## Building Cross-Classified Multilevel Structure for Credit Scorecard: A Bayesian Monte Carlo Markov Chain Approach

ALESIA KHUDNITSKAYA *(Technische Universität Dortmund, Ruhr Graduate School in Economics, Germany, Khudnitskaya@statistik.tu-dortmund.de)*

The asymmetry of information between loan grantors and borrowers will always be subject to detailed monitoring of customers credit worthiness. Therefore, this paper aims to introduce an improved type of credit scoring model - a cross-classified multilevel scorecard.

The model produces forecasts of probability of default on a loan for individual borrowers and could be implemented in decision-making process in retail banking. The credit scoring model is built within two-step: the clustered structure is created and then the model is fitted.

The clustered structure is a cross-classified multilevel, which has the individual applicants for a loan cross-classified by their living environments (microenvironment), occupational fields and infrastructure of shopping facilities in the area of their residence. Applicants for a loan are considered as lower-level units in this clustered set-up. The structural approach helps to model exposure to unobserved random effects which have a certain impact on default. We determine environment, occupation and infrastructure-specific effects as being random and include them at the second level of model hierarchy along with the other explanatory variables. These random effects bring additional information and used to model exposure to environment-specific risk factors, occupational hazards and infrastructure risks.

In the second step, the multilevel credit scoring model is fitted by applying Bayesian modelling. The Bayesian Monte Carlo Markov Chain approach is appealing here as we simulate random effects of particular microenvironment or occupation and are interested in making inference of the parameter estimates in the population of different microenvironments or population of various occupations. Combining prior information on random effects and the data we try to simulate a joint posterior distribution and make point or interval estimates of the model parameters of interest. Mainly, the question lies in calculation between clusters variation within each of the cross-classifications.

The non-informative and weakly informative prior distributions are used for the variance parameters of the random effects. The starting values are obtained after model was fitted in Stata. For convergence-checking purpose three chains in parallel are performed with average chain-length of 500.000 iterations.

The data used in the study is a random sample of 5956 observations which includes detailed information on customers who apply for a bank loan. The individual level data contains personal information including income, marital status, number of dependents, profession, age and others. Credit Reference Agency data provides detailed data on derogatory reports, credit enquiries and other Court records. Market descriptive data includes detailed information for the 5-digit area zip code in which applicant resides. The sample probability of default is 9.5%.

The obtained results confirm a reasonable advantage of applying cross-classified multilevel structure to set up a credit scoring model. There is a significant variance within clusters in each of the three cross-classifications. Compared to the conventional logistic regression, cross-classified multilevel scoring model gives more accurate predictions: smaller Brier score and information criteria statistics (BIC, AIC).

*Keywords*: multilevel statistical modelling, cross-classified structure, random-effects, credit scoring, Bayesian Monte Carlo Markov Chain, logistic regression