Credit Default Prediction Using Machine Learning

Project Report By:
Swaroop Itkikar
23118077

Introduction

This project aims to predict credit card default risk using various machine learning algorithms. The workflow involves data extraction, feature engineering, and the application of advanced classification models to address class imbalance and improve predictive accuracy.

Libraries and Tools

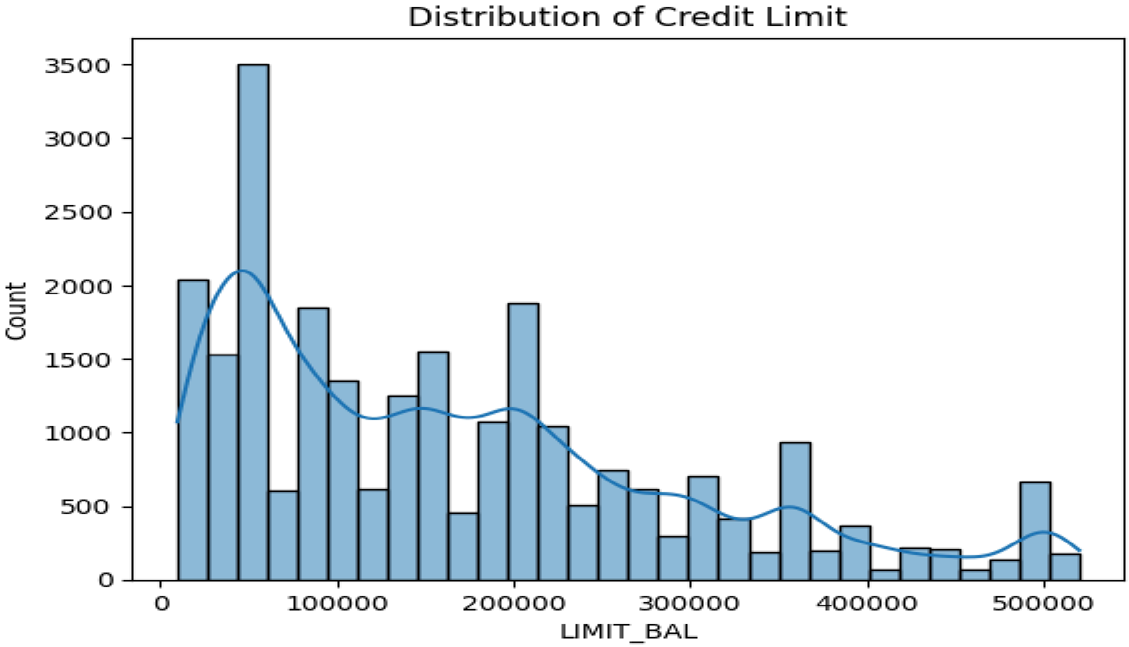
The following libraries and tools were used:

- pandas, numpy: Data manipulation and analysis
- matplotlib, seaborn: Data visualization
- scikit-learn: Machine learning algorithms and utilities
- imbalanced-learn (imblearn): Handling class imbalance (SMOTE)
- xgboost, lightgbm, catboost: Gradient boosting frameworks
- Other utilities: joblib, scipy, etc.

The dataset includes features such as marriage, sex, education, credit limit, age, payment history, bill amounts, and payment amounts. Additional engineered features include average bill amount, payment-to-bill ratio, repayment ratio, number of late months, maximum delay, average delay, repayment consistency, utilization ratio, delinquency streak, and principal component analysis (PCA) features1.

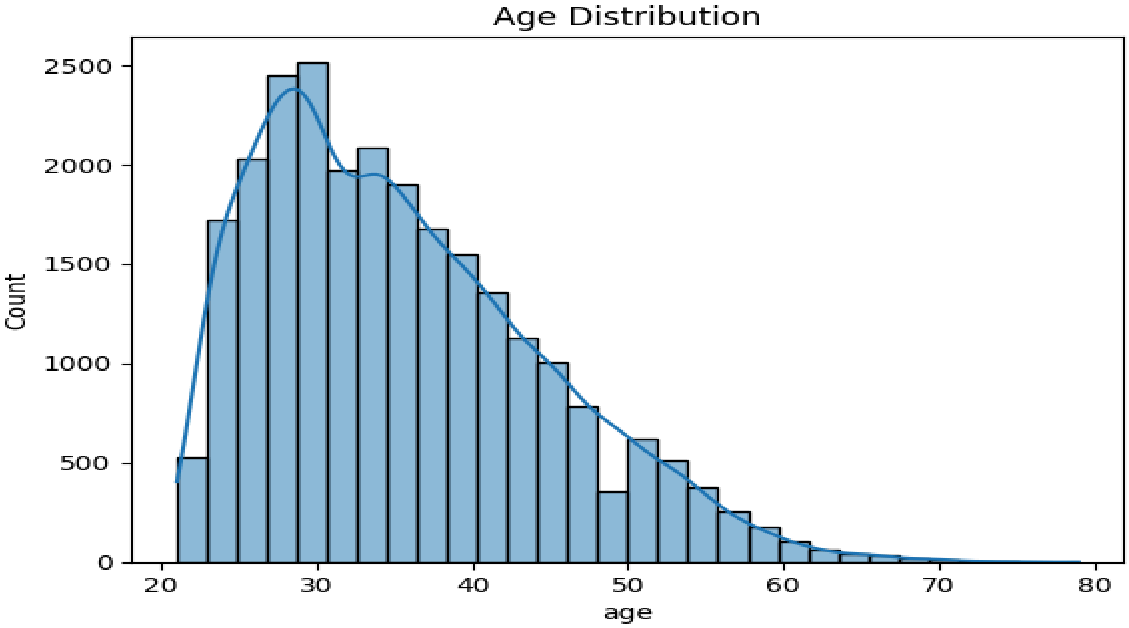
Data Exploration

a) Distribution of Credit Limit:



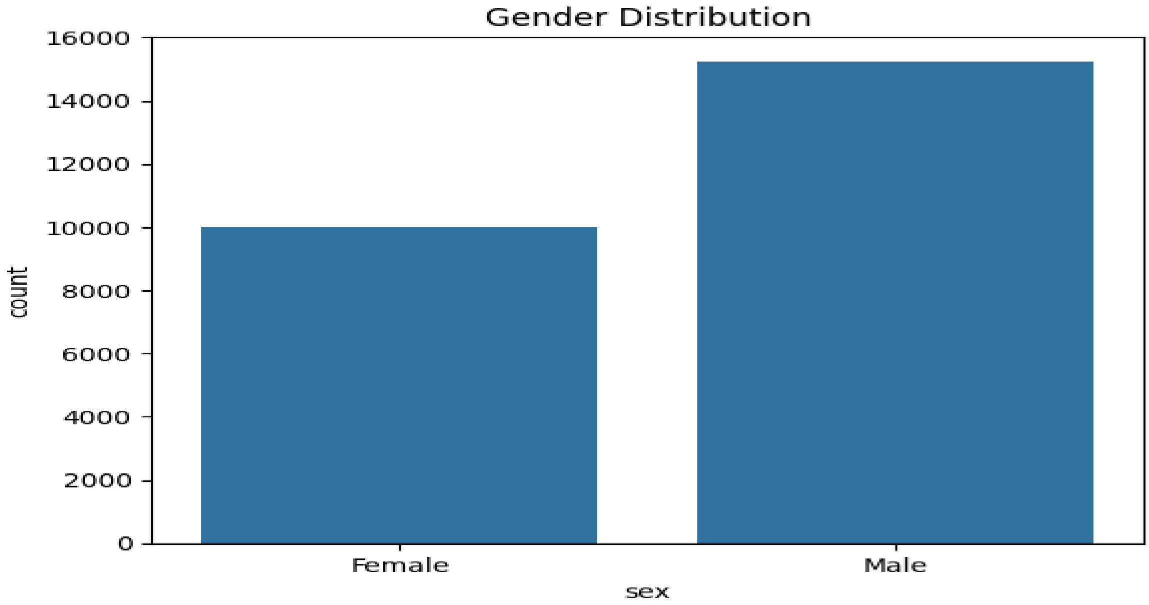
Right-skewed distribution. Most customer have credits limits between Rs 50,000 to Rs 300,000. Outliers exist up to Rs 1,000,000- could explore limiting or log scaling

b) Age Distribution:



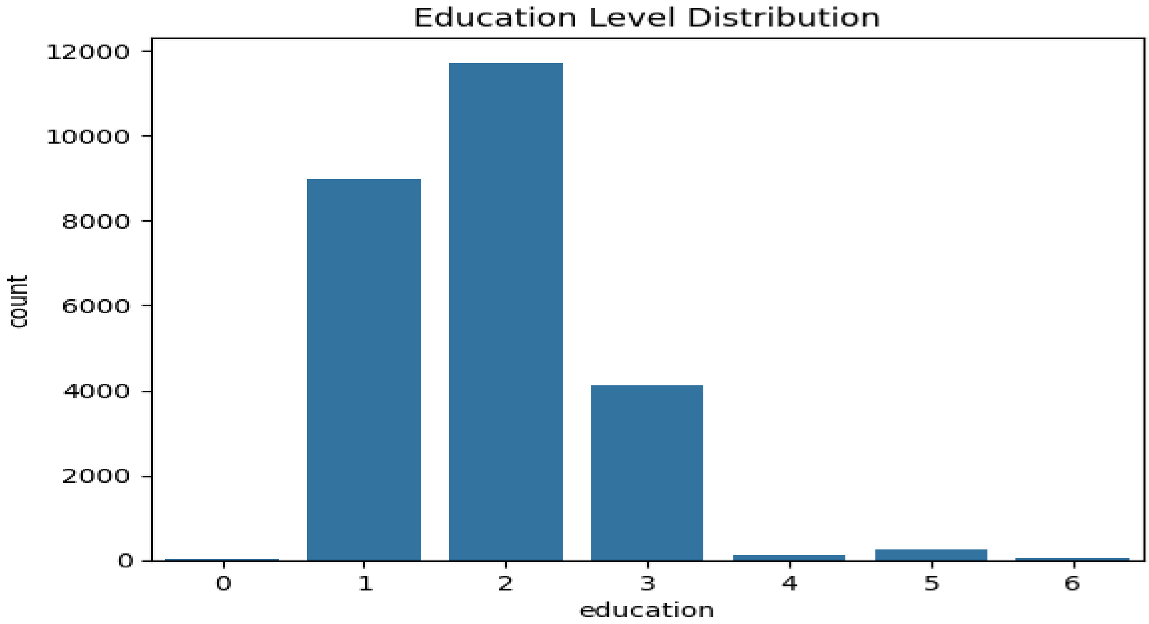
Majority customers are in the age group of 25–45 years.Few very old customers; may be less relevant to default prediction.

c) Gender Distribution:



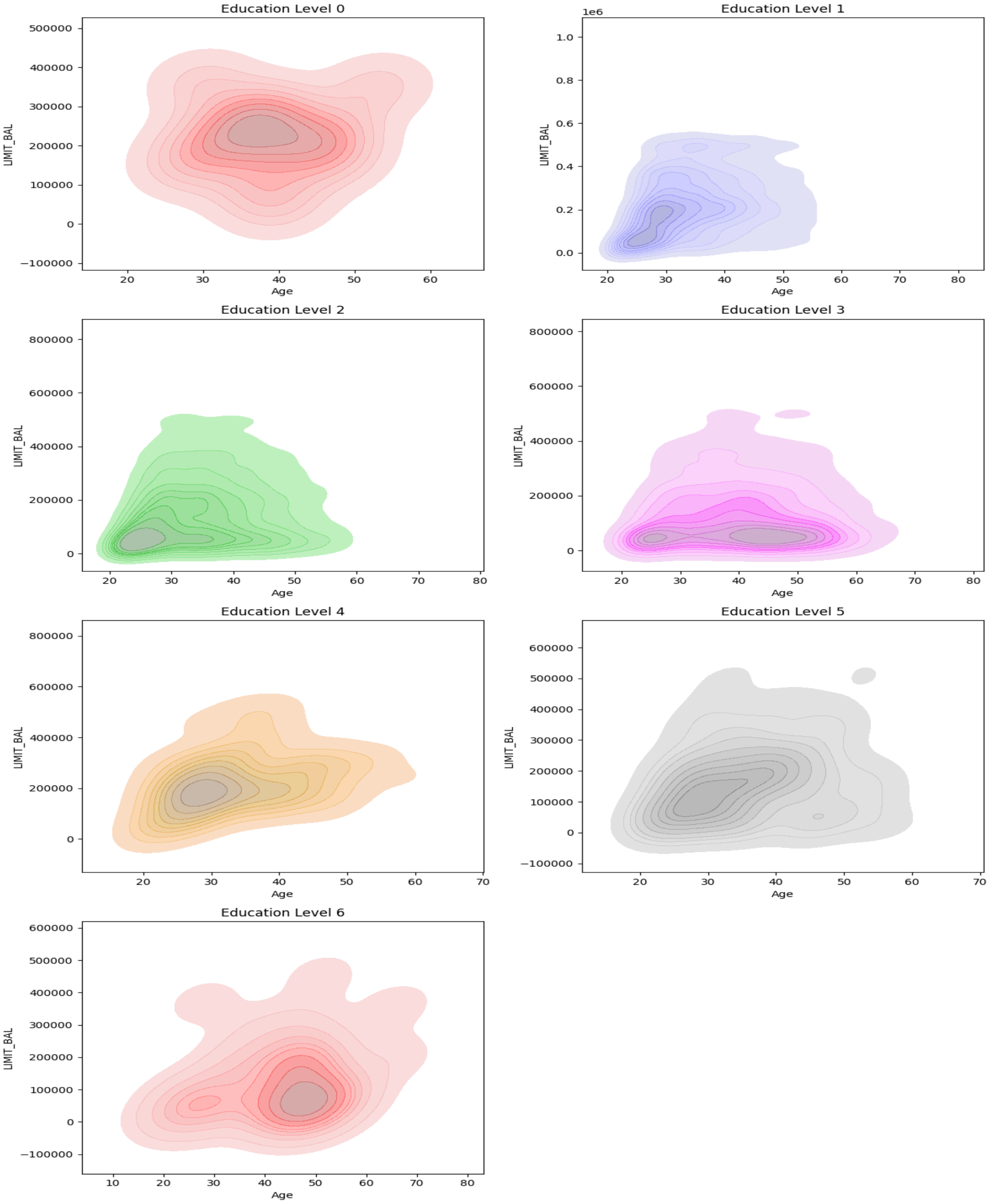
Slightly more male customers. Default rates may not differ significantly by gender, but worth testing.

d) Education Level Distribution:



Most customers fall under levels 1-3. Levels 5, 6 may be invalid or unknown categories -> clean or merge if needed.

e)

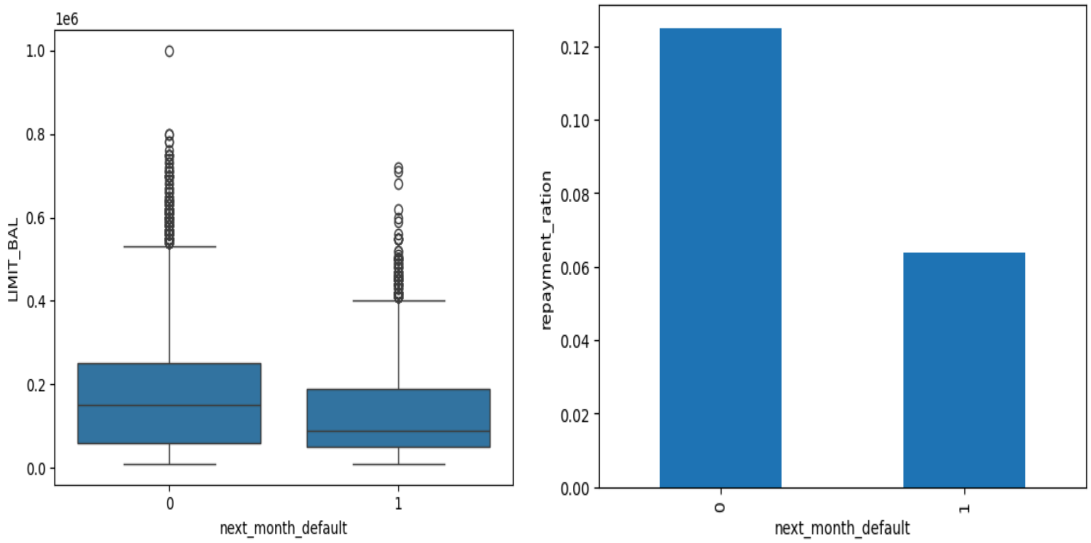


- Concentration Patterns:
  Most education levels show a central concentration of customers in the age range of roughly 25–50 years and credit limits between ₹100,000 and ₹400,000.

This suggests that the majority of credit card holders, regardless of education, fall within this age and credit limit band.

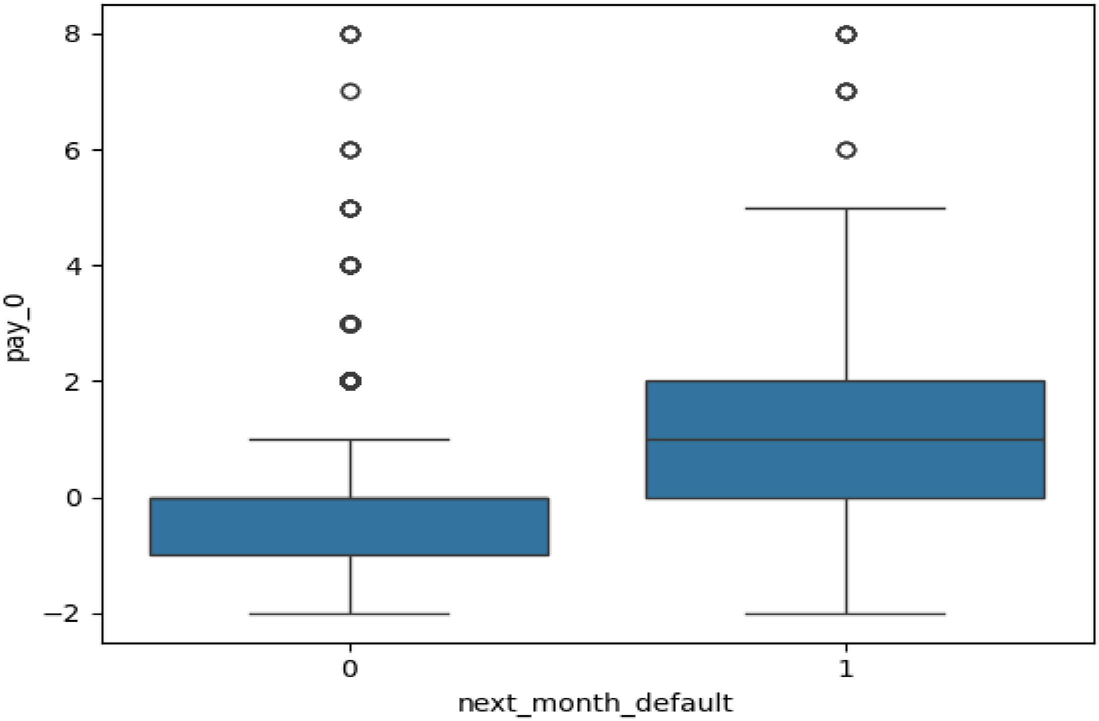
- Spread by Education Level:
    - Education Level 1 (top right) displays a broader spread in both age and credit limit, with some individuals having very high credit limits (up to ₹1,000,000) and ages extending into the 80s.
    - Other Levels (0, 2, 3, 4, 5, 6) have more tightly clustered distributions, with fewer outliers at high ages or very high credit limits.
    - Education Level 6 (bottom left) and Level 0 (top left) show a slightly wider age range, but still most customers are under 50.
- Credit Limit Trends:
  Across all education levels, higher credit limits are generally associated with middle-aged customers (30–50 years), while younger and older customers tend to have lower credit limits.
- Outliers:
  A few education levels (notably 1 and 3) have outliers with exceptionally high credit limits, but these are rare.
- Education as a Segmentation Variable:
  The distribution of credit limits and ages does not vary dramatically between education levels, though some levels (like 1) show more diversity and higher maximum credit limits. This might indicate that education level 1 corresponds to a group with higher earning potential or more established credit histories.

f)



Customers with high credit limits and consistent repayment show low default risk.

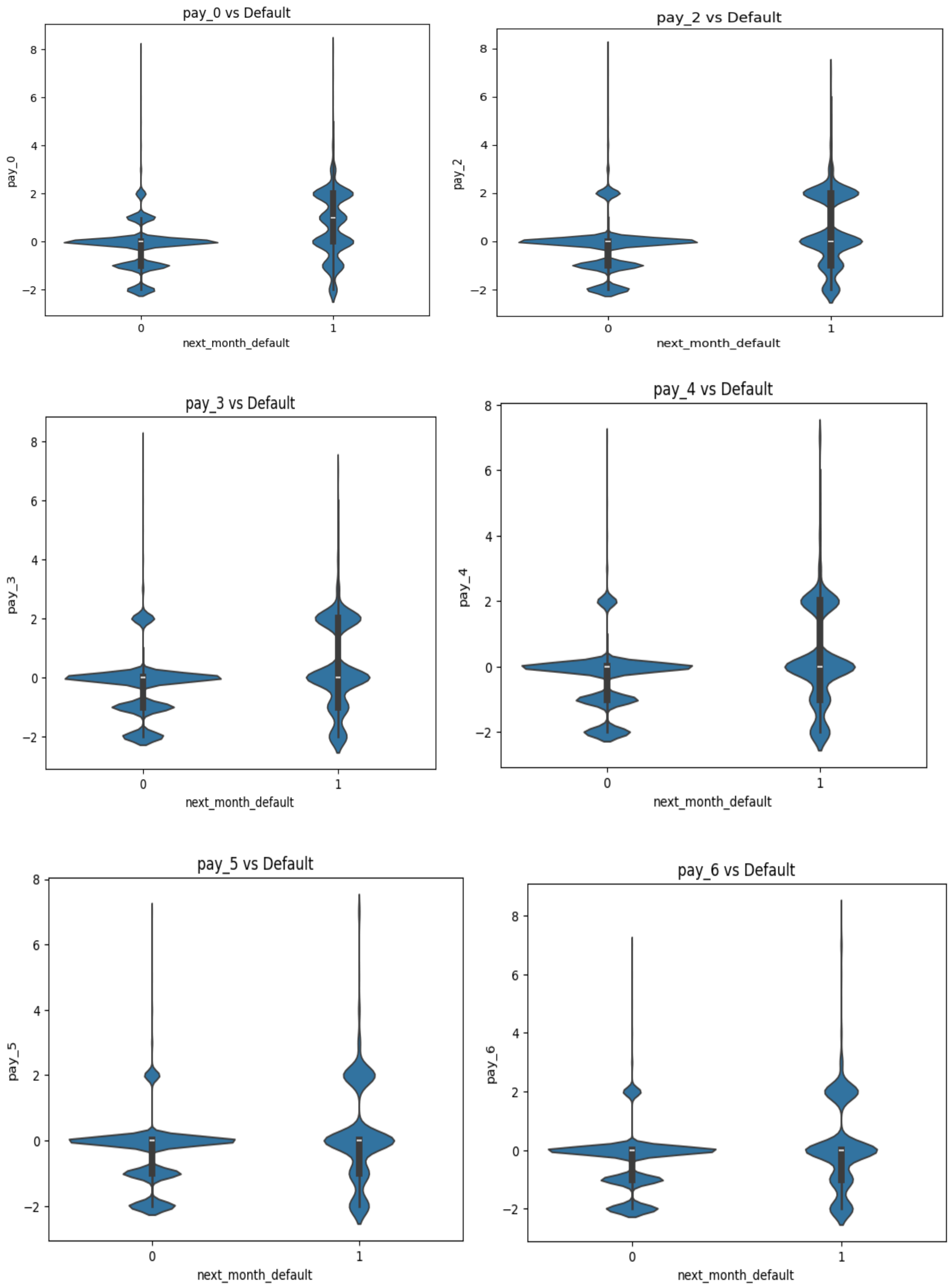
g) pay_0 vs next_month_default:



For non-defaulters (next_month_default = 0), the median payment status is around -1, which means that most customers in this group paid their bills in full and on time. The central box, representing the 25th to 75th percentile, stretches from -1 to 0. This indicates that the majority either paid in full or at least made a partial payment. The whiskers of the boxplot extend from about -2 to 1, showing that a few customers had no consumption (possibly no new charges), and some experienced a 1-month delay but still did not default. There are also a few outliers reaching up to 8, representing rare cases where customers delayed payments for up to 8 months but surprisingly did not default.

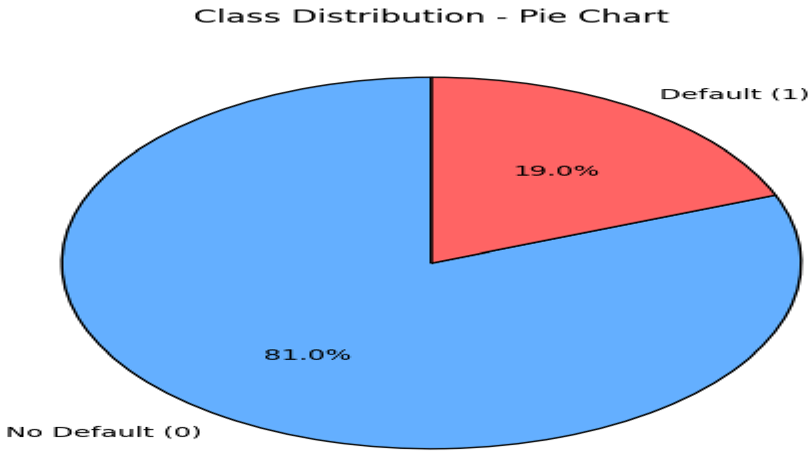
For defaulters (next_month_default = 1), the median payment status is around 1, suggesting that most of these customers had at least a 1-month overdue payment. The box (25th to 75th percentile) ranges from 0 to 2, which means that most defaulters were already experiencing payment delays in the recent months. The whiskers extend from about -2 to 5, indicating that while some defaulters did pay fully or partially (-1 or 0), many had significant delays. Outliers again reach up to 8, highlighting that severe delay behavior is present among some defaulters.

h)


pay_0 vs Default
pay_0
next_month_default
8
6
4
2
0
−2
0
1
pay_2 vs Default
pay_2
next_month_default
8
6
4
2
0
−2
0
1
pay_3 vs Default
pay_3
next_month_default
8
6
4
2
0
−2
0
1
pay_4 vs Default
pay_4
next_month_default
8
6
4
2
0
−2
0
1
pay_5 vs Default
pay_5
next_month_default
8
6
4
2
0
−2
0
1
pay_6 vs Default
pay_6
next_month_default
8
6
4
2
0
−2
0
1

- Defaulters tend to have higher values of pay_0, indicating delays or missed payments in the most recent month.
- Non-defaulters cluster around pay_0 = 0, which indicates on-time payments.
- pay_0 is one of the strongest indicators of default — it directly represents past delinquency.
- The higher and more spread-out distribution for class 1 makes this a powerful predictive feature.

i)Class Distribution:

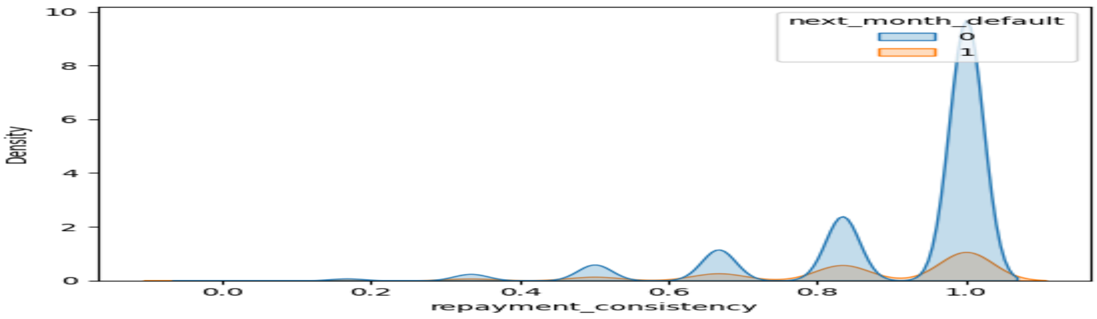


80.96% customers are non-defaulters,19.04% customers defaulted

Insight: Strong class imbalance.

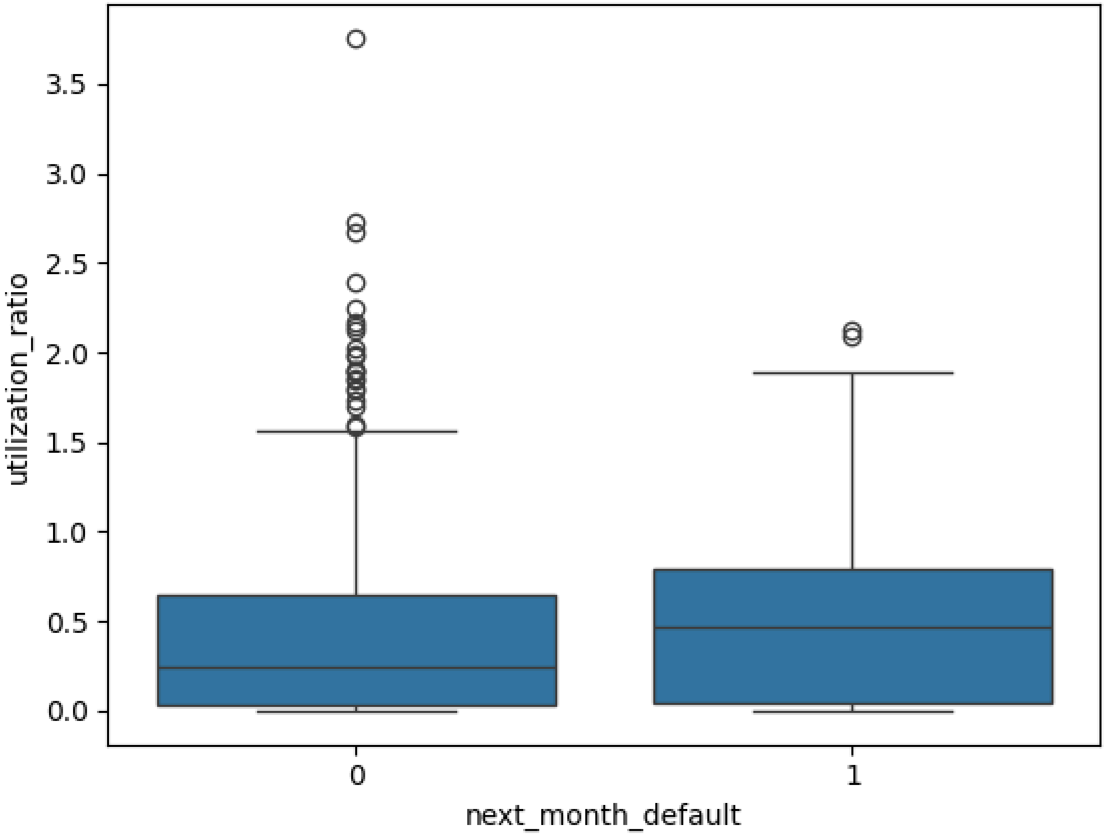
ML models may become biased toward predicting non-default unless balanced.

j) repayment_consistency:



high repayment consistency means less risk of default.

k) utilization_ratio:



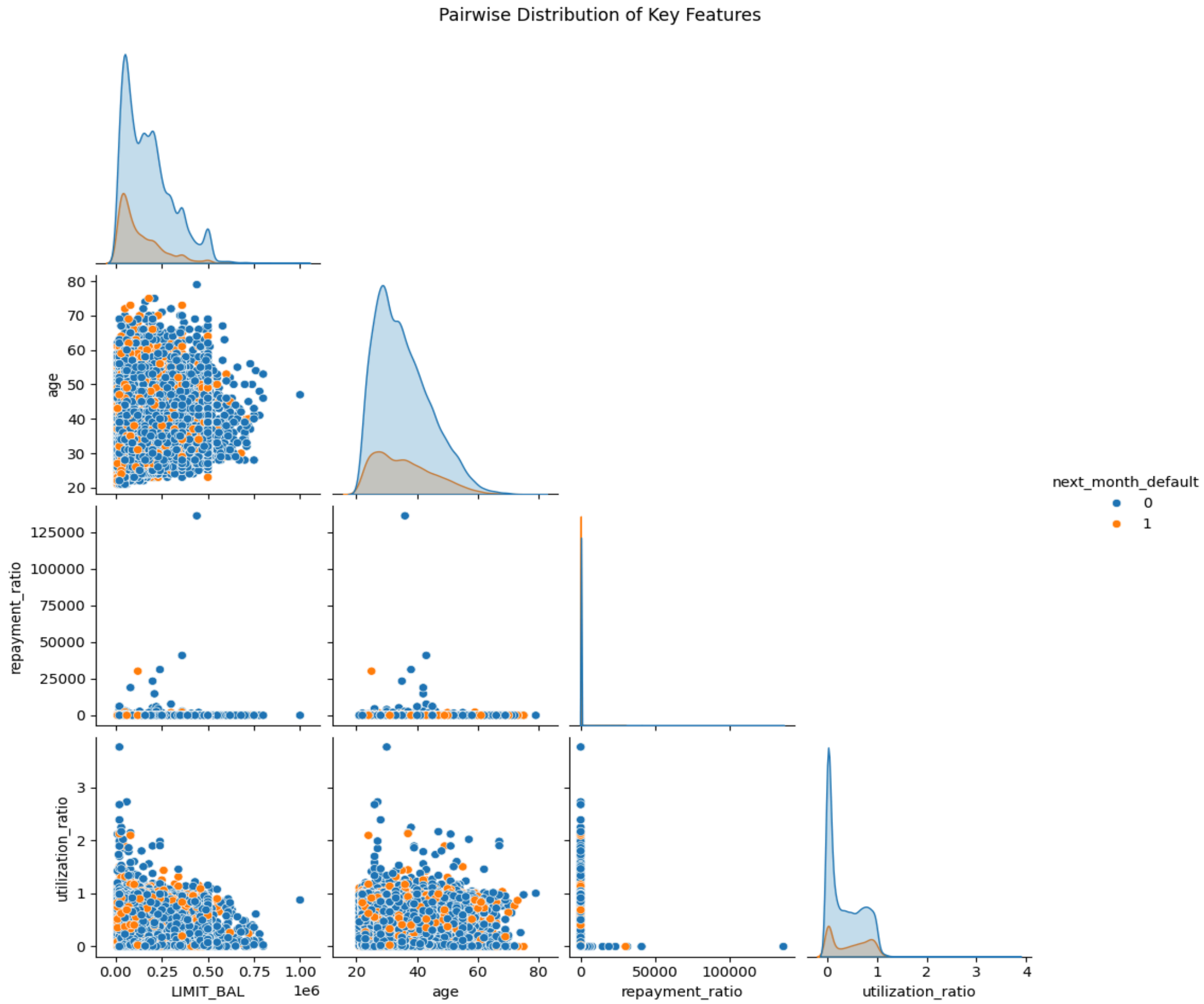
The median utilization ratio is higher for defaulters (1) compared to non-defaulters (0).

Defaulters tend to have less extreme outliers, whereas non-defaulters have more instances with extremely high utilization ratios (>2.5).

A higher utilization_ratio (i.e., using more of the available credit) is positively correlated with higher default risk.

This could be a strong predictor feature for default classification.

I)



LIMIT_BAL and utilization_ratio show some negative correlation, which makes sense — higher limits generally lead to lower utilization ratios.

age is somewhat evenly distributed, though younger users seem more scattered.

Most users (default and non-default) have very low repayment ratios, with some clear outliers.

The scatterplots don’t show clear linear separability between classes.

However, the distribution plots (top diagonal) show that:

Defaulters have a slightly higher utilization and lower repayment.

age and LIMIT_BAL have subtle impact but aren’t highly discriminative alone.

Multivariate modeling will likely be more effective than relying on individual features.

Feature Engineering

- **Missing Values: The 'age' column had 126 missing values; all other columns were complete.**
- **Feature Engineering: New features were created to capture repayment behavior, delay patterns, and utilization metrics.**

The project introduced several new features to better capture customer repayment behavior, delay patterns, and utilization metrics. Here’s a summary of the key engineered features and why they were included:

- AVG_Bill_amt: The average bill amount over the past 6 months. This helps capture a customer’s typical monthly liability and spending pattern, which is indicative of their credit usage behavior.
- PAY_TO_BILL_ratio: The ratio of total payments made to total bill amounts over 6 months. This feature reflects the customer’s repayment discipline; a lower ratio may indicate financial stress or a tendency to revolve credit.
- repayment_ratio: Similar to PAY_TO_BILL_ratio, this measures the proportion of bills that are actually repaid, giving further insight into repayment consistency.
- num_late_months: Counts the number of months with delayed payments (where the payment status is positive). This quantifies chronic lateness, which is a strong indicator of default risk.
- max_delay: The maximum delay recorded in the payment status features. This captures the worst-case delinquency experienced by the customer.
- avg_delay: The average of the payment delay status over the relevant months. This smooths out fluctuations and highlights persistent payment issues.
- repayment_consistency: Measures how consistently the customer makes timely repayments. Consistency is often valued in credit risk models because sporadic behavior can signal instability.
- utilization_ratio: The average bill amount divided by the credit limit. High utilization is a classic predictor of financial stress and default risk.
- delinquency_streak: The longest consecutive streak of late payments. This helps identify patterns of sustained financial trouble.
- utilization_maxdelay_mult: A composite feature multiplying utilization ratio by maximum delay, capturing the joint effect of high credit usage and severe delinquency.
- BILL_PCA1: The first principal component from a PCA on bill amounts, summarizing the most significant variance in billing patterns in a single feature for dimensionality reduction and noise reduction.

These features were engineered based on domain knowledge and exploratory data analysis, with the aim of making the model more interpretable and predictive by reflecting real-world credit scoring logic.

Removed Features and Their Rationale

Some features were removed or replaced during the feature engineering process:

- Raw monthly bill and payment columns (e.g., Bill_amt1, pay_amt1, etc.): After extracting aggregate and behavioral features (like averages, ratios, and delays), the raw monthly columns became redundant and risked introducing multicollinearity. Removing them simplified the feature set and reduced noise without losing predictive power.
- Highly correlated or low-variance features: Features that did not provide additional information (e.g., those with little variance across customers or high correlation with other engineered features) were candidates for removal to prevent overfitting and improve model generalization.
- Customer_ID: As a unique identifier, it carries no predictive value and was dropped before modeling1.

- **Dimensionality Reduction: PCA was applied to bill amount features to reduce dimensionality and capture key variance.**

Handling Class Imbalance

Synthetic Minority Over-sampling Technique (SMOTE) was used to balance the classes in the training data, as default cases are typical

---

Model Building

A variety of classifiers were imported and used, including:

- Logistic Regression
- Decision Tree
- Random Forest
- Gradient Boosting (HistGradientBoosting, AdaBoost, GradientBoosting)
- XGBoost
- LightGBM
- CatBoost
- Support Vector Machine (SVC)

- K-Nearest Neighbors
- Naive Bayes
- Voting Ensemble Classifier

Hyperparameter tuning was performed using GridSearchCV and RandomizedSearchCV. Model evaluation metrics included accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrix1.

Results and Discussion

Class	Precision	Recall	F1-Score	Support
0	0.90	0.79	0.84	4095
1	0.41	0.62	0.49	955
Overall Accuracy	—	—	0.76	5050
Macro Avg	0.66	0.71	0.67	5050
Weighted Avg	0.81	0.76	0.78	5050

Business Implications: Impact of False Positives and False Negatives

A key consideration in deploying a credit default prediction model is understanding the practical consequences of its errors, specifically false positives and false negatives. These outcomes have significant financial and reputational implications for the bank.

False Positives (Type I Error)

A false positive occurs when the model predicts a customer will default, but in reality, they would not have. In the banking context, this means:

- Lost Revenue and Opportunities: The bank may unnecessarily restrict credit, reduce credit limits, or decline new credit to good customers. This can lead to lost interest income and missed business opportunities.
- Customer Dissatisfaction: Customers wrongly flagged as high risk may feel unfairly treated, leading to dissatisfaction, damaged trust, and potential churn to competitors.
- Resource Strain: Unnecessary risk mitigation actions (e.g., more frequent reviews, stricter terms) consume valuable operational resources that could be better used elsewhere.

False Negatives (Type II Error)

A false negative occurs when the model fails to flag a customer who actually will default. For the bank, this is generally more costly:

- Direct Financial Loss: The bank extends credit to a customer who fails to repay, resulting in the loss of the principal and any expected interest. The financial impact of a false negative is typically much higher than a false positive—often by an order of magnitude.
- Increased Credit Risk: Accumulation of undetected defaulters can undermine the bank's risk profile, affecting capital adequacy and regulatory compliance.
- Reputational Damage: High default rates may erode investor and public confidence in the bank's risk management practices.

Balancing the Trade-Off

- In credit risk, false negatives are generally considered far more costly than false positives. For example, if the loss from a defaulted loan is the full loan amount, but the loss from denying a good customer is only the lost interest, then a single false negative can outweigh several false positives.

- The model’s classification threshold should be set to reflect the bank’s risk appetite, prioritizing recall (catching as many actual defaulters as possible) even at the expense of some false positives.
- The business must also consider the long-term effects on customer relationships and operational efficiency when calibrating the model.