Credit Default Prediction Report

Finance Club Open Project Summer 2025

Shubh Garg
23124035
BSBE

SECTION 01

Introduction and Project Overview

This project develops a machine learning model to predict credit card defaults (`next_month_default`) for a bank, using a dataset of ~25,000 training records and ~5,000 validation records. Features include credit limit (`LIMIT_BAL`), repayment status (`pay_0–pay_6`), bill amounts (`Bill_amt1–Bill_amt6`), payment amounts (`pay_amt1–pay_amt6`), and demographics (`age`, `sex`, `education`, `marriage`).

Key Deliverables:

- Comprehensive Jupyter notebook with EDA, preprocessing, feature engineering, modeling, and SHAP explainability
- Prediction file (`final_predictions.csv`) with Customer_ID and next_month_default for validation set
- This comprehensive report summarizing methodology, findings, and business implications

25K
TRAINING RECORDS

5K
VALIDATION RECORDS

22%
DEFAULT RATE

ROC-AUC
PRIMARY METRIC

SECTION 02

Exploratory Data Analysis

2.1 Data Characteristics

The training dataset contains 25,000 records with 24 features. Missing values were limited to age (~5%), which were imputed with the median value of 35 years. The dataset exhibits a class imbalance with a 22% default rate, necessitating resampling techniques like SMOTEENN.

2.2 Key Findings

📊 **Repayment Behavior**

Customers with pay_0=0 (no delay) have 10% default rate, while those with pay_0≥2 exceed 50% default rate

💳 **Credit Limits**

Lower credit limits (<50,000) show default rates up to 30%, compared to 15% for higher limits (>200,000)

👥 **Demographics**

Males have slightly higher default rate (24%) than females (20%), with education level showing similar patterns

SECTION 03

Preprocessing and Feature Engineering

3.1 Preprocessing Steps

`Missing Values`
Imputed age with training median (35)

`Outliers`
Capped at 99th percentile

`Encoding`
One-hot encoded categorical variables

`Class Balance`
Applied SMOTEENN (60% defaults, 40% non-defaults)

`Scaling`
Standardized using StandardScaler

3.2 Feature Engineering

Behavioral features were engineered to capture repayment and utilization patterns:

`delay_count`
Number of delayed months

`mean_util`
Average credit utilization

`bill_trend`
Slope of bill amounts over time

`delay_x_mean_util`
Interaction between delays and utilization

SECTION 04

Model Development and Evaluation

4.1 Model Comparison

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.774	0.433	0.598	0.502	0.814
Decision Tree	0.704	0.338	0.579	0.427	0.704
Random Forest	0.767	0.422	0.606	0.498	0.823
XGBoost	0.774	0.431	0.577	0.493	0.809
LightGBM	**0.785**	**0.450**	**0.570**	**0.503**	**0.823**

4.2 Threshold Selection

Multiple thresholds were evaluated to balance precision and recall:

0.3
Precision: 0.410
Recall: 0.650

0.4
Precision: 0.443
Recall: 0.577
SELECTED

0.5
Precision: 0.474
Recall: 0.550

0.6
Precision: 0.495
Recall: 0.518

A threshold of 0.4 was selected to maximize recall (0.577) while maintaining reasonable precision (0.443), prioritizing the detection of risky customers.

SECTION 05

Business Implications

⚠️ **Risk Mitigation**

High-risk customers (delay_count≥2, mean_util>0.7) can be flagged for interventions like credit limit reductions or payment reminders

💰 **Cost Trade-offs**

The 0.4 threshold increases recall but also false positives. Higher thresholds may reduce unnecessary interventions

🎯 **Customer Segmentation**

Younger customers with delays and high utilization require targeted credit education or stricter limits

📊 **Monitoring Dashboard**

SHAP insights can be integrated into real-time dashboards for dynamic risk assessment

SECTION 06

Key Learnings

Critical Success Factors

- **Risk Drivers:** Repayment delays (pay_0, delay_count) and credit utilization (mean_util) are the strongest predictors
- **Model Choice:** LightGBM outperforms simpler models due to its ability to handle complex, non-linear patterns
- **Interpretability:** SHAP provides clear insights enhancing stakeholder trust
- **Threshold Tuning:** Balancing recall and precision is critical for business alignment

SECTION 07

Future Work

`Temporal Features`
Incorporate seasonal patterns and macroeconomic indicators

`Ensemble Models`
Combine LightGBM and XGBoost for improved performance

`Real-Time Integration`
Deploy SHAP-powered live risk dashboard

`Cost-Sensitive Learning`
Incorporate misclassification costs for financial optimization

CONCLUSION

Project Summary

The LightGBM model, with a cross-validated ROC-AUC of **0.8208 ± 0.0055**, effectively predicts credit defaults by leveraging key behavioral features like `delay_count` and `mean_util`. The selected threshold of 0.4 balances risk detection (recall: 0.577) with actionable precision (0.443), supported by SHAP insights for model interpretability.

This solution equips the bank with a robust tool for managing credit risk, providing clear strategies for risk mitigation and future enhancements. The final predictions for the validation set have been successfully generated and saved in `final_predictions.csv`.