**Abstract No: MR-15**

## Case study of credit risk analysis and creditworthiness prediction at loan approval

B. N. Weralupitiya and R. V. Jayatillake*

Department of Statistics, University of Colombo, Sri Lanka
rasika@stat.cmb.ac.lk*

The credit risk is considered as the risk associated with a borrower's failure to pay the loan or interest amount on time. An increase in Non-Performing Loan (NPL) ratio directly affects the financial performance of the banks as well as the economy of the country as a whole. Therefore, this case study was carried out for a specific bank in Sri Lanka with the objectives of developing a predictive model to assess the creditworthiness of potential loan applicants at the approval and to identify factors associated with time to first default. The data used for this study consisted of bank loan details of 10,626 existing customers in their current loan portfolio and their repayment behavior over 2.5 years. It consisted of 11 continuous and 7 categorical variables including customer's demographic details, personal financial details and bank-specific ratios. Furthermore, it included 10 transaction variables which all are categorical. The univariate tests such as Mann Whitney Test and Chi-Square Independence Test and graphical analyses identified that apart from variables "Age at Approval" and "CRIB Status at the approval", all the other variables showed a significant relationship with the variable of interest, "Loan Status". As only 33% of the respondents were non-performers, the Synthetic Minority Oversampling Technique (SMOTE) was used to handle the class imbalance. Several machine learning techniques such as Logistic Regression, Random Forest, Support Vector Machine, and Artificial Neural Network were applied with and without SMOTE Sampling to achieve the optimal model by comparing the ROC-AUC value of each model. The Artificial Neural Network model applied with SMOTE sampling was found to be the best model with a ROC-AUC value of 91.6%. Furthermore, the study data consisted of the borrowers' default status in every quarter. Therefore, a discrete-time hazard survival model was developed to identify the predictors that affect most of the risk of first default. It was found that the risk of first default to be higher in early quarters and decreases over time with the best fitted discrete survival model. Particularly, "Security type", "Loan to value ratio", "Tenure", "Purpose of the loan" and "Interest rate" were some of the variables found as the most significant variables that associate with the risk of first default.

**Keywords:** Creditworthiness, Machine learning, Non-Performing loans, Synthetic minority over sampling technique, Survival analysis