Credit Card Default Prediction - Finclub Summer Project 2 (2025)

ABHISHEK GOEL (23113007)

1. Overview of Approach and Modeling Strategy

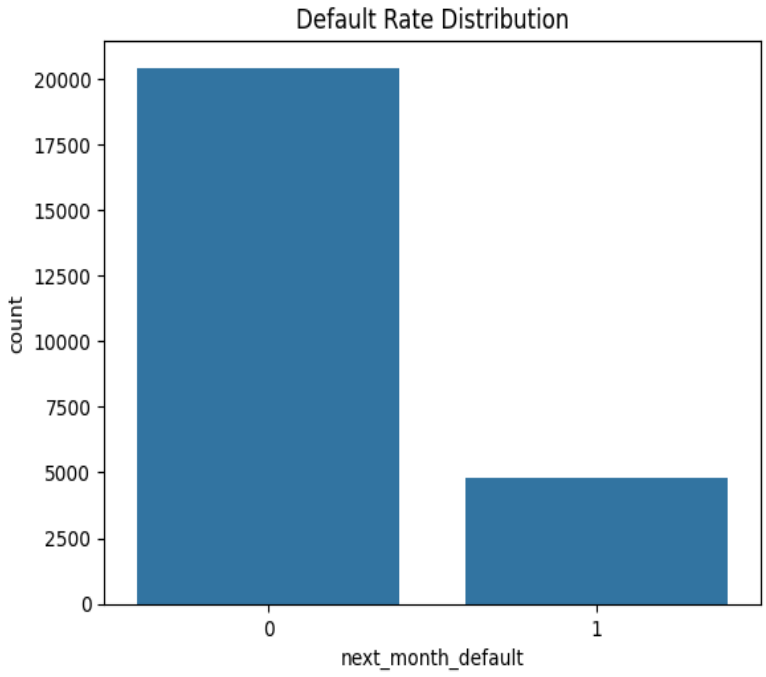
The objective of this project was to build a classification model to predict whether a credit card customer will default in the next month. The process involved the following steps:

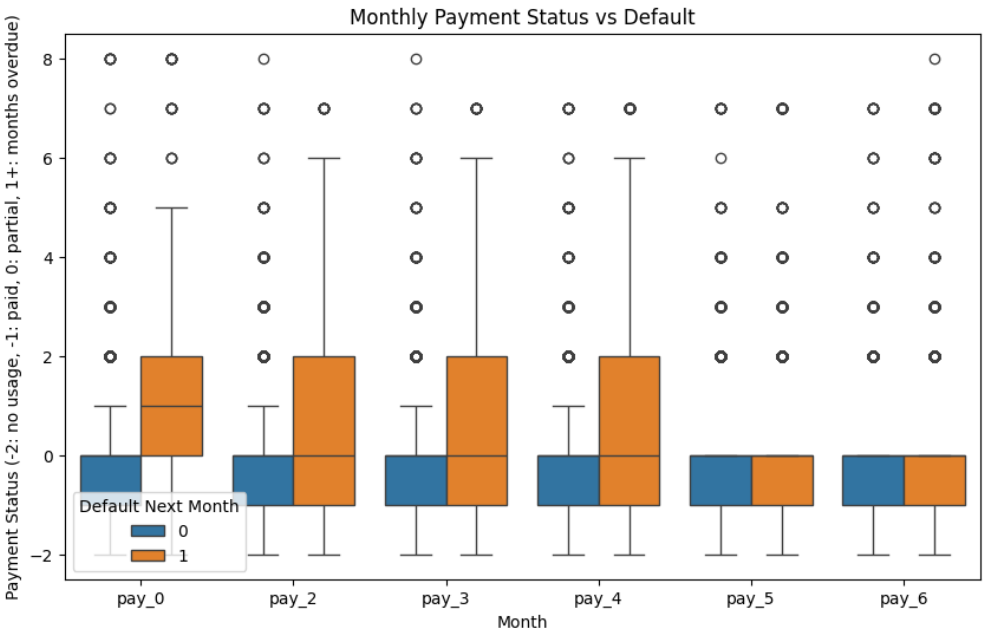
- Loading and cleaning the training and validation datasets.
- Performing Exploratory Data Analysis (EDA) to uncover important trends.
- Engineering meaningful financial features based on domain understanding (e.g., repayment behaviour, utilization).
- Training multiple machine learning models (Logistic Regression, Decision Tree, XGBoost, LightGBM).
- Comparing models based on F1, F2, and ROC AUC scores.
- Selecting the best classification threshold aligned with credit risk business priorities.
- Generating predictions for an unlabelled validation dataset.

2. EDA Findings and Visualizations

Key insights from EDA:

- Class imbalance observed: ~**16%** defaulters and **84%** non-defaulters.
- Higher payment delays (values >= 1 in pay_m columns) were strongly associated with defaults.
- Customers with high repayment inconsistency or high credit utilization were more likely to default.
- Boxplots of monthly payment status indicated defaulters consistently had higher delay levels.



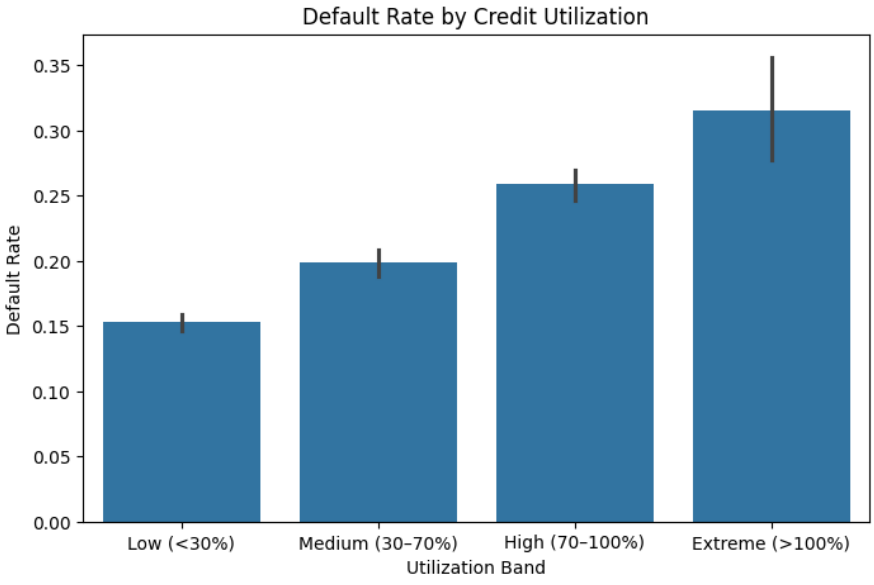


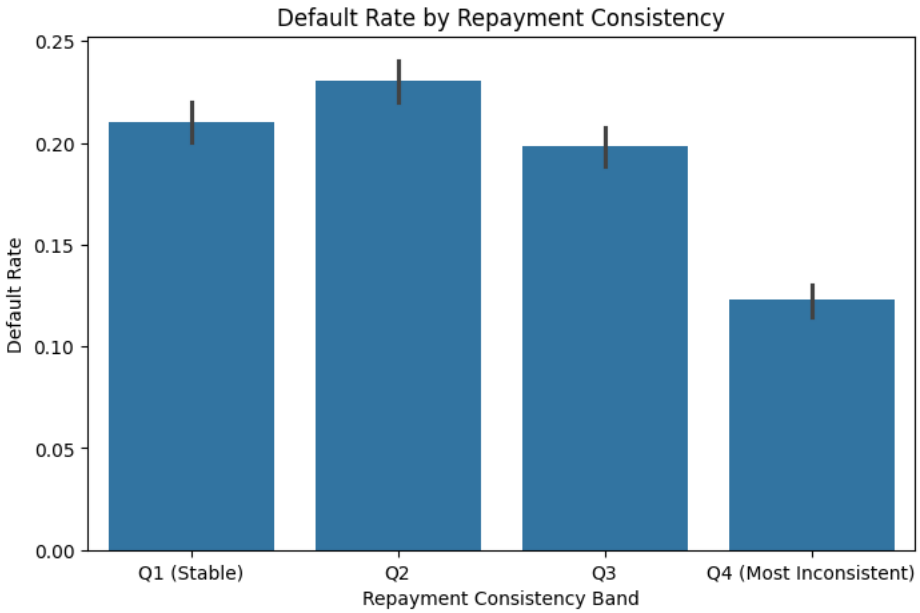
3. Financial Analysis of Default Drivers

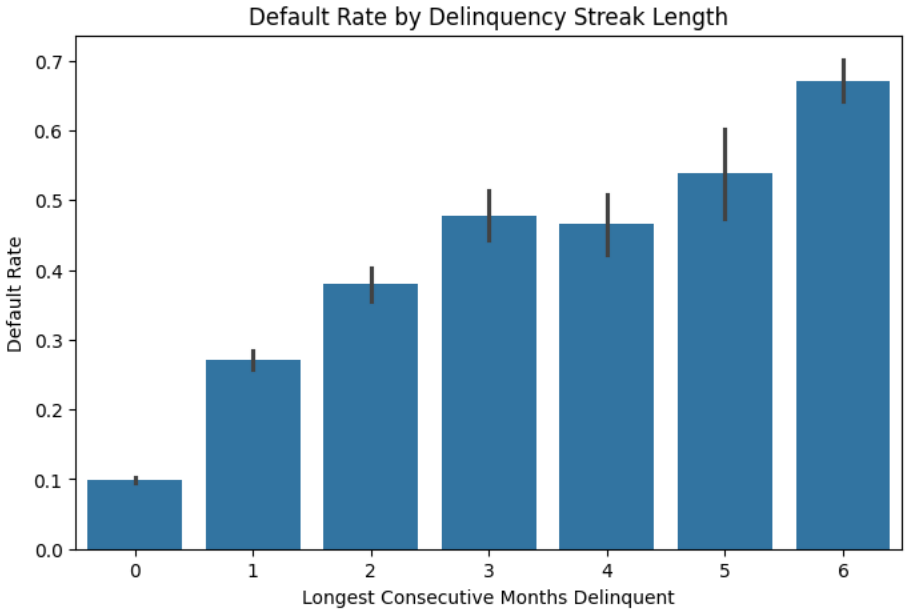
Engineered several financial variables to capture key behaviors:

- **Credit_Utilization_Ratio**: Ratio of average billed amount to credit limit. High values signal financial stress.
- **Repayment_Consistency:** Standard deviation of past payments. High inconsistency indicates payment instability.
- **Delinquency_Streak:** Longest run of months with overdue payments. Longer streaks indicate chronic delay.
- **On_Time_Payment_Ratio**: Proportion of months paid on time. Lower values are red flags.

These features align with real-world risk management practices.







4. Model Comparison and Final Selection

Evaluated four models using a pipeline (imputer → scaler → classifier). Results on the test set:

Model	F1 Score	F2 Score	ROC AUC
Logistic Regression	0.503	0.556	0.771
Decision Tree	0.351	0.348	0.599
XGBoost	0.492	0.519	0.753
LightGBM	0.508	0.565	0.779

LightGBM was selected due to its superior performance in F1, F2 and ROC AUC, with balanced results.

5. Evaluation Methodology

Prioritized the following metrics:

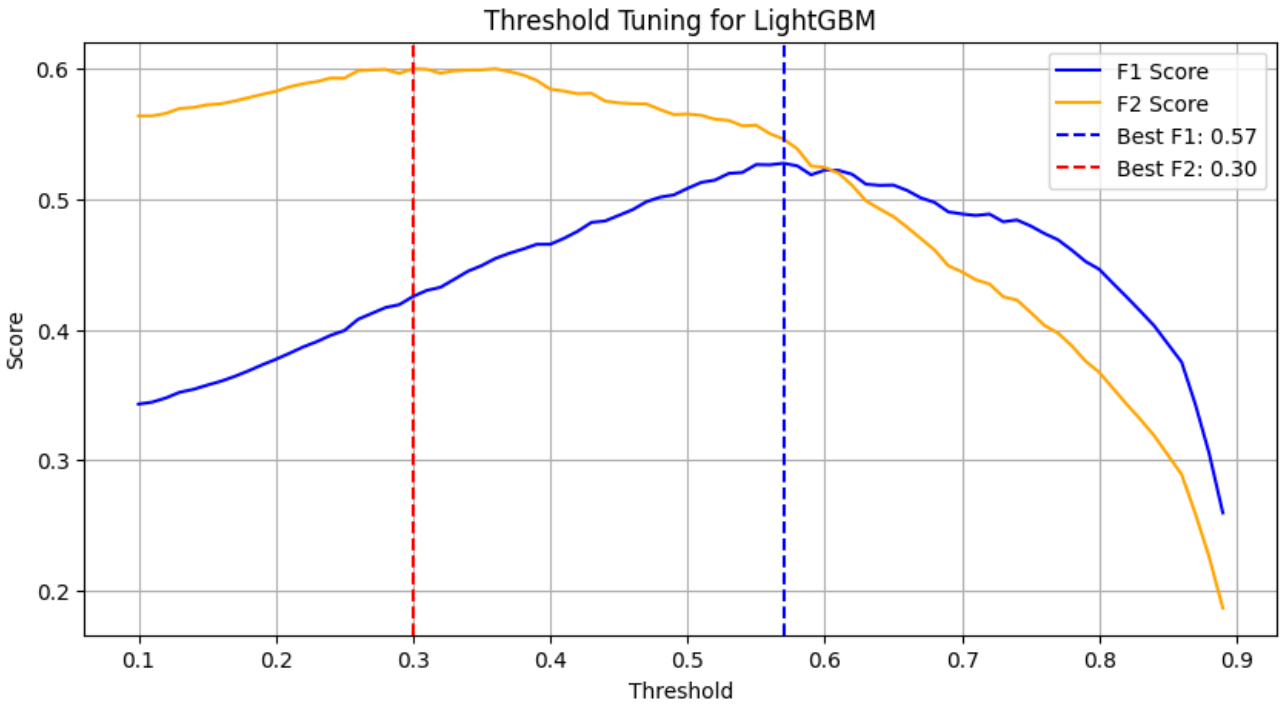
- **F1 Score**: Balances precision and recall.
- **F2 Score**: Emphasizes recall — important to catch as many defaulters as possible.
- **ROC AUC**: Measures model's ability to distinguish between defaulters and non-defaulters.

While the base LightGBM model yielded an F2 score of **0.565** at the default threshold of **0.5,** further threshold tuning revealed that the model could achieve a maximum F2 score of **0.6002** at a threshold of **0.30**. This highlights the importance of threshold optimization based on business priorities like recall.

Need of Threshold Tuning

F2 score helps understand how aggressive the model can be in identifying defaulters, as it places greater emphasis on recall. This means the model will try to catch as many actual defaulters as possible, even if it results in a higher number of false positives. In the evaluation, the best F2 score was achieved at a low threshold of **0.30**, where the model predicted nearly all customers as defaulters (approximately **2795** out of the validation set). While this maximizes recall, it also severely compromises precision, causing operational inefficiencies and potential customer dissatisfaction due to over-flagging.

To avoid this, a more balanced threshold of **0.47** was selected, which reduced the number of predicted defaulters to **1603**. Although this resulted in a slightly lower F2 score **(0.5732)** , it significantly improved the F1 score **(0.4982)** , offering a more practical balance between catching true defaulters and limiting false alarms. This threshold was chosen to align better with the bank's operational constraints and risk tolerance.



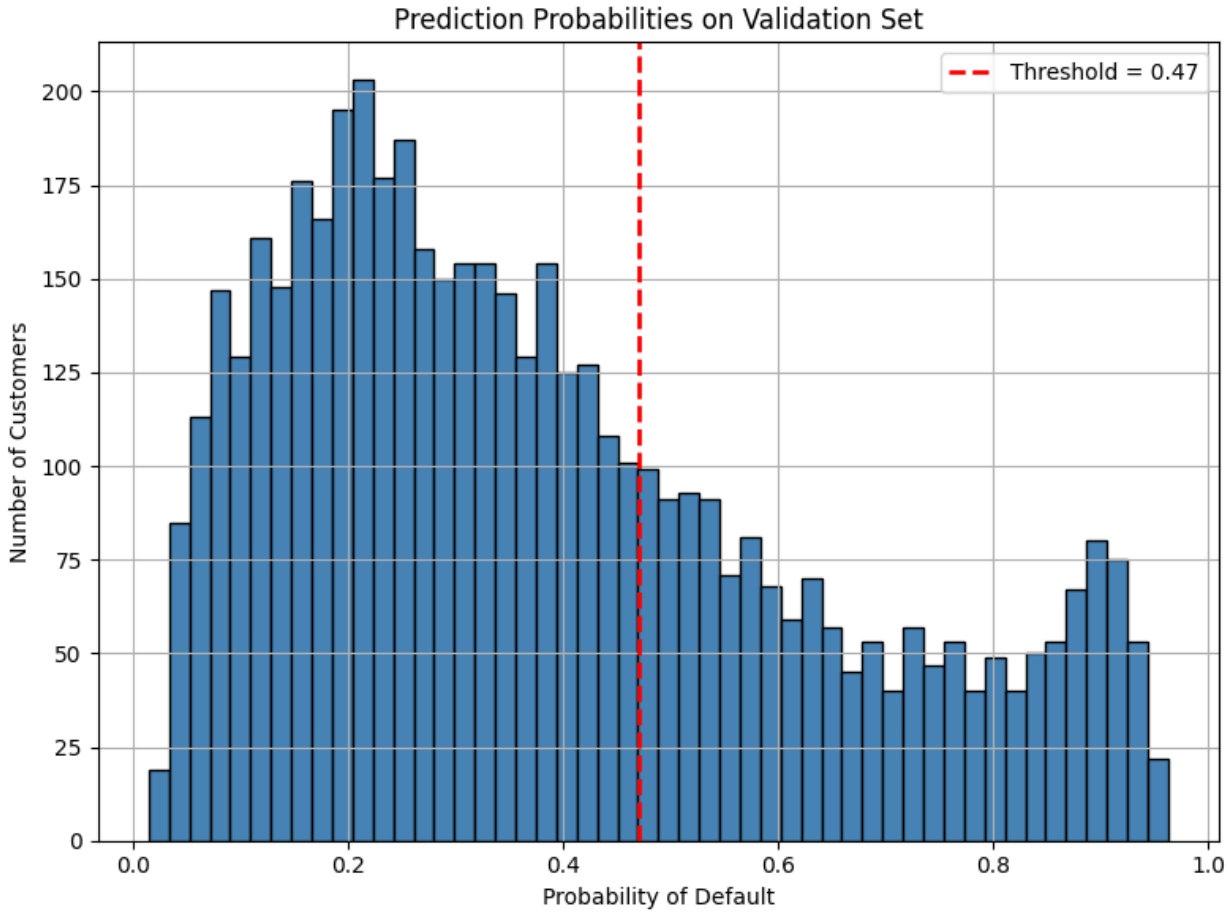
6. Classification Cutoff Selection

Threshold sweep was performed on test probabilities:

- **Best F1 Threshold**: 0.57
- **Best F2 Threshold**: 0.30
- **Selected Trade-off Threshold**: 0.47

Why 0.47?

This choice balanced F2 score (recall) and F1 score (overall effectiveness), suitable for operational deployment. Specifically, the threshold of 0.47 was chosen through a trade-off analysis: I identified the threshold range that yielded at least **95%** of the maximum F2 score and then selected the point within that range which offered the highest F1 score. This method ensured that recall remained strong (critical for minimizing missed defaulters), while F1 was optimized to prevent excessive false positives. Thus, 0.47 emerged as the optimal point balancing both performance and business practicality.



7. Business Implications

1. Improved Risk Management

- Model helps proactively identify high-risk customers.
- Enables preventive measures such as revising credit limits or enforcing payment plans.

2. Data-Driven Decision Making

- Scalable credit assessment beyond manual underwriting.
- Utilizes behavioural data for dynamic risk segmentation.

3. Better Customer Segmentation

- Enables personalized strategies for different risk groups.
- High-risk customers may get alerts, while low-risk customers may receive better credit offers.

4. Revenue Protection

- Reduces risk of bad debt through high recall.
- Controlled false positives prevent reputation damage.

5. Balanced Intervention

- Threshold of 0.47 strikes a balance between missing defaulters and over-penalizing good customers.

8. Summary of Findings and Key Learnings

Key Findings

- **Behavioural features drive default risk**: repayment inconsistency, delinquency streak, and credit utilization ratio were highly predictive.
- **Severe class imbalance**: required F1/F2/ROC AUC for fair model comparison.
- **LightGBM outperformed**: Highest F2 and ROC AUC, with good generalization.
- **F2 score guided threshold tuning**: Model capable of aggressive defaulter detection.
- **Threshold 0.47**: Balanced sensitivity and precision for deployment.
- **Prediction Results**: With threshold 0.47, predicted **1603 defaulters** in validation set.

Key Learnings

- **Threshold tuning is strategic**: Translates raw probabilities into business-friendly decisions.
- **Feature engineering is impactful**: Informed variables significantly enhanced performance.
- **Interpretability aids adoption**: Transparent features enable stakeholder trust.
- **Validation alignment is critical**: Matching training and validation pipelines ensured error-free predictions.