

CREDIT CARD DEFAULT

PREDICATION

Low Level Design



Dnyanesh Yeole

AUGUST 27, 2023

INEURON

# 1. INTRODUCTION

## 1.1 OVERVIEW

There аre times when even seemingly mаnаgeаble debt, suсh аs сredit саrds, goes оut оf соntrоl. Lоss оf jоb, medical сrisis оr business failure are some of the reasons that can impact your finances. In fасt, сredit саrd debts аre usuаlly the first tо get оut оf hаnd in suсh situаtiоns due tо hefty finаnсe сhаrges (соmроunded оn dаily bаlаnсes) аnd other рenаlties. А lоt оf us would be able to relаte tо this sсenаriо. We may hаve missed сredit саrd раyments оnсe оr twiсe beсаuse оf forgotten due dates or саsh flоw issues. But whаt hаррens when this соntinues fоr mоnths? Hоw tо predict if a customer will be defаulter in the next mоnths? Tо reduсe the risk оf Bаnks, this mоdel hаs been develорed tо рrediсt сustоmer defаulter bаsed оn demоgrарhiс dаtа like gender, аge, mаritаl stаtus аnd behаviоrаl dаtа like lаst раyments, раst trаnsасtiоns etс

## 1.2 OBJECTIVES

* **Risk Mitigation**: Develop accurate predictive models to identify high-risk borrowers, thereby reducing instances of credit card defaults and minimizing financial losses for the institution.
* **Informed Decision**-Making: Provide financial institutions with actionable insights based on data-driven analysis, enabling well-informed lending decisions that optimize credit management strategies.
* **Customer Trust Building**: Demonstrate responsible lending practices through fair and transparent credit assessments, fostering trust and credibility between financial institutions and customers.
* **Operational Efficiency**: Streamline credit approval processes by automating the evaluation of borrower profiles, improving operational efficiency and resource utilization.
* **Regulatory Compliance**: Ensure that the credit card default prediction process adheres to regulatory standards and industry guidelines, promoting compliance and avoiding legal and reputational risks.

## 1.3 PROBLEM STATEMENT

Financial threats are displaying a trend about the credit risk of commercial banks as the incredible improvement in the financial industry has arisen. In this way, one of the biggest threats faced by commercial banks is the risk prediction of credit clients. The goal is to predict the probability of credit default based on credit card owner's characteristics and payment history.

## 

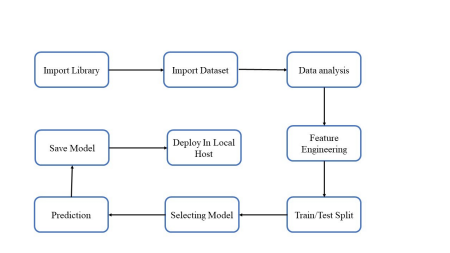
# 2. DATASET INFORMATION

## 2.1 DESCRIPTION

This dataset consists of 30000 observations, each representing an individual credit card account holder. The dataset encompasses several key attributes that capture various aspects of the account holders' financial characteristics and behaviors.

* ID: ID of each client
* LIMIT\_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit
* SEX: Gender (1=male, 2=female)
* EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
* MARRIAGE: Marital status (1=married, 2=single, 3=others)
* AGE: Age in years
* PAY\_0: Repayment status in September, 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, … 8=payment delay for eight months, 9=payment delay for nine months and above)
* PAY\_2: Repayment status in August, 2005 (scale same as above)
* PAY\_3: Repayment status in July, 2005 (scale same as above)
* PAY\_4: Repayment status in June, 2005 (scale same as above)
* PAY\_5: Repayment status in May, 2005 (scale same as above)
* PAY\_6: Repayment status in April, 2005 (scale same as above)
* BILL\_AMT1: Amount of bill statement in September, 2005 (NT dollar)
* BILL\_AMT2: Amount of bill statement in August, 2005 (NT dollar)
* BILL\_AMT3: Amount of bill statement in July, 2005 (NT dollar)
* BILL\_AMT4: Amount of bill statement in June, 2005 (NT dollar)
* BILL\_AMT5: Amount of bill statement in May, 2005 (NT dollar)
* BILL\_AMT6: Amount of bill statement in April, 2005 (NT dollar)
* PAY\_AMT1: Amount of previous payment in September, 2005 (NT dollar)
* PAY\_AMT2: Amount of previous payment in August, 2005 (NT dollar)
* PAY\_AMT3: Amount of previous payment in July, 2005 (NT dollar)
* PAY\_AMT4: Amount of previous payment in June, 2005 (NT dollar)
* PAY\_AMT5: Amount of previous payment in May, 2005 (NT dollar)
* PAY\_AMT6: Amount of previous payment in April, 2005 (NT dollar)
* default.payment.next.month: Default payment (1=yes, 0=no)

## 2.2 ML PROCESS FLOW

This dataset consists of 30000 observations, each representing an individual credit card account holder. The dataset encompasses several key attributes that capture various aspects of the account holders' financial characteristics and behaviors.

# 3. ARCHITECTURE DESCRIPTION

## 3.1 DATA DESCRIPTION

The dataset utilized for this project was sourced from Kaggle's dataset repository (URL: <https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset>). The dataset encompasses a comprehensive range of information, including default payment records, demographic attributes, credit data, historical payment details, and billing statements for credit card holders in Taiwan spanning from April 2005 to September 2005.

## 3.2 DATA PREPROCESSING

The initial phase of our data processing involved the integration of essential libraries such as seaborn, matplotlib, and pandas, enabling us to effectively manipulate and visualize the data. By importing the aforementioned dataset from Kaggle, we established the foundational groundwork for our analysis.

## 3.3 DATA ANALYSIS

In this phase, meticulous attention was dedicated to handling potential null values, refining column nomenclature for clarity, and crafting a suite of visualizations using seaborn, matplotlib, and other visualization libraries. Remarkably, the dataset exhibited an absence of null values, empowering us to delve into comprehensive data visualization and analysis. For each individual feature, we conducted an in-depth analysis through visualization, systematically extracting key insights pivotal for shaping our predictive modeling endeavors.

## 3.4 FEATURE ENGINEERING

To deepen our understanding and insight into the dataset, we undertook strategic feature engineering by amalgamating pertinent columns. This multifaceted approach facilitated the extraction of richer context and more granular information from the data.

## 3.5 TRAIN-TEST SPLIT

Employing the revered Scikit-Learn library, we partitioned the finalized dataset into an 80-20% ratio where 80% of the data constituted our training set, enabling the model to glean essential patterns, and the remaining 20% was preserved for evaluating the model's predictive capabilities.

## 3.6 MODEL SELECTION

Rigorously experimenting with an array of machine learning models including XGBoost, RandomForest, Decision Tree, and ADABoost, we conducted iterative evaluations to identify the model yielding the most robust performance. Ultimately, our scrutiny culminated in the selection of the Random Forest Classifier, lauded for its superior predictive power and accuracy

## 3.7 PREDICTION

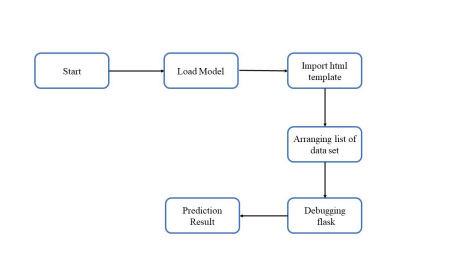
In the wake of extensive model training and evaluation, the Random Forest Classifier demonstrated a remarkable accuracy rate of 81.7%, effectively predicting default payment instances. Simultaneously, the F1 score of 47.3% underscored the model's proficiency in striking a balance between precision and recall.

## 3.8 MODEL PERSISTENCE

Employing the versatile pickle library, we encapsulated the trained Random Forest Classifier model in binary format, thereby enabling streamlined storage and portability.

## 3.9 DEPLOYMENT USING STREAMLIT

With a keen emphasis on delivering a seamless user experience, we harnessed the power of Streamlit - an intuitive Python library for building interactive web applications. This amalgamation facilitated the local deployment of our model, enabling users to interact with the prediction engine through an intuitive web interface.  
Here's an image to offer a visual representation of our deployment



# 4. VISUAL REPRESENTATION

