Prediction of Credit Card Defaults in Taiwan Based on Variable Selection Approach in Machine Learning Model

Siriprapa Tewee[1] Tanaporn Chaisinratham[1] Apichaya Buajee[1]

Lawitra Phanpanich[1] Siriyakon Saenpor[1] Chalermrat Nontapa[2*]

[1]*Data Science Research Center, Faculty of Science, Chiang Mai University*

[2]*Department of Statistic, Faculty of Science, Chiang Mai University*

(**chalermrat.n@cmu.ac.th*)

Abstract

This study aims to predict loan defaults using debtor behavior data from a Taiwanese bank, comprising 30,000 records available on Kaggle. Predicting loan default risk is important to financial institutions, as accurately predicting the likelihood of a borrower defaulting on their loans will help to reduce financial losses, thereby maintaining profitability and stability. The dataset includes key features such as credit amount, gender, education, marital status, age, monthly repayment status, billing statements, and repayment amounts from April to September 2005, along with the default status in the following month. To evaluate prediction performance, four machine learning models—Random Forest, XGBoost, CatBoost, and LightGBM. Each model represents a different approach to ensemble learning (e.g., bagging vs. boosting), which allows for a comprehensive comparison of their performance on the task at hand. These models were trained on 24,000 records and tested on 6,000. The results show that LightGBM also demonstrated strong performance in other metrics, with a Precision of 79.73%, Recall of 81.58%, and an F1-Score of 79.48%. LightGBM is the most effective model for predicting loan defaults, highlighting the advantages of gradient boosting techniques in credit risk assessment and financial decision-making.

คำสำคัญ: Loan Defaults, Financial, Variable Selection, LightGBM, XGBoost