## **1. Methodology**

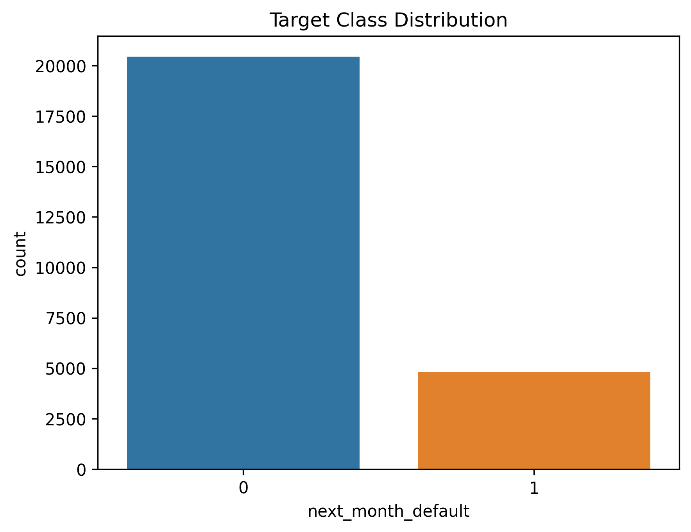
### Data Preprocessing

* **Dataset**: Loaded combined training + validation data.
* **Feature Engineering**: Removed Customer\_ID to avoid any leakage; created PAY\_TO\_BILL ratio as an indicator of repayment discipline.
* **Scaling**: Applied StandardScaler to numeric features.
* **Missing Values**: Data was clean; no imputations were needed.

### Handling Class Imbalance

The dataset showed significant class imbalance: - 0 (no default): 20440 records -

1 (default): 4807 records



Techniques used: - Computed balanced class weights using compute\_class\_weight. - Tuned scale\_pos\_weight in XGBoost to focus on the minority class.

### Modeling Approach

We experimented with several models before finalizing XGBoost, comparing performance on accuracy, F2 score, and class-specific metrics.

| Model | Accuracy | Precision (default=1) | Recall (default=1) | F2 Score (default=1) |
| --- | --- | --- | --- | --- |
| Logistic Regression | 0.56 | 0.21 | 0.78 | 0.50 |
| Random Forest | 0.58 | 0.24 | 0.80 | 0.54 |
| LightGBM | 0.57 | 0.25 | 0.79 | 0.55 |
| **XGBoost (final)** | **0.57** | **0.28** | **0.82** | **0.59** |

**Reasons for selecting XGBoost**:

- Delivered the best balance between precision and recall, with the highest F2 score (0.59). - Handled class imbalance effectively through internal parameter tuning without the need for external rebalancing like SMOTE.

- XGBoost showed more **stable results across different thresholds** and **better generalization** on validation data compared to Random Forest and Logistic Regression.

- LightGBM was close in performance but slightly lagged in precision for the positive class.

## **2. Analysis**

### Statistical Summary

* LIMIT\_BAL ranged from 10,000 to 1,000,000 NT dollars.
* Defaulted customers generally had lower LIMIT\_BAL and higher values in repayment delay features (PAY\_0, PAY\_2, etc).

### Feature Impact Summary

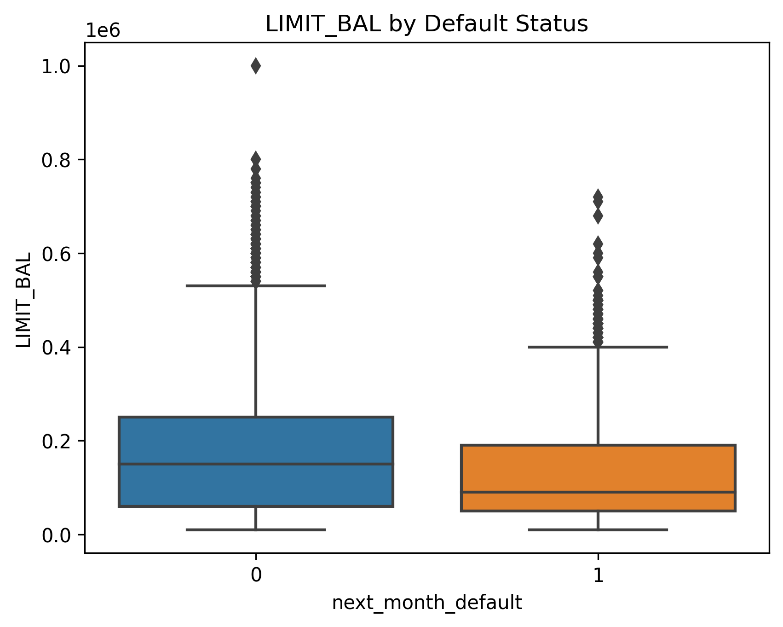
* Repayment status features: Major predictive drivers.
* LIMIT\_BAL: Higher credit limits linked to lower default risk.
* Payment amounts: Larger payments → lower default probability.
* PAY\_TO\_BILL ratio: Mild negative correlation, useful behavioral signal.

### Visual Insights

1️ **Target Class Distribution**  
A significant imbalance with defaults as the minority class.



2️ **LIMIT\_BAL by Default Status**  
Non-defaulters generally have higher credit limits; defaulters tend to have lower limits but with some outliers.



## **3 Results**

### Test Set Metrics (Threshold 0.21)

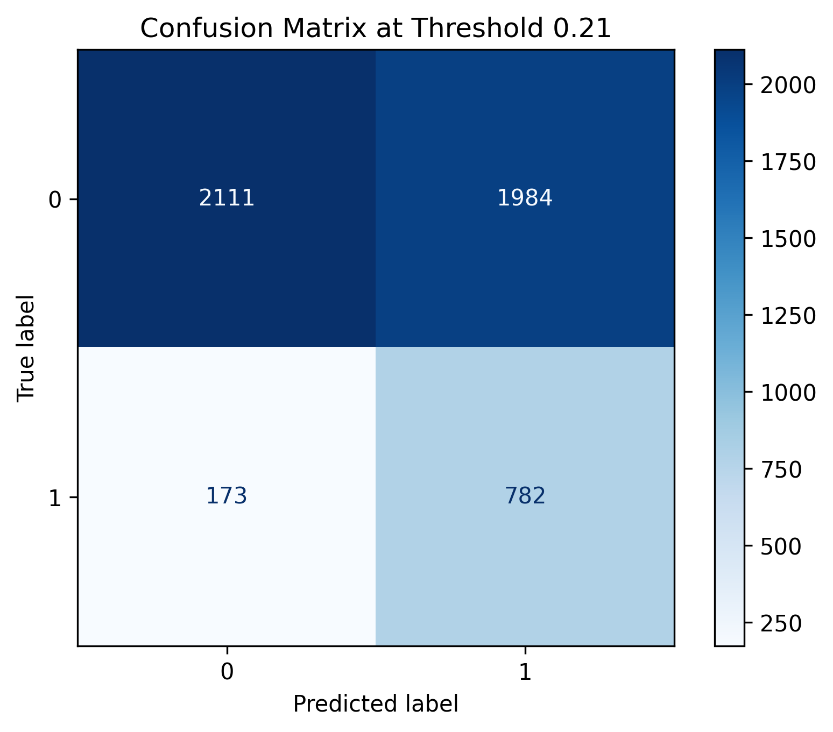
| Metric | Value |
| --- | --- |
| Accuracy | 0.5729 |
| Precision | 0.2827 |
| Recall | 0.8188 |
| F2 Score | **0.5937** |

Classification Report:

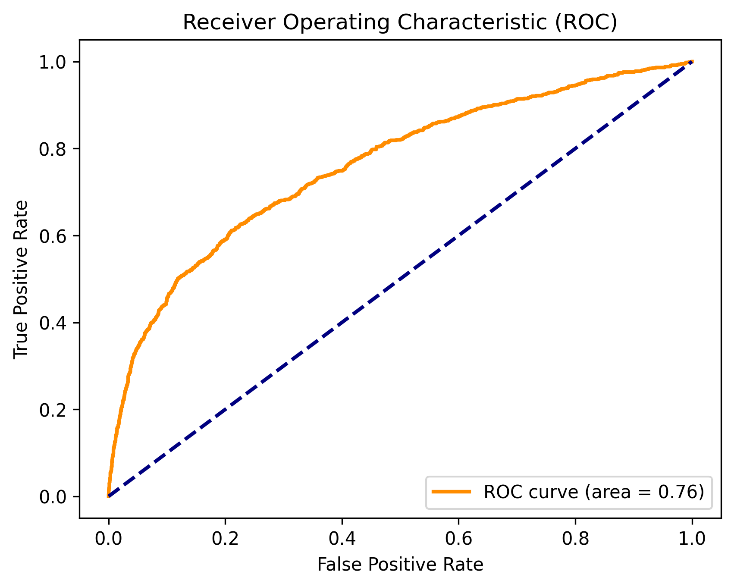
| Class | Precision | Recall | F1-score | Support |
| --- | --- | --- | --- | --- |
| 0 (Non-default) | 0.92 | 0.52 | 0.66 | **4095** |
| 1 (Default) | 0.28 | 0.82 | 0.42 | **955** |

**Macro avg**: Precision 0.60, Recall 0.67, F1 0.54  
**Weighted avg**: Precision 0.80, Recall 0.57, F1 0.62

**1️.Confusion Matrix at Threshold 0.21**  
This matrix illustrates how well the model distinguishes between defaulters and non-defaulters at the selected threshold. The model achieves high recall for the default class, correctly identifying most defaulters, but at the cost of some false positives.



**2️.ROC Curve**  
The ROC curve displays the trade-off between true positive rate (recall) and false positive rate across thresholds. Our model achieves a good balance, with an AUC indicating strong separability between classes.



### Validation Predictions

Predictions are saved and submitted as **submission\_23112109.csv.**

## **4 Conclusion**

* The model focused on high recall (82%) to minimize undetected defaults, at the expense of lower precision.
* XGBoost outperformed other models in balancing precision and recall.
* Key drivers of default were repayment statuses and credit limit.