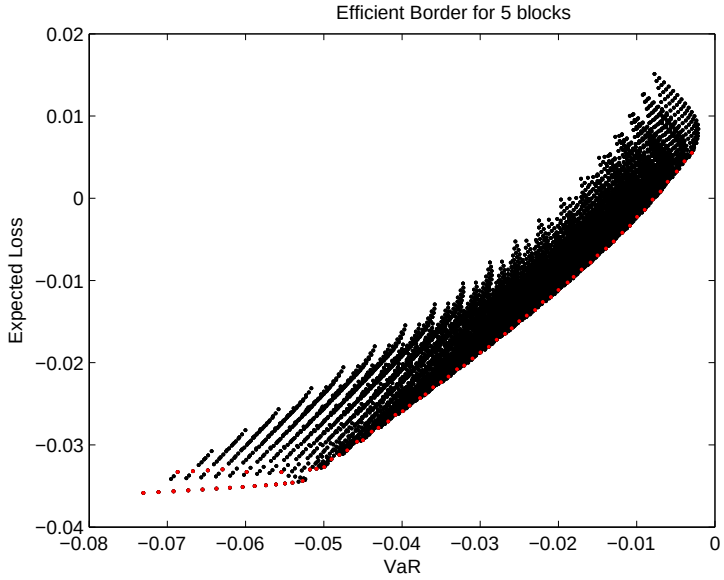

Figure 2: Brute Force for 5 blocks

furhter. Another interesting question is to look at the relationship between the efficient border and the interest rates charged for each block. We can also make the ecomony indices random variables and see how that changes the problem.