Serge Nane

Student ID: 21415521

DSC 680 – T302\_2261\_1

**Project Domain: Finance / Banking**

**Milestone 1:** Proposal and Data Selection

**Topic: Predicting Loan Default Risk Using Customer Financial Profiles**

**Business Problem**

Financial institutions face increasing challenges in managing credit risk, as loan default can significantly impact profitability and economic stability. Traditionally, lending decisions have relied on limited financial indicators such as credit scores, income levels, or employment duration. While these factors are important, they do not always capture the complex relationships between multiple borrower attributes that influence repayment behavior. The use of predictive analytics and machine learning can substantially enhance risk assessment by identifying subtle patterns and interactions within large financial datasets.

This project aims to develop and compare predictive models to estimate the probability of loan default using borrower demographic and financial indicators. The results will demonstrate how data-driven insights can help financial institutions make more informed credit decisions, reduce default rates, and improve portfolio risk management. A key emphasis will be placed on interpretability, ensuring that the models provide clear and actionable insights for financial sector decision-makers.

**Dataset**

The dataset for this project will be obtained from the publicly available LendingClub Loan Data on Kaggle, titled “**Loan Default Prediction Dataset**.” It contains detailed information on individual loan applications and repayment outcomes, including loan amount, interest rate, employment length, annual income, credit score, loan purpose, and debt-to-income ratio. The target variable indicates whether a borrower defaulted on their loan or fully repaid it.

With over 250,000 observations, this dataset offers sufficient depth and variability for robust predictive modeling. It will be divided into features representing borrower and loan characteristics, as well as a binary target variable reflecting repayment status. The dataset’s size and diversity make it well-suited for training, validating, and testing machine learning models while enabling meaningful exploratory data analysis and feature selection.

**Research Questions**

This study seeks to address several key questions related to credit risk modeling. First, can machine learning models accurately predict loan default risk using customer financial and demographic indicators? Second, which borrower characteristics most strongly influence repayment behavior? Third, how do Logistic Regression and Random Forest models compare in terms of predictive performance, interpretability, and practical value to financial institutions? Lastly, how can model results be applied to improve lending strategies and risk mitigation practices in real-world banking operations?

**Methods**

The analysis will begin with comprehensive data preprocessing to ensure the dataset is suitable for modeling. This will include the treatment of missing or invalid values, encoding of categorical features, normalization of continuous variables, and examination of class imbalance through resampling methods such as SMOTE. Exploratory Data Analysis (EDA) will be conducted to examine variable distributions, detect outliers, and identify correlations between borrower characteristics and default outcomes.

Modeling will involve two primary algorithms: **Logistic Regression** and **Random Forest**. Logistic Regression will serve as the baseline model due to its interpretability and ability to estimate the influence of each variable on the probability of default. Random Forest will be applied as a more complex ensemble method capable of capturing nonlinear relationships and interactions among predictors. Model evaluation will include metrics such as **accuracy, precision, recall, F1 Score, and ROC AUC** to comprehensively assess predictive performance. Visualization tools such as correlation matrices, confusion matrices, ROC curves, and feature importance charts will be used to support model interpretation.

While this project will primarily focus on these two algorithms, **XGBoost** may be considered for future analysis to explore potential performance enhancements through advanced gradient boosting techniques.

**Ethical Considerations**

Given that the dataset contains sensitive financial and demographic information, maintaining high ethical standards is crucial. Although the data are anonymized, all analyses will adhere to principles of data privacy and fairness. Potential algorithmic biases will be carefully examined to ensure that model outcomes do not disproportionately affect specific demographic groups, such as age or employment type. The emphasis on interpretability will help prevent misuse of automated decision-making, ensuring that models support rather than replace human expertise in credit evaluation. All results will be presented transparently and responsibly to foster trust in data-driven lending tools.

**Challenges and Limitations**

One of the anticipated challenges involves the issue of class imbalance, as loan default cases typically represent a smaller proportion of the data. This imbalance could lead to biased model predictions favoring the majority class. To address this, techniques such as SMOTE or adjusted class weights will be applied to ensure balanced model training. Another limitation relates to the trade-off between interpretability and performance. While Random Forest may outperform Logistic Regression in accuracy, it is inherently less transparent, which can make its outputs harder to justify in a financial context. Additionally, borrower behavior and external economic conditions may introduce noise or unpredictability that limits model generalization.

**References**

Kaggle. (n.d.). Loan Default Prediction Dataset. Retrieved from

<https://www.kaggle.com/datasets/>.

Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann.

Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decision Support Systems, 50(3), 559–569.