Credit Card Behaviour Score Prediction Using Classification and Risk-Based Techniques

Submitted by: Raghav Kumar | Enrollment Number: 23112079 | Program: Chemical Engineering
Submission Date: 16 June 2025

Abstract

This project aims to predict short-term credit card defaults using behavioral data of a bank's customers.

The target variable, **next_month_default**, indicates whether a customer will default on their next billing cycle.

We built a financially interpretable classification model for Banks by conducting detailed exploratory analysis and engineering features like credit utilization and repayment consistency. Class imbalance was addressed using SMOTE, and models such as Logistic Regression, XGBoost, and LightGBM were evaluated. Emphasis was placed on F2-score and recall to align with the bank's risk priorities.
The final model achieved strong predictive performance and provided valuable financial insights for managing credit risk and enabling early interventions.

1. Data Understanding and Preprocessing

The dataset provided contains 25,247 records and 27 columns, each representing a credit card customer's demographic information, behavioral history, and the binary target variable next_month_default. This target indicates whether a customer defaulted in the following billing cycle, with 1 representing default and 0 non-default.

1.1 Data Overview

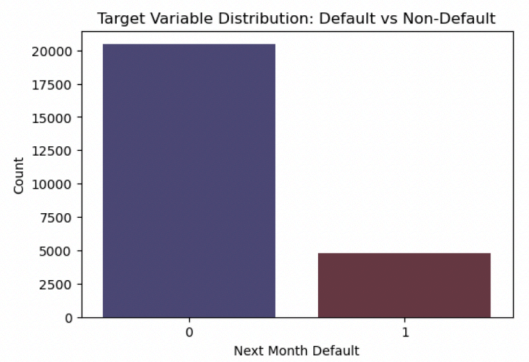
- Dataset Shape: 25,247 rows × 27 columns
- Missing Values: Only the age column had 126 missing values, which were imputed using the median.
- Data Types: Most columns were either int64 or float64. Categorical features such as sex, education, and marriage were explicitly converted to category type for memory efficiency and model compatibility.

1.2 Categorical Feature Handling

- sex: Mapped to 0 (female) and 1 (male), with all other values mapped to -1 (others).
- education: Consolidated into 4 categories, with values >4 grouped as 'others'.
- marriage: Consolidated into 3 categories, with values >3 grouped as 'others'.

1.3 Target Variable Distribution

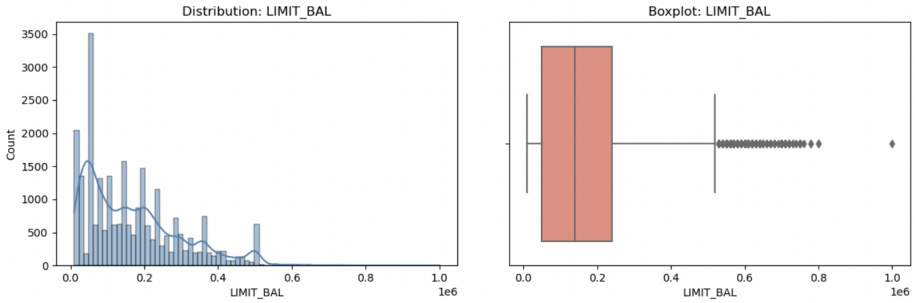
Exploratory analysis of the target variable revealed significant class imbalance, with roughly 80% of records labeled as non-default (0) and 20% as default (1). This imbalance **will require dedicated handling using techniques such as SMOTE**, class weighting, or threshold tuning during model training.



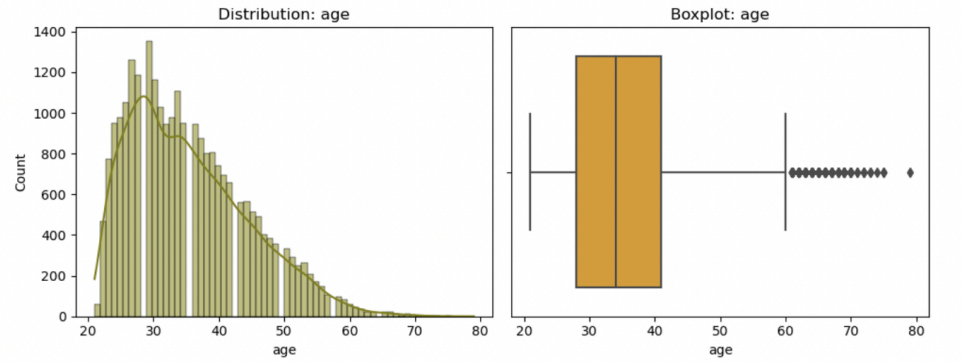
2. Exploratory Data Analysis (EDA)

2.1 Univariate Analysis

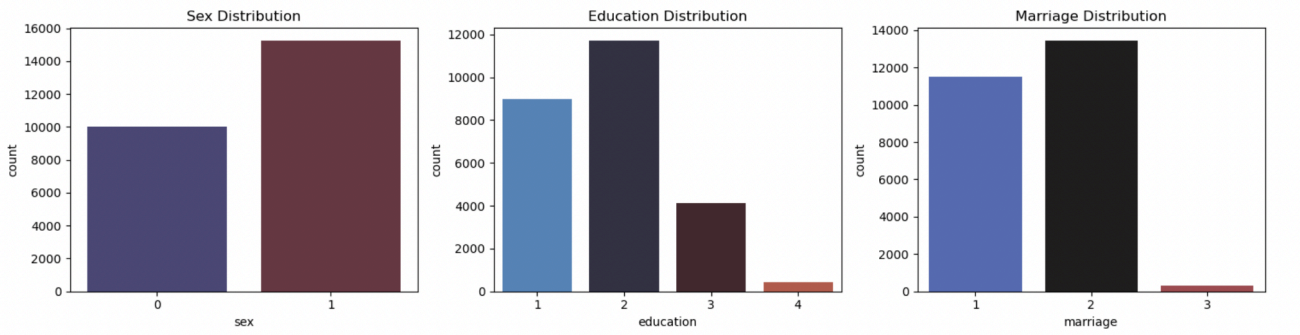
- Credit Limit (LIMIT_BAL):
  - The distribution is right-skewed, with most users having limits below ₹200,000.
  - Presence of outliers (e.g., limits above ₹500,000) suggests the potential need for log transformation or binning.



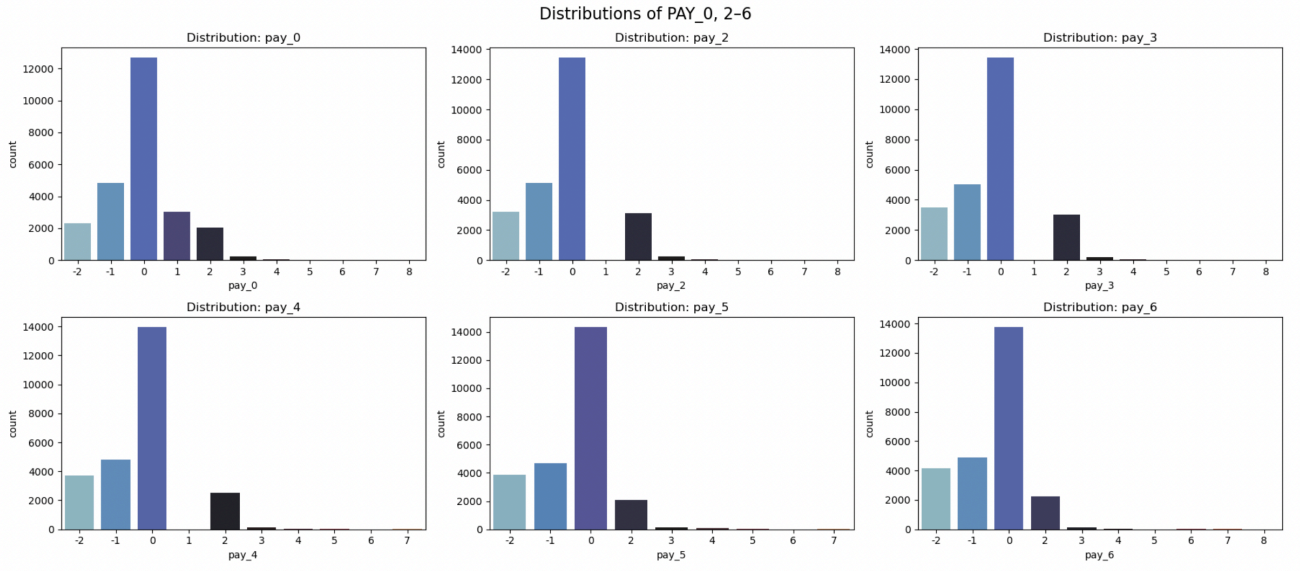
- Age:
  - Concentrated between 25 and 45 years, indicating a younger customer base.
  - Minor outliers exist (~70–80 years) but are within reasonable range.
  - Aged treated with median imputation and may be grouped into bins later.



- Categorical Variables:
  - Sex: Slightly more males than females.
  - Education: Majority are university graduates or postgraduates.
  - Marriage: Most customers are either married or single; category 3 (others) is sparse.



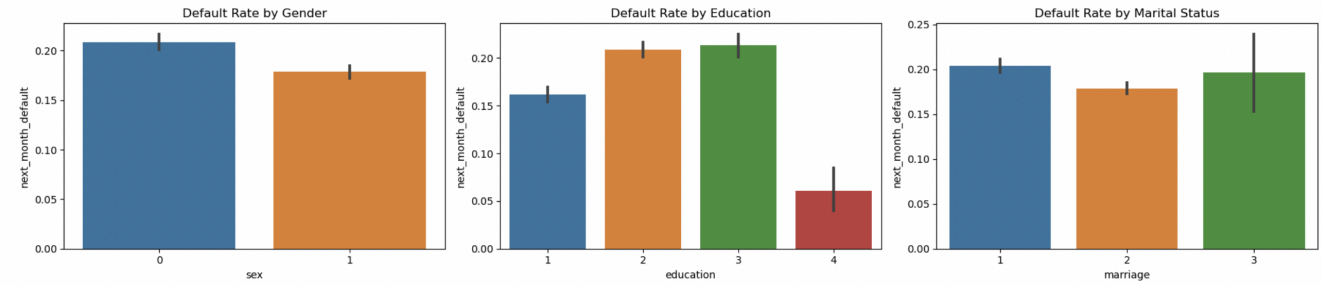
- Payment Status (pay_0 to pay_6):
  - Most values are 0, indicating **minimum or revolving payments** — a known risk pattern in credit.
  - Fewer entries are fully paid on time (-1), indicating potential repayment concerns.
  - Overdue values (1–3 months delay) are present but less frequent and consistent across months — a helpful signal for modeling delinquency trends.



2.2 Bivariate Analysis

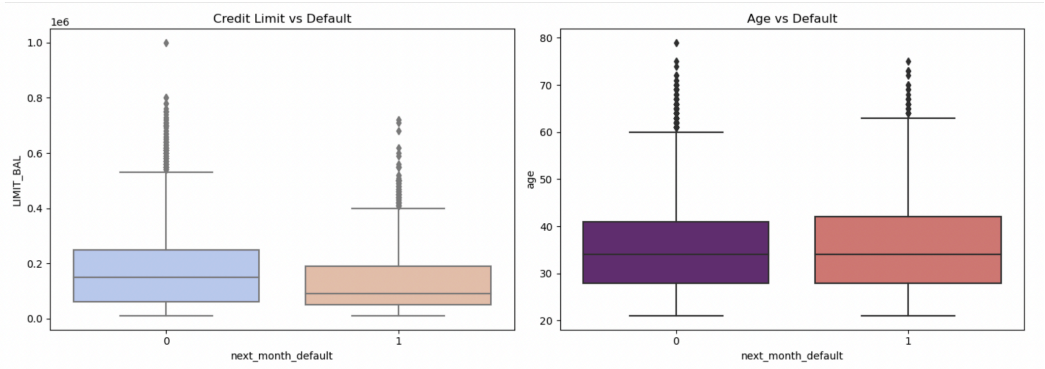
2.2.1 Demographic Trends

- Gender:
  - Female customers show a slightly higher default rate compared to males.
  - This may reflect differences in credit behavior or financial independence but remains marginal.
- Education:
  - Customers with high school education (3) show the highest default rate.
  - Lower education may be correlated with weaker financial literacy and poorer repayment discipline.
- Marital Status:
  - Default rate is similar for married and single customers.
  - Category 3 (Others) showed higher variance due to sparse data and was treated as "Other" during modeling.



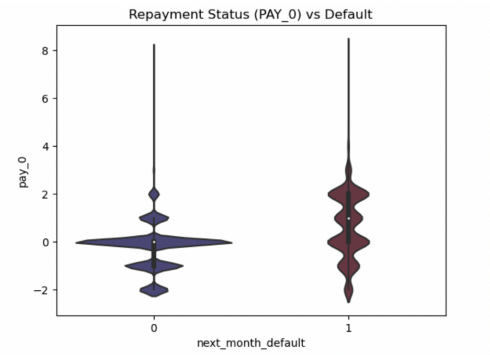
2.2.2 Financial Variable Distributions

- Credit Limit (LIMIT_BAL):
  - Right-skewed distribution; majority under ₹200,000. [*previously seen too]*
  - Defaulters tend to have a lower credit limit — median around ₹120,000 vs ₹160,000 for non-defaulters.
  - A predictive and financially interpretable feature.
- Age:
  - Most customers are in the 25–45 age range.
  - A mild trend suggests younger customers are more likely to default, possibly due to inexperience with credit.



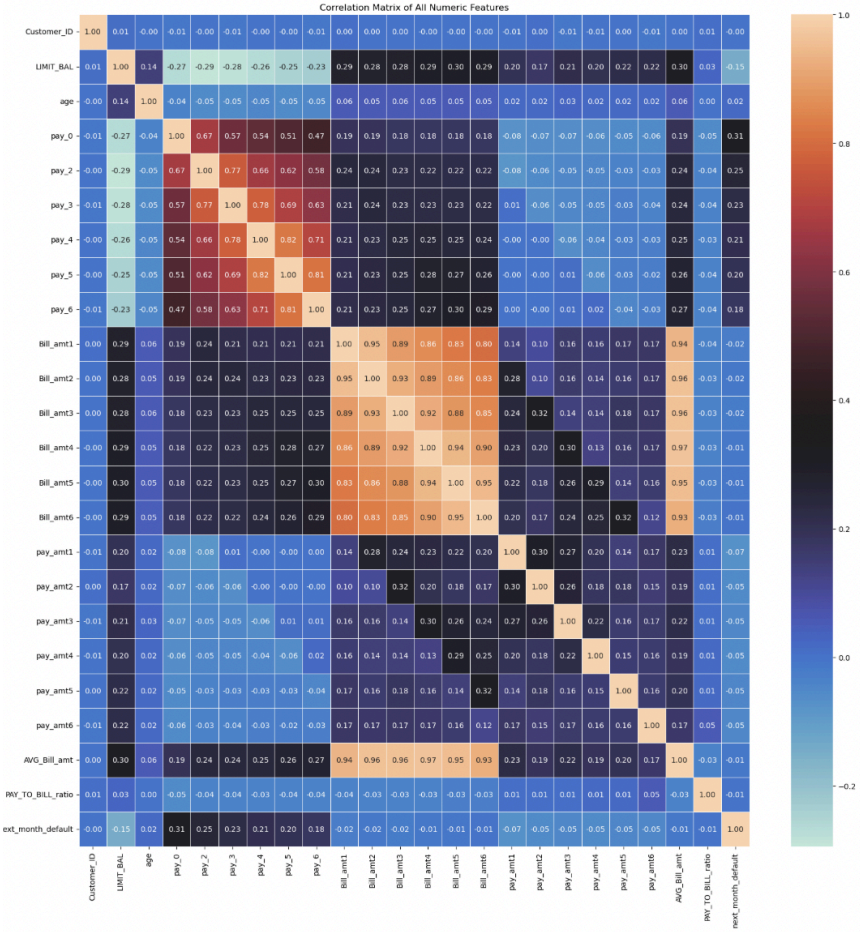
2.2.3 Repayment Behavior (PAY_0 to PAY_6)

- Values 0 (minimum payment made) dominate — a risky group relying on revolving credit.
- Fewer customers paid in full (-1), and a small but significant portion is overdue (1–3 months).
- Defaulting customers show consistently higher average PAY_x values, indicating chronic delinquency.
- **These features directly reflect repayment reliability and were highly predictive.**



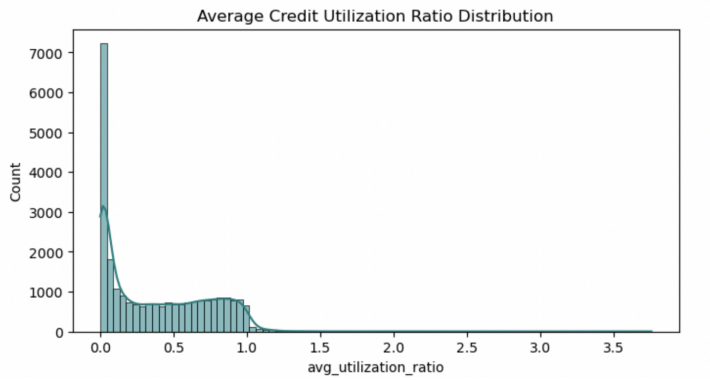
2.2.4 The Correlation Matrix

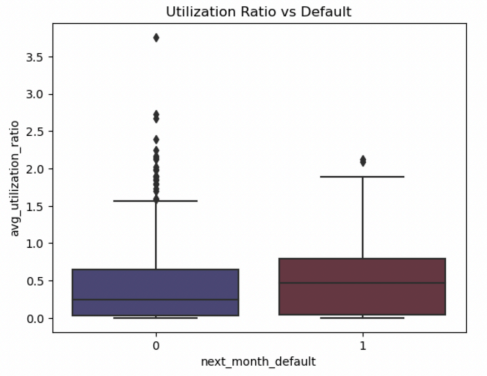
- pay_0 : Highest correlation (~0.31) with default — strongest individual predictor.
- Other pay_n : Gradual correlation decay, but all useful.
- LIMIT_BAL : Negatively correlated (~-0.15) — lower limits linked to higher default risk.
- Bill_amt* : Highly inter-correlated (0.85–0.98) — dimensionality reduction needed.
- PAY_TO_BILL_ratio : Low linear correlation (~-0.07) but possible non-linear behavior.



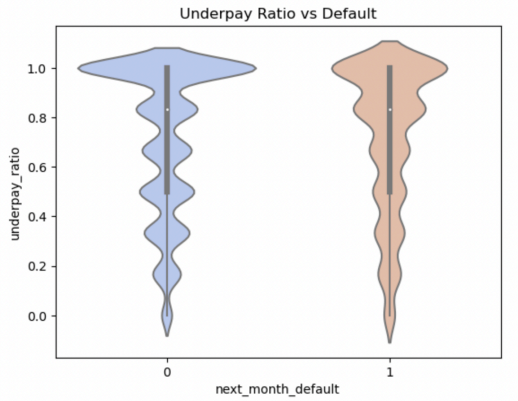
2.2.5 Financial Ratios and Behavioral Indicators

- **Credit Utilization Ratio**:
  - Calculated as average bill amount divided by credit limit.
  - Defaulters show significantly higher utilization — indicates stress or over-reliance on credit.
  - Right-skewed distribution — most users use <50% of credit.

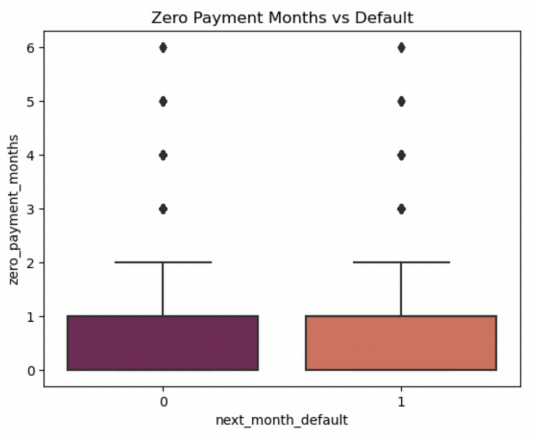




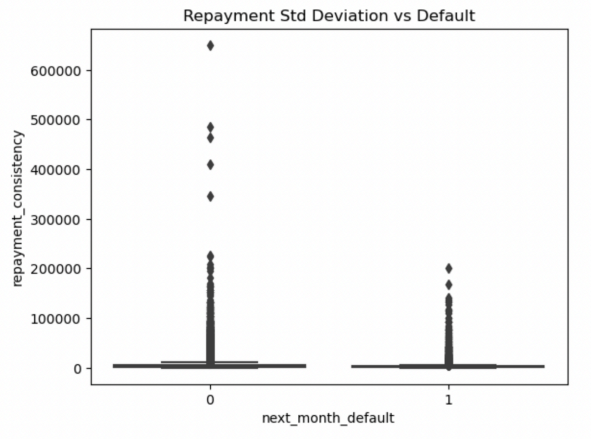
- **Underpayment Ratio**:
  - Fraction of months where payment < bill.
  - Defaulters consistently underpay more — strong link to poor repayment discipline.



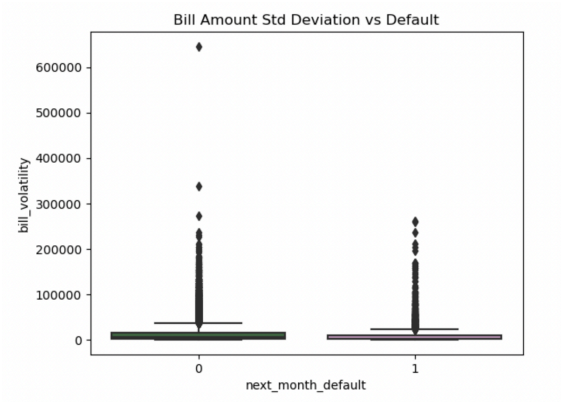
- **Zero Payment Months**:
  - Most customers have 0–2 such months.
  - Not highly discriminative alone but can serve as a complementary signal.



- **Repayment Consistency (Volatility in pay_amt1–pay_amt6):**
  - Defaulters show more irregularity in payment behavior (higher standard deviation).
  - Stability in repayment could be used as a trust signal.



- **Billing Volatility:**
  - Defaulters exhibit less billing volatility — possibly due to consistently high usage.
  - Implies over-reliance on available credit without fluctuation — a possible red flag.



3. Preparing the Dataset for Model Training

To enhance the model's predictive power and interpretability, extensive feature engineering and preprocessing were performed. This included handling outliers, crafting financially relevant features, and addressing data imbalance.

3.1 Outlier Handling and Transformation

Outliers in key numeric variables can distort model learning, especially in tree-based and linear models.

- Log Transformations were applied to skewed features like:
  LIMIT_BAL, repayment_consistency, and bill_volatility
- IQR-Based Capping was used to handle extreme values in:
  avg_utilization_ratio, underpay_ratio, and zero_payment_months

These transformations reduced skewness while preserving underlying distributional patterns.

3.2 Data Cleaning and Encoding

- Dropped: Non-informative identifier Customer_ID
- Categorical Encoding:
  - Rare categories were grouped into "Other" (e.g., education=4, marriage=3)
  - Categorical columns (sex, education, marriage) were cast to category dtype
  - Final model-ready format was created using One-Hot Encoding with drop_first=True to avoid multicollinearity

3.3 Financial Feature Engineering

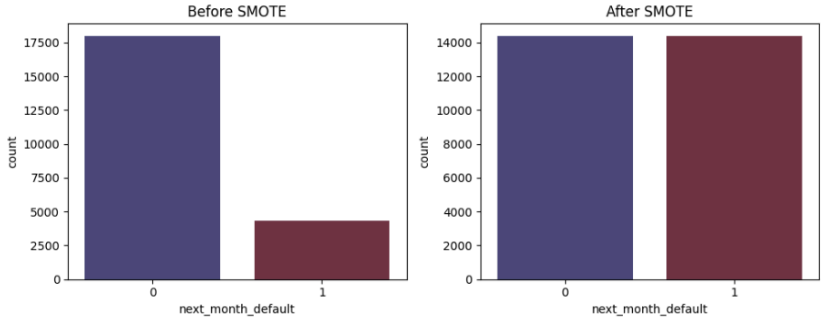
Seventeen domain-informed features were constructed to capture behavioral credit risk signals:

- avg_utilization_ratio : Avg. bill amount / credit limit — captures credit stress.
- underpay_ratio : Fraction of months with partial payments.
- zero_payment_months : Months where no payment was made.
- repayment_consistency : Std deviation of monthly payments.
- bill_volatility : Std deviation of bill amounts.
- max_delay : Maximum overdue value across PAY_0 to PAY_6.
- num_months_overdue : Total months with delay ≥ 1
- avg_delay_value : Mean delay among overdue months.
- delinquency_streak : Longest consecutive overdue months.
- avg_payment_ratio : Total payments / total bills over 6 months.
- total_overpaid_amount : Sum of overpayments (bill < payment).
- total_underpaid_amount : Sum of underpayments (bill > payment).
- monthly_payment_range : Max – Min payment across months.
- avg_payment_amount : Mean monthly payment amount.
- avg_bill_amount : Mean monthly bill amount.
- payment_std : Std dev of payment amounts.
- bill_std : Std dev of bill amounts.

These features not only improved model performance but also enhanced interpretability for risk managers.

3.4 Handling Class Imbalance

- First, **Highly correlated features (ρ > 0.90) were dropped** to prevent redundancy and overfitting.
- As the dataset is imbalanced (~20% defaulters), **SMOTE** (Synthetic Minority Oversampling Technique) **was applied to the training set** after imputing missing values (e.g., avg_payment_ratio was filled using median imputation).
- We've to make sure to **apply SMOTE after the Train-Test split** to make sure there is no **data leakage**.
- This **ensured a balanced distribution of both classes during training**, improving the model's ability to detect defaulters without sacrificing generalization.



4. Model Selection & Evaluation Journey

To build a reliable and business-aligned credit default prediction system, we adopted a structured modeling strategy — beginning with baseline models, progressing through performance tuning, and ultimately refining the system through ensemble learning and calibration. Throughout this process, our focus remained not only on maximizing technical metrics like the F2-score and AUC-ROC, but also on ensuring financial interpretability and real-world alignment with credit risk management practices. The journey was iterative, with each modeling stage evaluated against both statistical performance and the business consequences of misclassification — especially the high cost of overlooking defaulters. What follows is a comprehensive walk-through of this journey, highlighting how we arrived at the two most practical and deployable models.

4.1 Initial Modeling Phase — Baseline Trials

To establish baselines, we began with five core classification algorithms:

- Logistic Regression
- Decision Tree
- Random Forest
- XGBoost
- LightGBM
- Voting Ensemble

The following Metrics were obtained for each of the models :

```
Logistic Regression Metrics:
Accuracy:  0.39
Precision: 0.23
Recall:    0.90
F1 Score:  0.36
F2 Score:  0.56
AUC-ROC:   0.68
Confusion Matrix:
 [[1124 2964]
 [  98  864]]
```

```
Decision Tree Metrics:
Accuracy:  0.83
Precision: 0.55
Recall:    0.48
F1 Score:  0.51
F2 Score:  0.49
AUC-ROC:   0.76
Confusion Matrix:
 [[3704  384]
 [ 499  463]]
```

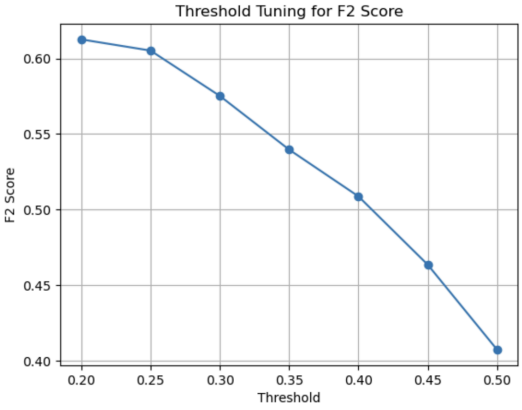
```
Random Forest Metrics:
Accuracy:  0.83
Precision: 0.58
Recall:    0.43
F1 Score:  0.49
F2 Score:  0.45
AUC-ROC:   0.78
Confusion Matrix:
 [[3794  294]
 [ 553  409]]
```

```
XGBoost Tuned Metrics:
Accuracy:  0.74
Precision: 0.39
Recall:    0.64
F1 Score:  0.49
F2 Score:  0.57
AUC-ROC:   0.77
Confusion Matrix:
 [[3137  951]
 [ 342  620]]
```

```
LightGBM Metrics:
Accuracy:  0.84
Precision: 0.60
Recall:    0.42
F1 Score:  0.49
F2 Score:  0.45
AUC-ROC:   0.78
Confusion Matrix:
 [[3815  273]
 [ 556  406]]
```

```
Voting Ensemble Metrics:
Accuracy:  0.82
Precision: 0.54
Recall:    0.50
F1 Score:  0.52
F2 Score:  0.51
AUC-ROC:   0.79
Confusion Matrix:
 [[3673  415]
 [ 479  483]]
```

Plotting, the graph between F2 score and Threshold of the Ensemble (which was giving the best trade-off between accuracy and F2 score among them), we found:



- Lowering threshold improves F2-score, which gives higher weight to recall (key in financial default scenarios).
- However, very low thresholds (e.g., 0.2) led to unacceptable trade-offs in accuracy.

***Financial rationale:** In credit risk, missing a defaulter is costlier than flagging a few safe users, hence F2-score optimization was prioritized.*

4.2 Hyperparameter Tuning

Proceeding to get the best Accuracy and F2-score trade off, we tuned the models by running **RandomizedSearchCV** over Random Forest, XGBoost, LightGBM and **GridSearchCV** over Decision Tree, Logistic Regression.

This yielded the following Metrics:

```
Random Forest Tuned Metrics:
Accuracy:  0.72
Precision: 0.37
Recall:    0.70
F1 Score:  0.48
F2 Score:  0.59
AUC-ROC:   0.79
Confusion Matrix:
 [[2939 1149]
 [ 289  673]]
```

```
XGBoost Tuned Metrics:
Accuracy:  0.19
Precision: 0.19
Recall:    1.00
F1 Score:  0.32
F2 Score:  0.54
AUC-ROC:   0.77
Confusion Matrix:
 [[   0 4088]
 [   0  962]]
```

```
LightGBM Tuned Metrics:
Accuracy:  0.68
Precision: 0.34
Recall:    0.74
F1 Score:  0.47
F2 Score:  0.60
AUC-ROC:   0.77
Confusion Matrix:
 [[2719 1369]
 [ 246  716]]
```

```
Decision Tree Tuned Metrics:
Accuracy:  0.78
Precision: 0.44
Recall:    0.59
F1 Score:  0.50
F2 Score:  0.55
AUC-ROC:   0.76
Confusion Matrix:
 [[3368  720]
 [ 394  568]]
```

```
Logistic Regression Tuned Metrics:
Accuracy:  0.28
Precision: 0.21
Recall:    0.98
F1 Score:  0.34
F2 Score:  0.56
AUC-ROC:   0.72
Confusion Matrix:
 [[ 476 3612]
 [  20  942]]
```

Final performances:

- Best LightGBM: F2-score = 0.60, AUC = 0.77
- Best Random Forest: F2-score = 0.59, AUC = 0.79

XGBoost and Logistic Regression, while decent, were outperformed by the two above in both stability and recall.

4.3 Final Refinement - Feature Selection & Borderline SMOTE

To further improve signal clarity, we:

- Applied **Borderline-SMOTE** to oversample near-decision-boundary examples.
- **Selected top 25 features** using mutual information.

Both LightGBM and Random Forest were retrained with this refined input. Recognizing that LightGBM's probability predictions were not well-aligned with true likelihood, we calibrated its probabilities using isotonic regression.

The following new Metrics were obtained:

```
Refined Random Forest (BorderlineSMOTE + KBest) Tuned Metrics:
Accuracy:  0.78
Precision: 0.45
Recall:    0.61
F1 Score:  0.52
F2 Score:  0.57
AUC-ROC:   0.78
Confusion Matrix:
 [[3370  718]
 [ 379  583]]
```

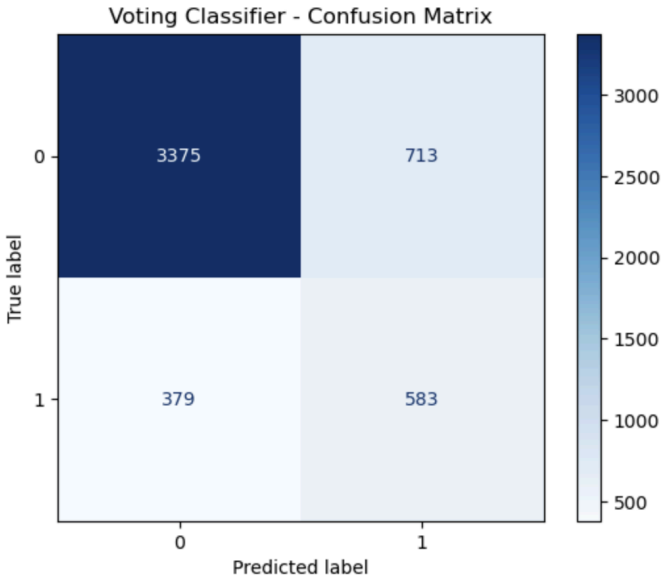
```
Calibrated LightGBM Tuned Metrics:
Accuracy:  0.84
Precision: 0.61
Recall:    0.38
F1 Score:  0.47
F2 Score:  0.41
AUC-ROC:   0.74
Confusion Matrix:
 [[3850  238]
 [ 594  368]]
```

4.4 Final Ensembles

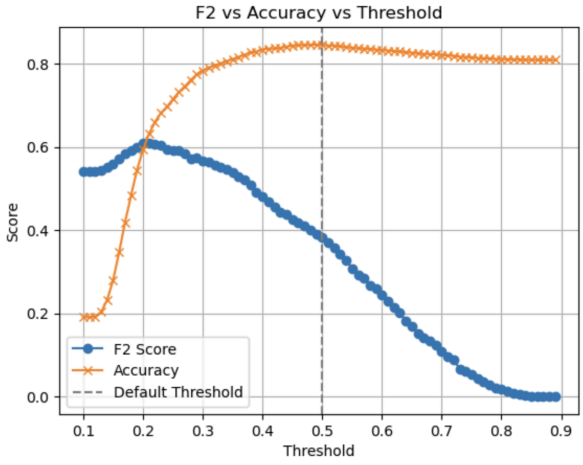
Upon combining the refined Random Forest with the Calibrated LightGBM using a VotingClassifier.

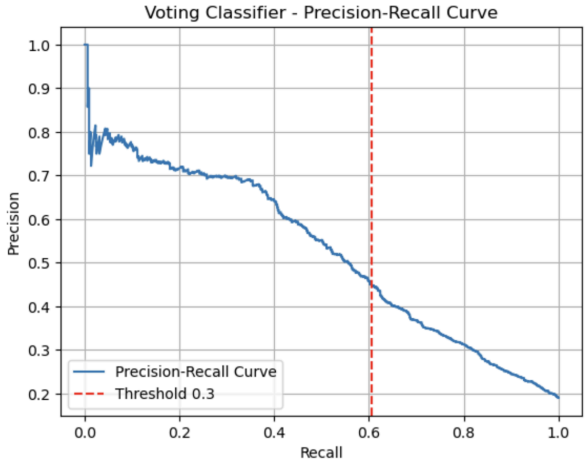
It resulted into a model with the following metrics:

```
Voting Classifier Metrics:
Accuracy     : 0.78
Precision    : 0.45
Recall       : 0.61
F1 Score     : 0.52
F2 Score     : 0.57
AUC-ROC      : 0.78
Confusion Matrix:
[[3375  713]
 [ 379  583]]
```



Upon observing its F2-Score vs Accuracy vs Threshold & Precision-Recall Curves:

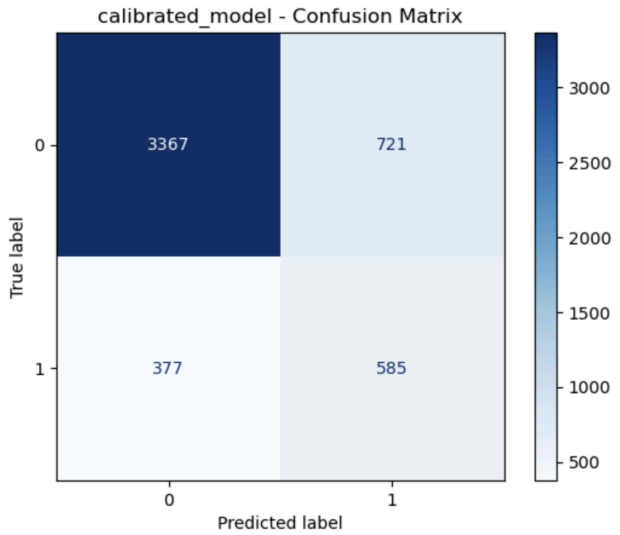




this **voting_model** has by far been our Best Model giving the best trade offs between the metrics with a considerably better **Threshold = 0.3** for testing:
F2-score = 0.57, Recall = 0.61, AUC = 0.78, **Accuracy = 0.78**

we further calibrated our voting_model using CalibratedClassifierCV, and the following similar metrics were obtained, but at* a **Threshold = 0.15 for testing*:*

```
calibrated_model Metrics:
Accuracy     : 0.78
Precision    : 0.45
Recall       : 0.61
F1 Score     : 0.52
F2 Score     : 0.57
AUC-ROC      : 0.78
Confusion Matrix:
[[3367  721]
 [ 377  585]]
```



calibrated_model *is just the calibrated version of voting_model** and its nature is justified as calibration applies Platt scaling (method='sigmoid') to adjust the probability outputs and ensures the model's confidence in its predictions aligns with the real-world outcomes.*

5. Validation Set Testing — Business Fit

We tested both final models on the validation dataset and studied the distribution of predicted probabilities.

- **voting_mode**l : It predicted 28.77% [threshold taken 0.30] as defaulters, which was greater than the actual (19.04%) real world scenario according to the training data.
- **calibrated_model** : It predicted 19.84% [threshold taken 0.25] as defaulters, which was closer to the actual (19.04%) real world scenario according to the training data.

If we see at different Thresholds:

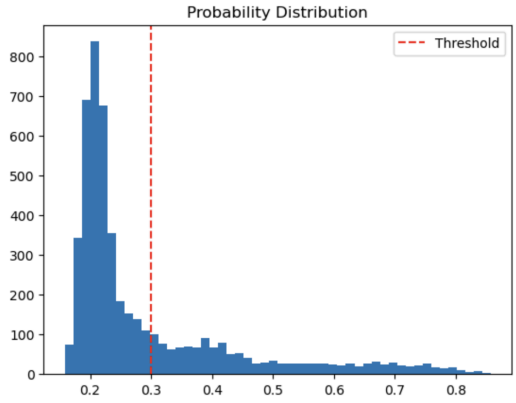
```
Threshold 0.25 → 38.64% predicted defaulters
Threshold 0.30 → 28.77% predicted defaulters
Threshold 0.35 → 23.27% predicted defaulters
Threshold 0.40 → 17.96% predicted defaulters
```

**voting_model*

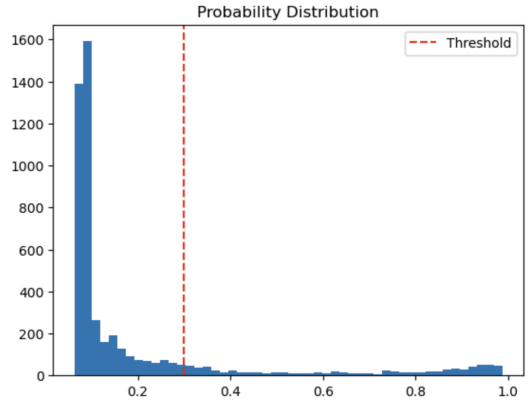
```
Threshold 0.25 → 19.84% predicted defaulters
Threshold 0.30 → 16.61% predicted defaulters
Threshold 0.35 → 14.27% predicted defaulters
Threshold 0.40 → 12.80% predicted defaulters
```

**calibrated_model*

*on seeing the probability distribution plots of both models:



*voting_model



*calibrated_model

- *We can see that the Probability is distribution of voting_model is spread-out as compared to its calibrated version calibrated_model.*
- *In the probability distribution plot, the voting_model's distribution has a bump-up while going towards right, which is smoothened out by calibration and is therefore reflected in the plot of calibrated_model. This probability distribution plot is the key differentiator between the two models.*
- *These are all results of Platt scaling onto the voting_model.*

6. Final Model Recommendation & Business Risk Alignment

6.1 Why F2-score?

When selecting the most appropriate evaluation metric for our credit default prediction task, we explored a variety of options—Accuracy, Precision, Recall, F1-score, F2-score, and AUC-ROC. But not all metrics serve the same business interests. After examining the real-world costs of misclassification, we centered our strategy around the F2-score, which gave greater priority to detecting actual defaulters over general correctness.

- Accuracy, while often the first go-to metric, proved deceptive in our case. Due to class imbalance (most customers don't default), a model could appear highly accurate just by predicting everyone as safe. But such a model would fail where it matters most—spotting real defaulters, and preventing potential financial loss.

- Precision tells us how trustworthy our positive predictions are, i.e., how many flagged defaulters actually default. However, precision alone ignores the ones we missed entirely—false negatives—which are far more dangerous in credit risk. Lending to someone who defaults can be much costlier than mistakenly flagging a reliable customer.

- On the flip side, Recall ensures we catch as many true defaulters as possible. It aligns more closely with our risk mitigation goals, but recall alone doesn't protect against excessive false positives, which could harm customer experience and reduce lending efficiency.

- F1-score, which tries to strike a balance between precision and recall, is a solid compromise. But in our scenario, where missing defaulters is disproportionately harmful, we needed a metric that leaned harder toward recall.

- This is where the F2-score came in. By giving recall more weight than precision, it helped us optimize for what mattered most—reducing the chance of undetected defaults, even if it meant slightly more conservative predictions.

- We used AUC-ROC as a secondary metric to ensure the model still had good general classification capability, but it did not influence threshold decisions or model selection directly.

In summary, F2-score wasn't just a technical preference—it reflected a business-aware modeling choice, guiding all our tuning, threshold calibration, and ensemble decisions with the end goal of minimizing real financial exposure.

**Business Implications* *of False Positives vs False Negatives:*

Error Type	Impact on Bank	Tolerance	Explanation
False Positive	Predicts default → denies credit to a good customer	Medium	Loss of potential revenue or customer dissatisfaction
False Negative	Predicts non-default → lends to a defaulter	Low	Leads to actual financial loss through missed repayments; must be minimized

If the bank is more tolerant of false positives than false negatives, meaning:* ***F2-score (which emphasizes recall)* *become the most critical evaluation metric. A model could've been more accurate by predicting the majority class (non-defaulters) of the imbalanced dataset, but that will lead to missing defaulters, hence the emphasis on F2-score.*

6.2 The Proposal

The final selections of Models is based on their behaviours of yielding results:
- Voting Model errs on the side of caution: prefers overflagging defaulters.
- Calibrated Model aligns with true defaulter rate: more realistic, usable for firms that can afford some missed defaulters.

Thus, this Analysis Yields **TWO USABLE MODELS** depending on the use case. Both models are valid for real-world use, depending on firm's **risk appetite**:

- Use **voting_model** if the business priority is **avoiding false negatives** (e.g., high-risk credit portfolios).
- Use **calibrated_model** if the firm values probabilistic realism and **balanced decisioning** (e.g., mass consumer lending, better customer experience).

Depending on the nature of use-case, any one of the two models can be selected and used confidently.