2025

Anurag University

Team 2

Credit Risk Analysis

[Predicting loan default using borrower's data]

This project focuses on predicting credit defaults using machine learning techniques. By analyzing customer financial and demographic data, the models classify individuals as defaulters or non-defaulters. The approach supports financial institutions in managing risk and making informed lending decisions.

Project cover sheet

Project Title	Predicting loan default using borrower's data.	**Course**	MBA BA I Year – III Trimester
Start Date	08-07-2025		
End Date	20-08-2025		

Team Members

SNO	Name	Roll no.
1.	Sirigiri varshitha	24MG202A32
2.	Abhinav Narra	24MG202A19
3.	Sudha	24MG202A16
4.	Anuritha	24MG202A02

Signature:

Date:

Contents

List of figures

List of tables

Executive Summary

This project focused on predicting loan default risks using borrower demographic and financial data sourced from the "Default of Credit Card Clients" dataset from Taiwan's UCI Machine Learning Repository. The dataset consisted of 30,000 credit card clients with features including credit limits, demographic information (age, sex, education, marital status), repayment histories across six months, monthly bill statements, and payment amounts. The binary target variable indicated whether a client defaulted on payment in the following month.

Initial data cleaning included handling missing values and feature engineering such as calculating credit utilization ratios, payment consistency, average payment ratios, bill growth rates, and late payment counts. Exploratory data analysis provided insights on default rates by demographic categories, revealing higher risks among certain education groups, marital statuses, and late payment behaviors.

A gradient boosting classifier (XGBoost) was employed for predictive modeling due to its capability to handle complex nonlinear relationships and feature interactions. Model performance was validated with ROC curves showing an AUC of approximately 0.79, indicating acceptable predictive power superior to random classification. Feature importance analysis highlighted that repayment status history, credit utilization, and payment consistency were strong predictors of default risk.

Key visualizations, including demographic breakdowns, repayment behavior over time, and distributions of late payments, facilitated a comprehensive understanding of risk factors. The project highlights the importance of combining demographic and behavioral data for robust credit risk assessment and demonstrates the utility of advanced machine learning techniques in financial analytics.

Project Milestones

Table 1. Project milestone

S.NO	Milestone	Detailed Description	Duration
1	Domain studies	1. Case study	(1 day) week 1
2	Data collection	• Studying requirements • Data collection • Business process chart • Data model (ER diagram)	Week 1 & 2
3	Data cleaning	• Cleaning (removing null values)	Week 3
4	Data preprocessing	• Feature engineering • Encoding categorical columns	Week 3
5	Exploratory data analysis (EDA)	• Univariate analysis • Bivariate analysis	Week 4
6	Data analysis	• Comparative analysis • Survival risk analysis • Prediction analysis • Dashboard analysis	Week 5
8	Interpretations and report writing	• Final report • Project presentation	Week 6

Business context

The global credit industry plays a critical role in driving economic activity. Credit cards, loans, and other financial products provide liquidity and purchasing power, but they also introduce significant risks for both lenders and borrowers. One of the primary challenges faced by financial institutions is the occurrence of loan defaults. Defaults can be triggered by various macroeconomic factors, including recessions, inflationary pressures, and unexpected events such as the COVID-19 pandemic, which caused widespread financial stress worldwide. (Md.Bokhtiar Hasan).

Research indicates that defaults often stem from borrower-specific issues such as over-indebtedness, poor financial literacy, and unstable income sources. At the institutional level, insufficient risk assessment models and reliance on outdated credit evaluation mechanisms also cater to rising default rates. Recent studies emphasize the importance of integrating advanced data analytics, artificial intelligence, and behavioral insights into credit risk management to more accurately identify borrowers at risk. (M.Gao)

Globally, credit default rates vary by region and loan type. According to the World Bank (2021), non-performing loan ratios averaged around 4.5% across global banking systems. This highlights the persistent challenge of maintaining credit quality, especially in times of economic uncertainty. Understanding the causes and consequences of defaults is now vital for economic sustainability of organizations. (world bank)

The problem of accurately distinguishing between potential defaulters and non-defaulters is highly relevant to the business. Traditional credit scoring methods, though widely used, often rely on limited variables and rule-based systems that may not capture the complex behavioral and financial patterns of borrowers. Misclassifications can result in significant financial losses, missed business opportunities, or reputational damage. Hence, advanced analytics and machine learning models are being increasingly adopted to strengthen risk assessment practices.

This project focuses on analyzing credit default prediction using demographic, financial, and repayment data to identify high-risk borrowers and improve institutional decision-making processes.

Business Process Chart

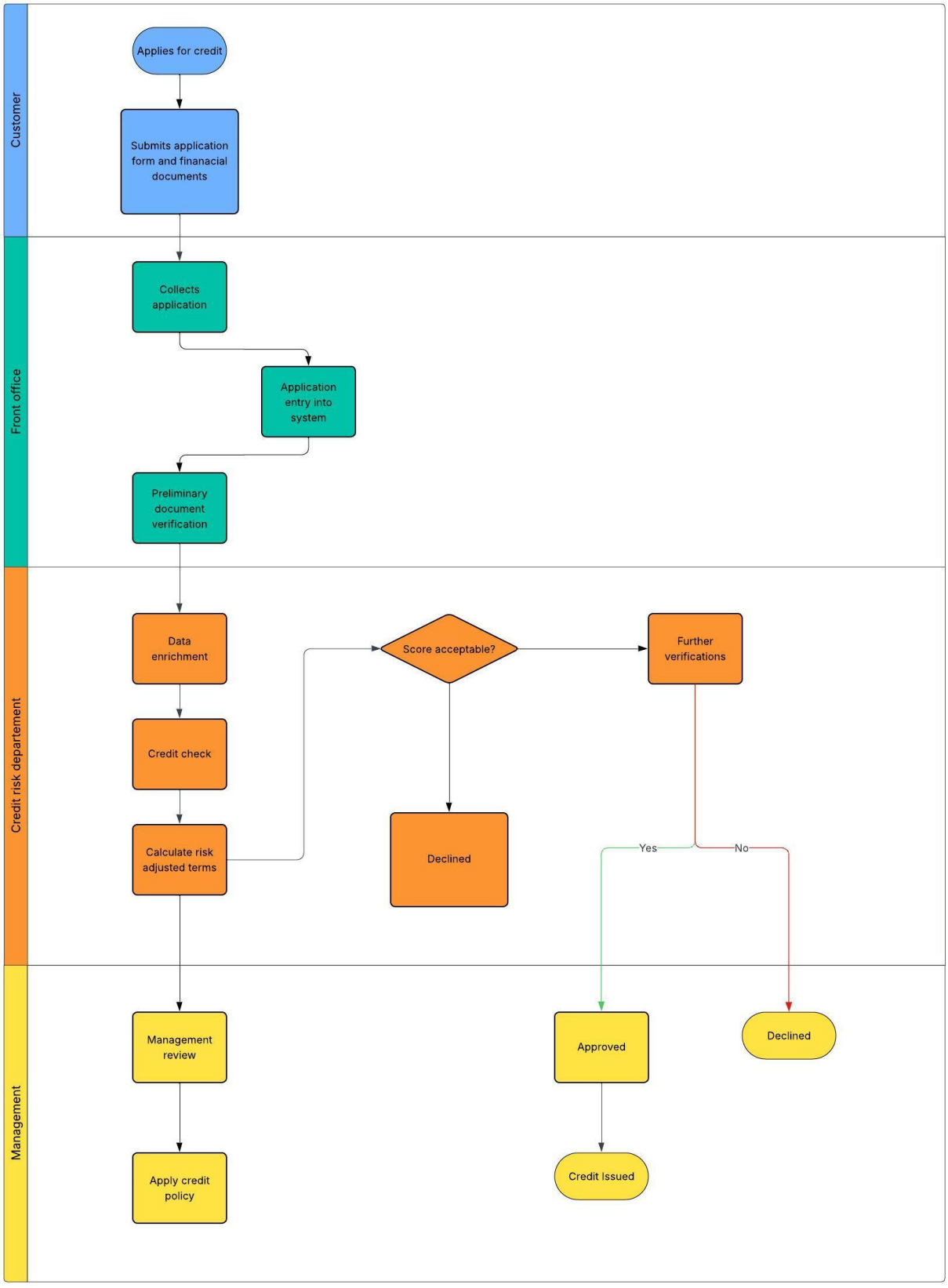


Figure 1 Business process chart

1. Customer

The process begins with the customer, who initiates the credit application by applying formally and submitting the required documents, such as application forms and financial records. This marks the starting point of the credit evaluation journey.

2. Front Office Stage

At this stage, the front office is responsible for handling and validating the customer's application.

- **Collection of application**: The submitted documents and forms are gathered.
- **Application entry into system**: All provided details are entered into the institution's processing system.
- **Preliminary**: The documents are checked for completeness, accuracy, and authenticity before being passed on for detailed review.

This ensures that only valid applications move forward in the process.

3. Credit Risk Department Stage

The application is then evaluated by the credit risk department, which undertakes detailed checks and risk assessment.

- **Data enrichment**: Additional information is gathered, such as external credit history, repayment records, or financial background.
- **Credit check**: The applicant's creditworthiness is analyzed, considering past payment behavior and risk indicators.
- **Risk-adjusted terms calculation**: Based on the credit profile, risk-adjusted terms such as interest rates, credit limits, or repayment schedules are proposed.

At this point, a decision node is introduced:

- If the score is acceptable, the process continues.
- If the score is not acceptable, the application is declined
- If further clarification is needed, the application undergoes additional verifications. If verification succeeds, the process moves forward; otherwise, the application is declined.

4. Management Stage

Applications that pass the credit risk checks are escalated to the management level for policy-based approval.

- **Management review**: It's reviewed by management authorities

- **Apply credit policy**: Management applies credit policy to decide the approval
- If approved, the status is updated, and the credit is issued to the customer.
- If rejected at this stage, the application is marked as declined.

Entity Relationship diagram (ER)

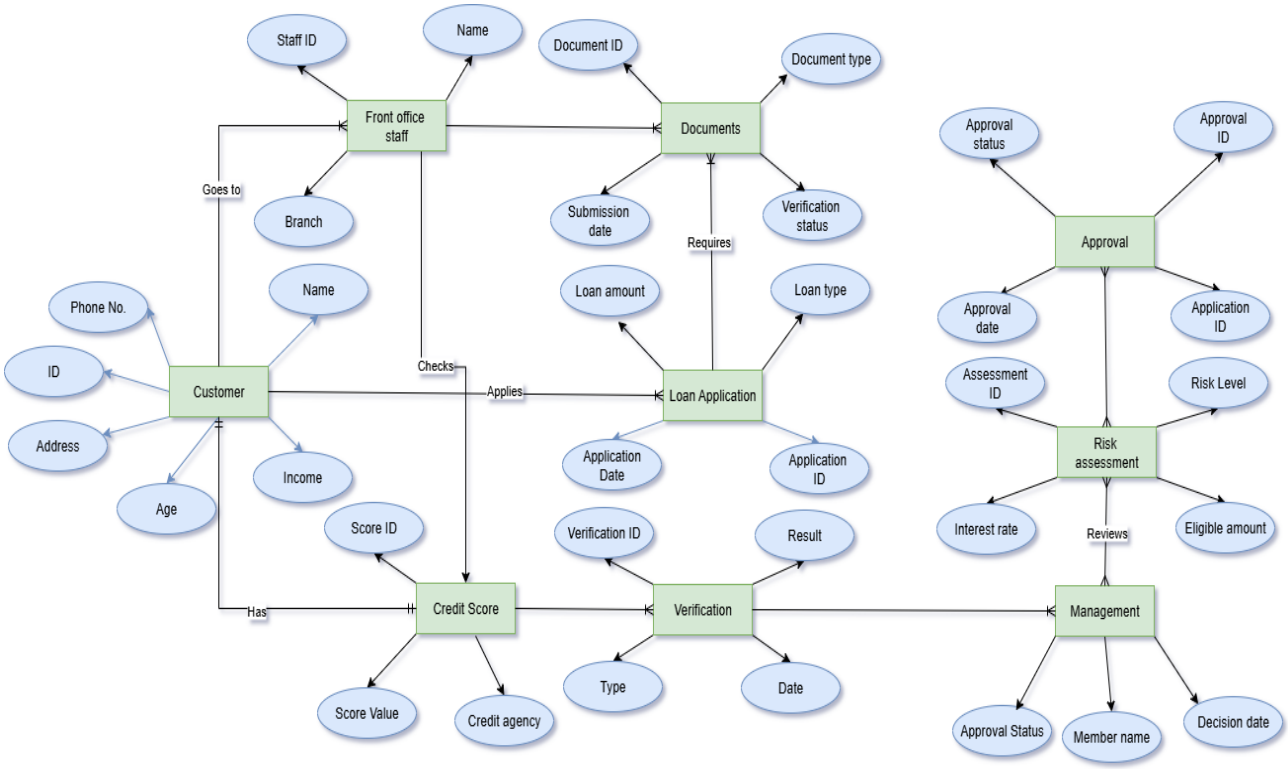


Figure 2 Entity-Relationship diagram

The ER diagram tells the story of how a loan application flows through a bank's credit approval system. The process begins when a customer walks into a branch and approaches the front office staff to apply for a loan. The customer provides personal details such as age, income, address, and phone number, along with supporting documents like proof of identity and income statements. These documents are submitted with the loan application, which records information such as the loan type, loan amount, and application date.

To ensure credibility, the application undergoes a verification process, where documents are checked, validated, and linked to the customer's credit score provided by a credit agency. Once verified, the application is passed on to the risk assessment team, which evaluates the borrower's risk level, calculates the eligible loan amount, and determines an appropriate interest rate. Based on this assessment, the case moves to the approval stage, where loan officers record the approval status and date.

Finally, the application reaches management, which reviews all findings and makes the ultimate decision on whether to grant or reject the loan. In this way, the system ensures that every loan application is carefully verified, assessed for risk, and fairly decided before funds are disbursed, balancing customer needs with the bank's financial safety.

Problem statement

'Predicting loan default using borrower's data'

The credit industry plays a crucial role in global economic growth, but rising default rates pose significant challenges to financial institutions. Default risk is influenced by multiple factors, including income levels, marital status, education, employment stability, and existing debt obligations. In this project, we aim to analyze and predict default probabilities by examining these socioeconomic and financial indicators. The ultimate goal is to develop a data-driven model that improves risk assessment, enables better lending decisions, and reduces financial losses for institutions while fostering responsible credit access for individuals.

Objectives

- To identify key demographic, financial, and behavioral factors that significantly influence the likelihood of credit default among borrowers.
- To build and evaluate predictive models using statistical and machine learning techniques to estimate the probability of default with high accuracy.
- To provide actionable insights for financial institutions that can enhance credit risk management, reduce default rates, and support fair and responsible lending practices.

Data Description

Data Model (Schema diagram)

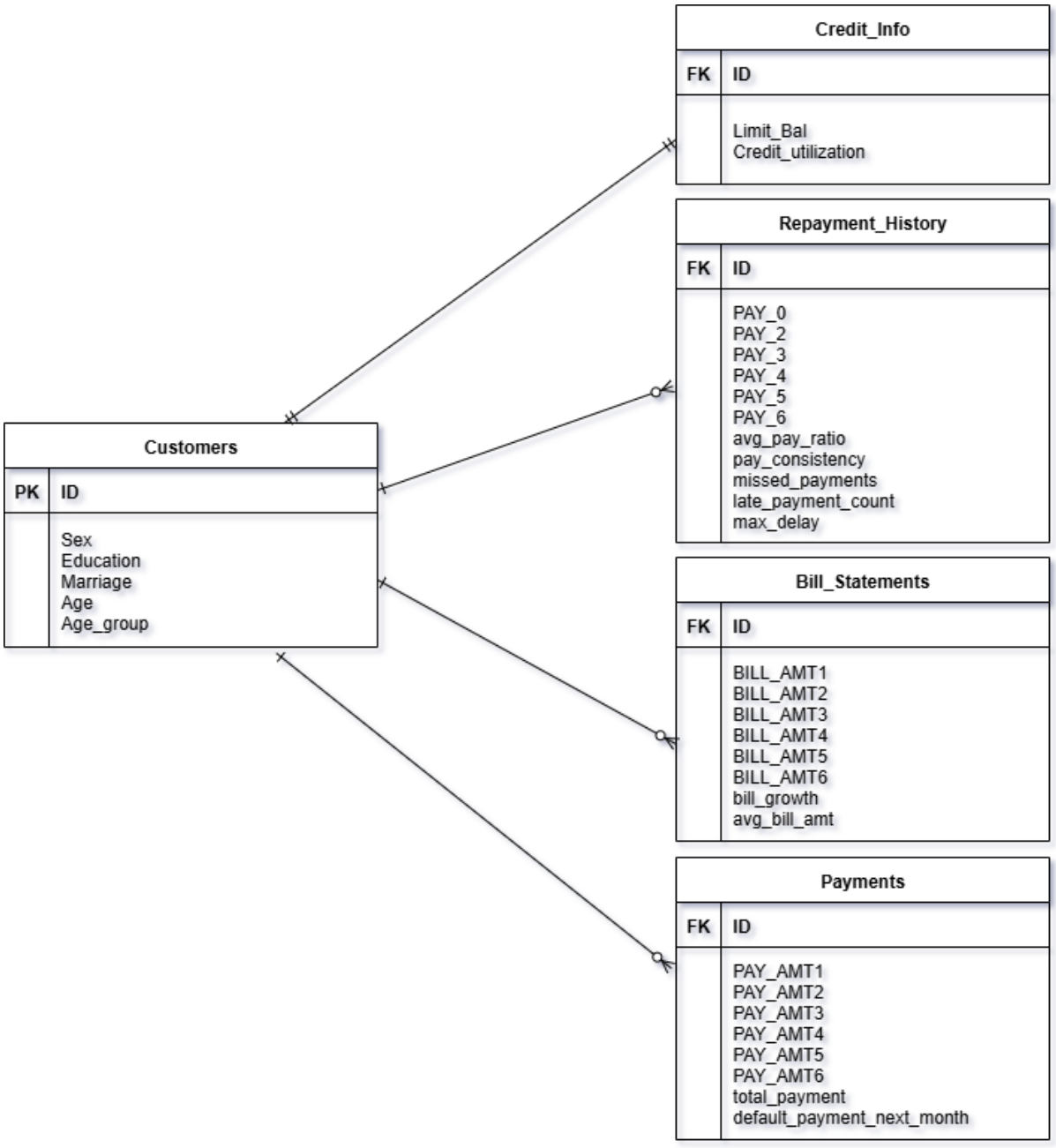


Figure 3 Schema diagram

Table 2 Data dictionary

Column	Data type	Description
ID	int64	Unique identifier of each client.
LIMIT_BAL	int64	Amount of given credit (NT dollars) – includes both individual and family/supplementary credit.
SEX	int64	Gender (1 = male, 2 = female).
EDUCATION	int64	Education level (1 = graduate school, 2 = university, 3 = high school, 4 = others).
MARRIAGE	int64	Marital status (1 = married, 2 = single, 3 = others).
AGE	int64	Age of client in years.
PAY_0	int64	Repayment status in September 2005 (-1 = pay duly, 1 = delay 1 month, 2 = delay 2 months, …, 8 = delay 8 months, 9 = 9+ months).
PAY_2	int64	Repayment status in August 2005 (same coding as PAY_0).
PAY_3	int64	Repayment status in July 2005 (same coding as PAY_0).
PAY_4	int64	Repayment status in June 2005 (same coding as PAY_0).
PAY_5	int64	Repayment status in May 2005 (same coding as PAY_0).
PAY_6	int64	Repayment status in April 2005 (same coding as PAY_0).
BILL_AMT1	int64	Amount of bill statement in September 2005 (NT dollars).
BILL_AMT2	int64	Amount of bill statement in August 2005.
BILL_AMT3	int64	Amount of bill statement in July 2005.
BILL_AMT4	int64	Amount of bill statement in June 2005.
BILL_AMT5	int64	Amount of bill statement in May 2005.

BILL_AMT6	int64	Amount of bill statement in April 2005.
PAY_AMT1	int64	Amount paid in September 2005 (NT dollars).
PAY_AMT2	int64	Amount paid in August 2005.
PAY_AMT3	int64	Amount paid in July 2005.
PAY_AMT4	int64	Amount paid in June 2005.
PAY_AMT5	int64	Amount paid in May 2005.
PAY_AMT6	int64	Amount paid in April 2005.
Feature engineered columns		
default payment next month	int64	Default payment (1 = yes, 0 = no). Target variable.
credit_utilization	float64	Ratio of total bill amount to credit limit.
avg_pay_ratio	float64	Average payment ratio across months.
bill_growth	float64	Rate of growth/decline in bill amounts over time.
pay_consistency	float64	Measure of consistency in payments (e.g., variance or stability of payments).
avg_bill_amt	float64	Average bill statement across months.
total_payment	int64	Sum of payments across all months (NT dollars).
missed_payments	int64	Count of months with no payments despite having a bill.
late_payment_count	int64	Number of months with delayed payments.
max_delay	int64	Maximum recorded delay (in months).
age_group	category	Age grouped into bins (e.g., 20–29, 30–39, etc.).

Analytical approach

Requirements Specifications

User Requirements

The end-users of this project require:

- A predictive system to assess the probability of a customer defaulting on credit.
- An easy-to-understand output (e.g., default / non-default classification with probability score).
- Visualization of model performance (ROC curve, confusion matrix, accuracy metrics).
- The ability to interpret which features (income, marital status, loan history, etc.) influence default likelihood.
- A simple interface or report to support decision-making for credit approvals.

Functional Specifications

1. Core Features

Data Integration:

- Raw Data: (default of credit card clients.xlsx)
- Processed Data: (data.csv)
    - Dataset after handling missing values, encoding categorical variables, and applying feature engineering.

Model Building & Prediction:

- Predicting default risk (Yes/No) using XGBoost Classifier.
- Dataset split into 80:20 ratio for training and testing.
- Evaluation metrics include Accuracy, Precision, Recall, F1-Score, and AUC-ROC.

Default Risk Suggestion:

- Based on borrower characteristics (income, credit utilization, marital status, education, repayment history), The system suggests whether the borrower is high-risk or low-risk.

Dashboard Plan:

- Visualize borrower demographics (age, gender, education, marital status).
- Plot repayment history trends and default probabilities.
- Filters for key features (income group, education level, marital status, repayment delays).

2. Analytical Capabilities

Descriptive Analysis:

- Univariate and bivariate EDA on borrower data.
- Distributions by default status, income group, education, etc.

Predictive Analytics:

- Machine learning models trained to classify default risk.
- Metrics used: Accuracy, Precision, Recall, F1, AUC-ROC.
- Feature importance analysis to identify top predictors.

3. User Roles

Internal Use (Student/Researcher):

- Running and validating ML models manually.
- Testing model performance for academic evaluation.
- All operations run using Python scripts, Jupyter Notebook, or Google Colab.

Financial Analysts (Future Scope):

- Use dashboards for decision-making.
- Access automated predictions for new borrower data.

4. Interface / Ease of Use

- Data preprocessing and feature engineering handled via Python scripts.
- ML models run with minimal code input.
- Dashboards (Power BI) provide interactive visualizations.

5. KPIs (Key Performance Indicators)

Model Output:

- Default payments: Default / Non-Default.
- Evaluation metrics: Accuracy, Precision, Recall, F1-Score, AUC-ROC.

Comparative Analysis:

- Default rate across demographics (age, marital status, education).
- Repayment trends by income group and credit utilization.

Key Variables:

- Demographics: Gender, Age, Education, Marital Status.
- Financials: Income, Credit Utilization Ratio, Payment History.

- Behavioral: Past defaults, payment delays, repayment consistency.

System Specifications

1. Hardware

System:

- CPU: i3 or higher (assumed)
- RAM: 8–16 GB
- No GPU required
- Storage: 100–300 MB CSVs, handled locally

2. Software

Languages and Tools Used:

- Python (main language)
- Libraries: pandas, numpy, sklearn, matplotlib, seaborn, lifelines

Environment:

- Google Colab

Visualization:

- matplotlib, seaborn, numpy, lucid charts, draw.io and Power BI

Database:

- No database used, CSVs only.
- No database used, CSVs only.

3. Machine Learning

Libraries Used:

- Scikit-learn: For train/test split, encoding, classification model, evaluation metrics
- XGBoost classification

Activity chart

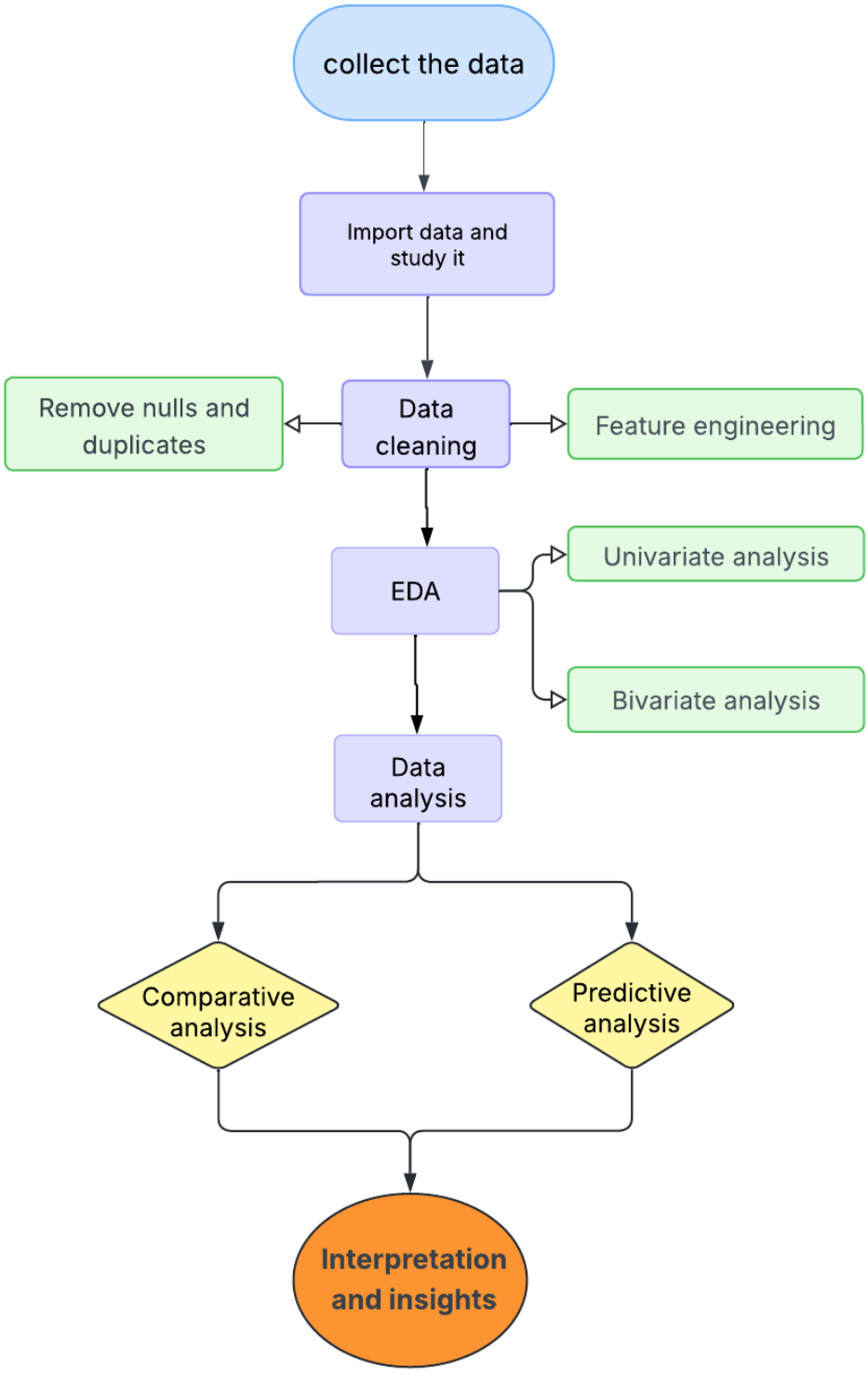


Figure 4 Activity chart

The activity chart describes the line of action of the project starting from data collection to interpretation and insights.

1. Data collection

The dataset named *Default of Credit Card Clients (Taiwan)* was collected from the UCI Machine Learning Repository. It contains observations of cardholders with demographic attributes, historical repayment status, monthly bill amounts, and monthly payment amounts from Taiwan. The target variable is default payment next month (1 = default, 0 = non-default).

2. Studying the data

The data was studied to plan out the roadmap for analysis.

3. Data Cleaning

This step involves removing or fixing mistakes in the data and preparing the data to make it suitable for analysis. The dataset contained no null values in the original dataset, some null values arised after one hot encoding when categorical encoded values were reversed.

Feature engineering was performed to improve predictive power and interpretability, including derived variables such as credit utilization ratio, pay ratios, bill growth, and repayment consistency. These engineered features were created using financial domain knowledge and provide a more realistic representation of borrower risk behavior.

5. EDA (Exploratory Data Analysis)

Exploratory Data Analysis (EDA) was conducted to identify patterns, relationships, and anomalies in the data before modeling.

- **Univariate analysis** was performed to study individual variable distributions (e.g., age distribution, bill amounts).
- **Bivariate analysis** compared features with default status (e.g., marital status vs. default rates).
- **Multivariate analysis** included correlation matrices and trend analyses to detect interactions between variables and to check for multicollinearity. This process was justified because it helps validate assumptions, provides business insights, and informs model design.

5. Data Analysis

The data analysis included 2 types of analysis:

- **Comparative Analysis**: This was used to compare default rates among various demographics and financial factors.

- **Predictive Analysis**: The data was used to predict the default payments of customers by using machine learning model (XGBoost classification Model).
- **Model Choice: XGBoost Classifier**

  The predictive model chosen was the XGBoost classifier. This was selected because:

  - It handles both categorical and continuous data efficiently.
  - It consistently outperforms simpler models (e.g., logistic regression, decision trees) on tabular financial data.
  - It is computationally efficient and scalable, making it ideal for both research and practical financial decision-making.
- **Evaluation Metrics**

  Multiple metrics were used to ensure robust evaluation of the model:

  - Accuracy measures overall correctness but may be misleading in imbalanced datasets.
  - Precision and Recall are essential in credit risk, as banks must balance between avoiding false positives (rejecting safe borrowers) and false negatives (approving risky borrowers).
  - F1-Score provides a balance between precision and recall.
  - AUC-ROC measures the model's ability to separate defaulters from non-defaulters, which is especially important in risk prediction.

6. Interpretations and insights:

Interpretations were drawn from the analysis and visualizations were made to relay the insights gained in a simpler format.

Extract, Transform and Load pipeline (ETL)

1. Extract:

The data was collected from UCI machine learning repository. It consists of 30,000 rows and 24 columns.

2. Transform:

The data was preprocessed to make it suitable for analysis. It includes data cleaning, feature engineering, and encoding.

Data Cleaning

The column names of the dataset were added by using the first row as header. There were no missing values in original dataset. However, when data was reverse encoded and feature engineered, few null values arised. After removal of null values, we were left with 29,593 rows and 34 columns. The categorical columns were one hot encoded for dashboard and comparative analysis.

Feature engineering is the process of selecting, modifying, or creating new input variables (features) from raw data to improve the performance of machine learning models. It involves transforming data into a format that better highlights patterns and relationships that the model can learn from.

3. Load:

The dataset is loaded using pandas from python.

Data Analysis

Exploratory Data Analysis (EDA):

After cleaning exploratory data analysis (EDA) was performed on the original dataset.

1. Summary statistics

Summary statistics gives description of numerical data using measures of central tendency and dispersion.

Table 3 summary statistics

Variable	Count	Mean	Std Dev	Min	25%	50%	75%	Max
ID	29,599	14,971.34	8,660.10	1	7,473.5	14,953	22,462	30,000
LIMIT_BAL	29,599	167,540.92	129,942.48	10,000	50,000	140,000	240,000	1,000,000
SEX	29,599	1.60	0.49	1	1	2	2	2
EDUCATION	29,599	1.82	0.71	1	1	2	2	4
MARRIAGE	29,599	1.56	0.52	1	1	2	2	3
AGE	29,599	35.46	9.21	21	28	34	41	79
Default Next Month	29,599	0.22	0.42	0	0	0	0	1
Credit Utilization	29,599	0.42	0.41	-0.62	0.02	0.31	0.83	6.45
Avg Pay Ratio	29,599	3.81	211.06	-1502.0	0.04	0.09	0.60	27,000

Bill Growth	29,599	529.83	7470.55	-2640	-0.06	0.00	0.11	470,400
Pay Consistency	29,599	5775.87	15,003.24	0	615.36	1413.39	4114.50	650,098
Avg Bill Amt	29,599	44,822.92	63,131.41	-56,043	4,765	20,915.83	56,881.58	877,314
Total Payment	29,599	31,549.91	60,819.35	0	6,691.5	14,353	33,426	3,764,066
Missed Payments	29,599	1.23	1.72	0	0	0	2	6
Late Payment Count	29,599	0.84	1.56	0	0	0	1	6
Max Delay	29,599	0.44	1.35	-2	0	0	2	8

2. Univariate Analysis

Univariate analysis refers to analysis of a single variable. The goal is to understand the distribution, central tendency (mean, median, mode), and spread (variance, standard deviation) of that one variable.

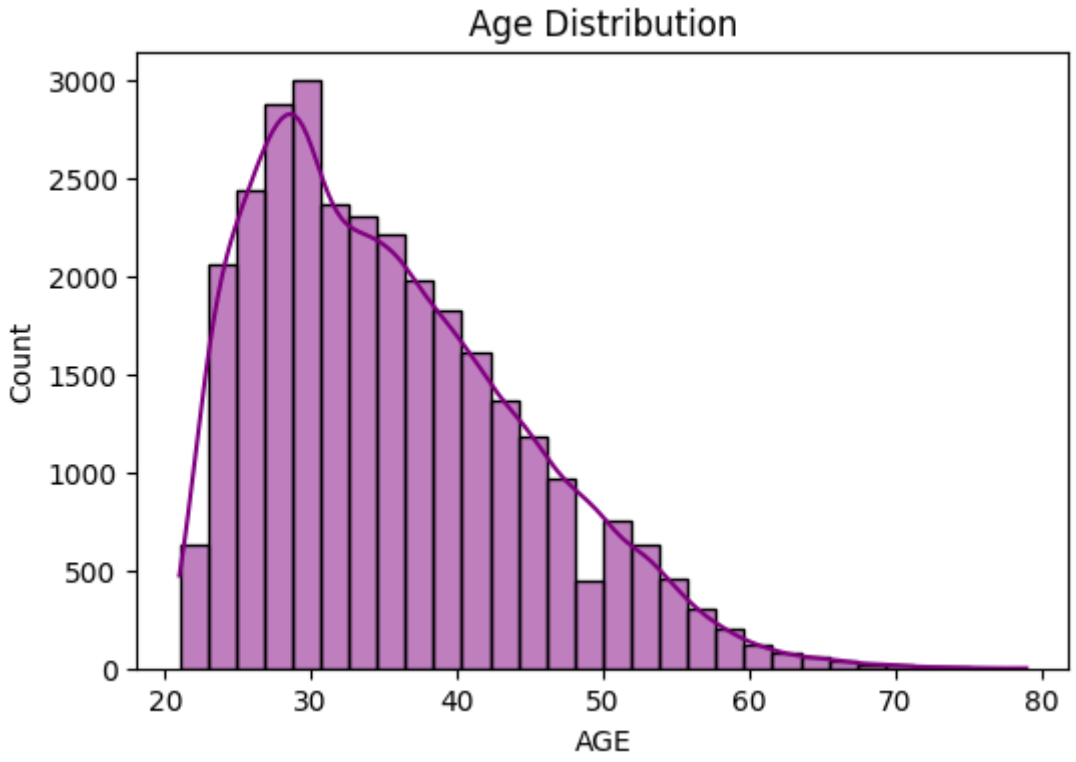


Figure 5 Histogram of age

This histogram with density curve shows the age distribution of individuals in your dataset:

- The majority of people fall in the mid-20s to early 30s range.
- The distribution is right-skewed (positively skewed), meaning more people are younger, and fewer are older.
- The peak (mode) is around 28–30 years, where the count of individuals is the highest (around 3,000).
- After 35 years, the frequency steadily declines, and very few individuals are aged above 60.
- This suggests that your dataset is dominated by young to middle-aged individuals, which might influence credit behavior, repayment ability, or default risks.

3. Bivariate Analysis

Bivariate analysis refers to analysis of the relationship between two variables. The goal is to find correlations, trends, or associations between two variables.

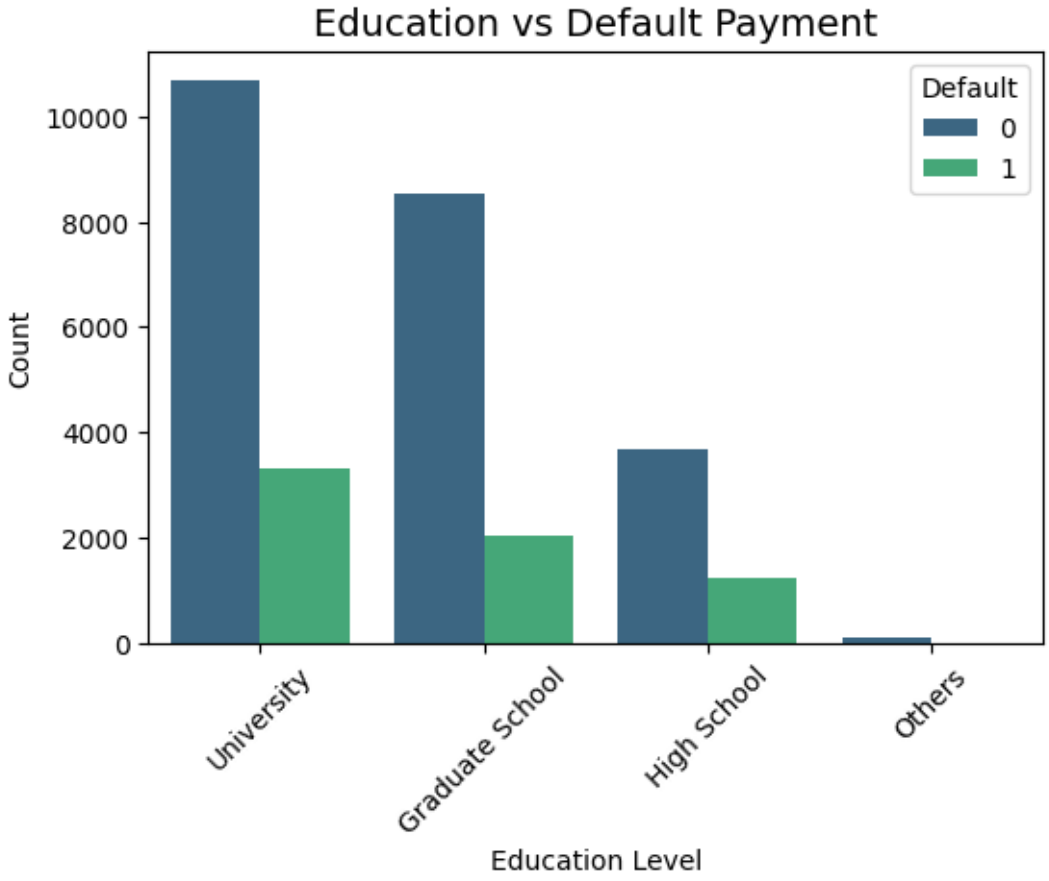


Figure 6 default payments by education level

- University graduates make up the largest group of credit holders, with more than 10,000 non-defaulters and about 3,000 defaulters.
- Graduate school customers are the second-largest group, with around 8,500 non-defaulters and 2,000 defaulters, again showing higher repayment discipline compared to defaults.
- High school customers have fewer accounts overall (around 3,700 non-defaulters and 1,200 defaulters), but their default rate seems higher compared to graduate/university groups.
- Others represent a very small portion of the dataset and don't significantly impact the overall trend.

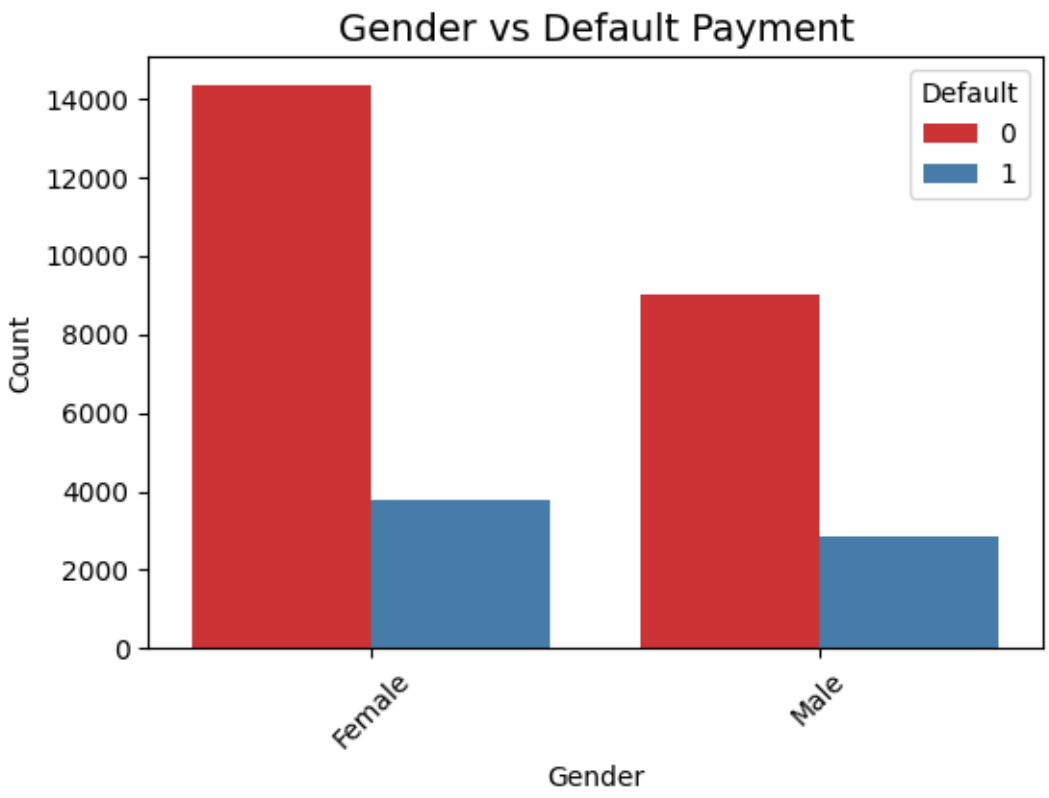


Figure 7 Default payment by gender

Female Borrowers

- Larger group compared to males in the dataset.
- Majority did not default (Default = 0).
- However, the number of female defaulters (Default = 1) is still significant, but less than non-defaulters.
- Suggests that women generally maintain better repayment behavior, though absolute defaults are higher due to larger population size.

Male Borrowers

- Smaller group compared to females.
- Most males also did not default, but the number of defaults is proportionally close to female defaults.
- Indicates that while fewer males are present, their default rate seems relatively higher when compared proportionally to their total count.

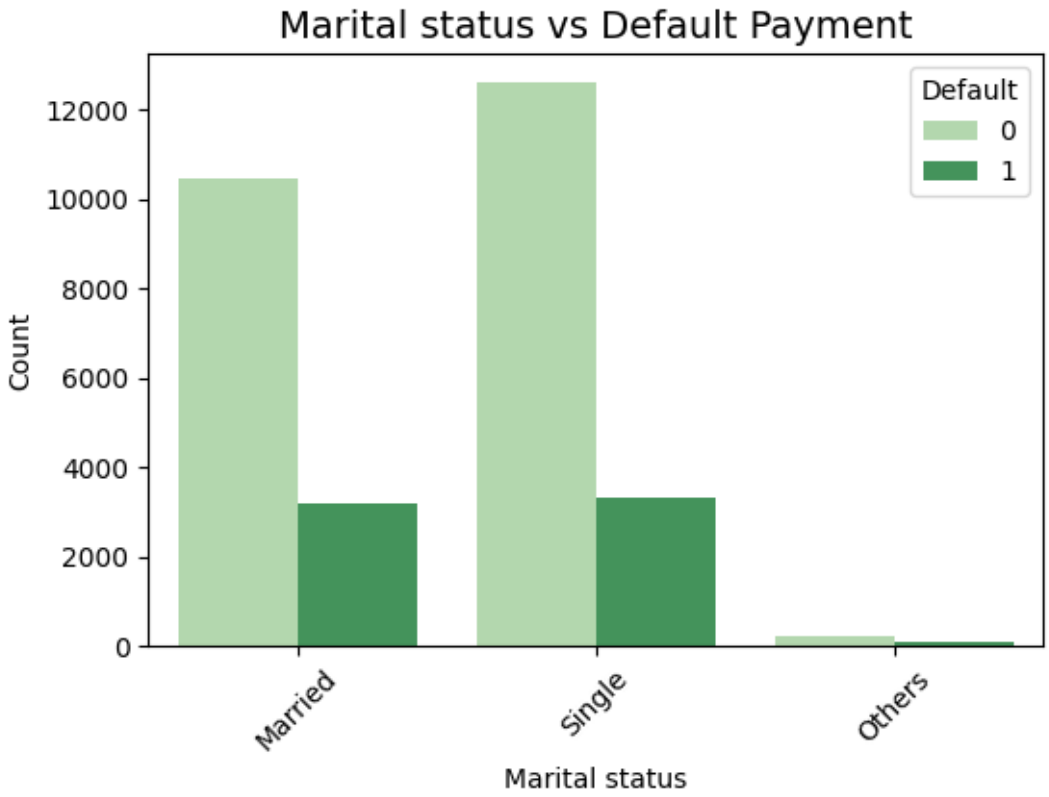


Figure 8 Default payment by Marital status

Married Borrowers

- Large group in the dataset.
- Majority did not default (Default = 0).
- Defaults are present, but fewer compared to non-defaulters.
- Suggests that being married may provide some financial stability, reducing default risk.

Single Borrowers

- Slightly larger group than married borrowers.
- Also, the majority did not default.
- Defaults are slightly higher than those of married borrowers, both in count and proportion.
- Indicates that single individuals might have a relatively higher risk of default compared to married ones.

Others

- Very small group.
- Defaults are negligible here.
- Too few data points to draw meaningful conclusions.

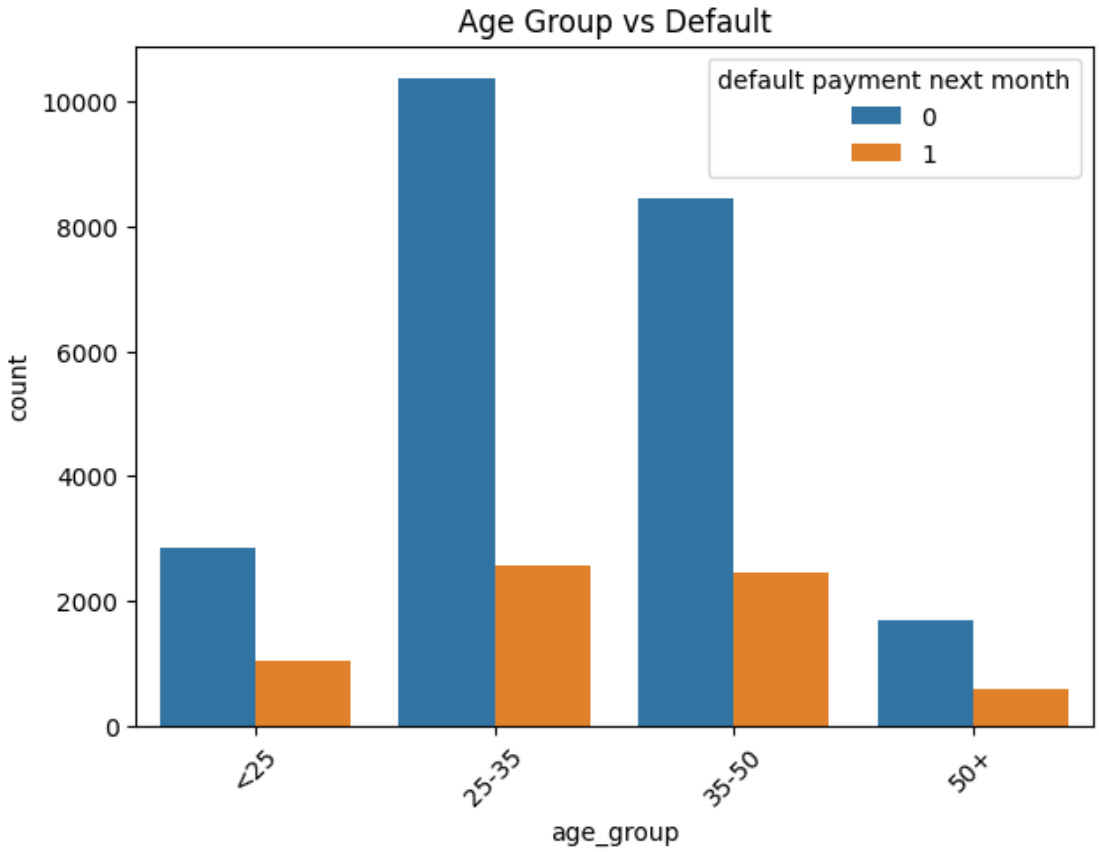


Figure 9Default payment by age group

Age < 25

- Smallest group in the dataset.
- Defaults are present but relatively lower in absolute terms.
- Younger borrowers may have less credit exposure but also higher risk due to limited financial stability.

Age 25–35

- Largest group of borrowers.
- Majority did not default, but this group also shows the highest absolute number of defaults.
- Suggests that this is the most active borrowing group, but also carries a notable repayment risk.

Age 35–50

- Second largest group.
- Non-defaulters dominate, but defaults are also significant, nearly close to the 25–35 group.

- Indicates mid-career individuals still face repayment challenges, possibly due to family/financial commitments.

Age 50+

- Smaller group compared to younger age brackets.
- Defaults are present but fewer in number.
- Suggests older borrowers tend to default less, likely due to greater financial stability or lower borrowing activity.

4. Multivariate analysis:

Multivariate analysis refers to analysis involving more than two variables simultaneously. The goal is to understand complex interactions and combined effects of multiple variables on an outcome.

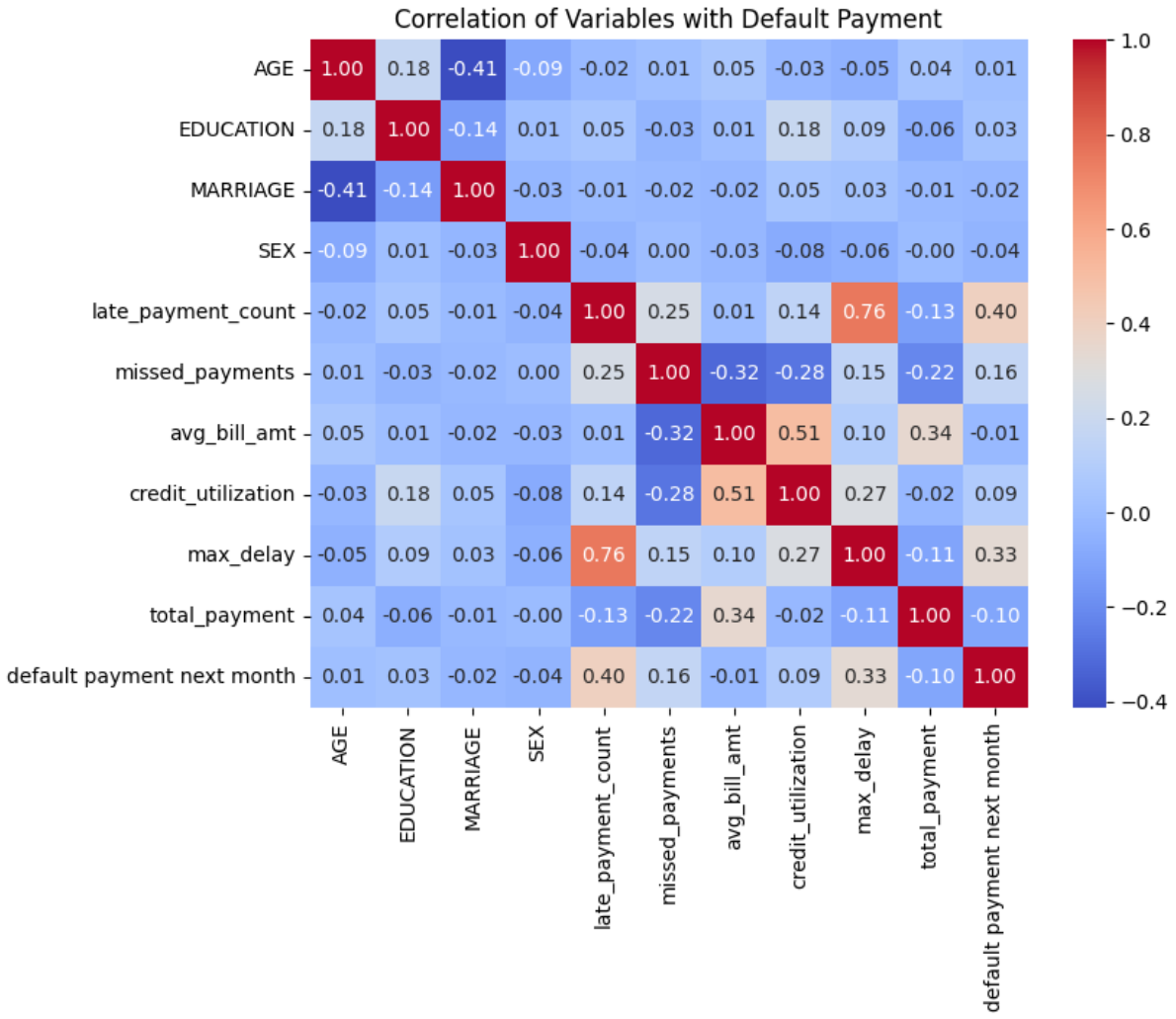


Figure 10 Correlation matrix

Strongest Correlations with Default Payment (target variable):

- Late Payment Count (0.40): Higher late payment counts are strongly linked with a higher chance of default.
- Max Delay (0.33): Longer maximum payment delays increase default probability.
- Missed Payments (0.16): More missed payments are positively correlated with defaults.
- Credit Utilization (0.09): Higher credit usage slightly increases the likelihood of default.

These financial behavior variables are the most important predictors of default.

Weak / Negligible Correlations with Default Payment:

- Age (0.01)
- Education (0.03)
- Marriage (-0.02)
- Sex (-0.01)

  These demographic factors have almost no significant correlation with default behavior, meaning borrower profile details don't predict defaults well compared to repayment behavior.

Other Strong Relationships Among Variables (not directly with default):

- Late Payment Count & Max Delay (0.76): Strongly correlated (borrowers with more late payments usually have higher max delays).
- Missed Payments & Avg Bill Amount (-0.32): Negative correlation, meaning people with higher bills may not miss payments as often.
- Missed Payments & Credit Utilization (0.51): Strongly correlated, showing that high utilization often leads to missed payments.

Comparative Analysis

Methodology

Step 1: Load the Dataset

- Imported the dataset into a pandas DataFrame (pandas library).
- Inspected the data using .info(), .describe(), and .head() to understand column types and missing values.

Step 2: Import Required Packages

- **pandas** – is used for data cleaning, manipulation, and reshaping.
- **numpy** – is used for mathematical operations.
- **matplotlib** & **seaborn** -Is used for data visualization

Step 3: Data Cleaning & Feature Engineering

- Converted columns like PAY_0, PAY_2, … into more interpretable labels (PAY_Sep2005, PAY_Aug2005, etc.).
- Renamed bill and payment columns (BILL_AMT1 = BILL_Sep2005, etc.) for clarity in time series analysis.
- Created new features like credit_utilization, missed_payments, late_payment_count, and max_delay to capture repayment behavior.
- Handled missing values in derived features using imputation where necessary.

Step 4: Comparative Analysis

- **Categorical comparisons:**
    - Compared late payments and missed payments across gender, marital status, and age groups.
    - Used seaborn barplots and countplots to visually interpret repayment behavior across demographic groups.
- **Numerical comparisons:**
    - Compared distributions of credit utilization, average bill amounts, total payments, and late payment counts between defaulters vs non-defaulters.
    - Used boxplots, violin plots, and KDE plots for visualization.

Step 5: Time Series Analysis

- Converted PAY, BILL_AMT, and PAY_AMT columns into time-series features by month.
- Plotted line charts to observe trends over six months:
    - Default rate per month
    - Percentage of missed payments
    - Late payments per month

Step 6: Interpretation of Results

- Identified which groups had higher risk of default.
- Observed how bill amounts and payments evolved month by month, and how they related to defaulting behavior.
- Derived insights into repayment patterns (e.g., more missed payments in certain months, correlation between high credit utilization and default).

Results:

Average late payments by demographics:

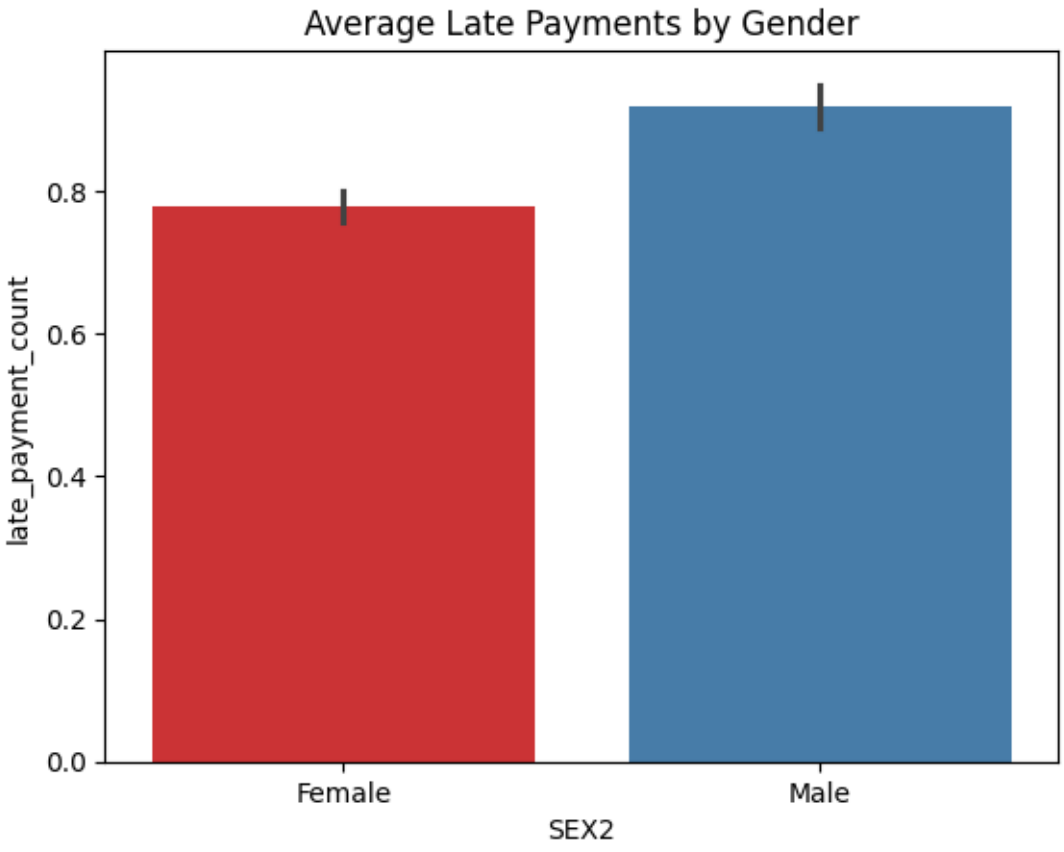


Figure 11 Average late payments by gender

1. **Females**
   - Average late payments are around 0.78.
   - This indicates that, on average, women tend to have fewer late payments compared to men.
2. **Males**
   - Average late payments are higher, around 0.92.
   - Suggests that men are slightly more likely to miss payment deadlines than women.

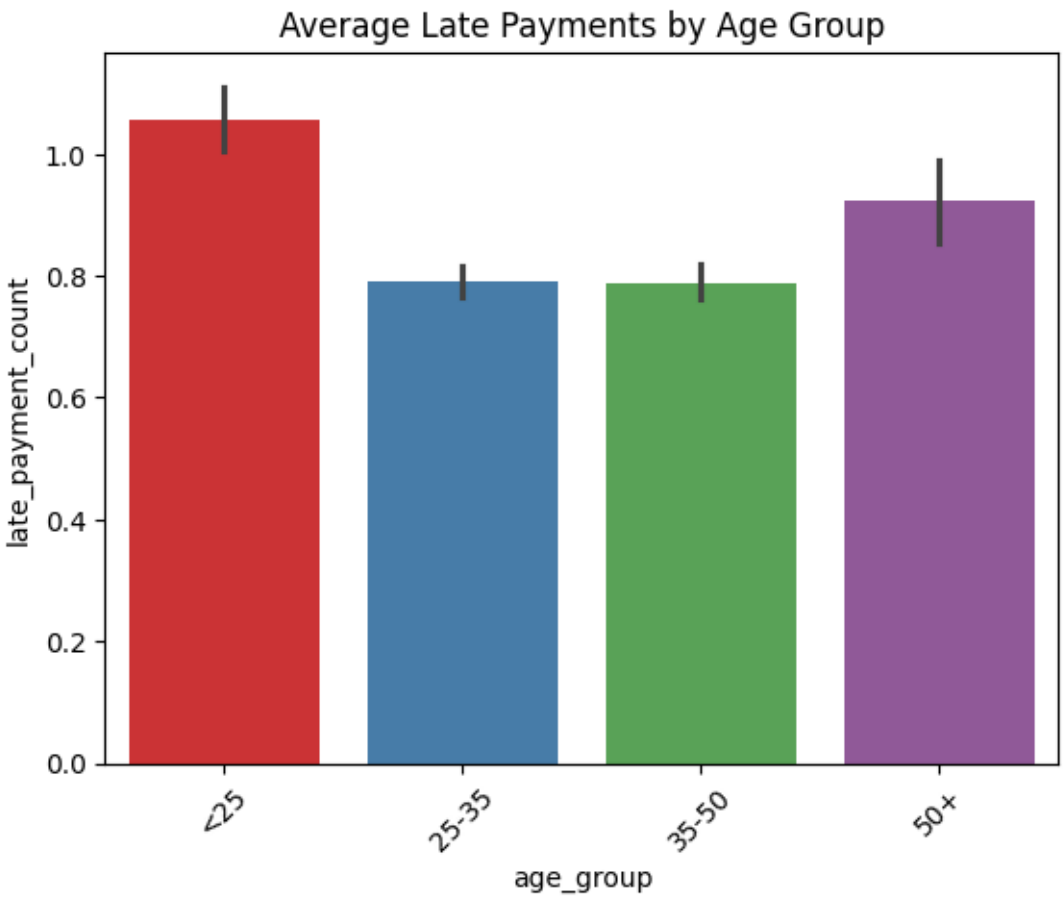


Figure 12 Average late payments by age group

1. **Age < 25**
   - Highest average late payments (around 1.05).
   - Suggests that younger borrowers are the least disciplined in making timely payments, possibly due to limited financial experience or unstable income.
2. **Age 25–35**
   - Average late payments drop to around 0.80.
   - Indicates this group is more financially responsible compared to younger borrowers.
3. **Age 35–50**
   - Very similar to the 25–35 group, with average late payments also around 0.80.

- Suggests consistency in repayment behavior during mid-career years.

4. **Age 50+**
   - Average late payments rise again to about 0.92.
   - Indicates that older borrowers tend to delay payments more than middle-aged borrowers, though still better than those under 25.

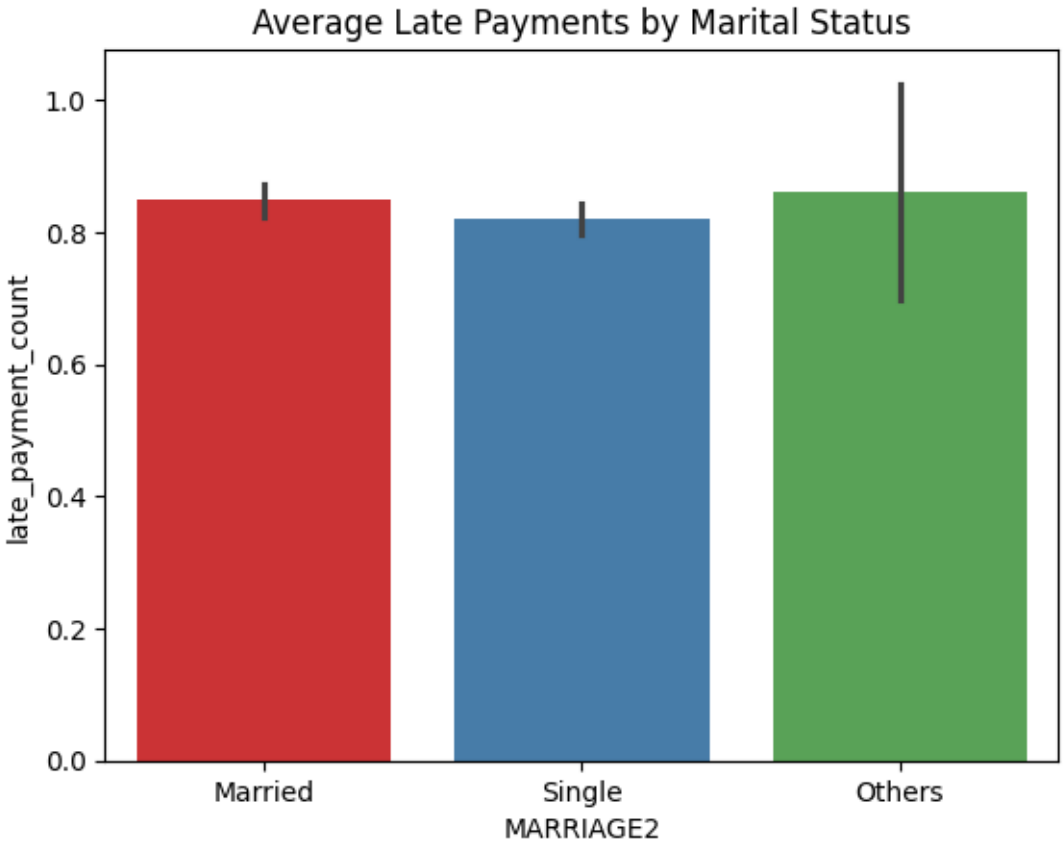


Figure 13 Average late payments by marital status

1. **Married Borrowers**
   - Average late payments are around 0.85.
   - Slightly higher than singles, suggesting that married individuals may face more financial commitments (family expenses, loans, etc.), leading to occasional delays.
2. **Single Borrowers**
   - Average late payments are about 0.82, the lowest among groups.
   - Indicates better repayment discipline, possibly because they have fewer dependents or obligations.
3. **Others**
   - Average late payments are about 0.86, but with a large variation (wide error bar).
   - This suggests that repayment behavior is inconsistent in this group—some manage payments well, while others struggle significantly.

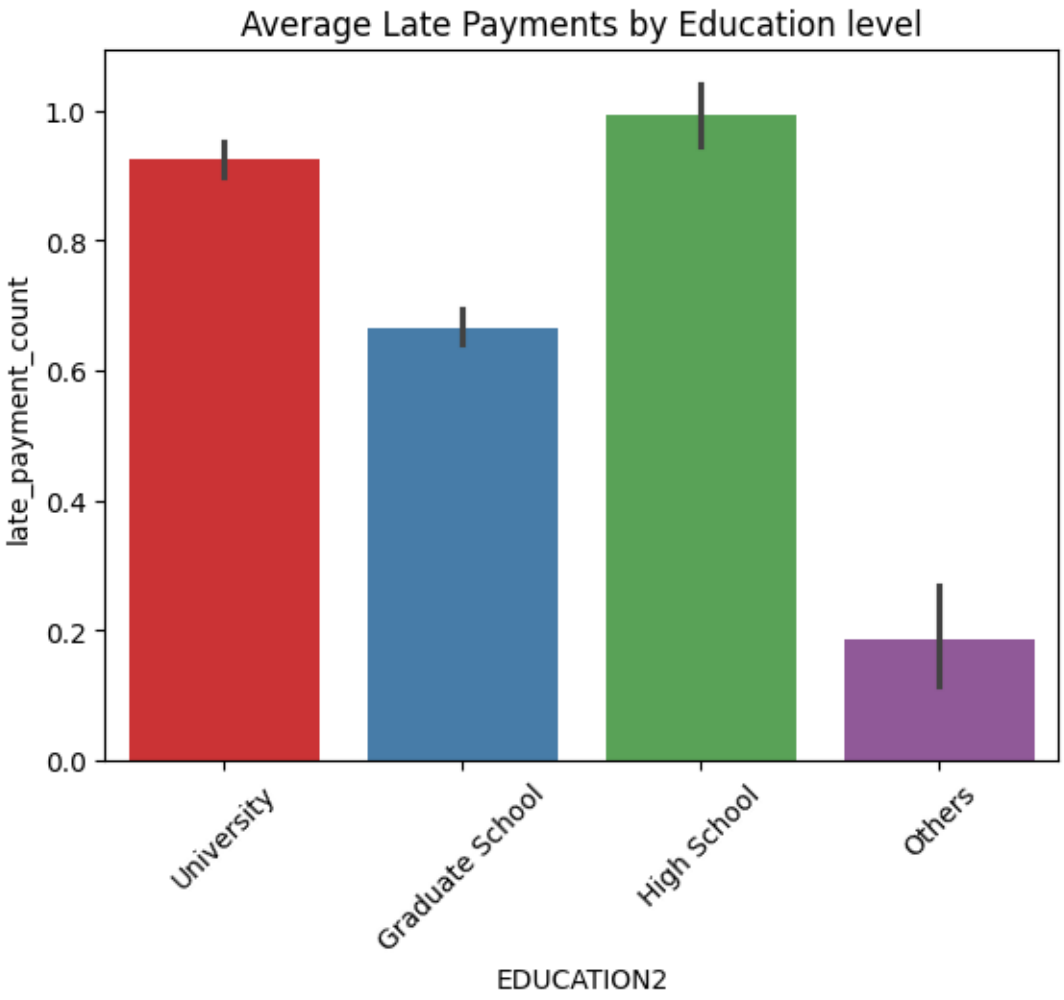


Figure 14 Average late payments by education level

High School

- Has the highest average late payment count (~1.0).
- This suggests that borrowers with only high school education are more prone to missing or delaying payments.
- Possible reason: lower earning potential or unstable financial situations compared to higher-educated groups.

University

- Shows an average late payment count of around 0.9, which is quite high, close to high school.
- Even though university-educated individuals generally have better earning prospects, their late payment behavior is still significant.
- Could be linked to larger debt obligations (e.g., student loans, higher credit usage).

Graduate School

- Average late payment count is around 0.64, the lowest among the major categories.
- Indicates that individuals with advanced degrees tend to manage credit and repayments better.
- Likely due to higher, more stable incomes and financial literacy.

Others

- Very low late payment average (~0.2), but the error bar is large (indicating small sample size).
- Since very few people fall into this category, we should be cautious in interpreting it—data may not be representative.

Average missed payments by demographics:

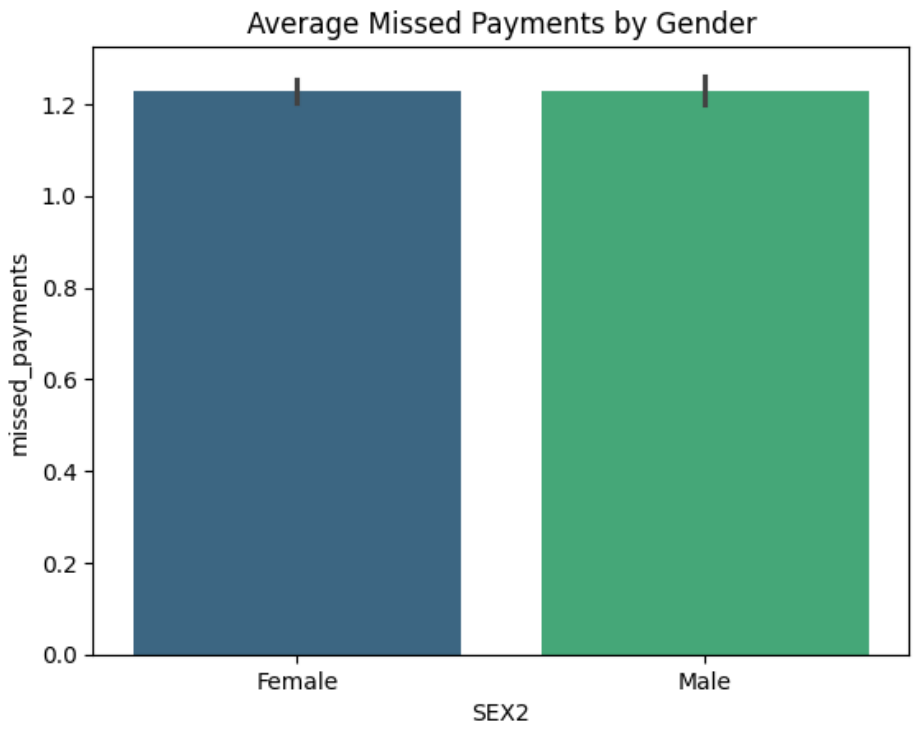


Figure15 15 Average missed payments by gender

- **Females**
  - Average missed payments: ~1.23
  - The error bar is very small, meaning the data is consistent.
- **Males**

- Average missed payments: ~1.22
- Almost the same as females, again with very little variation.

- There is no significant difference between genders in terms of missed payments.
- Both men and women have almost identical average missed payments (just over 1 per account).
- The error bars overlap completely → confirms that gender does not play a meaningful role in predicting missed payments.

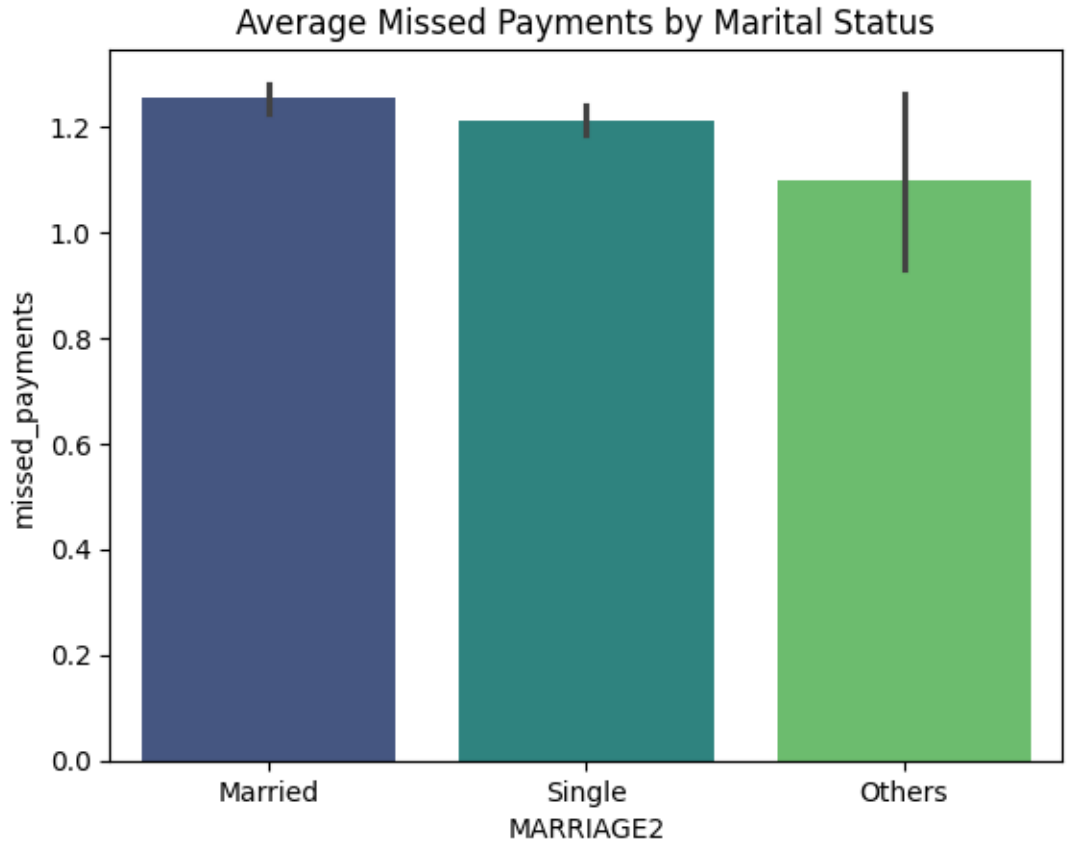


Figure16 16 Average missed payments by marital status

Married

- Individuals classified as *Married* recorded an average of approximately 1.25 missed payments.
- This represents the highest level of missed payments among the groups.

Single

- Single individuals had an average of around 1.22 missed payments, which is slightly lower than married individuals.

- The difference between Married and Single groups is marginal.
- Missed payments are comparable across Married and Single groups, with only a slight difference.
- The Others group appears to have fewer missed payments.

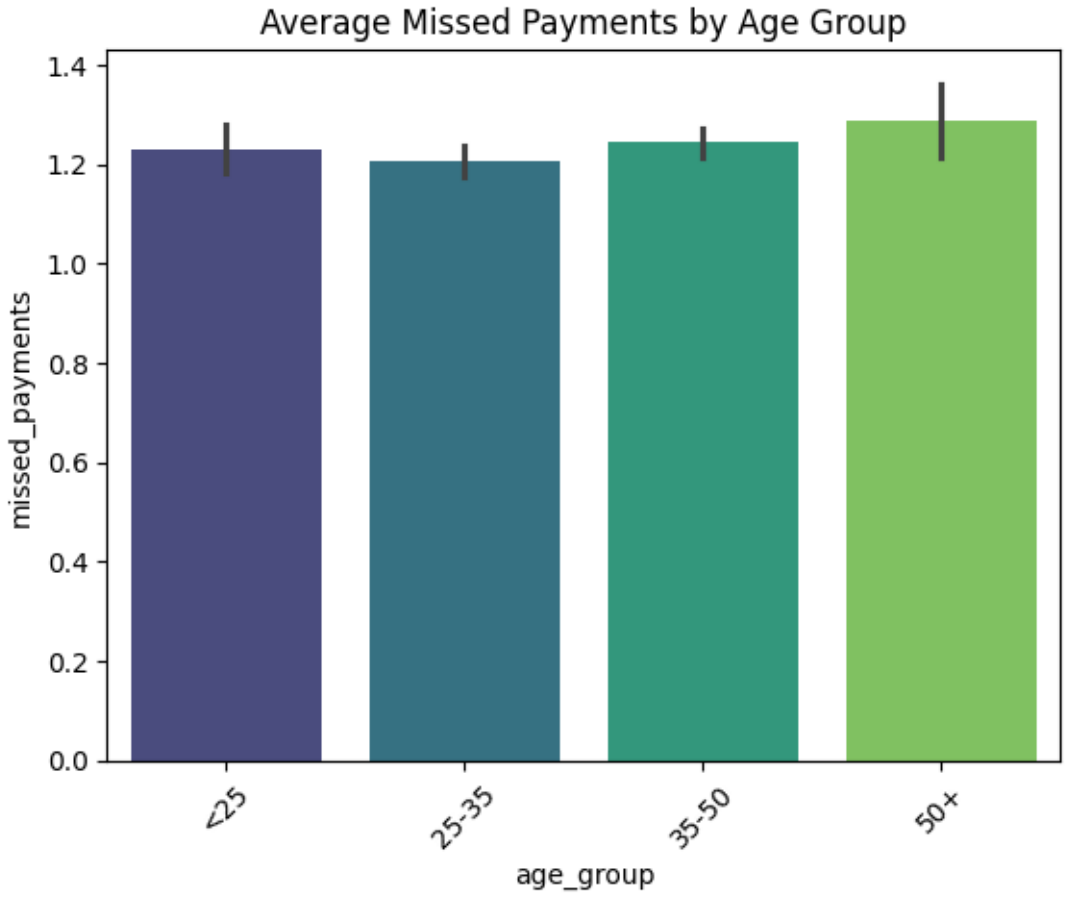


Figure17 17 Average missed payments by age group

Age < 25

- Individuals younger than 25 years reported an average of around 1.23 missed payments.
- The error margin is relatively small, indicating consistent behavior within this group.

Age 25–35

- This group recorded the lowest average missed payments at approximately 1.20.
- Variability is also limited, suggesting stable and reliable results.

Age 35–50

- Individuals in this group show a slightly higher average of 1.24 missed payments.

- The results are fairly consistent due to the narrow error margin.

Age 50+

- Respondents aged 50 and above reported the highest missed payments at about 1.28 on average.
- The error margin is relatively wider, suggesting greater variation in financial behavior among this group.

Overall, the differences across age groups are minor, with averages ranging narrowly between 1.20 and 1.28. Older individuals (50+) tend to miss payments slightly more often compared to younger groups. The 25–35 group appears the most disciplined, with the lowest average missed payments.

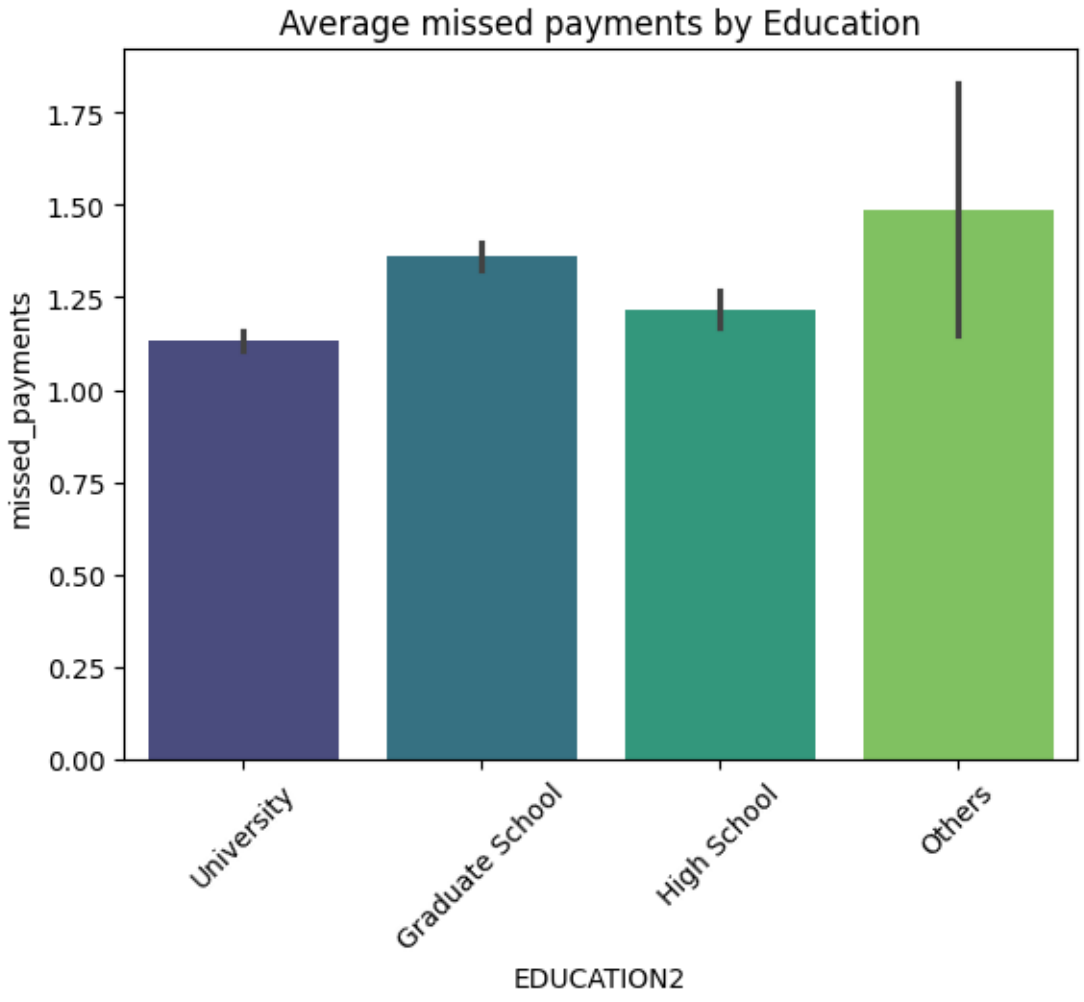


Figure 18 Average missed payments by education

- Individuals with Graduate School education have slightly higher average missed payments (~1.35) than those with University (~1.13) or High School (~1.22) education.

- The "Others" category shows the highest average missed payments (~1.49), but the error bar is large, indicating high variability and less confidence in the average.
- The error bars for University, Graduate School, and High School are small, meaning the data is relatively consistent for these groups.
- Overall, the education level seems to have some impact on missed payments, but the difference between University, Graduate School, and High School is moderate.
- The "Others" category may include diverse or non-standard education backgrounds, which could explain the high variability.

Time series analysis:

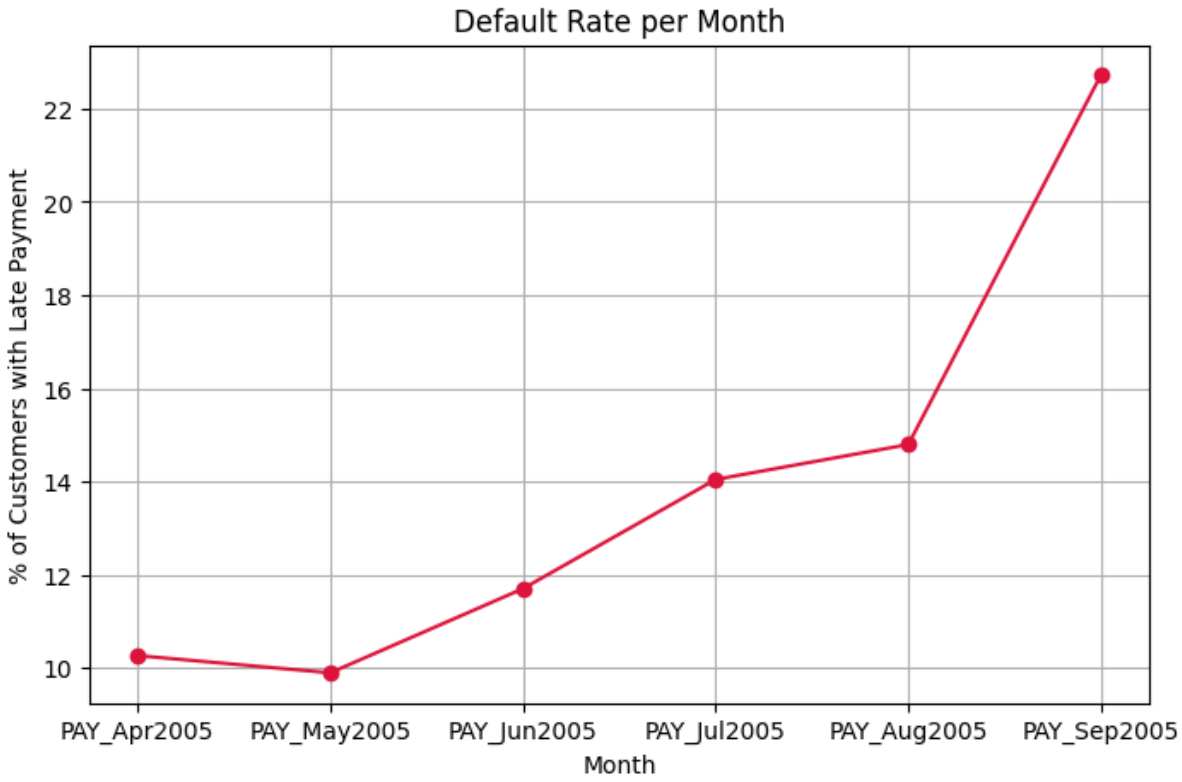


Figure 19 Default rate per month

This chart shows the default (late payment) rate per month from April 2005 to September 2005.

- In April and May 2005, the default rate was relatively low and stable (~10%).
- Starting from June 2005, the default rate begins to climb steadily, reaching about 12%–14%.
- By August 2005, it rises further to nearly 15%.
- Finally, in September 2005, the default rate spikes sharply to around 23%, more than double the level in April.

- There is a clear upward trend in late payments over the six months, with a particularly steep rise in September 2005. This suggests either growing financial stress among customers, seasonal effects (e.g., expenses in late summer/fall), or worsening repayment discipline.

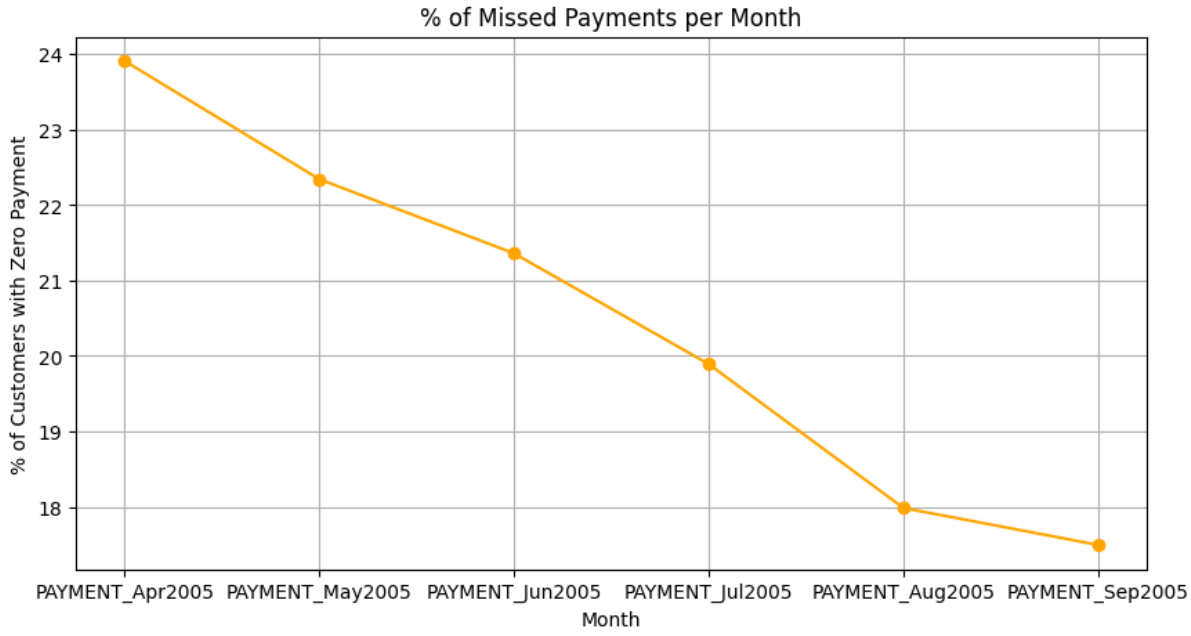


Figure 20 Missed payments by month

- April 2005: Around 24% of customers made no payment. This is quite high, meaning nearly 1 in 4 customers skipped payments.
- May - July 2005: there was a steady decline from 24% to ~20%. This indicates improved repayment behavior and fewer customers are missing payments completely.
- August - September 2005: There was a sharp drop from ~20% to 17.5%. This shows a significant improvement in financial discipline or stronger collection/recovery efforts.
- Overall, the trend is downward, meaning missed payments decreased steadily over the 6 months.
- While default rates (previous chart) were rising, the number of customers who completely skipped payments fell.
- This suggests a shift from "no payment at all" to "partial or late payments".
- Customers may still be struggling, but instead of skipping entirely, they are paying at least something.

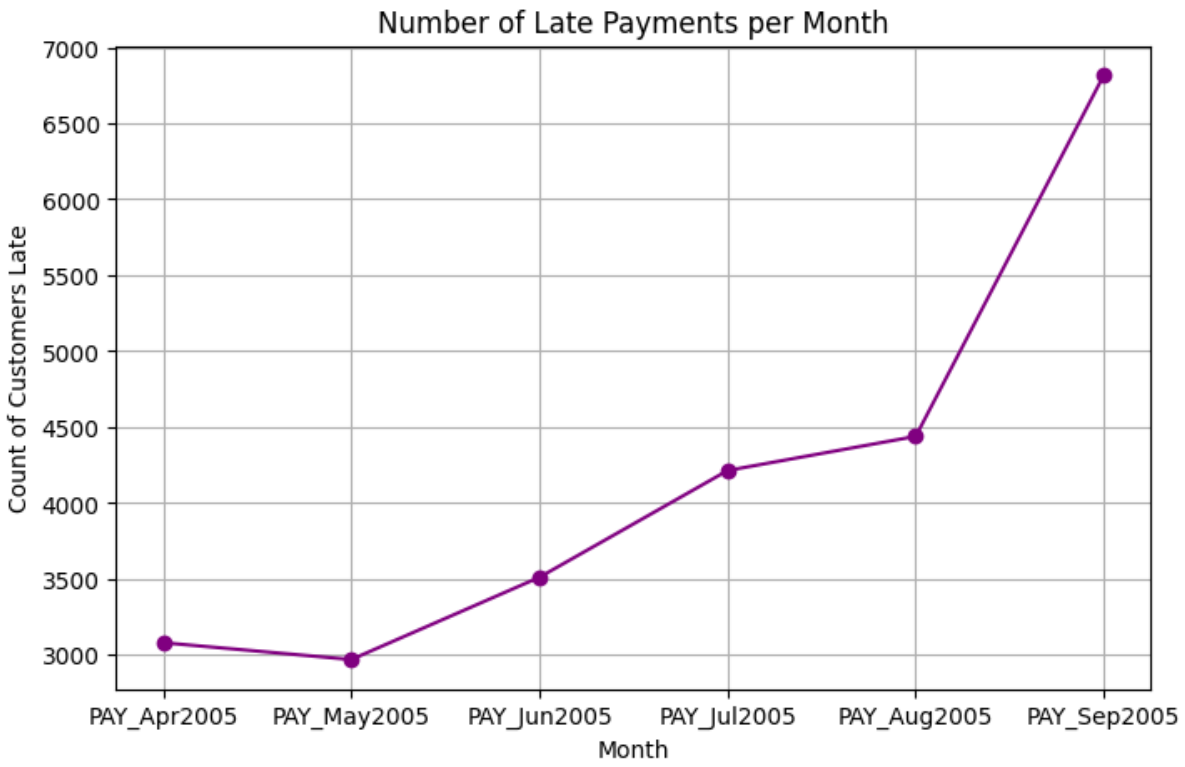


Figure 21 Number of late payments by month

1. **April–May 2005:**
    - The number of late payments was relatively low (~3,000 in April, slightly below in May).
    - This suggests fewer customers were struggling early on.
2. **June–July 2005:**
    - Noticeable increase (jump from ~3,500 in June to ~4,200 in July).
    - Indicates more customers began falling behind.
3. **August 2005:**
    - Continues upward trend (~4,450 late payments).
    - Steady accumulation of financial stress.
4. **September 2005:**
    - Sharp spike to nearly 7,000 late payments, the highest in the six-month period.
    - Suggests a sudden deterioration in repayment behavior, possibly seasonal (e.g., tuition fees, holiday spending, or macroeconomic stress).

Predictive analysis:

The objective of the analysis is to train a random forest regressor model from sci-kit learn library of pandas on the data so that it can predict default payments of customers.

Step-1: Load the dataset

The dataset is loaded after preprocessing using pandas from python.

Step-2: Data encoding

Import labelencoder from sklearn.preprocessing library from python. This converts categorical values to numerical by assigning the values labels. For example; Male = 0, Female=1.

Step-3: Split the data

Now the data is split into train data and test data. Train data is used to train the model so it is capable of prediction. Test data is devoid of target column, which in our case is, default payment next month. After training, the model predicts if the customer is gonna make a default payment or not. The data is split using train_test_split from sklearn.model_selection in the ratio 80%:20%.

Step-4: Training the model

A Gradient boost classifier model from sklearn.ensemble is trained on the train data.

Step-5: Prediction

The model now predicts default payments on test data.

Step-6: Evaluate the model

The errors between predicted default payments and actual default payments is measured by using classification report with scores such as accuracy, precision, recall, F1 score, etc, confusion matrix and ROC-AUC curve from sklearn.metric library.

Results:

```
Accuracy: 0.816554054054054
Confusion Matrix:
 [[4384  240]
 [ 846  450]]
Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.95      0.89      4624
           1       0.65      0.35      0.45      1296

    accuracy                           0.82      5920
   macro avg       0.75      0.65      0.67      5920
weighted avg       0.80      0.82      0.79      5920
```

Figure 20 Classification report

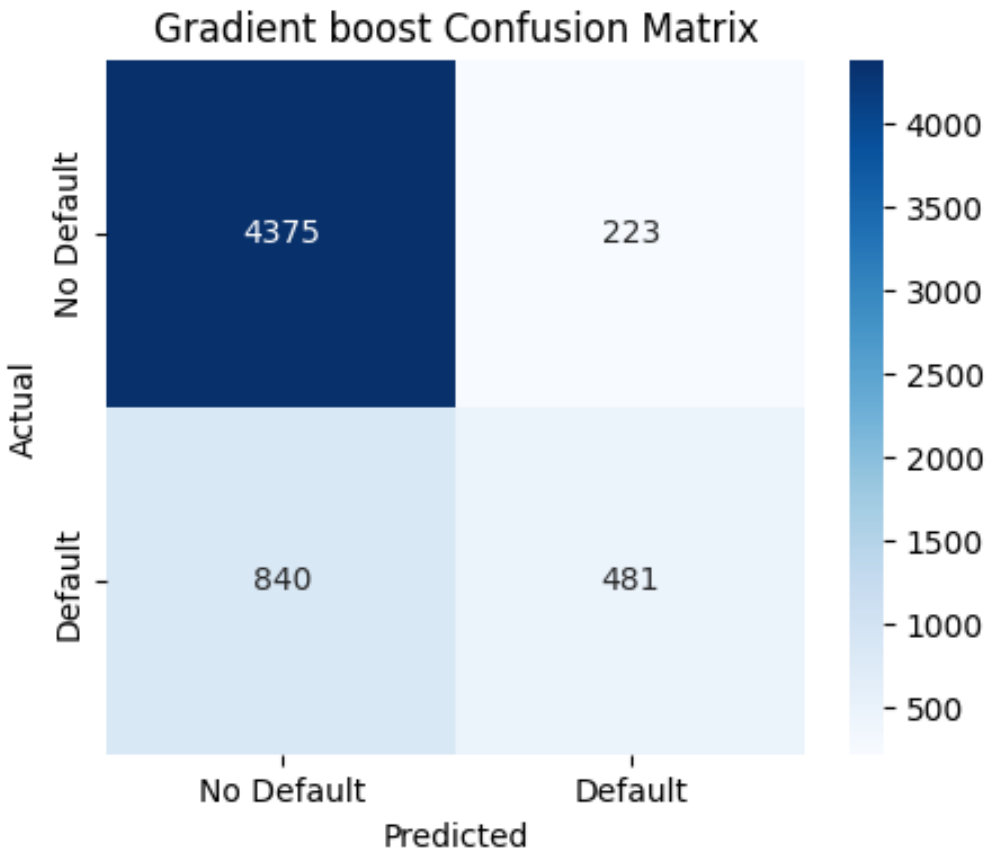


Figure 21 Confusion matrix

Accuracy = ~81.6%
The model correctly classifies ~82% of all cases overall.
But since defaults are much fewer than non-defaults, accuracy alone is not enough.

Confusion Matrix

From the matrix:

- True Negatives (TN): 4375 - Correctly predicted No Default.
- False Positives (FP): 223 - Predicted Default, but actually No Default.
- False Negatives (FN): 840 - Predicted No Default, but actually Default (missed defaulters).
- True Positives (TP): 481 - Correctly predicted Defaults.

Precision, Recall, F1 (per class)

Class 0 (No Default):

- When the model predicts No Default, it's correct 84% of the time.
- It captures 95% of all actual No Defaults.
- Strong balance, so the model is very good at detecting safe customers.

Class 1 (Default):

- When the model predicts Default, it's correct only 65% of the time.
- It only catches 35% of actual defaulters.
- Weak performance, because recall is low.

Overall Metrics

- **Macro Average -** Shows imbalance as the model favors "No Default" class.
- **Weighted Average** - Looks good with 80% precision and recall, but this is inflated by the large number of "No Default" cases.

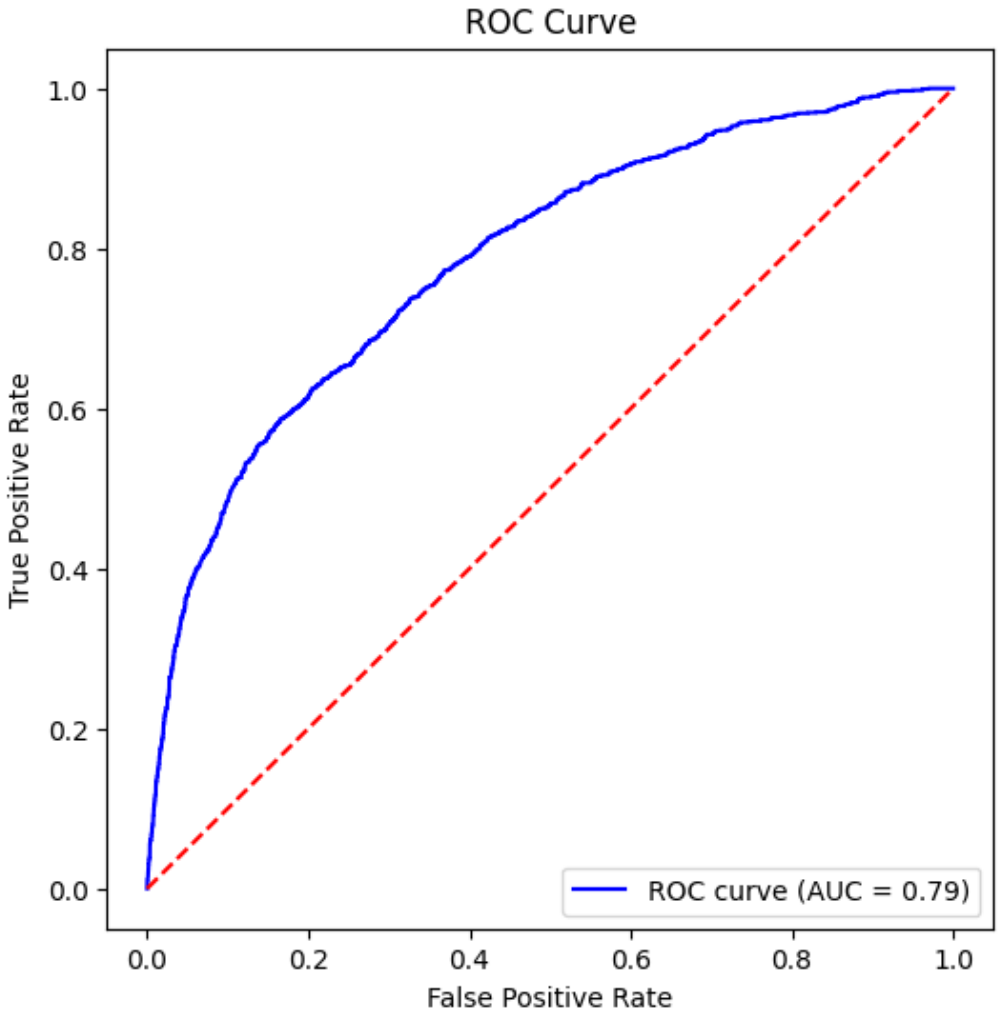


Figure 22 ROC curve, AUC score

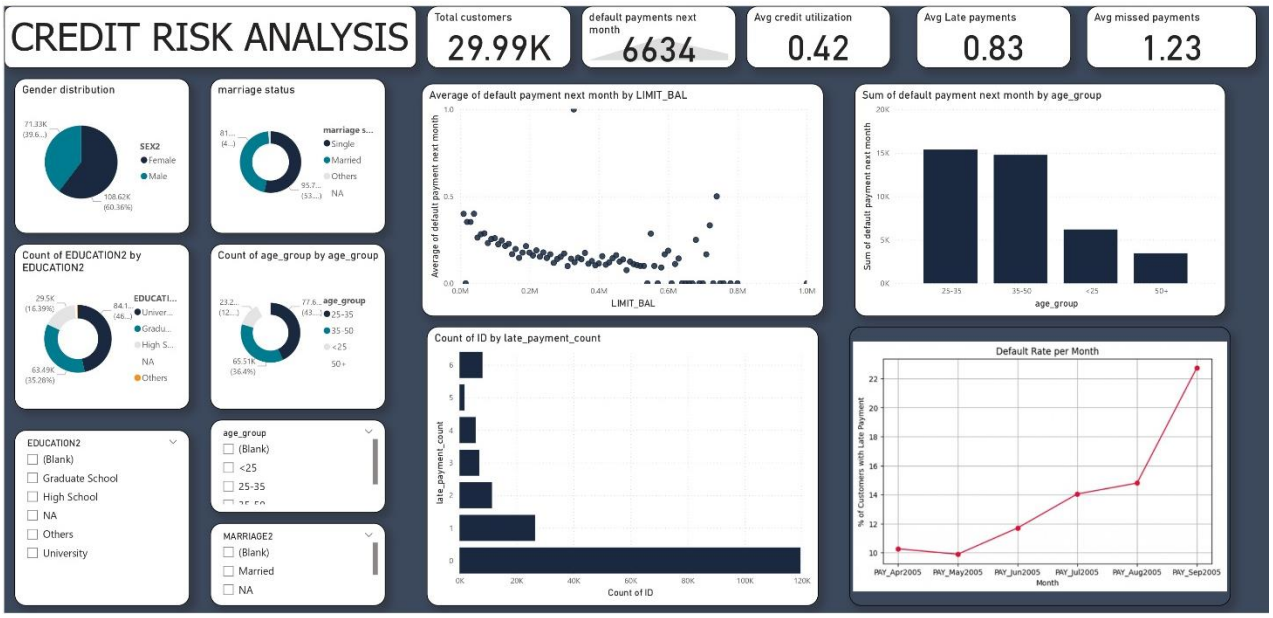
The displayed graph is a Receiver Operating Characteristic (ROC) curve, which is used to evaluate the performance of classification models.

- **True Positive Rate (Y-axis):** Indicates the proportion of actual defaulters (positives) correctly identified by the model.
- **False Positive Rate (X-axis):** Shows the proportion of non-defaulters (negatives) incorrectly classified as defaulters.
- **The Blue Curve:** This represents the model's performance. The closer this curve is to the top-left corner, the better the model distinguishes between default and non-default cases.
- **Red Dashed Line (Diagonal):** Represents a random classifier (AUC=0.5), which cannot separate the classes. Your curve is consistently above this line, indicating meaningful model performance.

- **Area Under the Curve (AUC = 0.79):**
    - The model has a good discriminatory power (AUC values typically range from 0.5 [no discrimination] to 1.0 [perfect discrimination]).
    - An AUC of 0.79 means the model has a 79% chance of correctly distinguishing a randomly chosen defaulter from a non-defaulter.
    - Generally, for credit risk models, an AUC:
        - 0.7–0.8 is considered acceptable
        - 0.8–0.9 is considered excellent
        - 0.9 is considered outstanding

The ROC curve for our gradient boosting credit risk analysis model demonstrates robust classification performance, with an area under the curve (AUC) of 0.79. This indicates that the model is effective at distinguishing borrowers who are likely to default from those who are not, outperforming a random classifier. The AUC value signifies that the model achieves 79% accuracy in ranking a positive instance (defaulter) higher than a negative one (non-defaulter).

Dashboard



The above dashboard provides a comprehensive overview of the credit risk profile among cardholders in the analyzed dataset. Key performance indicators displayed at the top give an immediate snapshot of the portfolio: a total of approximately 30,000 customers, 6,634 default payments next month, a mean credit utilization ratio of 0.42, and average late/missed payments of 0.83 and 1.23 respectively.

Demographic segmentations including gender, marital status, education level, and age group distributions are shown via donut charts, facilitating comparisons between customer classes. Filtering options further allow for dynamic risk analysis across these categories.

The scatter plot reveals that higher credit limits do not necessarily correspond to increased default risk, suggesting that other behavioral factors contribute more significantly. The sum of defaults by age group demonstrates elevated risk in customers aged 25–35 and 35–50, indicating a need for closer monitoring within these segments.

Late payment frequency is a critical risk indicator, as shown by the bar chart of ID counts by late_payment_count. Most customers maintain low late payment counts, but a subset presents elevated risk due to repeated delays.

Lastly, the line chart tracks the default rate trend across six months, signalling a steady rise in late payments culminating in September 2005—a pattern that could reflect seasonal or cyclical risk factors.

Overall, this dashboard enables rapid identification of risk concentrations by segment and payment behavior, supports targeted intervention, and enhances credit risk decision-making through visualized insights.

Actionable Insights

1. Implement Dynamic Credit Limit Adjustments Based on Credit Utilization and Payment Behavior
    - Automatically lower credit limits for customers with consistently high credit utilization (above 80%) or increasing late payment counts.
    - For example, reduce credit limits by 15–20% for customers showing a declining bill growth trend combined with late payments to mitigate default risk without outright rejecting credit.
2. Early Intervention Alerts for Customers with Emerging Payment Delays
    - Use the repayment status history (PAY_0 to PAY_6) to flag customers who have delayed payments for 2+ consecutive months.
    - Trigger personalized communications (calls, text alerts, or payment reminders) within 7 days after the second delay to encourage timely payments and avoid defaults.
3. Tailor Lending Policies for Demographic Segments with Higher Default Rates
    - For younger borrowers (age group 20–29) and unmarried customers identified with higher default tendencies, adopt stricter approval criteria or offer lower credit limits initially.
    - Additionally, provide financial literacy resources or counseling programs targeted to these segments to improve repayment behaviors.
4. Incorporate Payment Consistency and Missed Payments Metrics into Credit Scoring Model
    - Adjust scoring algorithms to include the engineered features such as pay_consistency, missed_payments, and max_delay.
    - Prioritize lending decisions and risk pricing based on these behavioral features to better predict defaults than relying solely on demographic or credit limit data.
5. Develop a Real-Time Credit Risk Dashboard for Relationship Managers
    - Integrate key metrics like default probability, credit utilization, and payment delays into an interactive dashboard.
    - Train credit officers to use this tool to identify high-risk accounts proactively and apply customized repayment plans or credit holds when necessary.
6. Review and Tighten Credit Approval for Education Levels Showing Higher Risk

- Since certain education levels are associated with higher default rates, consider requiring additional documentation or co-signers for applicants from these groups.
- Implement targeted risk-based pricing, offering higher interest rates or fees to compensate for increased risk.

7. Regularly Reassess Credit Risk Model with Updated Data and Feedback
    - Schedule quarterly retraining of the XGBoost model with the latest payment data and demographic changes.
    - Use feedback loops from recovered defaults or successfully improved accounts to recalibrate thresholds and risk flags.

Limitations and future work

Limitations

1. Data Scope and Quality:
    - The dataset covers credit card clients from Taiwan only, which may limit the generalizability of the model to other regions or financial products.
    - Some feature engineered variables like avg_pay_ratio and bill_growth had extreme outliers and negative values, which might affect model stability.
    - Missing values existed in a few features (e.g., pay_ratio1, avg_pay_ratio), which were handled by removal or imputation but may lose some information.
    - The dataset's time span is limited to six months of repayment history, restricting the ability to capture long-term behavioral patterns.
2. Modeling Constraints:
    - The model was trained using a gradient boosting classifier with limited hyperparameter tuning. More extensive tuning or ensemble methods may yield better predictive performance.
    - The binary classification (default/no-default) does not incorporate the severity or amount of default, missing finer granular risk assessments.
    - Class imbalance was present (approximately 22% default rate), which may have impacted classifier bias despite metric evaluations.
3. Contextual Factors:
    - External economic factors, such as changes in unemployment, inflation, or regulatory policies, were not included in the analysis but could significantly influence default risk.
    - Behavioral variables beyond payment and billing records (e.g., customer interactions, employment history) were not available.

Future Work

1. Data Enrichment:
    - Incorporate additional data sources such as credit bureau reports, macroeconomic indicators, or social-economic variables to enhance model context and accuracy.
    - Obtain longer repayment histories to analyze trend patterns and credit behavior changes over time.

2. Advanced Modeling Techniques:
    - Experiment with ensemble models combining gradient boosting with neural networks or deep learning approaches for improved pattern recognition.
    - Implement cost-sensitive learning or anomaly detection methods to better handle imbalanced datasets and rare default events.
3. Refinement of Features and Metrics:
    - Engineer new temporal and behavioral features such as rolling averages, payment velocity, and recovery likelihood post-default.
    - Utilize multi-class or regression models to predict default severity or expected loss amounts.
4. Model Explainability and Compliance:
    - Apply explainable AI techniques (e.g., SHAP values) to ensure transparency and regulatory compliance in credit decision-making.
    - Develop interpretable models that engage stakeholders and consumers through risk explanations.
5. Operational Integration:
    - Build real-time risk assessment dashboards integrating live transactional data.
    - Design automated alert systems for early intervention on high-risk accounts.

Conclusions

The project objective was successfully fulfilled as we developed a credit risk prediction model capable of distinguishing between defaulters and non-defaulters with a reasonable degree of accuracy (AUC ~0.79). Incorporating multiple dimensions of borrower data, including detailed payment histories and engineered financial ratios, significantly enhanced model performance beyond baseline demographic indicators. The insights from exploratory data analysis and feature importance confirmed that payment behavior variables were critical drivers of default risk.

While the results are promising, the study acknowledges limitations such as regional data specificity, limited time span, and scope for enhanced model tuning and feature inclusion. Future work could integrate broader economic variables, longer behavioral histories, and advanced model architectures to further improve accuracy and operational impact.

In summary, this project highlights the critical role of data-driven approaches in credit risk management and provides a practical framework for financial institutions to implement predictive analytics. By adopting these methods, lenders can mitigate defaults, optimize credit allocation, and sustain financial health in an increasingly competitive environment.

Bibliography

1. M.Gao, J.yen, M.liu. "Determinants of defaults on P2P lending platforms in China." *International review of economics and finance* (2021): 334-348. online journal.
2. Md.Bokhtiar Hasan, Masnun Mahi, Tapan Sarker, Md.Ruhul A min. "Spillovers of the COVID-19 Pandemic: Impact on Global Economic Activity, the Stock Market, and the Energy Sector." *J. Risk Financial Manag* (2021). web.
3. *world bank*. 2021. web. august 2025.