**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**

To first obtain an initial understanding of the data, the pandas info() method is used to obtain the information on the data types available and the total count of non-null values.

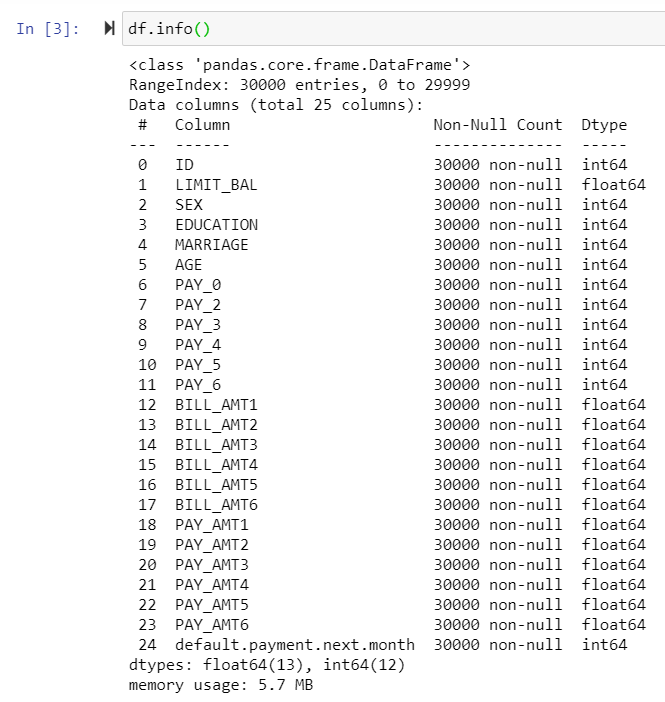


Figure 1: Information on Non-null count and Data type

From the results, we observe that there are no null values in the dataset as the total non-null count values of all the variables are maintained at 30,000. We see that SEX, EDUCATION, MARRIAGE, AGE, PAY\_0, PAY\_1, PAY\_2, PAY\_3, PAY\_4, PAY\_5, PAY\_6 are of type int64, which suggest that they are categorical variables, matching our previous understanding of the data.

We also observe that one of the columns about the repayment status is named “PAY\_0”, with the next column named “PAY\_2”. To avoid any future confusion and to facilitate better ease of understanding the data, the column **“PAY\_0” is renamed to  “PAY\_1”**.

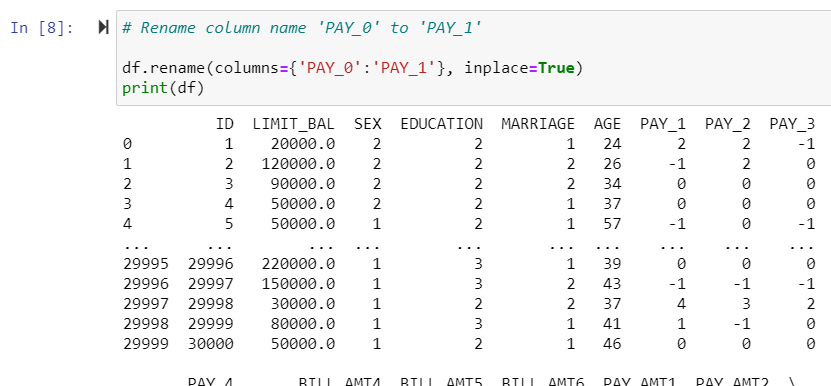
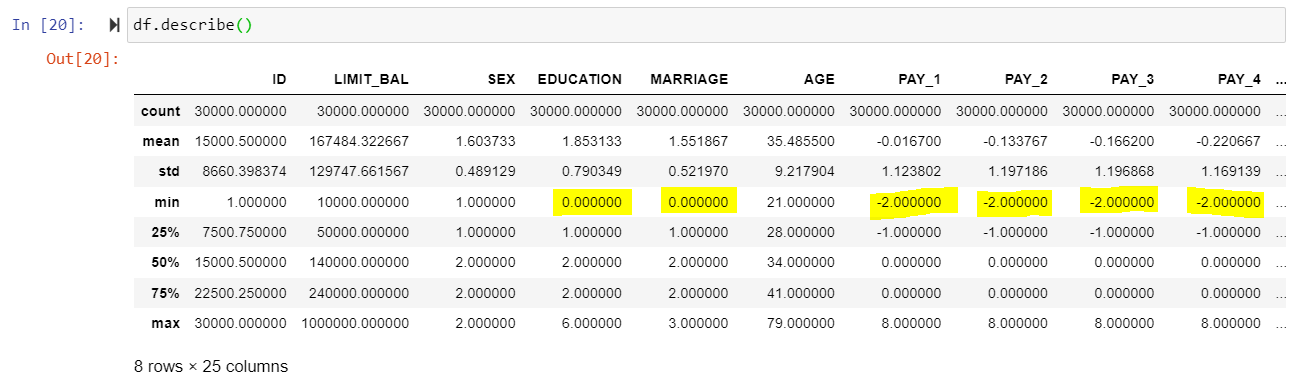


Figure 2: Renaming “PAY\_0” to “PAY\_1”

Subsequently, the pandas describe() method is used to obtain a statistical summary of the variables.



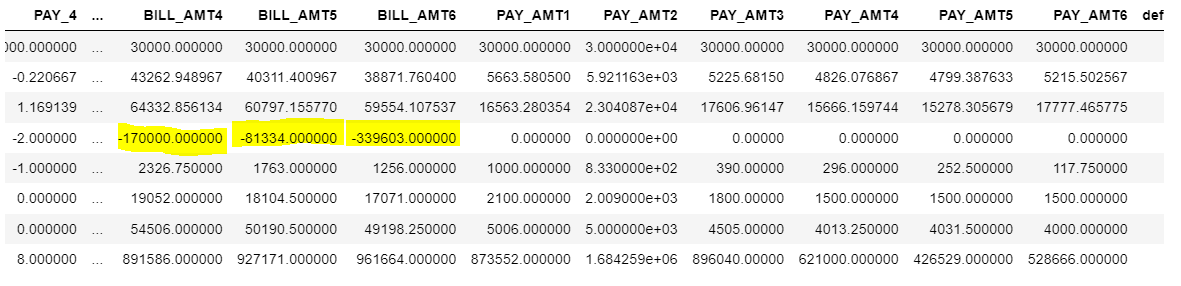


Figure 3: Statistical summary of the dataset, with some odd min values

The results showed that some minimum values were unusual as they did not match the descriptions from the original dataset in Kaggle. The oddities of the min values are summarized below:

1. **Education = 0**. The categories provided by Kaggle are: 1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown. The zero value could have been used to replace missing/null values. Based on a [source](https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset/discussion/34608) that has reached out to the dataset creator for comment, it is confirmed that 0, 5, 6 can be grouped under ‘unknown’.
2. **Marriage = 0**. The categories provided by Kaggle are: 1=married, 2=single, 3=others. Similarly, The zero value could have been used to replace missing/null values. Based on a [source](https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset/discussion/34608) that has reached out to the dataset creator for comment, it is confirmed that 3=divorced and 0=others. No changes need to be done to the dataset with Python.
3. **PAY\_X variables = -2**. The categories provided by Kaggle are: -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. We do not have a good guess for the real meaning of the -2  value, but it could be suggesting some form of loan repayment, similar to the -1 value. Based on a [source](https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset/discussion/34608) that has reached out to the dataset creator for comment, -2=‘No consumption’ and 0=use of revolving credit. We choose to group -2, -1, and 0 into one group under “Paid on Time”.
4. **Negative values of BILL\_X**. This could simply mean that the customers paid more than their actual loan amounts. However, there is no way to accurately guess this unless explained by the creator of the dataset.

For the inconsistency in the ‘EDUCATION’ column, we made the choice of grouping 0, 5 and 6 into one category called ‘unknown’. To do this, we replace the values 0 and 6 to 5. Values that are originally 5 are untouched.

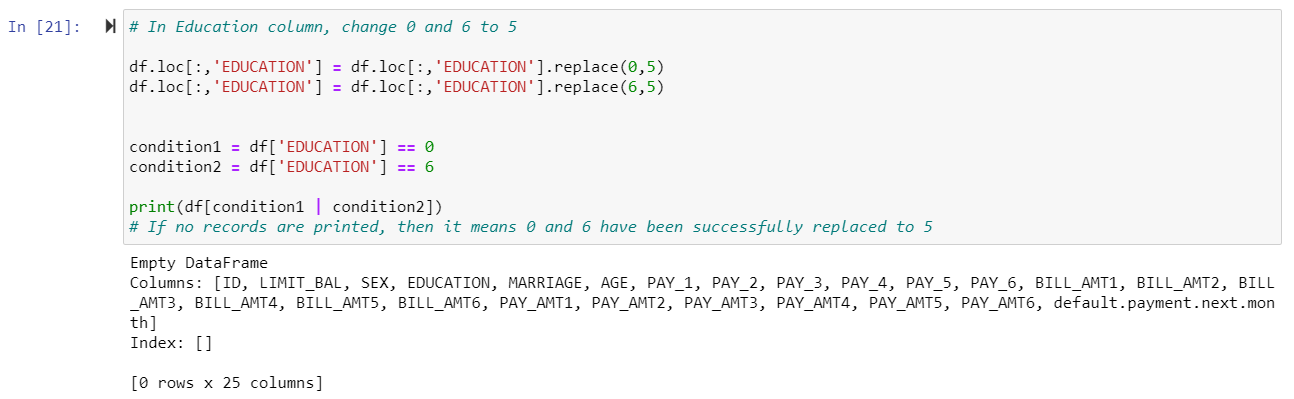


Figure 4: Replacing 0 & 6 to 5 in the “EDUCATION” column

After replacing the values, the display() method is used again. This time around it shows that the minimum value of the ‘EDUCATION’ column is now 1 and not 0, meaning that the 0 and 6 values have been successfully changed to 5. The categories for the ‘EDUCATION’ column are as follows: 1=graduate school, 2=university, 3=high school, 4=others, 5=unknown.

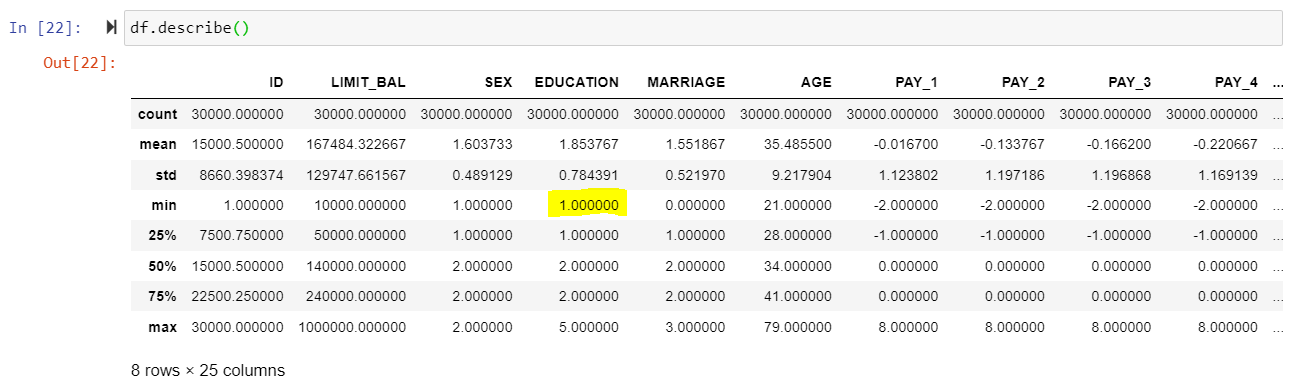


Figure 5: min of ‘EDUCATION’ no longer shows zero

To summarize, these are the additional information on the categories:

1. EDUCATION: (0, 5, 6 = unknown)
2. MARRIAGE: (3 = divorce, 0 = others)
3. PAY\_AMTX: (-2, -1, 0 = Paid on Time)

# **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>