Addressing Class Imbalance in Loan Default Prediction: A Cluster-Based Synthesis Approach

[1] Yue Yang, [1] Boon Giin Lee, [1] Anthony Graham Bellotti, [1] Qinglin Mao, [1] Honghao Zhang
[1] Yifan Yu

[1] School of Computer Science,
University of Nottingham Ningbo China
Ningbo, Zhejiang 315100, China

ABSTRACT

Prediction of loan default is essentially a binary classification problem. Nevertheless, most available public loan-level dataset is highly imbalanced where the default data account for less than 1% of all the loans. The issue of class imbalance has a significant impact on causing a model heavily biased on classification, which is not clearly presented in the common performance evaluation metrics. Balancing of both default and non-default data in a dataset through data synthetization technique, by increasing the data size of default data is an important means of addressing the model bias issue. This study proposes an integration of clustering and oversampling methods based on the foundation of state-of-the-art methods, including SMOTE, GAN, etc., to preprocess the dataset. The method is tested on the US Freddie Mac single family loan-level dataset. The dataset is divided into several clusters using hierarchical clustering, to address the "Simpson's paradox" issue. Then, the SMOTE and GAN (and other) methods are carried out to synthesize minority data where the data size of default data matches the data size of non-default data in each cluster. Additionally, this study explores the impact of preprocessing techniques on model performance. The study utilizes classifiers such as decision tree, multilayer perception, CatBoost classifier and XGBoost classifier to perform classification on the augmented dataset and use ensemble learning to improve the predictive performance of the classification models. The experimental results indicate that the default prediction using cluster-based data as training and testing samples achieved better performance, compared to the default prediction using whole data as training and testing sets. The proposed method indicates significant improvements of predicting minority class (default) which addresses the generalizability limitation of existing oversampling methods.