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

1. Introduction and Objective

The goal of this project is to build a predictive classification model to identify customers who are likely to default on their credit card payment in the upcoming month. Leveraging anonymized behavioural data for over 30,000 credit card holders, the objective is not only accurate prediction but also developing a financially interpretable solution that supports credit risk mitigation for Bank A. Our focus lies on maximizing the F2 score, prioritizing recall due to the critical importance of identifying true defaulters.

2. Exploratory Data Analysis (EDA)

2.1 Data Overview:

- Train dataset: ~25,000 records with a binary target variable (next_month_default)
- Validation dataset: ~5,000 records with the same features but no target
- Features include demographics, payment history (PAY_0 to PAY_6), bill/payment amounts, and derived features like AVG_Bill_amt, PAY_TO_BILL_ratio

2.2 Class Distribution:

- Dataset is imbalanced: ~78% non-defaulters and ~22% defaulters
- Addressed using SMOTE and class-weighted learning

2.3 Key Observations:

- High PAY_0 (payment delay) increases default likelihood
- Higher AVG_Bill_amt often correlates with lower default risk
- Low PAY_TO_BILL_ratio signals poor repayment behavior and higher risk

2.4 Hypothesis Testing Insights

A point-biserial correlation analysis was performed to assess the linear relationship between numerical features and the binary target (next_month_default). Key results:

- pay_amt3 to pay_amt6, and AVG_Bill_amt had statistically significant positive correlations ($p < 0.05$), confirming that higher payments reduce default risk.
- UTILIZATION_RATIO and Bill_amt5 had insignificant correlations ($p > 0.05$), indicating weak individual predictive power for linear models.

While these features may not be useful in logistic regression, they were retained in tree-based models which can capture non-linear interactions.

2.5 Visual Insights:

- Correlation heatmap revealed multicollinearity among **Bill_amtX** and **pay_amtX**

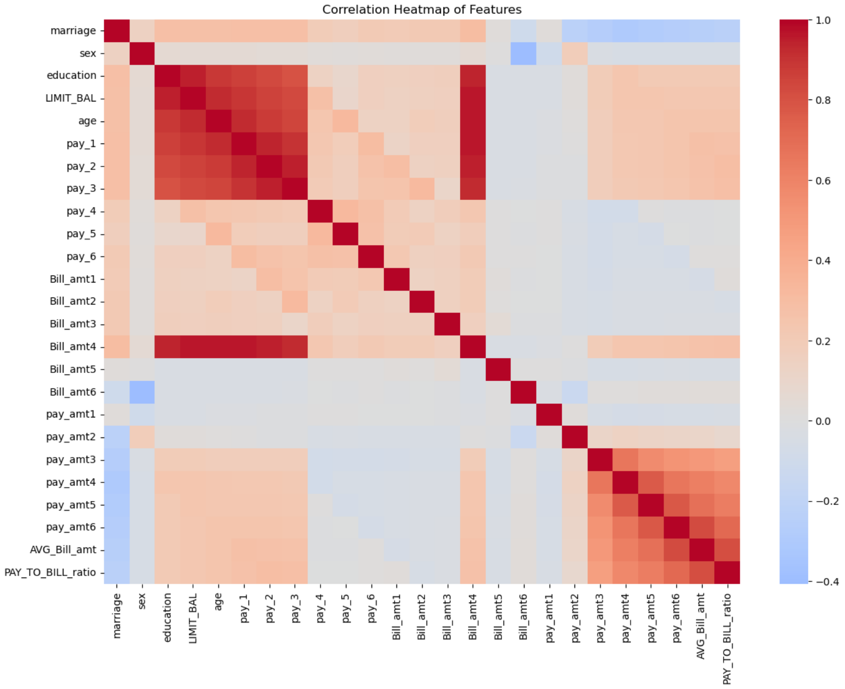

Fig-1: *Correlation heatmap of features*

- Boxplots demonstrated that higher delinquency and utilization relate to default risk

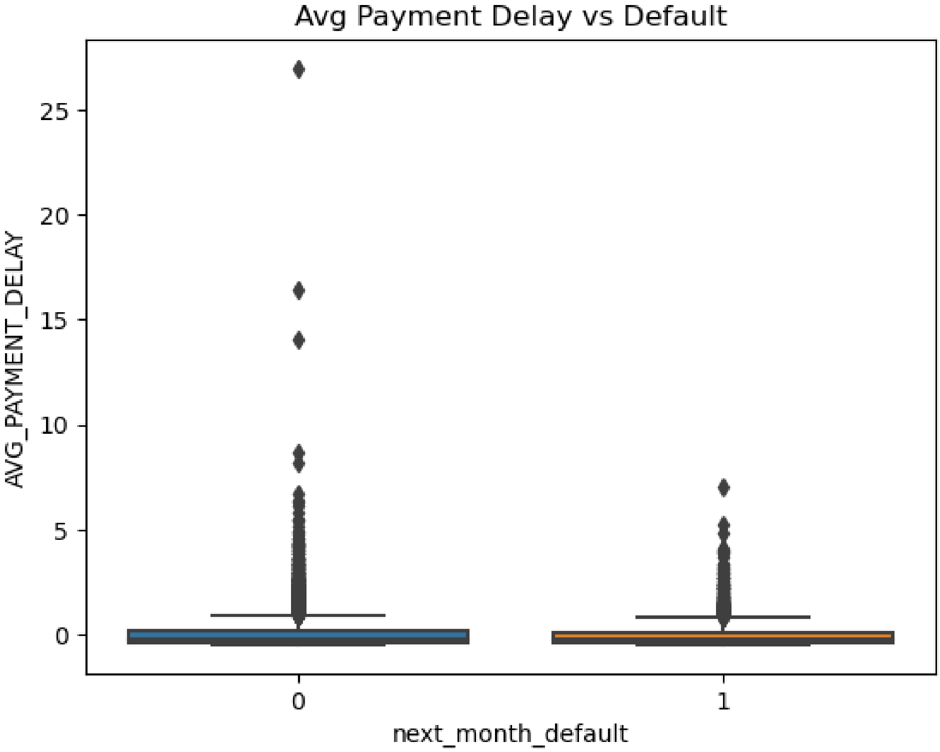

Fig-2: *Avg Payment Delay vs Default*

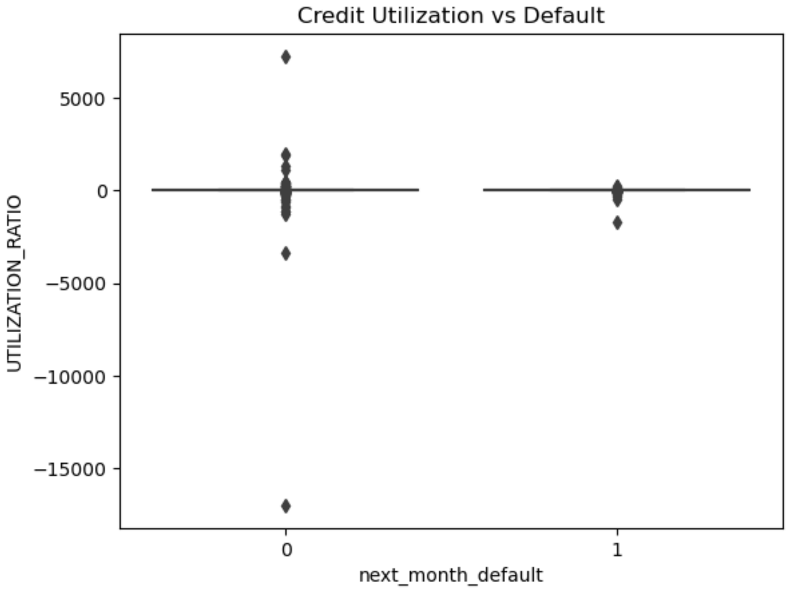

Fig-3: *Credit Utilization vs Default*

3. Financial Behaviour Insights

- PAY_x: Delayed months (PAY_x > 0) are strong default indicators
- PAY_TO_BILL_ratio: Low values imply underpayment — common among defaulters
- UTILIZATION_RATIO: Higher utilization usually flags customers nearing credit limit
- Demographics (education, marital status, age): Minor trends, not primary drivers

4. Feature Engineering

- **Delinquency metrics**: MAX_PAYMENT_DELAY, AVG_PAYMENT_DELAY, count of delayed months
- **Repayment ratios**: Monthly and average pay_amt / bill_amt, total repayment vs bill amounts
- **Aggregation**: AVG_Bill_amt, TOTAL_PAY_AMT, TOTAL_BILL_AMT
- **Encoding**: One-hot encoding for categorical features like education and marital status

5. Model Building & Selection

5.1 Algorithms Tested:

- Logistic Regression
- Decision Tree Classifier
- Random Forest
- XGBoost
- LightGBM

5.2 Class Imbalance Handling:

- SMOTE for oversampling defaulters
- Class weights for logistic and tree models

5.3 Cross-Validation Strategy:

- Stratified 5-Fold CV for stable performance estimation

5.4 Final Model Comparison and Selection:

Model	Accuracy	Precision	Recall	F1 Score	F2 Score	ROC-AUC
Random Forest	0.8897	0.9019	0.8745	0.8880	**0.8798**	0.9564
XGBoost	0.8886	0.9173	0.8542	0.8846	0.8661	0.9465
LightGBM	0.8875	**0.9224**	0.8462	0.8827	0.8604	0.9436
Decision Tree	0.8191	0.8166	0.8229	0.8198	0.8216	0.8191
Logistic Regression	0.6862	0.6948	0.6639	0.6790	0.6699	0.7551

Conclusion:
Random Forest outperformed all others with the best F2 score. It captures defaulters effectively, balancing precision and recall. Although LightGBM had better precision and speed, the slight recall trade-off made Random Forest more favourable under F2-focused evaluation.

6. Evaluation Metrics and Threshold Tuning

6.1 Metrics Used: Accuracy, Precision, Recall, F1 Score, **F2 Score**, ROC-AUC

6.2 Why F2 Score?

- Banks care more about **not missing defaulters** (Recall)
- F2 Score emphasizes Recall more than F1/Accuracy
- False negatives pose higher financial risk than false positives

6.3 Threshold Optimization and Trade-Off:

- LightGBM's final training reported the best F2 Score of **0.8833** at a custom threshold of **0.3435**
- Classification report:
  - Accuracy: 0.8568
  - Precision: 0.8298
  - Recall: 0.9023
  - ROC-AUC: 0.9372
  - This balance ensured very few defaulters were missed (high recall) while precision and accuracy remained acceptable.

Trade-off Insight:

- Precision dropped slightly, Accuracy decreased marginally
- Recall increased significantly, improving F2
- Visualized below:

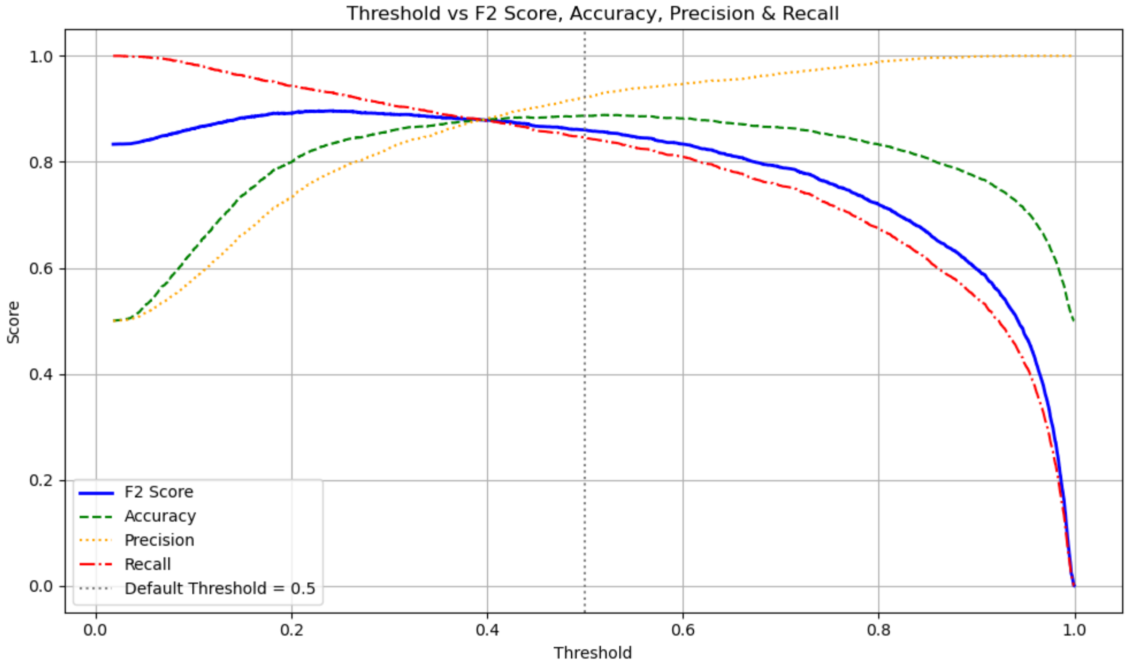

Fig-4: *Threshold vs F2 Score, Accuracy, Precision & Recall*

Threshold Optimization Summary

To align with the bank's goal of minimizing credit risk by catching potential defaulters, we analyzed the trade-offs across a range of thresholds. The following plot illustrates how F2 Score, Accuracy, Precision, and Recall evolve:

1. **F2 Score (Blue Line)**: Peaks around threshold 0.35, since it gives more weight to Recall. Beyond this, performance drops as Recall or Precision weakens.
2. **Accuracy (Green Dashed Line)**: Increases with the threshold up to ~0.5, then stabilizes. Higher thresholds reduce false positives but risk more false negatives.
3. **Precision (Orange Dotted Line)**: Rises steadily — at high thresholds, the model is more confident, improving Precision.
4. **Recall (Red Dash-Dot Line)**: Starts high and drops. At low thresholds, more defaulters are caught, but at the expense of more false positives.

Business Interpretation The optimal F2 score is achieved at a threshold of ~0.3435. This value maximizes Recall while retaining acceptable Precision. Since the bank's priority is to reduce the number of undetected defaulters (false negatives), this trade-off is ideal. The bank can tolerate some false positives, but missing true defaulters poses a greater financial risk.

This reasoning guided our threshold selection for LightGBM and final model predictions.

7. Results on Training Data

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**F2 Score**
Logistic Reg.	0.6862	0.6948	0.6639	0.6790	0.6699
Decision Tree	0.8191	0.8166	0.8229	0.8198	0.8216
Random Forest	0.8897	0.9019	0.8745	0.8880	**0.8798**
XGBoost	0.8886	0.9173	0.8542	0.8846	0.8661
LightGBM	0.8875	0.9224	0.8462	0.8827	0.8604

8. Predictions on Validation Set

- Validation set (~5,000 records) was transformed and scored using LightGBM (threshold=0.3435)
- Output format: Customer_ID, Predicted_Default
- Breakdown:
  - Predicted Defaulters: **1177**
  - Non-Defaulters: **3839**

9. Business Implications

- **False Negatives (missed defaulters):** High financial risk — leads to uncollected dues
- **False Positives:** May restrict good customers, but preferable to missed defaulters

- **Usage:**
  - Trigger credit review / early intervention
  - Adjust interest rates based on risk
  - Prioritize collection resources

10. Summary & Key Learnings

- F2 Score fits credit risk goals better than Accuracy or F1
- Behavioral variables like PAY_X, repayment ratio, and utilization drive risk
- Threshold tuning ensures fewer defaulters are missed
- Random Forest was the strongest performer overall; LightGBM was fast and precise