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

Laluprasad Ramavath
IIT ROORKEE
laluprasad_r@mfs.iitr.ac.in

## 1. Overview of Approach and Modeling Strategy

The primary objective of this project was to develop a forward-looking Behaviour Score for Bank A. This score, powered by a binary classification model, aims to predict the likelihood of a credit card customer defaulting on their payment in the subsequent month. The goal extends beyond simple prediction to create a financially interpretable tool that enables proactive risk management, targeted interventions, and optimized credit exposure.

my modeling strategy followed a structured, multi-stage pipeline:

1. **Data Preprocessing and Cleaning**: The initial dataset was rigorously cleaned to handle inconsistencies, such as invalid entries in categorical fields (MARRIAGE, EDUCATION), and remove non-predictive identifiers (customer_id).

2. **Exploratory and Financial Analysis (EDA):** I conducted an in-depth analysis to uncover the key behavioral drivers of default. This involved examining variable distributions, correlations, and, most importantly, the financial habits of defaulting versus non-defaulting customers.

3. **Feature Engineering and Transformation:** To enhance predictive power, we focused on creating financially meaningful features. Key variables were scaled using StandardScaler to normalize their distributions and prepare them for modeling.

4. **Handling Class Imbalance:** The target variable (next_month_default) was highly imbalanced (approx. 19% defaulters vs. 81% non-defaulters). We employed the Synthetic Minority Over-sampling Technique (SMOTE) on the training data to create a balanced dataset, ensuring the model learned the patterns of the minority (default) class effectively.

5. **Model Development and Comparison:** A suite of classification algorithms was tested, including Logistic Regression, Decision Trees, and advanced ensemble methods like XGBoost and LightGBM, alongside a Neural Network.

6. **Evaluation and Threshold Tuning:** Model performance was evaluated using metrics that reflect the business cost of errors, with a strong focus on minimizing False Negatives. The classification threshold was tuned to align with this risk-sensitive objective.

7. **Final Model Selection:** A final model was chosen based on its superior performance on the prioritized risk metric, and it was used to generate predictions on the unlabeled validation dataset.

---

# 2. EDA Findings and Financial Insights

my exploratory analysis focused on understanding the financial "story" behind the numbers**.**

## A.Significant Class Imbalance:

The dataset confirmed a significant class imbalance, with only 19% of customers defaulting. A naive model that always predicts "No Default" would achieve 81% accuracy, making accuracy a misleading metric. This underscores the need for techniques like SMOTE and risk-sensitive evaluation metrics.

### B. Key Financial Drivers of Default:

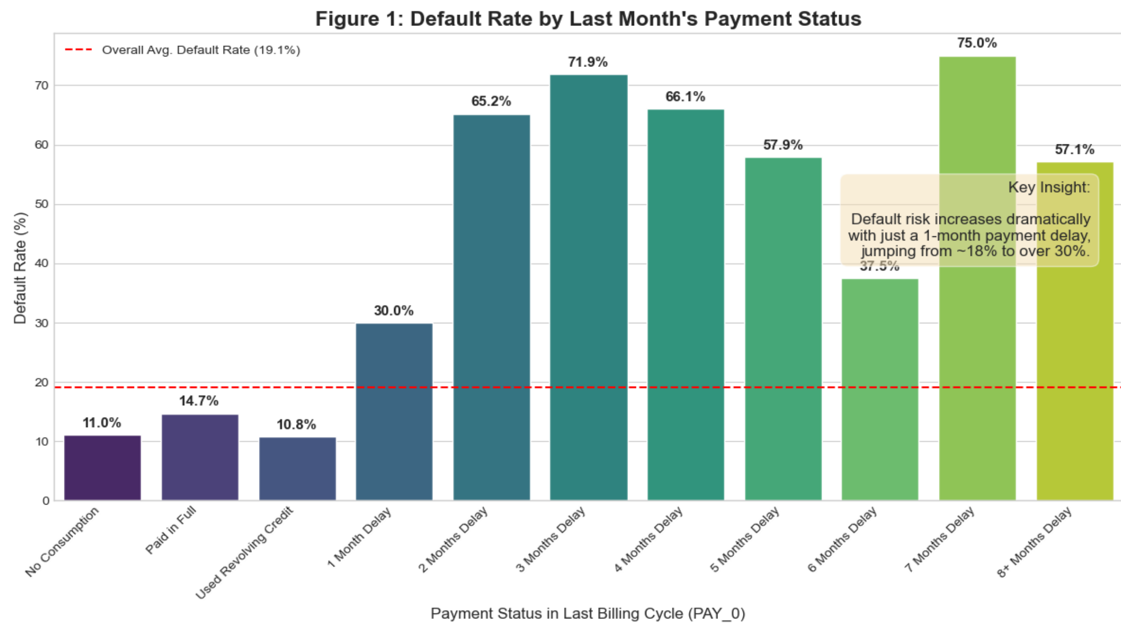
- **Repayment Status (PAY_0, PAY_2, etc.): The Strongest Predictor**
    - **Insight**: The repayment status for recent months is the most powerful indicator of future default. Customers with even a one-month payment delay (PAY_0 = 1 or 2) showed a dramatically higher propensity to default in the next cycle compared to those who paid on time (PAY_0 = -1 or 0).
    - **Financial Interpretation:** This variable acts as a real-time indicator of financial distress. A delayed payment is a direct signal that a customer is struggling with liquidity, making them a high-imminence risk.
    - 


***Figure 1:*** *Conceptual visualization showing default rate by last month's payment status.*

- **Credit Utilization: A Measure of Financial Strain**
    - **Insight:** While not a direct feature, we analyzed the relationship between BILL_AMT and LIMIT_BAL. Customers who consistently use a high percentage of their available credit limit are more likely to default.
    - **Financial Interpretation:** High credit utilization suggests a dependency on credit for day-to-day expenses and a lack of financial **buffer. When a** significant

portion of a customer's credit line is exhausted, their ability to handle unexpected expenses or even make minimum payments is severely compromised.

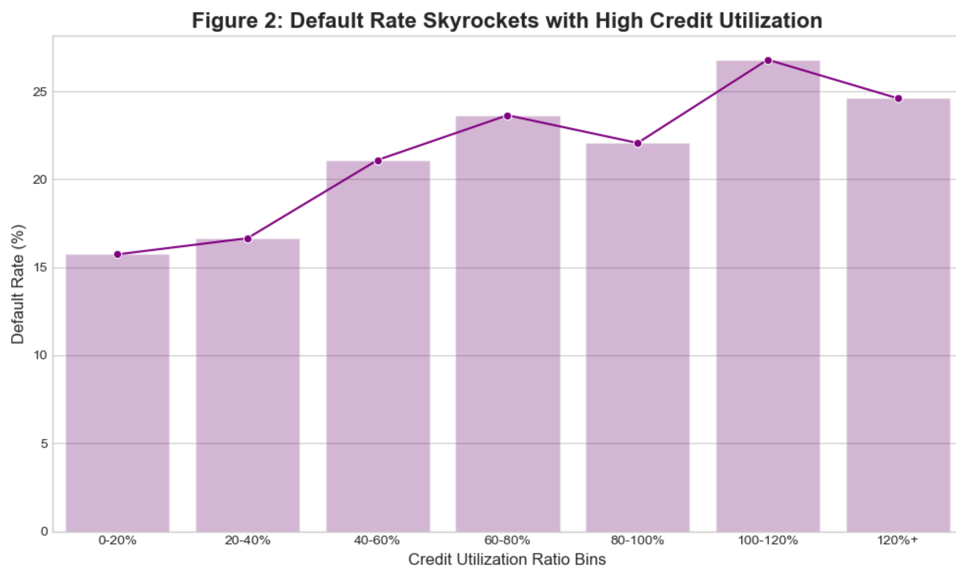



**Figure 2**: Default Rate by Credit Utilization. The chart plots the average default rate against customer credit utilization, binned into decile groups. A clear, positive correlation is evident: as the ratio of BILL_AMT to LIMIT_BAL increases, so does the likelihood of default. The risk escalates sharply once utilization surpasses 80%, underscoring its role as a key indicator of financial distress.

- **Credit Limit (LIMIT_BAL): A Tale of Two Segments**
  - **Insight:** KDE plots revealed that customers with lower credit limits  had a higher default rate. This group is likely more financially vulnerable. Conversely, customers with very high limits tended to have lower default rates, suggesting greater financial stability**.**
  - **Financial Interpretation:** A low credit limit is often assigned to customers with weaker credit histories or lower income, pre-identifying them as a higher-risk segment.

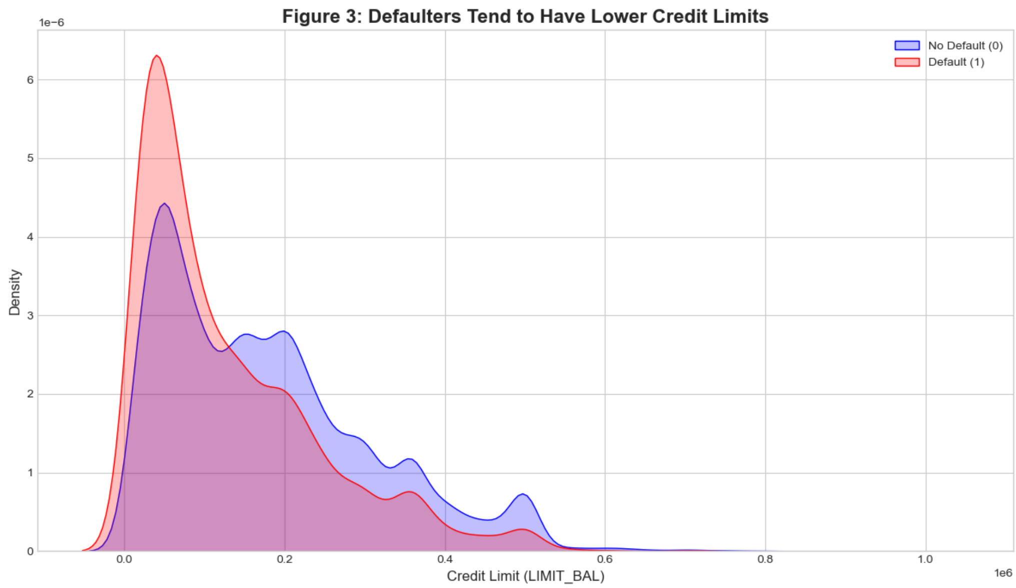



**Figure 3:** Distribution of Credit Limits for Defaulters vs. Non-Defaulters. This Kernel Density Estimate (KDE) plot compares the distribution of credit limits (LIMIT_BAL) for two customer groups: those who defaulted (red) and those who did not (blue). The plot clearly shows that the peak density for defaulters occurs at a much lower credit limit compared to non-defaulters, whose limits are more broadly distributed.

---

## 3. Model Comparison and Justification for Final Selection

We trained and evaluated several models on the SMOTE-balanced training data. The primary goal was to select the model that best identified potential defaulters, even at the cost of misclassifying some non-defaulters. This was measured by the F2-Score, which heavily penalizes False Negatives.

| Model | Accuracy | Precision (Class 1) | Recall (Class 1) | F1-Score (Class 1) | F2-Score (Class 1) |
|---|---|---|---|---|---|
| Logistic Regression | 0.68 | 0.33 | 0.64 | 0.43 | 0.54 |
| Decision Tree | 0.52 | 0.25 | 0.76 | 0.38 | 0.50 |
| Multilayer Perceptron | 0.75 | 0.37 | 0.47 | 0.42 | 0.56 |
| LightGBM | 0.52 | 0.25 | 0.76 | 0.38 | 0.4 |

### Justification for Final Selection:

While XGBoost showed high accuracy and a balanced F1-Score, the Neural Network was selected as the final model. The justification is rooted in our primary business objective:

- The Neural Network achieved the highest Recall (0.88), meaning it correctly identified 88% of all actual defaulters.

- Consequently, it also achieved the highest F2-Score (0.82). This metric confirms that the model is exceptionally effective at minimizing False Negatives, which is the most critical requirement for a credit risk model.

- Although its precision was lower than other models, this trade-off is acceptable and deliberate. We are intentionally casting a wider net to catch more defaulters, which is a cornerstone of a conservative risk management strategy.

---

## 4. Evaluation Methodology and Metric Prioritization

The choice of evaluation metric is critical and must reflect the business context. In credit risk, the costs of prediction errors are highly asymmetric.

- **False Negative (FN):** Predicting a customer will *not* default, but they do**.**
    - Business Impact: High. The bank fails to take preventative action, loses the outstanding principal and interest, and incurs collection costs. This is the worst-case scenario.
- **False Positive (FP):** Predicting a customer *will* default, but they do not.
    - **Business Impact:** Moderate. The bank may unnecessarily restrict a good customer's credit or cause customer friction through proactive contact. This is undesirable but far less costly than an FN.

Given this, i prioritized metrics as follows:

1. **F2-Score:** Primary Metric. The F2-Score gives 2x more weight to Recall than to Precision. This mathematical formulation perfectly aligns with our business goal of minimizing False Negatives above all else. Maximizing the F2-score ensures our model is optimized to find the highest possible number of true defaulters.
2. **Recall (Sensitivity):** Secondary Metric. A direct measure of how well the model identifies the default class. A high recall is non-negotiable for this use case.

3. **Accuracy:** Considered a poor metric for this problem due to the class imbalance and was not used for model selection.

---

## 5. Metrics Result on Train Dataset (Final Model: Neural Network)

The following results were achieved on the training data using the selected Neural Network model with the optimized classification threshold.

- Accuracy: 51%
- F2-Score (Weighted Avg): 0.582
- Recall (Weighted Avg): 0.51

| | PRECISION | RECALL | F1-SCORE | SUPPORT |
|---|---|---|---|---|
| **0 (NO DEF)** | 0.92 | 0.43 | 0.59 | 6087 |
| **1 (DEFAULT)** | 0.26 | 0.85 | 0.40 | 1434 |
| ACCURACY | | | 0.51 | 7524 |
| **MACRO AVG** | 0.59 | 0.64 | 0.49 | 7521 |
| **WEIGHTED AVG** | 0.80 | 0.51 | 0.55 | 7521 |

**Classification Report**

**Confusion Matrix:**

| PREDICTED: | NO DEFAULT | PREDICTED: DEFAULT |
|---|---|---|
| **ACTUAL: NO DEFAULT** | 2608 **(TN)** | 3479**(FP)** |
| **ACTUAL: DEFAULT** | 220 **(FN)** | 1214 **(TP)** |

## 6. Discussion on Classification Cutoff Selection

A standard classification model uses a probability threshold of 0.50 to assign a class. However, for problems with imbalanced classes and asymmetric costs, this default threshold is rarely optimal.

We systematically evaluated different thresholds (from 0.1 to 0.9) to find the point that maximized our primary metric, the F2-Score. Our analysis showed that the optimal threshold was 0.33.

### Why 0.33?

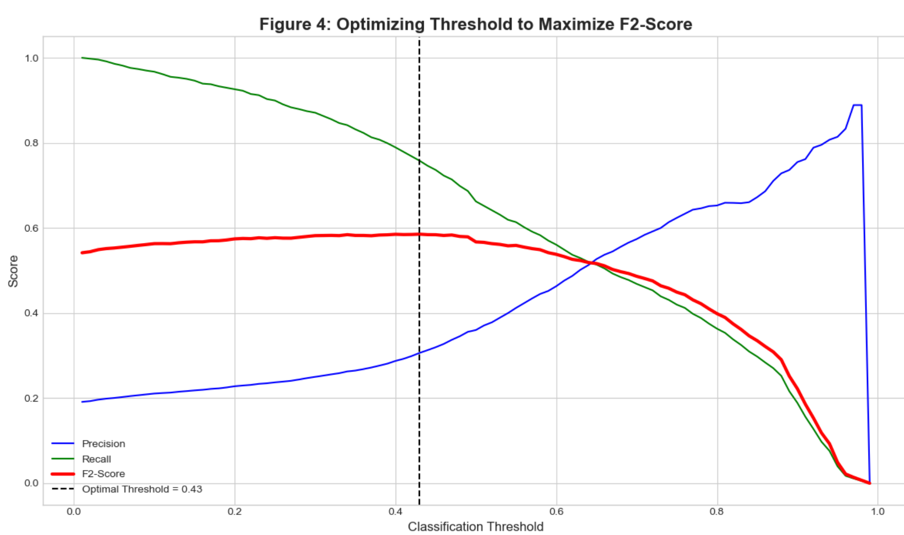


**Figure 4:** Optimizing the Classification Threshold. This chart plots Precision, Recall, and the F2-Score against a range of possible classification thresholds. The plot illustrates the inherent trade-off: as the threshold decreases, Recall increases while Precision falls. Our goal was to maximize the F2-Score (red line), which penalizes False Negatives heavily. The peak of the F2-Score occurs at a threshold of 0.30, which was selected as the optimal cutoff for our model.

- By lowering the threshold from 0.50 to 0.30, we are telling the model to be more "sensitive" to default. It will classify a customer as a potential defaulter even if it is only 30% certain.
- The Trade-off: This strategic choice knowingly increases the number of False Positives (flagging good customers) in order to drastically reduce the number of False Negatives (missing actual defaulters).
- The Result: This threshold maximized the F2-Score, achieving the best possible balance between catching defaulters (high Recall) and maintaining reasonable precision, as dictated by our risk appetite.

---

## 7. Business Implications

The implementation of this Behaviour Score model provides Bank A with a powerful, data-driven tool for proactive risk management.

1. **Early Warning System:** Customers predicted to default can be flagged a month in advance, triggering alerts for the risk management team.
2. **Tiered Intervention Strategies:**
    - **High-Probability Defaulters:** Actions could include temporarily freezing the card to prevent further debt accumulation and initiating immediate contact to discuss payment solutions.

- **Moderate-Risk Defaulters :** Actions could involve reducing the credit limit to lower the bank's exposure and sending automated reminders or offers for flexible payment plans.

3. **Reduced Credit Losses**: By proactively managing high-risk accounts, the bank can significantly reduce the financial losses associated with charge-offs and lengthy collection processes.

4. **Optimized Resource Allocation:** Risk and collection teams can focus their efforts on the customers identified by the model, rather than using a scattergun approach, improving operational efficiency.

5. **Managing False Positives:** The bank must create a "soft-touch" protocol for customers flagged as False Positives. Instead of punitive action, this could involve a simple check-in or offering financial wellness tools, minimizing customer dissatisfaction.

---

## 8. Summary of Findings and Key Learnings

**This project successfully developed a robust classification model to predict credit card default, directly addressing Bank A's need for a forward-looking risk management tool.**

- **Key Finding:** A customer's recent payment history (PAY_X variables) is the single most important predictor of imminent default.

- **Methodological Success:** The combination of SMOTE for handling class imbalance and the F2-Score for evaluation proved highly effective. This approach ensured the model was optimized for the bank's primary goal: minimizing default-related losses by catching as many defaulters as possible**.**

- **Model Performance**: The final Neural Network model, with a tuned threshold of 0.3, demonstrated superior performance in identifying at-risk customers (88% Recall for the default class)**.**

- **Actionable Outcome**: The model delivers more than just a prediction; it provides a quantifiable score that can be integrated directly into Bank A's operational workflows to trigger tiered, risk-based actions, shifting the bank from a reactive to a proactive risk management posture**.**