

# Default of Credit Card Clients

Course: MIS 451 - Machine Learning for Business

Lecturers: Mr. Dang Thai Doan & Ms. Huynh Gia Linh

Date: 6/6/2025

Presented by: Group 4 Pieces



# Team Member



**Thanh Thao**  
2232300157



**Gia Tue**  
2132300511



**Hieu Ngan**  
2032300513



**Thanh Giang**  
2132300593

# Table of Contents



Introduction



Model Development



Data Collection



Interpretation &  
Business Insights



Exploratory Data  
Analysis (EDA)



Team Member Roles &  
Responsibilities



Data Cleaning &  
Transformation

# Introduction

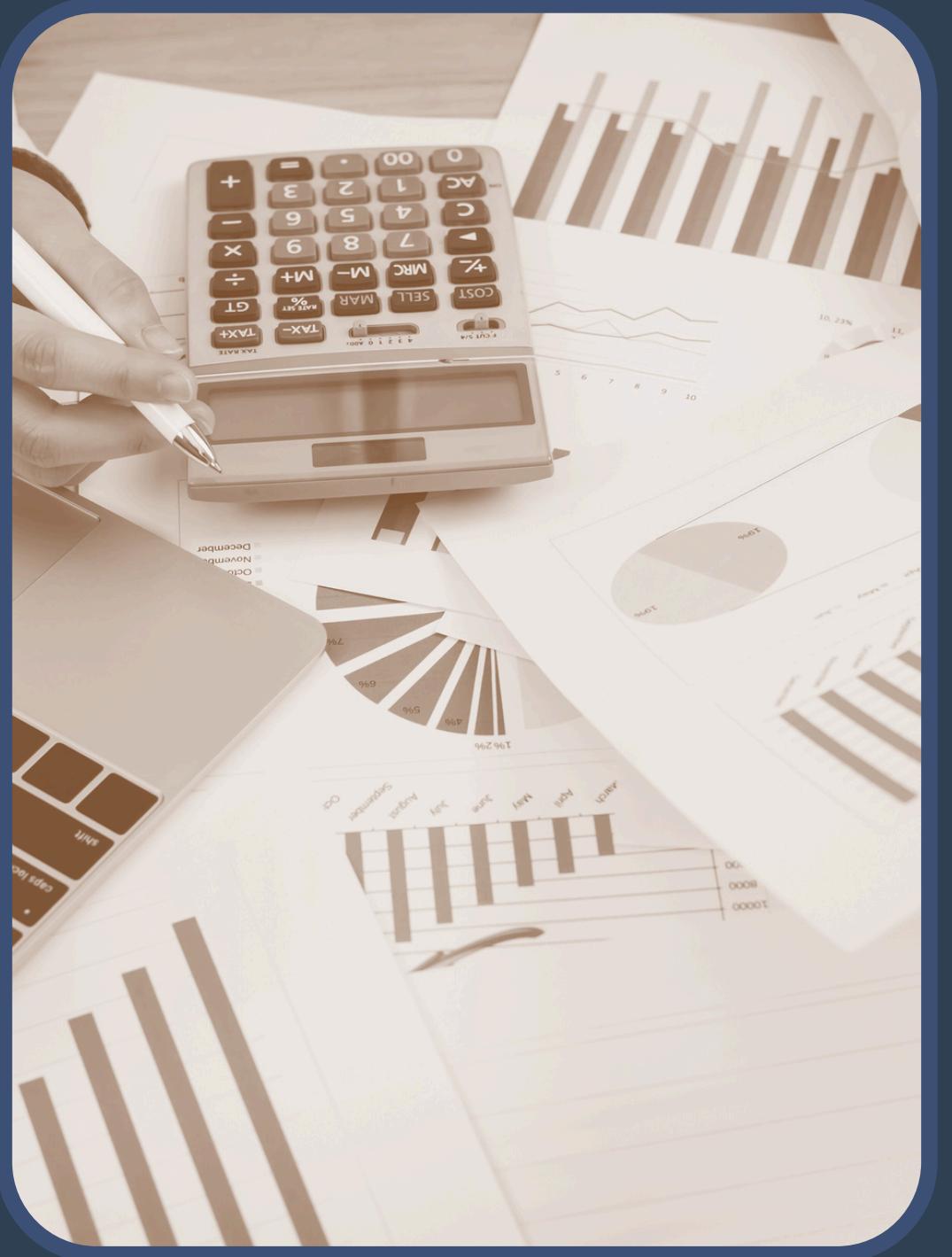
**Dataset:** UCI Default of Credit Card Clients (30,000 records, 23 features)

## Business Problem:

Credit card default prediction is crucial for risk management in the banking sector

Identifying **high-risk clients in advance** allows financial institutions to **take early preventive actions**

**Objective:** Build and evaluate classification models to predict whether a customer will default on their next credit card payment



# Data Collection

## Dataset Source:

- UCI Machine Learning Repository
- Dataset: Default of Credit Card Clients in Taiwan
- 30,000 records, 23 features

## Feature Types:

- Demographics: Gender, age, education, marital status
- Financial Info: Credit limit, bill amounts, payments
- Repayment History: Monthly payment status (April–September 2005)

## Target Variable:

- `default.payment.next.month` → 1 = default, 0 = no default



# Data Collection

Variable	Description
LIMIT_BAL	Credit limit (maximum amount the client can borrow)
SEX	Gender (1 = Male, 2 = Female)
EDUCATION	Education level (1 = Graduate, 2 = University, etc.)
MARRIAGE	Marital status (1 = Married, 2 = Single, etc.)
AGE	Client's age in years
PAY_0 to PAY_6	Repayment status for past 7 months (-1 = paid early, 0 = on time, >0 = delay in months)
BILL_AMT1 to BILL_AMT6	Bill statement amount for each month
PAY_AMT1 to PAY_AMT6	Amount actually paid for each month
default.payment.next.month	Target variable: 1 = default, 0 = no default

# Exploratory Data Analysis (EDA)

## EDA Steps:

- Checked missing values → None found
- Class distribution: ~23% default, ~77% non-default → imbalanced
- Visualized gender, education, and marriage status
- Heatmap of correlations:
  - Strongest predictors: PAY\_0 (0.32), PAY\_2, PAY\_3
  - Demographics had weak correlation

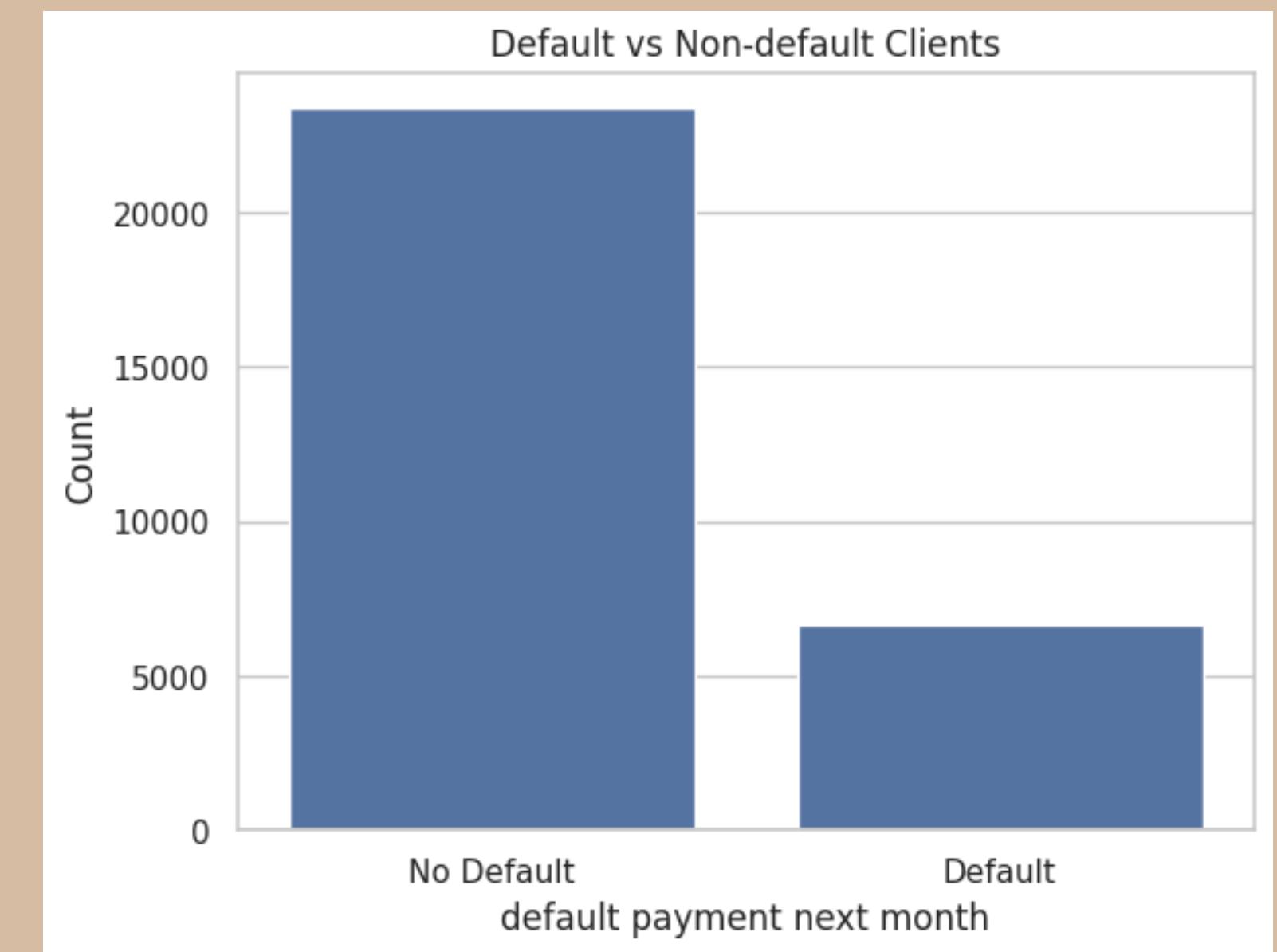
## Key Insight:

Late recent payments are strong indicators of default.



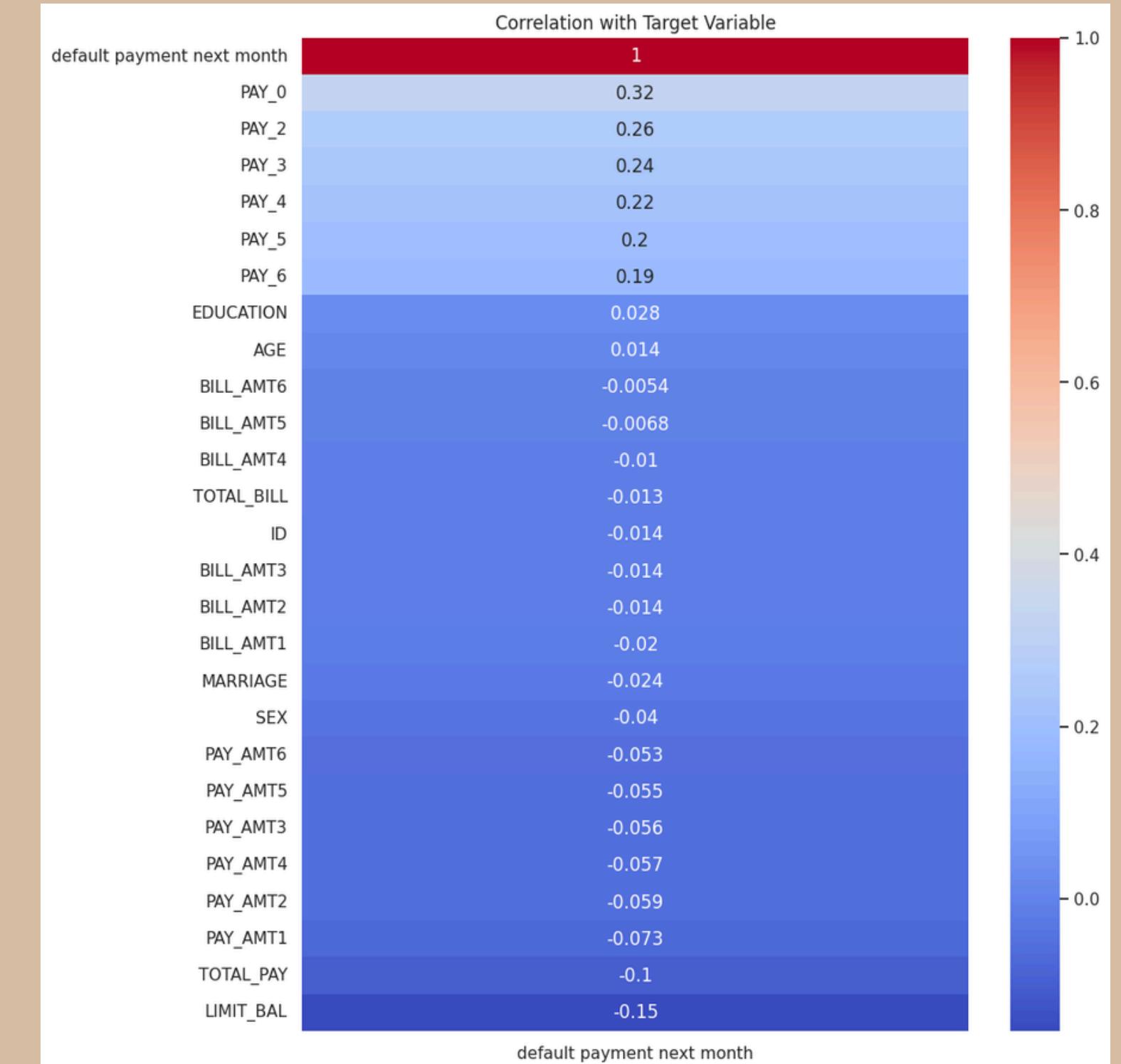
# Exploratory Data Analysis (EDA)

Significant class imbalance, avoid biasd learning



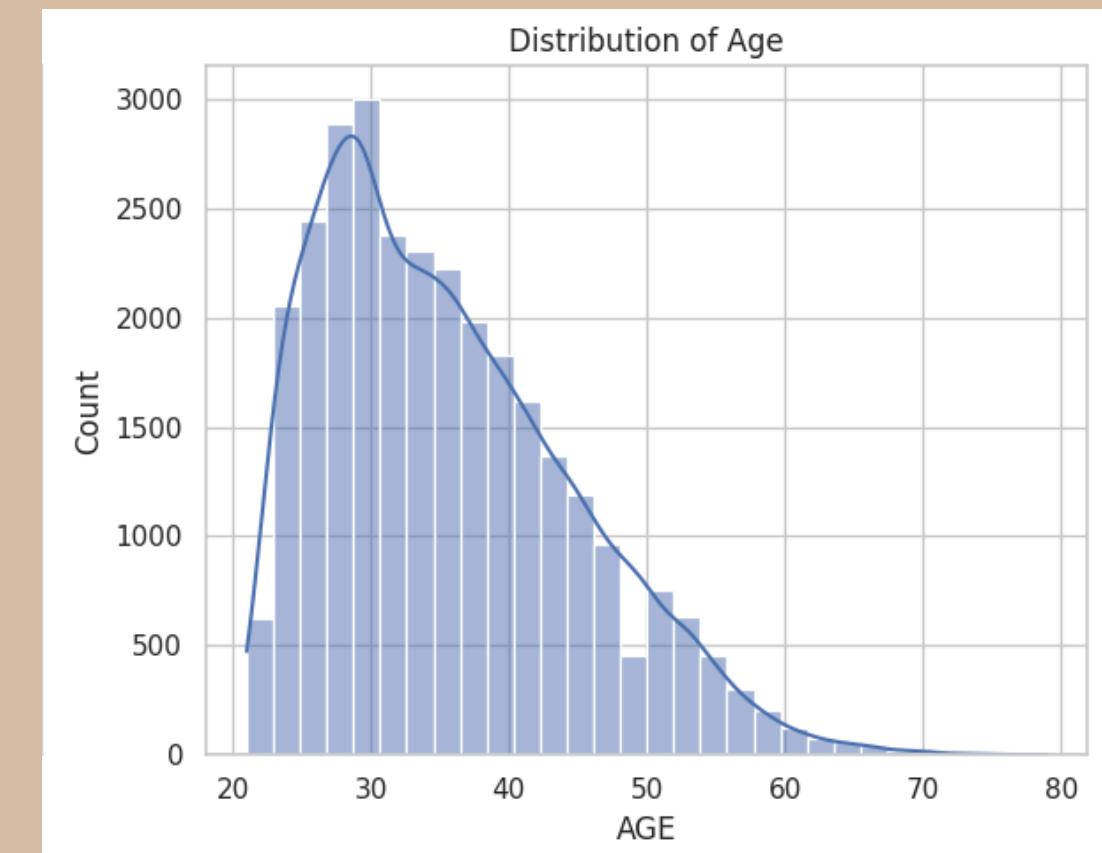
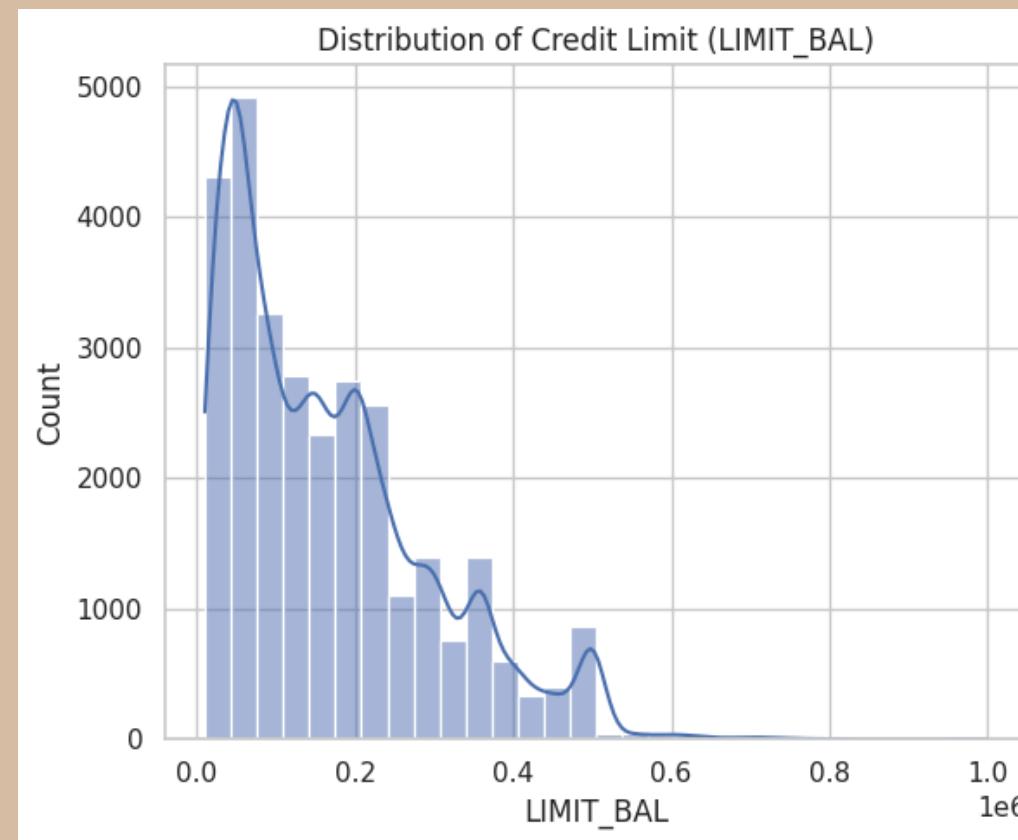
# Exploratory Data Analysis (EDA)

Late payments predict default, higher credit limits lower risk, data needs cleaning

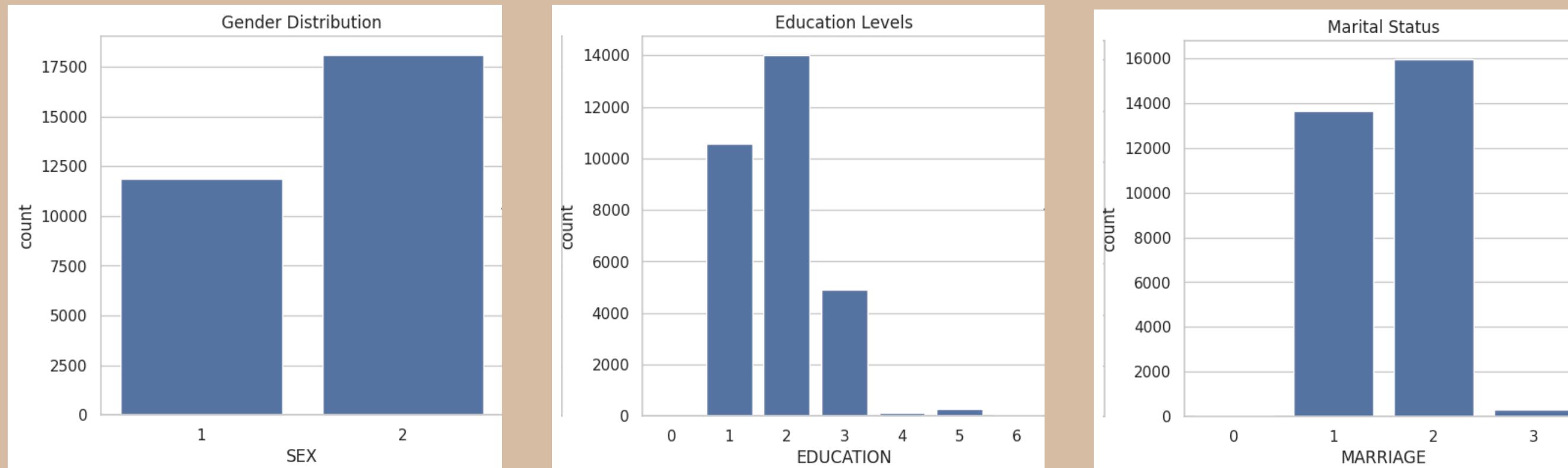


# Exploratory Data Analysis (EDA)

- Mostly credit limit, 200,000 NT\$, long tail on the right
- 25-40 years old, fewer older customers, left-skewed



# EDA



- 60% female (code 2), 40% male (code 1)
- College/graduate degree (code 1, 2), rae values (0, 4, 5, 6), group as 'Others'
- Married/single (code 1, 2), unusual values (0, 3), group as 'Others'

# Data Cleaning & Transformation

## Cleaning Steps:

- Grouped invalid values in education & marriage → 'Others'
- No missing data, no duplicates
- Transformation:
- One-hot encoding for categorical variables

## Feature engineering:

- TOTAL\_BILL\_AMT: Sum of 6 months' bills
- TOTAL\_PAY\_AMT: Sum of 6 months' payments
- Feature selection using ANOVA F-test
- Standardized all features using StandardScaler
- Applied SMOTE to balance class in training set



# Data Cleaning & Transformation

```
[ ] # Combine rare/unexpected values  
data_bank['EDUCATION'] = data_bank['EDUCATION'].replace({0: 4, 5: 4, 6: 4})  
data_bank['MARRIAGE'] = data_bank['MARRIAGE'].replace({0: 3})
```

```
[ ] # One-Hot Encode Categorical Variables  
data_encoded = pd.get_dummies(data_bank, columns=['SEX', 'EDUCATION', 'MARRIAGE'], drop_first=True)
```

```
[ ] # Total amount billed and paid across 6 months  
data_encoded['TOTAL_PAY'] = data_encoded[[f'PAY_AMT{i}' for i in range(1, 7)]].sum(axis=1)  
data_encoded['TOTAL_BILL'] = data_encoded[[f'BILL_AMT{i}' for i in range(1, 7)]].sum(axis=1)
```

```
[ ] from sklearn.feature_selection import SelectKBest, f_classif
```

```
[ ] # 4. Apply SMOTE to the TRAINING data only (to balance classes)  
smote = SMOTE(random_state=42)  
X_train_resampled, y_train_resampled = smote.fit_resample(X_train_scaled, y_train)
```

# Model Development

## Model Evaluation - Metrics

### Metrics Used:

- Accuracy: Overall correctness
- F1-score: Focused on default class
- ROC-AUC: How well model separates default vs non-default (The higher the AUC, the better the model is at classification)



# Model Development

## Model Evaluation - Results

Model	Accuracy	F1 (Default)	ROC-AUC
Logistic Regression	64%	44	7.089
Random Forest	78%	50	7.356
SVM	77%	52	7.469
<b>MLP Neural Network</b>	<b>80%</b>	<b>52</b>	<b>7.572</b>

Best model: MLP Neural Network

# Interpretation & Business Insights

## Top Predictive Features:

- PAY\_0, PAY\_2, PAY\_3 → late payments = higher risk
- LIMIT\_BAL and TOTAL\_PAY\_AMT → higher values = lower risk

## Insights:

- Payment history is more predictive than demographic data.
- Financial institutions can use these models to:
- Detect risky clients early
- Adjust credit policies or offer support
- Improve customer retention and reduce loss



# Team Member Roles & Responsibilities

**Member**

Bùi Gia Tuệ

Nguyễn Thanh Giang

Trần Tiến Thảo Hiếu Ngân

Bùi Thị Thanh Thảo

**Responsibilities**

Designed presentation slides and implemented model code in Google Colab

Focused on model development and selection, including testing different classifiers

Conducted final review and enhancement of all project steps, including improvement suggestions

Provided feedback and edits for the slides; served as the primary presenter



From: Team 4 Pieces

To: Class

# Thank you very much!

