Building a Behaviour Score: Credit Card Default Prediction Using Classification and Risk-Based Techniques

(Finance Club Open Project Summer 2025)

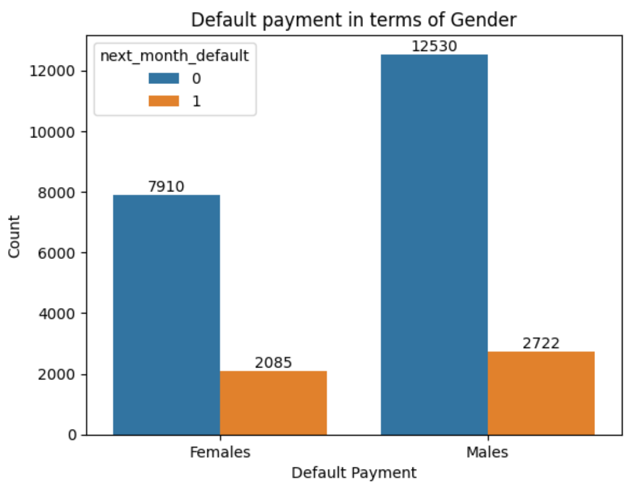
Yash Arora ,22112117

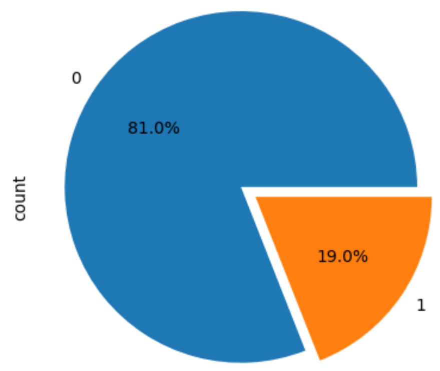
1. Approach Overview and Modeling Strategy

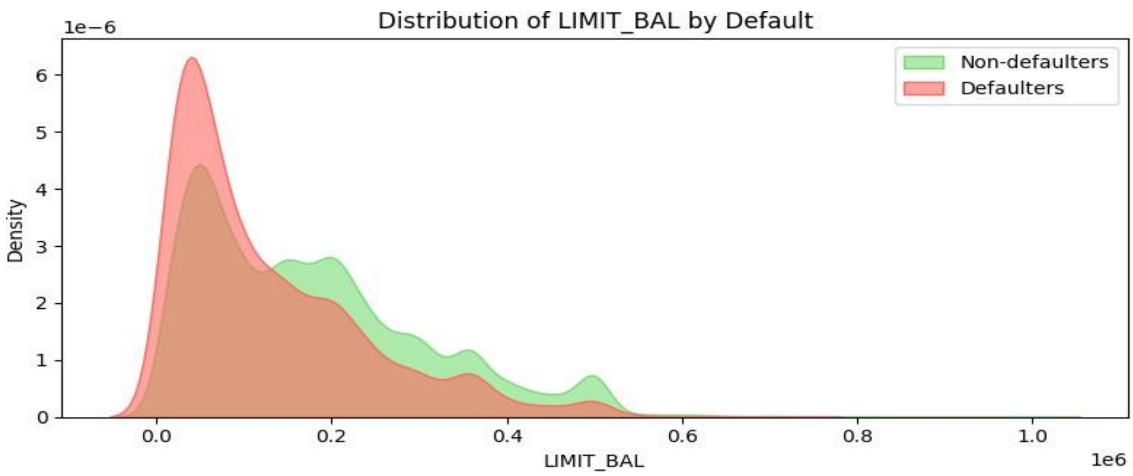
- Exploratory Data Analysis (EDA) for understanding feature importance
- Feature
  engineering for extracting meaningful parameters from raw data.
- Dataset balancing with SMOTE to overcome class imbalance.
- Dimensionality reduction with PCA for model optimization.
- Logistic Regression classification modeling and comparison with other models.
- Recall, F1, and F2 score evaluation with recall as the priority due to business risk.

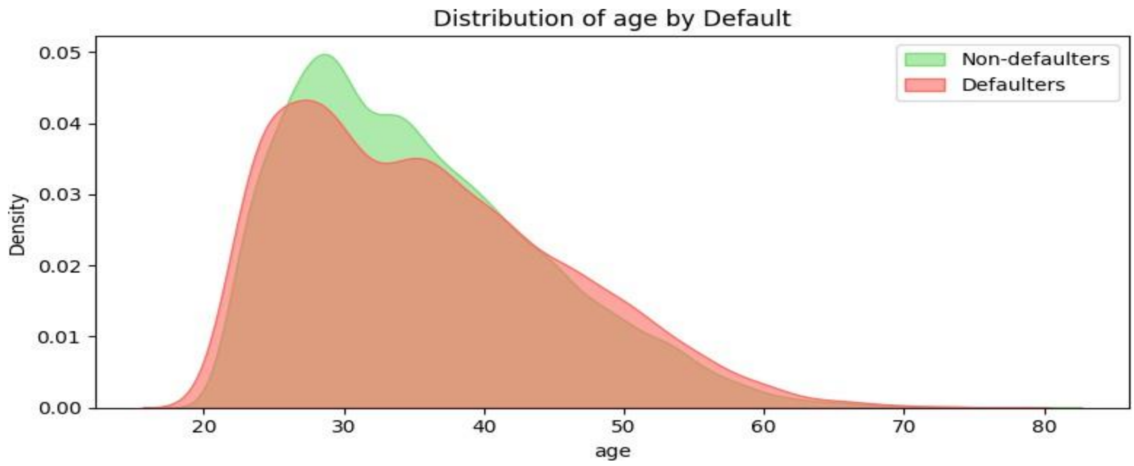
2. EDA Findings and Visualizations

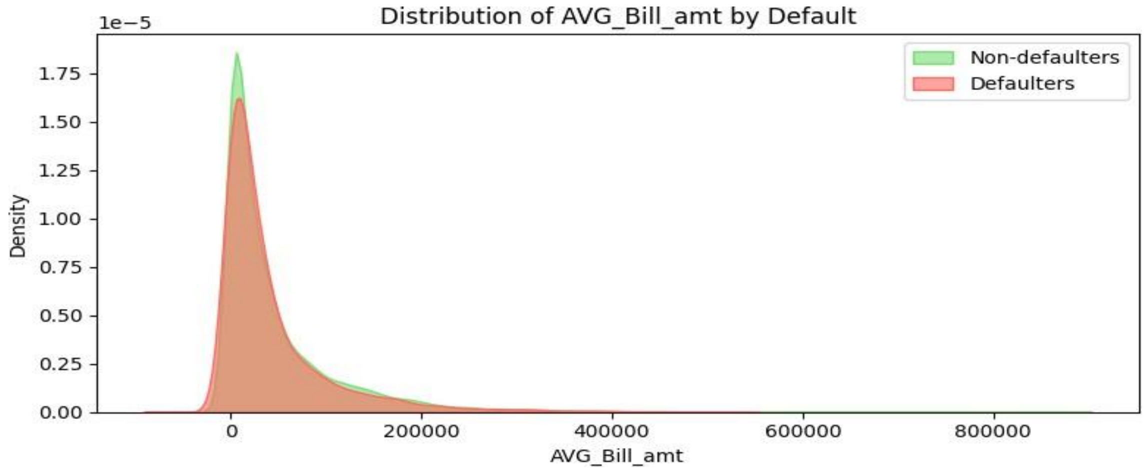
Distributional Patterns: The majority of features such as bill amounts, payments, and limits are right-skewed. Payment history shows trends in behaviour associated with the likelihood of default.



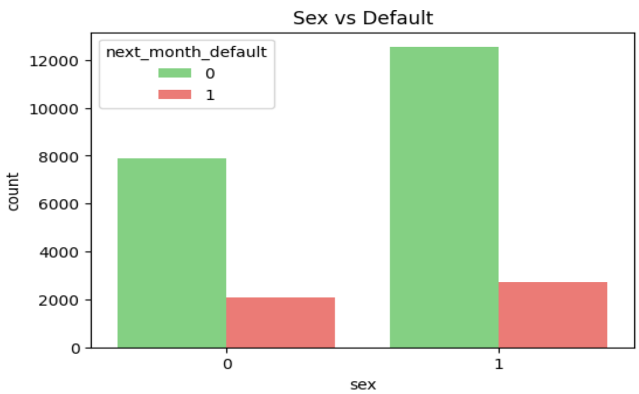


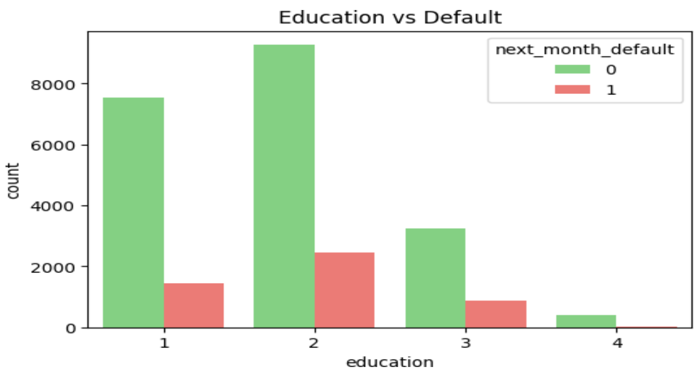


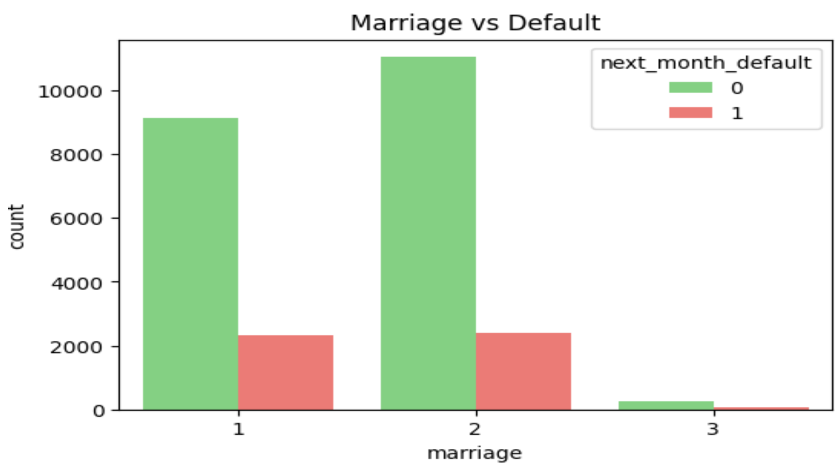




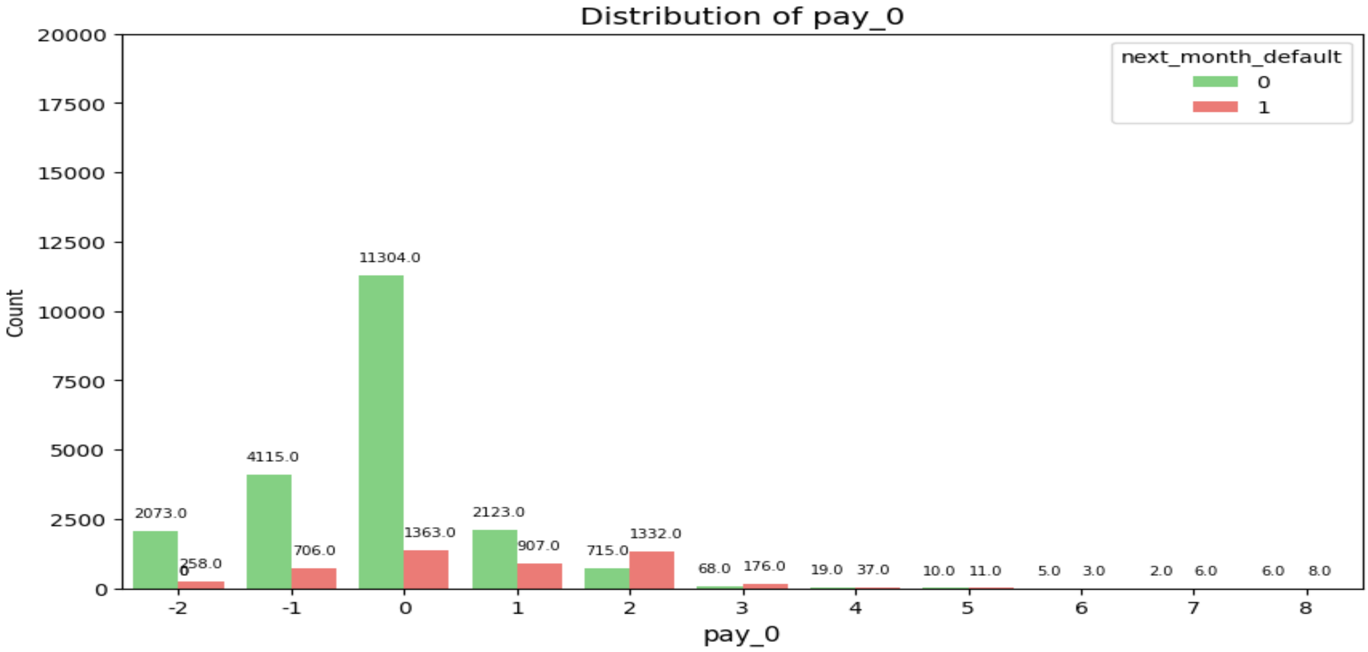
For Categorial features

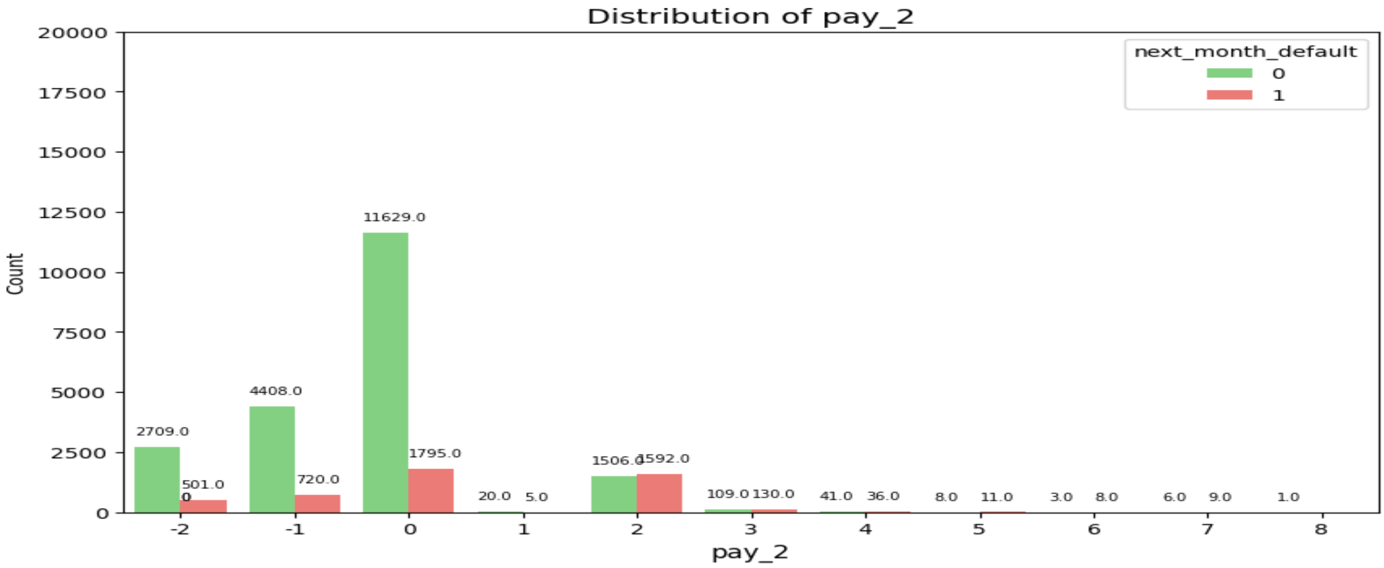


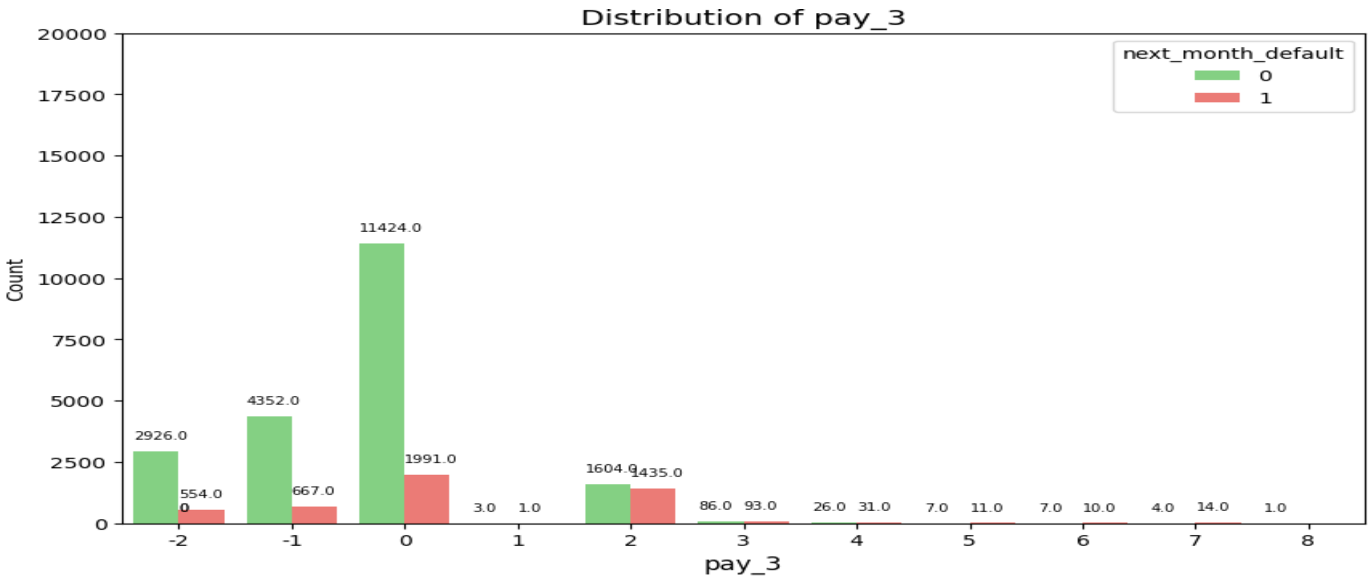


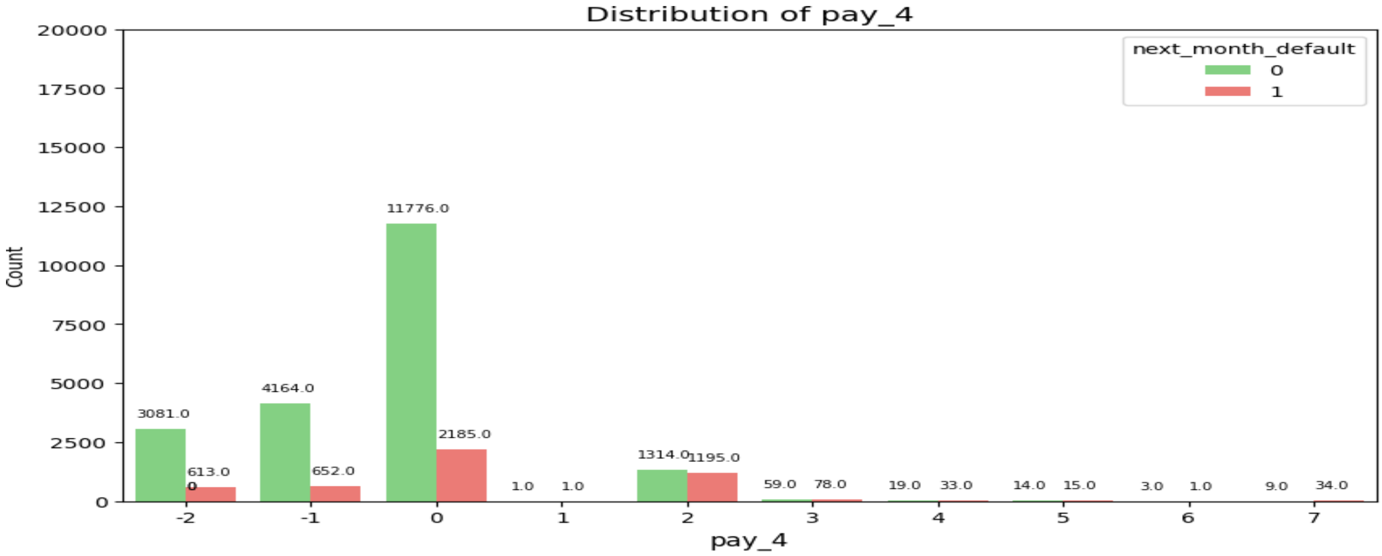


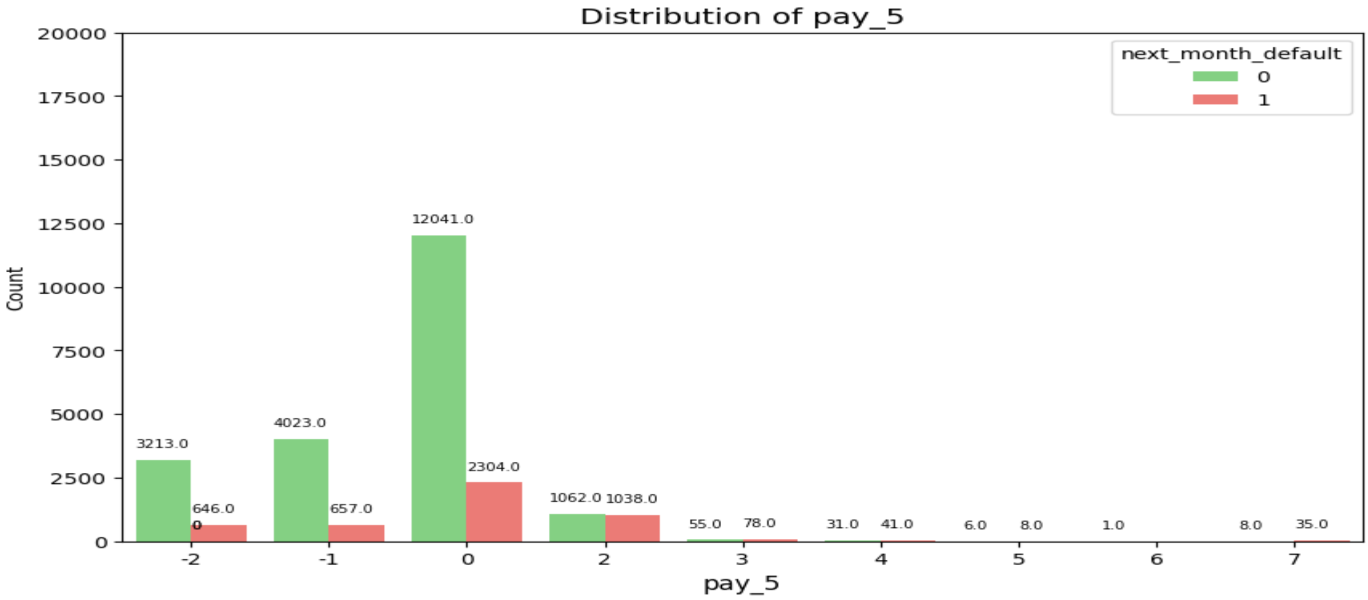
PAY_X status vs default

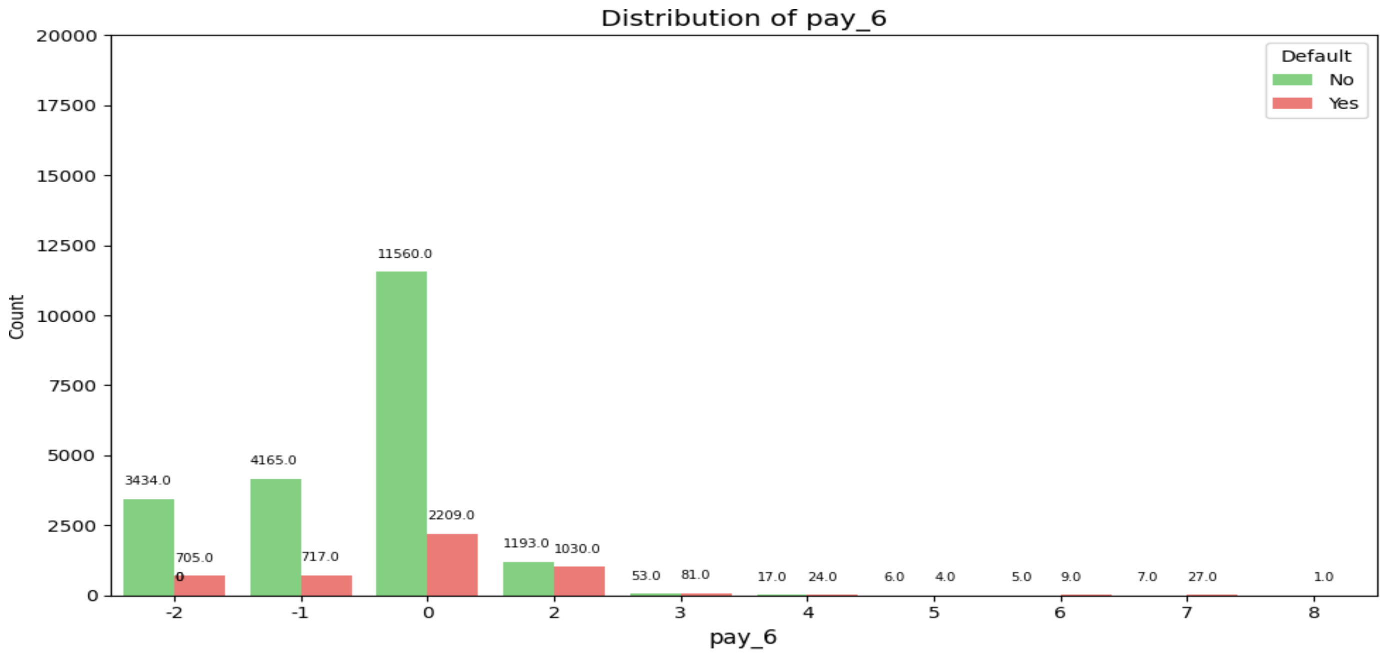




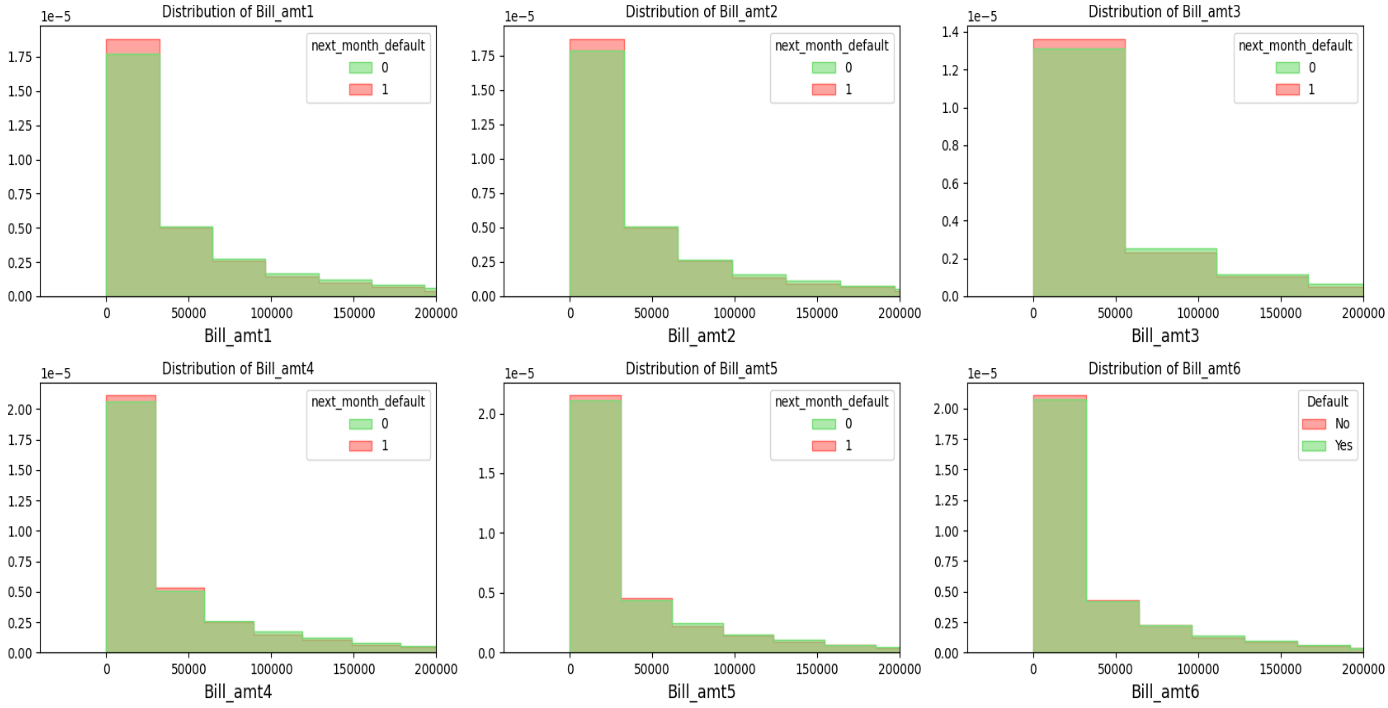



Distribution of pay_4
next_month_default
0
1
Count
pay_4
3081.0
613.0
4164.0
652.0
11776.0
2185.0
1.0
1.0
1314.0
1195.0
59.0
78.0
19.0
33.0
14.0
15.0
3.0
1.0
9.0
34.0

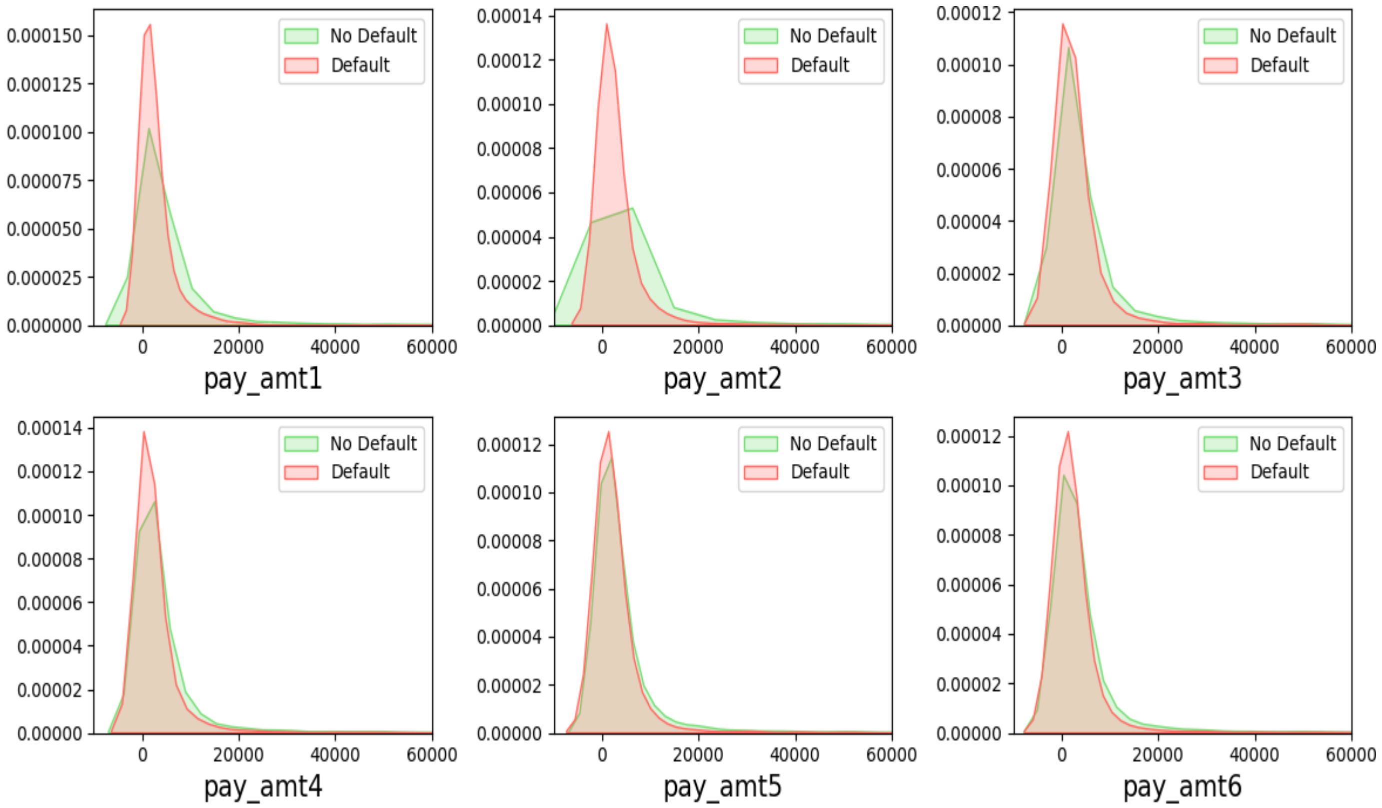

Distribution of pay_5
next_month_default
0
1
Count
pay_5
3213.0
646.0
4023.0
657.0
12041.0
2304.0
1062.0
1038.0
55.0
78.0
31.0
41.0
6.0
8.0
1.0
8.0
35.0


Distribution of pay_6
Default
No
Yes
Count
pay_6
3434.0
705.0
4165.0
717.0
11560.0
2209.0
1193.0
1030.0
53.0
81.0
17.0
24.0
6.0
4.0
5.0
9.0
7.0
27.0
1.0

Distribution of bill_amt_X :

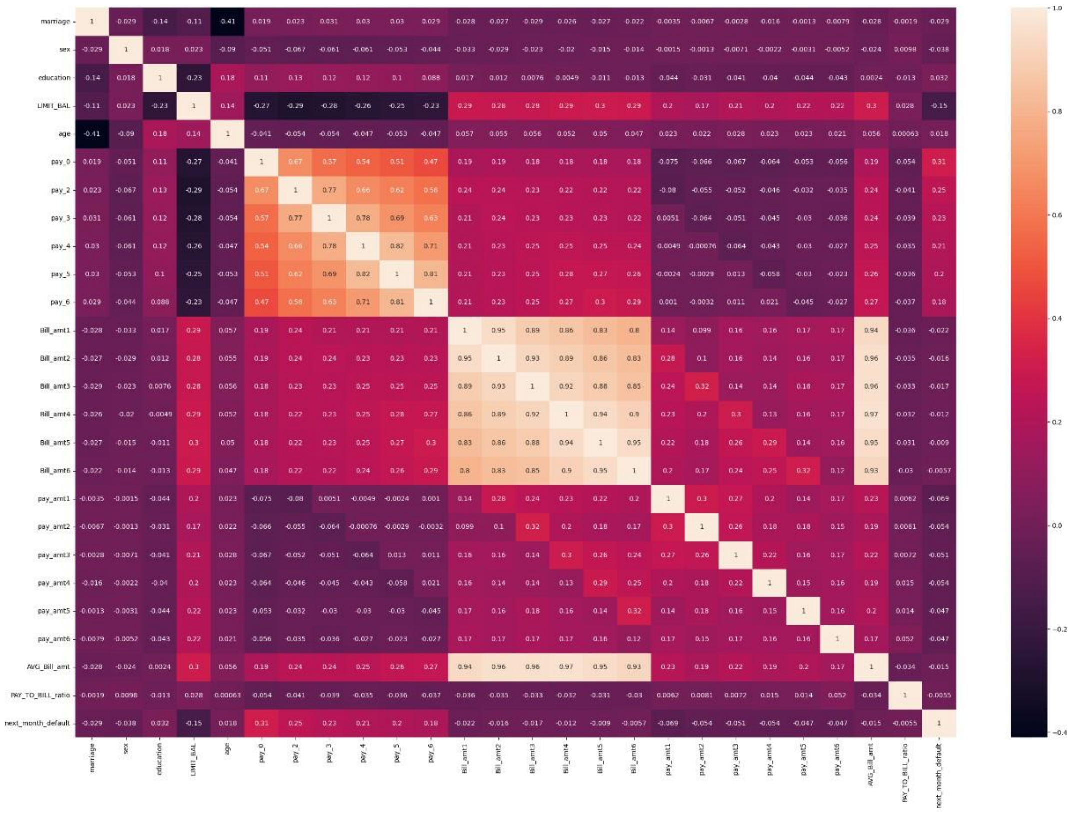


PAY_AMT_X DISTRIBUTION:



Correlation Heatmap:

Found that Strong correlation between `pay_X` variables, showing redundancy which made it appropriate to use PCA.



The correlation heatmap provides an overview of how features in the dataset relate to each other as well as to the target variable next_month_default.

Important Findings: There is a steady pattern in the repayment behavior of customers across months, as evidenced by the moderately positive correlations between the repayment history variables (pay_0 to pay_6).

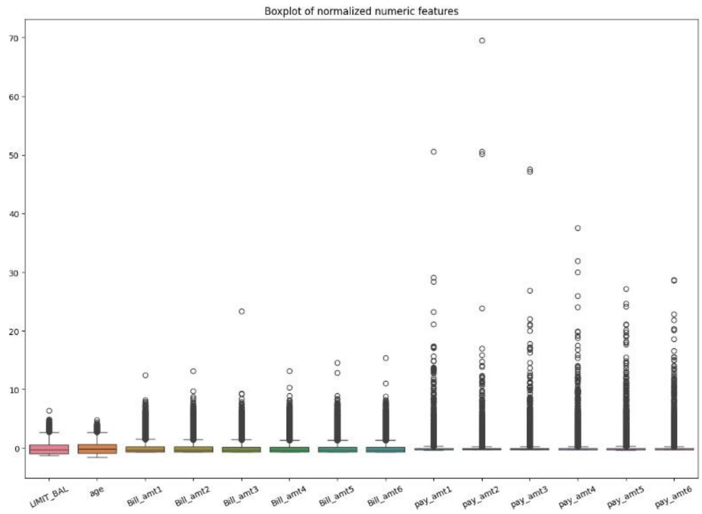
The high inter-correlation (above 0.85) between the bill amounts (bill_amt1 to bill_amt6) also suggests that they represent a similar financial signal, which may enable dimensionality reduction or aggregation.

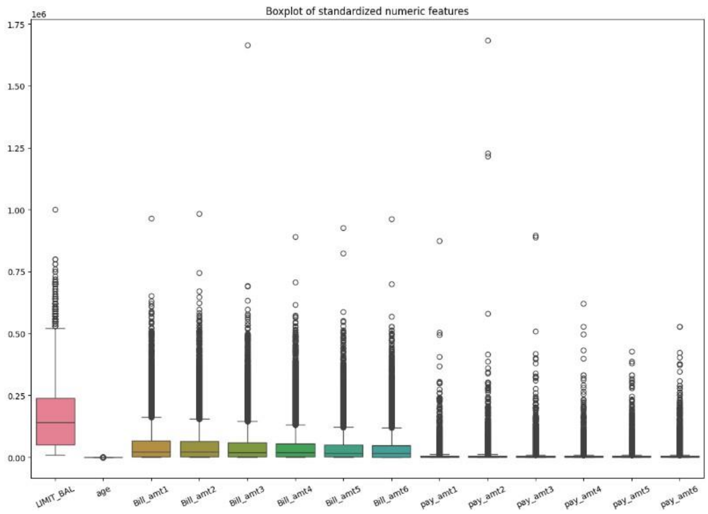
Payment amounts (pay_amt1 to pay_amt6) also show a positive correlation, albeit to a lesser degree, suggesting that customers' repayment patterns are consistent over time.

As is typical of unbalanced or noisy financial data, there is not much of a correlation between next_month_default and the majority of individual features. But:

The most recent repayment status, or pay_0, has the strongest positive correlation with default, indicating that it is a significant predictor.

Age and LIMIT_BAL have marginally negative associations with default, suggesting that clients with higher credit limits and older age are slightly less likely to default.





Key Observations:

* Clients with continuously high overdue payment amounts had a greater default rate.
* High credit usage and past-due payments were both strong predictors of default.
* The majority of defaulters had medium to low credit limits, defying the belief that only high credit customers are higher risk.

Data splitting: Separated the features and target variables in x and y class.

Split the Training dataset into train and validate keeping ratio of y class same for better results.

Preprocessing:

- Plotting 'LIMIT_BAL', 'age', 'AVG_Bill_amt' ,histogram , they are not normally distributed.
- Then used standard scaler and minMaxscaler to scale those features

DATA BALANCING:

Used SMOTE(Synthetic Minority Oversampling Technique) to oversample minority class (defaulters) to increase recall.

Rationale: Generates synthetic examples rather than simple duplication

Result: Balanced dataset for improved model training

```
Training set class proportion after SMOTE Oversampling
```

1 to 2 of 2 entries Filter

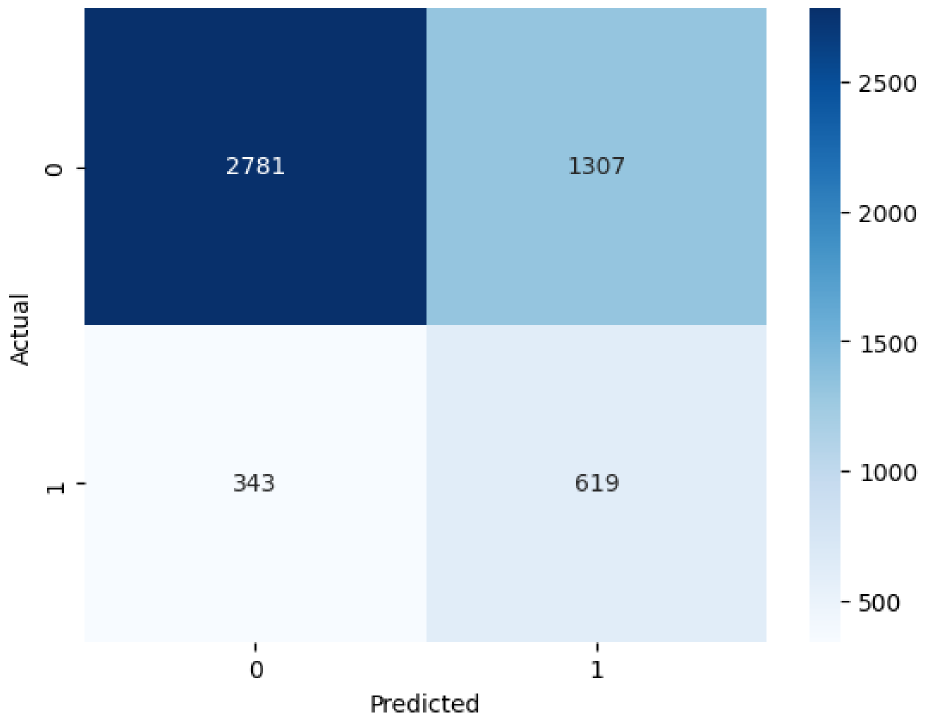
index	Number	Percentage
Non-defaulters	16352	50.00%
Defaulters	16352	50.00%

Show 25 per page

Model Development s Comparison:

1. Logistic Regression

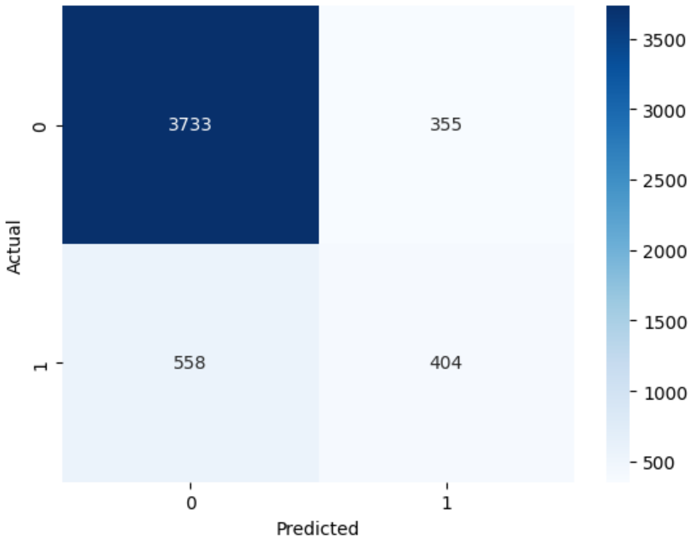
- A linear model that estimates probabilities and is easy to interpret.
- High precision (0.89) for non-defaulters; low (0.32) for defaulters.
- Defaulter recall: 0.64; F1-score: 0.43.
- F2-score: 0.538 — favors recall, thus decent for default detection.



○

2. Random Forest

- Ensemble method combining multiple decision trees with bagging.
- Reduces overfitting and improves accuracy by averaging predictions.
- Defaulter precision: 0.530; recall: 0.42.
- Overall accuracy: 0.820; stable across classes.
- F2-score: 0.531

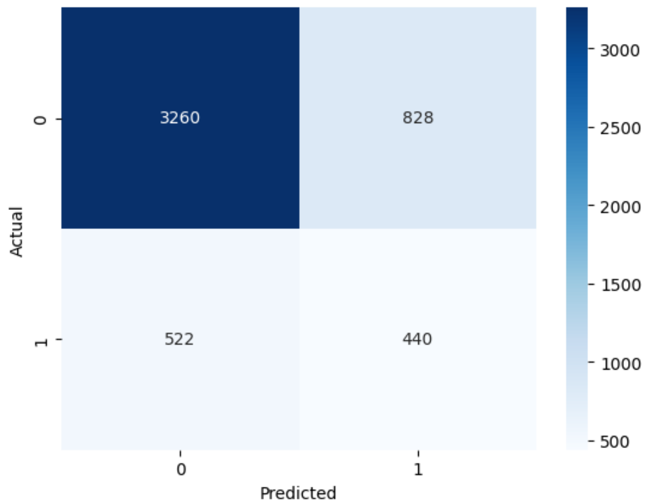


-

---

3. Decision Tree

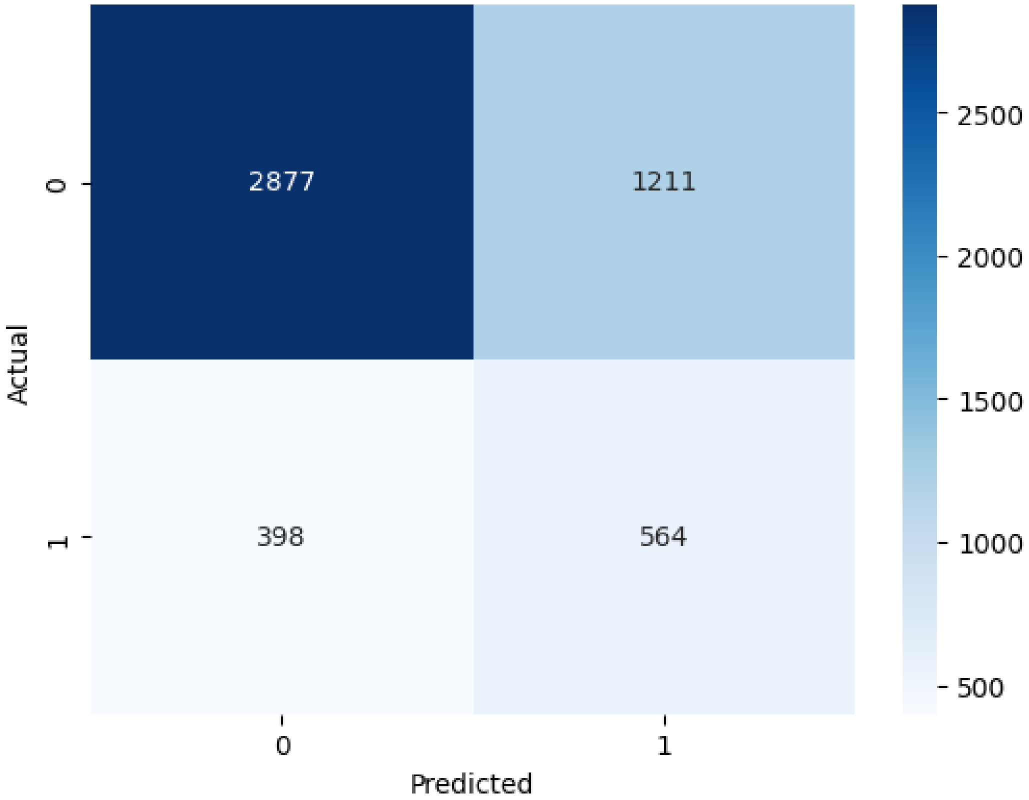
- Tree-based model that splits data based on feature thresholds.
- Prone to overfitting, but very interpretable.
- Defaulter recall: 0.457, precision: 0.347.
- F2-score: 0.546



---

4. K-Nearest Neighbors (KNN)

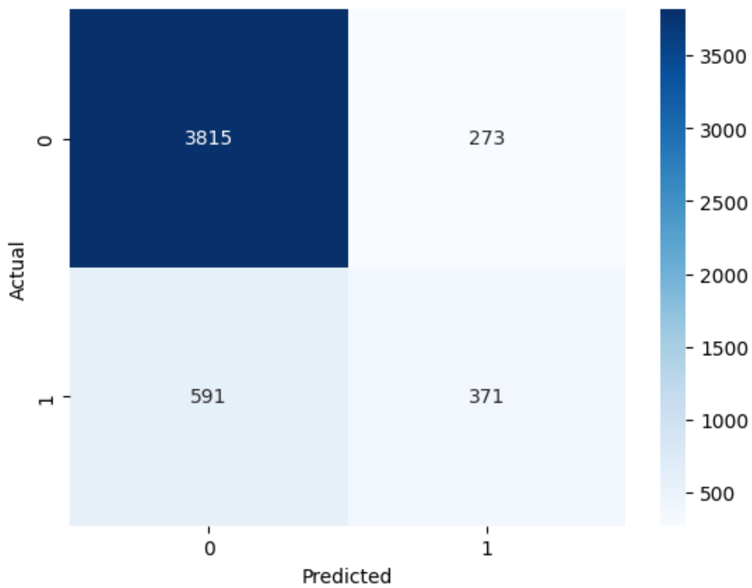
- Instance-based, non-parametric model; classifies based on majority vote of nearest neighbors.
- No training phase; sensitive to scale and irrelevant features.
- Lowest F1/F2 performance for defaulters.
- Recall (0.59) is okay, but poor precision (0.315).
- F2-score: 0.520



---

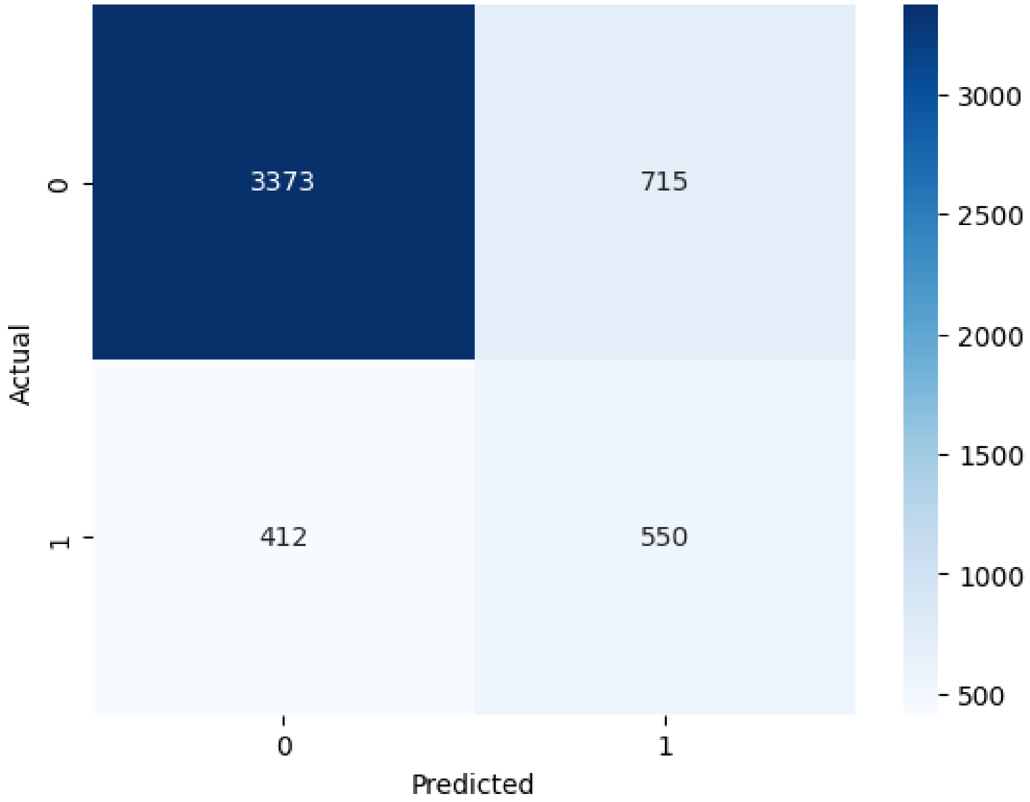
5. XGBoost

- Gradient boosting framework; builds trees sequentially to reduce residual errors.
- Handles missing values and overfitting well; very popular for tabular problems.
- Best recall (0.596) and strong precision (0.420) for defaulters.
- Highest F2-score: 0.550



1. AdaBoost

- Gradient boosting framework; builds trees sequentially to reduce residual errors.
- Handles missing values and overfitting well; very popular for tabular problems.
- Best recall (0.596) and strong precision (0.420) for defaulters.
- Highest F2-score: 0.550



Metrics Used:

Accuracy: General performance measure but can be deceptive on unbalanced data.
Recall: Given higher priority because losing a defaulter is more expensive than a false alarm.
F1 Score: Balanced precision and recall.
F2 Score: Similar to F1 but assigns more importance to recall — essential for financial risk.

Model Comparison and Justification for Final Selection:

BEST MODEL *__: Adaboost__*

Selected on the basis of **F1** score. Given the business context — where both **precision** and **recall** for defaulters are critical — we chose **F1-score for the minority class (defaulters)** as the **primary metric** to select the best model. F1-score balances the trade-off between **false positives** and **false negatives**, making it more appropriate than accuracy in imbalanced classification problems like credit default.

The **Random forest model** achieved the **highest F1-score (0.506)** for defaulters (class 1), making it the most balanced and reliable model for identifying likely defaults while avoiding too many false alarms.

Parameter Tuning:

To further optimize performance, hyperparameter tuning was applied:

Random Forest:

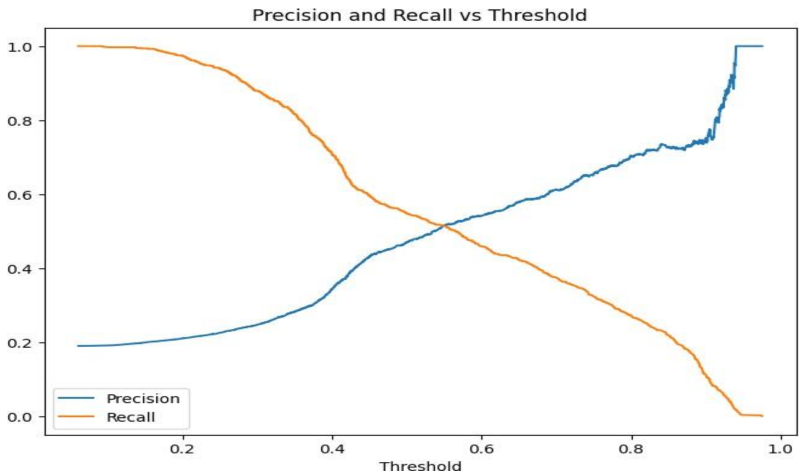
- Tuned n_estimators, max_depth, min_samples_split, and min_samples_leaf.

- **XGBoost**:
  - Tuned learning_rate, max_depth, n_estimators, and scale_pos_weight to improve F1-score under class imbalance.

Optimization was done using GridSearchCV improving the F1 performance by controlling overfitting and increasing model generalization.

Threshold Selection:

- While the default threshold of 0.5 was initially used, **threshold tuning was explored** by analyzing precision-recall trade-offs.



Best threshold: 0.5

Best F1-score at threshold: 0.5G

Business Implications

- **Improved Risk Management**: Early identification of potential defaulters enables proactive intervention and reduces financial losses.
- **Actionable Insights:** Key drivers like payment history and credit utilization guide smarter credit decisions.
- **Targeted Strategies:** High-risk customers can be offered alternate repayment plans or closer monitoring.
- **Optimized Lending:** F1-score-based model ensures a balance between detecting defaulters and retaining creditworthy clients.

Summary

- This project focused on predicting credit card defaults using classification models.
- After performing EDA and feature engineering, multiple models were evaluated.
- Random Forest emerged as the best model based on F1 score, balancing precision and recall effectively.
- The classification threshold was fine-tuned to maximize F1 performance.
- The final model offers strong predictive power and can assist financial institutions in reducing default risk while maintaining lending efficiency.