Predicting credit default risk to reduce financial losses

By Abigael Kariuki

February 2026

Outline

- ❑ Business Problem
- ❑ Stakeholders
- ❑ Dataset
- ❑ Exploratory Data Analysis (EDA)
- ❑ Modelling
- ❑ Limitations
- ❑ Business Recommendations
- ❑ Conclusions

Business problem

Banks and financial institutions face significant losses when customers default on credit payments.

The goal of this project is to predict the likelihood of default next month, allowing lenders to:

- Reduce losses
- Improve risk-based pricing
- Intervene early with high-risk customers

We will build and evaluate models to predict default, using borrower demographics and credit behavior, with a focus on identifying defaulters.

Stakeholders

Primary stakeholders:

- Credit risk and risk analytics teams in banks and financial institutions

Secondary stakeholders:

- Product managers responsible for pricing and credit limits
- Collections teams prioritising outreach of potential customers
- Senior management monitoring portfolio risk

Dataset

❖ The dataset used is the UCI Credit Card Default dataset, which contains 30,000 observations of credit card clients.

❖ It includes:
  - Demographic information eg age, gender, education, etc
  - Credit limits
  - Historical outstanding debt amounts
  - Repayment behavior over six months
  - Whether the customer defaulted in the previous months and others

Variables

Target Variable

❖ Default status

1 = Default
0 = No default

- This is a binary classification problem.

Features (independent variables)

❖ There are 25 total feature variables in our dataset.

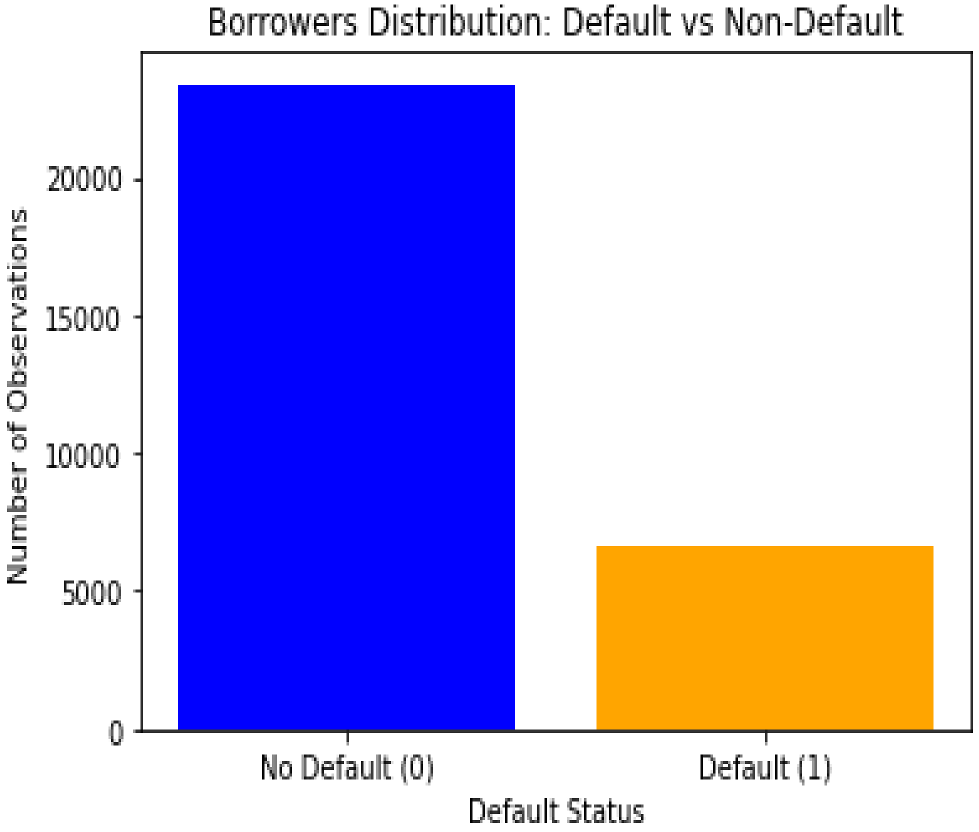
❖ All of them were used as predictors, excluding the unique identifier (`ID`). A few include:

-Credit limit (LIMIT_BAL)
-Demographic variables (AGE, EDUCATION, SEX, MARRIAGE)
-Repayment status history (PAY_0 to PAY_6)
-Monthly bill statement amounts (BILL_AMT1 to BILL_AMT6)
-Monthly payment amounts (PAY_AMT1 to PAY_AMT6)

Exploratory Data Analysis (EDA)

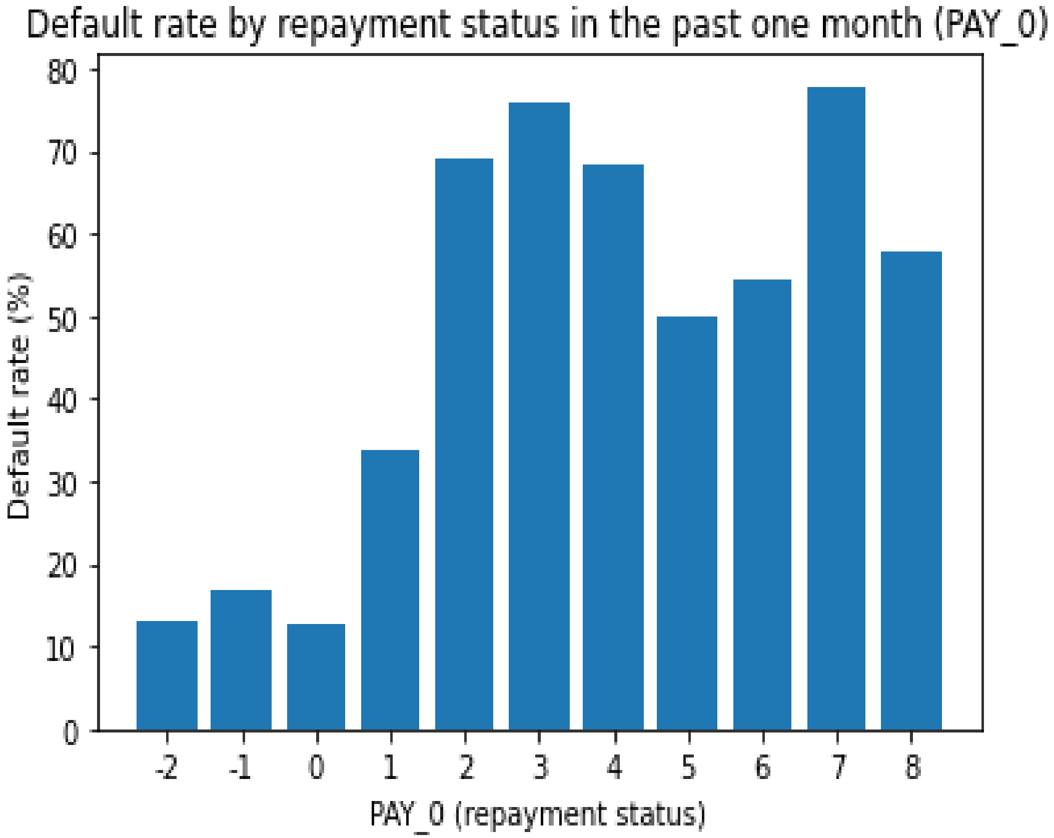
Distribution of borrowers: majority are non-defaulters, resulting in an imbalanced dataset

- The dataset has **30,000 borrowers**. The distribution of borrowers is as follows:
  - **Non-defaulters:** 23,364 (78%)
  - **Defaulters:** 6,636 (22%)
- This shows the dataset is highly imbalanced, with non-defaulters making up the majority.
- This creates a challenge for the model, as it may favor predicting non-defaulters.
- To address this, class weights are applied in the modelling to prioritize detecting defaulters and reduce false negatives *(i.e., borrowers who default but are predicted as non defaulters).*



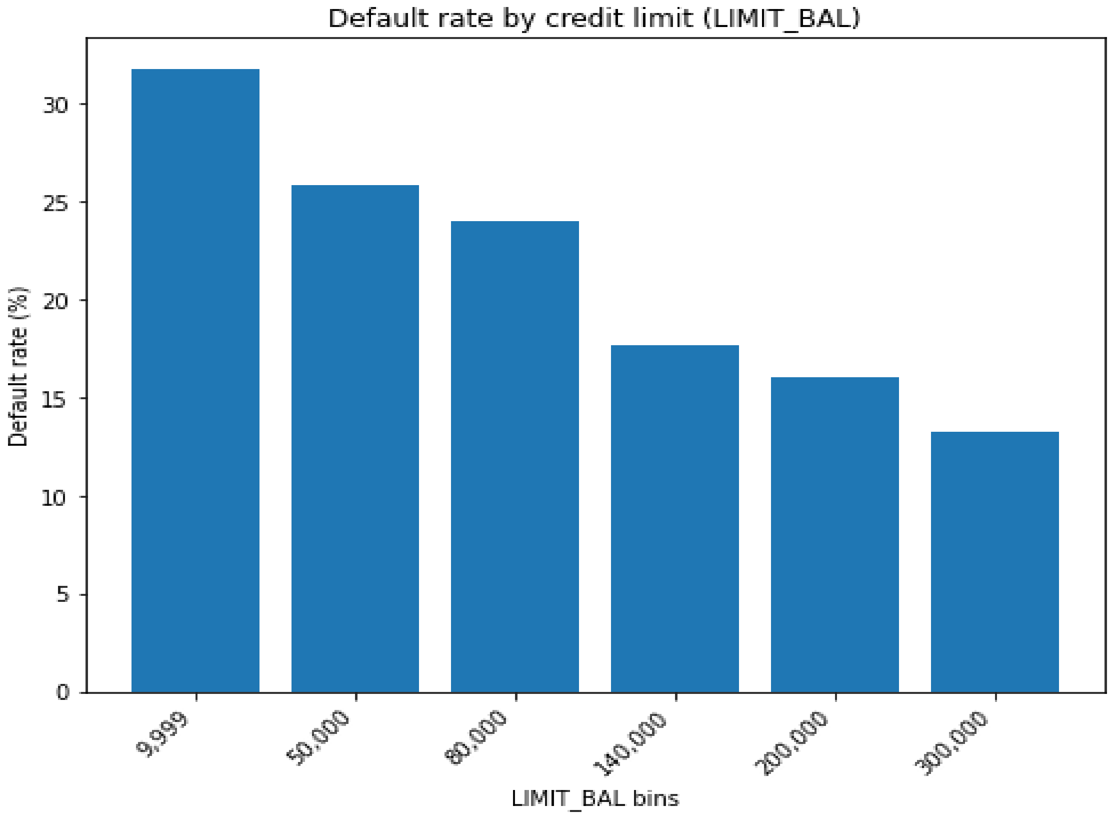
Credit default rate by repayment status: it increases significantly as repayment delays become more severe

- Borrowers with recent loan payments (status 0) have default rates below **10%**.
- Borrowers with recent payment delays (e.g., status 6 and above) show sharply higher default probabilities.
- Severe delays (status 6 and above) show default rates above **50–70%**.
- **Therefore, recent payment delays (delinquency) are a strong indicator of future loan defaults**



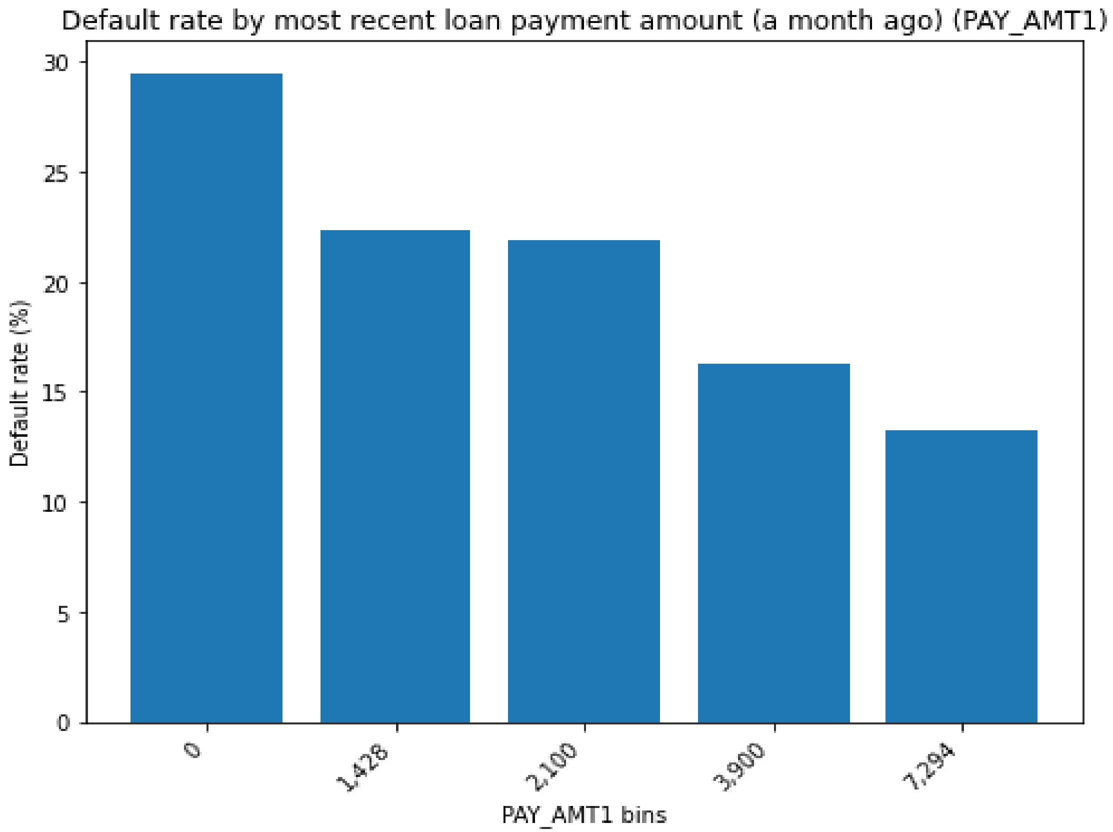
Credit default rate by credit limit: higher limits are associated with lower default risk

- Borrowers with higher credit limits tend to have lower default rates
- Those with lower credit limits tend to default more frequently, signaling that borrowers with lower financial capacity are more likely to default.
- Credit capacity is a key indicator of financial stability, with **higher credit limits correlating to a lower likelihood of default.**



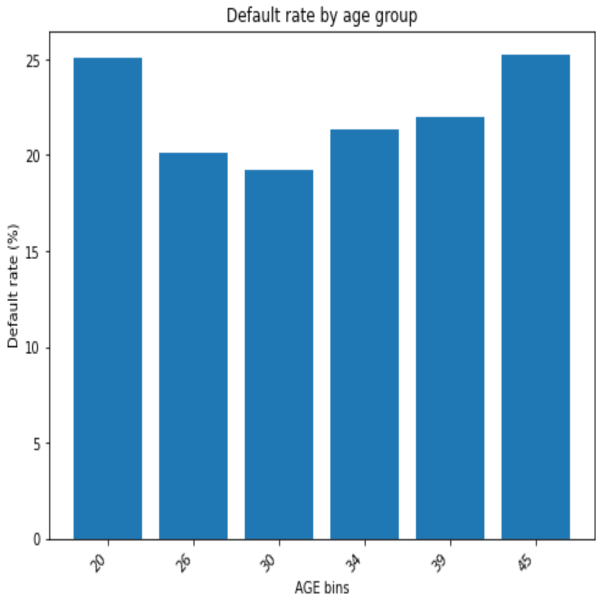
Credit default rate by recent loan payment amount: default risk decreases as payment amounts increase

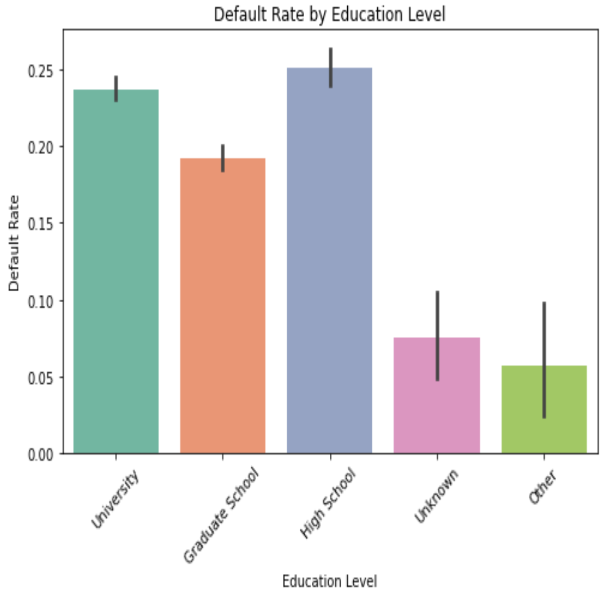
- Borrowers making smaller recent payments show higher default rates.
- As recent payment amounts increase, default risk steadily declines.
- **Customers making larger payments demonstrate stronger repayment capacity.**
- Repayment capacity is therefore an important indicator of credit stability.

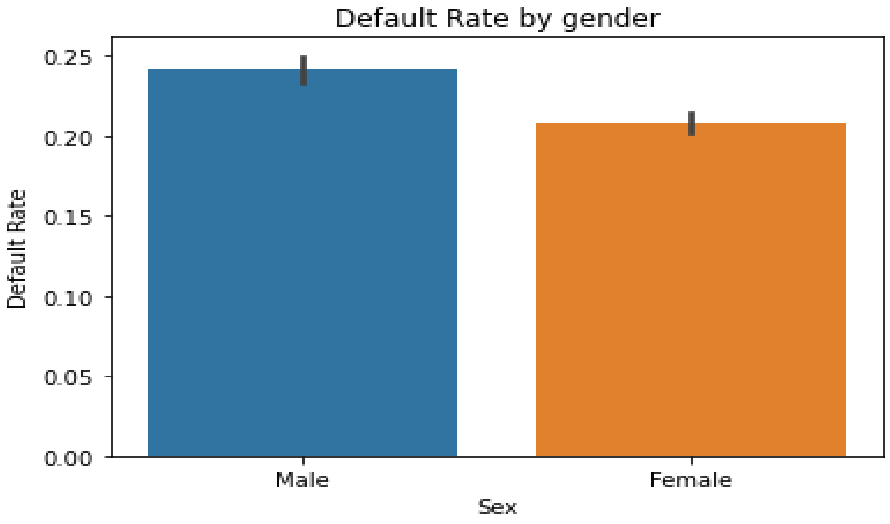


Credit default rate by borrower characteristics: demographic factors show limited variation in default risk

- Default rates vary slightly across age, gender, education, and marital status.
- Differences between demographic groups are not substantial.
- No strong pattern suggests that demographics alone drive default risk.
- **Behavioral factors (repayment history and payment behavior) remain more influential predictors of default.**

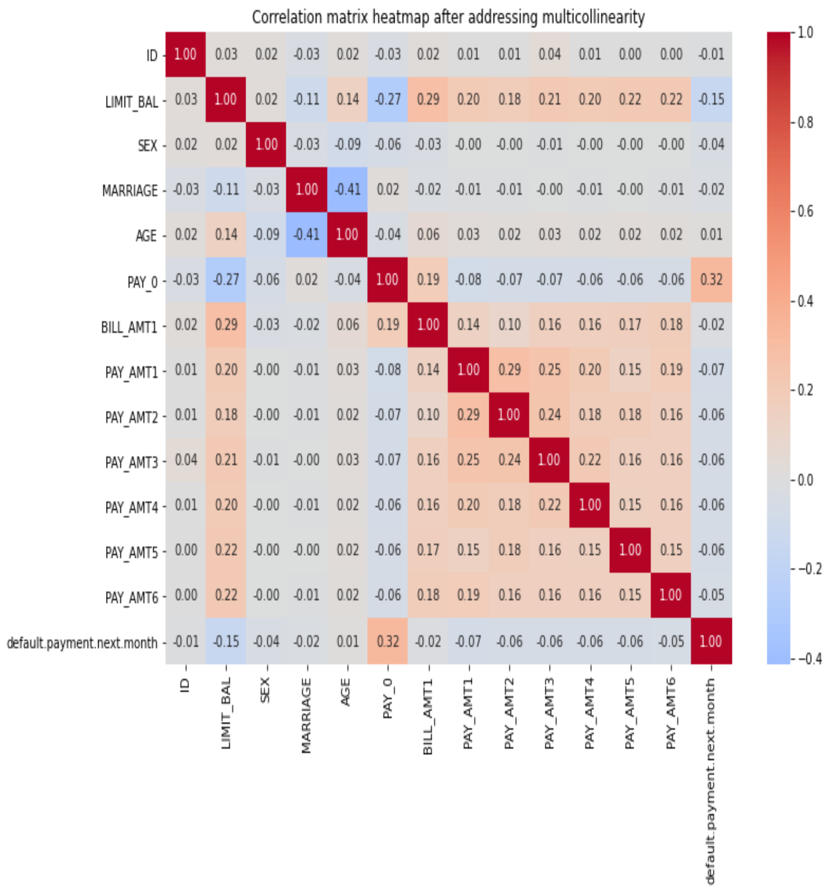






Addressing multicollinearity

- A multicollinearity check revealed strong correlations between several variables, particularly those linked to their past values.
- Highly correlated variables included:
    - **Repayment status**: PAY_2–PAY_6
    - **Bill amounts**: BILL_AMT2–BILL_AMT6
- To reduce multicollinearity, we retained:
    - **PAY_0** (most recent repayment status)
    - **BILL_AMT1** (most recent bill amount)
- After removing redundant variables, 12 features were used in the modelling process.
- This reduction in collinearity enhances **model stability** and **interpretability**


Correlation matrix heatmap after addressing multicollinearity

Modelling

- **Objective:** Predict the probability of a borrower defaulting.
- We focused on four models:
  - Logistic regression
  - Logistic regression with class weights (balanced)
  - Decision tree
  - Decision tree with class weights (balanced)

Model 1: Logistic Regression (baseline, without class weights)

The model was trained to predict borrower defaults. Evaluation metrics: **accuracy**, **precision**, **recall**, and **F1-score**

Focus was on maximizing **recall** to minimize false negatives (defaulters misclassified as non-defaulters)

The model achieved 81.1%, but it missed 75% of the actual defaulters (**low recall of 25%).**

Precision for defaulters was 70%, meaning when predicted, the model is correct 70% of the time.

False negatives are significant, with 995 defaulters incorrectly classified as non-defaulters.

Given the low recall, we added **class weights** to the model to help better detect and prioritize defaulter

Classification report

Accuracy = 81.1%

	Precision	Recall	F1-score
Non-defaulters(0)	0.82	0.97	0.89
Defaulters (1)	0.70	0.25	0.37

Confusion matrix

4,533 (TN)	140 (FP)
995 (FN)	332 (TP)

Model 2: Logistic Regression (with class weights)

The model was trained with **class weights** to handle class imbalance, focusing on improving recall for defaulters, which increased from 25% to 63%.

It correctly identified 831 out of 1,327 defaulters, cutting **false negatives nearly by half** (from 995 to 496)

Accuracy decreased from 81% in the baseline model to 7%; expected when prioritizing the minority class.

Reducing missed defaulters is more important than maximizing overall accuracy.

This model **outperformed** the baseline model.

We then did **decision tree model** to capture non-linear relationships, complex feature interactions and further improve recall for defaulters.

Classification report

Accuracy = 76%

	Precision	Recall	F1-score
Non-defaulters (0)	0.87	0.69	0.77
Defaulters (1)	0.36	0.63	0.46

Confusion matrix

3,219 (TN)	1, 454 (FP)
496 (FN)	831 (TP)

Model 3: Decision tree model (without class weights)

The model achieved an accuracy of 73%, with a recall of 40% for defaulters.

Recall for defaulters improved from 25% in the baseline logistic regression to 40%, compared to 63% in the balanced logistic regression model, reducing missed defaulters from 995 to 795

This model performed worse than the balanced logistic regression in handling class imbalance.

Therefore, we proceeded to retrain the model with class weights to better handle imbalance and improve recall.

Classification report

Accuracy = 72.7%

	Precision	Recall	F1-score
Non-defaulters(0)	0.83	0.82	0.82
Defaulters (1)	0.39	0.40	0.39

Confusion matrix

3,835 (TN)	838(FP)
795 (FN)	532 (TP)

Model 4: Decision tree model (with class weights)

The Decision Tree model (with class weights) achieved 73% accuracy, with a defaulters recall of 41%, identifying 538 out of 1,327 defaulters and missing 789 high-risk borrowers.

Compared to the logistic regression(with class weights whose recall was at 63%, the balanced Decision Tree performs worse at detecting defaulters.

Although it captures nonlinear relationships, increasing model complexity did not improve minority class detection.

Classification report

Accuracy = 73.4%

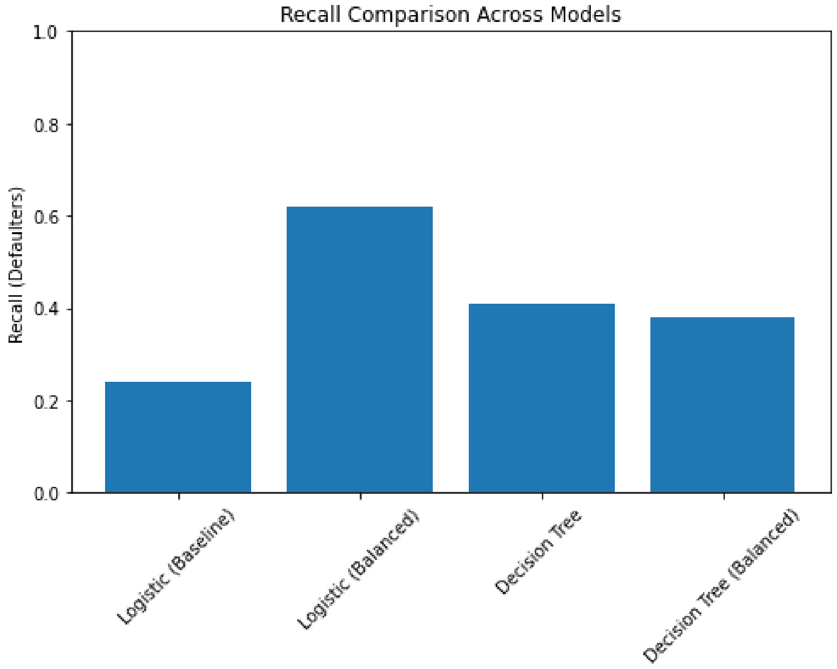
	Precision	Recall	F1-score
Non-defaulters(0)	0.83	0.83	0.83
Defaulters (1)	0.40	0.41	0.40

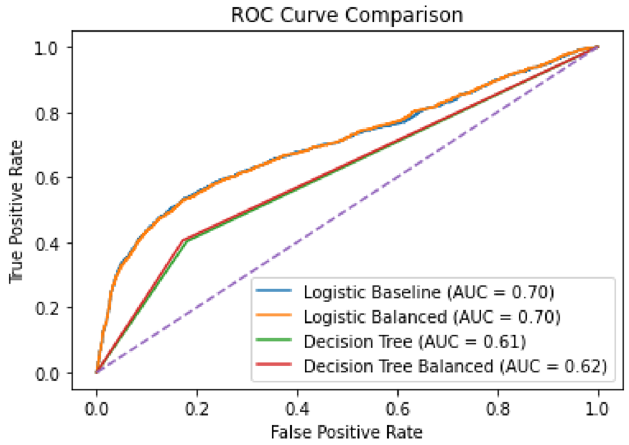
Confusion matrix

3,870(TN)	803 (FP)
789 (FN)	538 (TP)

Final model selection

- Evaluation approaches:
  - **Recall**: Prioritized for detecting defaulters. **Logistic regression (with class weights)** achieved the highest recall at **62%**.
  - **ROC AUC curve**: Measures the model's ability to separate defaulters. **Logistic regression (with class weights)** had the highest AUC at **0.71**.
- The **logistic regression (with class weights)** model significantly improved defaulter detection, outperforming the others.
- Given the business goal to minimize missed high-risk borrowers, it was selected.
- Its performance metrics are a recall of 62% for defaulters, a precision of 37%, an accuracy of 68%, and an AUC of 0.71

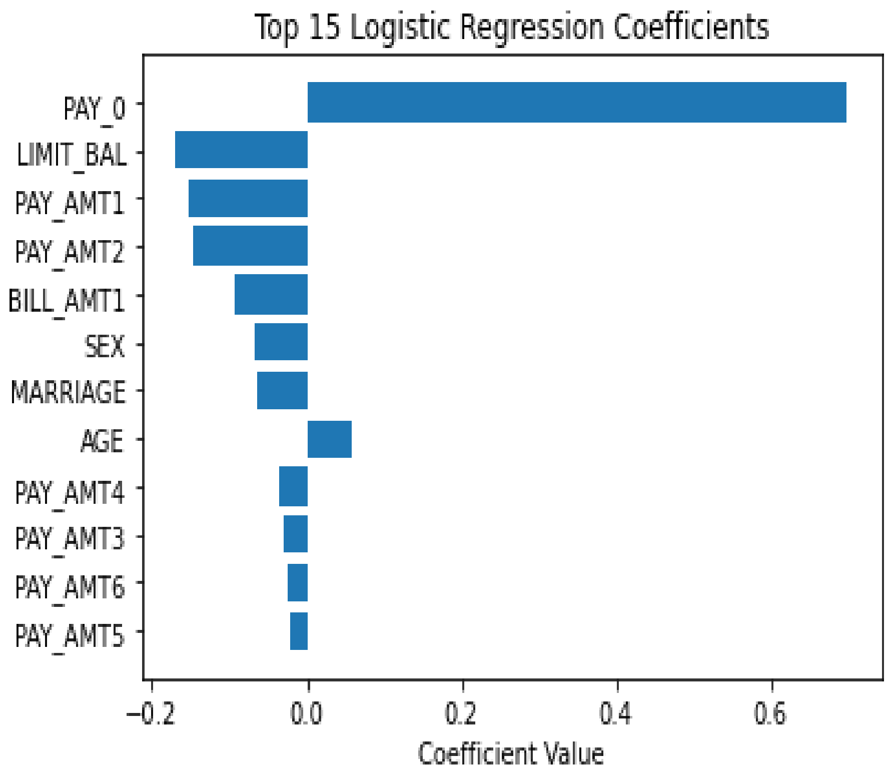




Features importance of Logistic regression (with class weights)

Following model selection, we analysed feature importance to identify key predictors of default risk.

- The most influential factor is recent repayment status (PAY_0), where higher values indicate more severe delays and increased default risk.
- Conversely, higher recent payment amounts (PAY_AMT1, PAY_AMT2) and credit limits (LIMIT_BAL) are linked to lower default risk, suggesting that better repayment behavior and higher credit capacity reduce the likelihood of default.



Conclusions

- Recent repayment behavior is the strongest driver of default risk.
- Higher credit limit reduces default risk (negative coefficient).
- Recent payment delays are a strong indicator of future default, highlighting the need for early intervention strategies for at-risk customers.
- Credit limit adjustments should be dynamic, rewarding good repayment behavior and minimizing exposure to high-risk borrowers.

Recommendations

- Closely monitor customers with recent missed or delayed payments, as they exhibit significantly higher default risk.
- Implement early intervention strategies such as reminder notifications or repayment support for at-risk customers.
- Always review credit limits since customers with stronger credit profiles and higher limits tend to have lower default risk.
- Use dynamic credit limit adjustments as part of ongoing risk management, rewarding consistent repayment behavior while reducing exposure to higher-risk borrowers.
- Deploy the model as a risk-screening tool to flag high-risk accounts for further review rather than relying solely on automatic rejection.

By proactively identifying early warning signs, banks can reduce credit losses and strengthen portfolio risk management.

Thank you!