# Loan Default Prediction Model - Detailed Report

## 1. Data Overview and Preprocessing

The dataset is highly imbalanced, with default (1) cases making up only 11.61% of the total:

- Non-default (0): 88.39%

- Default (1): 11.61%

- Total samples: 255,347 (Training set: 204,277; Validation set: 51,070)

Preprocessing steps:

- Categorical encoding using LabelEncoder:

- 'Education': {"Bachelor's": 0, "High School": 1, "Master's": 2, "PhD": 3}

- 'EmploymentType': {'Full-time': 0, 'Part-time': 1, 'Self-employed': 2, 'Unemployed': 3}

- 5 other similar fields

- Missing value imputation: median for numerical, most frequent for categorical.

- ID fields (e.g., LoanID) were removed to avoid information leakage.

## 2. Undersampling Strategy

To address class imbalance, we used RandomUnderSampler to downsample the majority class (non-default) at a 3:1 ratio.

Original class distribution in training set:

- Default (1): 23,722

- Non-default (0): 180,555

After undersampling:

- Default (1): 23,722

- Non-default (0): 71,166

- Total: 94,888 samples

## 3. Model Training and Hyperparameter Optimization

We used XGBoostClassifier and optimized hyperparameters using RandomizedSearchCV with 5-fold cross-validation.

The scoring metric was a custom F2-score, prioritizing recall over precision.

Best parameters found:

- subsample: 0.7

- n\_estimators: 500

- min\_child\_weight: 2

- max\_depth: 4

- learning\_rate: 0.03

- gamma: 0.1

- colsample\_bytree: 0.8

Final model was trained using these parameters on the resampled training set.

Best F2-score from CV: 0.5588

## 4. Classification Threshold Optimization

To better identify defaults, we optimized the classification threshold based on the precision-recall curve.

Optimal threshold: 0.2238

- Precision: 0.2125

- Recall: 0.7447

- F2-score: 0.4962

## 5. Validation Set Performance

Using the optimized threshold, the model was evaluated on the validation set:

Class 0 (Non-default):

- Precision: 0.95

- Recall: 0.64

- F1-score: 0.76 (Support: 45,139)

Class 1 (Default):

- Precision: 0.21

- Recall: 0.74

- F1-score: 0.33 (Support: 5,931)

Overall metrics:

- Accuracy: 0.65

- Macro average F1-score: 0.55

- Weighted average F1-score: 0.71

- ROC-AUC: 0.7591

- Validation set F1-score: 0.3307