**Data Exploratory Analysis & Modelling of Credit Default Risk Dataset from UCI**

**By:** Helen Tan Meng Zhen

Contents

[**1.** **Introduction** 3](#_Toc80351190)

[1.1 Summary 3](#_Toc80351191)

[1.2 Problem Statement & Objective 3](#_Toc80351192)

[**2.** **Dataset** 4](#_Toc80351193)

[2.1 Data Description 4](#_Toc80351194)

[2.2 Data Cleaning 4](#_Toc80351195)

[**3.** **Data Visualization** 6](#_Toc80351196)

[3.1 Age vs Default 7](#_Toc80351197)

[3.2 Loan Amount vs Default 9](#_Toc80351198)

[3.3 Income Range vs Default 10](#_Toc80351199)

[**4.** **Data Modelling** 12](#_Toc80351200)

[4.1 Logistic Regression 13](#_Toc80351201)

[4.2 Decision Tree – CART 14](#_Toc80351202)

[4.3 Random Forest 15](#_Toc80351203)

[4.4 XGBoost 15](#_Toc80351204)

[4.5 Neural Network 17](#_Toc80351205)

[4.6 Model Comparison 18](#_Toc80351206)

[**5.** **Business Solution & Recommendation** 18](#_Toc80351207)

[**6.** **Conclusion & Future Study** 19](#_Toc80351208)

# **Introduction**

## **1.1 Summary**

In this report, a simple credit risk default dataset is analyzed to explore the relationship between the occurrence of a credit default (target variable) and the various predictor variables, with the aim to predict whether a loan will be repaid. Subsequently, the machine learning models are trained and tested, after which their resulting accuracies are compared against each other. Finally, this report will give a brief recommendation on possible business solutions to be considered henceforth.

## **1.2 Problem Statement & Objective**

The major risks faced by banks include credit, operational, market and liquidity risk.

Credit risk refers to the risk of loss resulting from the failure of a client of counterparty to meet its contractual obligations toward the bank. In other words, the risk of a default or a non-payment by the client/borrower. This is a major risk to banks, as banks generate revenue mainly from interest on loans and a high credit risk would in turn negatively impact business. Based on the annual report in 2019 for UBS, a prominent bank based in Switzerland, credit risk has been flagged as one of the primary risks that the bank faces.

Furthermore, in wake of the global financial crisis comes stringent regulatory requirements that demands that banks have thorough knowledge of their clients and their associated credit risk and have transparent, auditable risk management frameworks. Banks that are vulnerable to credit risk are targets for hefty fines by regulators, and risk reputational loss in the eyes of investors. Beyond being a mere compliance exercise, better credit risk management can also potentially be an opportunity for banks to improve overall performance. Hence, it is imperative that banks place great focus on managing their credit risk.

Among the possible methods for credit risk management is for banks to develop reliable Artificial Intelligence (AI) models. With the skyrocketing growth of machine learning model development by start-ups and fintechs, the application of AI to predict the probability of a default has been a hot topic in the banking industry. Touted to accurately predict potential credit defaults and serve as an early warning system, it is important for banks to keep abreast of technology developments in the industry, and develop reliable models in order to improve decision-making efficiency and make better business decisions.

Based on an online credit risk dataset, this report aims to analyze the relationship between the occurrence of a credit default and the available variables, and **predict whether a loan will be repaid**.

# **Dataset**

## **2.1 Data Description**

The dataset used in this report contains credit card information of clients of an unknown bank in Taiwan, a public online dataset from the UCI Machine Learning Repository made available on [Kaggle](https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset). It provides information on default payments, demographic variables, bill statements and payment history, in the period from April 2005 to September 2005.

This dataset contains a total of 30,000 records with each record corresponding to a single credit card client. It has a total of 25 attributes, with 23 predictive variables, 1 non-predictive variable (ID), and 1 target variable (default.payment.next.month). A tabular summary of the data is shown below:

|  |  |  |  |
| --- | --- | --- | --- |
| Total Number of Records: 30,000,  Number of Attributes: 25 (23 predictive - green, 1 non-predictive - blue, 1 target - red) | | | |
| Column Name | Description | Column Name | Description |
| ID | ID of each client | BILL\_AMT1 | Amount of bill statement in September 2005 (in NT dollar) |
| LIMIT\_BAL | Amount of given credit in NT dollars | BILL\_AMT2 | Amount of bill statement in July 2005 (in NT dollar) |
| SEX | Gender (1=male, 2=female) | BILL\_AMT3 | Amount of bill statement in June 2005 (in NT dollar) |
| EDUCATION | Education level (1=Graduate school, 2=University, 3=others, 4=others, 5=unknown, 6=unknown) | BILL\_AMT4 | Amount of bill statement in May 2005 (in NT dollar) |
| MARRIAGE | Marital status (1=married, 2=single, 3=others(changed to divorced during data cleaning)) | BILL\_AMT5 | Amount of bill statement in April 2005 (in NT dollar) |
| AGE | Age in years | BILL\_AMT6 | Amount of bill statement in September 2005 (in NT dollar) |
| PAY\_0  (Changed to PAY\_1 during data cleaning) | Repayment status in September 2005 (-1=pay duly, 1=payment delay for 1 month, 2=payment delay for 2 months, … 8=payment delay for 8 months, 9=payment delay for 9 months and above) | PAY\_AMT1 | Amount of previous payment in September 2005 (in NT dollar) |
| PAY\_2 | Repayment status in August 2005 (same scale as above) | PAY\_AMT2 | Amount of previous payment in August 2005 (in NT dollar) |
| PAY\_3 | Repayment status in July 2005 (same scale as above) | PAY\_AMT3 | Amount of previous payment in July 2005 (in NT dollar) |
| PAY\_4 | Repayment status in June 2005 (same scale as above) | PAY\_AMT4 | Amount of previous payment in June 2005 (in NT dollar) |
| PAY\_5 | Repayment status in May 2005 (same scale as above) | PAY\_AMT5 | Amount of previous payment in May 2005 (in NT dollar) |
| PAY\_6 | Repayment status in April 2005 (same scale as above) | PAY\_AMT6 | Amount of previous payment in April 2005 (in NT dollar) |
|  | | default.payment.next.month | Did the client default payment? (1=Yes, 0=No) |

Table 1: Summary of variables in the dataset and their meanings (Source: [Kaggle](https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset))

In a real-world scenario, understanding the risk profiles of a potential client would no doubt require much more than these variables. Hence, working with a small dataset such as this would enable us to get an initial understanding of how predicting credit risk can be done, and similar methods may be employed when dealing with larger and more complicated datasets.

## **2.2 Data Cleaning**

xxx

# **Data Visualization**

xxx

## **3.1 Age vs Default**

xxx

## **3.2 Loan Amount vs Default**

Xxx

## **3.3 Income Range vs Default**

xxx

# **Data Modelling**

## **4.1 Logistic Regression**

xxx

## **4.2 Decision Tree – CART**

xxx

## **4.3 Random Forest**

Xxx

## **4.4 XGBoost**

Xxx

## **4.5 Neural Network**

xxx

## **4.6 Model Comparison**

Xxx

# **Business Solution & Recommendation**

# **Conclusion & Future Study**

**Readings:**

<https://www2.deloitte.com/cn/en/pages/risk/articles/artificial-intelligence-for-credit-risk-management.html>

<https://www.sas.com/en_sg/insights/risk-management/credit-risk-management.html>

<https://home.kpmg/xx/en/home/insights/2020/07/a-new-risk-management-playbook-for-banks.html>