Cover

TO BE DECIDED​

Wu Yifei, 58243516​

Wu Shengjun, 58065781​

Wong Tsz Yiu, 58218046​

Guanze Chen, 58155621​

Guan Yanying, 58000752​

YU Xinyu, 58188431​

ZHAO Yisong, 57965957

IS5740, Semester A 2023/2023​

Predicting Credit Card Default in Taiwan​

Table of Contents

[1. Introduction 4](#_Toc151894633)

[1.1. Background and Importance of Predicting Credit Card Defaults 4](#_Toc151894634)

[1.2. Objectives of the Report 4](#_Toc151894635)

[2. Data Description and Dataset Overview 4](#_Toc151894636)

[3. Data Exploration & Analysis 6](#_Toc151894637)

[3.1. Target/Dependent Variable – Client’s Default Status in October 2005 (TARGET) 6](#_Toc151894638)

[3.2. Independent Variable – Categorical Features 1 6](#_Toc151894639)

[3.2.1. Categorical Features 1 – Gender (SEX) 6](#_Toc151894640)

[3.2.2. Categorical Features 1 – Marital Status (MARRIAGE) 7](#_Toc151894641)

[3.2.3. Categorical Features 1 – Education Level (EDUCATION) 8](#_Toc151894642)

[3.3. Independent Variable – Numerical Features 1 9](#_Toc151894643)

[3.3.1. Numerical Feature 1 – AGE 9](#_Toc151894644)

[3.3.2. Numerical Feature 1 – CREDIT\_LIMIT 10](#_Toc151894645)

[3.4. Independent Variable - Categorical Feature 2 - Repayment Status (PAY\_XXX) 12](#_Toc151894646)

[3.5. Feature Correlation 14](#_Toc151894647)

[3.5.1. Independent Variable – Numerical Feature 2 15](#_Toc151894648)

[3.5.1.1. Numerical Feature 2 – Bill Statement Amount (BILL\_AMT\_XXX) 15](#_Toc151894649)

[3.5.1.2. Numerical Feature 2 – Previous Payment (PAY\_AMT\_XXX) 15](#_Toc151894650)

[4. Problem of Interest Identified from Preliminary Analysis 15](#_Toc151894651)

[4.1. First Problem - Developing an Accurate Predictive Model 16](#_Toc151894652)

[4.2. Second Problem - Assessing the Impact of Various Attributes 16](#_Toc151894653)

[5. Data Pre-processing 16](#_Toc151894654)

[5.1. Treatment of Missing Values and Data Cleaning 16](#_Toc151894655)

[5.2. One-Hot Encoding for Categorical Feature 16](#_Toc151894656)

[5.3. Splitting the Dataset: Training & Validation Sets 16](#_Toc151894657)

[6. Model Building and Analytics 16](#_Toc151894658)

[6.1. Logistic Regression 16](#_Toc151894659)

[6.1.1. Full Logistic Regression 16](#_Toc151894660)

[6.1.2. Feature Selection - Forward & Backward Logistic Regression 17](#_Toc151894661)

[6.2. Evaluation of Model Performance 18](#_Toc151894662)

[6.2.1. Interpretation of Performance Metrics 18](#_Toc151894663)

[6.2.2. Model Comparison – Logistic Regression Models 18](#_Toc151894664)

[6.3. Full Logistic Regression Analysis 19](#_Toc151894665)

[6.3.1. Statistical Significance of Variables 19](#_Toc151894666)

[6.3.2. Feature Importance of Variables 21](#_Toc151894667)

[6.3.3. Maximum Likelihood Estimates 21](#_Toc151894668)

# **1. Introduction**

## **1.1. Background and Importance of Predicting Credit Card Defaults**

This project delves into the realm of credit card default predictions. The ability to anticipate such defaults is of paramount importance in the banking sector, given the substantial impact on financial stability. Our methodology aims to develop a predictive model that can accurately identify potential defaulters based on clients’ demographics data and historical credit card usage information. An accurate prediction model can assist the banking institutes in achieving an effective risk management and making informed lending decisions

## **1.2. Objectives of the Report**

The primary objective of this report is twofold: to develop a predictive model that can accurately identify potential defaulters, and to underscore the model's strategic value in effective risk management and lending decisions.

**Effective Risk Management & Informed Lending Decisions:**

Predicting credit card defaults is not merely about detecting risk—it's about financial stability. Since defaulting clients can result in significant financial losses, due to the unpaid debts and administrative costs associated with collections. If the banks are able to identify high-risk clients, they can reduce these potential cost by

* **Implement Early Intervention Strategies**: Proactive engagement through timely reminders and financial advisory services can help clients maintain fiscal discipline.
* **Take Proactive Measures**: For clients identified as high-risk, banks can mitigate exposure to non-payment by adjusting credit limits, revising lending terms, or, in certain cases, temporarily suspending credit facilities. These measures collectively contribute to reducing the risk of default, mitigate the impact of defaults as well as lowering the bank's exposure to financial losses.

Conversely, for clients with a low-risk profile, banks can extend more favourable commercial terms and superior customer service, thereby fostering loyalty and potentially increasing profitability. In essence, the predictive model serves as a pivotal tool for financial institutions to balance opportunity with risk, ensuring sustainable growth and customer satisfaction.

# **2. Data Description and Dataset Overview**

Our dataset is about Client defaults in the credit card domain, it contains the demographic information, and credit card usage patterns of Taiwanese clients from April to September in 2005. The target variable denotes whether they were defaulted on their credit card payments in the subsequent month (October 2005). There are in total 30,000 observations and 24 attributes.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Attributes Name | | Role | Level | Description |
|  |  | |  |  |
| Target/Dependent Variables | | | | |
| TARGET | | **Target** | **Binary** | **Default Status in October 2005 (1=yes, 0=no)** |
| Independent Variables 1 – Demographics | | | | |
| SEX | | Input | Nominal | Gender (1=male, 2=female) |
| MARRIAGE | | Input | Nominal | Marital status (1=married, 2=single, 3=others) |
| EDUCATION | | Input | Nominal | Education Level (1=graduate school, 2=university, 3=high school, 4=others) |
| AGE | | Input | Interval | Age in years |
| Independent Variables 2 - Credit Card Usage | | | | |
| CREDIT\_LIMIT | | Input | Interval | Amount of credit limit in NT dollars |

**Repayment status:**

-2: **No credit-card consumption**; the account had a zero balance and no transactions.

-1: **Paid in full**; the account had a balance that was paid in full.

0: the **minimum payment was made**, but the full balance wasn't paid off.

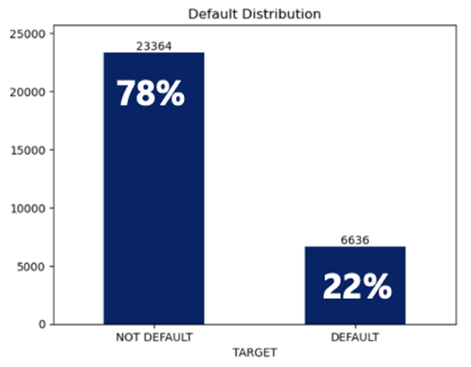
1: **Payment delay** for 1 month; and for 2, 3, etc., indicating the number of months delayed.

|  |  |  |  |
| --- | --- | --- | --- |
| PAY\_XXX  (SEP to APR) | Input | Nominal | **Repayment status** in September to April 2005, with categories ranging from timely repayment to delays of several months. |
| BILL\_AMT\_XXX (SEP to APR) | Input | Interval | Amount of bill statement in September to April 2005 (NT dollar) |
| PAY\_AMT\_XXX (SEP to APR) | Input | Interval | Amount of previous payment in September to April 2005 (NT dollar) |

# 3. Data Exploration & Analysis

## 3.1. Target/Dependent Variable – Client’s Default Status in October 2005 (TARGET)

|  |  |  |
| --- | --- | --- |
| Target | Number | Percentage |
| Default | 6,636 | **22.12%** |
| Not Default | 23,364 | **77.88%** |



There are 30,000 observations in the dataset, the Target variable indicates whether a client defaulted or not in October 2005. The distribution of the Target variable is as follows:

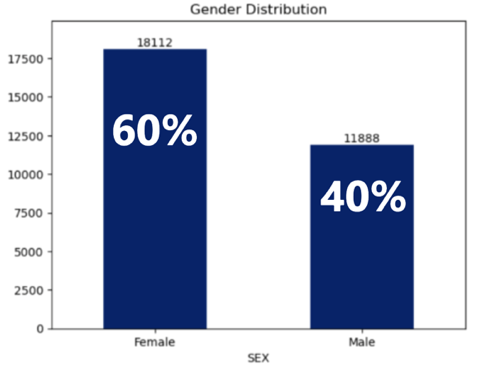
* **Default**: There are 6,636 clients classified as Default, which represents approximately 22.12 % of the total observations, who have failed to meet the credit agreement terms.
* **Not Default**: Conversely, there are 23,364 clients identified as Not Default, constituting approximately 77.88% of the observations. These clients adhere to the credit terms successfully.

The histogram illustrates this distribution, emphasizing the disparity between the two categories. The **average default rate in the dataset is around 22%,** highlighting the incidence of credit default.

## 3.2. Independent Variable – Categorical Features 1

### 3.2.1. Categorical Features 1 – Gender (SEX)

|  |  |  |
| --- | --- | --- |
| SEX | Number | Percentage |
| Male (1) | 11,888 | **39.63%** |
| Female (2) | 18,112 | **60.37%** |

****

The histogram illustrate the Gender distribution, there are more female than male in the dataset, with females making up around 60% compared to males around 40% in the observations.

**Pivot Table – Gender (SEX) against TARGET variable**

|  |  |  |  |
| --- | --- | --- | --- |
| TARGET | Default | Not Default | Default Percentage |
| 1. SEX | | | |
| 1.1. Male | 2,873 | 9,015 | **24.17%** |
| 1.2. Female | 3,763 | 14,349 | **20.78%** |

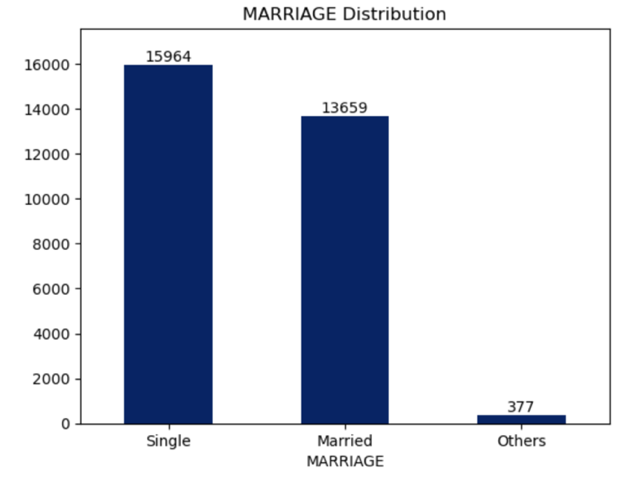
The pivot table further explores the relationship between Gender (SEX) and the Target variable.

* **Males** have a default rate of 24.17%, with 2,873 out of 11,888 male clients in default.
* **Females** have a lower default rate of 20.78%, with 3,763 out of 18,112 female clients in default.

These findings suggest that **Gender (SEX) may play a role in predicting default**, with males exhibiting a higher likelihood of default within this dataset.

【**Disclaimer**】However, it is important to note that *t*he observed gender differences in default rates do not imply a causal relationship. Gender may be correlated with other factors that directly influence the likelihood of default. To adequately assess the impact of gender on default probability, it would be necessary to control for other variables through multivariate analysis.

### 3.2.2. Categorical Features 1 – Marital Status (MARRIAGE)



|  |  |  |
| --- | --- | --- |
| MARRIAGE | Number | Percentage |
| Married (1) | 13,659 | **45.53%** |
| Single (2) | 15,964 | **53.21%** |
| Other (3) | 377 | 1.26% |

The histogram illustrate the distribution of the marital status of the clients, where the majority are Single (53.21%), followed by Married (45.53%), and a small fraction categorized as 'Other' (1.26%).

**Pivot Table – Marital Status (MARRIAGE) against TARGET variable**

|  |  |  |  |
| --- | --- | --- | --- |
| TARGET | Default | Not Default | Default Percentage |
| 2. MARRIAGE | | | |
| 2.1. Married | 3,206 | 10,453 | **23.47%** |
| 2.2. Single | 3,341 | 12,623 | **20.93%** |
| 2.3. Other | 89 | 288 | 23.61% |

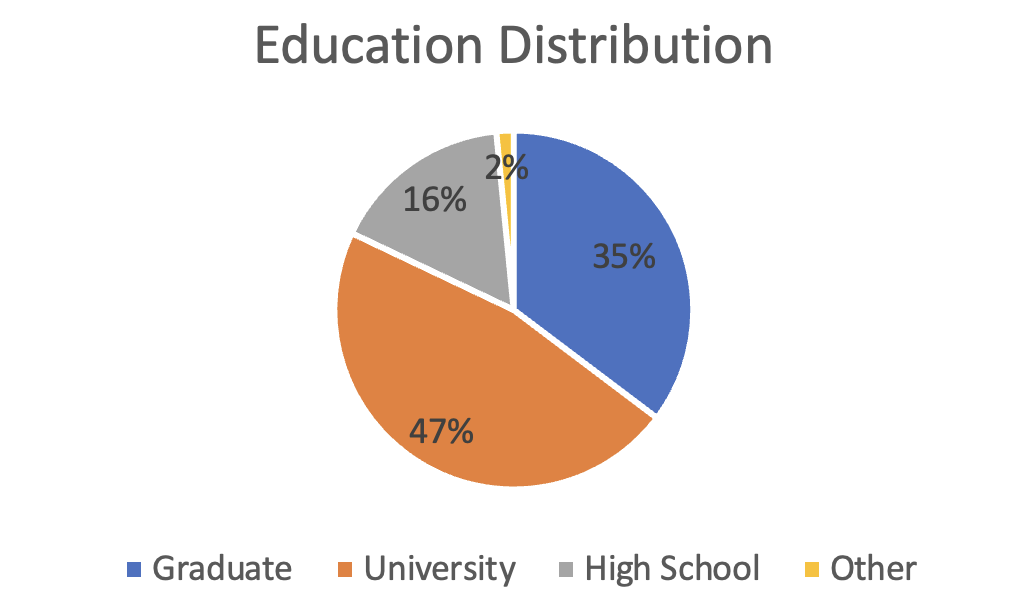
The pivot table further explores the association between Marital Status and the likelihood of default.

* **Married** clients have a higher default rate of 23.47%.
* **Single** clients have a slightly lower default rate of 20.93.
* For the category labelled **Other**, the default rate is 23.61%, but with ONLY 89 out of 377 clients. It is essential to note that this group has considerably fewer observations compared to the other, which could limit the reliability and statistical significance of the default rate observed.

This data indicates that **Marital Status (MARRIAGE) could be a factor in predicting default**, as Married clients show a higher likelihood for default compared to single clients. 【**Disclaimer**】

### 3.2.3. Categorical Features 1 – Education Level (EDUCATION)

|  |  |  |
| --- | --- | --- |
| EUDCATION | Number | Percentage |
| Graduate (1) | 10,585 | **35.3%** |
| University (2) | 14,030 | **46.8%** |
| High School (3) | 4,917 | **16.4%** |
| Other (4) | 468 | 1.6% |



The pie chart illustrates the Education Distribution, which shows that the majority of clients have University level education (46.8%), followed by Graduate level (35.3%), High School (16.4%), and a small percentage falling into the Other category (1.6%).

**Pivot Table – Education Level (EDUCATION) against TARGET variable**

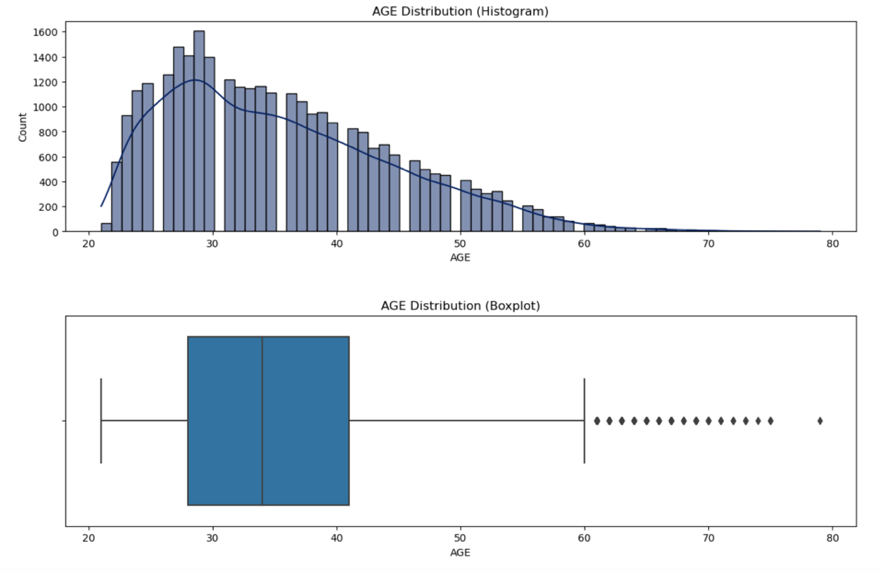
|  |  |  |  |
| --- | --- | --- | --- |
| TARGET | Default | Not Default | Default Percentage |
| 3. EDUCATION | | | |
| 3.1. Graduate | 2,036 | 8,549 | **19.23%** |
| 3.2. University | 3,330 | 10,700 | **23.73%** |
| 3.3. High School | 1,237 | 3,680 | **25.16%** |
| 3.4. Other | 33 | 435 | 7.05% |

* **Graduate**: The default rate for clients with a graduate level of education is 19.23%.
* **University**: The default rate for clients with a university level of education is 23.73%.
* **High School**: Clients with a high school level have the highest default rate of 25.16%.
* Other**:** Although this group has a lowest default rate at around7.05%, there are considerably fewer observations compared to the other categories, which could limit the reliability and statistical significance of the default rate observed for this group.

This pattern reveals an inverse relationship between the level of education and the likelihood of default, clients with a higher education levels tend to have a lower default rates. This data indicates that **Education level could be a factor in predicting Default**, as clients with Graduate education level show a lower likelihood for default compared to clients with University and High School education level. 【**Disclaimer**】

## 3.3. Independent Variable – Numerical Features 1

### 3.3.1. Numerical Feature 1 – AGE

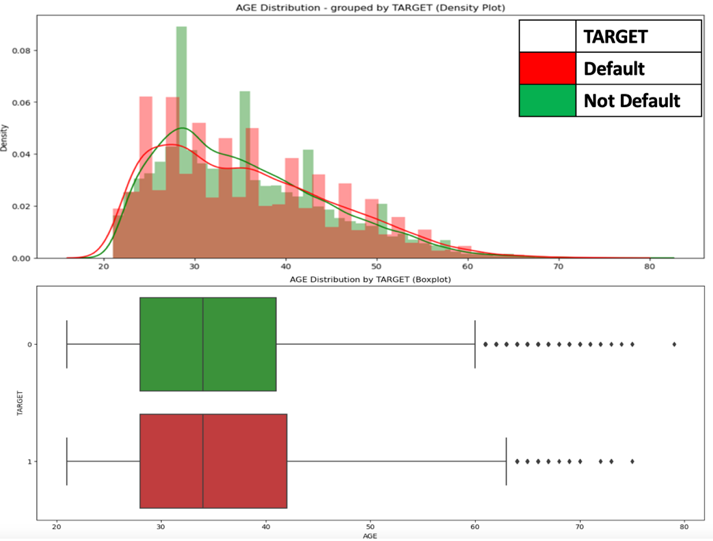


|  |  |
| --- | --- |
| Stats | AGE |
| Mean | 35.48 |
| Standard Deviation | 9.21 |
| MIN | 21 |
| 25% | 28 |
| 50% | 34 |
| 75% | 41 |
| Max | 79 |
| Skewness | 0.732246 |
| Kurtosis | 0.044303 |

The age distribution exhibits a rightward skew, with a skewness value of +0.73, a greater concentration of younger clients within the dataset and a gradual decline in frequency as age increases. The interquartile range, spanning from the 25th percentile (age 28) to the 75th percentile (age 41), further highlighting the majority of a younger demographic. This is aligned with the histogram and boxplot visualizations.

While the boxplot signal the presence of outliers particularly in the higher end of the age range, the kurtosis measure of 0.044, which is close to 0, suggests that the distribution's tails are not significantly different from those of a normal distribution, and the likelihood of outliers is not higher than normal.

**Density Plot and Box Plot of Age Distribution split by Target variable**



When the age distribution is split by the target variable, both the density plot and the boxplot reveals closely aligned distributions in both categories. In the density plot, where the areas under the curve for both groups show highly overlap, indicating similar distribution shapes. The boxplots also illustrate that the median ages is roughly the same for both groups, and both have a similar range of ages and presence of outliers, further confirming their similarity.

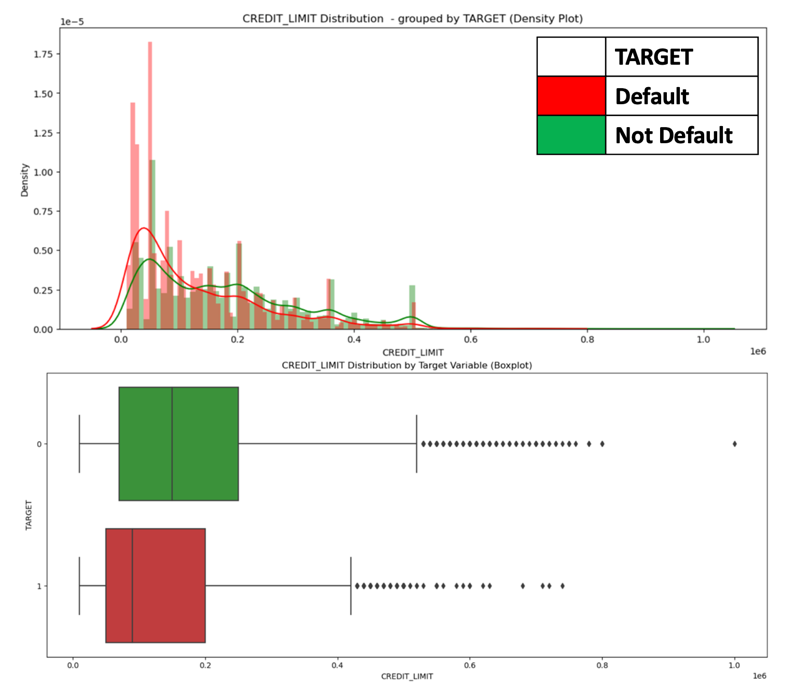
Given the parallel nature of the age distributions across both target categories, it appears that **Age, by itself, might not be a decisive predictor of credit default in this dataset**. The similarity in age characteristics between the two groups suggests that other variables might be more influential in determining the likelihood of credit default. 【**Disclaimer**】

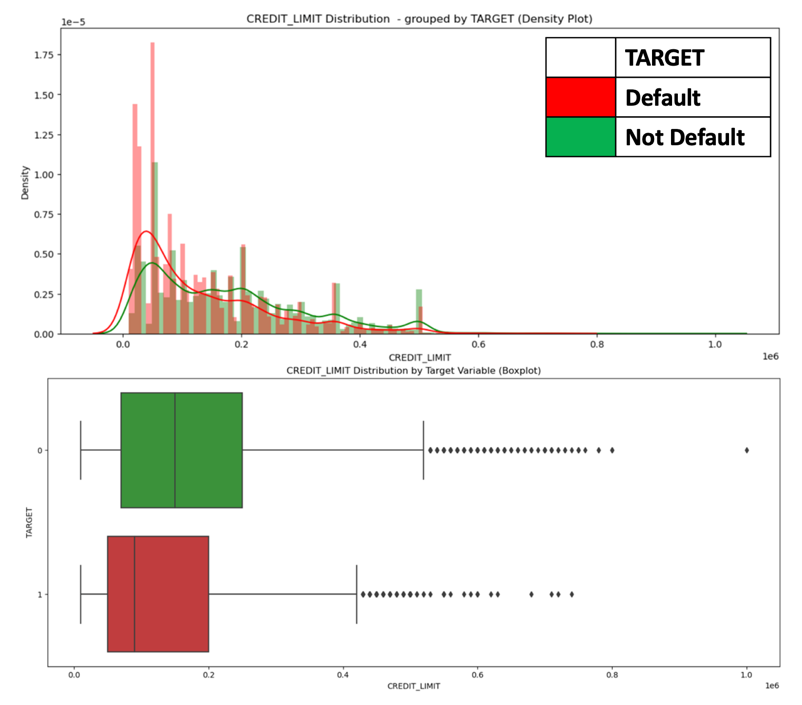
### 3.3.2. Numerical Feature 1 – CREDIT\_LIMIT

|  |  |
| --- | --- |
| Stats | CREDIT\_LIMIT |
| Mean | 167,484 |
| Standard Deviation | 129,748 |
| MIN | 10,000 |
| 25% | 50,000 |
| 50% | 140,000 |
| 75% | 240,000 |
| Max | 1,000,000 |
| Skewness | 0.992867 |
| Kurtosis | 0.536263 |

The histogram shows that the distribution of credit limits is right-skewed, with a large number of clients having lower credit limits and fewer clients having very high credit limits. The boxplot indicates a median credit limit that is relatively low compared to the maximum credit limit, confirming the skewness seen in the histogram. The presence of outliers suggests that there are accounts with unusually high credit limits.

**Density Plot and Box Plot of Credit Limit Distribution split by Target variable**





When separated by the target, the density plot shows that the default category having more weight in the lower credit limit area. This might imply that client with lower credit limits have a higher risk of default. The boxplot also shows that the median credit limit is lower for clients who default. It is consistent with the idea that lower credit limits may be associated with a higher risk of default.

**Pivot table - CREDIT\_LIMIT (10,000 – 100,000) against Target**

|  |  |  |  |
| --- | --- | --- | --- |
| TARGET | Not Default | Default | Default Percentage |
|  |  |  |  |
| CREDIT\_LIMIT |  |  |  |
| 10,000 | 296 | 197 | **39.96%** |
| 16,000 | 2 | 0 | **0** |
| 20,000 | 1278 | 698 | **35.32%** |
| 30,000 | 1042 | 568 | **35.28%** |
| 40,000 | 138 | 92 | **40.00%** |
| 50,000 | 2480 | 885 | **26.3%** |
| 60,000 | 592 | 233 | **28.24%** |
| 70,000 | 521 | 210 | **28.73%** |
| 80,000 | 1204 | 363 | **23.17%** |
| 90,000 | 485 | 166 | **25.50%** |
| 100,000 | 776 | 272 | **25.96%** |

The pivot table present the default percentage across the selected credit limit intervals (10,000 to 100,000). It shows that the default percentage of the clients with a lower credit limit range is generally high than the average default rate of 22%. This table reinforces the observation from the visualizations that lower credit limits are associated with a higher percentage of defaults. To conclude, This data indicates that the **Credit Limit may serves as a significant factor in predicting Default**. As clients with a lower credit limit show a higher likelihood for default. 【**Disclaimer**】

## 3.4. Independent Variable - Categorical Feature 2 - Repayment Status (PAY\_XXX)

**Frequency distribution of Repayment Status across April to September 2005**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Repayment Status | PAY\_APR | PAY\_MAY | PAY\_JUN | PAY\_JUL | PAY\_AUG | PAY\_SEP |
| -2 (No consumption) | 4895 | 4546 | 4348 | 4085 | 3782 | 2759 |
| -1 (Paid in full) | 5740 | 5539 | 5687 | 5938 | 6050 | 5686 |
| 0 (Min Pay) | 16286 | 16947 | 16455 | 15764 | 15730 | 14737 |
| 1 (times delay) | 0 | 0 | 2 | 4 | 28 | 3688 |
| 2 | 2766 | 2626 | 3159 | 3819 | 3927 | 2667 |
| 3 | 184 | 178 | 180 | 240 | 326 | 322 |
| 4 | 49 | 84 | 69 | 76 | 99 | 76 |
| 5 | 13 | 17 | 35 | 21 | 25 | 26 |
| 6 | 19 | 4 | 5 | 23 | 12 | 11 |
| 7 | 46 | 58 | 58 | 27 | 20 | 9 |
| 8 | 2 | 1 | 2 | 3 | 1 | 19 |

**Sum up the number of clients who had one or more instances of payment delays (Status: 1-8),**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Repayment Status | PAY\_APR | PAY\_MAY | PAY\_JUN | PAY\_JUL | PAY\_AUG | PAY\_SEP |
| -2 (No consumption) | 4895 | 4546 | 4348 | 4085 | 3782 | 2759 |
| -1 (Paid in full) | 5740 | 5539 | 5687 | 5938 | 6050 | 5686 |
| 0 (Min Pay) | 16286 | 16947 | 16455 | 15764 | 15730 | 14737 |
| Sum of Clients with Payment Delays(# >=1) | 3079 | 2968 | 3510 | 4213 | 4438 | 6818 |

To enhance the interpretability and clarity in data visualization of the Repayment Status variable. We consolidate the number of clients with one or more instances of payment delays into a single category, The aim is to capture a comprehensive picture of those who exhibit any degree of payment difficulty. providing a clear differentiation between those who maintain their payment schedules (pay in full, min pay) and those who do not, streamlining the analysis by focusing on the occurrence of risk behavior rather than its frequency

Minimum Payment (0): The largest group of clients are those who make the minimum payment. There is a slight decreasing trend in this category.

Sum of Clients with Payment Delays (1 and above): shows a increasing trend in the number of clients with payment delays over time, indicating worsening individual financial condition and probably unsuccessful interventions by the credit institution to help clients get back on track with their payments.

No Consumption (-2): The constant decrease from April to Sep might suggest a trend towards more credit card accounts being used.

Paid in Full (-1): The numbers remain relatively stable across the months for the accounts paid in full, indicating a consistent repayment behavior among a large group of the customers.

**Pivot Table: Repayment Status variable (PAY\_XXX) against Target Variable**

|  |  |  |  |
| --- | --- | --- | --- |
| TARGET | Not Default | Default | Default Percentage |
|  |  |  |  |
| PAY\_SEP |  |  |  |
| -2 (No consumption) | 2394 | 365 | **13.23%** |
| -1 (Paid in full) | 4732 | 954 | **16.78%** |
| 0 (Min Pay) | 12849 | 1888 | **12.81%** |
| Sum of Clients with Payment Delays(#>=1) | 3389 | 3429 | **50.29%** |
|  |  |  |  |
| PAY\_AUG |  |  |  |
| -2 (No consumption) | 3091 | 691 | **18.27%** |
| -1 (Paid in full) | 5084 | 966 | **15.97%** |
| 0 (Min Pay) | 13227 | 2503 | **15.91%** |
| Sum of Clients with Payment Delays(#>=1) | 1962 | 2476 | **55.79%** |

When utilizing the pivot table of the repayment status (September and August as example) against the Target. Clients who has a status of no consumption (-2), pay in full (-1), or make minimum payments (0) in September and August show default rates significantly lower than the average default rate of 22%, as compared with around **13 to 18%.** (Recall those categorical demographics variable that is lower than the average default rate is **around 20**% only, i.e. Female, Single, Graduate)

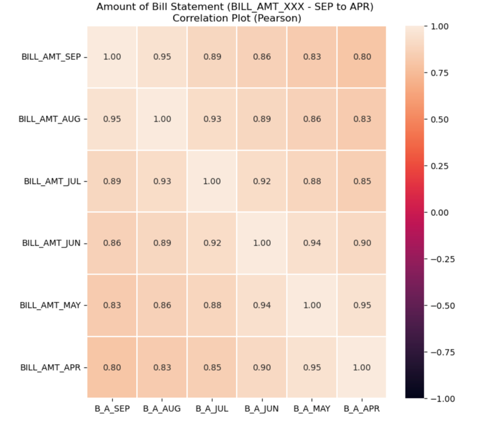
On the other hand, for the group of clients with payment delays (status >=1). This group shows an exceptionally high default percentage— around 50% for September and August. These figures show an obvious contrast which can infer that **Repayment Status could potentially serve as a VERY strong differentiator** when assessing the risk of default. Clients with even a single payment delay are more likely to default in the future, suggesting that proactive measures could be implemented for such clients to mitigate risk. 【**Disclaimer**】

## 3.5. Feature Correlation

To explore the relationship among the variables, we use a heat map along with a correlation matrix, and computed the Pearson correlation coefficients for all variable pairs. Highly correlated variables have a lighter colour and coefficients close to 1, indicating a stronger positive correlation.

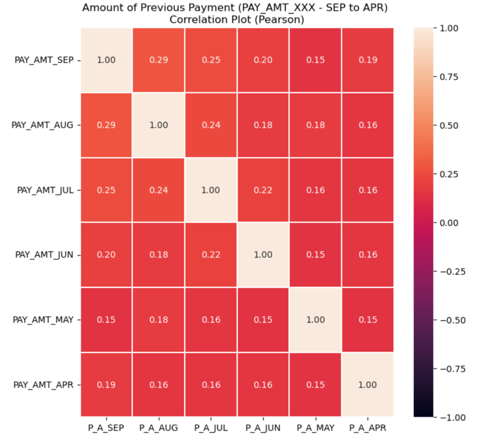
### 3.5.1. Independent Variable – Numerical Feature 2

#### 3.5.1.1. Numerical Feature 2 – Bill Statement Amount (BILL\_AMT\_XXX)



The Bill Statement variable show very high correlations with each other, often exceeding 0.9. This might suggested that the bill amounts are likely to be consistent over time, unless any significant changes in spending behavior. These **High Correlations** indicates a **Significant** **Multicollinearity** problem among these variables. Given these results, particular attention would be paid to the Bill Statement variables when building the logistic regression model. We might consider removing most of the variables, OR combining them into a single metric, i.e. Creating an average.

#### 3.5.1.2. Numerical Feature 2 – Previous Payment (PAY\_AMT\_XXX)



While for the Previous Payment variable (PAY\_AMT\_XXX), the result show a low correlation with each other, within the range of 0.1 to 0.3. suggests that there is no concerning level of multicollinearity among these variables. Each one is providing relatively unique information that is not redundant with the information from the others.

# **4. Problem of Interest Identified from Preliminary Analysis**

Through our initial exploration of the dataset, we have highlighted two primary problems that our predictive model must address:

## **4.1. First Problem - Developing an Accurate Predictive Model**

The first problem is to build a reliable model capable of predicting credit card defaults among clients with high accuracy. The complexity arises from the need to navigate through a multitude of variables and their interrelations, which could significantly influence the prediction outcome.

## **4.2. Second Problem - Assessing the Impact of Various Attributes**

The second problem is to understand how different attributes affect the probability of a client defaulting. Our preliminary analysis indicates that factors such as repayment status, credit limit, and demographic variable each play a role. Quantifying the impact of these attributes will be crucial for refining the predictive model.

# **5. Data Pre-processing**

Before diving into the modelling, we ensured our dataset was primed and ready for analysis.

## **5.1. Treatment of Missing Values and Data Cleaning**

We first tackled any missing values and clean up the data. Details are in the **Appendix**.

## **5.2. One-Hot Encoding for Categorical Feature**

For categorical data like "SEX," "EDUCATION," and "MARRIAGE," we applied one-hot encoding. This technique was also used for the various Repayment Statuses "PAY\_XXX" from September to April, transforming them into a format suitable for our model.

## **5.3. Splitting the Dataset: Training & Validation Sets**

70% for training our model and 30% for validation to test our model's predictions.

# **6. Model Building and Analytics**

## **6.1. Logistic Regression**

Logistic regression is a powerful statistical method for binary classification, exceling in estimating the probability of an event occurrence by fitting data to a logistic function, which makes it an ideal choice for our analysis where the target variable is binary (Default vs. Not Default).

### 6.1.1. Full Logistic Regression

**A close-up of a test

Description automatically generated**

Likelihood ratio ChiSq’s probability is less than 0.05 (<.0001), the model is statistically significant.

The Full logistic regression model includes all features to assess their individual contribution to the Target variable. This Full model serves as a benchmark for subsequent feature selection methods.

### 6.1.2. Feature Selection - Forward & Backward Logistic Regression

Feature selection was carried out using both Forward & Backward logistic regression techniques, detailed in the **Appendix**. The summary table visually distinguishes which variables were selected.

|  |  |  |  |
| --- | --- | --- | --- |
| Feature Selection | Full Logistic | Forward Logistic | Backward Logistic |
| Demographics | | | |
| 1. SEX |  |  |  |
| 2. MARRIAGE |  |  |  |
| 3. EDUCATION |  |  |  |
| 4. AGE |  |  |  |
| Credit Card Usage | | | |
| 5. CREDIT\_LIMIT |  |  |  |
| 6. PAY\_SEP |  |  |  |
| 7. PAY\_AUG |  |  |  |
| 8. PAY\_JUL |  |  |  |
| 9. PAY\_JUN |  |  |  |
| 10. PAY\_MAY |  |  |  |
| 11. PAY\_APR |  |  |  |
| 12. PAY\_AMT\_SEP |  |  |  |
| 13. PAY\_AMT\_AUG |  |  |  |
| 14. PAY\_AMT\_JUL |  |  |  |
| 15. PAY\_AMT\_JUN |  |  |  |
| 16. PAY\_AMT\_MAY |  |  |  |
| 17. PAY\_AMT\_APR |  |  |  |
| 18. BILL\_AMT\_SEP |  |  |  |
| 19. BILL\_AMT\_AUG |  |  |  |
| 20. BILL\_AMT\_JUL |  |  |  |
| 21. BILL\_AMT\_JUN |  |  |  |
| 22. BILL\_AMT\_MAY |  |  |  |
| 23. BILL\_AMT\_APR |  |  |  |

For the **AGE** **variable**, our earlier data exploration highlighted that its distribution is quite similar across the target categories. It was not selected by both reduced logistic regression. This supports our initial observation of AGE's limited differential power in predicting defaults.

Most of the **Repayment Status variables** (PAY\_XXX) (except the one in August) are selected, highlighting their significant role as indicators of credit default. While only one **Bill Statement variable** (BILL\_AMT\_XXX) is included in both reduced logistic regression which is consistent with our earlier concerns about multicollinearity among these variables that can affect the stability and interpretation of the coefficient estimates.

Next, it is important to evaluate the performance of the reduced model (Forward and Backward) against the Full model to ensure that the simplification does not significantly compromise the model's accuracy and predictive ability.

## 6.2. Evaluation of Model Performance

### 6.2.1. Interpretation of Performance Metrics

**Accuracy Rate (1-Misclassification Rate):** Given the slight imbalance of the dataset, where 78% of observations are 'Not Default', this figure sets a baseline for model performance. In a scenario where an AI model which learns nothing, but making the predictions based solely on the majority class, it would probably achieve an accuracy rate of around 78% in the validation set. Therefore, this serves as a benchmark – if our model has a accuracy rate smaller than 78%, we are performing even worse than the naïve classifier (AI model learns nothing) that only predicts the most common outcome.

**Sensitivity (TP/(TP+FN)):** Domain knowledge informs us that the cost of failing to identify a default client (a false negative) is higher than incorrectly classify a non-default client as default (a false positive). Therefore, sensitivity becomes a critical metric. It measures the proportion of actual default that our model correctly identify. A high sensitivity rate is desirable as it reflects the model's ability to capture the majority of the Default cases.

### 6.2.2. Model Comparison – Logistic Regression Models

|  |  |  |  |
| --- | --- | --- | --- |
| Performance | Full Logistic Regression | Forward Logistic | Backward Logistic |
| Training | | | |
| 1. Accuracy Rate | **82.18%** | **82.152%** | **82.128%** |
| Validation | | | |
| 1. Accuracy Rate | **82.08%** | **82.035%** | **82.035%** |
| 2. Average Squared Error | **0.13558** | **0.13557** | **0.13559** |
| 3. Sensitivity: TP/(TP+FN) | **0.3630** | **0.3640** | **0.3640** |
| 4. Specificity: TN/(TN+FP) | **0.9509** | **0.9501** | **0.9501** |

(**Appendix**) The Full Logistic Regression model shows the best performance on the training set. For the validation set, in general all the models have a accuracy rate of around 82% that is significantly higher than the baseline of (78%), indicating that they have learnt something which is better than the naïve classifier. The Full model achieves the highest accuracy rate and specificity in the validation set. However, it is notable that the Forward Logistic Regression model has the lowest average squared error and the highest sensitivity, which is our primary metric of interest given the higher cost associated with false negatives. The Backward Logistic Regression model has the same sensitivity as the Forward model.

Despite the Full Logistic Regression model's marginally better accuracy and specificity, the Forward and Backward models has a similar performance as the Full model while utilizing fewer variable. They successfully eliminate the AGE variable and limit the number of Bill Statement variables, streamlining the model without a significant loss in performance.

(AUC-ROC)

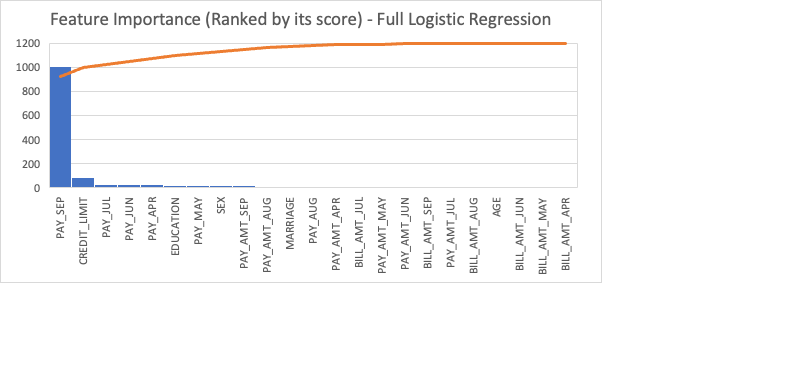
## 6.3. Full Logistic Regression Analysis

### 6.3.1. Statistical Significance of Variables

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Full Logistic | Forward Logistic | Backward Logistic | Full Logistic  Pr > ChiSq |
| Demographics | | | | |
| 1. SEX |  |  |  | **< 0.0001** |
| 2. MARRIAGE |  |  |  | **0.0011** |
| 3. EDUCATION |  |  |  | **< 0.0001** |
| 4. AGE |  |  |  | **0.566** |
| Credit Card Usage | | | | |
| 5. CREDIT\_LIMIT |  |  |  | **< 0.0001** |
| 6. PAY\_SEP |  |  |  | **< 0.0001** |
| 7. PAY\_AUG |  |  |  | **0.4061** |
| 8. PAY\_JUL |  |  |  | **0.0007** |
| 9. PAY\_JUN |  |  |  | **0.0023** |
| 10. PAY\_MAY |  |  |  | **0.0041** |
| 11. PAY\_APR |  |  |  | **0.0021** |
| 12. PAY\_AMT\_SEP |  |  |  | **< 0.0001** |
| 13. PAY\_AMT\_AUG |  |  |  | **0.0002** |
| 14. PAY\_AMT\_JUL |  |  |  | **0.4624** |
| 15. PAY\_AMT\_JUN |  |  |  | **0.1894** |
| 16. PAY\_AMT\_MAY |  |  |  | **0.1418** |
| 17. PAY\_AMT\_APR |  |  |  | **0.0378** |
| 18. BILL\_AMT\_SEP |  |  |  | **0.365** |
| 19. BILL\_AMT\_AUG |  |  |  | **0.5606** |
| 20. BILL\_AMT\_JUL |  |  |  | **0.0542** |
| 21. BILL\_AMT\_JUN |  |  |  | **0.637** |
| 22. BILL\_AMT\_MAY |  |  |  | **0.7748** |
| 23. BILL\_AMT\_APR |  |  |  | **0.8291** |

(**Appendix**) Variables have been coloured based on their p-values to denote statistical significance. Variables with p-values less than 0.05, indicating statistical significance, are marked in yellow, and vice versa. The result aligns closely with the feature selection outcomes of the reduced models. Significant variables such as Sex, Marriage, Education, Credit Limit and Repayment Status variable have a low p-values, indicating a strong association with the Target outcome. For the Bill Amount variable in July, which has a p-value slightly above 0.05, we marked it in green. This highlight that while its p-value is above the standard threshold for significance, it is considered marginally significant and we'll treat it as having passed the threshold for the purposes of our analysis.

### 6.3.2. Feature Importance of Variables



(Appendix) The feature importance scores and the corresponding graph illustrates the predictive power of each variable, with the Repayment Status variables notably standing out (Among the top ranks—five out of the first eight influential predictors are Repayment Status variable). The most recent Repayment Status, specifically for September (PAY\_SEP), not only stands at the forefront, but also dominates other variables in terms of feature importance as shown in the graph, suggesting the critical significance of a client’s latest payment behavior as a strong indicator of credit default.

Following the Repayment Status variables, Credit Limit (2nd) and others demographic factors such as Education (6th) and Sex (8th) are another influential predictor, which also exhibits a significant influence on the model's outcome.

### 6.3.3. Maximum Likelihood Estimates

|  |  |  |
| --- | --- | --- |
| Variable | Estimate | Exp (Est) |
| 1. Repayment Status (September): 0 (The Min pay is made) | **-1.0996** | **0.333** |
| 2. Repayment Status (September): -2 (No Credit Card Consumption) | **-0.9119** | **0.402** |
| 3. Repayment Status (September): 2 (2 months delays) | **+1.1855** | **3.272** |

In examining the maximum likelihood estimates, the coefficients and their corresponding odds ratios reveal the strength and direction of each variable’s association with the probability of default. For the Repayment Status in September (PAY\_SEP), different categories provide various insights:

Repayment Status (September): "Min pay is made" (PAY\_SEP\_0) (binary variable after one-hot encoding): With a coefficient of -1.0996 and an odds ratio of 0.333, the model suggests that one increase in this variable (when the minimum payment is made by the client in September), the odds of defaulting in October (Target variable) decrease substantially by |1-0.333| = 66.7%. This indicates a strong negative association between making the minimum payment in September and the likelihood of default (in October).

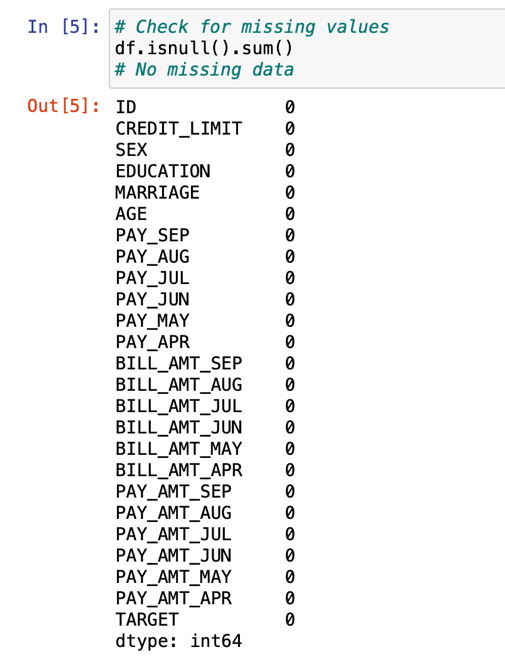
Repayment Status (September): "No Credit Card Consumption" (PAY\_SEP\_-2) (binary variable after one-hot encoding): The coefficient for this status is -0.9119, and the odds ratio is 0.402. This model suggests that one increase in this variable (clients who did not use their credit card for consumption) in September, the odds of defaulting in October (Target variable) decrease substantially by |1-0.402| = 59.8%. This also indicates a strong negative association between having a Repayment Status of no credit-card consumption in September and the likelihood of default (in October).

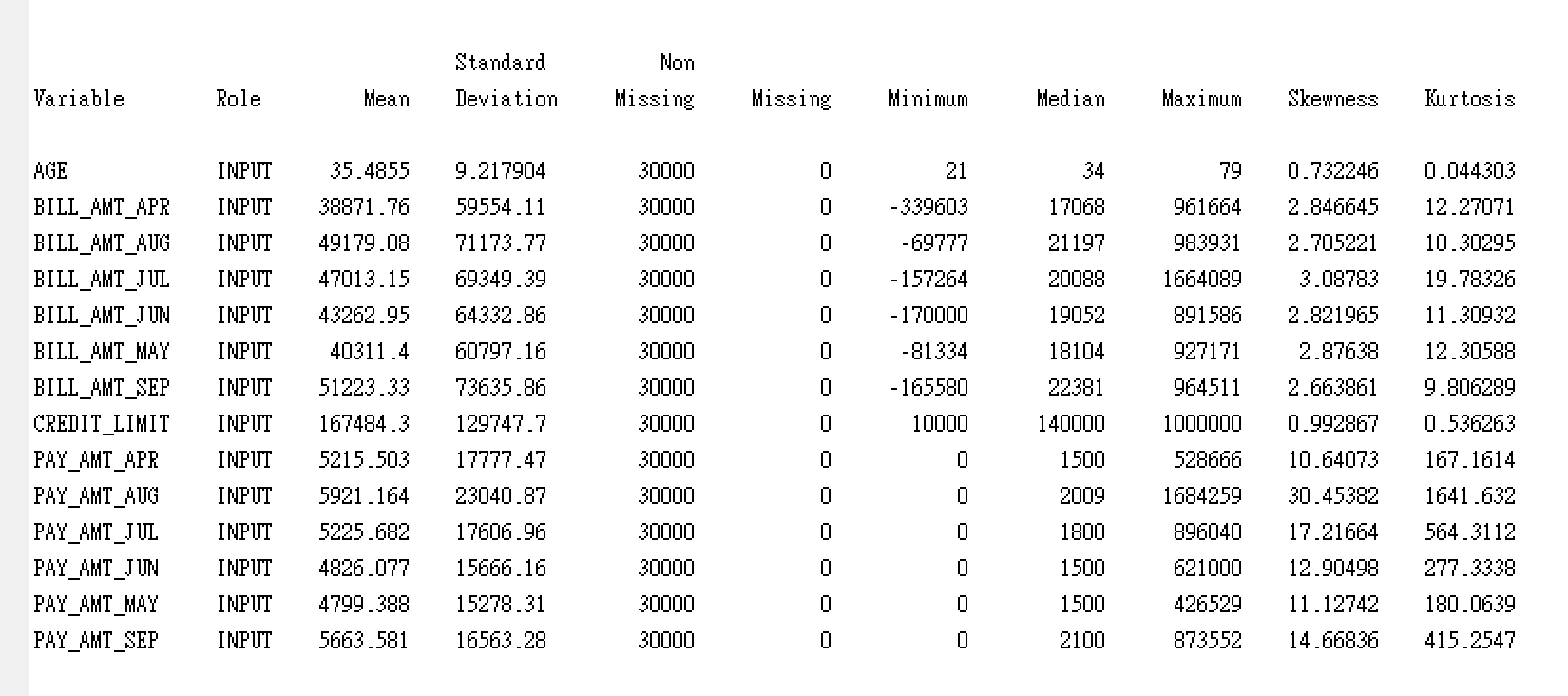
Repayment Status (September): "2 months delay" (PAY\_SEP\_2): Conversely, a coefficient of +1.1855 and an odds ratio of 3.272 indicate that clients with a 2-month delay in payment in September are over three times as likely to default on their credit card payment in October, a substantial increase in risk.

1. References

2. Appendix

12.1. Checking Missing Data

A table of numbers and text

Description automatically generated

No missing data is found.

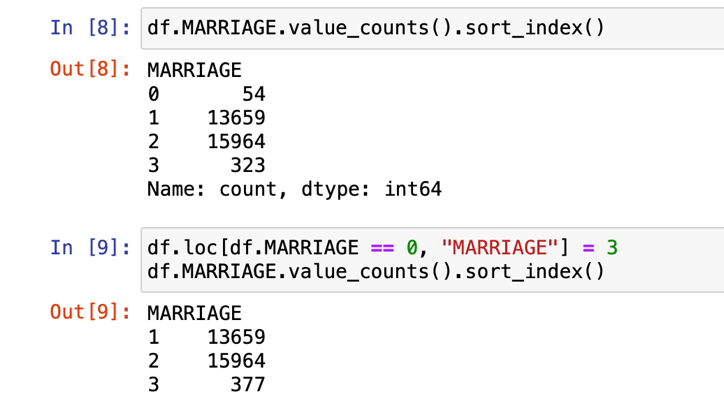
12.2. Data Cleaning

According to the original variable description in the official documentation of the dataset,

MARRIAGE: Marital status (1=married, 2=single, 3=others)

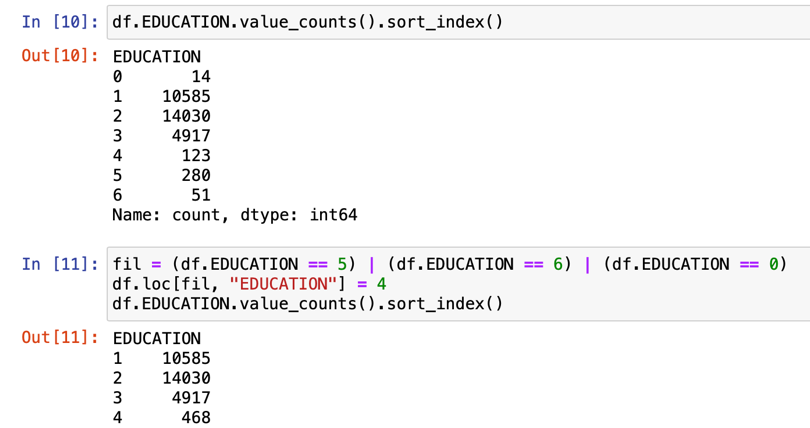
EDUCATION: Educational level (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)

12.2.1. Categorical variable - MARRIAGE



It is not expected to have observations labelled as “0” in “MARRIAGE”, however it is sensible to categorized them as "Other" – Group 3 in “MARRIAGE”.

12.2.2. Categorical variable – EDUCATION



Similarly, for the observations with mislabelled and undocumented “0”, “5” and “6” value in “EDUCATION”, it can be categorized them as “Other” – Group 4 in “EDUCATION:.

12.1 Detailed Procedure of Analysis

12.2 Additional Model Insights and Visualizations

12.3 Dashboard Creation Process

12.4