Enhanced Credit Card Default Prediction Using SMOTE, LSTM, and Stacked Ensemble Methods

Project Proposal

Muhammad Khalid Al Ghifari
Department of Informatics
Syiah Kuala University
Banda Aceh, Indonesia
khalid.22@mhs.usk.ac.id

Andrie Fadhlullah Wahby
Department of Informatics
Syiah Kuala University
Banda Aceh, Indonesia
andri33@mhs.usk.ac.id

Project Category: Finance & Commerce

***Abstract*—Credit card default prediction is a core problem in financial risk management, where inaccurate predictions may cause substantial financial losses. Real-world credit datasets are typically imbalanced, leading many machine learning models to achieve high accuracy but poor default detection. This proposal aims to improve default prediction by combining class imbalance handling using Synthetic Minority Over-sampling Technique (SMOTE), temporal modeling of payment behavior with Long Short-Term Memory (LSTM) networks, and stacked ensemble learning. Experiments will be conducted on the UCI Credit Card Default dataset to evaluate improvements in recall, F1-score, and AUC while maintaining acceptable overall performance.**

***Index Terms*—credit card default, class imbalance, SMOTE, LSTM, ensemble learning**

I. Introduction

Credit card default prediction plays a critical role in financial institutions, as it directly affects credit risk management and lending decisions. The rapid growth of consumer credit usage increases the exposure of financial institutions to default risk, making accurate and reliable prediction models essential. However, most real-world credit datasets exhibit severe class imbalance, where non-default cases significantly outnumber default cases.

In imbalanced settings, conventional machine learning models tend to optimize overall accuracy by favoring the majority class. This behavior results in poor recall for default cases, even when reported accuracy appears high. In practical applications, failing to identify defaulters is more costly than incorrectly flagging low-risk customers, which highlights the inadequacy of accuracy-centric evaluation.

Recent studies on the UCI Credit Card Default dataset demonstrate that advanced models such as deep neural networks can achieve accuracy above 80%, yet recall for default cases often remains below 40%. This imbalance between accuracy and recall motivates the need for approaches that explicitly prioritize default detection.

A. Motivation

This project addresses credit card default prediction as an applied machine learning problem in the financial domain. The primary motivation is to reduce false negatives by improving the detection of high-risk customers. This study explores whether combining class balancing techniques, temporal modeling of payment behavior, and ensemble learning can produce models that better align with real-world risk management objectives.

II. Related Work

Early research on credit scoring relied on statistical models such as logistic regression and linear discriminant analysis due to their interpretability and solid theoretical foundations [1], [2]. However, these methods assume linear relationships and often fail to capture complex interactions among financial variables.

Machine learning techniques such as support vector machines, random forest, and gradient boosting have been introduced to address non-linearity and feature interactions [3], [4]. While these models often improve predictive accuracy, several empirical studies report that recall for default cases remains limited in highly imbalanced datasets.

To mitigate class imbalance, sampling-based approaches such as SMOTE have been proposed [5]. Previous studies show that SMOTE can improve minority class learning in financial risk prediction, yet its effectiveness in combination with modern ensemble models for credit card default prediction has not been thoroughly explored [6].

Moreover, most existing studies treat payment behavior as static features, ignoring temporal dependencies across billing cycles. LSTM networks, designed to model sequential data, have demonstrated success in financial time-series prediction but remain underutilized in credit card default research [7]. The main research gap lies in the lack of integrated frameworks that jointly address class imbalance, temporal behavior, and model diversity.

III. Method

A. Dataset and Preprocessing

This study uses the UCI Credit Card Default dataset [8], which contains 30,000 customer records and a binary target variable indicating default payment. Approximately 22% of the observations correspond to default cases, resulting in a significant class imbalance. The dataset will be divided into training and testing sets using stratified sampling. Feature scaling will be applied using standard normalization.

B. Baseline Models

Six baseline models will be implemented to establish reference performance: linear discriminant analysis, logistic regression, support vector machine, random forest, XGBoost, and a deep neural network. These models follow configurations commonly reported in prior studies and serve as benchmarks for evaluating proposed improvements.

C. Proposed Models

To address class imbalance, SMOTE will be applied to the training data to generate synthetic minority class samples. Temporal payment behavior will be modeled using an LSTM-based architecture that processes monthly payment information as sequential data. In addition, a stacked ensemble will be constructed by combining XGBoost, random forest, and logistic regression as base learners, with logistic regression acting as a meta-learner.

IV. Intended Experiments

The experimental design aims to evaluate the contribution of each proposed component. First, baseline models will be replicated to verify consistency with reported results in the literature. Second, SMOTE-enhanced models will be compared against non-balanced counterparts to assess improvements in recall. Third, the LSTM model will be evaluated to determine its ability to capture temporal payment patterns. Finally, a stacked ensemble will be tested to examine whether combining diverse models improves overall performance.

Model evaluation will prioritize recall, F1-score, and AUC-ROC due to the imbalanced nature of the dataset. Confusion matrices will be analyzed to assess the trade-off between default detection and false alarms.

References

[1] D. J. Hand and W. E. Henley, "Statistical classification methods in consumer credit scoring," *J. R. Stat. Soc. A*, vol. 160, no. 3, pp. 523–541, 1997.

[2] I.-C. Yeh and C.-H. Lien, "The comparisons of data mining techniques for credit default prediction," *Expert Systems with Applications*, vol. 36, no. 2, pp. 2473–2480, 2009.

[3] J. Cervantes et al., "A comprehensive survey on support vector machine classification," *Neurocomputing*, vol. 408, pp. 189–215, 2020.

[4] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. KDD*, 2016.

[5] N. V. Chawla et al., "SMOTE: Synthetic minority over-sampling technique," *JAIR*, vol. 16, pp. 321–357, 2002.

[6] N. Suhadolnik et al., "Machine learning for enhanced credit risk assessment," *Journal of Risk and Financial Management*, vol. 16, no. 12, 2023.

[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[8] I.-C. Yeh, "Default of credit card clients," *UCI Machine Learning Repository*, 2016.

[9] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.