# Credit Card Default Prediction High-Level Document

## 1. Introduction:

### 1.1 Background:

The Credit Card Default Prediction project aims to leverage machine learning techniques to predict credit card defaults, contributing to effective risk management for financial institutions.

### 1.2 Objectives:

- Develop machine learning models for credit card default prediction.
- Evaluate model performance using different algorithms and handling imbalanced datasets.

## 2. Dataset Overview:

### 2.1 Data Source:

The dataset is obtained from the https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset

2.2 Details of the dataset:

LIMIT_BAL: This feature represents the credit limit assigned to the individual's credit card. It indicates the maximum amount of credit the person can utilize.

SEX: This feature represents the gender of the credit card holder. While gender itself may not directly impact credit card fault detection, it can be considered as a demographic factor that might have some influence on creditworthiness.

EDUCATION: This feature indicates the educational background of the credit card holder. It can provide insights into the person's level of education, which might indirectly correlate with their financial stability and ability to manage credit.

MARRIAGE: This feature represents the marital status of the credit card holder. Similar to gender, marital status can be a demographic factor that could potentially impact credit card fault detection.

AGE: This feature denotes the age of the credit card holder. Age can be an important factor in assessing creditworthiness as it often correlates with financial responsibility and stability.

PAY_0, PAY_2, PAY_3, PAY_4, PAY_5, PAY_6: These features represent the repayment status of the credit card for the past six months. The values indicate the payment status (e.g., -1 represents payment delay for one month, 0 represents payment on time, 1 represents payment delay for two months, and so on). These features are crucial in determining the payment behaviour of the individual over time.

BILL_AMT1, BILL_AMT2, BILL_AMT3, BILL_AMT4, BILL_AMT5, BILL_AMT6: These features represent the amount of bill statement for the respective months. They provide information about the outstanding balance on the credit card at specific points in time.

PAY_AMT1, PAY_AMT2, PAY_AMT3, PAY_AMT4, PAY_AMT5, PAY_AMT6: These features represent the amount of payment made by the credit card holder for the respective months. They indicate the actual payments made to reduce the outstanding balance.

default payment next month: This is the target variable or the dependent variable that indicates whether the credit card holder defaulted on their payment in the following month (1 for default, 0 for no default).

### 2.3 Data Exploration:

Performed initial data exploration, including displaying the top and last 5 rows of the dataset, checking null values, and obtaining dataset information.

## 3. Data Preprocessing:

### 3.1 Handling Imbalanced Data:

Considered the issue of imbalanced data and explored techniques for addressing it.

### 3.2 Feature Matrix and Response Vector:

Stored feature matrix (X) and response (target) vector (Y) for model training.

## 4. Train-Test Split:

### 4.1 Splitting the Dataset:

Divided the dataset into training and test sets for model evaluation.

## 5. Model Development:

### 5.1 Handling Imbalanced Dataset:

#### 5.1.1 Under sampling:

Utilized under sampling techniques to balance the class distribution.

#### 5.1.2 Oversampling:

Implemented oversampling techniques to balance the class distribution.

### 5.2 Model Selection:

#### 5.2.1 Logistic Regression:

Developed a Logistic Regression model for credit card default prediction.

#### 5.2.2 Decision Tree Classifier:

Implemented a Decision Tree Classifier for credit card default prediction.

#### 5.2.3 Random Forest Classifier:

Utilized a Random Forest Classifier for credit card default prediction.

## 6. Model Evaluation:

### 6.1 Performance Metrics:

Evaluated model performance using metrics such as accuracy, precision, recall, and F1-score.

## 7. Model Saving:

### 7.1 Save the Model:

Saved the trained models for future use or deployment.

## 8. Conclusion:

Summarized key findings, challenges, and the effectiveness of different models in predicting credit card defaults.

## 9. Future Work:

Outlined potential areas for improvement or extension of the project.

## References:

Cited relevant literature, datasets, and methodologies used in the project.