# Loan Default Prediction Project Report

## 1. Introduction

This project aims to build a classification model that predicts whether a loan applicant is likely to default, using the Lending Club Loan Dataset. The model can help lenders identify high-risk applicants and reduce the likelihood of defaults.

## 2. Objectives

- Build a classification model using LightGBM and SVM.  
- Preprocess the dataset by handling missing values and class imbalance using SMOTE.  
- Evaluate the model using classification metrics: Precision, Recall, and F1 Score.  
- Generate insights and recommendations for lenders.

## 3. Personal Findings

- LightGBM performed better than SVM in terms of classification metrics on the test set.  
- SMOTE effectively handled class imbalance, which significantly improved model performance.  
- Feature scaling was crucial for SVM performance but less impactful for LightGBM.  
- The target class (loan\_status) had many subcategories, but simplifying it to a binary classification task (default vs non-default) made the problem more tractable.

## 4. Problems Faced

- The dataset had numerous missing values and irrelevant columns, which required extensive cleaning.  
- Imbalanced classes made initial model training ineffective.  
- Encoding high-cardinality categorical variables posed a risk of overfitting.

## 5. Counter Actions Taken

- Dropped columns with excessive missing values and low information gain.  
- Used forward fill (ffill) for imputing missing values in relevant features.  
- Applied Label Encoding for categorical variables and used SMOTE to balance the classes.  
- Standardized features to improve SVM performance.

## 6. Conclusion

The LightGBM model emerged as the most effective in identifying loan defaulters. This project demonstrates the importance of preprocessing and handling class imbalance when working with real-world financial datasets. These insights can be leveraged by lenders to refine their risk assessment strategies and minimize losses.