Credit Card Default Prediction Report

Shreyansh Kuntal
shreyansh_k@ph.iitr.ac.in

22324018

Contents

1 Overview and Modeling Strategy 2

2 Exploratory Data Analysis (EDA) 2

2.1 Categorical Features 2

2.2 Numerical Features 5

3 Feature Engineering 14

4 Model Training and Selection 15

4.1 Handling Imbalanced Data 15

4.2 Metric Used 15

4.3 Classification Threshold 15

4.4 Model Comparison 16

4.5 Business Implications 17

5 Summary and Learnings 17

1 Overview and Modeling Strategy

The goal of this project is to predict whether a customer will default on their credit card payment next month. The project involved the following structured approach:

- Performed exploratory data analysis (EDA) to identify behavioral and financial patterns among variables and assess their relationship with credit default risk.
- Handled class imbalance using SMOTE to ensure the model learns from both default and non-default cases effectively.
- Engineered domain-specific features like credit utilization and delinquency streaks.
- Evaluated and compared classification models such as Logistic Regression, Decision Trees, Random Forest and Gradient Boosting using F2 score on test data. Gradient boosting gave maximum F2 score.
- Tuned classification thresholds on the best model to reflect the bank's credit risk tolerance.

2 Exploratory Data Analysis (EDA)

2.1 Categorical Features

Marriage Status

- The dataset includes four marriage categories: Married (1), Single (2), Others (3), and No description (0).
- The bar plot and proportion analysis show that married individuals exhibit a higher rate of default (approximately 20.3%) compared to singles (17.9%). This could be attributed to additional financial responsibilities such as rent or mortgage payments, childcare, family healthcare, or education expenses that strain disposable income.
- Figure 1 shows the distribution of defaults across different marriage categories.

Sex

- As shown in Figure 2, the proportion of female defaulters is higher (20.9%) compared to male defaulters (17.8%).
- Male customers form a larger share of the dataset, but females are slightly more likely to default.
- Possible financial reasons include:
  - Lower average income and credit limits for women.

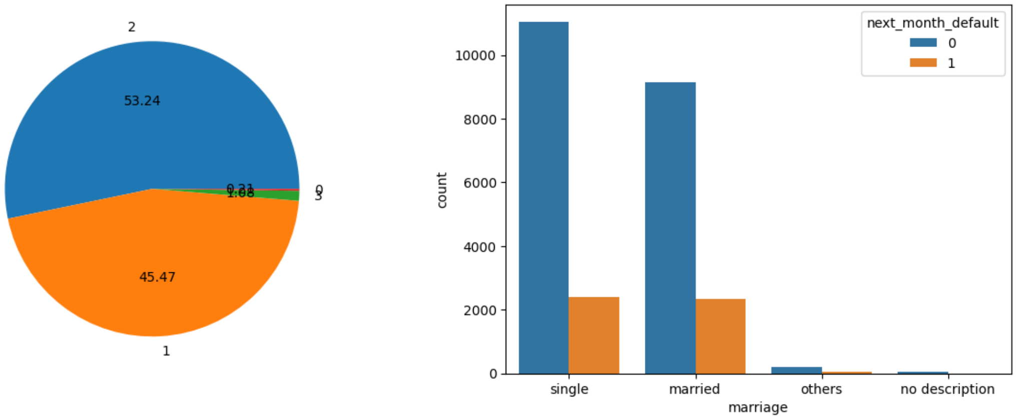


Figure 1: Default distribution by marriage status.

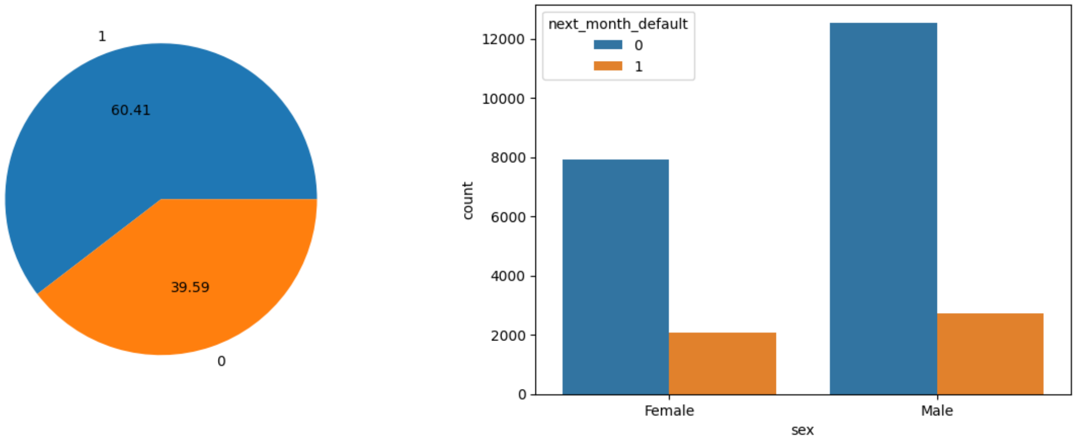


Figure 2: Default distribution by sex

- Higher financial burden among single female households.
- Greater vulnerability to economic stress due to caregiving responsibilities.

- These trends indicate a need for credit risk models to consider sex-based financial disparities.

Education Level

- As observed in Figure 3, defaulters are distributed across all education levels.
- Customers with **graduate school education** (coded as 1) have the **lowest default proportion** (16.2%), suggesting better financial stability or awareness.
- In contrast, **high school** (21.3%) and **university** (20.9%) educated customers show a higher likelihood to default.
- Possible factors:

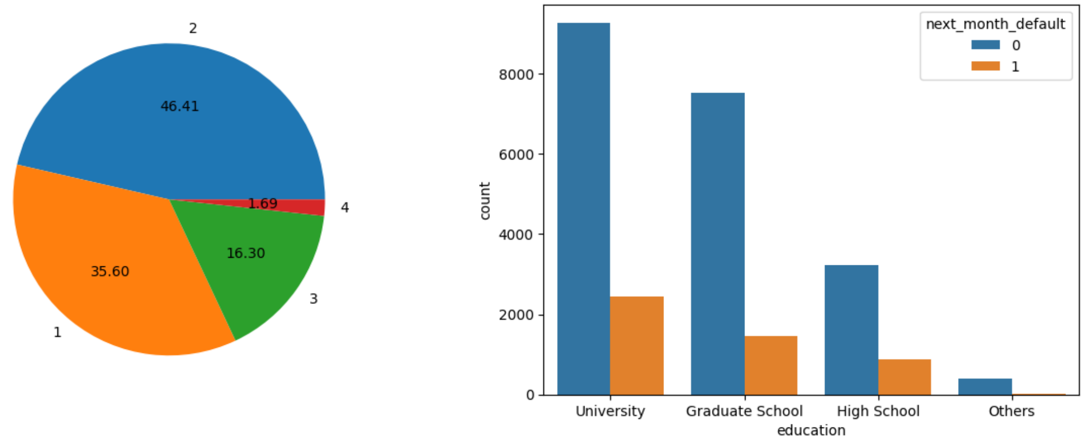


Figure 3: Default distribution by education level

- Graduate degree holders may have higher income or financial literacy.
- High school graduates might lack access to well-paying jobs and credit management training.

- The "Others" category shows the least defaults but is underrepresented in the dataset, making trends less reliable.

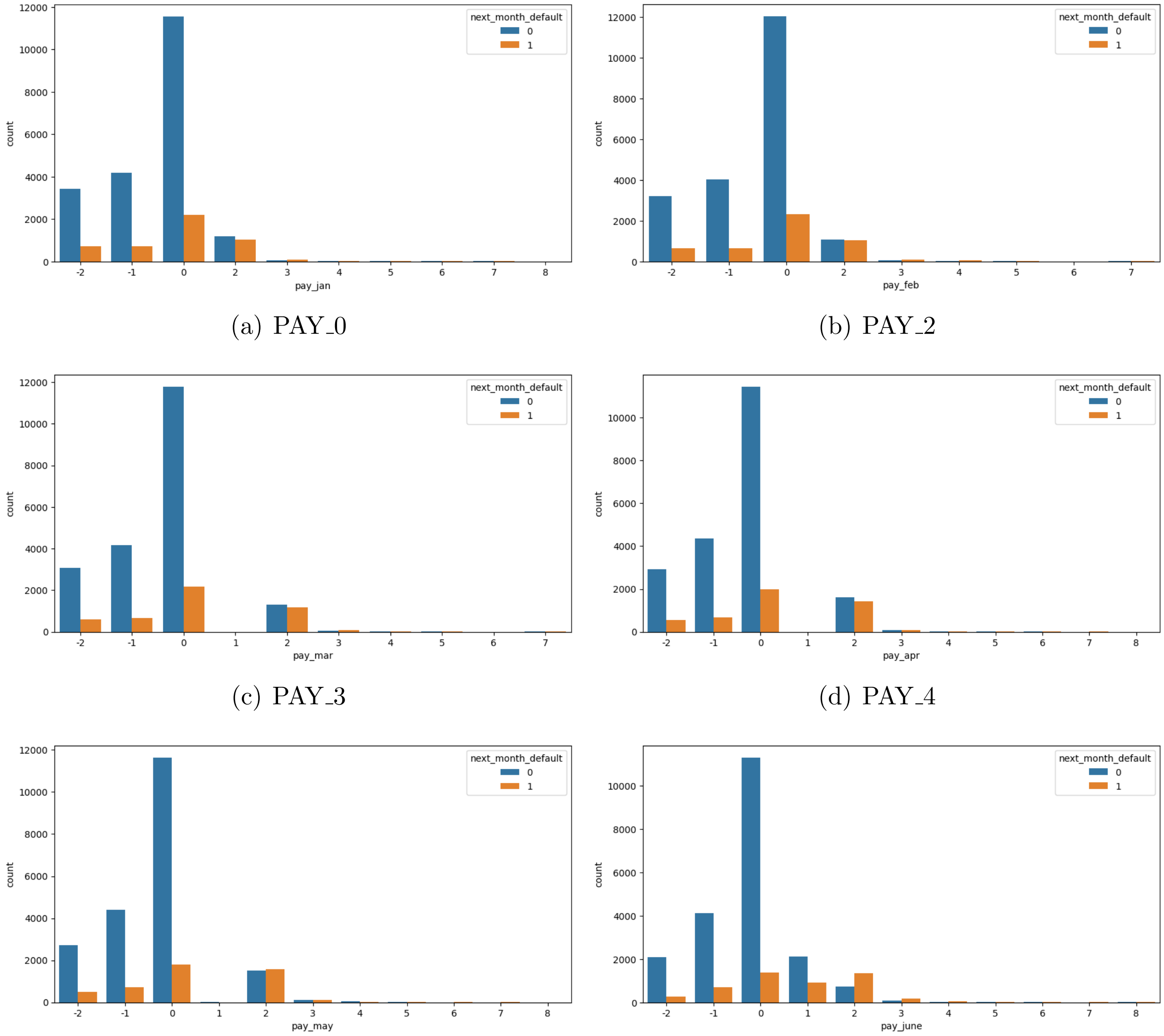
Payment Status (PAY_0 to PAY_6)

- These variables represent the repayment status of the customer in each of the six months preceding the observation. The values indicate:
  - **-2**: No bill generated (no credit consumption)
  - **-1**: Bill generated and fully paid in the same month
  - **0**: Partial or minimum payment made (revolving credit)
  - $\geq$**1**: Payment delayed by corresponding months (e.g., 2 = 2-month delay)

- **Key observations from the plots:**
  - A large proportion of **defaulters** fall under the status **PAY_m = 0**, suggesting that customers who only make partial or minimum payments are at higher risk of default.
  - The next most frequent status for defaulters is **PAY_m = 2**, indicating delayed payments are common among defaulters.
  - **Non-defaulters** mostly have statuses **-1** or **0**, reflecting on-time full payments or partial payments.
  - Very few defaulters are observed with status **-2** or **-1**, showing that either no credit usage or timely full repayment correlates with lower default risk.

- From PAY_0 to PAY_6, there is a noticeable trend: the share of defaulters increases in higher delay categories (e.g., PAY_m ≥ 3), particularly in recent months.

The payment status distribution for each month is visualized in **Figure 4**.



(a) PAY_0

(b) PAY_2

(c) PAY_3

(d) PAY_4

(e) PAY_5

(f) PAY_6

Figure 4: Distribution of repayment statuses from PAY_0 (most recent) to PAY_6

2.2 Numerical Features

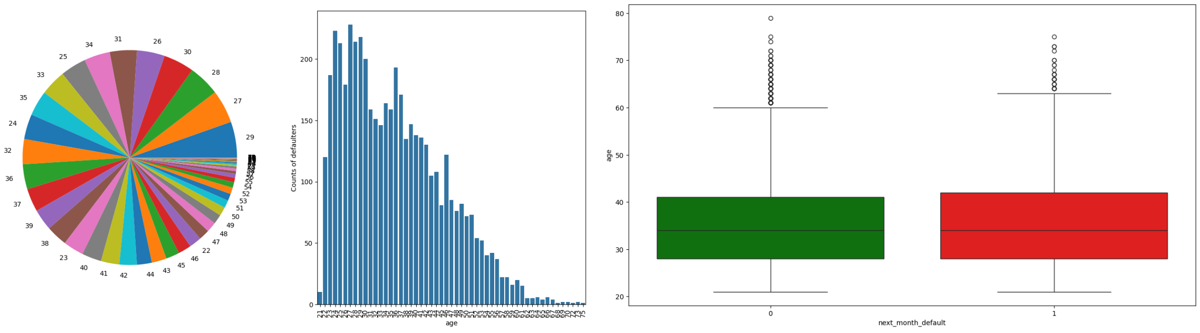
AGE

- This variable indicates the age of the customer in years.
- **Summary statistics:**
  - Mean: 35.4 Std Dev: 9.17

- Min: 21   Max: 79
  - 25th percentile: 28   Median: 34   75th percentile: 41

- **Observation:** The distribution of age among defaulters shows a roughly Gaussian shape centered around 27–34 years.
  - The highest concentration of defaulters appears near 27 years, possibly because individuals in their late 20s are early in their careers with lower income stability, higher risk appetite, and limited credit experience.
  - The average (mean) age of 35.4 reflects a moderate skew caused by older customers with larger but less frequent defaults. This aligns with real-world credit behavior, where risk generally reduces with age due to increased income and credit maturity.
  - Younger customers (less than 27) may have smaller credit limits and hence fewer defaults, while older customers (more than 40) may exhibit more financial responsibility, resulting in fewer defaults.
  - The long tail toward older ages is also visible in both histogram and boxplot, highlighting presence of outliers or rare older defaulters.

Visualizations of age distribution and dispersion are provided in **Figure 5**.



(a) Distribution of age among defaulters

(b) Boxplot showing spread and outliers

Figure 5: Age distribution and summary statistics

Monthly Bill Amount (BILL_AMT_MONTH)

- These variables represent the billed amount for each of the past six months, from `BILL_AMT1` (June) to `BILL_AMT6` (January).
- **Observation:**
  - The pairplot and correlation matrix (**Figures 6 and 7**) indicate strong linear correlations between consecutive monthly bills.
  - Customers who had high bill amounts in one month also tend to have high bills in the previous months, suggesting a consistent spending pattern over time.

  - This implies that many defaulters exhibit stable or habitual credit usage behavior rather than impulsive spending.
  - From a credit risk perspective, persistent high billing amounts over time without corresponding payments may be early indicators of financial stress.

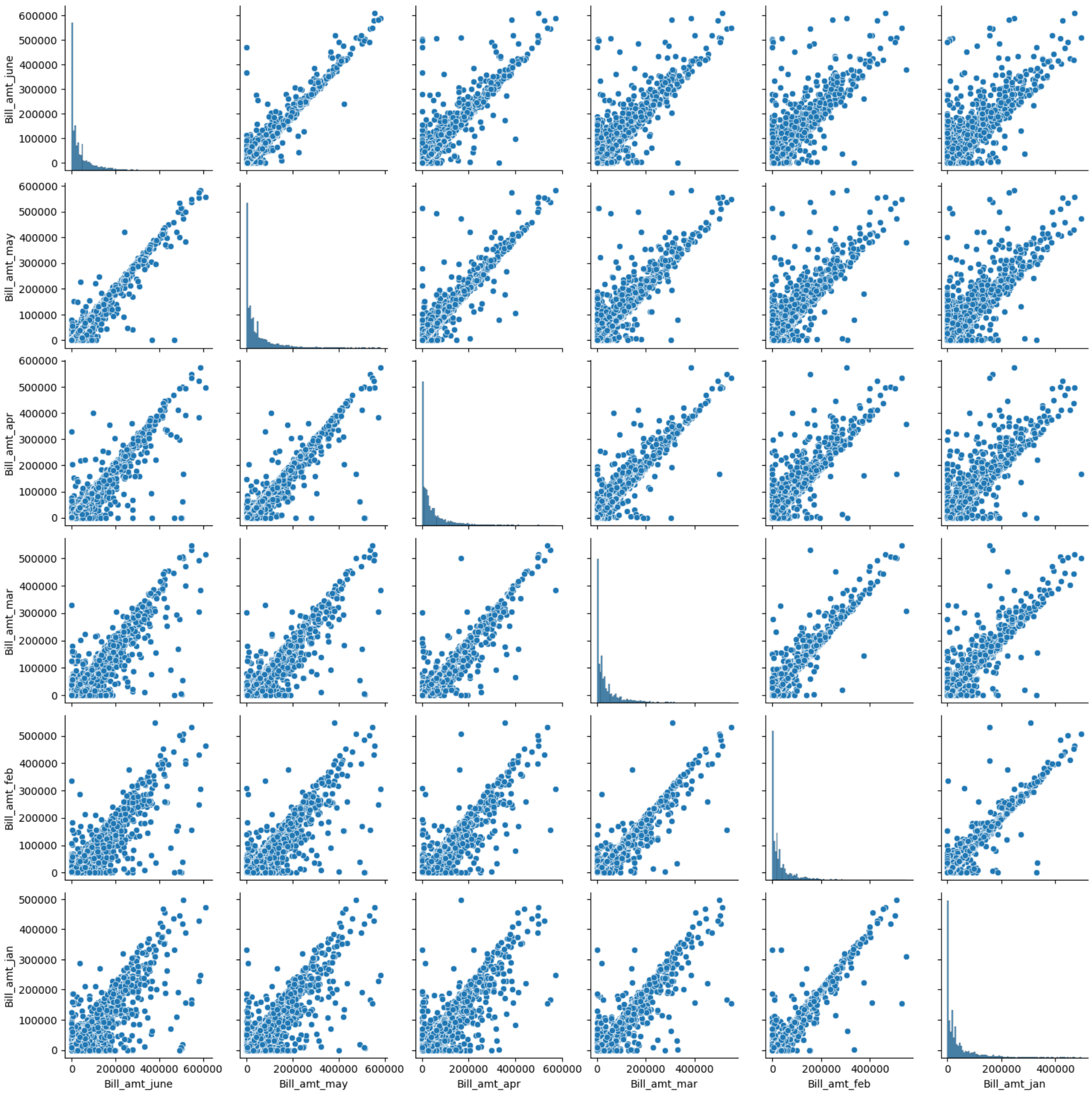


Figure 6: Pairplot showing linear relationships between monthly bill amounts of defaulters

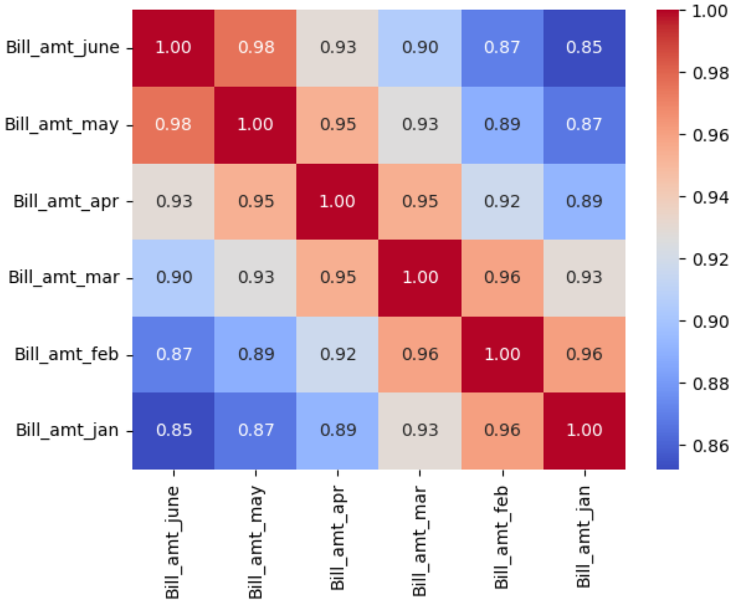


Figure 7: Correlation matrix among monthly bill amounts (January to June) of defaulters

Monthly Payment (PAY_AMT_MONTH)

- These variables represent the amount paid in each of the past six months, from `PAY_AMT1` (June) to `PAY_AMT6` (January).
- **Observation:**
  - As shown in the pairplot and correlation matrix (**Figures 8 and 9**), payment behavior is notably inconsistent across months.
  - The lack of correlation indicates that defaulters tend to pay arbitrarily—large payments one month may be followed by negligible or no payments in subsequent months.
  - This irregularity may reflect lack of structured repayment schedules.
  - From a financial perspective, this makes it a challenge for lenders to estimate creditworthiness based solely on payment amounts.

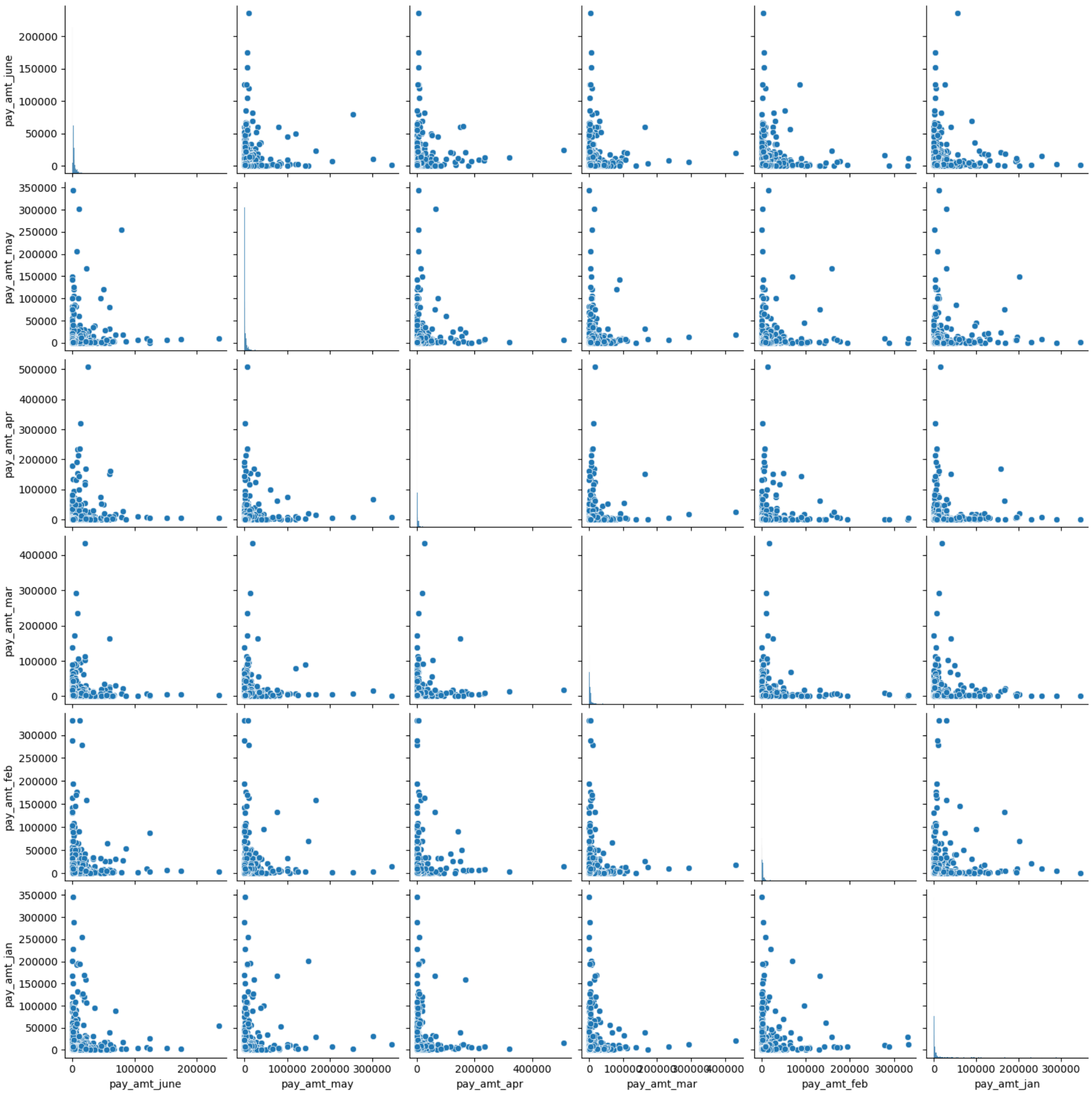


Figure 8: Pairplot showing weak or no correlation between monthly payment amounts of defaulters

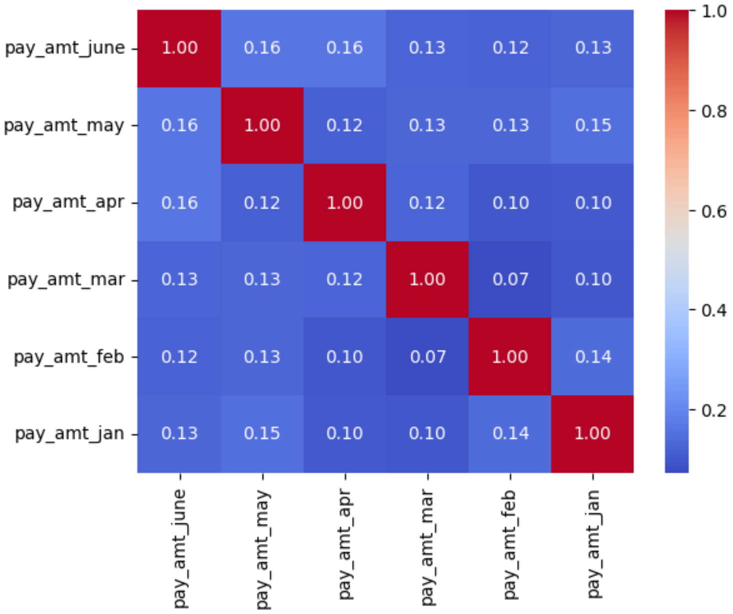


Figure 9: Correlation matrix among payment amounts (January to June) of defaulters

Average Monthly Bill Amount

- This variable represents the average of the six `BILL_AMT` variables, reflecting the mean monthly expenditure over January to June.
- **Observation:**
  - As shown in **Figure 10**, the medians of average bill amounts between defaulters and non-defaulters are comparable.
  - However, the maximum bill amounts for non-defaulters are significantly higher, suggesting that higher credit limits (and hence higher spending) are granted to financially stable customers.
  - The interquartile range (IQR) also appears broader for non-defaulters, hinting at a more diverse but still well-managed spending pattern.
  - From a financial risk perspective, lower average bill amounts in defaulters may indicate lower credit limits, imposed due to prior defaults.

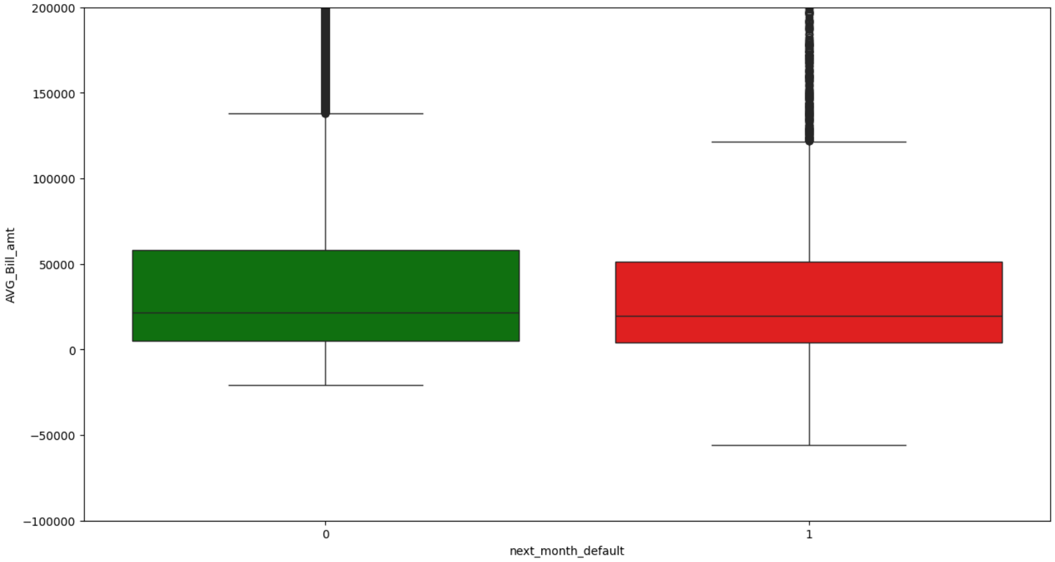


Figure 10: Box plot showing average monthly bill amount across default classes

PAY_TO_BILL Ratio

- This variable measures the ratio of payment amount to billed amount for each month, averaged across six months. It captures how much of their dues customers typically pay.
- **Observation:**
  - As seen in **Figure 11** and **Figure 12**, most defaulters exhibit a ratio clustered close to zero, indicating minimal payments against their bills.
  - In contrast, non-defaulters display a wider spread, with some even exhibiting negative ratios (advance payments).
  - A low PAY_TO_BILL ratio among defaulters suggests payment behavior that leans toward revolving debt increasing credit risk.
  - Financially, a persistent low ratio reflects high debt accumulation and poor repayment discipline—clear indicators for default risk.

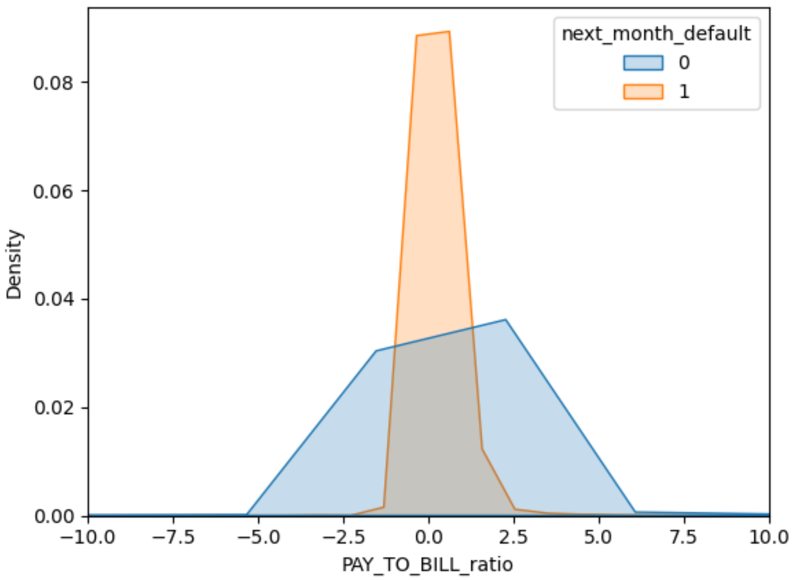


Figure 11: KDE plot showing PAY_TO_BILL ratio distribution by default class

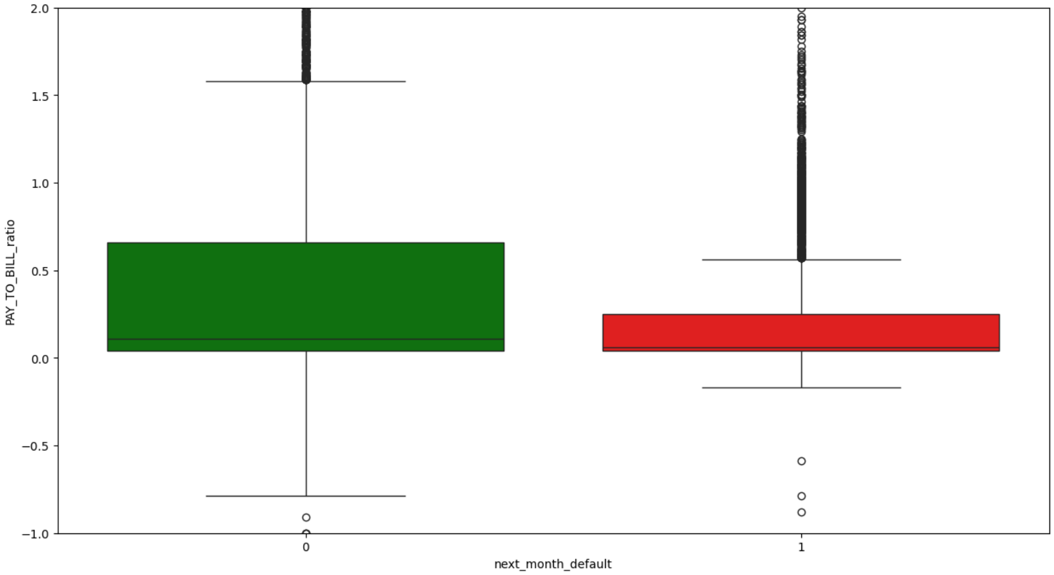


Figure 12: Box plot of PAY_TO_BILL ratio across default status

Credit Utilization Ratio (AVG_util_ratio)

- The Credit Utilization Ratio is calculated by dividing the total outstanding bill amount by the total credit limit assigned to the customer. A high utilization ratio suggests that the customer is using a large portion of their available credit which may indicate financial stress or reliance on borrowed funds. A low utilization ratio suggests smart credit use and is viewed more favorably by banks.
- **Observation:**

- As shown in **Figure 13**, the box plot highlights that defaulters typically have a higher median and average credit utilization ratio compared to non-defaulters.
- A high utilization ratio often signals financial strain or dependency on credit, both of which elevate the risk of future default.
- Non-defaulters, with lower utilization, tend to manage their credit lines more conservatively, a behavior generally rewarded by credit scoring models.

- **Financial Insights:**
  - Financially, consistently high utilization reduces credit flexibility and is a red flag for lenders — it reflects not just need but often an inability to repay timely.

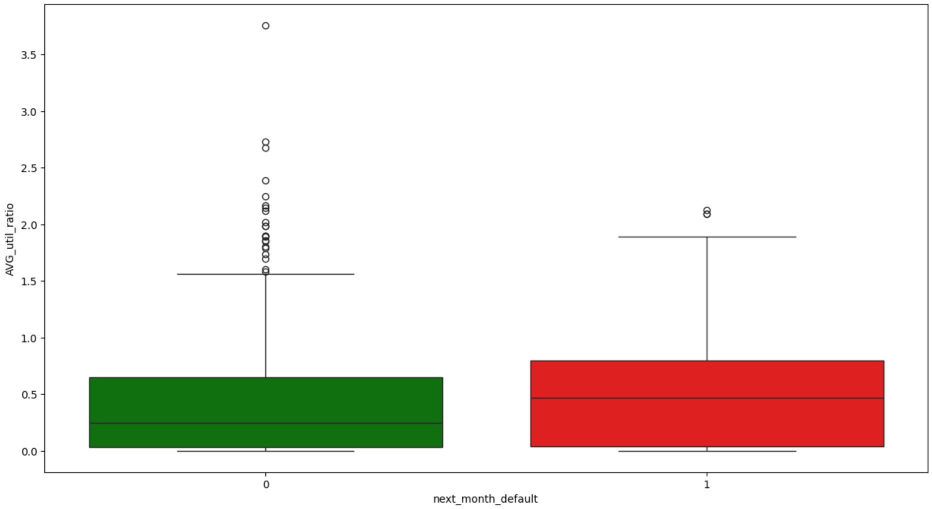


Figure 13: Box plot of average Credit Utilization Ratio by default status

Repayment History — Delinquency Streak

- The Delinquency Streak is defined as the longest number of consecutive months in which a customer has failed to make timely payments (`pay_x` ≥ 1).
- **Observation:**
  - As shown in **Figure 14**, the box plot reveals that the maximum delinquency streak for non-defaulters is zero (excluding outliers), indicating consistent payment behavior.
  - Defaulters exhibit streaks ranging from 0 to 6 months, with the interquartile range (IQR) stretching from 0 to beyond 3 months.
  - The broader spread and higher median among defaulters highlight their prolonged payment lapses, making this a strong predictive feature.
- **Financial Insights:**

- Delinquency streaks shows financial mismanagement. Customers with repeated missed payments shows a declining ability or willingness to repay debt.
- From a credit risk perspective, a single late payment might be manageable, but long streaks increase losses to lender leading to stricter credit terms.

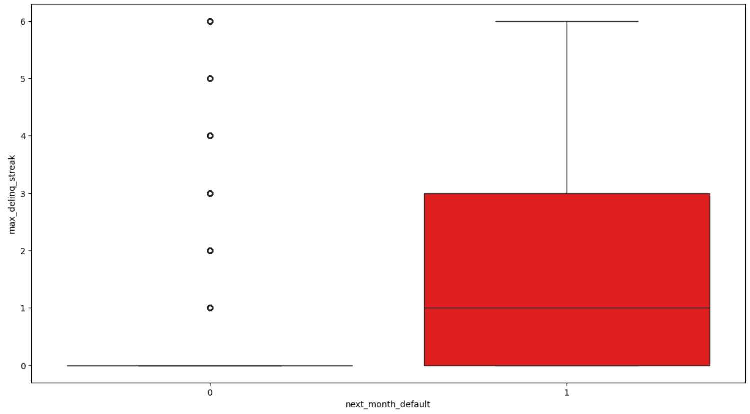


Figure 14: Box plot of Delinquency Streak by default status

3 Feature Engineering

In order to improve model performance and embed domain knowledge, we engineered new features that capture meaningful financial behavior beyond raw variables. Two such high-value features are described below.

1. Delinquency Streak

Definition: The longest consecutive streak of months during which the customer delayed payments (i.e., months where `PAY_x` ≥ 1). **Formula:**

$$\text{Delinquency Streak} = \max\left(\text{length of consecutive months with } PAY\_x \geq 1\right)$$

Further analysis and visualizations of this feature are provided in the EDA section.

2. Credit Utilization Ratio (AVG_util_ratio)

Definition: The average proportion of the assigned credit limit used over six months. **Formula:**

$$\text{Credit Utilization Ratio} = \frac{1}{6}\sum_{i=1}^{6}\left(\frac{BILL\_AMT_i}{LIMIT\_BAL}\right)$$

This feature's financial relevance is discussed in the EDA section along with supporting plots.

4 Model Training and Selection

4.1 Handling Imbalanced Data

The original dataset had a significant imbalance between defaulters and non-defaulters, which can bias models toward the majority class. To fix this, we applied Synthetic Minority Oversampling Technique (SMOTE), a popular method that synthetically generates new instances of the minority class.

- **Original dataset shape:** 25,121 samples
- **Resampled dataset shape after SMOTE:** 40,674 samples

4.2 Metric Used

F2-score is used as the primary evaluation metric. The F2-score places more emphasis on recall than precision, making it especially suitable for credit risk problems.

- **Recall-focused:** F2-score is designed to minimize false negatives by giving more weight to recall. This means we aim to catch as many defaulters as possible.
- **Bank appetite:** In the context of credit risk, missing a defaulter (false negative) is usually more costly than incorrectly flagging a non-defaulter (false positive).
- **Trade-offs:** We are willing to tolerate a slightly higher false positive rate in exchange for reducing missed defaulters.

This approach reflects real-world financial practices, where banks prefer conservative models that prioritize early risk detection.

4.3 Classification Threshold

Understanding the Cost of Errors

Case	**Meaning**	**Business Risk**
False Positive (FP)	Predict default, but customer actually repays	Lost opportunity: rejected a good customer → lost revenue
False Negative (FN)	Predict non-default, but customer actually defaults	High financial loss: approved credit to a risky customer

Table 1: Business implications of classification errors

Interpretation

- A **false positive** may hurt customer trust and reduce profits due to unnecessarily rejecting good customers.
- A **false negative** may result in a direct financial loss by approving credit to risky customers.
- In credit risk modeling, minimizing **false negatives** is generally prioritized, even at the cost of more false positives.

To balance this trade-off, we analyzed how the F2 score varies with different classification thresholds. While the F2 score peaks at a very low threshold, using such a threshold may lead to a large number of false positives, degrading customer experience.

Instead, we selected a threshold of **0.50**, which still maintains a strong F2 score while offering a more balanced trade-off between false positives and false negatives. This threshold aligns with the business's conservative risk policy.

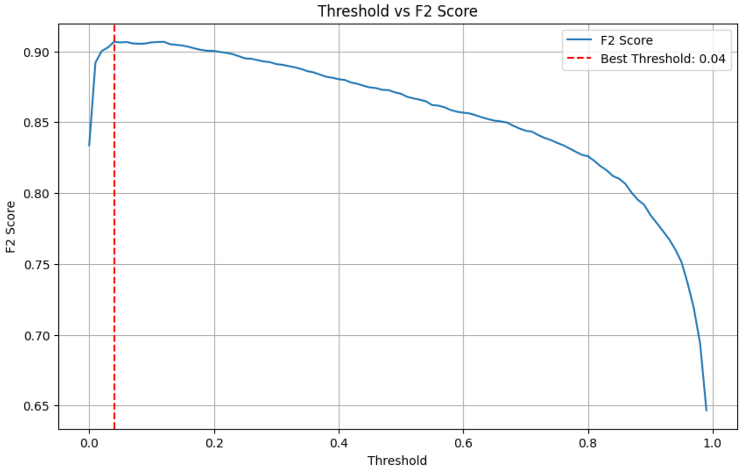


Figure 15: F2 Score vs Classification Threshold

4.4 Model Comparison

We evaluated multiple classifiers on the resampled dataset using several performance metrics on the chosen threshold of **0.50**. The results are summarized below:

Classifier	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**F2 Score**	**ROC-AUC**
Logistic Regression	0.758	0.790	0.702	0.744	0.718	0.836
Decision Tree	0.782	0.775	0.797	0.785	0.792	0.782
Random Forest	0.856	0.877	0.829	0.852	0.838	0.935
XGBoost	**0.881**	**0.894**	**0.864**	**0.879**	**0.870**	**0.949**

Table 2: Performance comparison of different classifiers on test data

Model Selection: Among all models, **XGBoost** achieved the highest F2 score, making it the optimal choice for our objective of minimizing false negatives. It also performed strongly across other metrics such as ROC-AUC and Recall, further justifying its selection.

4.5 Business Implications

Scenario	**Implication**
Too many False Positives	Good customers are rejected, resulting in lost revenue and poor customer experience.
Too many False Negatives	Risky customers are approved, potentially causing large financial losses.
Threshold Tuning	Helps align model behavior with the bank's risk appetite and business goals.

5 Summary and Learnings

- The project aimed at predicting credit card default using a combination of exploratory data analysis, engineered features, and machine learning classification.

- **Feature Importance:** Figure 16 shows SHAP-based feature importance. The most influential feature is the `max_delinq_streak`, which captures a customer's longest consecutive month streak of missed payments — clearly differentiating defaulters.

- **Behavioral Indicators:** Other highly ranked features include recent `pay_amt` values (especially `pay_amt_june`, `may`, and `jan`), supporting the idea that recent payment history is a strong predictor of default.

- **Demographics:** `Marriage` and `Sex` variables are also ranked highly. This suggests demographic patterns exist in repayment behavior. For example, `marriage_married` being significant might reflect financial responsibilities influencing payment regularity.

- **Engineered Features:** The introduced features — `AVG_util_ratio` appear in the middle of the importance rankings and `max_delinq_streak` on top showing it's meaningful contribution.

- **Limit and Spending Behavior:** `LIMIT_BAL` and monthly `Bill_amt` figures indicate consistent spending patterns but are not as strong alone in predicting default as delinquency and payment data.

- **Model Performance:** XGBoost was chosen based on its highest F2-score, aligning with business goals of minimizing false negatives. SMOTE effectively balanced class distribution for better recall.

- **Business Takeaway:** Delinquency patterns and recent payment history are the most actionable signals. Financial institutions can flag early warning signs by tracking these patterns and apply stricter lending criteria.

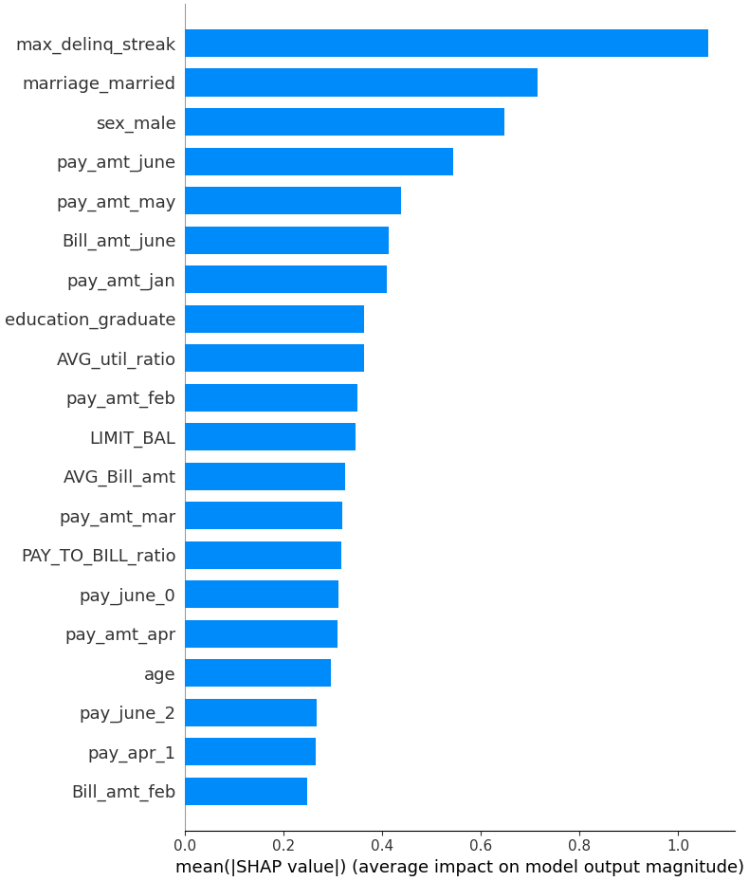


Figure 16: SHAP feature importance: average absolute impact on model output