The impact of low number of events and class imbalances in PD model development

Abstract

Estimating the risk parameters for banking portfolios that exhibit a low-default or zero-default nature (referred to as low-default portfolios or LDPs) poses a common challenge to financial institutions worldwide. Frequently, default events are minimal and extremely underrepresented in the data, causing the event classes to be unbalanced (also called class imbalance).

For industry practitioners who develop models for risk parameters based on LDPs and class-imbalanced data, it is not well researched (in a general context) what impact the low number of events, and the class imbalance problem will have on some aspects of the model performance:

- the accuracy of predictions
- the discrimination ability of the model
- classification cut-offs

In this research, we used a simulation design that mimics real-world LDP examples and different class imbalanced scenarios to generalise the likely model development outcomes should a model be developed on such data. More specifically, we investigated the impact of the size of the development data set, the event rate and the strength of the association between the response and the predictors on various measures such as the F1/P4 score, the Gini coefficient and the optimal cut-off used for classification.

Some general results are that, for a fixed event rate and a given degree of association between the response and the predictors, a maximum median F1 or P4 score can be achieved. We also observed that the degree of association plays a crucial role, where a higher F1 score is achieved in the case of a stronger association. Furthermore, the level of class imbalance heavily influences the maximum achievable F1 or P4 score, where a very small event rate can greatly reduce the classification accuracy.

In the results it is striking that the Gini coefficient for all considered event rates seems to converge to the same value as the sample size increases. This is in contrast with the behaviour of the F1 score, which converges to a unique value for each event rate.

When considering the optimal cut-off for classification, there is an optimal cut-off for a specified event rate, which is somewhat independent of the sample size for a given degree of association between the response and the predictors.

In conclusion, this research could result in deriving practical guidelines and thresholds of achievable model performance metrics given the characteristics of the development data.

Authors & Affiliations

Prof Willem D Schutte[1,2], Prof Charl Pretorius[1], Dr Neill Smit[1], Mrs Leandra Van der Merwe[3], Mr Robert Maxwell[1]

[1]North-West University, Potchefstroom, South Africa. [2]Absa, Sandton, South Africa. [3]North-West University, Vanderbijlpark, South Africa