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

## Overview of Approach and Modeling Strategy

The goal of this project is to develop a predictive model for Bank that flags credit card customers who are likely to default in the following month. To achieve this, we used a supervised classification approach, applying multiple machine learning models and performing detailed feature analysis to ensure interpretability and reliability.

**Steps:**

- Data loading and inspection
- Exploratory Data Analysis (EDA)
- Data cleaning and preprocessing
- Feature transformation (e.g., encoding and mapping categories)
- Model building using classification algorithms (Logistic Regression,Random Forest, XGBoost, etc.)
- Model evaluation and selection using relevant metrics
- Interpretation of results for business relevance

## EDA Findings and Visualizations

- The dataset contains over 30,000 credit card customers with behavioral and demographic variables.
- Target variable (next_month_default) shows class imbalance — most customers do not default. This means that if we make model from this data it will overfit.
- Age columns have 126 null values.
- No non-numeric features.
- Descriptive Statistics: Computed mean, median, min, max, std dev for numeric features.
- Variable distribution analysis through histograms revealed skewness in repayment and bill amount features.

## Data Cleaning

- Missing data in Age column is replaced by median. Median imputation is less sensitive to outliers and provides a central tendency estimate, making it a reliable choice for skewed data
- All fully duplicated rows were removed to ensure data uniqueness and integrity.
- The education column contained undefined or rare codes such as 0, 4, 5, and 6. So they all are mapped to 4. And similarly with Marriage 0,3 is mapped to 3 . This will Reduces category noise and groups all ambiguous cases into one interpretable category.

- The column pay_0 was renamed to pay_1 to maintain a logical and chronological naming convention from pay_1 to pay_6. This improves clarity during analysis and modeling by aligning all payment status variables sequentially.

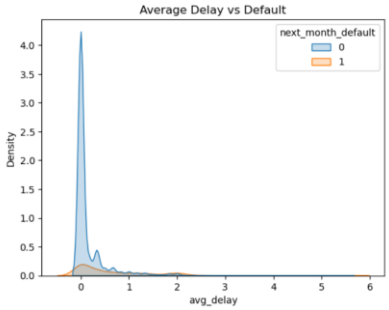
# Advance EDA Findings

- Approximately 20% default rate, confirming imbalance.
- LIMIT_BAL had a right-skewed distribution.High limits slightly reduce default risk; however, it is not strongly correlated.
- Low credit limit, younger age, and late payment patterns are more common in defaulters.
- Repayment Status (pay_0, pay_2, etc.): Strong positive correlation with default. Customers with delayed payments in past months are more likely to default again.
- Strong positive correlation between consecutive pay_X values.
- avg_delay had a meaningful correlation with next_month_default
- LIMIT_BAL: Negatively correlated with default. Lower limit users are riskier.
- Bill Amount & Repayment Amount: Have weak negative correlation, but when analyzed in ratios (e.g., PAY_TO_BILL_ratio), they reflect financial discipline.
- Payment Amounts (pay_amt1 to pay_amt6): Positive payments reduce risk; missed payments increase likelihood of default.
- Bivariate Analysis:
  - Correlation heatmap with the target variable.
  - Distribution plots of key variables segmented by default status.

# ❖Behavioral Feature Analysis

❖ Late repayments are a strong signal of credit stress.
❖ High and rising bills without proportional repayment is a red flag.
❖ Lower or inconsistent repayment often correlates with default.
❖ Customers with very low limits tend to default more.

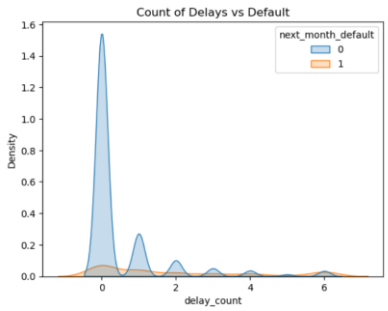
## 1. Average Delay vs Default




This grapgh shows:

- Customers who had higher average delays in the past are more likely to default next month. This makes avg_delay a strong behavioral predictor of credit risk.

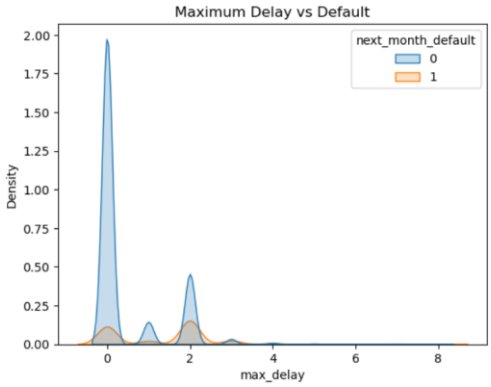
## 2. Count of Delayed Payments vs Default




This graph shows:

- Customers who have delayed their payments more frequently are at a higher risk of defaulting in the next month.
- The count of delayed payments is an important behavioral feature to predict credit default.

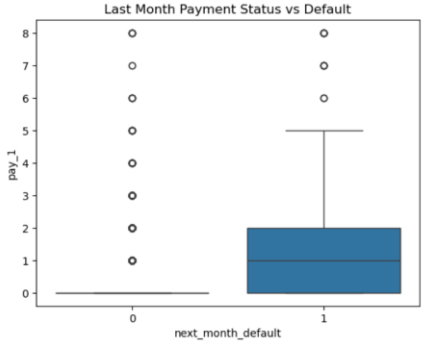
## 3. Maximum Delay vs Default




This graph shows:

- Customers who have had even one significant payment delay are more likely to default in the next month. This makes maximum delay a strong risk indicator

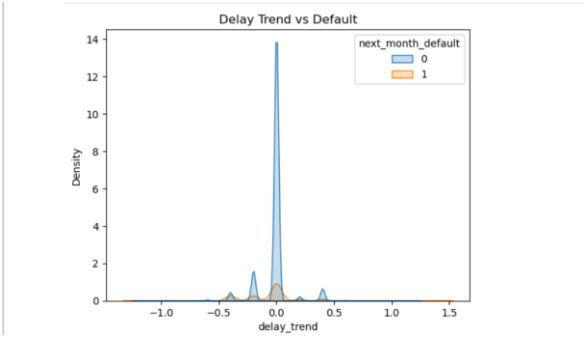
## 4. Recent Delay (Last Month) vs Default




This graph shows:

- Defaulters typically had higher last-month delay values. Non-defaulters mostly paid on time or early. So, the most recent payment behavior is highly predictive of future default.

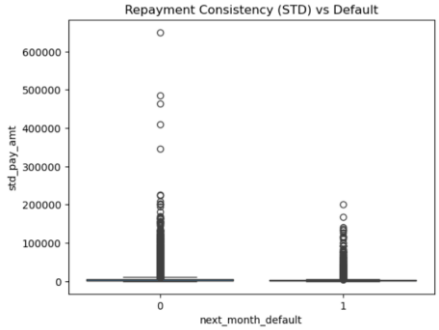
## 5. Trend of Delay (Increasing or Decreasing)




This graph shows:

- Worsening delay trends over months indicate a higher risk of default, even if past delays weren't severe. This trend metric can act as an early warning signal.

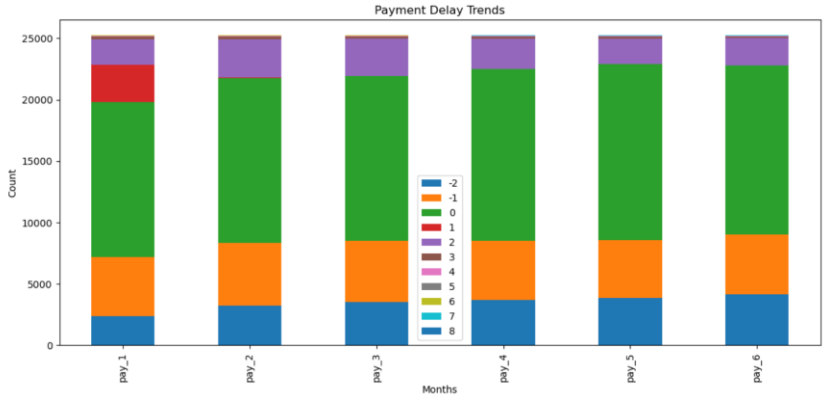
## 6. Repayment Consistency (STD) vs Default




This graph shows:

- Inconsistent payment behavior is linked to higher default risk.
- Stable payment patterns (low STD) are characteristic of reliable customers.

## 7. Payment Status Trend




This graphs shows:

- Majority of payments are made on time:
  - ❖ The green bar (0) dominates in all months, indicating most users pay on time.
- Stable payment trend over time:
  - ❖ The stacked pattern is consistent across pay_1 to pay_6.
  - ❖ No significant worsening or improvement in overall user behavior.
- Delays exist but are fewer in number:
  - ❖ Orange to purple segments (1 to 4) show delayed payments.
  - ❖ Very few users have extreme delays (6, 7, 8) — these are the brown/gray thin layers at the top.

# Financial Insights and Analysis

## (a) Credit Utilization Ratio

- Reflects how much of their credit the customer is actually using.
- Higher utilization (e.g., >80%) is a red flag — it often indicates that the customer is dependent on credit and might be financially stressed.
- A low utilization ratio (e.g., <30%) shows healthy borrowing behavior.
- Customers with a consistently high utilization ratio and low repayments are likely to default.

## (b) Delinquency Streak

- A “delinquency streak” refers to the number of consecutive months where the customer missed or delayed payments
- One-time delay might not be a big issue, but repeated delays are highly predictive of future default.
- A streak of 3+ months is a strong signal of long-term credit distress.

## (c) Payment Consistency (Standard Deviation of Payments)

- Consistent payers usually have predictable income and spending habits.
- Highly variable payers may have unstable income, seasonal cash flow, or financial distress.

# Feature Selection

- Helps surface behavioral patterns from raw data.
- Improves model interpretability and performance by encoding real-world financial intuition.
- In our project, we employed a systematic feature selection process combining both domain knowledge and statistical methods to identify the most impactful features for predicting the target variable
- The final feature set reflects a balanced selection of demographic, financial, and behavioral characteristics.

# Handling Imbalance Dataset

- Handle the imbalance dataset with SMOTE to reduce overfitting due to imbalance dataset.

# Data Preprocessing

- Standardization ensures models aren't biased by feature scales.
- Used StandardScaler to normally distribute the data.
- Then train test split have been done.

# Evaluation Methodology

- Recall (Sensitivity) — Highest Priority
  - In credit risk, false negatives (predicting a customer will not default when they actually will) are very costly.
  - A high recall ensures the bank catches as many defaulters as possible before issuing or continuing credit, minimizing potential losses
- F2 Score — Recall-Weighted Measure
  - Gives more weight to recall than precision, aligning with business needs to minimize missed defaulters.
  - Ideal when missing a defaulter is much worse than incorrectly flagging a few non-defaulters. F2 score reflects this trade-off in real banking risk scenarios.
- F1 Score — Balanced View
  - It ensures the model does not just catch many defaulters (recall), but also that the defaulters it predicts are actually correct (precision),avoiding too many false alarms.
- In this project, recall, F1 score, and F2 score were prioritized over raw accuracy because:
  - Catching defaulters early is more important than avoiding a few false alarms.
  - Business losses are driven by missed defaulters, not overly cautious predictions.

# Selection of classification cutoff

- In this project,instead of using the default threshold of 0.5 (used by predict()), we first obtained predicted probabilities using predict_proba():
- For each threshold, evaluation metrics were calculated—particularly focusing on F2-score, which gives more weight to recall. This is suitable in domains where missing a positive case (false negative) is costlier than raising a false alarm.
- A loop structure was used to evaluate F2-score at each threshold. The threshold yielding the best F2-score was selected for every model.
- The F2-score was prioritized over F1 because in our context, recall (minimizing false negatives) was more important than precision.
- The selection was based on empirical performance(F2 Scores) rather than relying on standard defaults. This ensures the model is more aligned with the actual business or application goals.

# Model Comparison and Justification for Final Selection

## 1. Logistic Regression

- **Pros:** Simple, interpretable, coefficients directly explain impact of features.
- **Cons:** Linear assumptions limit performance on complex patterns.
- **Performance:** Moderate recall and F1 score; low F2 score compared to ensemble models.

## 2. Random Forest

- **Pros:** Handles feature interactions well, robust to noise, reduces overfitting.
- **Cons:** Less interpretable than single tree; requires tuning.
- **Performance:** Good F1 and F2 scores, good balance of recall and precision and better than Logistic Regression.

## 3. LightGBM (Light Gradient Boosting Machine)

- **Pros:** Extremely fast training speed and low memory usage.
- **Cons:** Less interpretable than simpler models like logistic regression.
- **Performance:** Better performance than Random Forest and Logistic Regression. High F2 score, strong recall, and fast training

## 4. XGBoost (Extreme Gradient Boosting)

- **Pros:** High predictive power, handles missing values, supports regularization.
- **Cons:** Slightly complex to tune, less transparent than logistic regression.
- **Performance**: Best performing model — highest F2 score, strong recall, good generalization across train/test sets.

### Justification for Final Model Selection: XGBoost

- Chosen Model: XGBoost was selected as the final model.
- Reason:
  - It provided the highest F2 score, aligning with our business goal of maximizing recall.
  - Showed strong ability to capture complex patterns in customer behavior.
  - Performs well with imbalanced data, which is critical in credit default problems.

# Metrics result on train dataset

| Model | Accuracy | Recall | F1-Score | F2-Score | ROC-AUC |
|---|---|---|---|---|---|
| Logistic Regression | 0.58 | 0.97 | 0.70 | 0.84 | 0.83 |
| Random Forest | 0.72 | 0.94 | 0.78 | 0.868 | 0.89 |
| LightGBM | 0.81 | 0.93 | 0.83 | 0.886 | 0.937 |
| XgBoost | 0.80 | 0.93 | 0.82 | 0.887 | 0.9366 |

**Final model: XGBoost**, due to its:

- Superior recall and F2 score
- Final F2 Score: 88.7% and Accuracy: 80%
- At last just to increase accuracy , I have tried many thresholds with XgBoost and found out that better accuracy(0.86) along with better F2 score(0.872) is coming.-->(Just a trial, don't consider this F2 Score for evaluation as final F2 Score is mentioned above)

# Business Implications

- The model enables early identification of high-risk customers, allowing Bank A to:
  - Adjust credit limits or freeze accounts preemptively.
  - Offer restructuring or advisory services.
  - Trigger early warning systems for accounts showing risky behavior.
  - Interpretability of selected features aids in compliance and regulatory reporting.

### 1. Reduced Credit Losses

- Early identification of high-risk customers allows the bank to take preventive actions (e.g., limit reduction, early collection efforts).
- This directly reduces Non-Performing Assets (NPAs) and charge-offs.

### 2. Better Risk-Based Decision Making

- With probability scores and interpretable risk factors, the bank can:
  - Approve/reject loans more confidently
  - Offer differentiated interest rates or limits
- Leads to smart lending with optimized risk-reward balance

### 3. Improved Customer Targeting & Segmentation

- Customers can be segmented into:
  - Low-risk (upsell/loyalty benefits)
  - Medium-risk (monitor or offer flexible payment options)
- High-risk (restrict, restructure, or collect early)
- Enables personalized risk strategies instead of one-size-fits-all.

### 4. Operational Efficiency

- Automating the risk scoring system saves manual effort.
- Credit teams can focus on borderline or high-risk cases, not review every customer manually.

# Summary of Findings and Key Learnings

- Detailed exploratory data analysis, we uncovered strong behavioral signals such as credit utilization ratio, payment irregularities, and overdue streaks that significantly correlate with default risk.
- Feature engineering enhanced these patterns, while model experimentation showed that ensemble methods like XGBoost performed best, especially in optimizing recall and F2 score — critical metrics in minimizing business loss.
- By tuning the classification threshold, we aligned model behavior with real-world risk tolerance.
- Importantly, we learned that financial modeling is not just about prediction accuracy, but about interpretable risk insights, actionable thresholds, and aligning technical decisions with business goals such as customer segmentation, early intervention, and loss prevention.
- The project reinforced the value of domain-driven feature engineering and metric prioritization tailored to credit risk.