**3.0 Methodology**

By estimating a customer's probability of missing payments, credit default prediction models help financial organisations manage risk. Machine learning provides strong tools for creating these predictive models as structured data becomes more widely available. This section explores the procedures and methods used to develop a model that uses a variety of machine learning approaches to forecast credit default. Data preparation, feature engineering, model selection, evaluation metrics, and model performance analysis are all covered in this part.

### 3.1 About the Dataset

The UCI Machine Learning Repository is the source of this dataset, which sheds light on the 2006 credit card debt problem that affected Taiwanese banks. Many banks gave credit cards to unqualified applicants in an effort to gain market share, which led to widespread debt buildup and higher cardholder delinquency rates. Customers' trust in personal finance was eroded by this practice, which presented serious problems for banks and consumers alike.

### 3.1.1 Column Information

Repayment status (PAY\_0 to PAY\_6) from April to September 2005 is one of the important features in the dataset that provides information about customer behaviour and financial history. Values in this field reflect whether payments were made on time, late, or significantly past due. Payment amounts (PAY\_AMT1 to PAY\_AMT6) reflect prior payments made each month, whereas bill amounts (BILL\_AMT1 to BILL\_AMT6) reflect the monthly statement balances in NT dollars. These characteristics are crucial for assessing credit risk and comprehending repayment patterns.

### 3.2 Data Preparation

### 3.2.1 Loading and Initial Analysis

### Loading the dataset and performing a preliminary examination to comprehend its structure are the first steps in the data preparation process. The first rows are shown using df.head(), exposing a number of characteristics pertaining to client demographics and credit history. *SEX, EDUCATION, MARRIAGE,* and *default* the goal variable that indicates if a consumer has defaulted are important characteristics. This stage helps us find key features for the credit default prediction model by giving us a brief summary of the data. Running df.info() also provides useful information on the counts of non-null values, the data types for each feature, and any possible missing or inconsistent data that should be cleaned up before modelling.

### 3.2.2 Dropping Unnecessary Columns

### It's possible that some of the dataset's columns are irrelevant to the prediction task. In this instance, the ID field has no predictive value and only functions as a unique identifier. Using df.drop('ID', axis=1, inplace=True), it is removed from the dataset because it has no direct correlation to the likelihood of default. By removing superfluous variables from the dataset, this phase lowers the likelihood of overfitting and frees the model to concentrate on more significant characteristics. Additionally, eliminating unnecessary columns improves data processing and storage, which is very advantageous in machine learning workflows.

### 3.2.3 Encoding Categorical Variables

### Since many machine learning algorithms need numerical input, categorical features must be encoded. The categorical columns in this dataset are MARRIAGE, EDUCATION, and SEX. LabelEncoder is used to convert these labels into numerical numbers. For instance, SEX categories like "male" and "female," respectively, are changed to 0 and 1. In a similar manner, categories in MARRIAGE and EDUCATION are represented numerically. Additionally, the target variable default undergoes label-encoding, which transforms it into a binary format appropriate for classification tasks. To guarantee interoperability with machine learning models and to standardise the data format, these variables must be encoded.

### 3.2.4 Exploratory Data Analysis (EDA)

### The purpose of exploratory data analysis is to learn more about the dataset's distribution, relationships, and any problems. Essential metrics, such as the mean, standard deviation, min, and max values for every feature, are provided via basic statistical summaries (df.describe()). These summaries show each variable's range and can indicate the existence of outliers or skewed distributions that could impair model performance.

### For every variable, histograms are plotted in order to better evaluate feature distributions. Skewness, kurtosis, and other anomalies in the data, including extreme values or concentrations around particular spots, can be found with the use of these visualisations. For instance, to enhance model interpretability and performance, adjustments can be necessary if a variable's histogram has a significant skew.

### To investigate the connections between characteristics and spot possible multicollinearity problems, a correlation heatmap is created. We can see which features are significantly associated with the target variable and with each other by computing the correlation matrix and displaying it as a heatmap. While features with high inter-correlation may be redundant, those with a significant correlation to the target variable may be very useful in the model. Therefore, the heatmap acts as a guide for feature selection by highlighting which variables might need to be changed, removed, or combined.

### EDA provides information about feature distributions, correlations, and possible problems with data quality. The subsequent stages of the data preparation process, including feature scaling, managing skewed variables, and selecting feature engineering tactics, are guided by this knowledge. By minimising noise and optimising the information available to the model, these data preparation procedures together produce a refined dataset that is better suited for machine learning, ultimately increasing the predictive ability and resilience of the model.

### 3.3. Feature Selection and Engineering

### 3.3.1 Feature Importance using Random Forest

### A RandomForestClassifier is used to compute feature importance in order to identify which attributes have the most impact on the target variable. The dataset is used to train the classifier, and a bar chart displays the feature importance results. This graphic makes it evident which characteristics are most important in forecasting credit default. Features that have a major influence on the model's predictions are indicated by higher significance ratings. We may improve the model's efficiency and interpretability by concentrating on these important aspects. While keeping crucial predictive information, less significant traits may be eliminated to cut down on noise and the possibility of overfitting.

### 3.3.2 Outlier Detection

### Machine learning models' accuracy and stability may be impacted by outliers. Box plots are created for every feature in order to highlight data points that significantly depart from the normal range. Errors in data entry, measurements, or special circumstances can all lead to outliers. To lessen their influence, techniques like adjustment (such as log scaling) or removing extreme values may be used, depending on the distribution and model performance. When outliers are handled correctly, the model maintains predicted accuracy and generalises well to new data.

### 3.3.3 Target Variable Analysis

It is crucial to examine the distribution of the target variable, especially in order to spot any class imbalance in the dataset. Whether there are an unequal number of default and non-default cases in the dataset is indicated by a count plot for the target variable (default). Model performance can be distorted by class imbalance, frequently favouring the majority class. In order to balance the dataset and increase the model's capacity to forecast minority class outcomes, methods like modifying class weights or using resampling approaches (like SMOTE) may be employed if there is a considerable imbalance.

### 3.4 Data Splitting and Scaling

### 3.4.1 Splitting Data

### A 70-30 split is used to separate the dataset into training and test sets. This guarantees that 30% of the data is set aside for testing and 70% is used to train the models. In order to ensure that any class imbalance in the original data is reflected proportionately in both the training and test data, stratification is used to maintain the target variable's distribution across both sets. As a result, assessment metrics become more trustworthy.

### 3.4.2 Feature Scaling

### StandardScaler is used to scale each feature to a mean of zero and a standard deviation of one in order to standardise feature data. Because it increases convergence speed and model stability, this is crucial for models like logistic regression that are sensitive to feature scales.

**3.5 Modeling and Evaluation**

Three machine learning models—Logistic Regression, Random Forest Classifier, and XGBoost Classifier—are used in this investigation to forecast credit card default. These models were selected because to their diverse strengths, which include robust performance on unbalanced datasets, interpretability, and insights into feature relevance. To counteract any potential class imbalance that can skew models towards the majority class, balanced class weights are given to each model after it has been trained on the dataset. To give a comprehensive evaluation of each model's performance, evaluation measures including accuracy, precision, recall, and F1-score are computed using confusion matrices and classification reports.

**3.5.1 Logistic Regression**

For binary classification applications, logistic regression is frequently used due to its simplicity and interpretability. As a linear model, it assigns probabilities to each instance by assessing the link between the attributes and the chance of credit default. Standardisation approaches are used to train the model using scaled data, which enhances the performance of distance-based models such as Logistic Regression. A confusion matrix that includes the number of true positives, true negatives, false positives, and false negatives is used to assess the predictions on the test set. The categorisation report also displays important parameters including F1-score, recall, accuracy, and precision. These measures offer a thorough analysis of the model's effectiveness in accurately separating defaults from non-defaults, facilitating a deeper comprehension of its predictive power.

**3.5.2 Random Forest Classifier**

In order to increase accuracy and decrease overfitting, Random Forest is an ensemble technique that constructs several decision trees and aggregates their predictions. Because it can handle high-dimensional data and offer insights regarding feature relevance, this model is especially useful for identifying the characteristics most closely linked to default risk. To guarantee that minority classes are treated fairly, the balanced class weight parameter is included. A confusion matrix and a classification report that summarises accuracy, precision, recall, and F1-score are used to assess Random Forest's performance, same like in Logistic Regression. Compared to single-decision-tree models, the model's ensemble nature improves stability and accuracy, and its feature importance rankings provide insightful information about how different qualities affect default risk.

**3.5.3 XGBoost Classifier**

An improved gradient boosting technique called XGBoost is used because of its great efficiency and predicted accuracy, particularly when applied to structured data. XGBoost is a boosting technique that works especially well with imbalanced datasets since it creates trees in a sequential fashion, each one attempting to fix the mistakes of the one before it. Hyperparameters are used to optimise the model in order to avoid overfitting and guarantee that it recognises intricate patterns in the data. A confusion matrix and classification report are used to assess XGBoost's predictions on the test set following training on the scaled dataset. XGBoost is a strong option for credit default prediction because of its high precision and recall scores, which show that it is able to strike a compromise between forecast accuracy and sensitivity to the minority class.

All things considered, the combination of these models offers a thorough assessment of the dataset. XGBoost provides high accuracy and robustness against imbalances, Random Forest provides insights into feature relevance, and Logistic Regression provides interpretability. Based on the precise needs of accuracy, interpretability, and model resilience, the optimum model for credit default prediction can be found by comparing their performances.

### 3.6 Confusion Matrix and Model Comparison

A confusion matrix that shows the numbers of True Positives, True Negatives, False Positives, and False Negatives is created for each of the three models—Logistic Regression, Random Forest Classifier, and XGBoost Classifier—to provide a thorough analysis of performance. This matrix, which shows the accuracy of forecasts for defaulters versus non-defaulters, is a crucial tool for evaluating each model's predictive quality. In particular, the matrix sheds light on each model's capacity to differentiate between true defaulters and non-defaulters, which is critical for assessing model trustworthiness in this situation. We can determine which model is most effective at lowering misclassification—specifically, False Positives and False Negatives—by comparing the confusion matrices. This has a direct bearing on risk management.

**3.7 ROC Curve and AUC Score**

Plotting the True Positive Rate (sensitivity) against the False Positive Rate, the Receiver Operating Characteristic (ROC) curve illustrates the trade-offs between sensitivity and specificity and offers a visual depiction of each model's performance. Because it exhibits a better balance between accurately recognising genuine positives and minimising false positives, a model with a high Area Under the Curve (AUC) score is typically more robust.

**3.8 Conclusion**

In summary, this work uses data from the UCI Machine Learning Repository to show how machine learning models, such as Logistic Regression, Random Forest, and XGBoost, can be used to forecast credit default. These models provide useful insights into credit risk by utilising data preparation strategies like encoding, feature selection, and addressing imbalances. A thorough evaluation is provided by metrics such as accuracy, precision, recall, and AUC, with XGBoost demonstrating the highest performance because of its capacity to manage imbalances and intricate patterns. This emphasises how crucial model tuning and selection are to financial organisations' ability to anticipate loan default effectively.