Research Proposal

The credit card issuers continuously face credit debt crises, especially under severe economic stress, like the COVID-19 pandemic effect. How to lower the risk of credit default is their persistent business focus. This project aims to determine potential defaulters based on the client's previous card activities, socioeconomic and demographic status using multiple machine learning algorithms. Ideally, we can find a reasonable model with good accuracy for card issuing companies and their shareholders.

We will use the dataset from the UCI machine learning repository. They were collected in Taiwan between April 2005 to September 2005, including 30000 instances with 24 attributes. The credit line, gender, education, marital status, past payments history, bill statement, and the amounts of previous payments are the predictor variables. Default, whether the client will default or not, is the target variable. We will focus on answering the following key business questions in this study:

I. How does the probability of credit card default vary by different predictors?
II. Which predictor or predictors are the most essential for the default payment?
III. Which classification algorithms is the best choice for our prediction?
IV. Are there other vital predictors not listed in this dataset when we develop a credit card default model?

To answer the above questions, we will design a workflow comprising exploratory data analysis, feature engineering, data preprocessing, modeling development, and model evaluation. Several classification algorithms will be tested, from the simple K-nearest neighbor (KNN) classifier, Logistic regression, to the popular ensemble algorithms like random forest, gradient boosting, Ada boosting, and LightGBM boosting. Standard classification metrics, accuracy, recall, f1-score, ROC, and AUC, will be applied to evaluate the classification models. Finally, we will summarize all critical findings in our reports and presentation.