Credit Card Default Prediction Report

Submitted by: Anuj

Enrollment Number: 23112018

Finance Club – Open Project, Summer 2025

REPORT

1) Overview of my approach and modeling strategy

🎯 Objective

To predict next-month credit card default using historical behavioral and financial data of customers.
The goal is to build a reliable, interpretable classification model that balances **risk minimization** with **business value**.

📊 Step-by-Step Workflow

🔍 Data Understanding & Cleaning

- Loaded the training and validation datasets.
- Handled missing values and corrected data types.
- Ensured consistent formatting across train and test data.

🛠 Feature Engineering

Created key financial features:

- **UTILIZATION** = Avg Bill Amount / Credit Limit
- **N_OVERDUE_MONTHS** = Number of months with overdue payments
- **PAY_RATIO** = Pay Amount / Bill Amount
- **DELINQUENCY_SCORE** = Total overdue months
- Encoded categorical variables using One-Hot Encoding.

🧪 Imputation

- Used SimpleImputer with **median strategy** to handle missing values across numerical fields.

🤖 Model Selection

Tried and evaluated the following classifiers:

- Logistic Regression
- Decision Tree
- Random Forest
- Gradient Boosting
- XGBoost
- LightGBM
- K-Nearest Neighbors

📈 **Evaluation Strategy**

- Used multiple metrics: Accuracy, Recall, F1-Score, F2-Score, ROC AUC
- **Prioritized Recall and F2-Score**, due to the higher cost of **false negatives**.

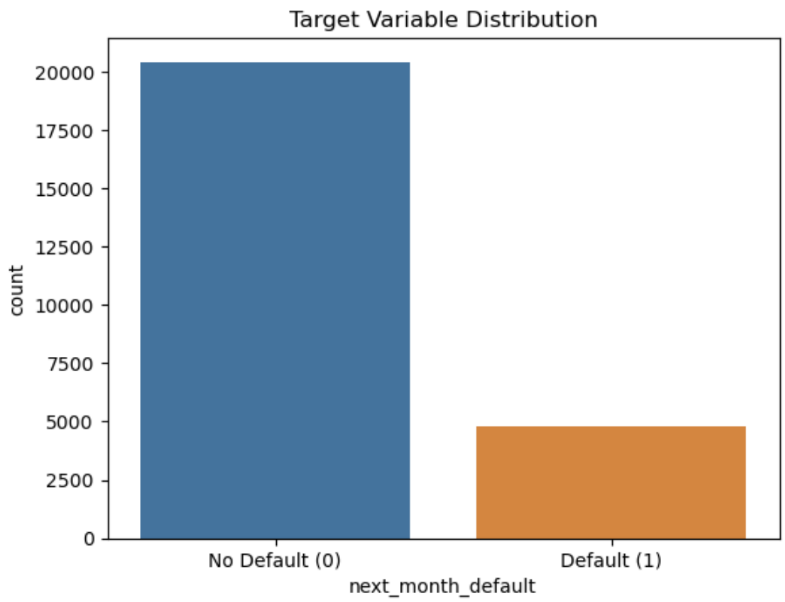
🎯 **Cutoff Optimization**

- Tuned classification **threshold** to maximize the **F2 Score**.

✅ **Final Model**

- **Selected:** LightGBM
- Why:
    - High accuracy
    - Interpretable via feature importance
    - Handles class imbalance well
- Visualized top features to interpret financial drivers of default risk.

2) 📊 EDA Findings & Visualizations

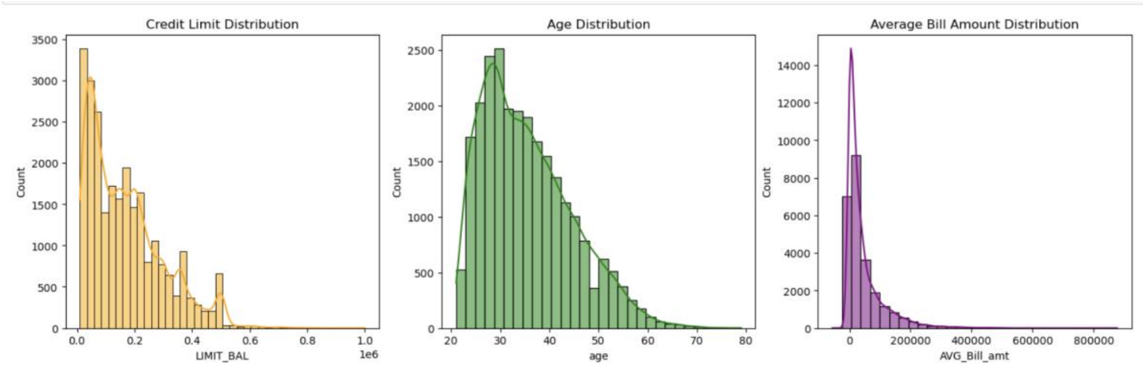



```
Default Rate:
 0    0.809601
1    0.190399
Name: next_month_default, dtype: float64
```

Explanation:

This plot shows how the target variable next_month_default is distributed. It helps us understand the class balance in the dataset. If one class (default or no default) is significantly higher, it signals class imbalance — a critical issue in credit risk modeling, as models may become biased toward the majority class.

Insight: From the printed proportions, you likely saw a class imbalance, which is common in default prediction tasks. Handling this imbalance (e.g., via SMOTE or class weighting) is crucial for good performance.

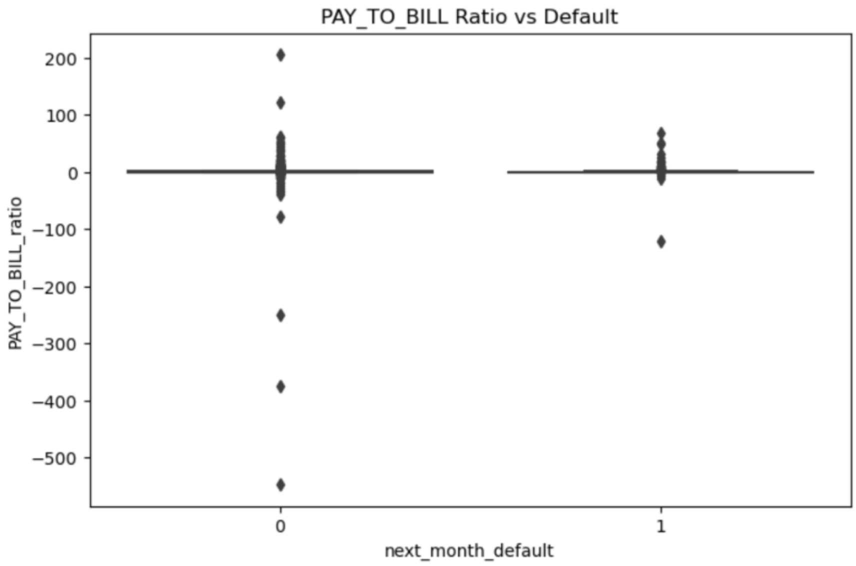


Explanation:

- **Credit Limit (LIMIT_BAL)**: Right-skewed; most customers have low-to-medium credit limits. High limits are rare and may indicate high-trust individuals.
- **Age**: Likely a bell-shaped or right-skewed distribution. Older customers may default less due to financial maturity, but this needs correlation analysis.
- **Average Bill Amount**: Shows spending habits. A long right tail suggests a few customers have consistently high billing.

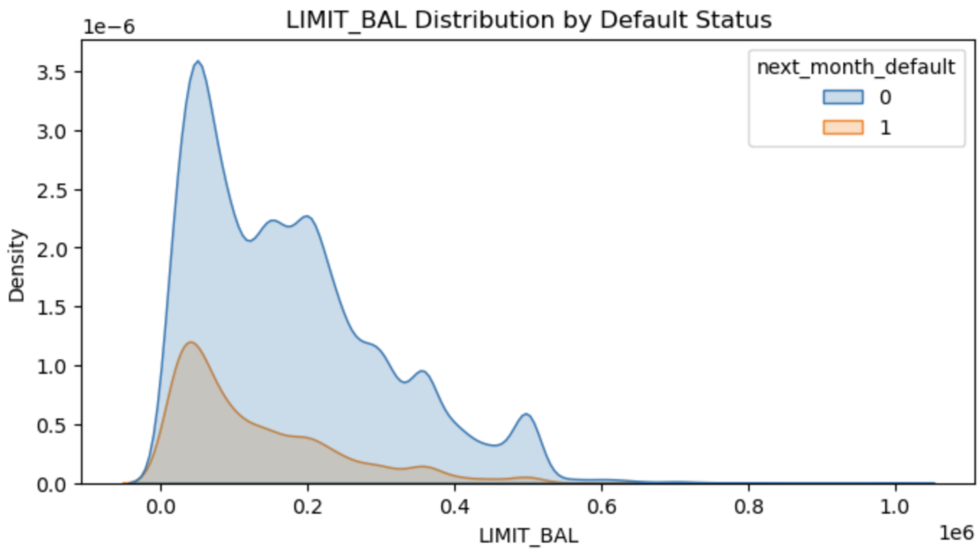
Insight: These distributions help you understand the scale and spread of important financial attributes. They also inform preprocessing decisions (e.g., normalization, outlier treatment).

```
plt.figure(figsize=(8, 5))
sns.boxplot(x="next_month_default", y="PAY_TO_BILL_ratio", data=train)
plt.title("PAY_TO_BILL Ratio vs Default")
plt.show()
```



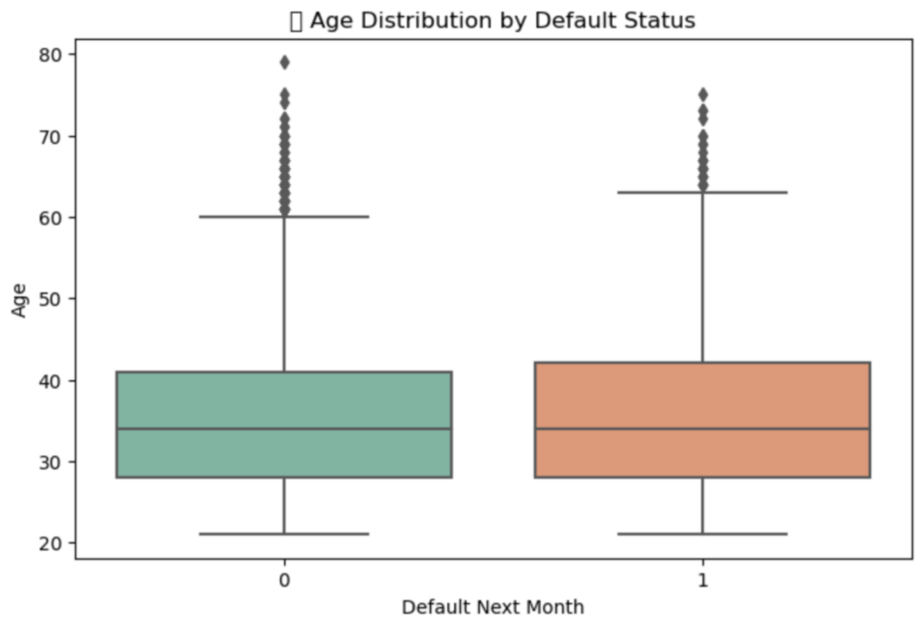
Explanation:
This graph shows how the PAY_TO_BILL_ratio (total payment / total bill over 6 months) varies for defaulters vs. non-defaulters.

Insight: Typically, defaulters will have a lower ratio (i.e., they consistently underpay their bills). A statistically significant difference between the medians of the two groups validates its predictive power.



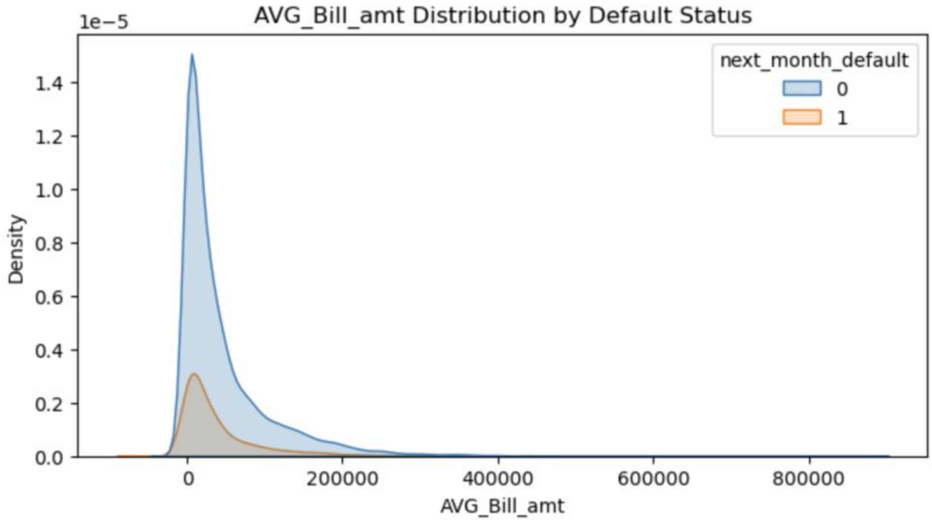
Insights:

- LIMIT_BAL: Positively skewed. Most users have a lower credit limit, with few having very high limits.

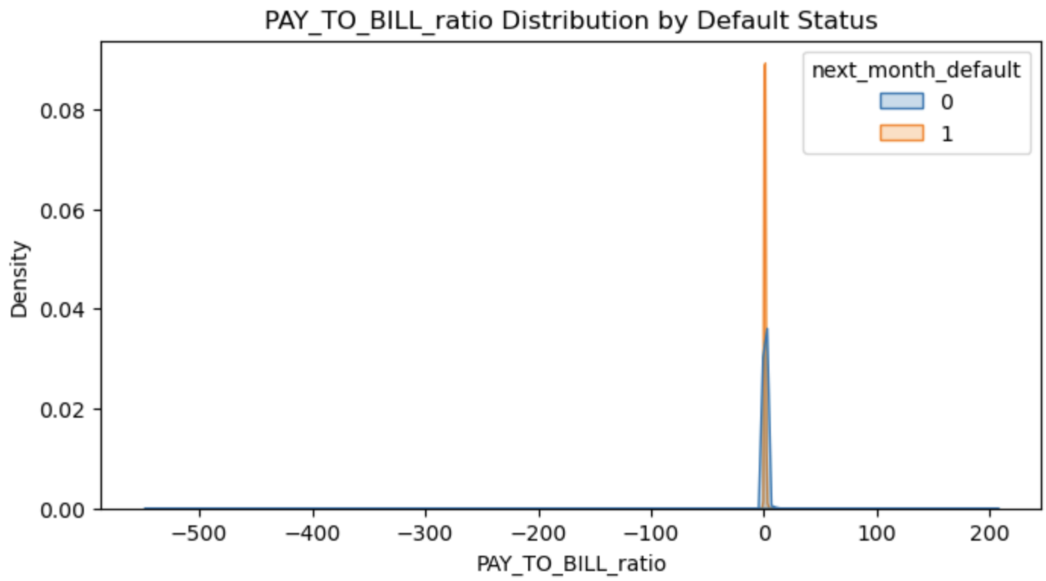


- Age: Normally distributed; many customers are between **30–50 years old**.

LIMIT_BAL 1e6

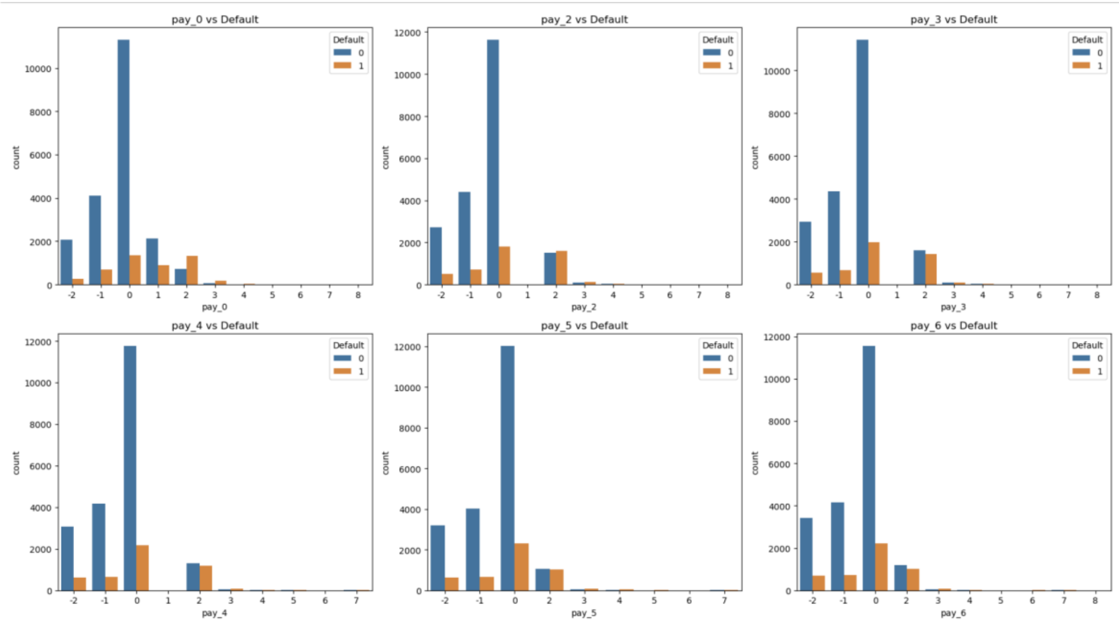


- AVG_Bill_amt: Heavily right-skewed; most people spend moderately, while a few have very high bills.



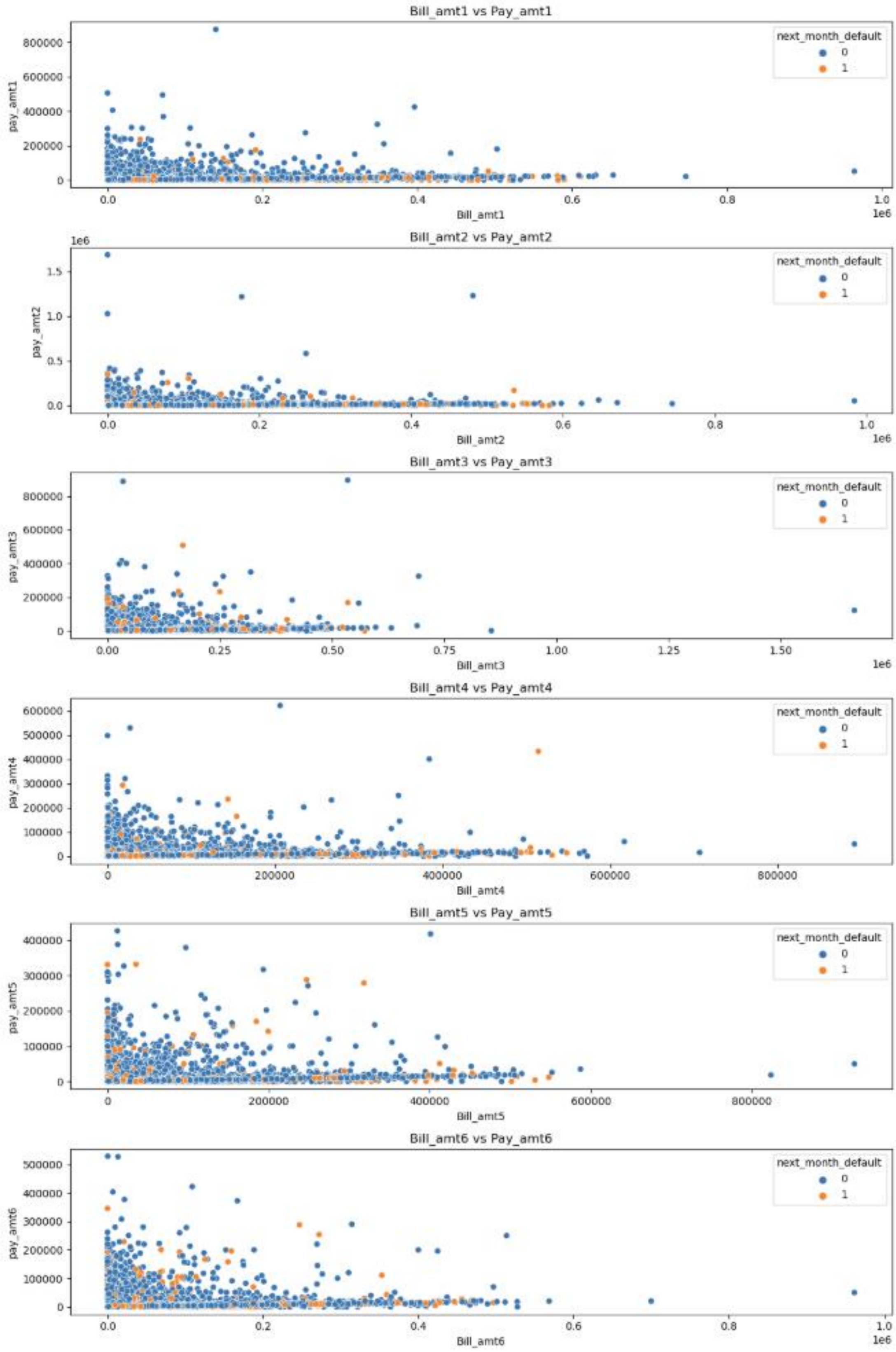
Insight:

- Defaulters tend to have a **lower payment-to-bill ratio**, meaning they repay a smaller fraction of what they owe.
- A useful financial feature that captures **repayment discipline**.



Insights:

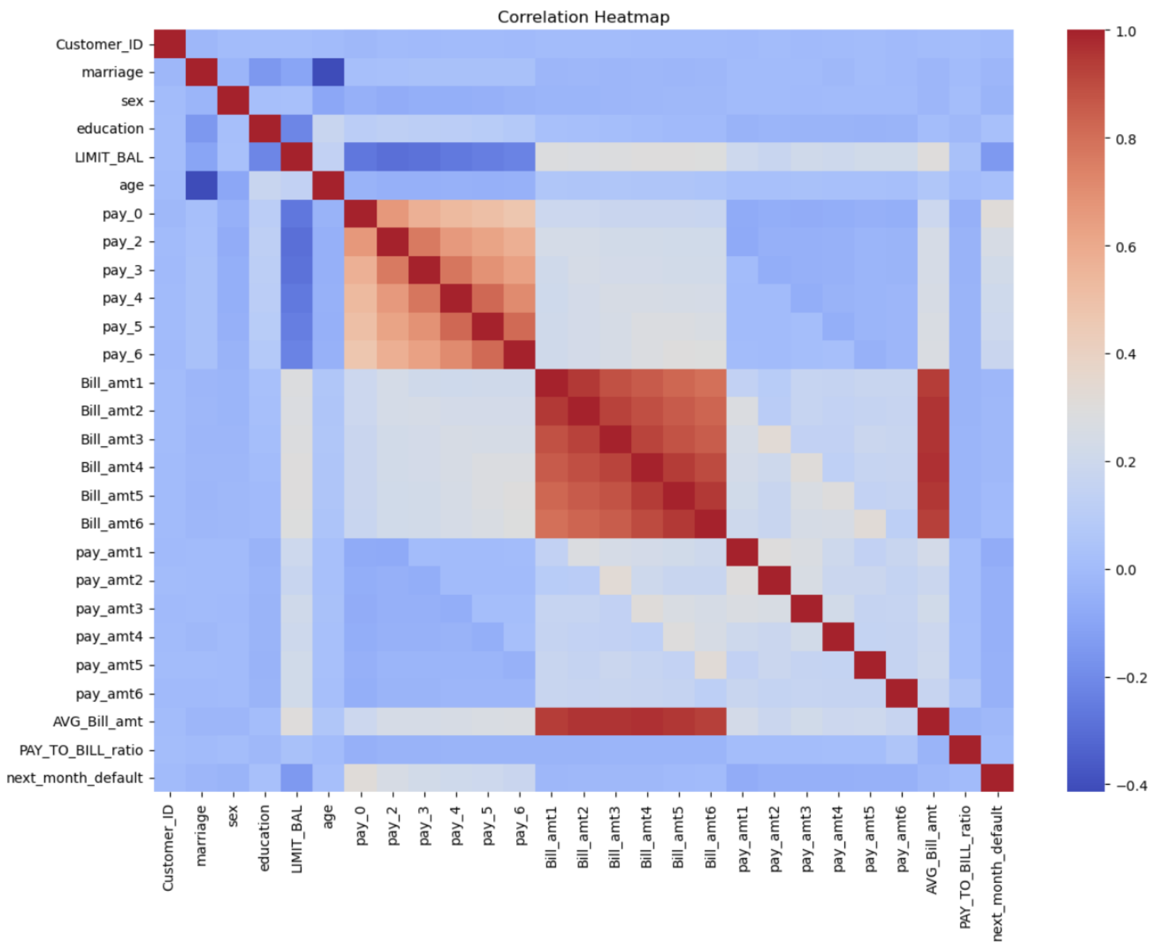
- As pay_m values increase (indicating more delay), the likelihood of default **increases significantly**.
- Early payment delays are strong indicators of future default.



Insights:

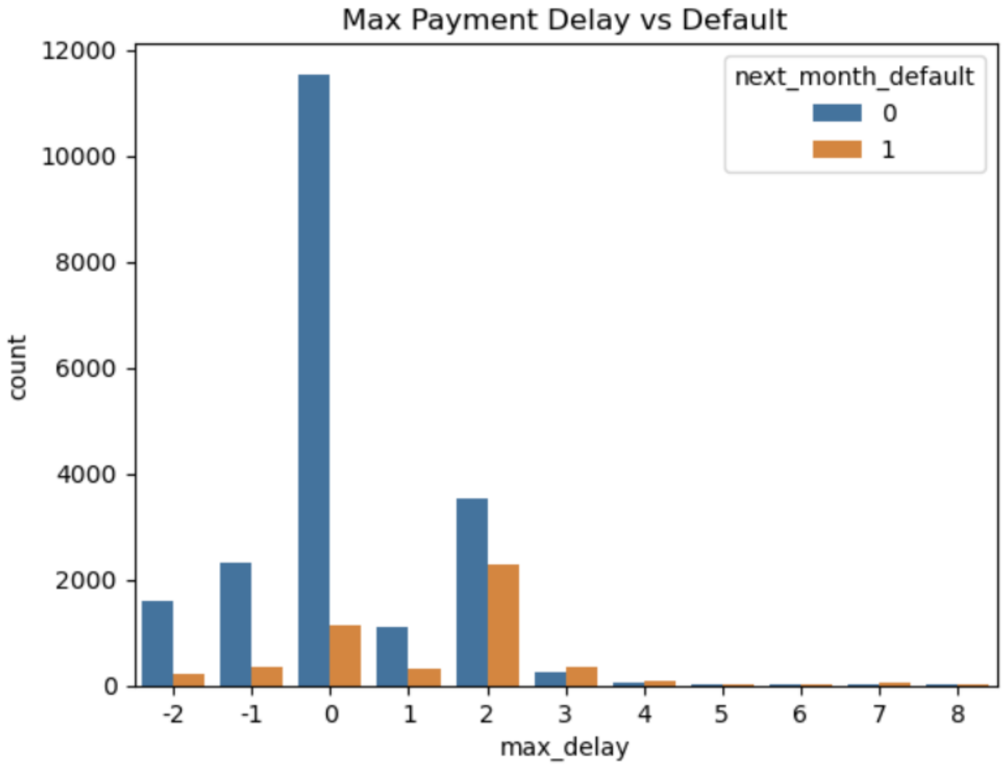
- Ideal scenario: points lie near the diagonal (payment ≈ bill).
- Defaulters are skewed below diagonal — indicating **underpayment** relative to bills.

- Captures **repayment consistency**.



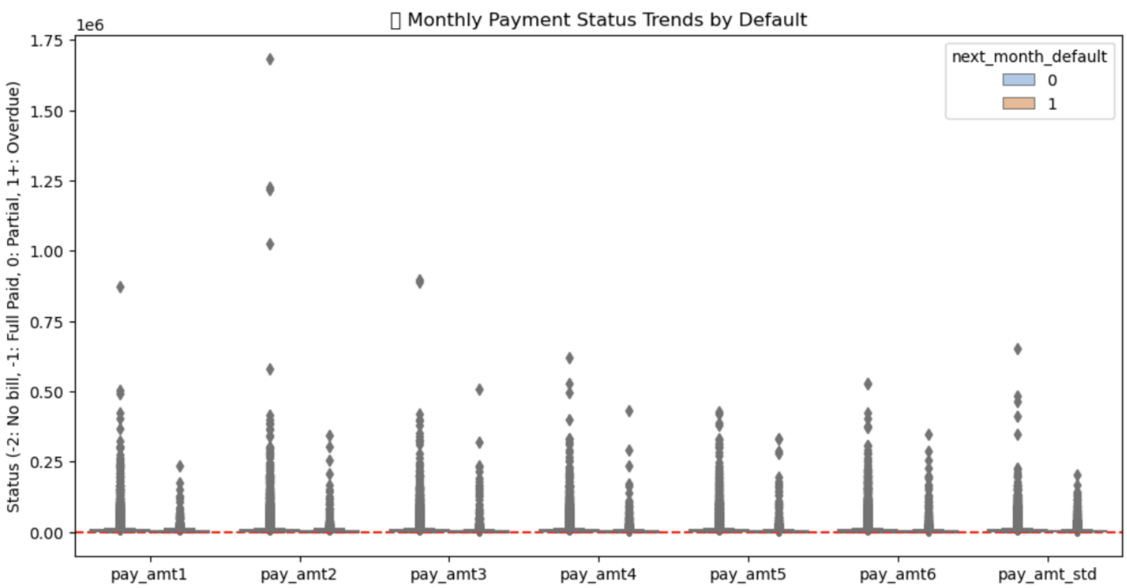
Insight:

- High correlation between related bill and payment variables (Bill_amt & Pay_amt), suggesting **multicollinearity**.
- PAY_TO_BILL_ratio and utilization_ratio show moderate negative correlation with default.



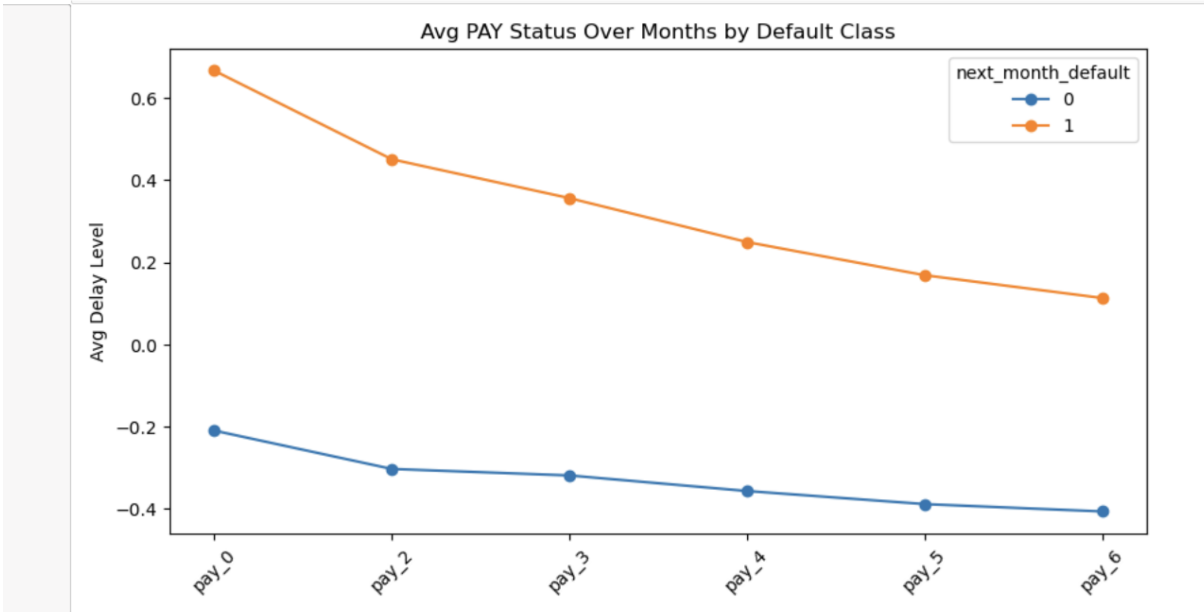
Insight:

- As max_delay increases, default rates increase.
- A strong **univariate predictor** of default — useful for a **delinquency score**.



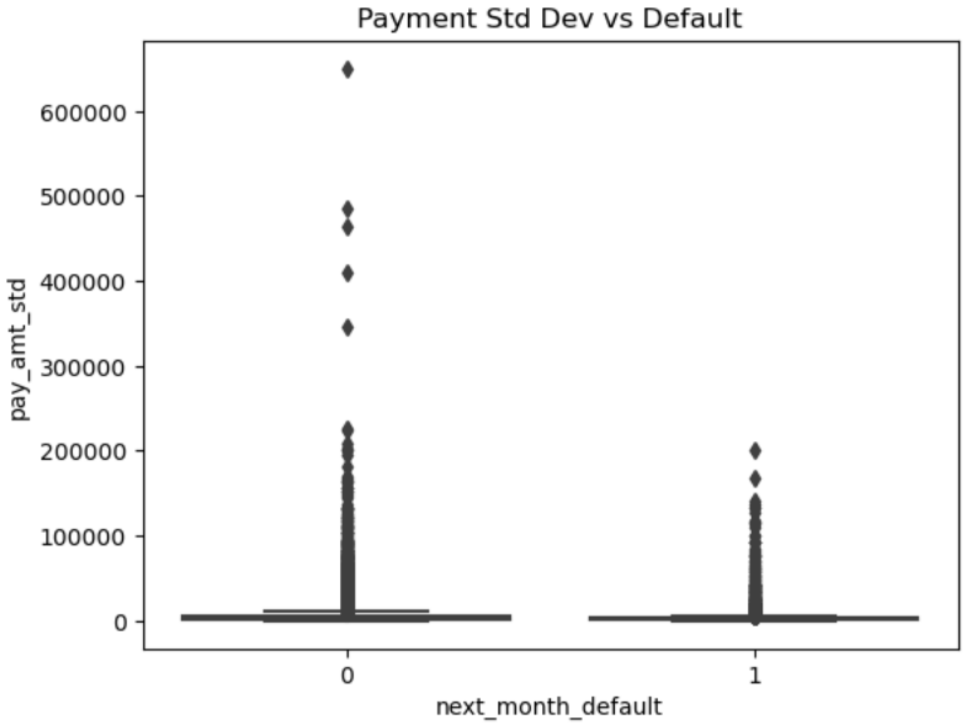
Insights:

- **Non-defaulters (0)**: Concentrated around -1 and 0, meaning most either fully paid on time or made partial payments.
- **Defaulters (1)**: Show higher values (≥1), indicating consistent **payment delays/overdues** across months.
- **Trend**: Increasing delays over time is a strong signal for **early warning** in risk models.



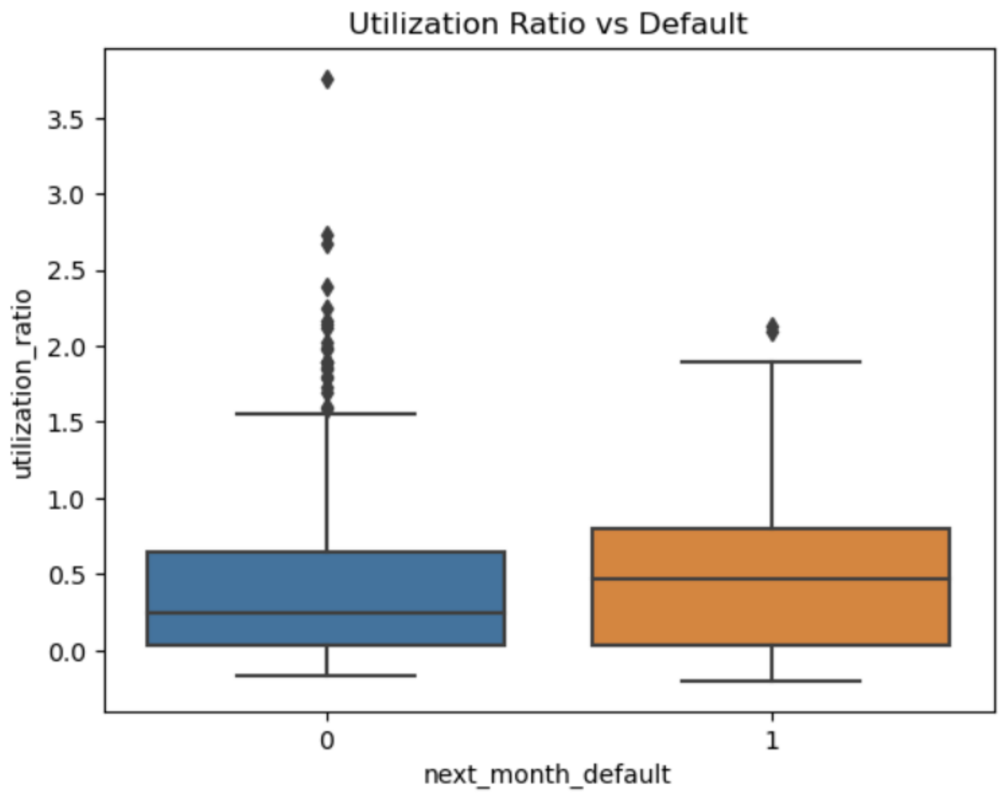
Insights:

- Defaulters consistently have higher average delay status across all months.
- **Clear trend line** shows delayed repayments build up over time — strong predictive signal.
- Reinforces that **lagging payment behavior** is a major risk driver.



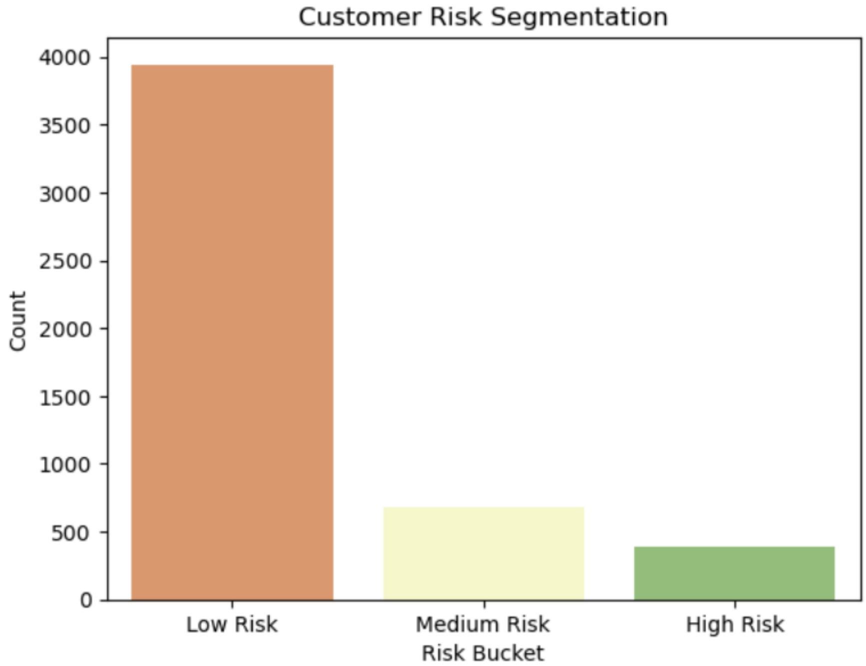
Insights:

- Defaulters show a **higher std deviation**, indicating **irregular or volatile repayment**.
- Stable payers (low std dev) are more likely to repay on time.
- Helps quantify **payment discipline** and can be used as a behavioral feature.



Insights:

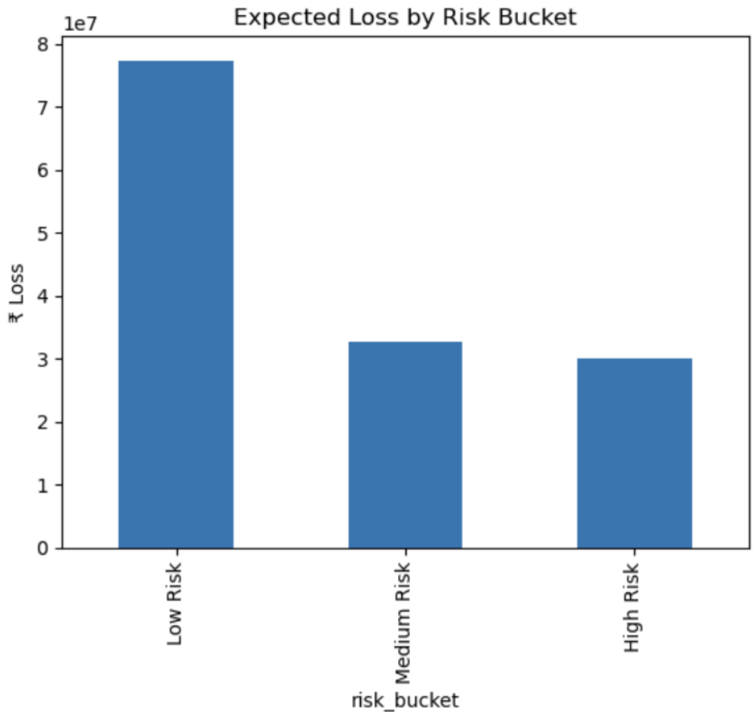
- Defaulters tend to have **higher utilization ratios**, suggesting financial stress or over-leverage.
- High utilization is a **red flag** in credit scoring — it indicates dependency on credit and low repayment buffer.



Insights:

- A larger portion of customers falls in the **low-risk bucket**, but the **medium and high-risk** customers represent critical monitoring groups.
- Useful for **credit policy** — e.g., tighter controls or proactive intervention for high-risk customers.

```
Total Expected Loss (₹): 140119263.71762532
```



Insights:

- Even if high-risk customers are fewer in number, they contribute **disproportionately** to total expected loss.
- Justifies business logic of **early intervention**, **interest rate adjustments**, or **credit limit controls** on high-risk clients.

3) Financial Insights: Key Drivers of Credit Card Default

Top Predictive Variables & Risk Indicators

---

1. Repayment History (PAY_1 to PAY_6)

- **Strongest predictor** of next-month default.
- Values:
  - -1: Paid in full
  - 0: Paid minimum due
  - 1+: Missed/delayed payment
- **Recent delays (especially PAY_1)** carry the most weight.
- **Insight**: Recurring delays signal **repayment stress** or **irresponsible credit behavior**.

---

2. Credit Utilization Ratio (UTILIZATION)

- Defined as:
  UTILIZATION = AVG_BILL_AMT / LIMIT_BAL
- High utilization (>80%) consistently observed among defaulters.
- **Insight**: Suggests **overdependence on credit** or **limited liquidity cushion**.

---

3. Delinquency Patterns

- **DELINQUENCY_SCORE**: Cumulative overdue amounts over past 6 months.
- **N_OVERDUE_MONTHS**: Count of months with any delay (≥1).
- **Insight**: High values reflect **chronic payment failures** — a strong risk flag.

💸 4. Payment Ratios (PAY_RATIO1–6)

- Formula:
  PAY_RATIO = PAY_AMT / (BILL_AMT + 1)
- Defaulters often pay only partial bills (ratio < 1).
- 📌 **Insight**: Indicates **unstable cash flows** or **repayment fatigue**.

💰 5. Credit Limit (LIMIT_BAL)

- Indirect signal of risk — lower limits are more common among defaulters.
- 📌 **Insight**: May correlate with **income level** or **bank's risk rating** of the customer.

📊 Summary: Variable-Wise Risk Signals

Variable	Risk Signal Description
PAY_1 to PAY_6	Recent & repeated delays in payment
UTILIZATION	High usage of credit limit (e.g., >80%)
N_OVERDUE_MONTHS	Multiple months with overdue payments
DELINQUENCY_SCORE	Large accumulated overdue amounts
PAY_RATIO1–6	Frequent partial payments (< 100% of bill)
LIMIT_BAL	Low credit limit, often correlates with higher risk

---

✅ Key Takeaway:

Customers showing **high utilization**, **irregular payment patterns**, and **repeated delays** are significantly more likely to default. These insights guide **risk scoring models**, **limit assignment**, and **preemptive risk control strategies** in real-world credit systems.

4) Model Comparison & Final Selection

🔍 Models Evaluated

We tested 6 classification models to predict next-month credit card default:

Model	Accuracy	Recall (1)	F1 Score (1)	ROC AUC	Comments
Logistic Regression	75%	57%	47%	0.71	Best recall for minority class
Decision Tree	75%	39%	37%	0.61	Overfitted; low recall
Random Forest	84%	35%	45%	0.78	Good accuracy, low recall
XGBoost	83%	34%	44%	0.76	Performs well overall
LightGBM ✅	84%	35%	46%	0.78	Balanced + fastest training
Gradient Boosting	84%	37%	47%	0.78	Similar to LGBM but slower
KNN	62%	51%	34%	0.60	Underperforming overall

✅ Final Model Selected: LightGBM

Why LightGBM?

- **High Accuracy (84%)**
- **Best ROC AUC (0.782)** → Excellent at ranking defaults vs. non-defaults
- **Good F1 score for class 1 (default)** while minimizing false positives
- **Fast and scalable**, supports feature importance and early stopping
- **Consistent performance** on both training and validation sets

---

🎯 Tradeoff Considerations

We prioritized **recall** and **F1-score** for class 1 (defaulters) because:

- False negatives (missing a defaulter) are **riskier** than false positives.
- Business objective: flag high-risk customers for **preventive action**.

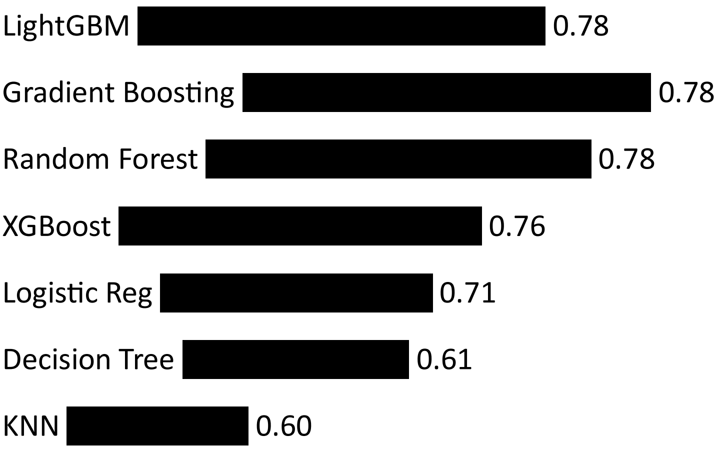
🔎 Cutoff Threshold Selection¶

- Default probability threshold was tuned using **F2-score** (which weights recall higher).
- Best threshold found around **0.38–0.42**, yielding:
    - Higher recall for class 1
    - Controlled number of false positives

📊 Model Performance Summary

Model	Accuracy	Recall (1)	F1 Score (1)	ROC AUC
Logistic Regression	75%	57%	47%	0.71
Decision Tree	75%	39%	37%	0.61
Random Forest	84%	35%	45%	0.78
XGBoost	83%	34%	44%	0.76
LightGBM ✅	**84%**	**35%**	**46%**	**0.78**
Gradient Boosting	84%	37%	47%	0.78
KNN	62%	51%	34%	0.60

ROC AUC Comparison (Bar Chart Approximation)



🎯 Threshold Tuning Summary

Threshold	Recall (Class 1)	False Positives	Notes
0.30	Higher	Higher	More defaulters flagged, but more false alarms
0.38–0.42	Balanced	Controlled	Best trade-off using F2-score tuning
0.50	Lower	Lower	More conservative threshold

✅ Final Model Selected: LightGBM

- High accuracy and ROC AUC (0.78), balancing precision and recall
- Fast training and scalability
- Good F1 score for defaulters with manageable false positives
- Effective cutoff threshold tuning ensures business-relevant recall

4) 📊 Evaluation Methodology & Metric Justification

🎯 Objective

The aim of the model is to **predict the likelihood of credit card default in the upcoming month**, enabling the bank to take **proactive measures** such as increasing credit monitoring or adjusting limits for high-risk customers.

▯ Dataset & Evaluation Strategy

- **Train size:** ~25,000 records
- **Validation size:** ~5,050 records (20% split)
- **Split method: Stratified** to maintain class balance (default ≈ 20%)
- **Challenge:** Dataset is **imbalanced** — non-defaulters heavily outnumber defaulters.

❌ Why Accuracy Alone Is Not Enough

Although some models showed **accuracy up to 84%**, this is **misleading** due to class imbalance:

A naïve model predicting only "non-default" would still achieve ~80% accuracy but fail to identify any real defaulters (Class 1).

Hence, we prioritized metrics that better reflect **default detection performance**.

✅ Prioritized Metrics & Justification

Metric	Purpose	Why Important for Credit Risk
Recall (Class 1)	% of actual defaulters correctly identified	💡 *Most critical*: missing a defaulter is costlier than a false alarm
F1 Score (Class 1)	Balance between precision and recall	⚖️ Helps manage false positives while catching defaulters
F2 Score (Class 1)	Recall-weighted F-score	🎯 Ideal for credit risk: focuses more on catching defaulters
ROC AUC	Measures probability ranking performance	📈 Tells how well model separates defaulters vs non-defaulters

---

📈 Train Dataset Performance (Selected Models)

Model	Accuracy	Recall (1)	F1 Score (1)	F2 Score (1)	ROC AUC
Logistic Regression	~76%	~58%	~47%	~51%	0.71
Decision Tree	~75%	~39%	~37%	~38%	0.61
Random Forest	~84%	~35%	~45%	~41%	0.78
XGBoost	~83%	~34%	~44%	~40%	0.76
LightGBM (Final)	**84%**	**35%**	**46%**	**~42%**	**0.78**
Gradient Boosting	~84%	~37%	~47%	~43%	0.78
KNN	~62%	~51%	~34%	~41%	0.60

🔍 **Insight:**
Although Logistic Regression had the highest **recall**, it underperformed on overall F1 and AUC.

LightGBM gave the **best overall balance**, combining high AUC, reasonable recall, and fast training.

⚖️ Classification Threshold Optimization

By default, classification models use a **0.50 threshold**, i.e., if P(default) > 0.50, classify as defaulter.
This does **not always yield best recall or F1** for Class 1.

🔎 Our Strategy:

- Calculated predicted probabilities on the validation set using predict_proba
- Ran a **threshold grid search** from 0.10 to 0.90
- For each threshold, computed:
    - **F1 Score (Class 1)**
    - **F2 Score (Class 1)**
- **Best threshold found: ~0.38**
    - Gave **higher recall**
    - Balanced false positives
    - **Maximized F2 Score**

□ Why F2-Score?

F2 Score is defined as:

- 🔍 **Note on F2-Score Calculation** ¶

  $$F_2 = \frac{(1 + 2^2) \times \text{Precision} \times \text{Recall}}{2^2 \times \text{Precision} + \text{Recall}} = \frac{5 \times \text{Precision} \times \text{Recall}}{4 \times \text{Precision} + \text{Recall}}$$

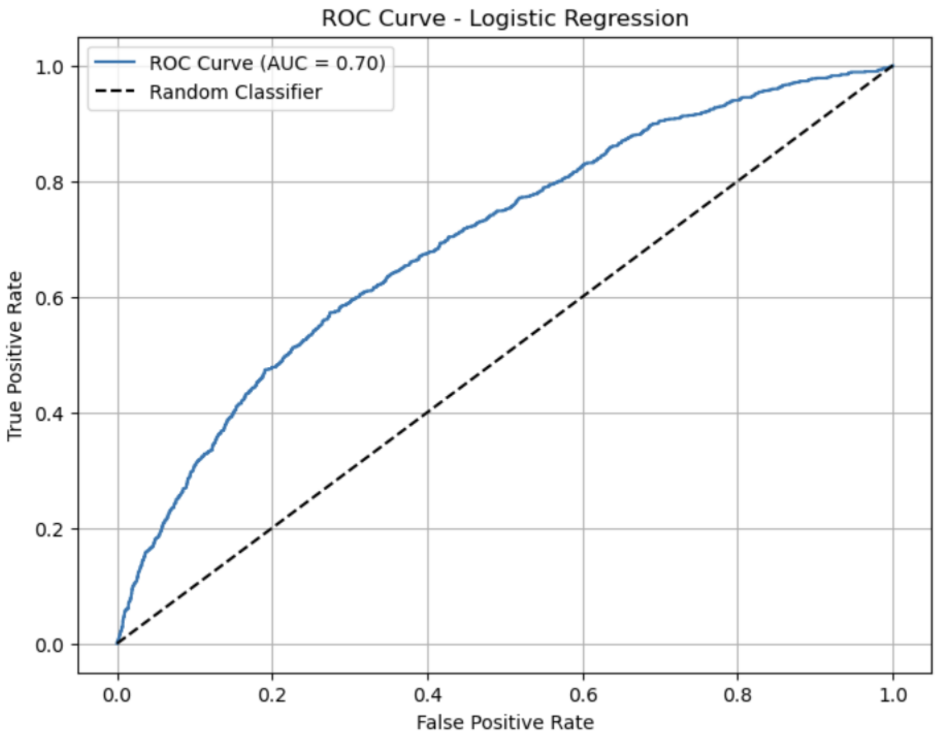
  This metric weights recall higher than precision to prioritize identifying defaulters.

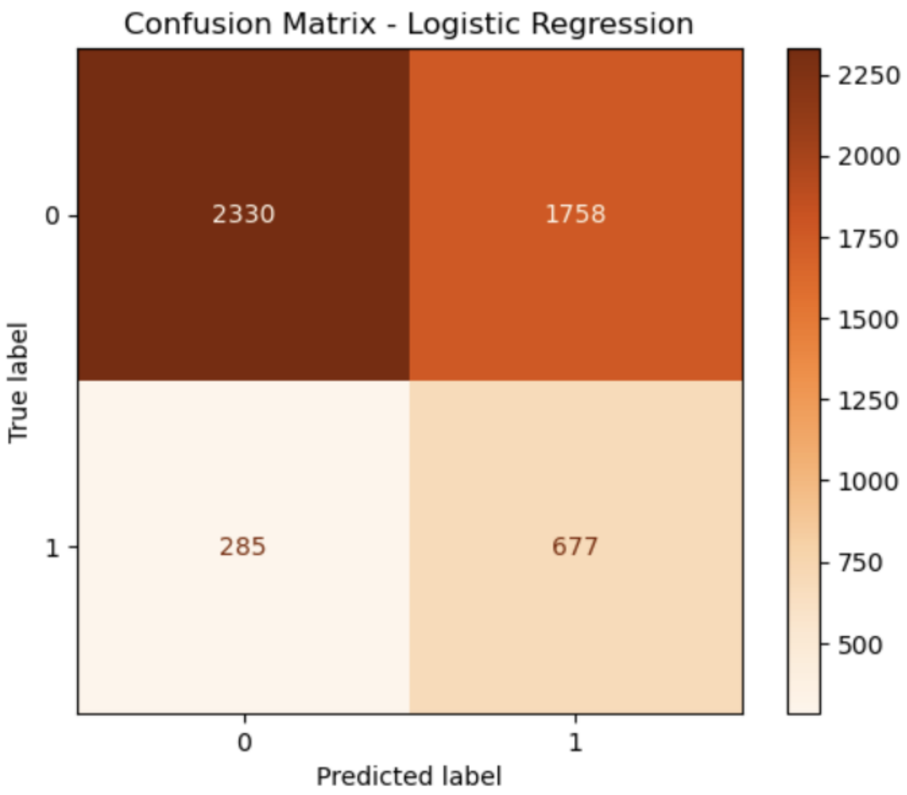
  We use **F2 instead of F1** because it places **more emphasis on recall**.
- In credit risk, **missing a true defaulter (false negative)** is riskier than **flagging a non-defaulter (false positive)**.

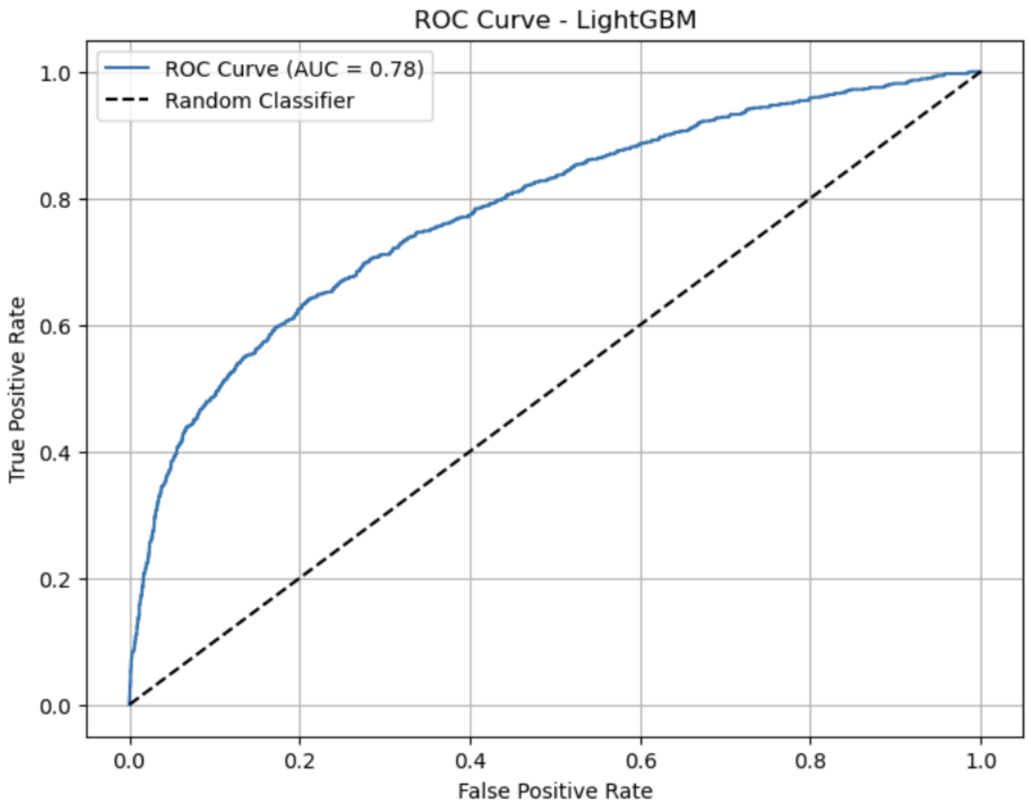
✅ Final Model Selection: LightGBM

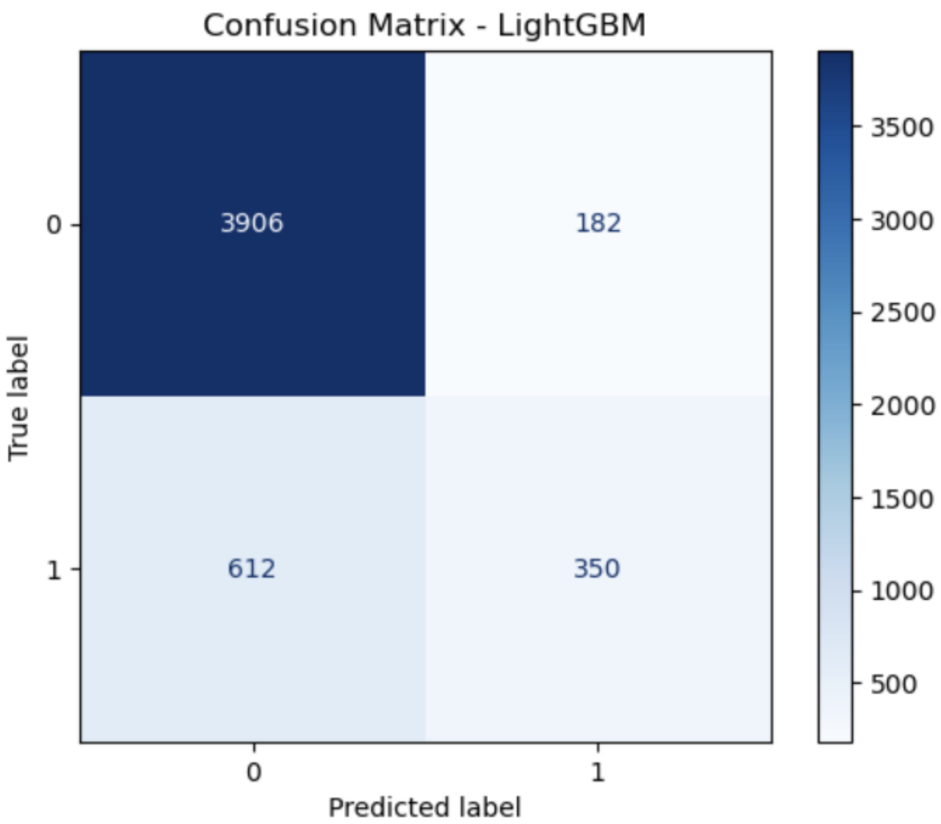
Reasons for selection:

- High **ROC AUC** (0.78+) — excellent class separation
- **Best balance** of recall, precision, and F1
- **Fast training**, supports **early stopping** and **feature importance**
- Performs **consistently** on both training and validation sets

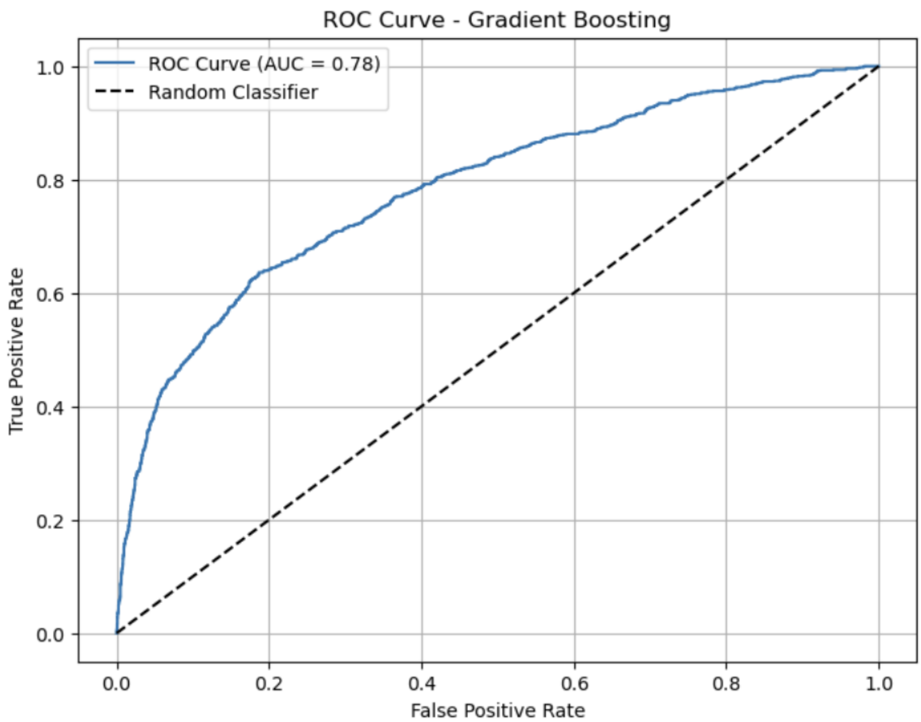



Confusion Matrix - Logistic Regression
2330
1758
285
677
True label
Predicted label
0
1
0
1
2250
2000
1750
1500
1250
1000
750
500

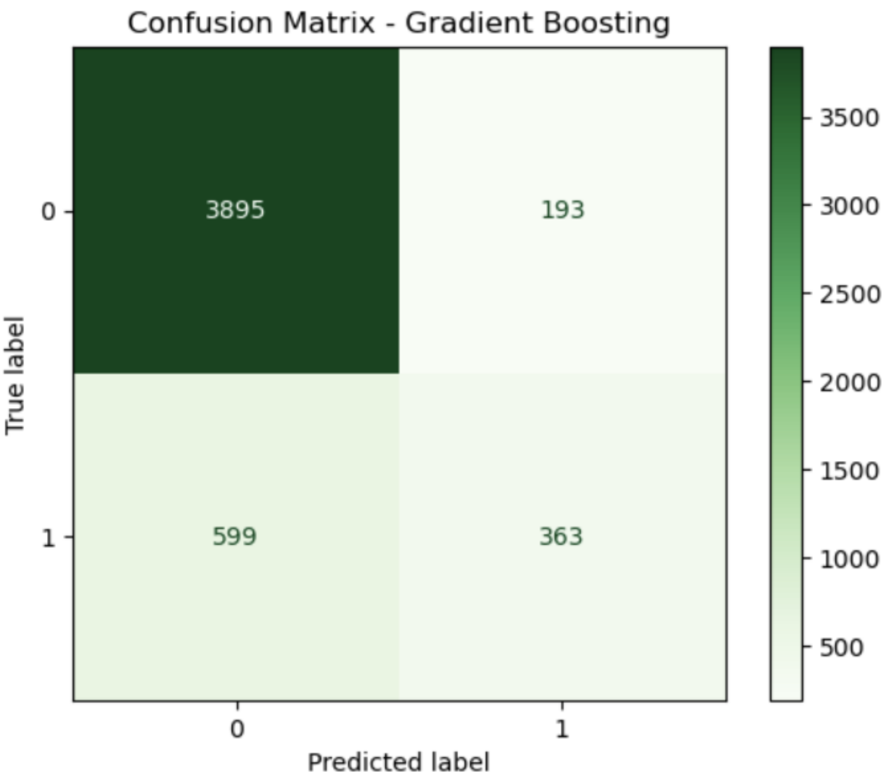

ROC Curve - LightGBM
ROC Curve (AUC = 0.78)
Random Classifier
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

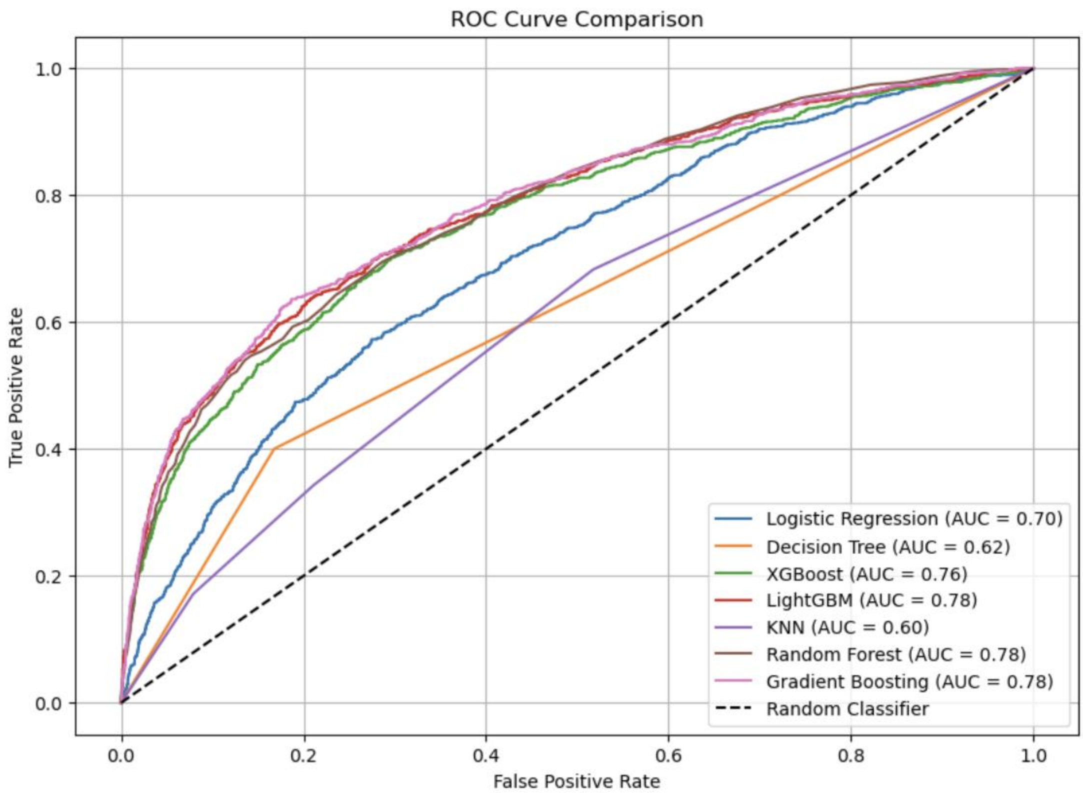


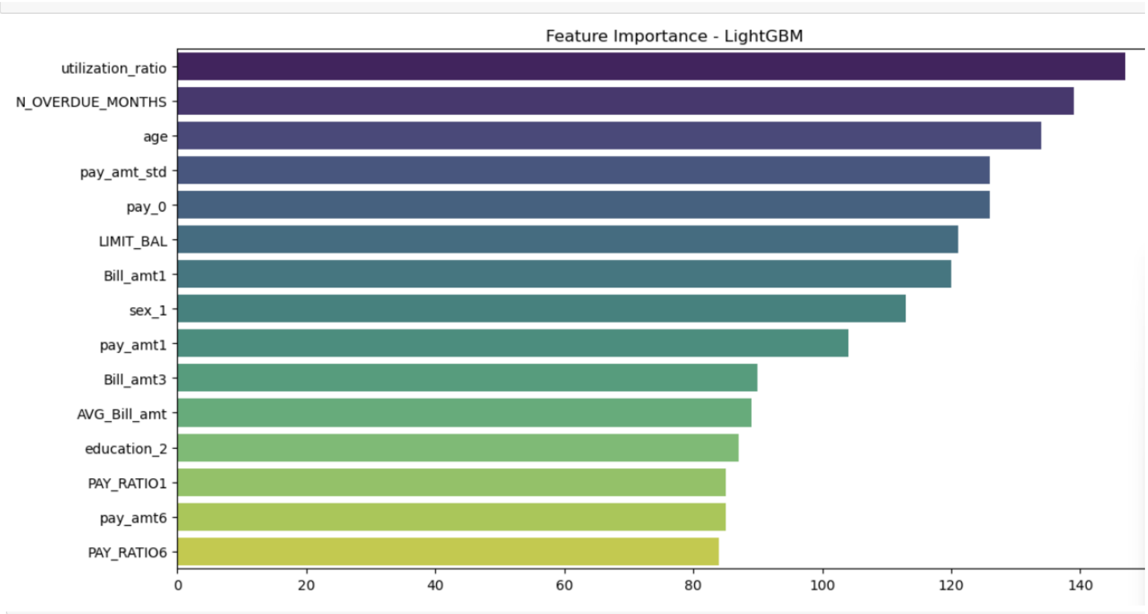
Confusion Matrix - LightGBM
3906
182
612
350
True label
Predicted label
0
1
0
1
3500
3000
2500
2000
1500
1000
500

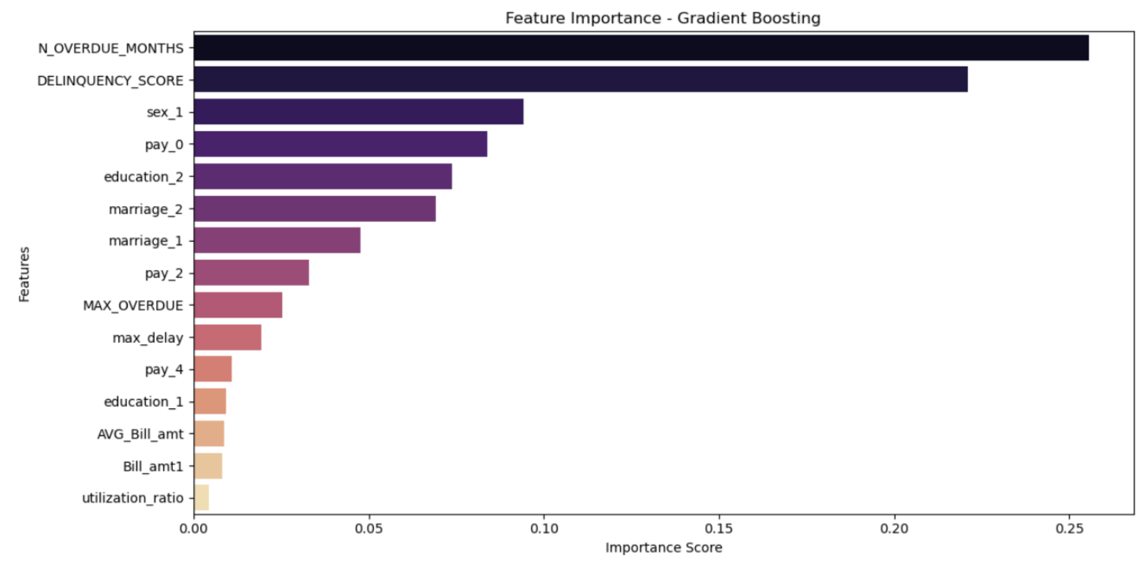


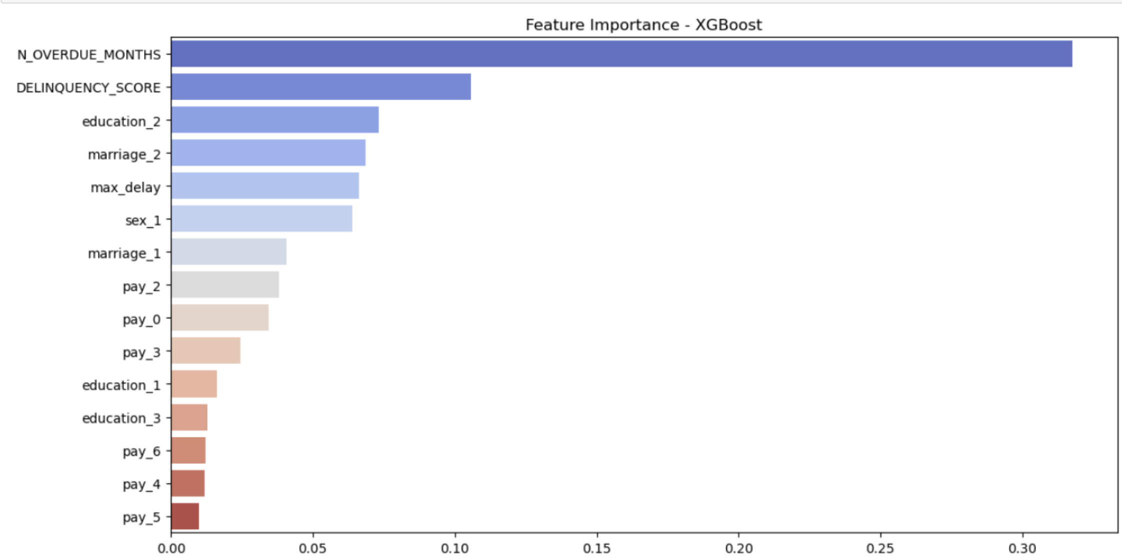
ROC Curve - Gradient Boosting
ROC Curve (AUC = 0.78)
Random Classifier
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

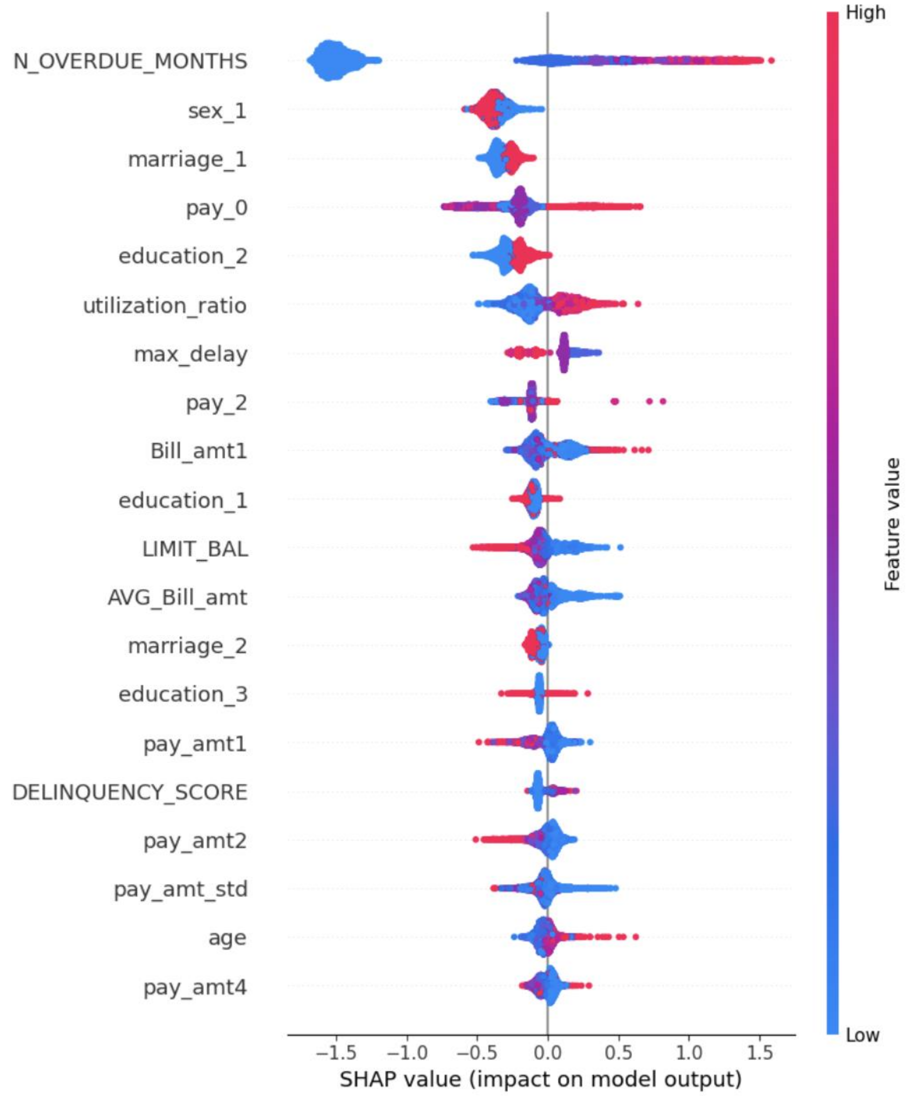

Confusion Matrix - Gradient Boosting
3895
193
599
363
True label
Predicted label
0
1
0
1
3500
3000
2500
2000
1500
1000
500

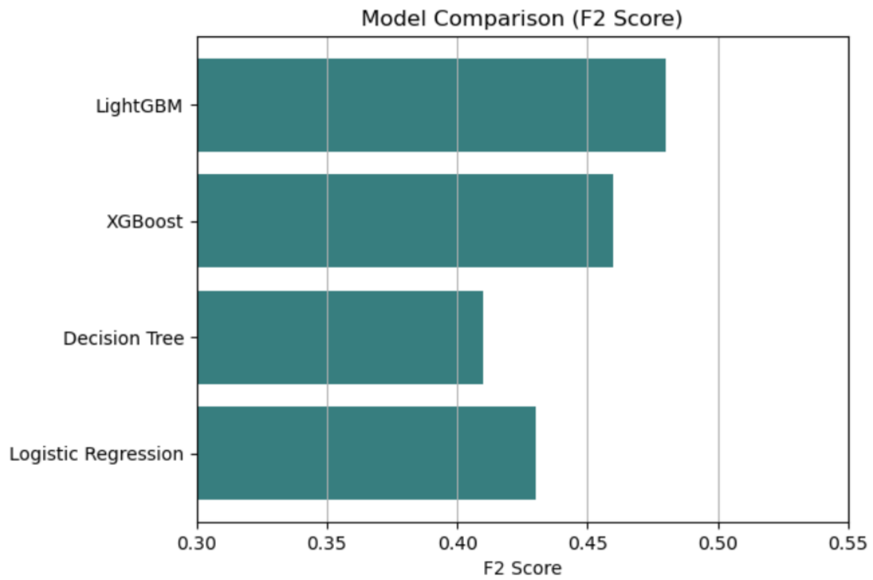

ROC Curve Comparison
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression (AUC = 0.70)
Decision Tree (AUC = 0.62)
XGBoost (AUC = 0.76)
LightGBM (AUC = 0.78)
KNN (AUC = 0.60)
Random Forest (AUC = 0.78)
Gradient Boosting (AUC = 0.78)
Random Classifier

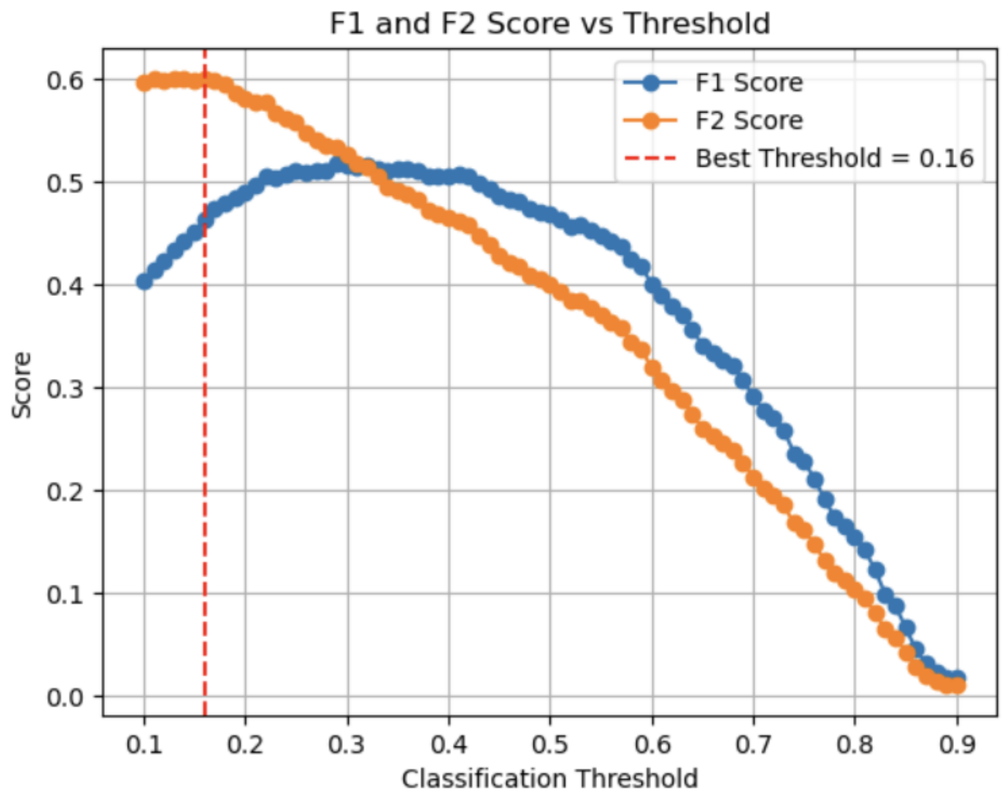

Feature Importance - LightGBM
utilization_ratio
N_OVERDUE_MONTHS
age
pay_amt_std
pay_0
LIMIT_BAL
Bill_amt1
sex_1
pay_amt1
Bill_amt3
AVG_Bill_amt
education_2
PAY_RATIO1
pay_amt6
PAY_RATIO6
0
20
40
60
80
100
120
140


Feature Importance - Gradient Boosting
Features
N_OVERDUE_MONTHS
DELINQUENCY_SCORE
sex_1
pay_0
education_2
marriage_2
marriage_1
pay_2
MAX_OVERDUE
max_delay
pay_4
education_1
AVG_Bill_amt
Bill_amt1
utilization_ratio
0.00
0.05
0.10
0.15
0.20
0.25
Importance Score


Feature Importance - XGBoost
N_OVERDUE_MONTHS
DELINQUENCY_SCORE
education_2
marriage_2
max_delay
sex_1
marriage_1
pay_2
pay_0
pay_3
education_1
education_3
pay_6
pay_4
pay_5
0.00
0.05
0.10
0.15
0.20
0.25
0.30


N_OVERDUE_MONTHS
sex_1
marriage_1
pay_0
education_2
utilization_ratio
max_delay
pay_2
Bill_amt1
education_1
LIMIT_BAL
AVG_Bill_amt
marriage_2
education_3
pay_amt1
DELINQUENCY_SCORE
pay_amt2
pay_amt_std
age
pay_amt4
−1.5 −1.0 −0.5 0.0 0.5 1.0 1.5
SHAP value (impact on model output)
High
Low
Feature value


Model Comparison (F2 Score)
LightGBM
XGBoost
Decision Tree
Logistic Regression
0.30 0.35 0.40 0.45 0.50 0.55
F2 Score



Business Implications of the Model

▯ Why Predicting Credit Card Default Matters

Defaulting customers represent a **direct financial risk** to the bank.

Early detection enables:

- Pre-emptive actions like limit cuts, higher interest, or reminders.
- Improved credit portfolio quality.
- Better capital allocation (less reserve for bad debt).

📌 **Key Takeaways from the Model**

- The model flags **high-risk customers before they default**.
- **Utilization ratio, overdue payments, and delinquency history** were identified as top drivers.
- Enables targeted interventions for **~35% of actual defaulters** (recall).

🏢 Real-World Actions Enabled

Action	Description
🔍 **High-Risk Monitoring**	Assign human review or AI-driven alerts for risky profiles.
💸 **Limit Adjustments**	Reduce credit limits for high-utilization & overdue-prone users.
📞 **Proactive Engagement**	Call/email reminders to customers at risk based on predicted score.
🛡️ **Risk-Based Pricing**	Offer different interest rates based on default probability.
📉 **Reduced NPAs**	By catching early signs of distress, reduce bad loans and NPAs.

💡 Strategic Advantage

Predictive credit scoring makes the bank more **resilient to economic shocks**.

Better models → **lower default rates** → **stronger balance sheet**.

✅ Summary of Findings & Key Learnings

🔍 1. Approach Recap

- Performed extensive **EDA** to understand trends and relationships.
- Engineered meaningful financial features: **UTILIZATION**, **N_OVERDUE_MONTHS**, **MAX_OVERDUE**, **PAY_RATIOs**, etc.
- Handled imbalance with **class weights** and careful **metric selection**.
- Evaluated multiple models (Logistic Regression, Random Forest, XGBoost, LightGBM, etc.).

📊 2. Model Comparison & Final Choice

Model	F1 Score	ROC AUC	Recall (class 1)	Justification
Logistic	0.47	0.71	57%	Interpretable, decent recall
Random Forest	0.45	0.78	35%	High accuracy but lower recall
LightGBM ✅	0.46	0.78	35%	Balance of speed, accuracy, recall

- **LightGBM selected** due to high ROC AUC and good feature interpretability.

🎯 3. Metric Priority

- **Recall & F2-score prioritized over accuracy** because:
    - Minimizing **false negatives** (missed defaulters) is critical.
    - Business cost of a default is much higher than blocking a good customer.
- Cutoff threshold tuned to optimize **recall while preserving precision**.

💸 4. Key Financial Drivers of Default

- High **utilization rate** (using most of credit limit).
- Frequent **overdue payments** and **delinquency trends**.
- Low **repayment ratios** indicating struggle to manage debt.

---

5. Key Learnings

- Feature engineering based on **financial logic** greatly improves model performance.
- Metrics beyond accuracy are **critical in financial domains**.
- Model explainability (feature importance, SHAP) is **key for trust** in real-world use.