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

**By:** Helen Tan Meng Zhen

Contents

[**1.** **Introduction** 3](#_Toc80978432)

[1.1 Summary 3](#_Toc80978433)

[1.2 Problem Statement & Objective 3](#_Toc80978434)

[**2.** **Dataset** 4](#_Toc80978435)

[2.1 Data Description 4](#_Toc80978436)

[2.2 Data Cleaning 5](#_Toc80978437)

[**3.** **Data Visualization** 8](#_Toc80978438)

[3.1 Correlation Between Variables 9](#_Toc80978439)

[3.2 Amount of Credit Given vs Default Status 10](#_Toc80978440)

[3.3 Gender vs Default Status 12](#_Toc80978441)

[3.3 Education Level vs Default Status 13](#_Toc80978442)

[3.4 Age vs Default Status 15](#_Toc80978443)

[3.5 Marital Status vs Default Status 16](#_Toc80978444)

[**5.** **Data Modelling** 17](#_Toc80978445)

[4.1 Logistic Regression 17](#_Toc80978446)

[4.2 Decision Tree – CART 17](#_Toc80978447)

[4.3 Random Forest 18](#_Toc80978448)

[4.4 XGBoost 18](#_Toc80978449)

[4.5 Neural Network 18](#_Toc80978450)

[4.6 Model Comparison 18](#_Toc80978451)

[**6.** **Business Solution & Recommendation** 18](#_Toc80978452)

[**7.** **Conclusion & Future Study** 18](#_Toc80978453)

# **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.

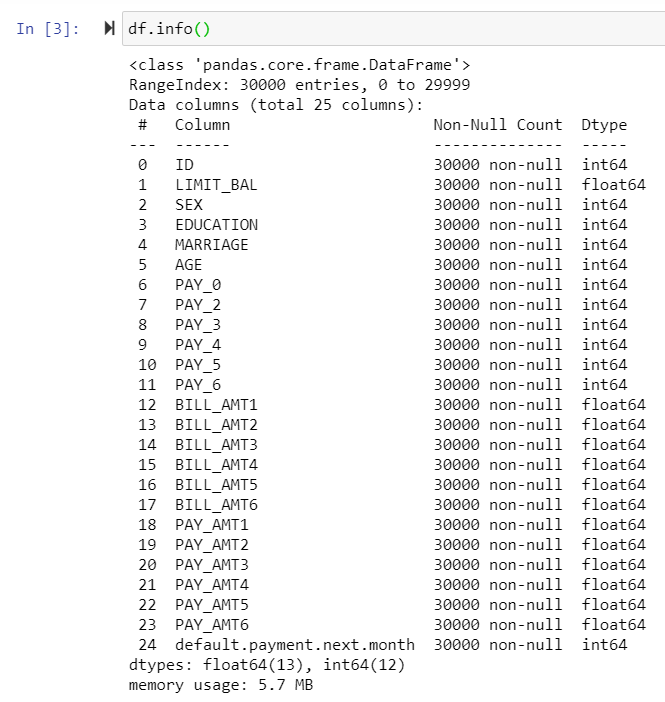


Figure 1: Information on Non-null count and Data type

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”**.

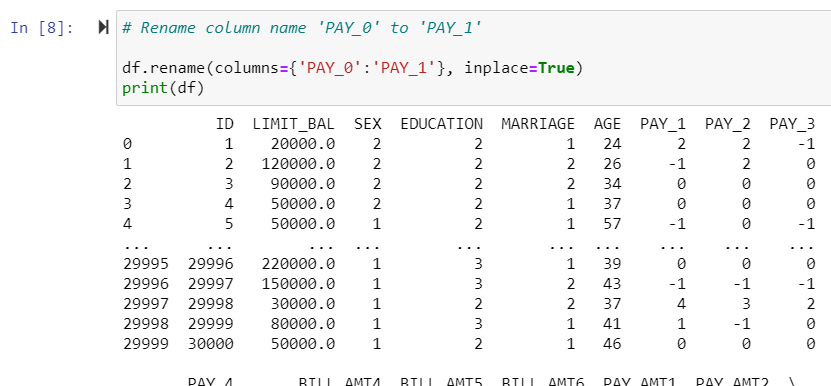
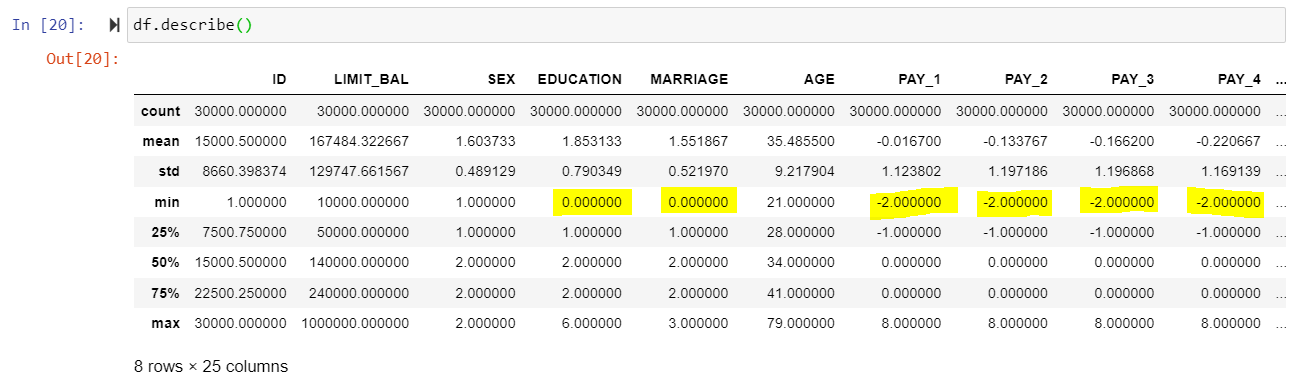


Figure 2: Renaming “PAY\_0” to “PAY\_1”

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



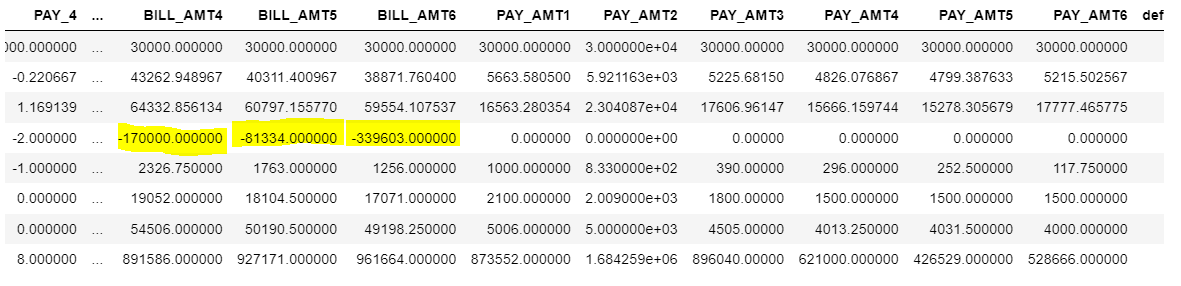


Figure 3: Statistical summary of the dataset, with some odd min values

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.

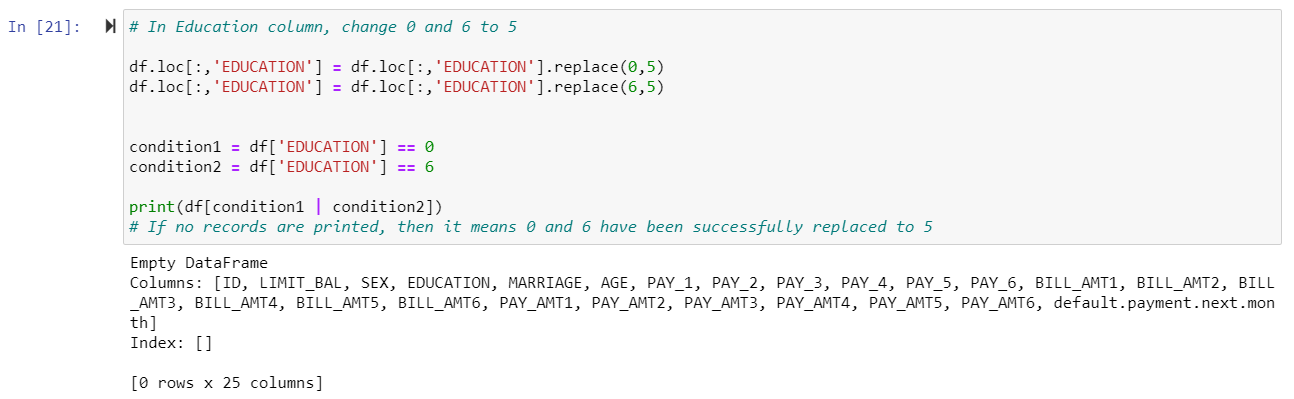


Figure 4: Replacing 0 & 6 to 5 in the “EDUCATION” column

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.

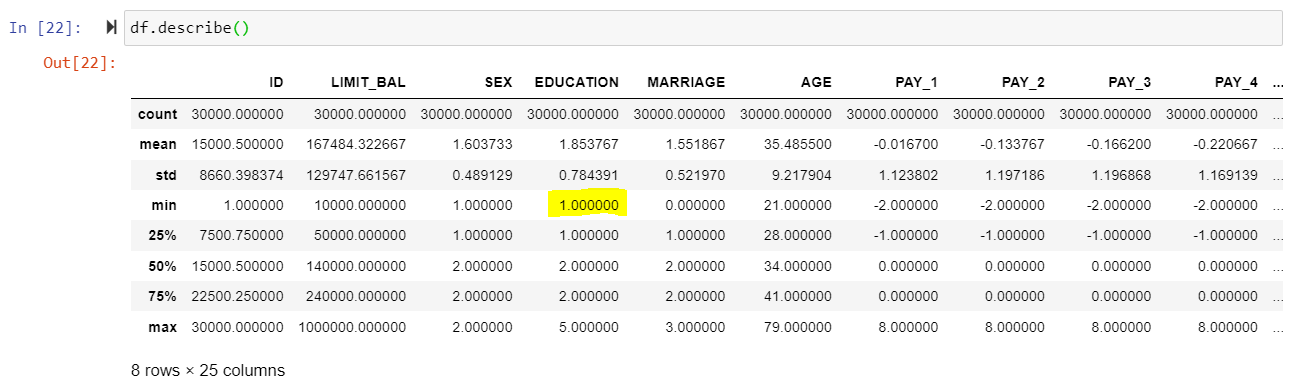


Figure 5: min of ‘EDUCATION’ no longer shows zero

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**

In this section, data exploratory analysis is performed to investigate the relationship between the target variable and the various predictor variables to determine which ones would be a good predictor for whether a person will default payment in the next month or not.

Firstly, we plot the distribution of the target variable which is credit default. From the barplot below, we observe that the majority of the clients in this dataset repaid their loans on time as the number of non-default records vastly out-number those with defaults. Out of the total 30,000 records, 23364 clients repaid their loans while the remaining 6636 clients defaulted.

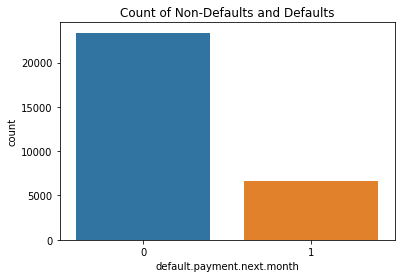
1. 

Figure 12: Count of non-defaults (default.payment.next.month = 0) and defaults (default.payment.next.month = 1)

The prediction of the occurrence of a credit default in the next month is most likely to depend demographic variables such as age, gender, education and marital status, as well as the repayment history of the client which can be gleaned from the variables like the limiting balance, credit repayment status from the months of April to September. Logically, the ID variable would not be a predictor variable as it is just an identifier number attached to each person. The upcoming sections will explore their relationship with the target variable in more detail.

## **3.1 Correlation Between Variables**

The following heat map depicts the correlation between all the variables in the dataset. From the heatmap, we observe that none of the predictor variables are strongly correlated with the target variable (default.payment.next.month), as majority of the variables have a correlation coefficient value of less than 0.01 with the target variable, with the highest value only being 0.32 from PAY\_1. The six ‘PAY\_’ variables are observed to have a weak positive correlation with the target variable.

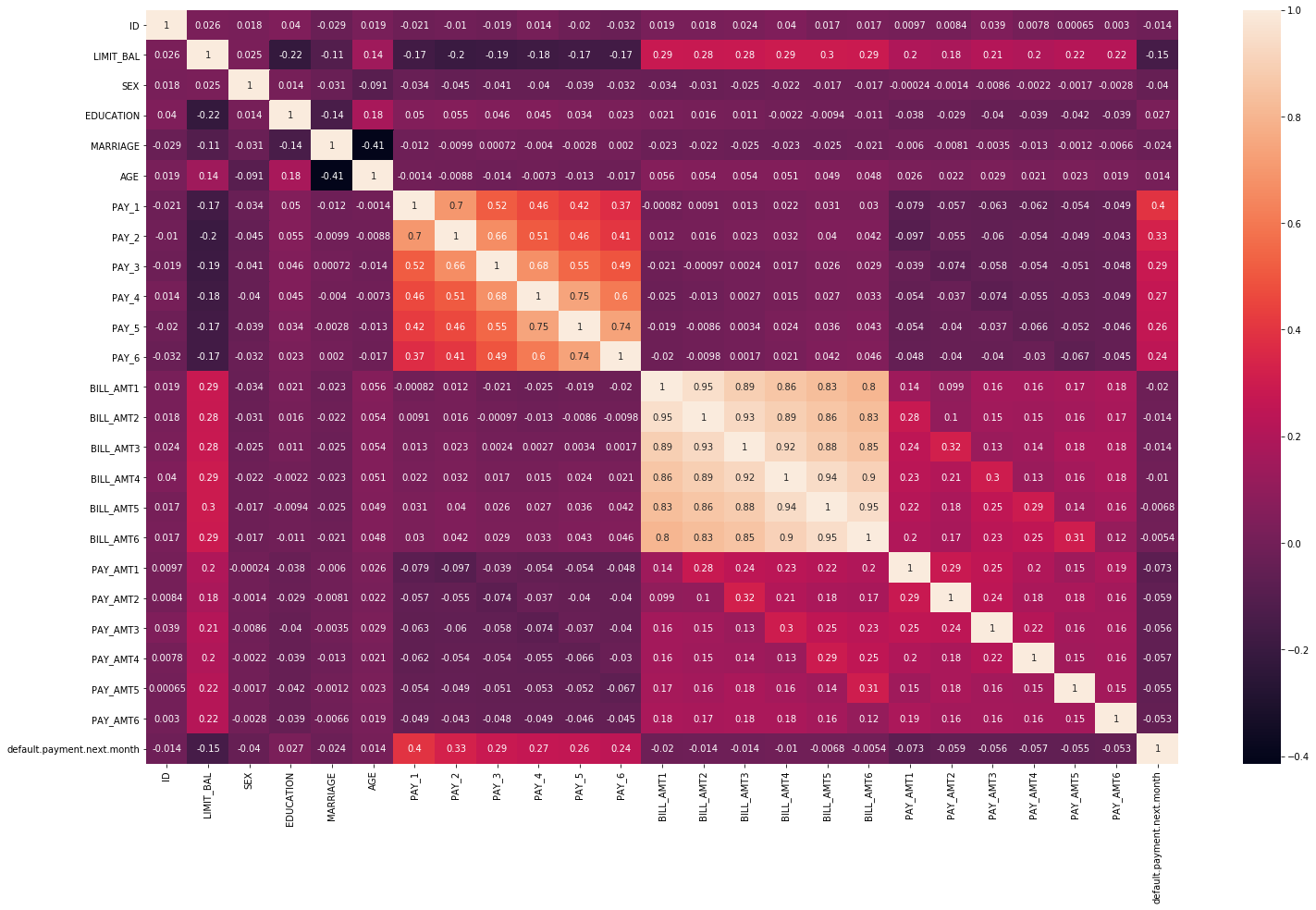


Figure 13: Heatmap depicting the correlation between the variables

Another obvious observation is that all the “BILL\_AMT” variables have a high positive correlation between each other, as seen from the lightest area in the middle of the heatmap, with correlation coefficient values of between 0.80 - 0.95.

Similarly, all “PAY\_” variables have a comparatively good positive correlation between each other as well, with the correlation coefficient values between the range of 0.47 - 0.82. For PAY\_1, it has the highest correlation with PAY\_2 (0.67). For PAY\_2, it has the highest correlation with PAY\_3 (0.77). For PAY\_3, it has the highest correlation with PAY\_4 (0.78), and so on. This implies that the **current repayment status of the current month is most highly dependent on the repayment status of the previous month.**

## **3.2 Amount of Credit Given vs Default Status**

This section explores the relationship between the amount of credit limit given to a client (LIMIT\_BAL) and the client’s credit default status, to check if credit limit amount is a good predictor for an occurrence of credit default. We first generate a statistical summary of the “LIMIT\_BAL” column with the pandas describe() method.

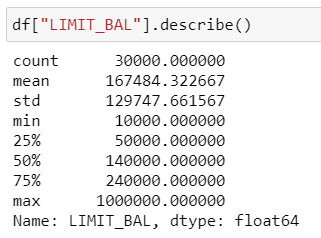
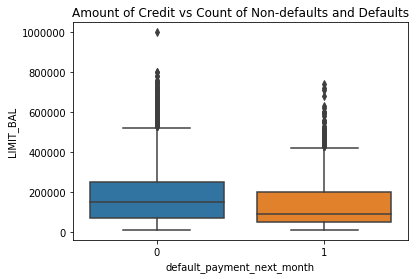


Figure 14: statistical summary of the “LIMIT\_BAL” column

From the statistical summary, we can understand that most clients are given a credit limit value of between NT$ 50,000 - NT$ 240,000, as indicated by the 25 and 75 percentile of the given credit limit values. On average, clients in this dataset are given a credit amount of NT$ 167,484.

A boxplot of the given credit limit amount is plotted against the default status. The credit limit range of clients who did not default and clients who defaulted are shown to be roughly similar. Focusing on the boxplot of clients who defaulted (orange boxplot), we see that the majority of clients who defaulted have taken a credit limit of between NT$ 50,000 - NT$200,000.



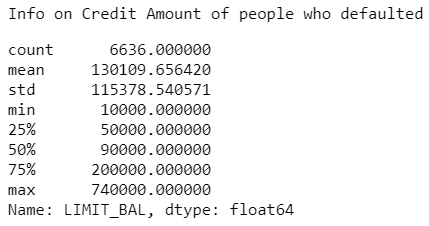
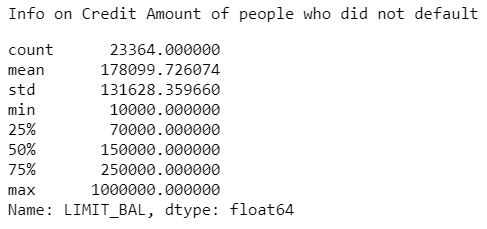


Figure 15: Boxplot of credit limit against default payment status

In an attempt for a closer examination of LIMIT\_BAL, we split the values of LIMIT\_BAL into bins. A total of 8 bins were created with a bin interval of (1,000,000 - 1,000)/8 = 125,000. A barplot of the number of clients is plotted against the amount of credit limit for each bin.

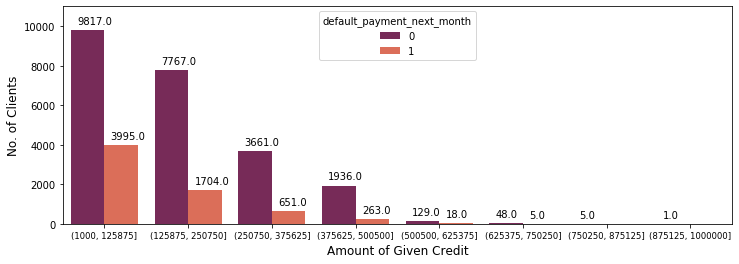


Figure 16: No. of clients who did not default and who defaulted, grouped by amount of credit given

Subsequently, the percentage of default for each group of credit limit amount is plotted.

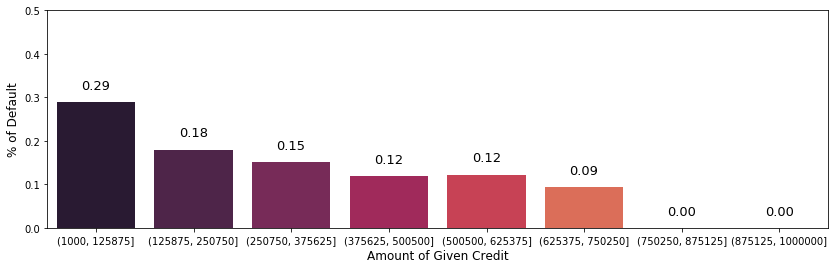


Figure 17: Percentage default by each group of given credit limit

From the barplot, it is clear that the highest chance of default (29%) is from the clients who have a credit limit amount from NT$ 1000 - NT$ 125875. We also observe a trend where the percentage of default decreases as the credit limit amount increases. **This suggests that the higher the amount of credit limit, the lower the probability of defaulting.**

## **3.3 Gender vs Default Status**

This section explores the relationship of a client’s gender and their default status to investigate if gender might be a good predictor for credit default. We first look at the distribution of males and females in the dataset by plotting the count of each gender. We observe that there are more females than males in the dataset (males = 11888, females = 18112).

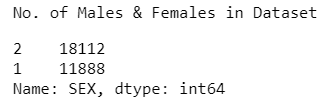
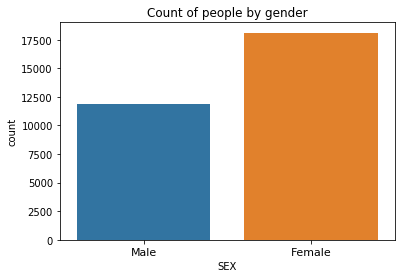


Figure 18: Number of males & females in the dataset

A barplot of the number of clients who did not default and who defaulted are plotted for both males and females. We observe that for both genders, there are significantly more people who paid their loans on time compared to people who did not.

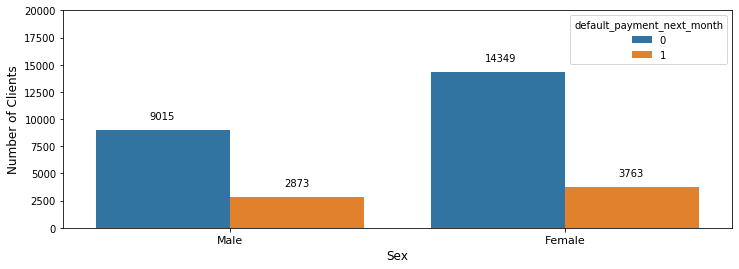


Figure 19: No. of people who did not default (blue) & who defaulted (orange), grouped by gender

The proportion of men who defaulted is 2873 / (2873+9015) = 0.242, while the proportion of women who defaulted is 3763 / (3763+14349) = 0.208. In other words, in this particular dataset, 24% of men defaulted while. 21% of women defaulted. **This suggests that men might have a higher chance of credit default than women.**

## **3.3 Education Level vs Default Status**

This section explores the relationship between Education level and Default status. We first plot a barplot of the number of clients by their education level. We see that the majority of the clients’ education level is at the university level (14030), followed by graduate school (10585), high school (4917), unknown (345), followed by others (123).

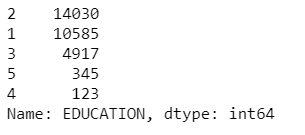
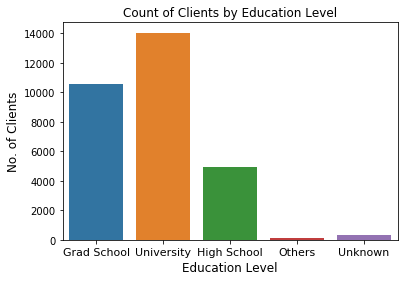


Figure 20: Total no. of Clients by Education level (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown.)

We examine the number of clients who did not default and who defaulted for each education level. We observe that across all education levels, clients who did not default outnumber those who defaulted.

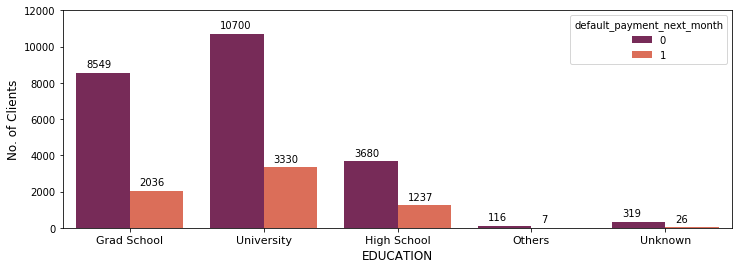


Figure 21: No. of clients who did not default and who defaulted, grouped by education level

Subsequently, we derive the percentage of default from each group and present it in the barplot below.

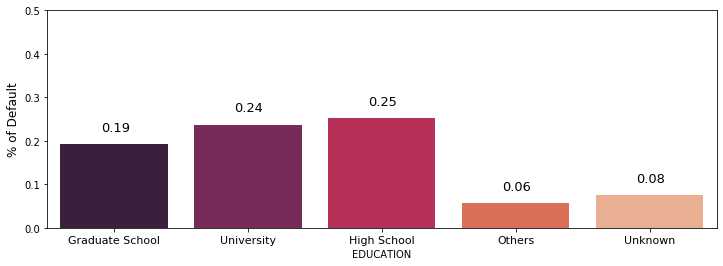


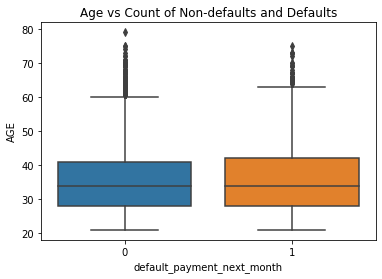
Figure 22: Percentage default by each Education level group

The results show that the education level with the highest chance of credit default is at the High School level, with 25% of clients with high school level education defaulting the most. Considering only the 3 most prominent education levels which is Graduate School, University and High School, the chance of default decreases from high school (25%) to university (24%) then to graduate school (19%). **This demonstrates that clients with a higher level of education would have a lower chance of default.**

Such a result is in line with our understanding as well. as we understand that people with a higher education qualification would generally be more adept at handling their finances. If possible, it would be ideal if future analyses can clarify the meanings of “others” and “unknown” with the dataset creator, to see if they are education levels above or below the high school level, to further cement the conclusion that higher education translates to lower chance of default.

## **3.4 Age vs Default Status**

This section explores the relationship between Age and Default status. A boxplot of the clients’ age is plotted against the default status. We see that generally the age range of clients who did not default is similar to that of those who defaulted, with the majority in the age range of 28 years old - 41 years old.



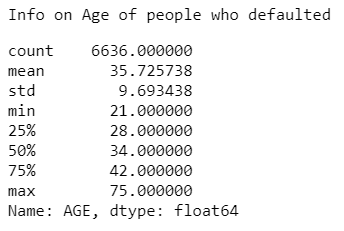
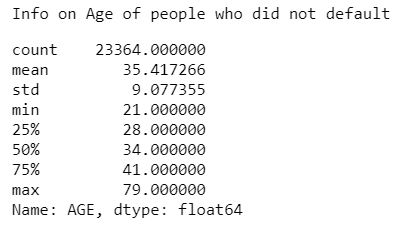


Figure 23: Boxplot of age against default payment status

## **3.5 Marital Status vs Default Status**

We focus on the relationship between clients’ marital status and their default status in this section. Firstly, we plot the total number of clients for each marital status, and see that the majority of the clients are single (15,964), followed by married clients (13,659), then divorced clients (323) and finally clients with other marital status (54).

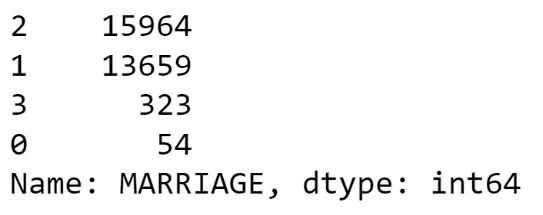
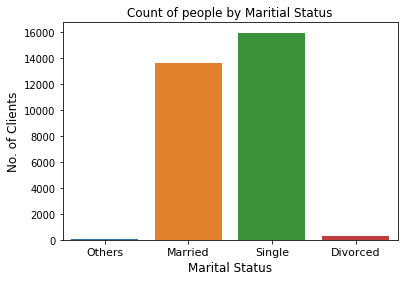
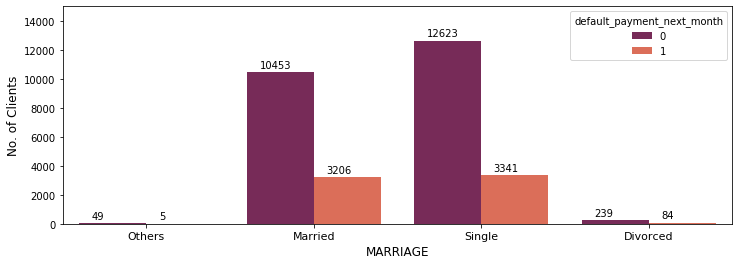


Figure 24: Total no. of Clients by Marital Status (0=Others, 1=Married, 2=Single, 3=Divorced.)

Subsequently, we examine the number of clients’ defaults for each marriage category. Across all categories, most clients did not default.

Figure 25: No. of clients who did not default and who defaulted, grouped by marital status

We derive the percentage default for each marriage category and present it in the bar plot below.

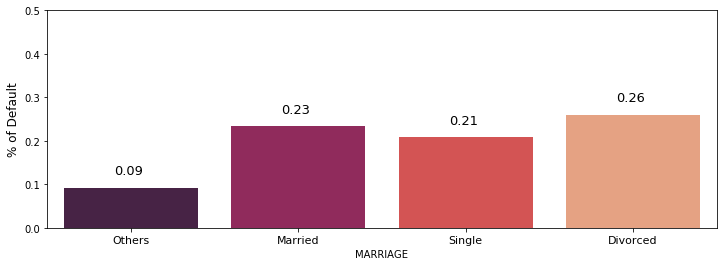


Figure 26: Percentage default by clients from each marriage category

From the results, it shows that clients who are divorced have the highest percentage of default (26%). Considering only the known groups, this is followed by married clients (23%) and the group with the lowest chance of defaulting are the single clients (21%). **This shows that divorced clients may be more susceptible to credit default, while single clients are most likely to pay their loans on time.**

# **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>