

Data Preprocessing and Exploratory Data Analysis (EDA)

● Data Preprocessing

本專題採用 **UCI Default of Credit Card Clients Dataset**，共 30,000 筆觀測值與 24 個變數，用以預測客戶違約率並給出信用評分，以下是原資料集變數資訊：

此資料集採用二元變數「違約付款」（是 = 1，否 = 0）作為應變數。此研究回顧了相關文獻，並使用以下 23 個變數作為自變數：

- (1) **LIMIT_BAL**: 給定信用額度（包含個人與家庭額度），單位：新台幣
- (2) **SEX**: 性別（1 = 男性，2 = 女性）
- (3) **EDUCATION**: 教育程度（1 = 研究所，2 = 大學，3 = 高中，4 = 其他）
- (4) **MARRIAGE**: 婚姻狀況（1 = 已婚，2 = 單身，3 = 其他）
- (5) **AGE**: 年齡（年）
- (6 - 11) **PAY_***: 還款狀態（2005 年 4 月至 9 月），PAY_0 = 2005 年 9 月的還款狀態；PAY_2 = 2005 年 8 月的還款狀態；... PAY_6：2005 年 4 月的還款狀況。（-1 = 準時付款，0 = 無延遲，1~9 = 延遲月數）
- (12 - 17) **BILL_AMT***: 帳單金額（新台幣），BILL_AMT1 = 2005 年 9 月的帳單金額；BILL_AMT2 = 2005 年 8 月的帳單金額；...；BILL_AMT6 = 2005 年 4 月的帳單金額。
- (18 - 23) **PAY_AMT***: 實際繳款金額（新台幣），PAY_AMT1 = 2005 年 9 月的還款金額；PAY_AMT2 = 2005 年 8 月的還款金額；...；PAY_AMT6 = 2005 年 4 月的還款金額。

為確保資料品質與後續模型的穩定性，我們進行以下前處理步驟：

1. **移除非特徵欄位：**
刪除識別用欄位 ID，該欄僅作為樣本編號，與違約行為無直接關聯。
2. **修正類別欄位異常值：**
將 EDUCATION 中非法值（0, 5, 6）重新分類為「其他」（4），並將 MARRIAGE 中的非法值 0 改為「其他」（3）。
3. **類別型轉換：**

SEX（性別）、EDUCATION（教育）、MARRIAGE（婚姻）均轉為類別型資料，以利後續 one-hot 編碼。各參數意義如下：

欄位名稱	說明
SEX	性別（1 = 男性，2 = 女性）
EDUCATION	教育程度（1 = 研究所、2 = 大學、3 = 高中、4 = 其他）
MARRIAGE	教育程度（1 = 研究所、2 = 大學、3 = 高中、4 = 其他）

4. PAY 還款狀態欄位修正：

還款狀態欄位 PAY_0 至 PAY_6 中的數值代表當月還款情況，部分樣本出現 -2（無明確定義），此處合併為 -1，視為「準時或無需付款」。

數值	意義
-1	按時付款或無應付款
0	使用信用額度但無延遲
1 - 9	延遲對應月數（例如 2 表示延遲兩個月）

5. 去除重複樣本：

共刪除 35 筆重複資料，最終保留 29,965 筆有效樣本。

6. 數值標準化：

針對連續變數進行 StandardScaler() 標準化，產生第二份資料表（uci_default_cleaned_scaled.csv），以使用於 Logistic Regression、SVM、Neural Network 等模型。

為了強化訓練模型時的表現，我們設計下列 9 項衍生特徵(Derived Features)：

衍生欄位名稱	說明	數值意義範圍
CREDIT_UTILIZATION	信用使用率（最近一期帳單金額 ÷ 額度）	0 - 1，越高代表越接近用滿信用額度
PAY_TO_BILL_RATIO	長期繳清率（六期繳款總額 ÷ 六期帳單總額）	0 - 1，0 表示未繳款或無帳單，1 表示全額繳清
BILL_CHANGE_MEAN	帳單變化程度（六期帳單金額變化平均值）	越大代表帳單金額波動大
PAYMENT_STD	繳款金額標準差	越大代表繳款不穩定
MAX_DELAY	最嚴重延遲月數（PAY_0~6 最大值）	-1 表準時，數值越高表示延遲越久
LAST_DELAY	最近一期延遲月數	-1 表準時，1 - 9 代表延

	(PAY_0)	遲月數
RECENT_PAY_RATIO	最近一期繳款比例 (PAY_AMT1 ÷ BILL_AMT1)	0 表無繳款或無帳單
TOTAL_BILL	六期帳單總額	新台幣金額
TOTAL_PAY	六期繳款總額	新台幣金額

● Exploratory Data Analysis (EDA)

■ 原始特徵分析

1. 數值統計摘要：

顯示各數值型欄位的分布 (count 、 mean 、 std 、 min 、 max)。

	count	mean	std	min	max
LIMIT_BAL	29965.0	167442.005006	129760.135222	10000.0	1000000.0
SEX	29965.0	1.603738	0.489128	1.0	2.0
EDUCATION	29965.0	1.842750	0.744513	1.0	4.0
MARRIAGE	29965.0	1.557283	0.521431	1.0	3.0
AGE	29965.0	35.487969	9.219459	21.0	79.0
PAY_0	29965.0	0.075021	0.990735	-1.0	8.0
PAY_2	29965.0	-0.006641	1.035798	-1.0	8.0
PAY_3	29965.0	-0.029067	1.024849	-1.0	8.0
PAY_4	29965.0	-0.074821	0.987241	-1.0	8.0
PAY_5	29965.0	-0.113799	0.941698	-1.0	8.0
PAY_6	29965.0	-0.127082	0.949184	-1.0	8.0
BILL_AMT1	29965.0	51283.009778	73658.132403	-165580.0	964511.0
BILL_AMT2	29965.0	49236.366294	71195.567392	-69777.0	983931.0
BILL_AMT3	29965.0	47067.916069	69371.352323	-157264.0	1664089.0
BILL_AMT4	29965.0	43313.329885	64353.514373	-170000.0	891586.0
BILL_AMT5	29965.0	40358.334390	60817.130623	-81334.0	927171.0
BILL_AMT6	29965.0	38917.012281	59574.147742	-339603.0	961664.0
PAY_AMT1	29965.0	5670.099316	16571.849467	0.0	873552.0
PAY_AMT2	29965.0	5927.983180	23053.456645	0.0	1684259.0
PAY_AMT3	29965.0	5231.688837	17616.361124	0.0	896040.0
PAY_AMT4	29965.0	4831.617454	15674.464538	0.0	621000.0
PAY_AMT5	29965.0	4804.897047	15286.372298	0.0	426529.0
PAY_AMT6	29965.0	5221.498014	17786.976864	0.0	528666.0
default payment next month	29965.0	0.221258	0.415101	0.0	1.0

2. 目標變數分布與類別變數分析：

違約 (1) 佔 22%，正常 (0) 佔 78%，屬中度不平衡資料。**SEX** 性別項目女性佔 60%，違約率略低於男性。**EDUCATION** 教育程度愈高違約

率愈低。**MARRIAGE** 項目單身者違約率較低。**AGE** 年齡集中於 30 歲上下，老年族群違約率稍高。

違約率(Default Rate): 22.13%

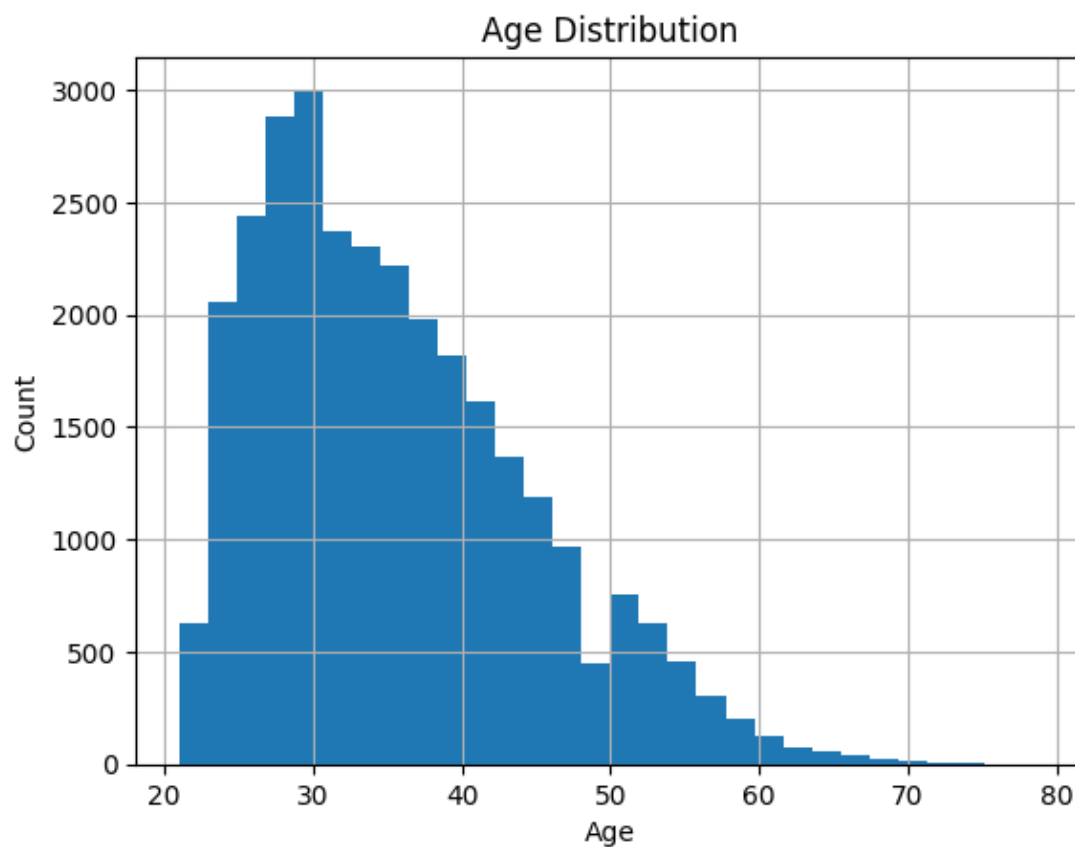
	Target	Count	Percent
0	0	23335	77.874187
1	1	6630	22.125813

	SEX	Count	Percent	Default Rate
0	1	11874	39.63%	24.16%
1	2	18091	60.37%	20.79%

	EDUCATION	Count	Percent	Default Rate
0	1	10563	35.25%	19.24%
1	2	14019	46.78%	23.74%
2	3	4915	16.40%	25.17%
3	4	468	1.56%	7.05%

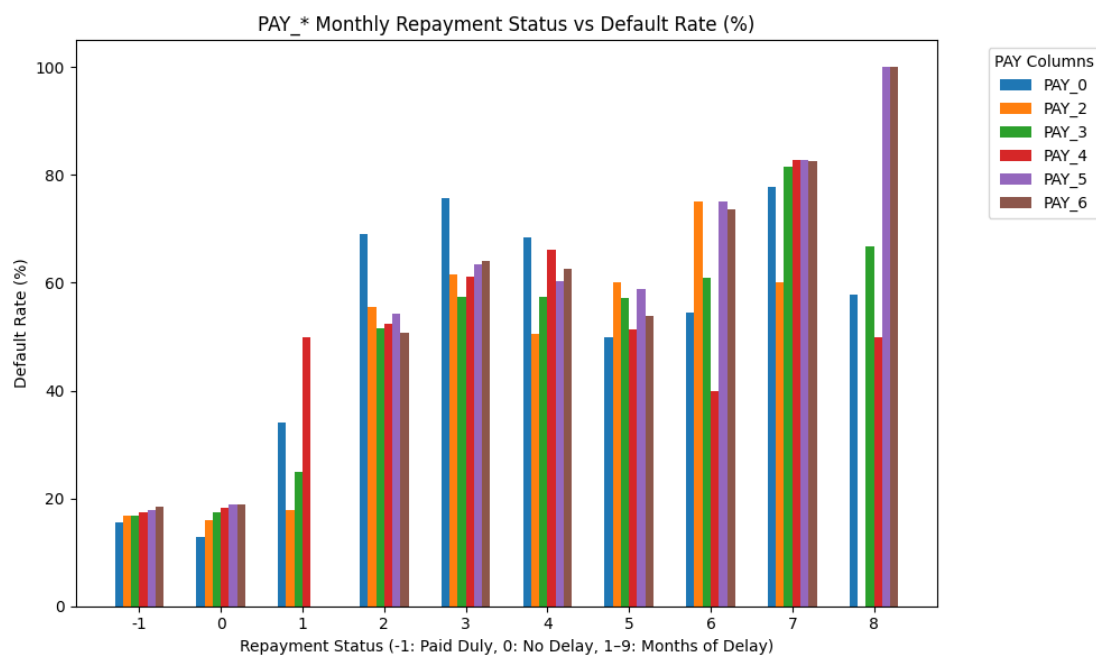
	MARRIAGE	Count	Percent	Default Rate
0	1	13643	45.53%	23.46%
1	2	15945	53.21%	20.95%
2	3	377	1.26%	23.61%

	AGE_GROUP	Count	Percent	Default Rate
0	20-29	9603	32.05%	22.87%
1	30-39	11226	37.46%	20.27%
2	40-49	6456	21.55%	22.94%
3	50-59	2341	7.81%	24.86%
4	60-69	314	1.05%	28.34%
5	70+	25	0.08%	28.00%



3. 還款行為：

延遲月份越多（PAY_0 - PAY_6 數值越高），違約率顯著上升。



■ 衍生特徵分析

1. 數值統計摘要：

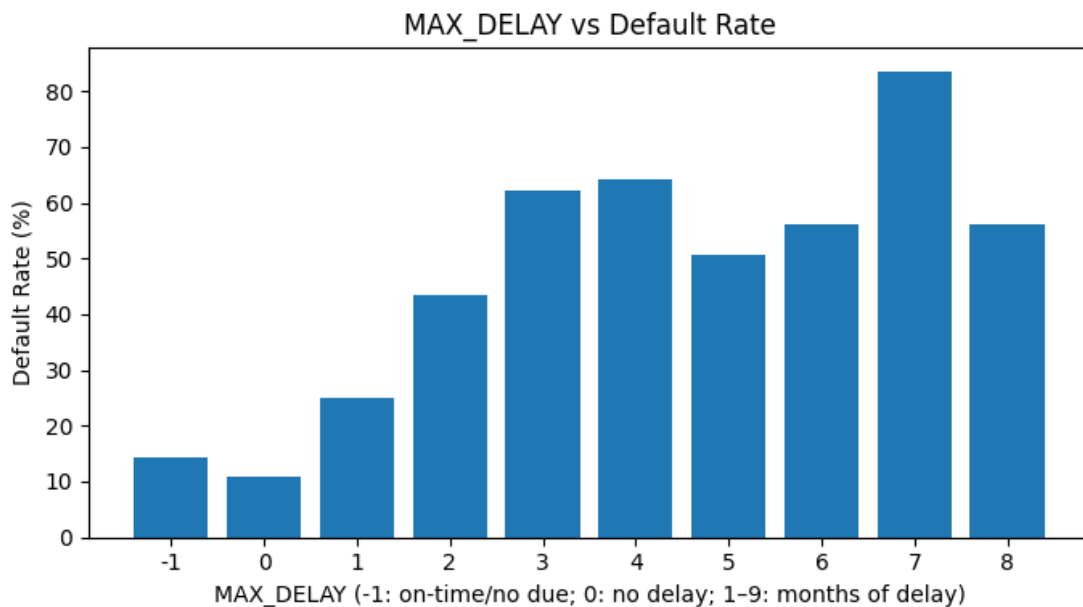
顯示各數值型欄位的分布（count、mean、std、min、max）。

	count	mean	std	min	max
CREDIT_UTILIZATION	29965.0	0.415148	0.386960	0.000000	1.000000e+00
PAY_TO_BILL_RATIO	29965.0	0.381252	7.675470	-546.928571	7.970000e+02
BILL_CHANGE_MEAN	29965.0	7817.530886	13718.221494	0.000000	7.094932e+05
PAYMENT_STD	29965.0	5811.698835	15013.888412	0.000000	6.500983e+05
MAX_DELAY	29965.0	0.509227	1.237792	-1.000000	8.000000e+00
LAST_DELAY	29965.0	0.075021	0.990735	-1.000000	8.000000e+00
RECENT_PAY_RATIO	29965.0	-2.229884	231.253172	-35436.000000	1.145367e+04
TOTAL_BILL	29965.0	270175.968697	379674.444976	-336259.000000	5.263883e+06
TOTAL_PAY	29965.0	31687.783848	60853.841129	0.000000	3.764066e+06

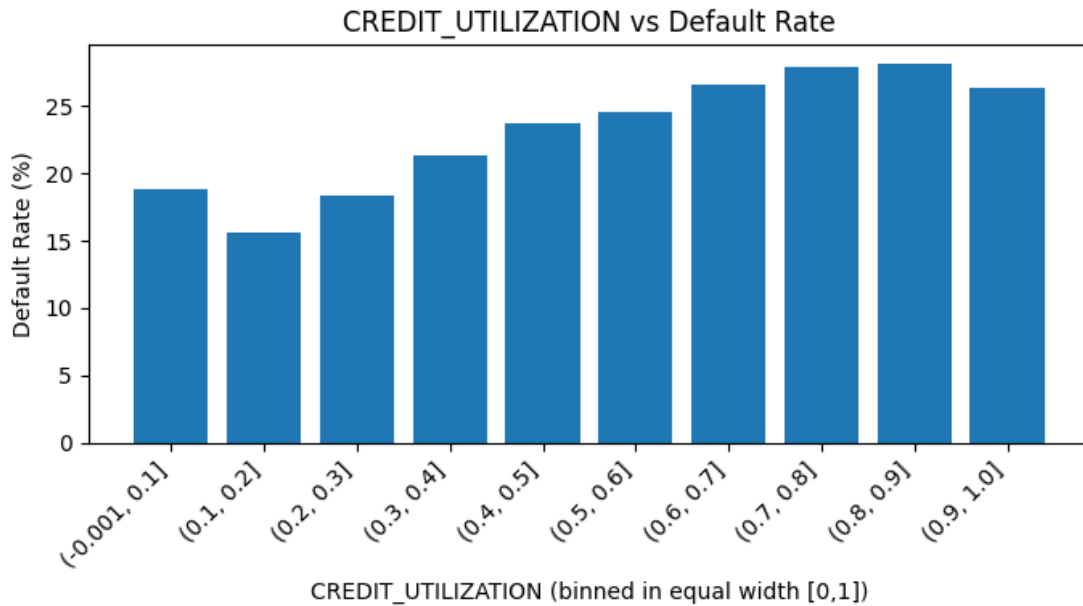
2. 違約率關係：

選取三個最具影響力的衍生變數進行比較：

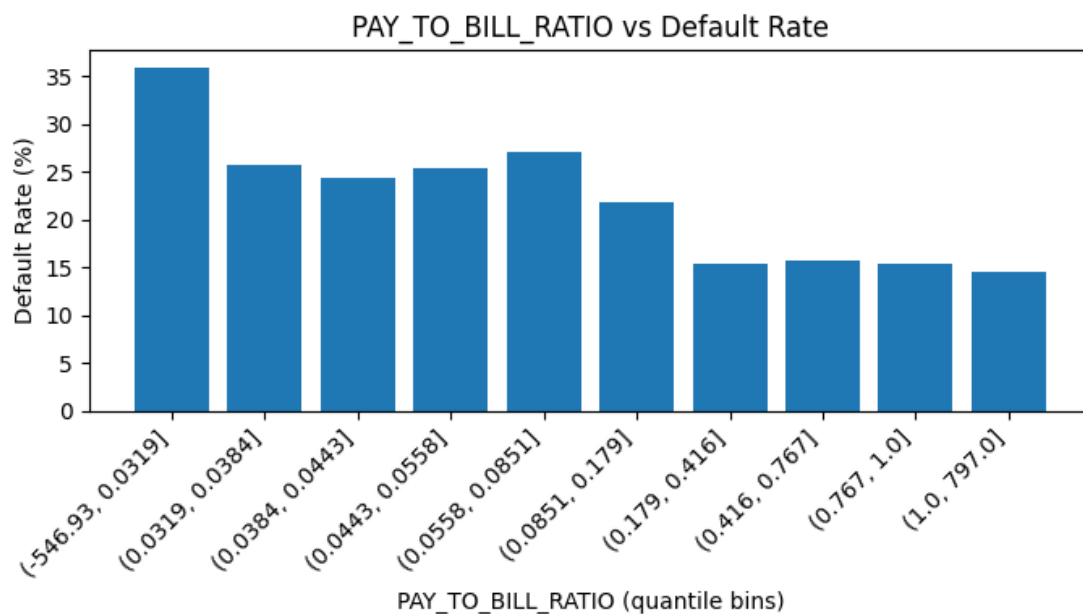
◆ **MAX_DELAY:** 最大延遲月數越高違約率較高。



◆ **CREDIT_UTILIZATION:** 高使用率者違約率較高。



◆ **PAY_TO_BILL_RATIO**: 長期繳清率越低違約率越高。



3. 特徵相關性：

分別將全部特徵（Original + Derived）與衍生特徵（Derived）繪製 Correlation Heatmap 進行觀察，結果顯示：

- ◆ BILL_AMT1-6、TOTAL_BILL 高度正相關；
- ◆ PAY_AMT1-6、TOTAL_PAY 正相關；
- ◆ 衍生特徵如 CREDIT_UTILIZATION、MAX_DELAY、PAY_TO_BILL_RATIO 與原始特徵相關性低，表示成功捕捉新資訊。

