REPORT

Overview of Approach and Modeling Strategy

The goal of this analysis is to predict credit default risk based on various customer demographic and financial behavior features. Initial data preprocessing steps focused on ensuring data integrity by addressing missing values in the age column and correcting erroneous negative values in avg_bill_amt and pay_to_bill_ratio. To tackle class imbalance, the SMOTE technique was applied, which helped generate synthetic samples for the minority class, enabling the model to generalize better and avoid bias toward the majority class.

The further workflow begins with a thorough exploratory data analysis (EDA) to uncover key patterns and risk indicators. The exploratory data analysis phase uncovered behavioral and demographic patterns associated with default. For ex: financial indicators such as lower credit limits, higher average bill amounts, low PAY_TO_BILL ratios, and payment delays (especially PAY_0 > 1) were strong predictors of default. These insights led to the creation of new derived features such as total_payment, credit utilization ratio, delinquency streak etc.

Multiple classification models were evaluated such as logistic regression, decision tree, random forest, LightGBM and XGboost. I have used approaches like GridSearchCV, feature engineering and threshold optimization to further refine the model. These steps were crucial in enhancing model robustness, especially in improving recall and F2-score.

Finally, best model was used for prediction on the validation dataset with filtering through threshold selection.

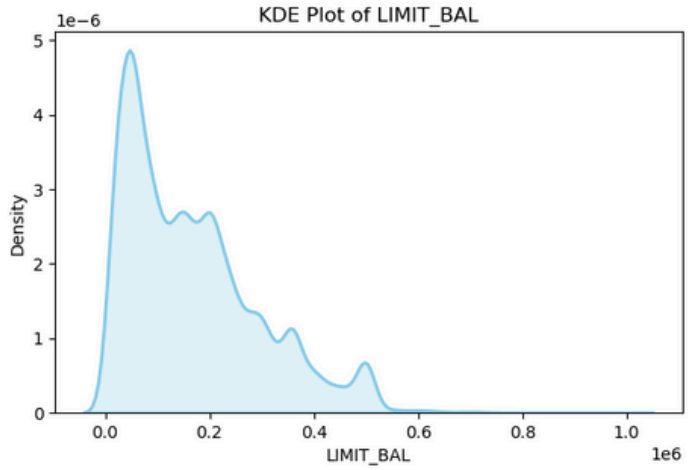
EDA findings and visualizations

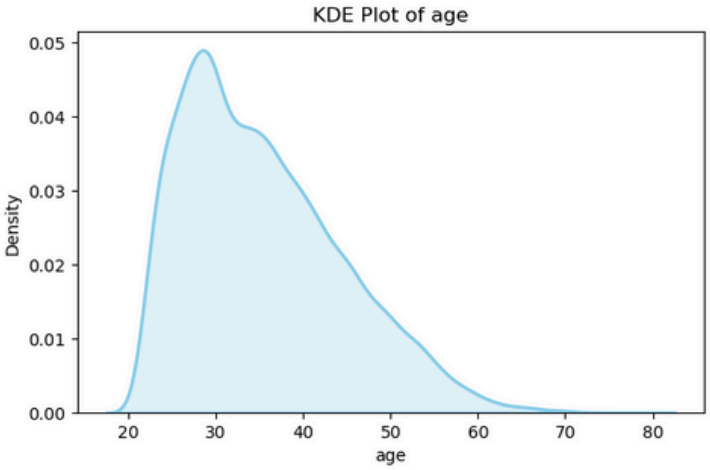
1. Feature Distributions (KDE Plots – LIMIT_BAL, Age, AVG_Bill_amt):

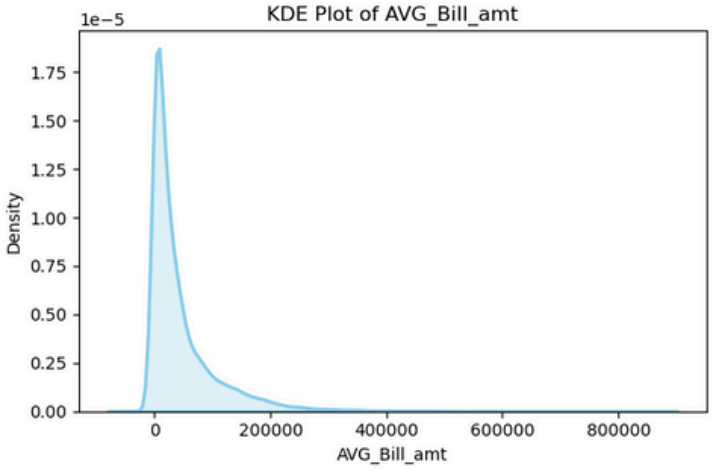
- The KDE plot of LIMIT_BAL shows a strong right-skew, indicating most customers have relatively low credit limits, with a few having very high ones.

The Age distribution is left-skewed, concentrated between 25–45 years, tapering off toward 70+.

- The AVG_Bill_amt is highly right-skewed with a dense cluster of customers having low average bill amounts, and very few with high bills. skewness factor for each is 0.99, 0.73 and 2.73 respectively.
- The skewness indicates that most customers fall into modest credit usage and age brackets, while extreme values (like very high bills or limits) are rare. These skews may require transformation (e.g., log scale or box-cox) for modeling.

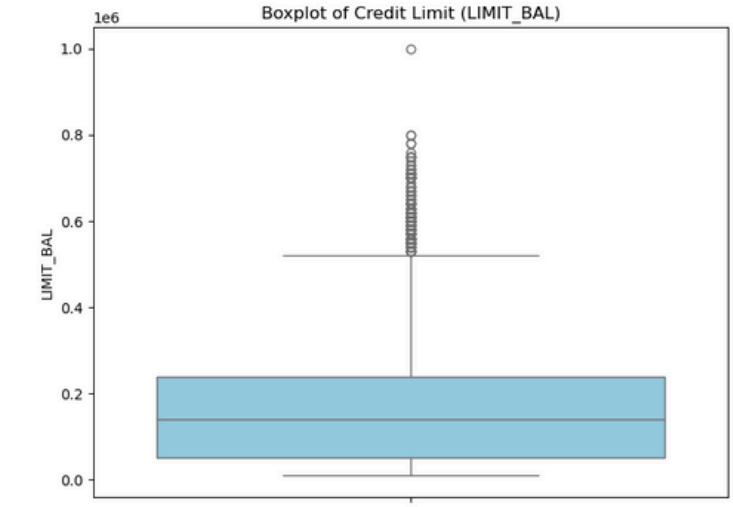


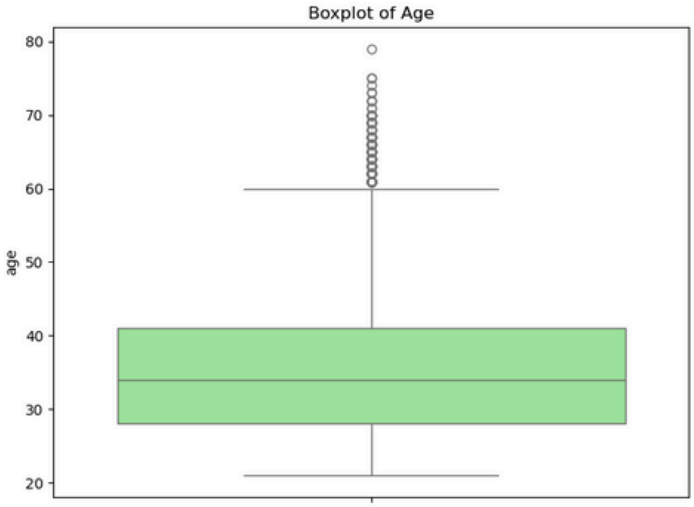


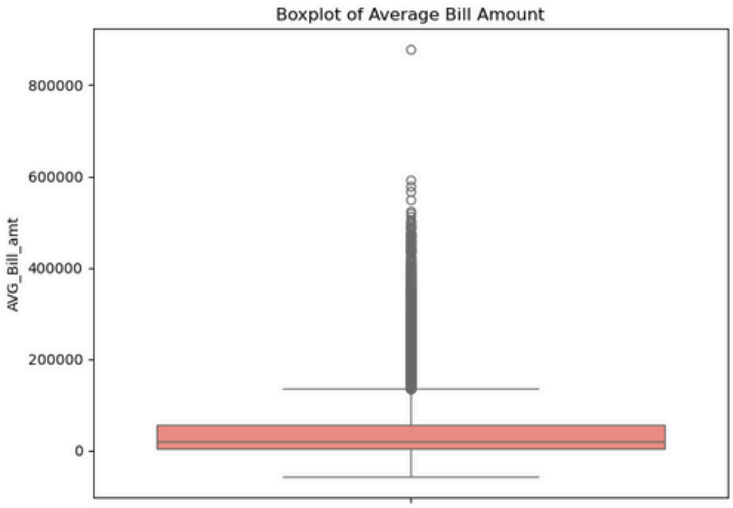


2. Outlier Detection via Boxplots.

- Boxplots for all three features show significant presence of outliers.
  LIMIT_BAL and AVG_Bill_amt especially show extreme outliers far from the upper quartile, while Age shows moderate outliers above age 60.



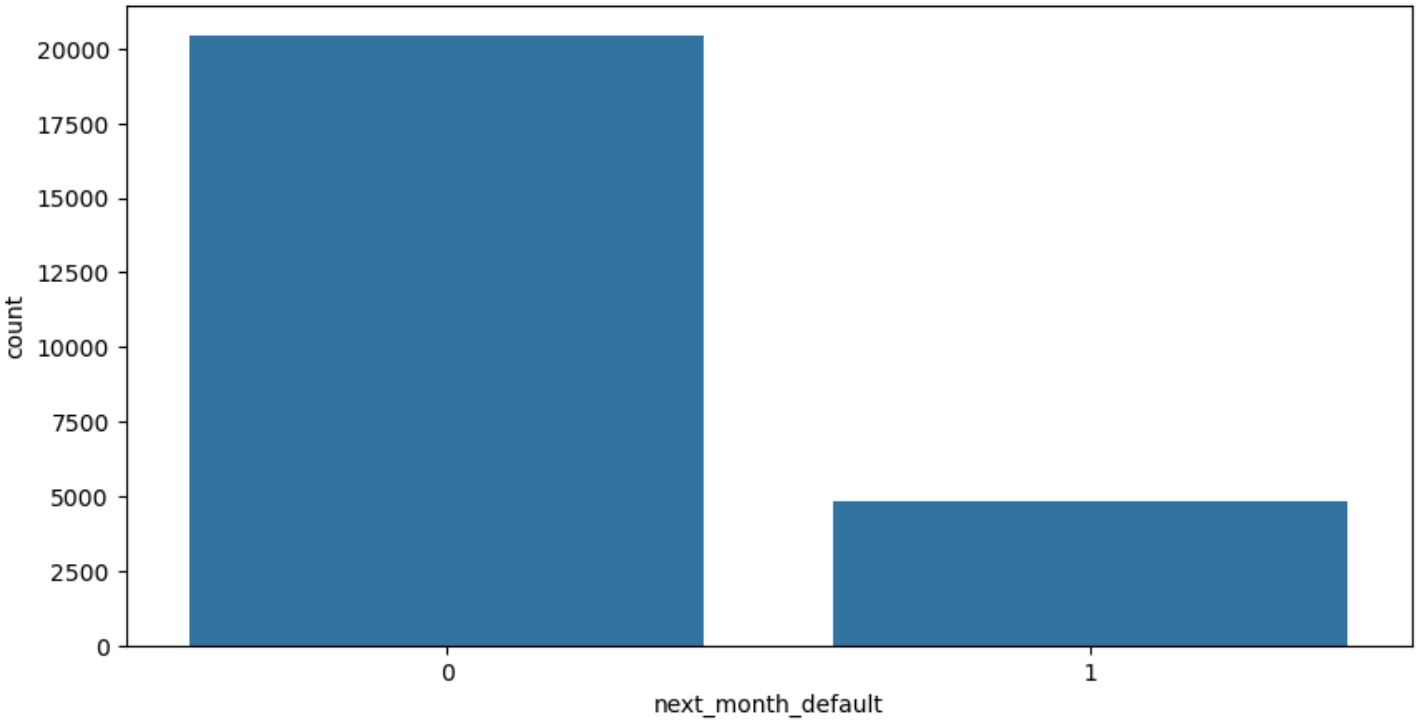




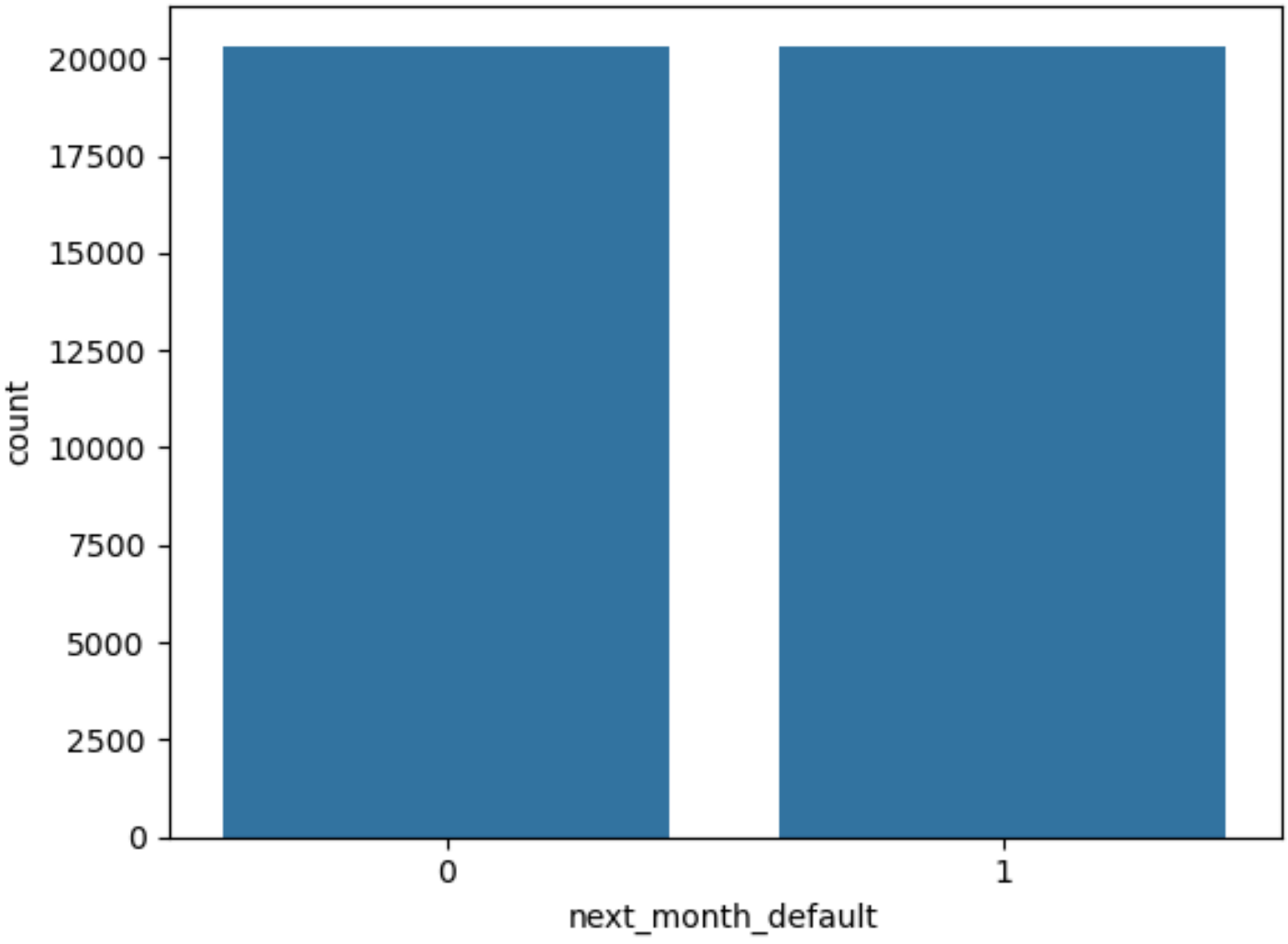
3. Class Distribution:

- The dataset shows significant class imbalance, with non-defaulters far outnumbering defaulters. This imbalance can bias the model toward predicting the majority class.
- To address this, I applied **SMOTE (Synthetic Minority Over-sampling Technique)** to balance the classes.

Before SMOTE:

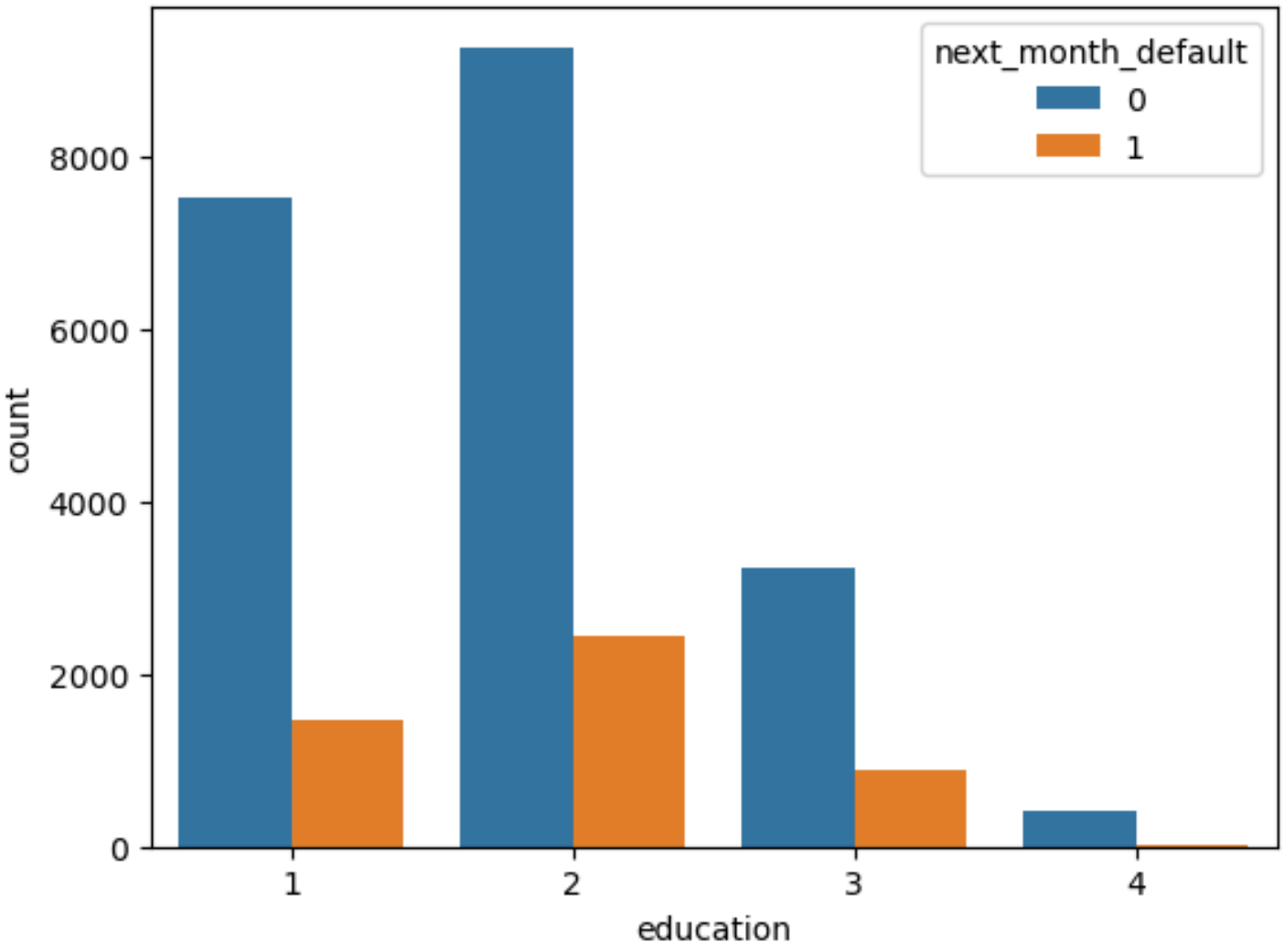


After applying SMOTE:



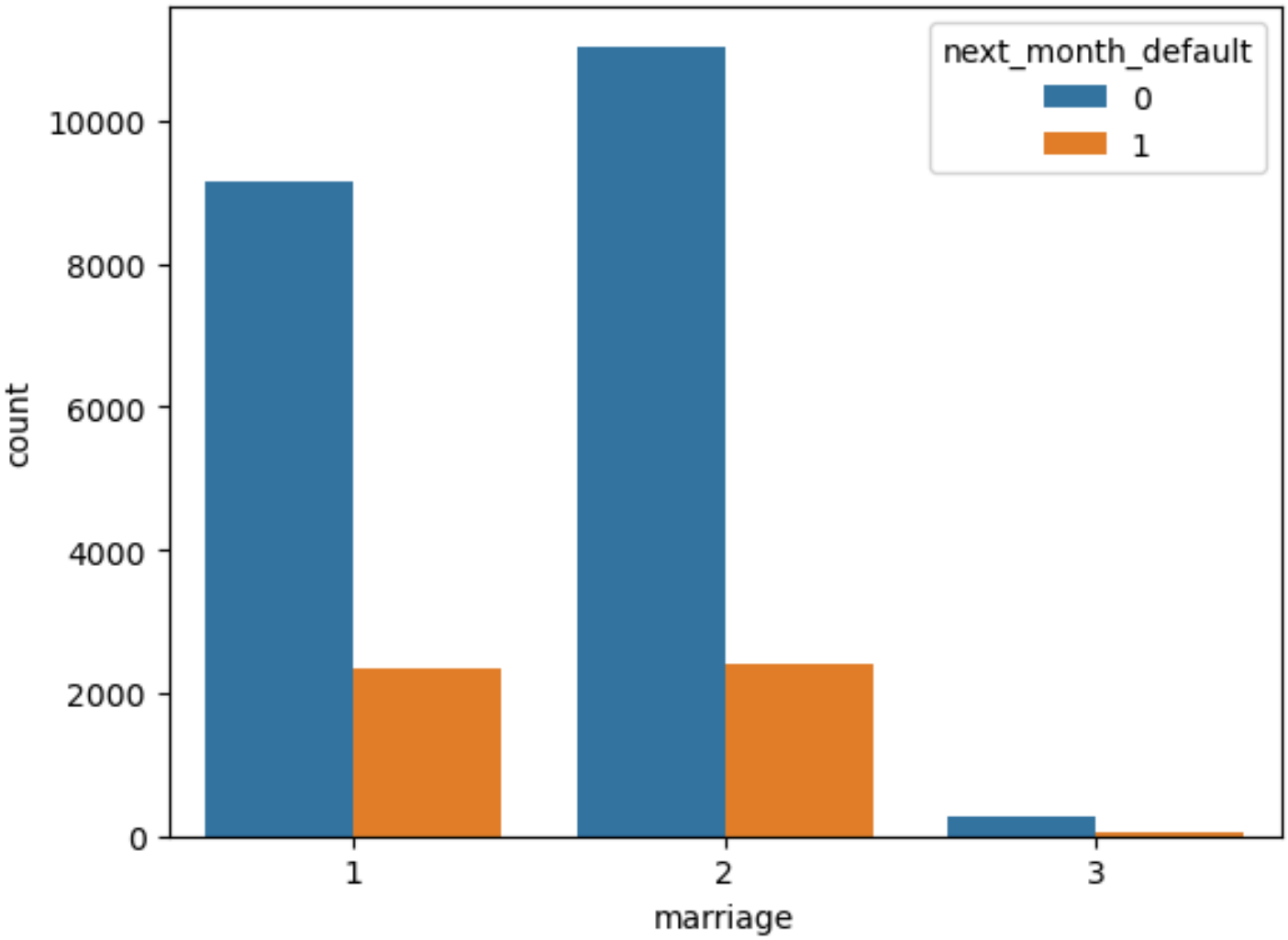
4. Education Level vs Default

- Customers with education level 2 (university) and 1 (graduate school) have the highest number of defaults in absolute terms. Education level 4 (others) shows fewer defaults but also has low total counts.



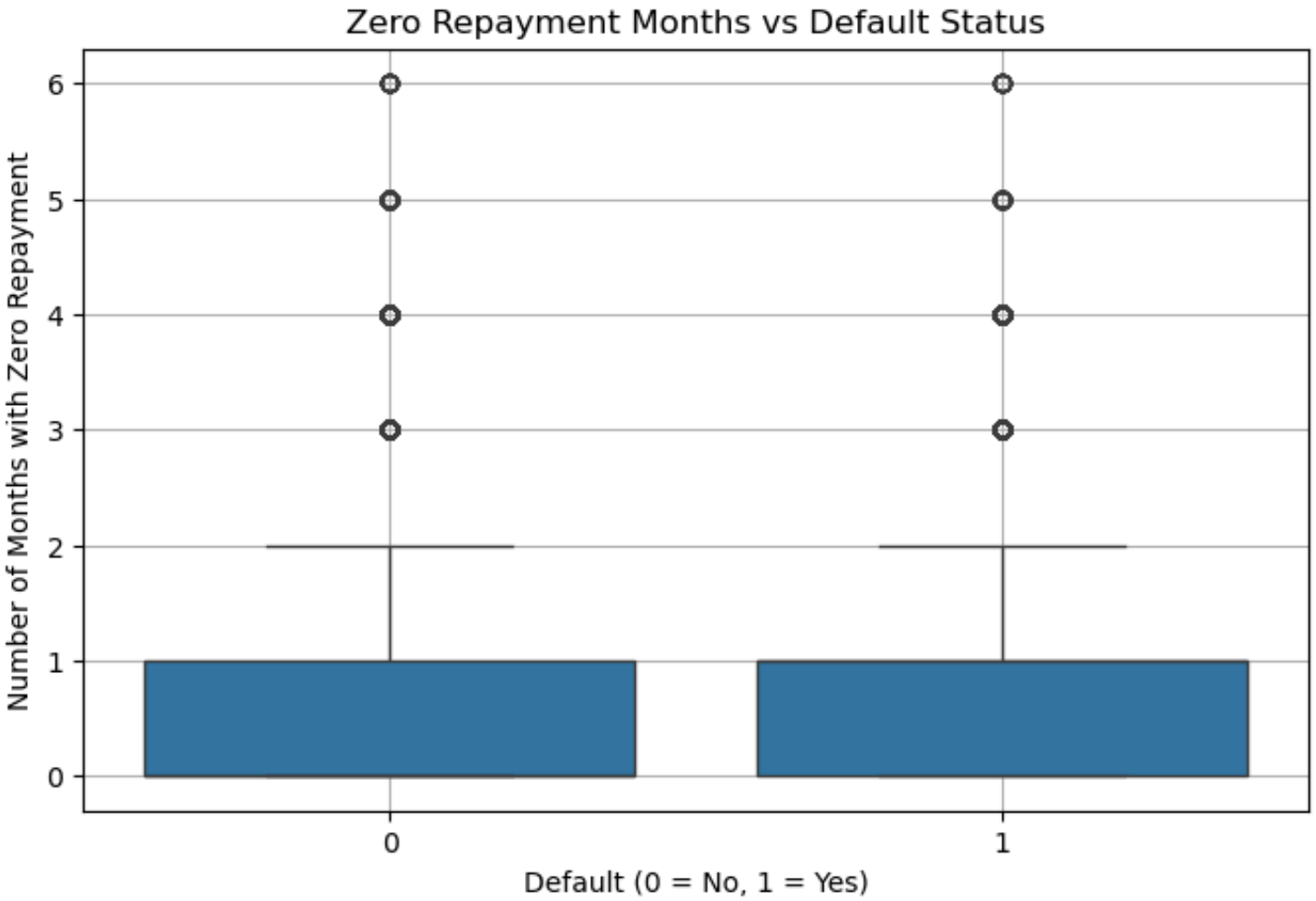
5. Marital Status vs Default

- Married customers (2) have a higher default count than singles (1), even though their total customer base is similar. Very few divorced (3) customers are present.



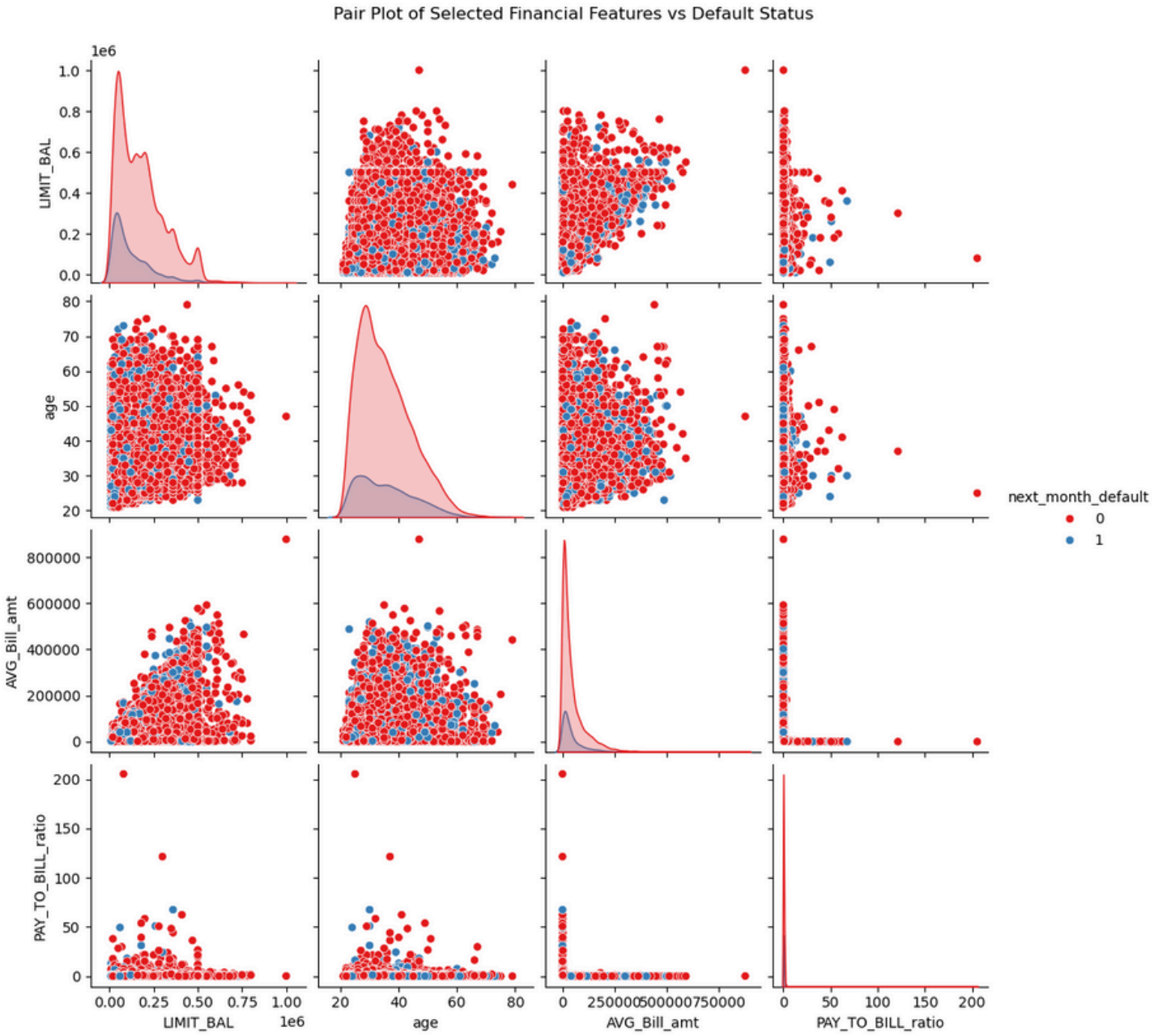
6. Zero Repayment Months vs Default.

- Both defaulters and non-defaulters show similar distributions in terms of the number of months with zero repayment. The number of such months ranges commonly between 0 and 3 for both groups.



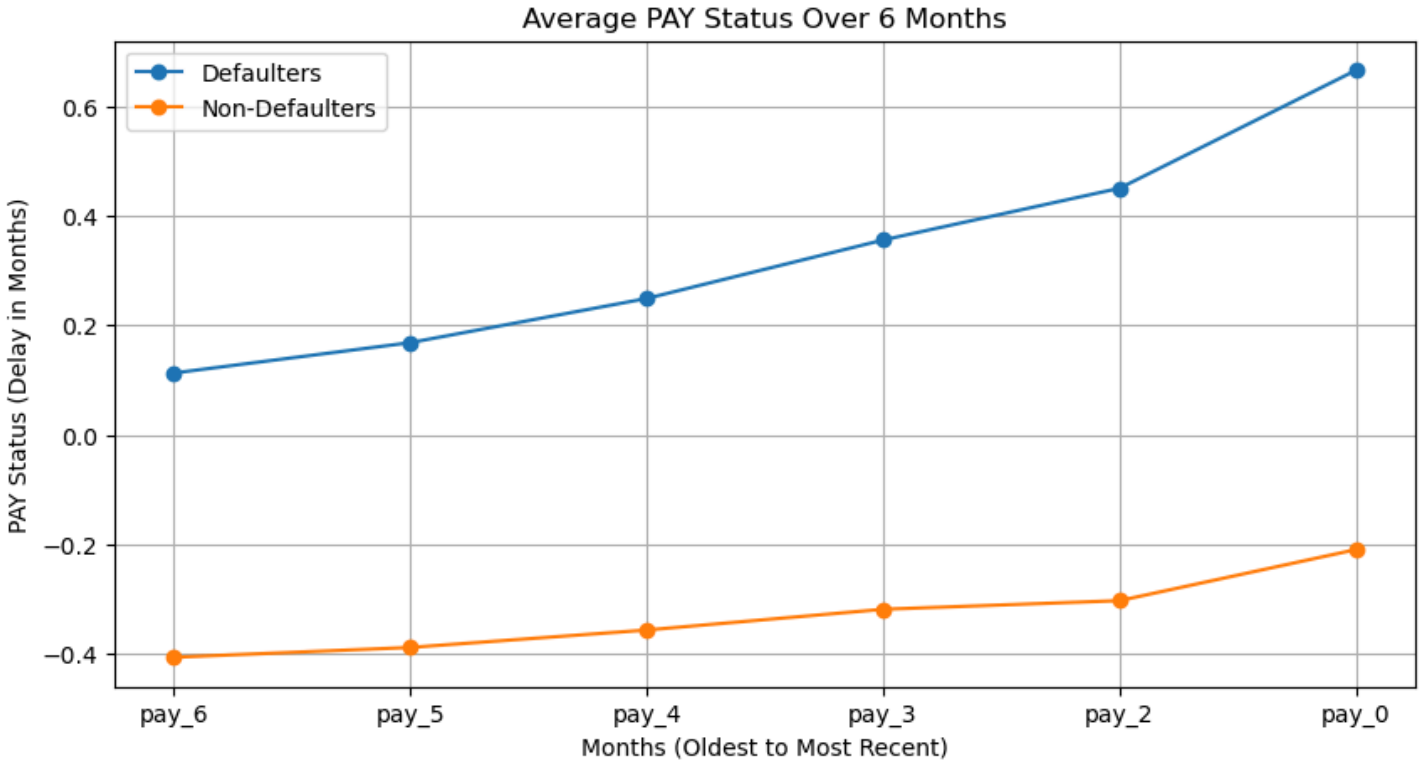
7. PAIR PLOT

- Moderate positive correlation between LIMIT_BAL and age, suggesting higher credit limits for older customers.
- LIMIT_BAL and AVG_BILL_amt show positive correlation, indicating customers with higher limits tend to have higher average bills.
- PAY_TO_BILL_ratio appears largely independent of other variables with most values clustered at low ratios.



8. Average PAY Status Over 6 Months

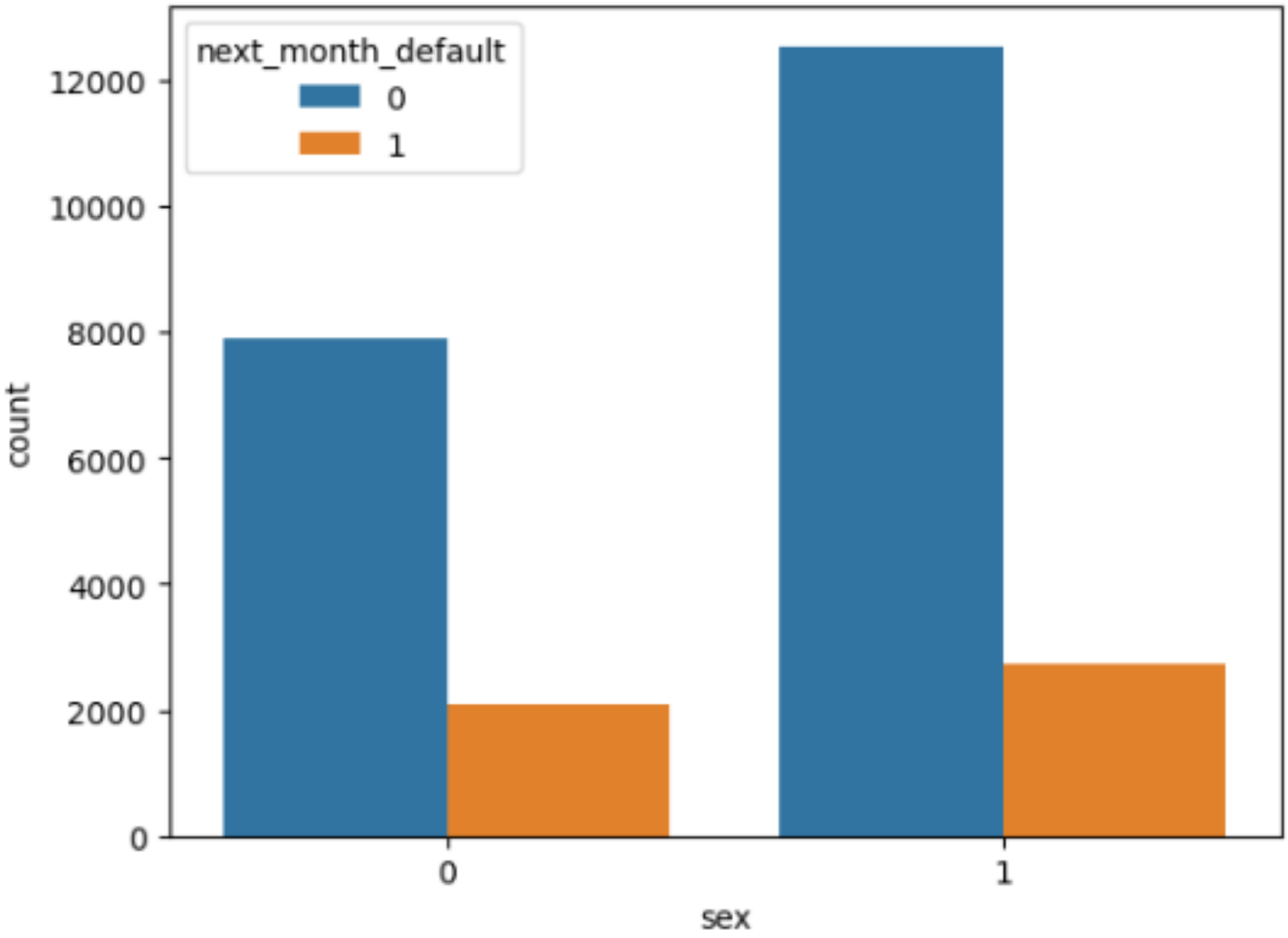
- **Defaulters (blue line)** show consistently positive PAY status values, indicating delayed payments or missed payments across all months
- **Non-defaulters (orange line)** maintain negative PAY status values, suggesting early payments or credits



Financial Drivers of Default – Insights & Interpretations

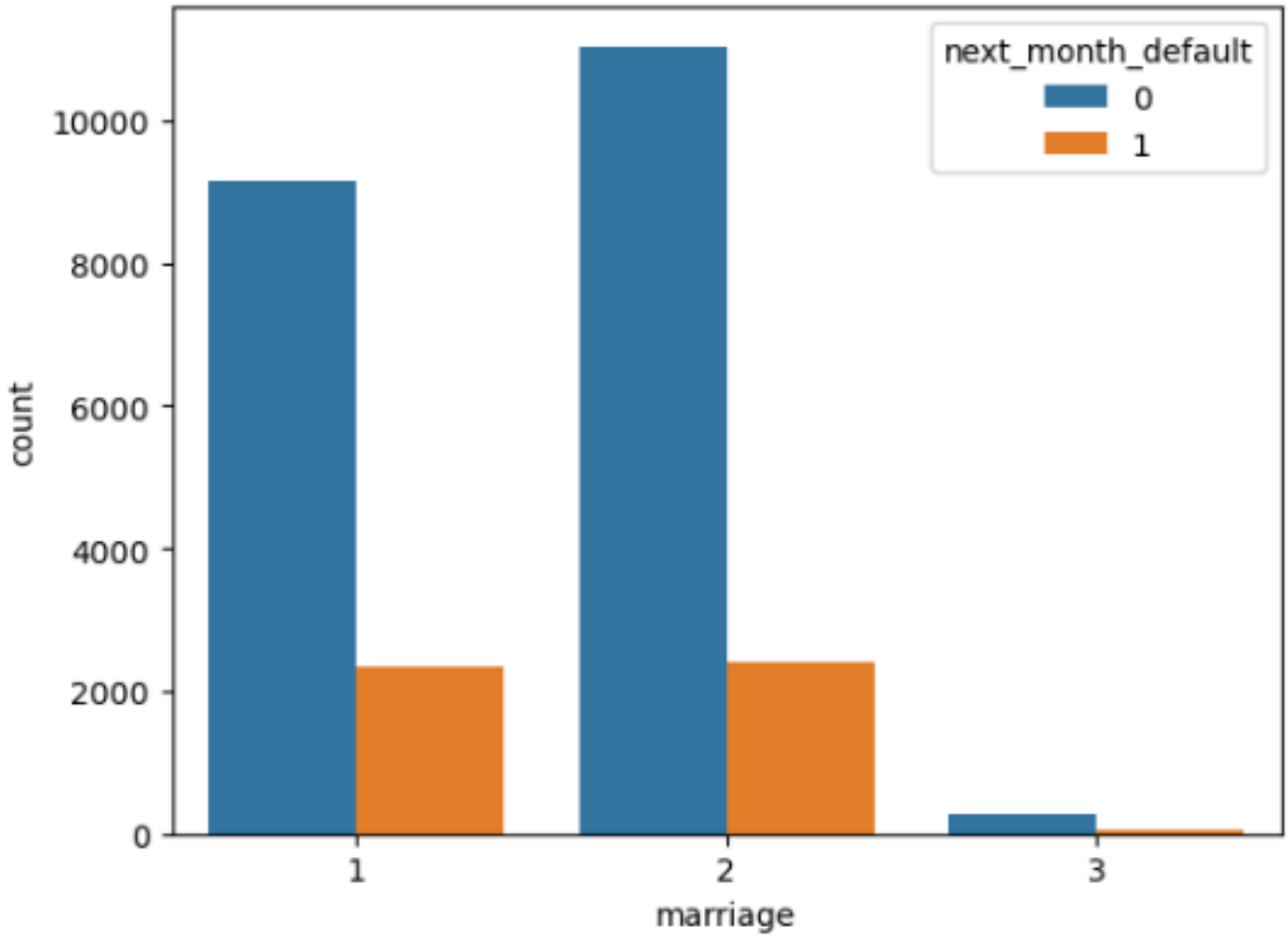
1. Gender

- **Observation:** The default rate among women (20.86%) is slightly higher than among men (17.84%).
- **Interpretation:** Gender has a slight effect on default risk. Female customers might need tailored credit education or monitoring.



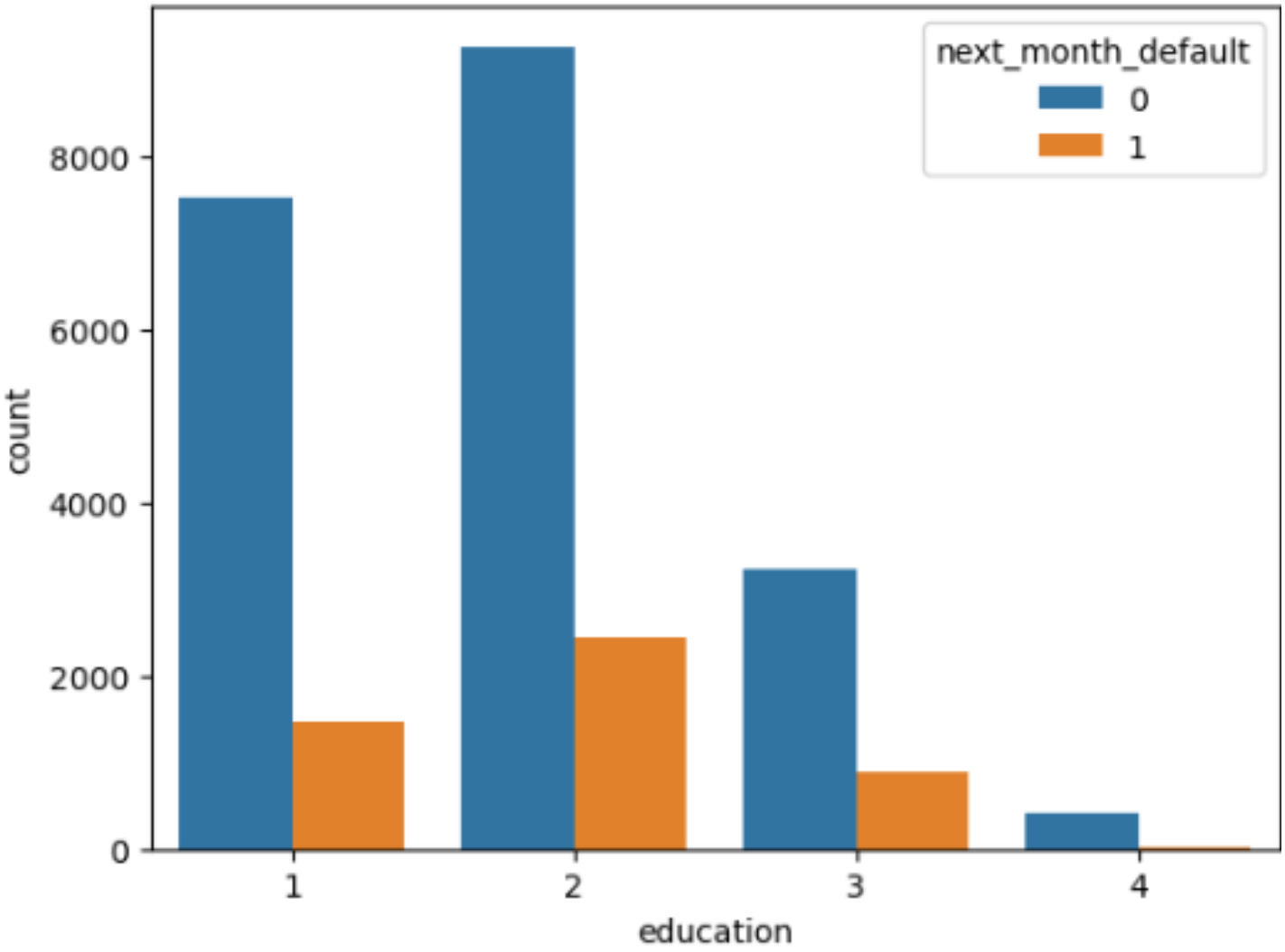
2. Marital Status

- **Observation:** Married individuals have a higher chance (20.37%) of defaulting compared to single individuals (17.88%).
- **Interpretation:** This could be linked to increased financial responsibilities like family or housing expenses.



3. Education Level

- **Observation:** Customers with university and high school education show higher default tendencies.
- **Interpretation:** Default risk isn't limited to less educated segments. Even educated individuals might need better financial planning tools.

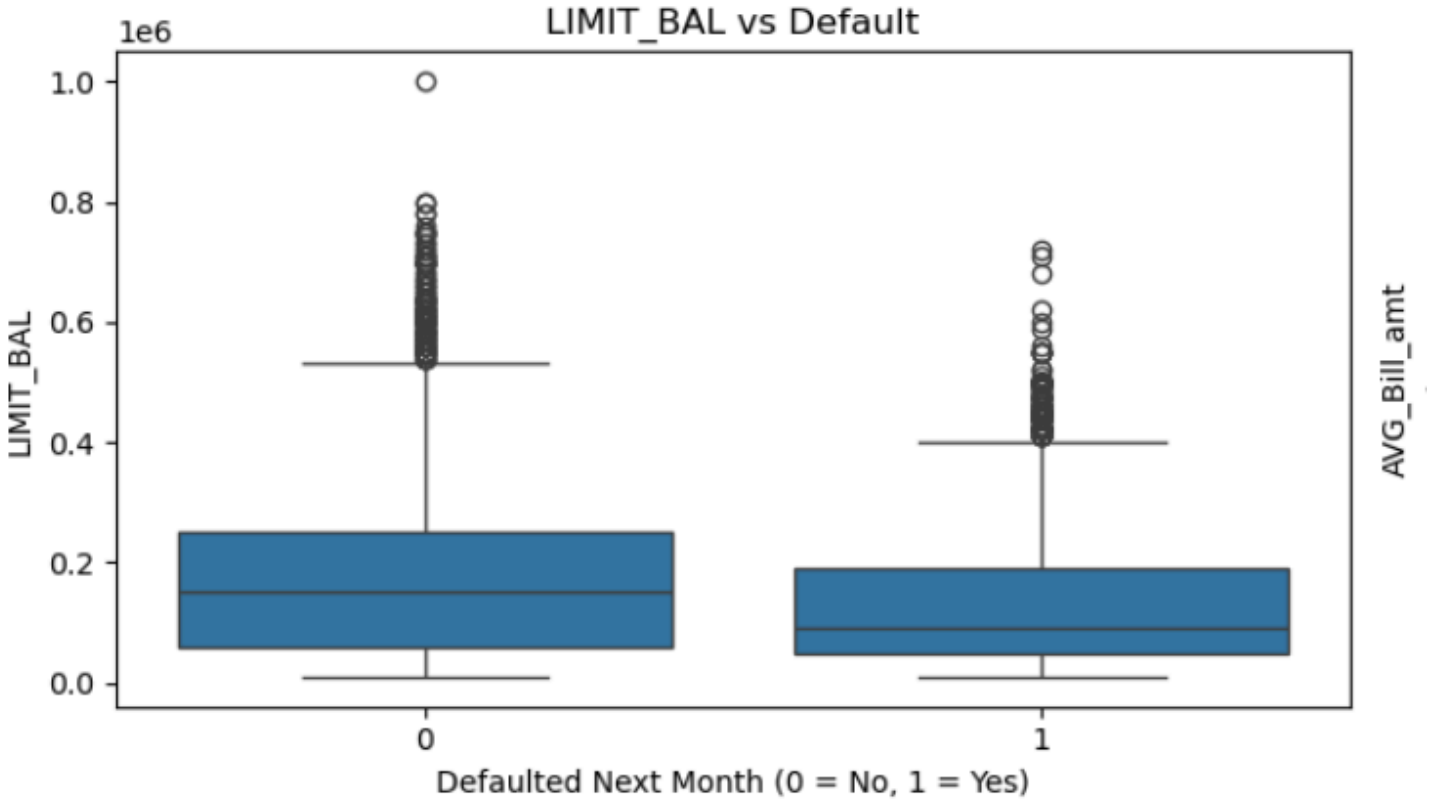


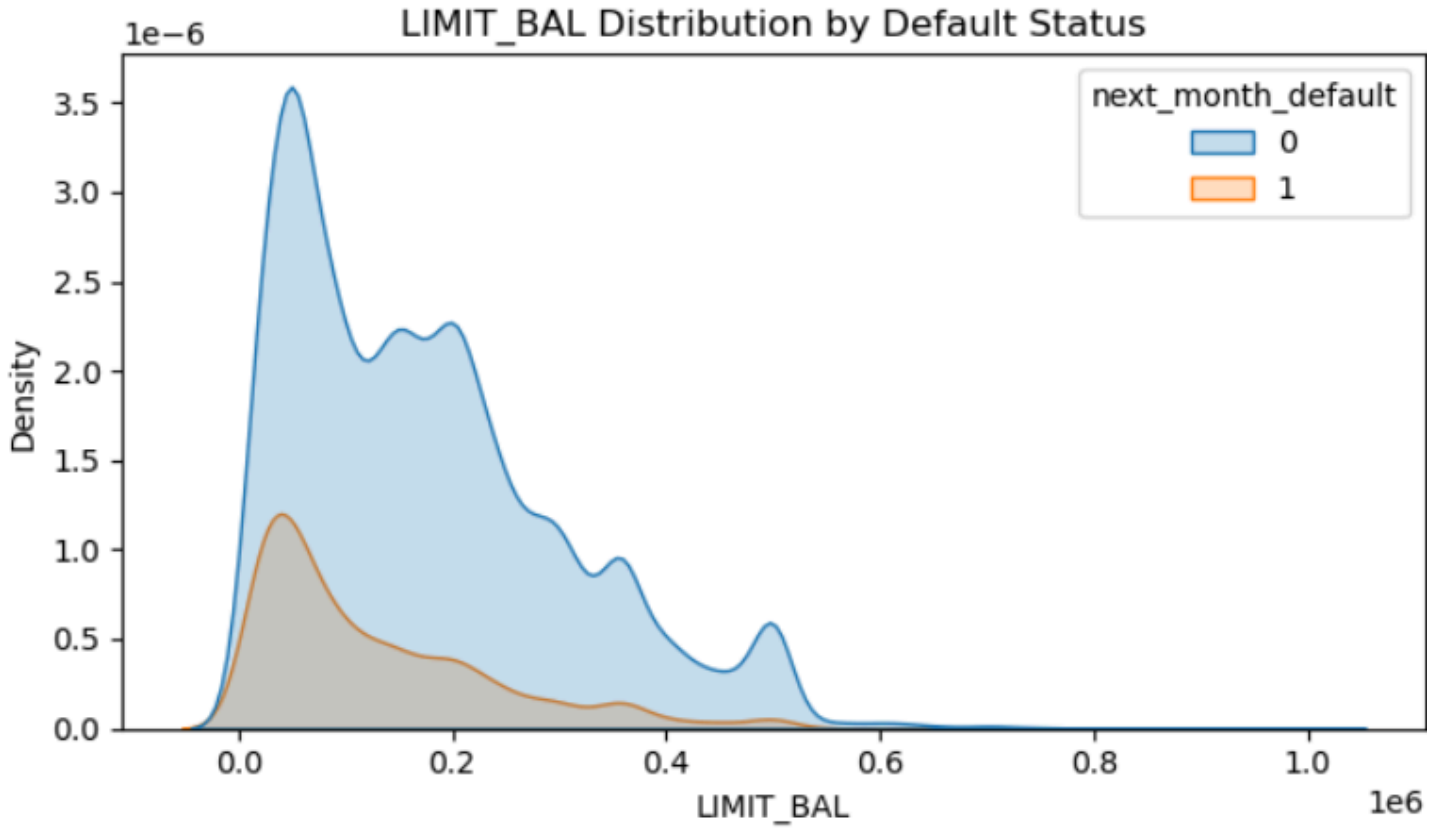
4. Age Distribution

- **Observation:** Both senior citizens and young individuals show a higher chance of default.
- **Interpretation:** Younger individuals may lack experience in managing credit; older individuals might face reduced income flow or fixed obligations.

5. Credit Limit (LIMIT_BAL)

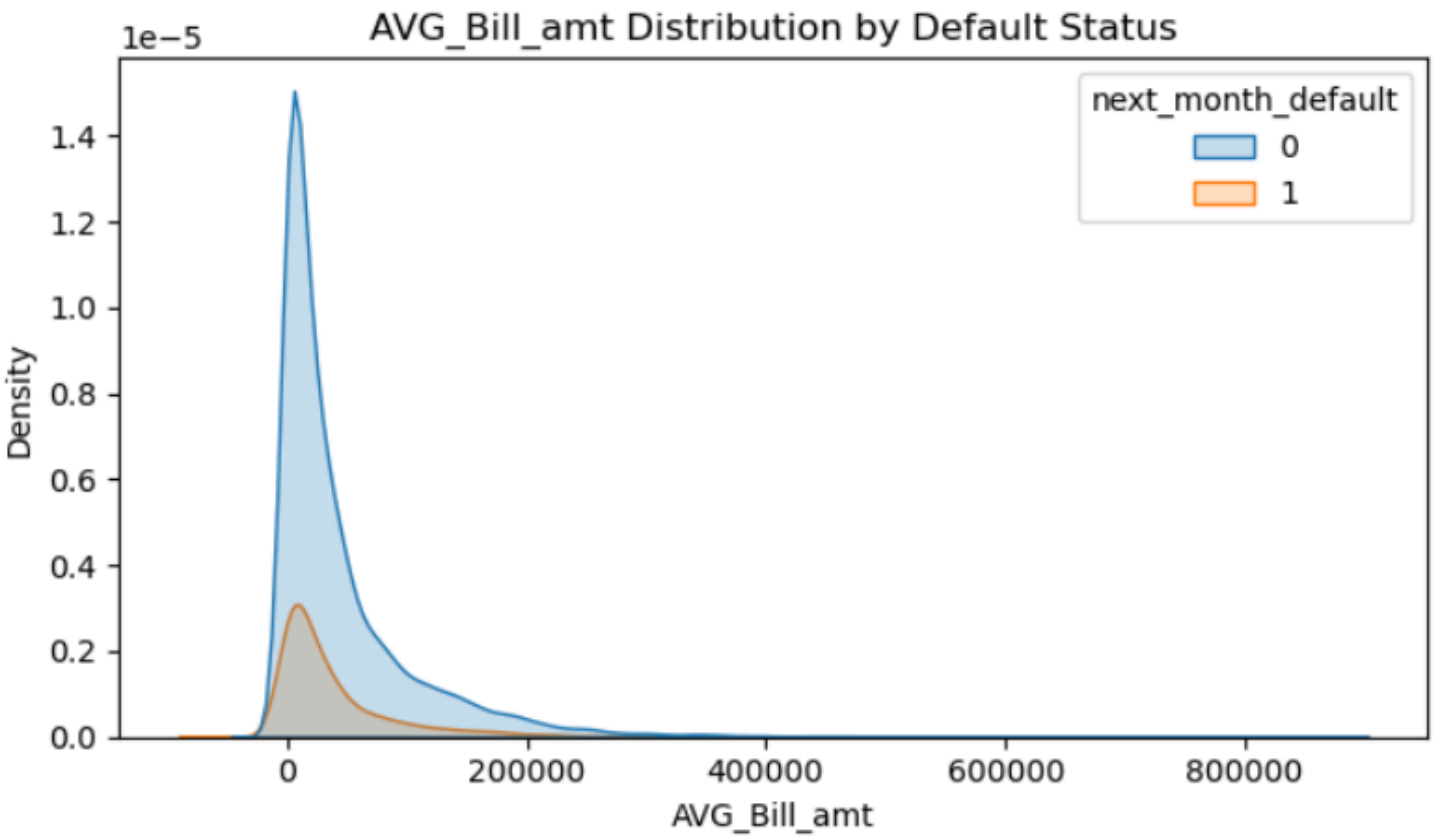
- **Observation:** Defaulters generally have lower credit limits and also distribution is right-skewed.
- **Interpretation:** Credit limit is a strong financial indicator. Lower credit may be associated with higher risk.

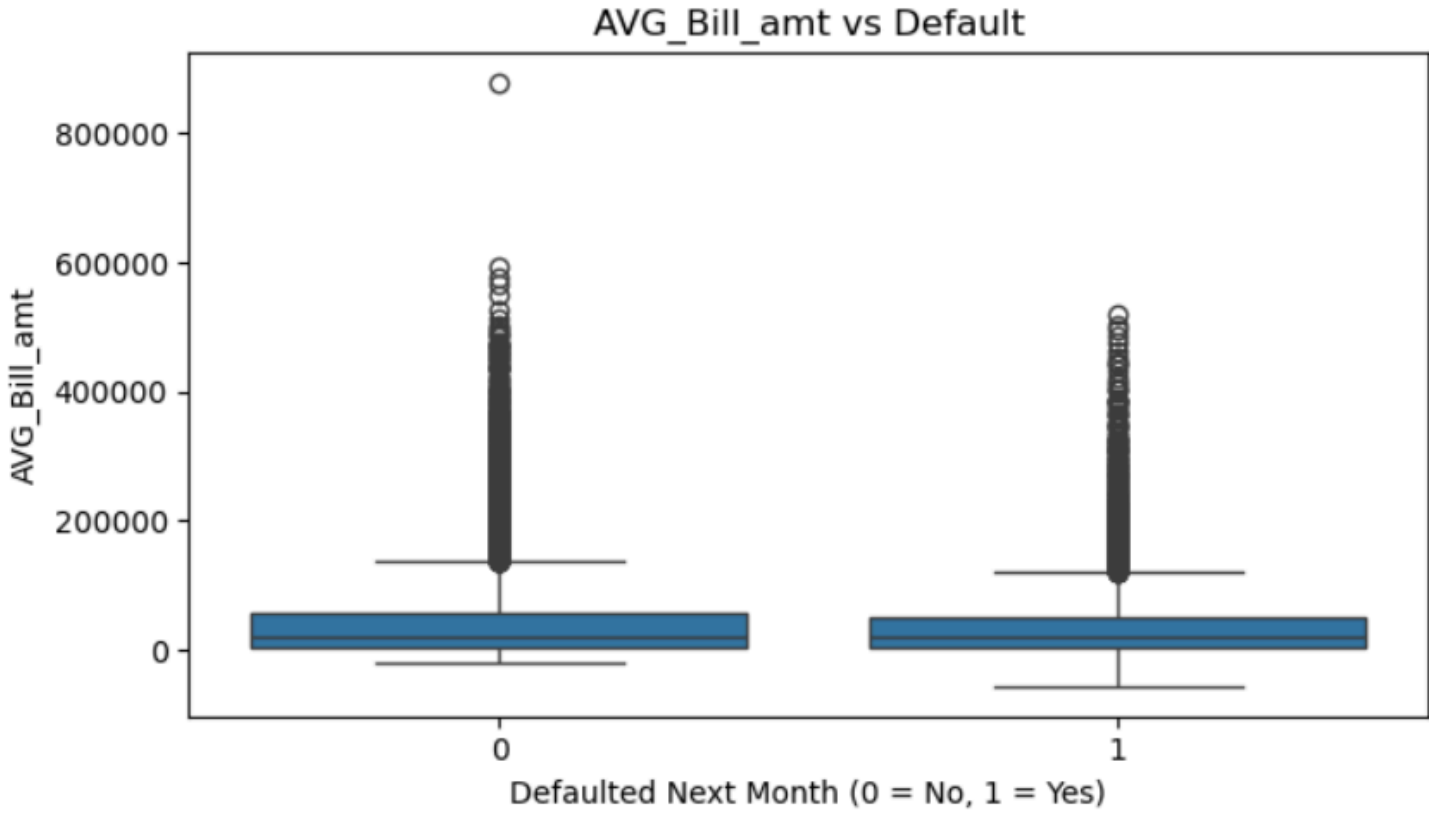




6. Average Bill Amount (AVG_BILL_AMT)

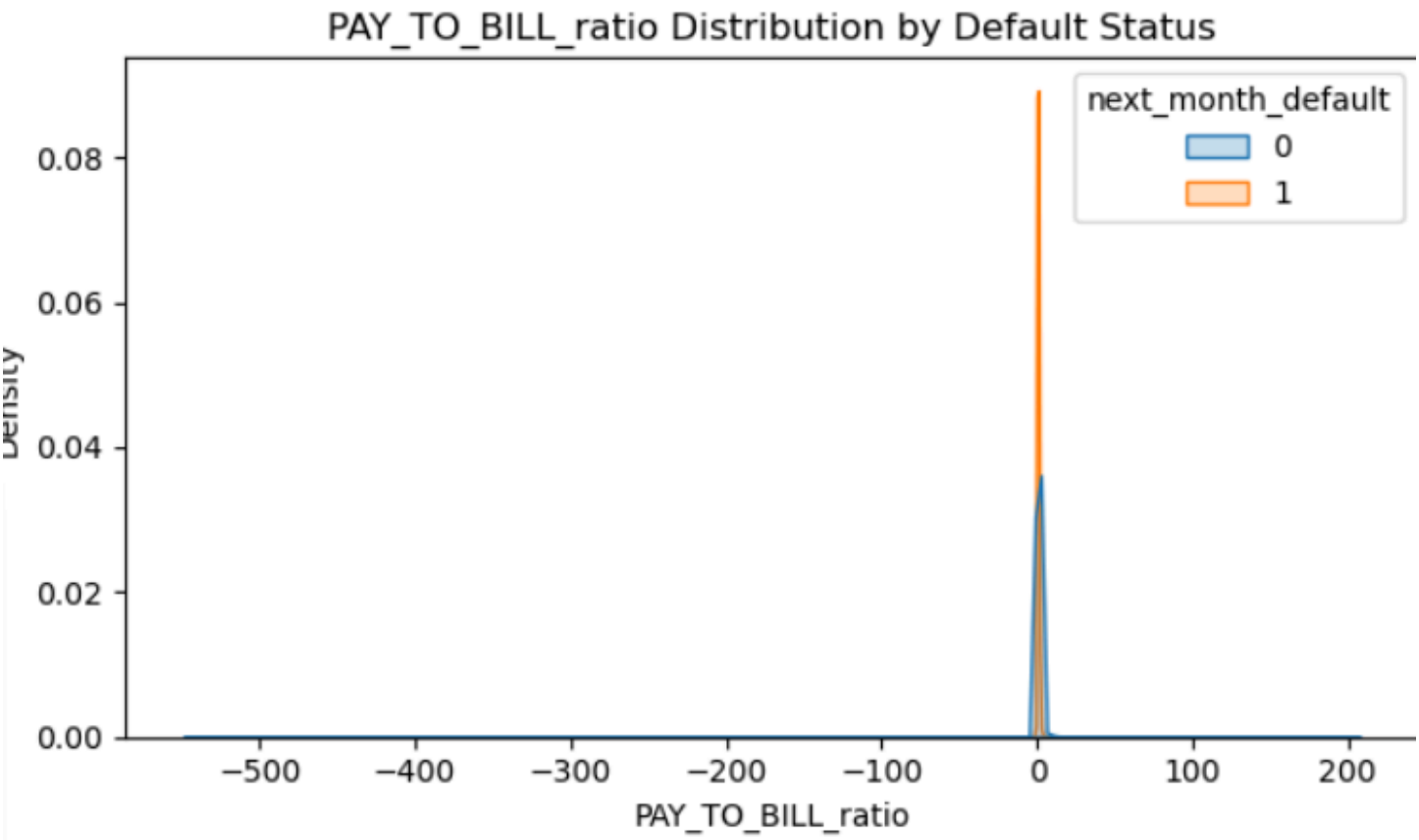
- **Observation:** Defaulters tend to have slightly higher average bills compared to non-defaulters while both defaulters and non-defaulters have similar distributions.
- **Interpretation:** Higher bill amounts might indicate overspending behavior, increasing the chance of default.

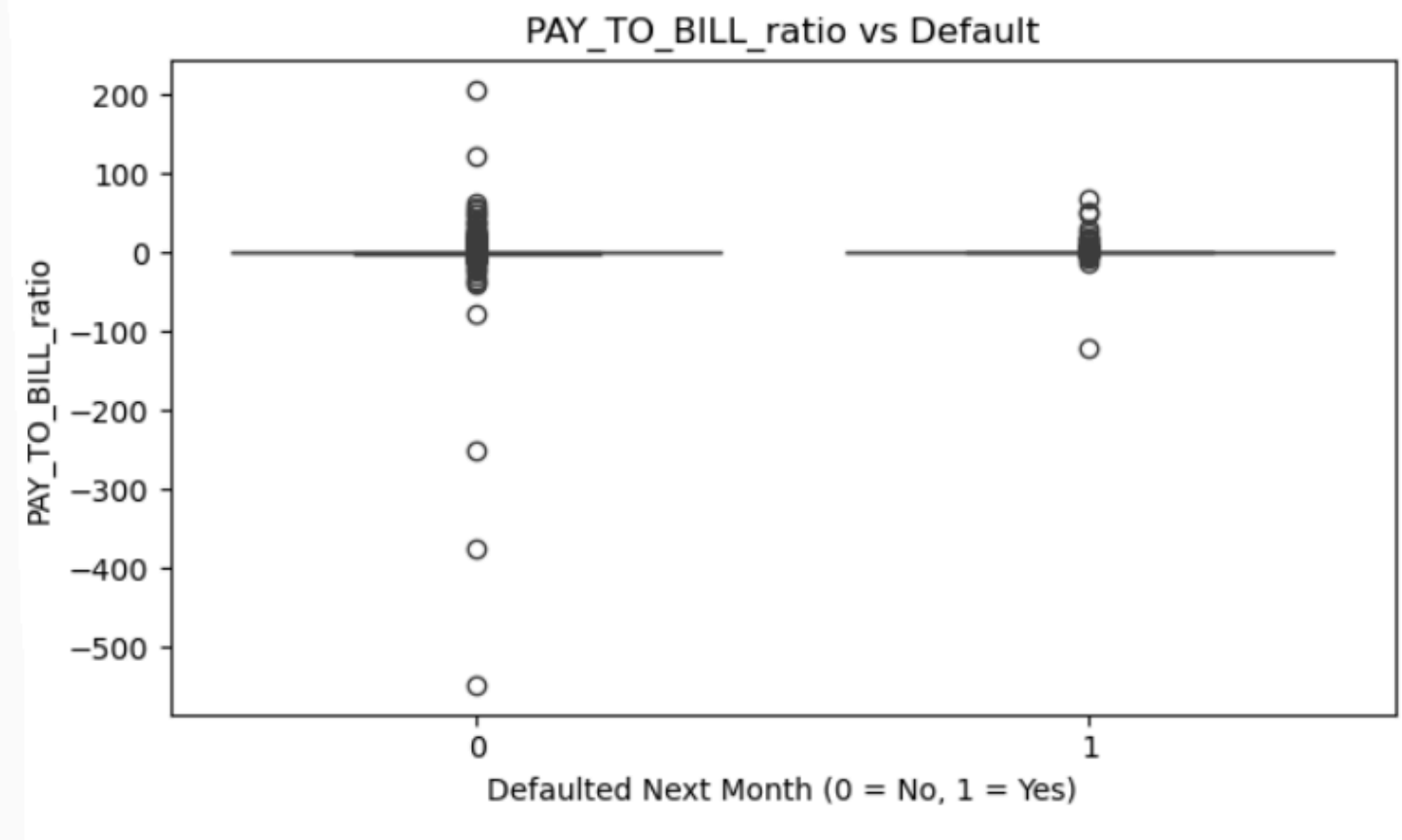




7. Repayment Ratio (PAY_TO_BILL_RATIO)

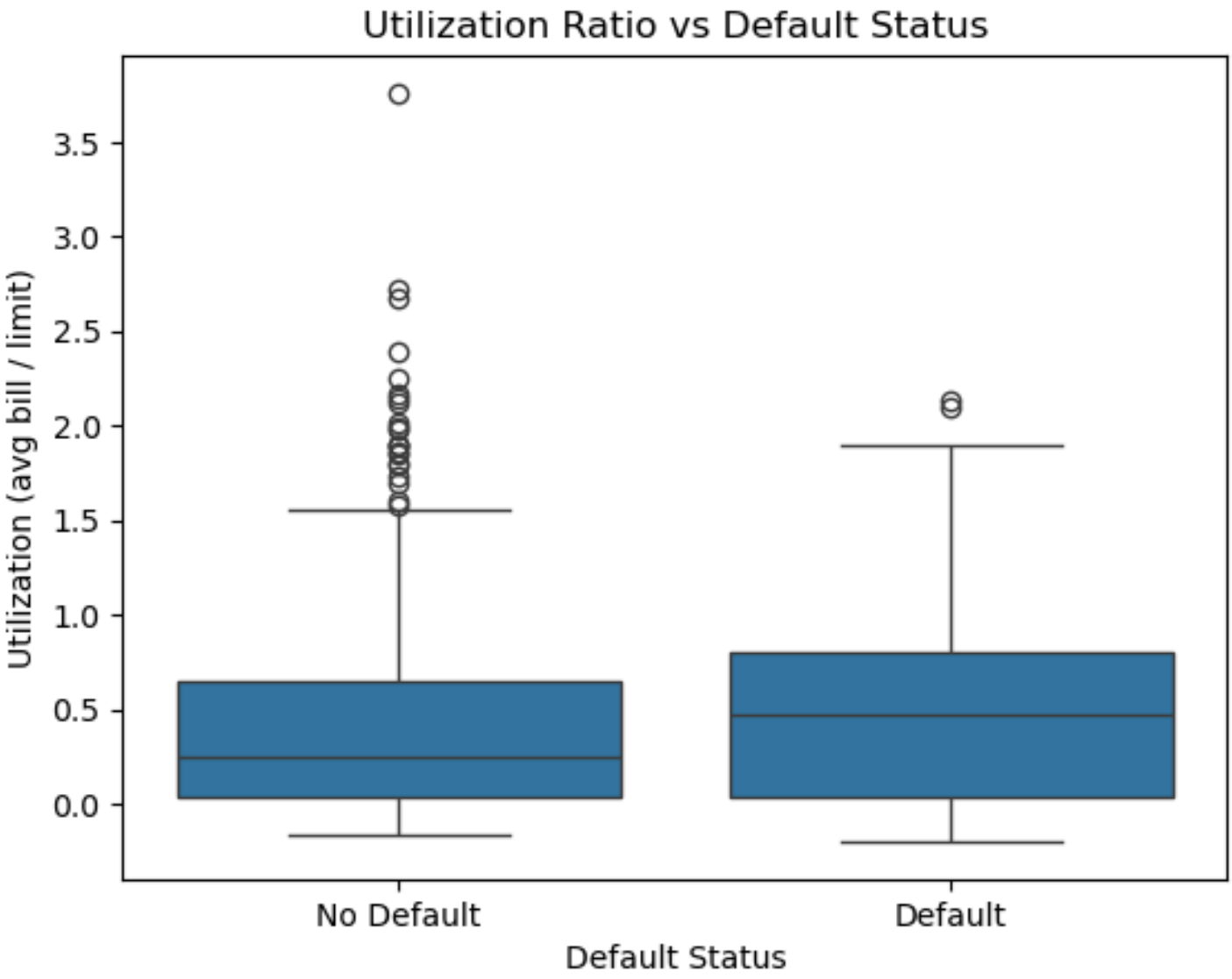
- **Observation:** Many defaulters have a repayment ratio near zero, indicating very low or no repayment activity.
- **Interpretation:** This is a major red flag. Poor repayment behavior is one of the most predictive indicators of default.





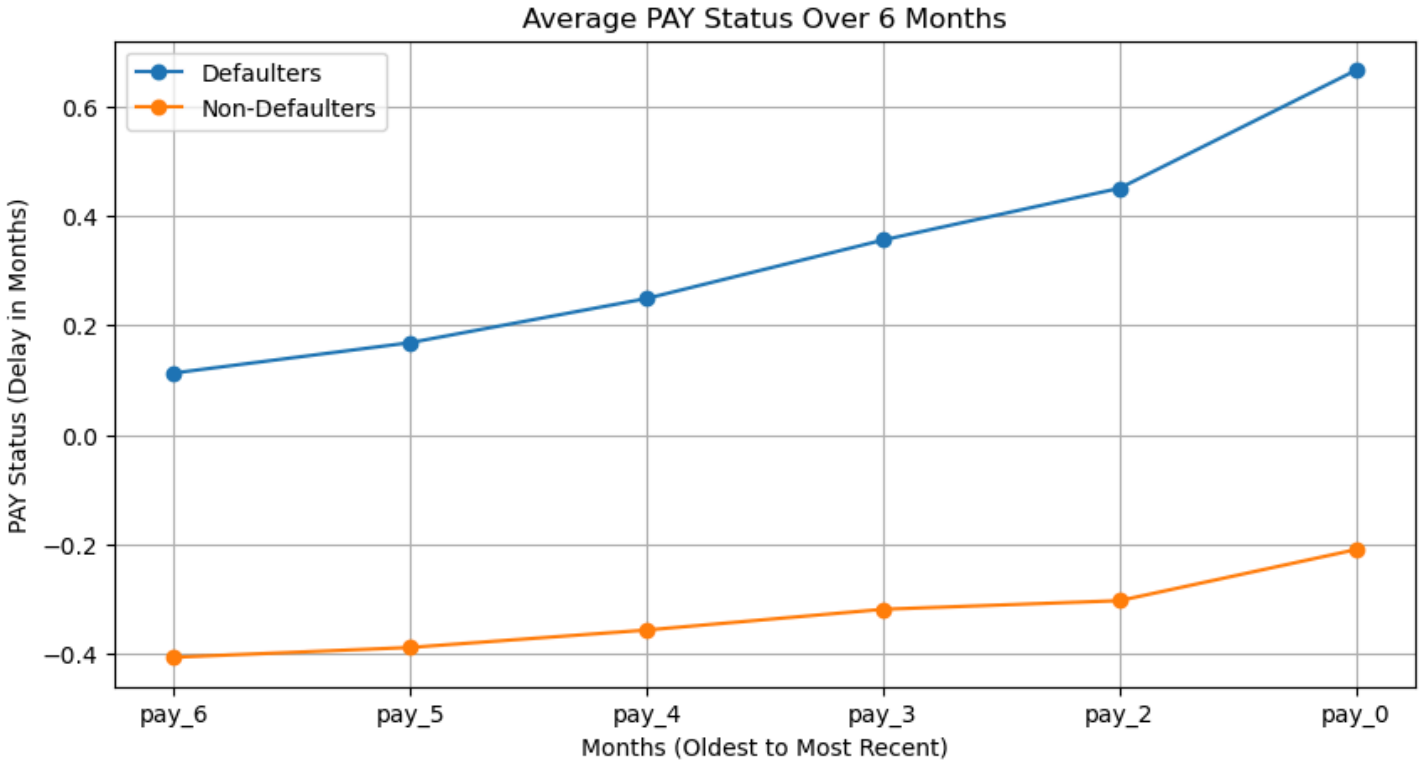
8. Over-limit Usage

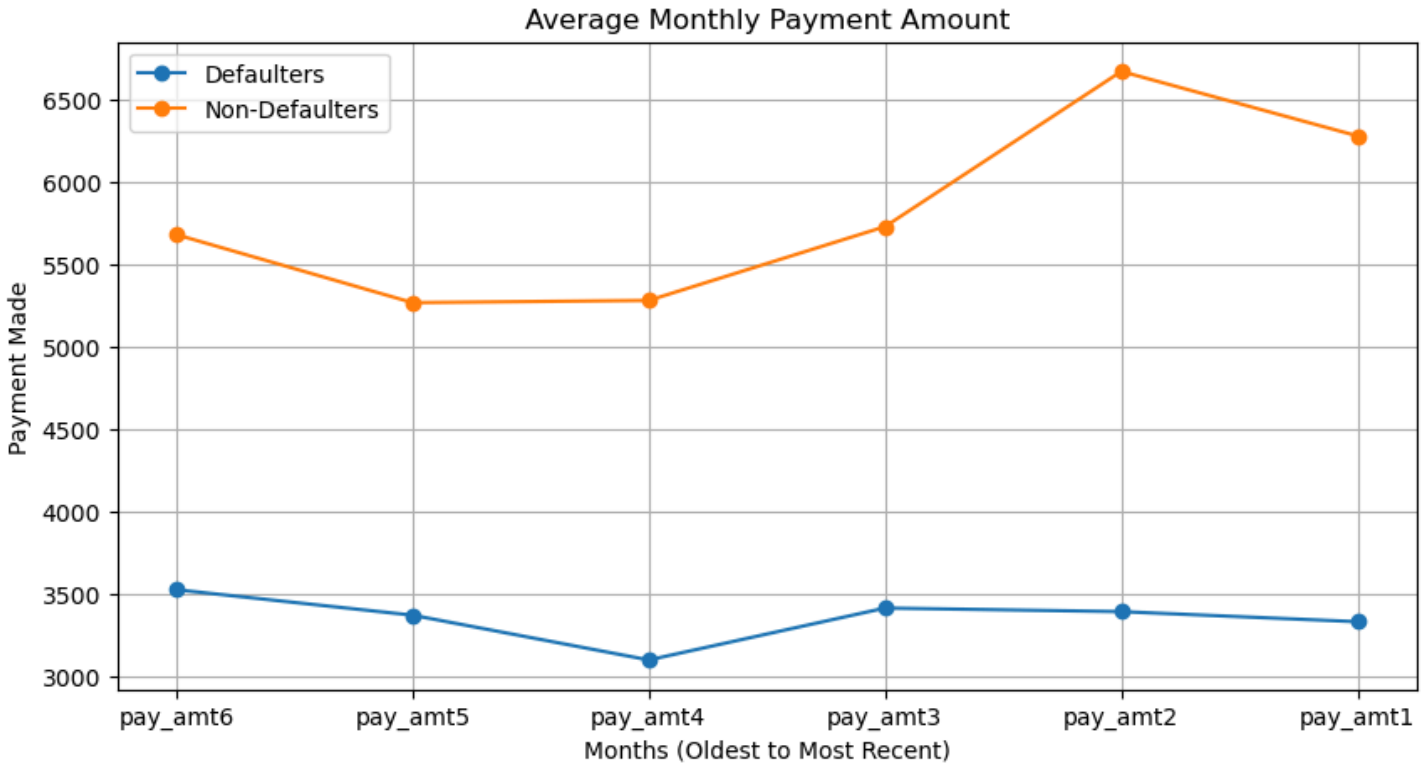
- **Observation:** Non-defaulters show more extreme outliers in over-limit usage but still avoid defaulting.
- **Interpretation:** Over-limit usage alone isn't predictive. Financial recovery ability plays a crucial role.



9. Payment Delay History (PAY_0) and amount of payment.

- **Observation:** Customers with PAY_0 > 1 show a default rate exceeding 38% and defaulters show a clear pattern of escalating payment delays - this is a classic early warning signal.
- **Interpretation:** Escalating payment delays are clear early warning signals for default and defaulters likely making only minimum payments, indicating cash flow stress.

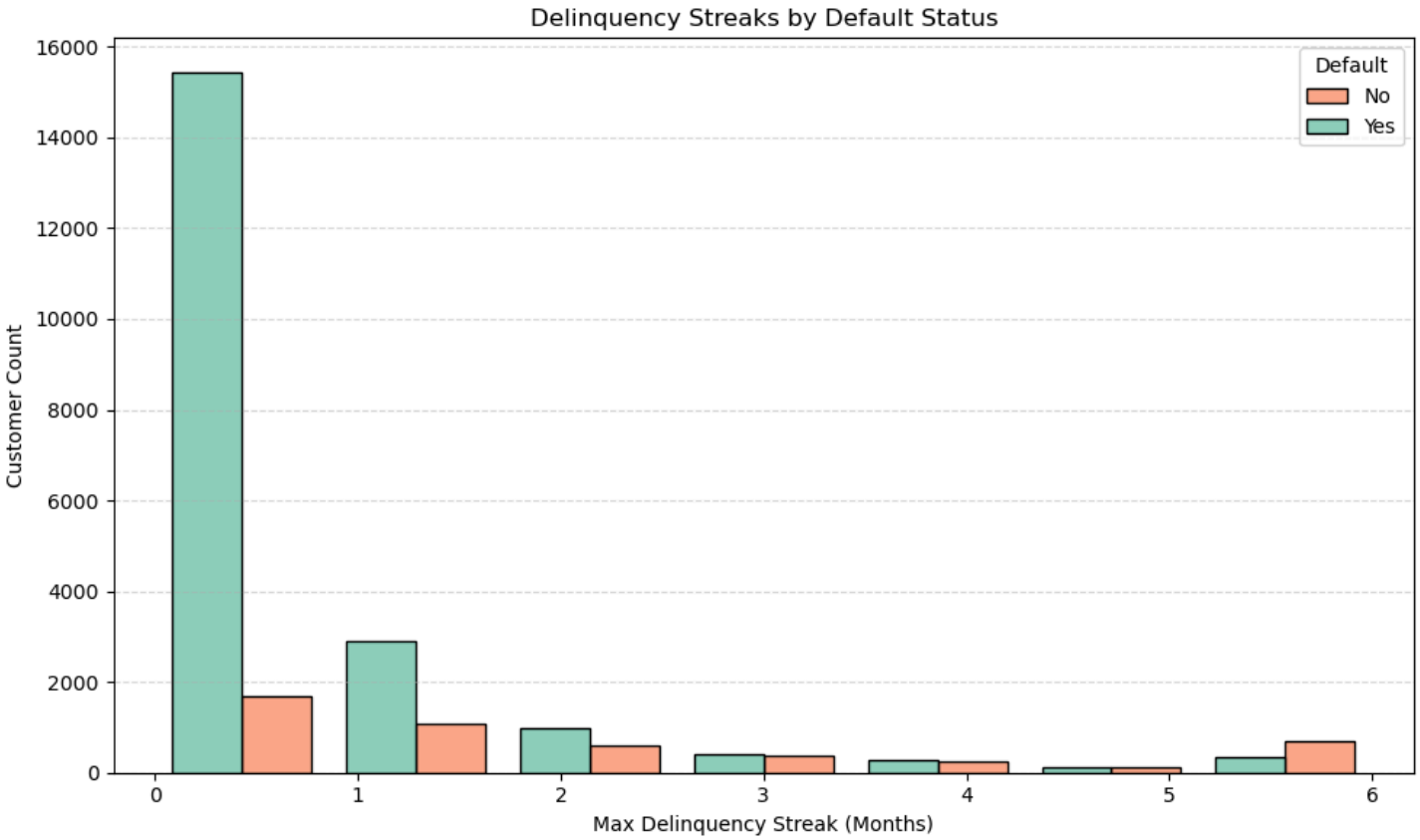




10. Default Rates by Max Delinquency Streak:

- **Observation:**
  Customers with 0-month delinquency streak have a default rate of 13.2%, whereas those with a 1–2 month streak show a significantly higher default rate of 49.6%. The default rate peaks at 59.2% for customers with 4 or more months of delinquency.
- **Interpretation:**
  There is a clear upward trend in default probability with increasing delinquency history.

This suggests that "max delinquency streak" is a strong predictor of default risk. Customers with a longer history of delayed payments are much more likely to default in the upcoming month, making it a critical feature for risk modeling.



Key financial insights:

1. Lower credit limits and repayment ratios are significant predictors of default.
2. Over-limit usage alone isn't predictive - it's the customer's ability to recover and manage these situations that determines default risk.
3. Consistently low payment amounts combined with delays signal high default risk.

Model comparison and justification for final selection.

1. Logistic Regression

- **Summary**: Logistic Regression offered a fast and interpretable baseline. However, its assumption of linearity limited its ability to capture complex relationships.

- **Limitation**: Despite decent recall (51%), it missed 49% of defaulters, leading to high Type II errors, which is risky in credit risk applications.
- **Justification**: It is not suitable as the final model due to poor performance on critical metrics like F2 Score and recall.

2. Decision Tree

- **Summary**: Highly interpretable and capable of modeling non-linearities, Decision Trees allow for simple rule-based insights.
- **Limitation**: The model overfit the training data and produced the worst recall (32%).
- **Justification**: Due to high variance and underperformance on defaulter detection, the model has limited practical use in this context despite its transparency.

3. Random Forest

- **Summary**: This ensemble model improved stability and reduced overfitting.
- **Limitation**: Despite a higher precision (56%), recall remained low (38%), making it less effective for minimizing false negatives.
- **Justification**: A moderate choice, but its increased computational cost..

4. XGBoost

- **Summary**: A powerful boosting algorithm delivering high accuracy and **59% precision**, XGBoost consistently outperforms simpler models.
- **Limitation**: The model's recall was low (32%), and hyperparameter tuning was more complex.
- **Justification**: Strong contender, especially for precision-sensitive applications, but not ideal when recall and F2-score are priority, as in credit risk.

5. LightGBM

- **Summary**: LightGBM provided the best trade-off across all metrics, including highest accuracy (0.84), highest precision (60%), and best Weighted F2 Score (0.8280).
- **Limitation**: Recall (34%) is still suboptimal, but better than other models in this category. Some interpretability is sacrificed for performance.
- **Justification**: It offers the best balance between identifying defaulters, minimizing false negatives, and maintaining reasonable interpretability.

Final Verdict

LightGBM is selected as the final model due to its top-ranked F2 score, balanced precision-recall profile, fast training speed, and ability to handle large datasets efficiently. While recall remains a challenge across all models, LightGBM offers the most reliable trade-off for business impact in default prediction scenarios.

Model	Accuracy	Precision (Class 1)	Recall (Class 1)	Type I error	Type II error	Weighted F2 Score
Logistic Regression	0.75	0.37	0.51	12%	49%	0.7573
Decision Tree	0.82	0.51	0.32	7%	68%	0.8121
Random Forest	0.83	0.56	0.38	7%	62%	0.8247
XGBoost	0.83	0.59	0.32	5%	68%	0.8248
LightGBM	0.84	0.60	0.34	5%	66%	0.8280

Evaluation Methodology

1. Recall (Sensitivity) – High Priority

- Recall measures the proportion of actual defaulters (Class 1) that were correctly identified.
- In credit risk, false negatives (missed defaulters) are more costly than false positives. A model that misses a defaulter may result in granting credit to someone likely to default, directly impacting financial stability.
- Prioritized to minimize Type II errors, ensuring more defaulters are correctly flagged, even at the cost of some false alarms.

2. F2 Score – Primary Metric

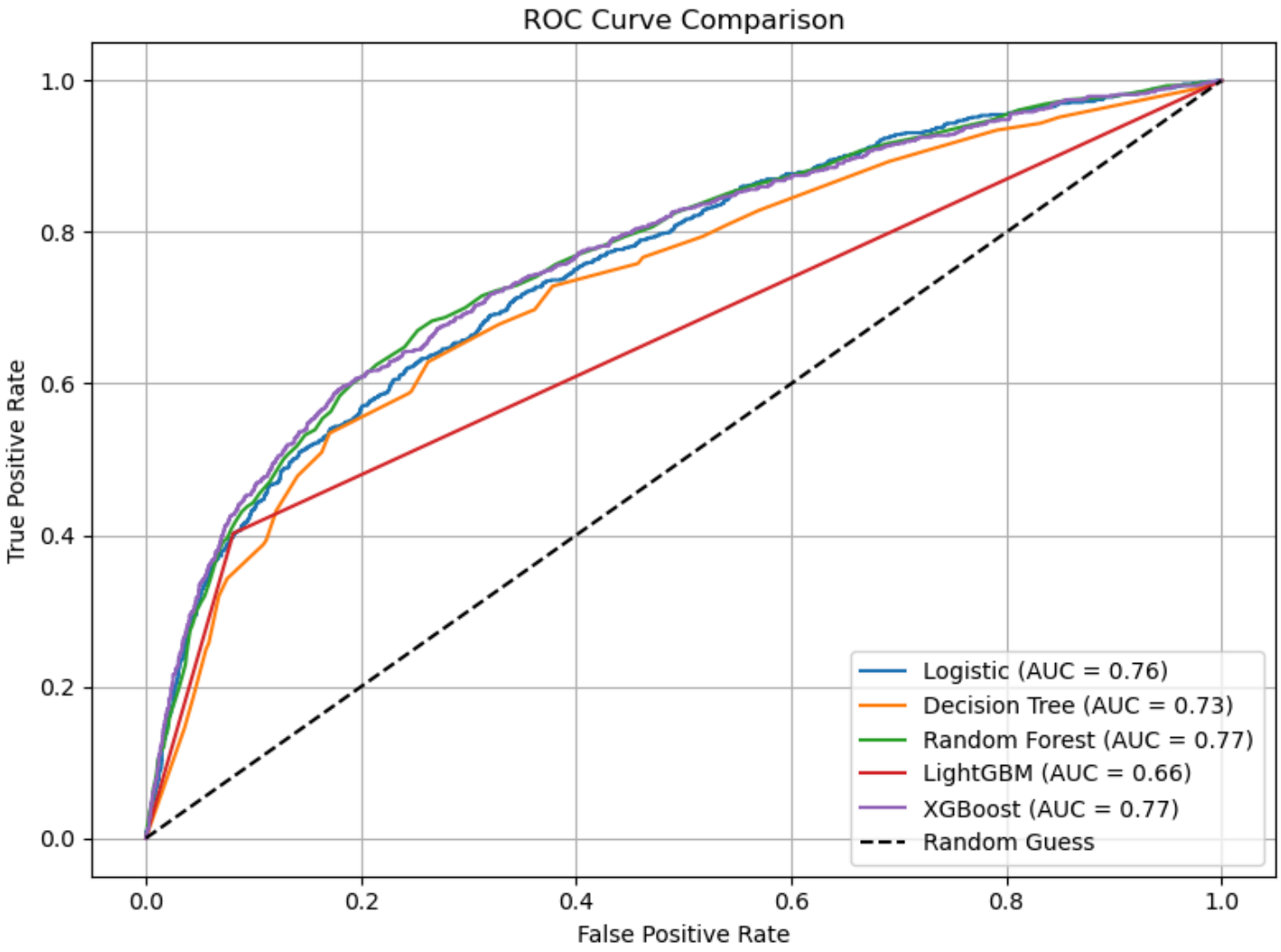
- F2 score is a weighted harmonic mean of precision and recall, placing more weight on recall (Recall is twice as important as Precision).
- The F2 score was used as the key selection metric, aligning with the business goal of catching as many defaulters as possible while still maintaining reasonable precision.

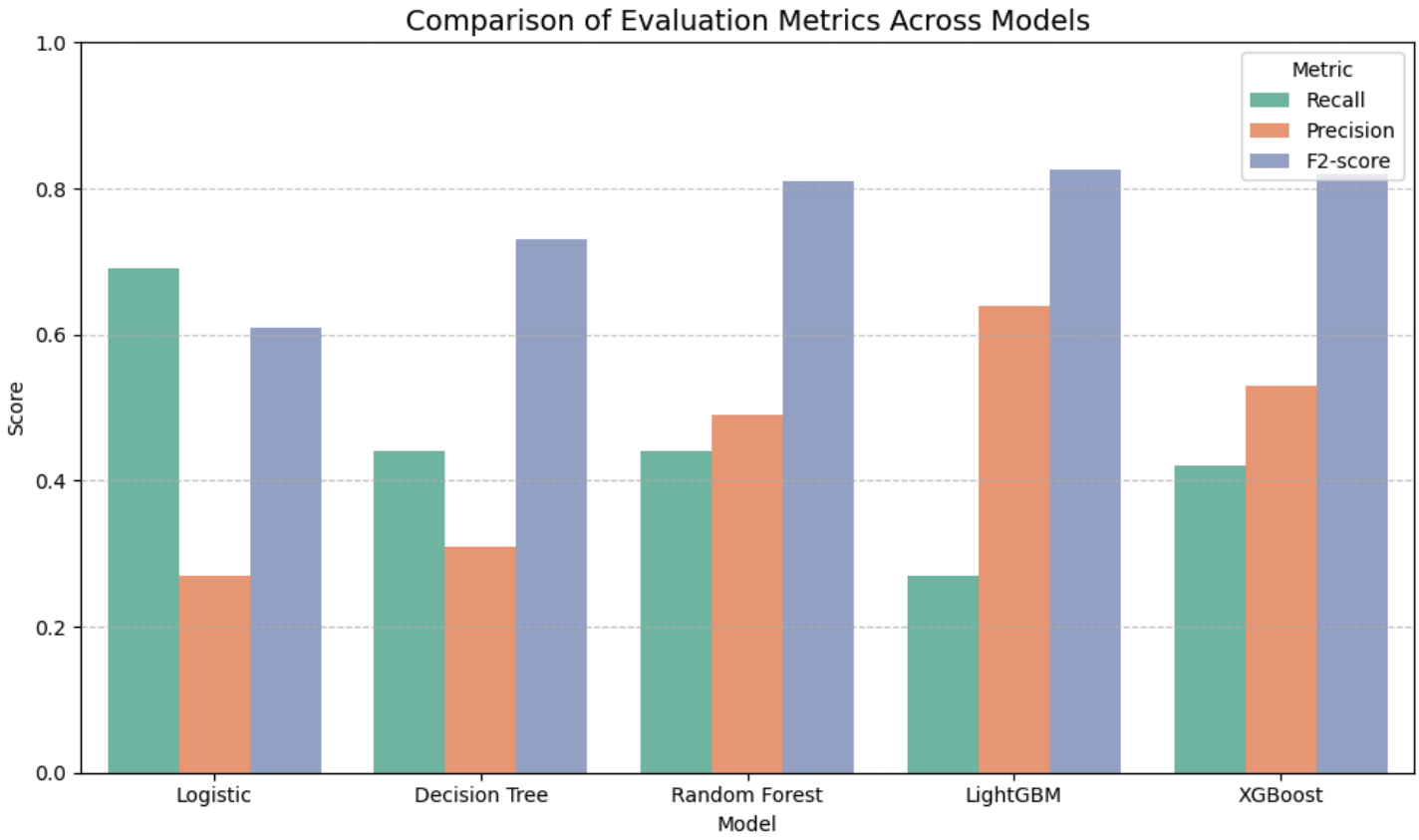
3. AUC-ROC Curve – Discrimination Power

- AUC (Area Under the Curve) measures how well the model distinguishes between the classes (defaulter vs. non-defaulter) across thresholds.
- A robust model should have a high AUC, indicating strong discrimination ability irrespective of the threshold.

4. Confusion Matrix – Diagnostic Tool

- The confusion matrix shows the actual vs. predicted classifications, breaking down true positives, true negatives, false positives, and false negatives.
- Essential for understanding error types, especially when tuning thresholds to balance precision and recall.



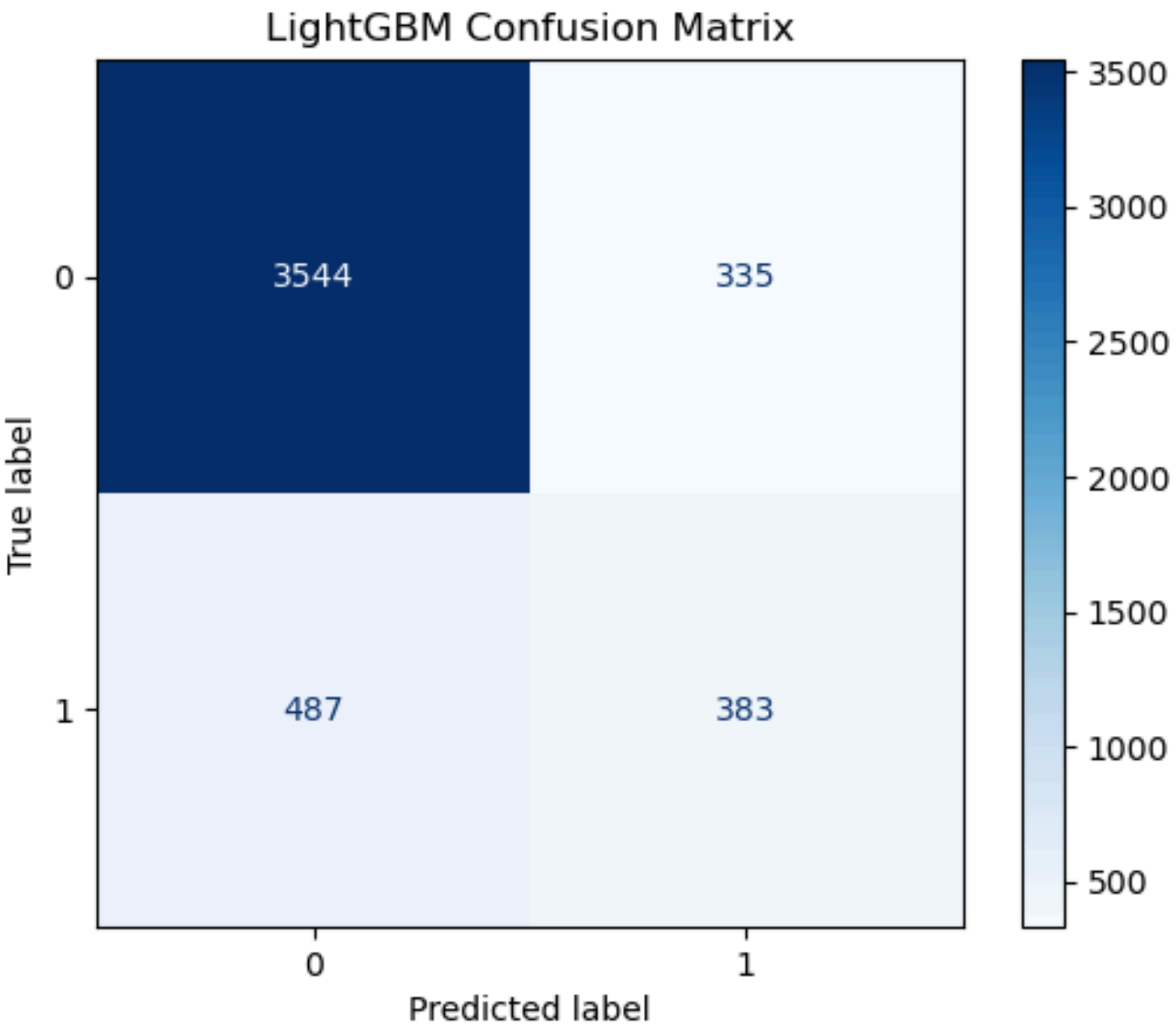


Metrics result on train dataset

To thoroughly assess the performance of the final selected model, LightGBM, multiple classification metrics were computed on the train/test split. The focus remained on minimizing false negatives by optimizing for recall and maximizing the weighted F2 score, which places greater emphasis on recall over precision.

Instead of relying on the default threshold of 0.5, a custom search was conducted across thresholds (0.1 to 0.99).

For each threshold, the weighted F2 score was computed to identify the point where the model achieves the best trade-off between high recall and acceptable precision, we got 0.6 as the best threshold.


LightGBM Confusion Matrix
3544
335
487
383
0
1
True label
0
1
Predicted label
3500
3000
2500
2000
1500
1000
500

```
Classification Report:

              precision    recall  f1-score   support

           0       0.86      0.95      0.90      3879
           1       0.60      0.34      0.43       870

    accuracy                           0.84      4749
   macro avg       0.73      0.64      0.67      4749
weighted avg       0.82      0.84      0.82      4749

Best Threshold: 0.60
Weighted F2 Score: 0.8280
```

Business implications

- **Delayed Payments as Early Warning:** Escalating delays in monthly repayments (e.g., PAY_0 > 1) signal elevated default risk (>38%), enabling proactive risk flags.
- **Minimum Payment Behavior:** Defaulters often make only minimal or zero payments, indicating cash flow stress and a strong likelihood of eventual default.
- **Repayment Ratio as Risk Signal:** Low PAY_TO_BILL ratios are prominent among defaulters, suggesting under-repayment even when bills are high.
- **Demographic Risk Pockets:** Younger and senior customers show higher default tendencies, calling for targeted financial education and credit counseling.
- **Marital Status Risk:** Married individuals have a slightly higher default probability (20.37%) than singles.

Summary of Findings and Key Learnings

- **LightGBM Selected as Best Model**: Achieved highest weighted F2 score (0.8280) and strong overall precision–recall balance, making it optimal for reducing false negatives in credit risk.
- **Threshold Tuning Boosted F2**: Customized threshold (0.60) significantly improved recall for defaulters while maintaining overall model stability.

- **Zero Payments Common Across Classes**: Months with zero repayments were frequent in both classes, hence not a reliable default indicator alone.
- **Credit Limits & Defaults Correlated**: Defaulters typically had lower credit limits, linking risk to initial creditworthiness.
- **Over-limit Usage Not Sole Indicator**: High usage didn't necessarily lead to default — recovery behavior and financial management matter more.