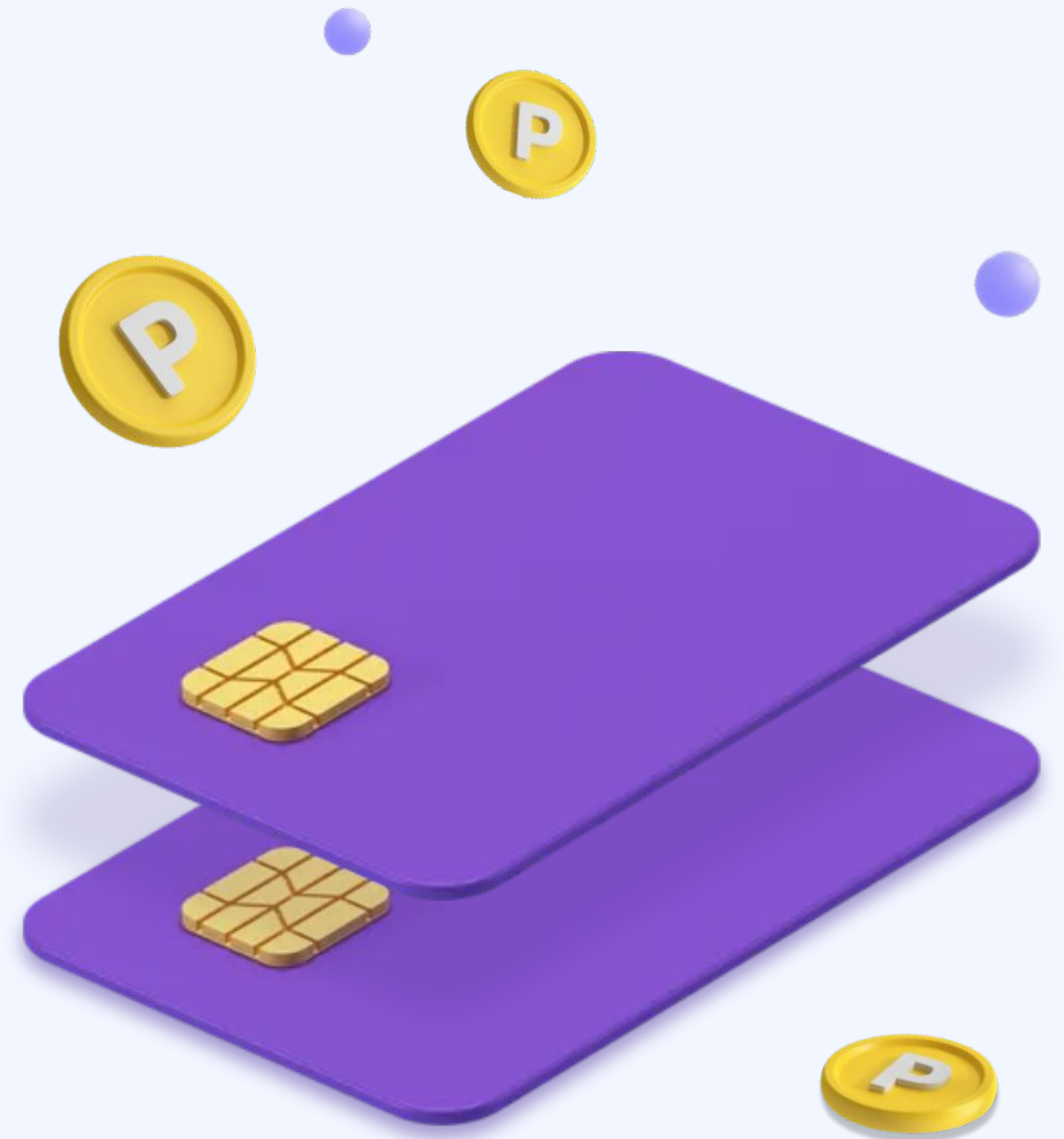


Predicting Loan Default Risk

By Melaku Mohammed and Christian Abrams



Problem Statement

- Banks make thousands of lending decisions daily, and accurately identifying high-risk borrowers is crucial for preventing financial losses and maintaining stable credit systems.
- Lenders need reliable risk signals for decision-making, especially when defaults are uncommon

Our Goal: Predict whether a borrower will default next month using financial & demographic data.



Duration

● April 2005 – September 2005

Location

● Credit card clients in Taiwan

Dataset: Credit Card Clients in Taiwan from April-Sept 2005

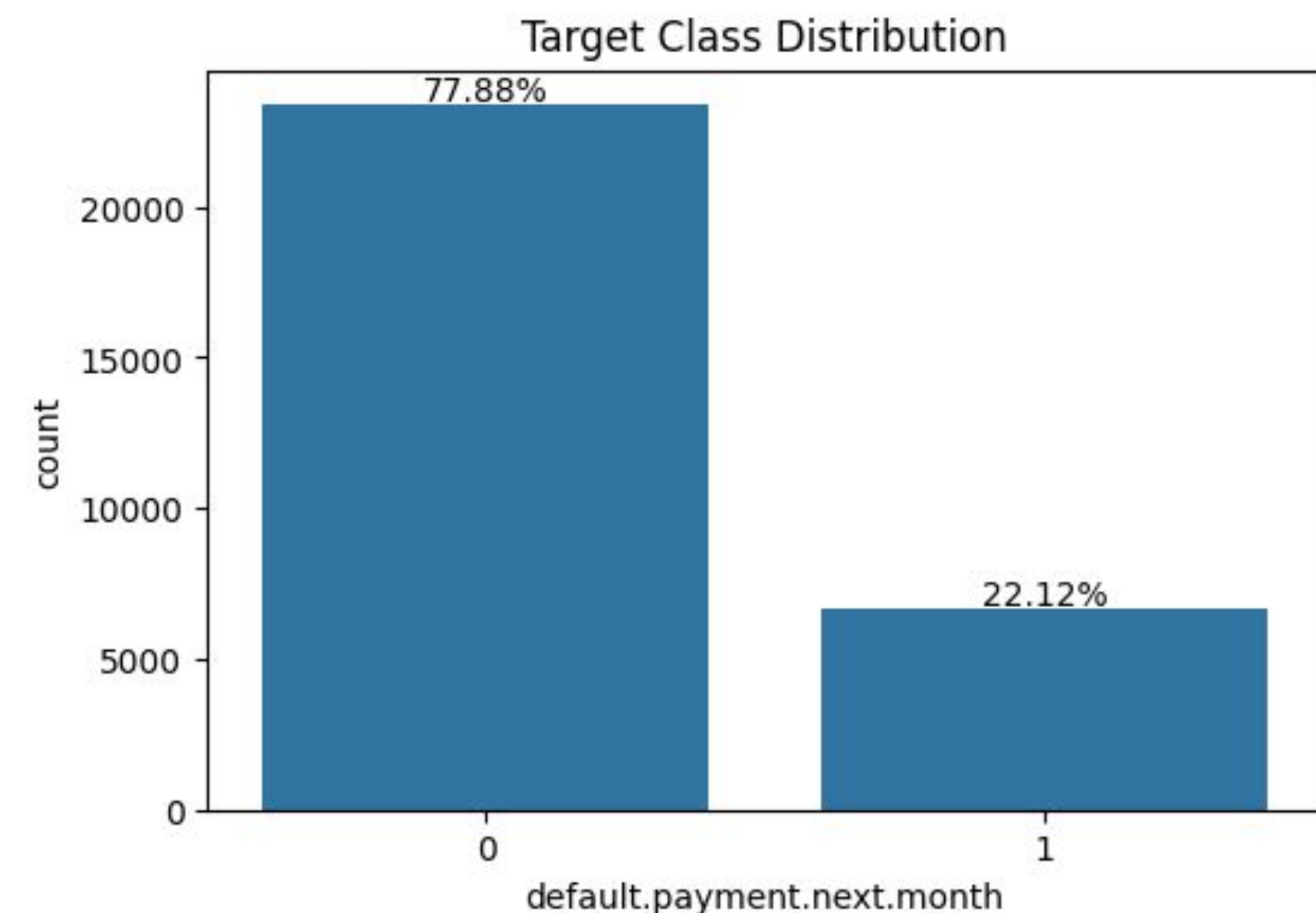
Size: 30,000 borrowers

25 Features:

- *Demographics* (SEX, EDUCATION, MARRIAGE, AGE)
- *Repayment status* (PAY_0–PAY_6)
- *Bill amounts* (BILL_AMT1–BILL_AMT6)
- *Payment amounts* (PAY_AMT1–PAY_AMT6)
- *Target:*
default.payment.next.month (1 or 0)

Target Class Distribution

The bar chart below shows a strong imbalance between those don't default and those who do, which is why SMOTE was needed



Methodology

Data Cleaning

- Filled missing values (e.g., Marriage)
- Removed outliers (e.g., bill amounts)
- Scaled features (For LR)
- Identified major class imbalance

EDA

- Examined repayment patterns (PAY_0–PAY_6)
- Identified & ranked strong predictors of default.

03

01

02



Modeling

- Logistic Regression
- Random Forest
- XGBoost
- LightGBM

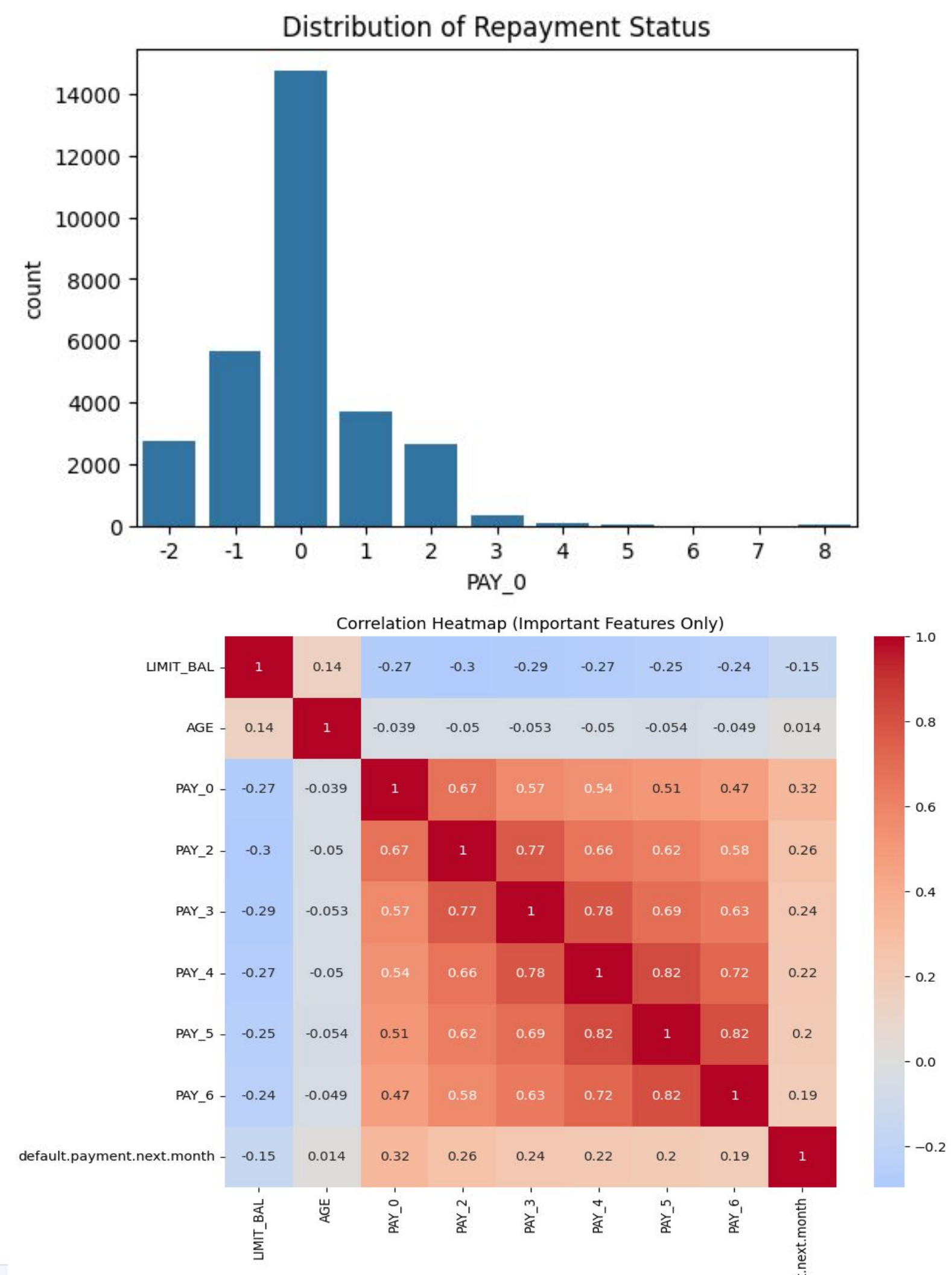
Evaluation Metrics

- Accuracy
- Precision
- Recall
- F1-Score
- AUC

03

04

EDA



Repayment Status Distribution:

- Shows the borrower's **most recent repayment status**, where 0 = paid on time, negative values = early/adjusted payment, and positive values = months late.
- Most borrowers are **on time** (0), while serious delinquency ($\text{PAY_0} \geq 1$) is rare but strongly linked to default risk.
- These rare late-payment cases highlight the class imbalance problem and explain why resampling (SMOTE) is needed for models to learn default patterns.

Correlation Heatmap:

- Repayment history** (PAY_0 – PAY_6) shows the **strongest correlations** with default, borrowers who are late in one month are often late in others.
- LIMIT_BAL** has a **moderate negative correlation** with default, meaning borrowers with higher credit limits tend to default less often.
- AGE** and other demographics show minimal correlation, reinforcing that financial behavior matters far more than demographic traits for predicting default.

Key Predictors

Repayment patterns overwhelmingly predict borrower default.

PAY_0

#1

Most Recent Repayment Status

- Strongest single predictor across all models.
- Late payments here signal early financial distress.

BILL_AMT1

#2

Most Recent Bill Statement

- Indicates current debt burden and spending behavior.
- Higher outstanding balances correlate with higher risk.

PAY_2

#3

Repayment Status Two Months Prior

- Very high correlation with PAY_0 and with default risk.
- Captures persistent or repeated delinquency.

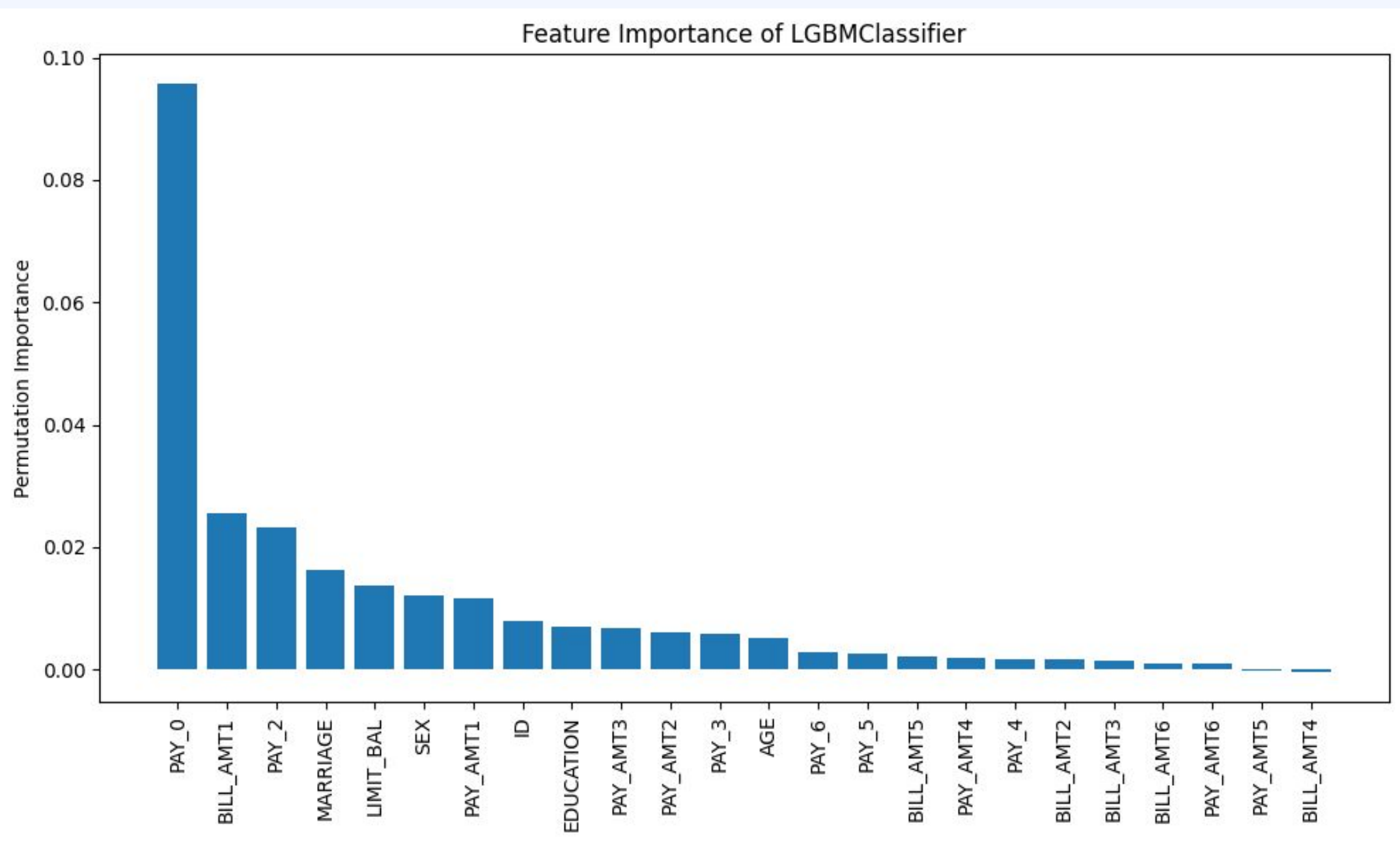
Model Results

Models Tested: Logistic Regression, Random Forest, XGBoost, LightGBM, Binary Classification Neural Network

Enhanced Models: Each trained again using SMOTE to address the strong class imbalance (22% default vs. 78% non-default).

Best Overall Model: LightGBM + SMOTE

- Achieved the best balance of recall (85%), precision (86%), and AUC (76%) among all models.
- Recall for defaulters improved substantially
 - Helped with high-risk borrowers that were initially missed.



Conclusions

1. Borrower default can be predicted effectively using **financial** and **behavioral** data.
2. **Repayment history** is the strongest signal of future default across all models.
3. **SMOTE** works really well for improving **recall** and allowing models to learn minority default patterns.
4. **LightGBM + SMOTE** offers the best overall balance and is the most practical model for real-world credit risk systems.



Thank You.

