Loan Default Risk Analysis

(Dataset : Loan Dataset)

(Last updated : 13-July-2025)

Project’s Agenda

- Dataset Overview
- Conduct Exploratory Data Analysis (EDA)
- Visualization
- Chi^2 test & Cramér’s V Test
- Loan Prediction
- Conclusions

Dataset Overview

01

Total Records:

32,586

02

Target Variable:

loan_status_clean

✓ Non-Default: 25,586 (~79%)

✗ Default: 6,819 (~21%)

03

Features:

Numerical: age, income, loan amount, interest rate, employment years

Categorical: home ownership, loan intent, loan grade, etc

Exploratory Data Analysis (EDA)

01

Target Variable Distribution:

Default: 6,819 (≈ 21%)

Non-Default: 25,589 (≈ 79%)

02

Distribution Checks:

customer_age, customer income, loan_amount, loan_int_rate.

Detected right skew in all except interest rate(normal)

03

Data Types:

Categorical: home ownership, loan intent, loan grade.

Numerical: age, income, interest rate, loan amount

04

Outliers detected:

In customer_income, loan_amnt, and customer_age

05

Initial Patterns Identified:

Younger borrowers = slightly higher default rate

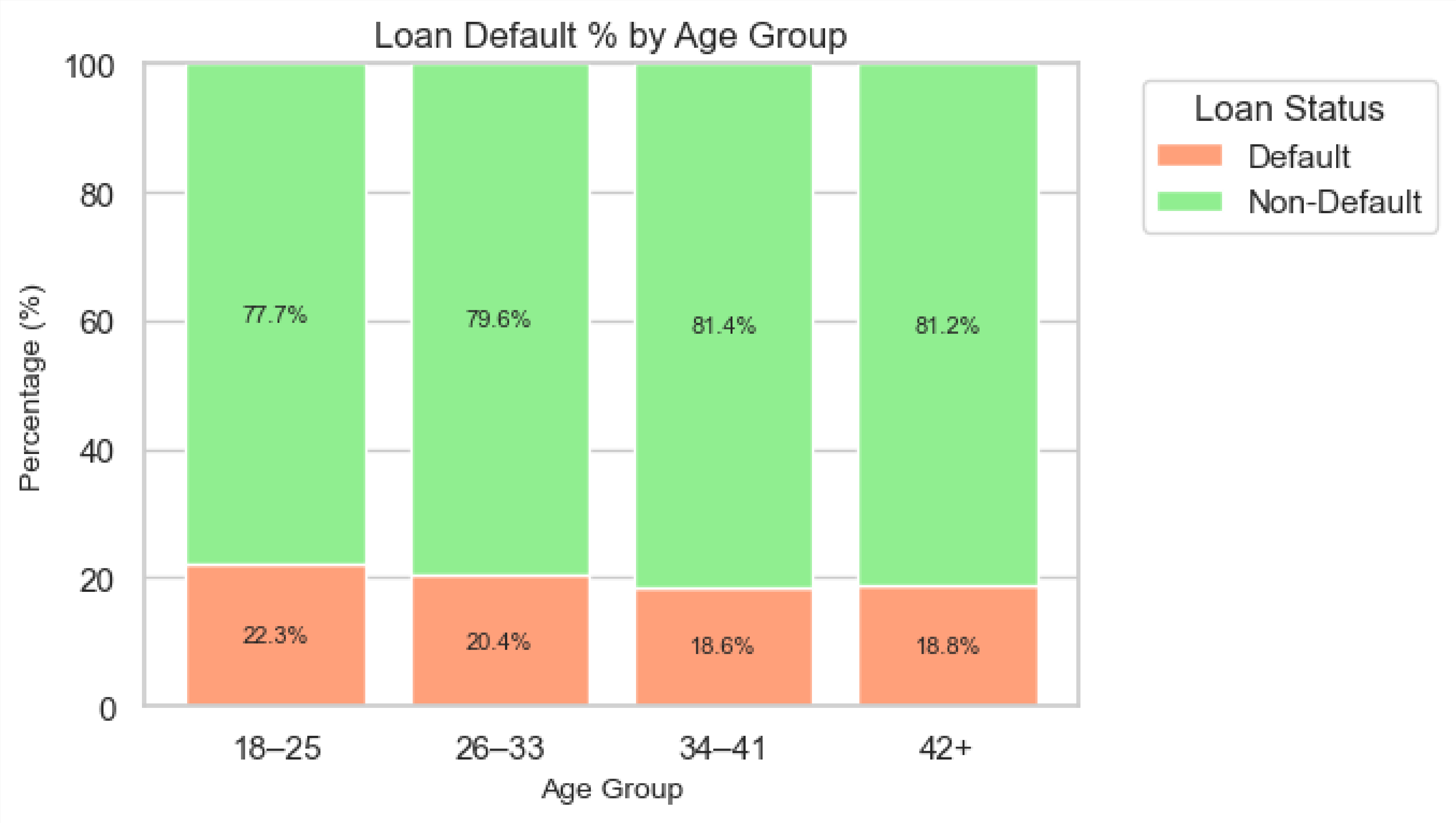
High income = lower risk of default

Larger loan amounts (20k+) associated with higher default

Loan Purpose affects risk — e.g., Medical & Debt Consolidation = high default

Visualization

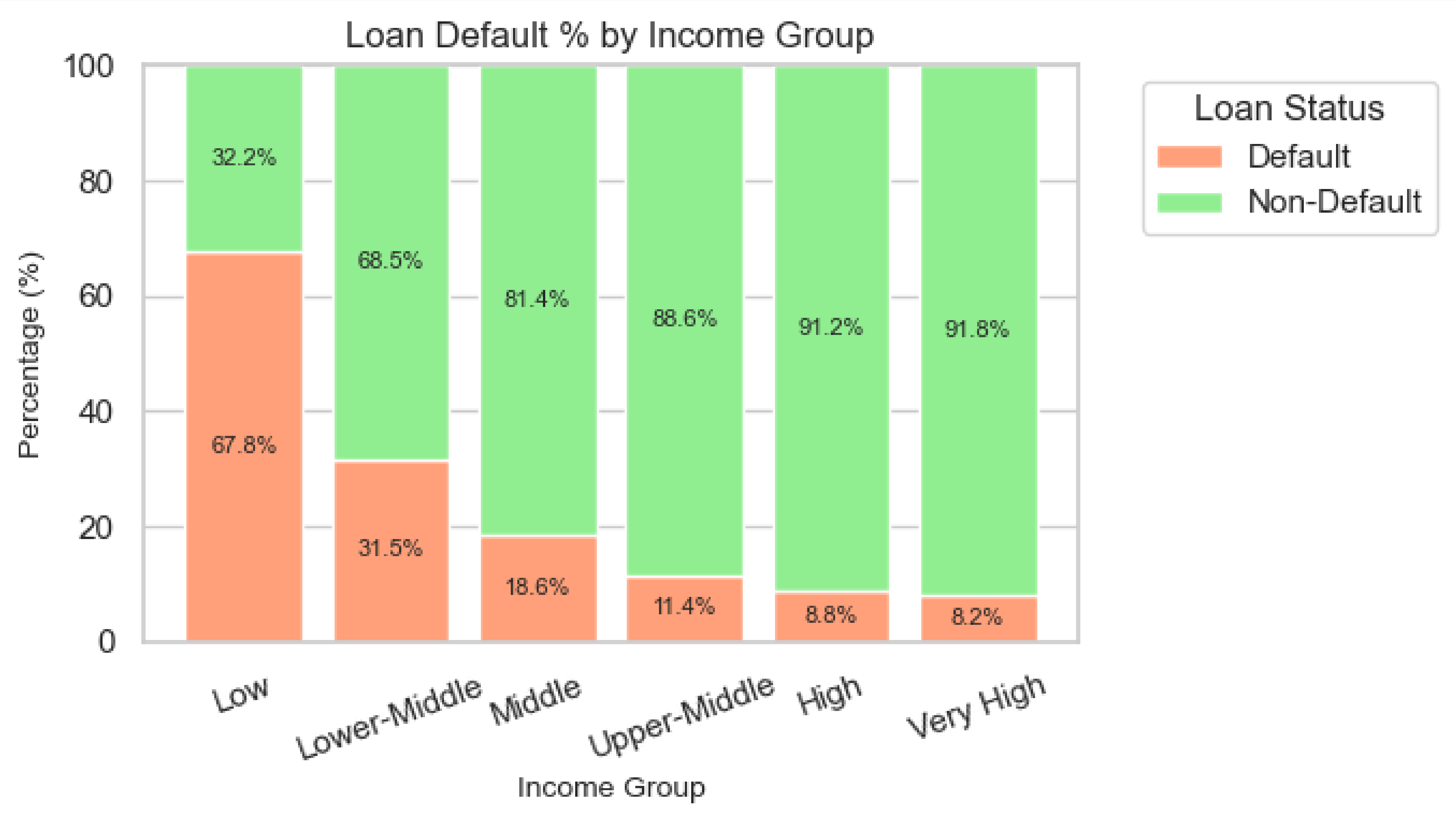
Age Group vs Loan Status



Insight:

- Default rate slightly decreases as age increases — younger borrowers (18–25) show highest risk.
- Borrowers aged 34+ are more stable — lowest default rates and better repayment behavior.

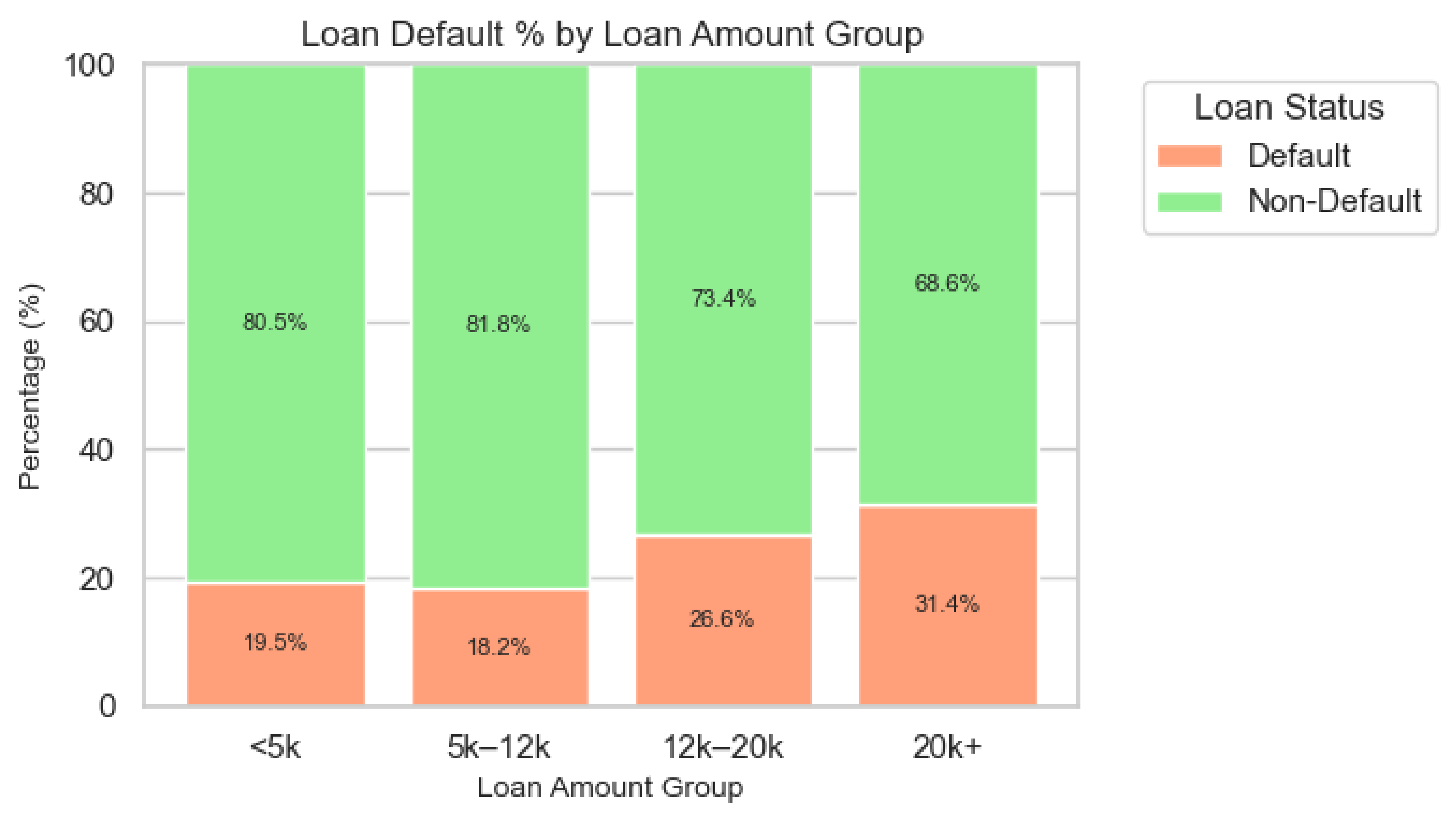
Income Group vs Loan Status



Insight:

- Default rate drops steadily as income increases.
- Low-income borrowers show highest defaults — financial stress.

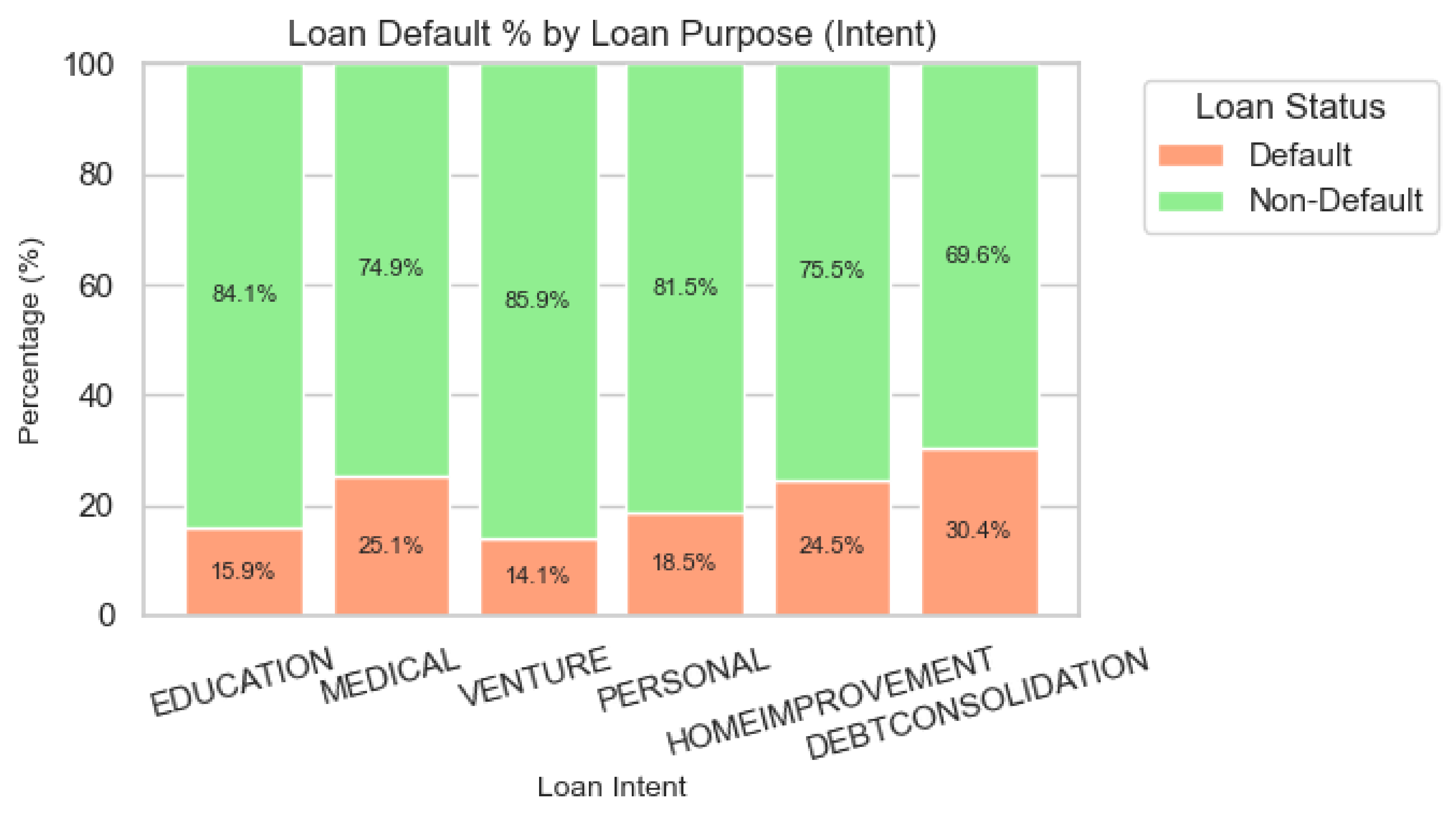
Loan Amount Group vs Loan Status



Insight:

- Up to 12k: Default rate stays low and steady — borrowers likely manage repayment well.
- Above 12k: Default risk rises quickly — larger loans bring more pressure and missed payments.

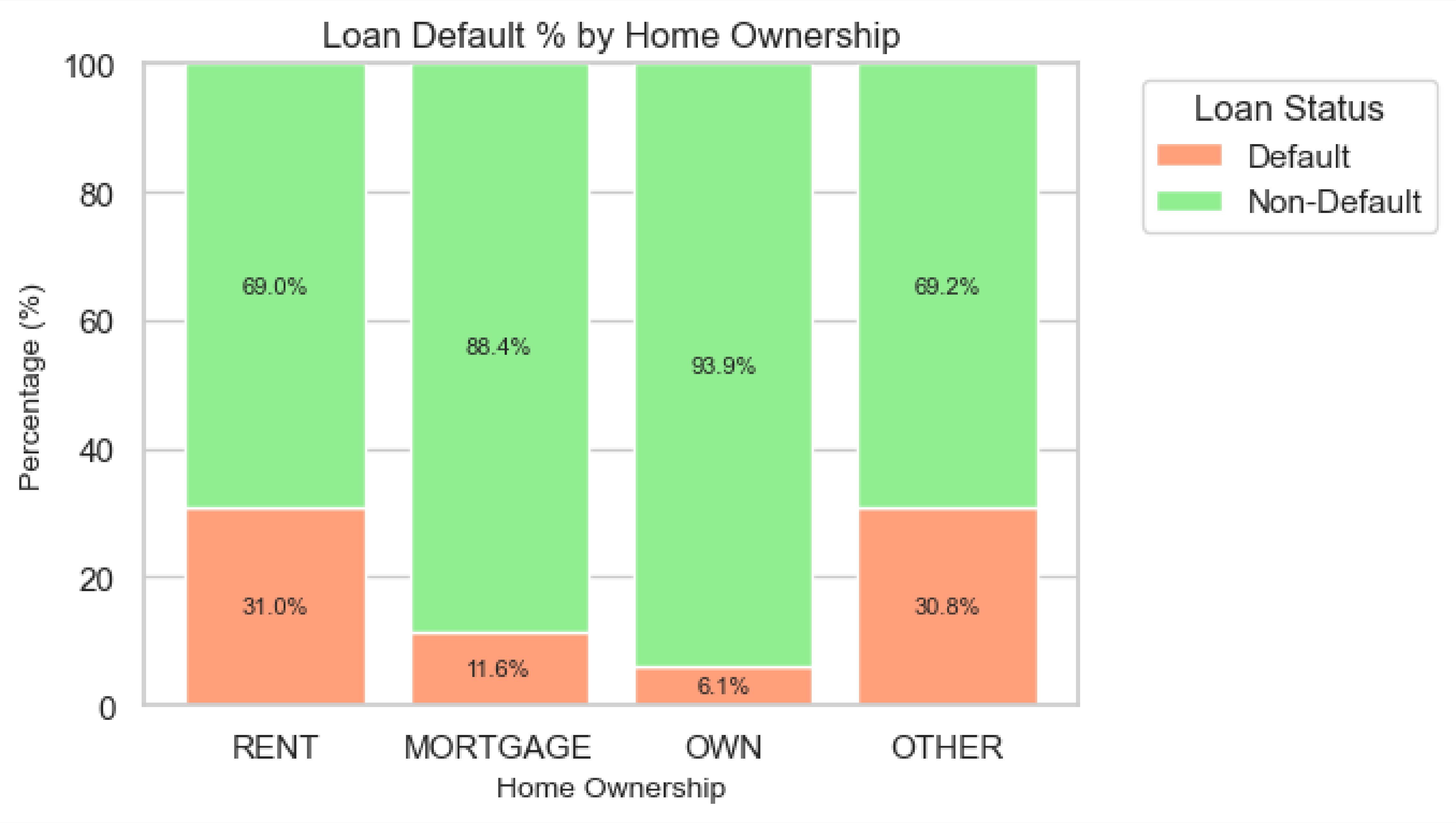
Loan Intent vs Loan Status



Insight:

- Lowest default rates in Education and Venture loans — borrowers tend to repay better.
- Highest risk seen in Debt Consolidation, Medical, and Home Improvement — likely tied to urgent or unstable financial situations.

Home Ownership vs Loan Status

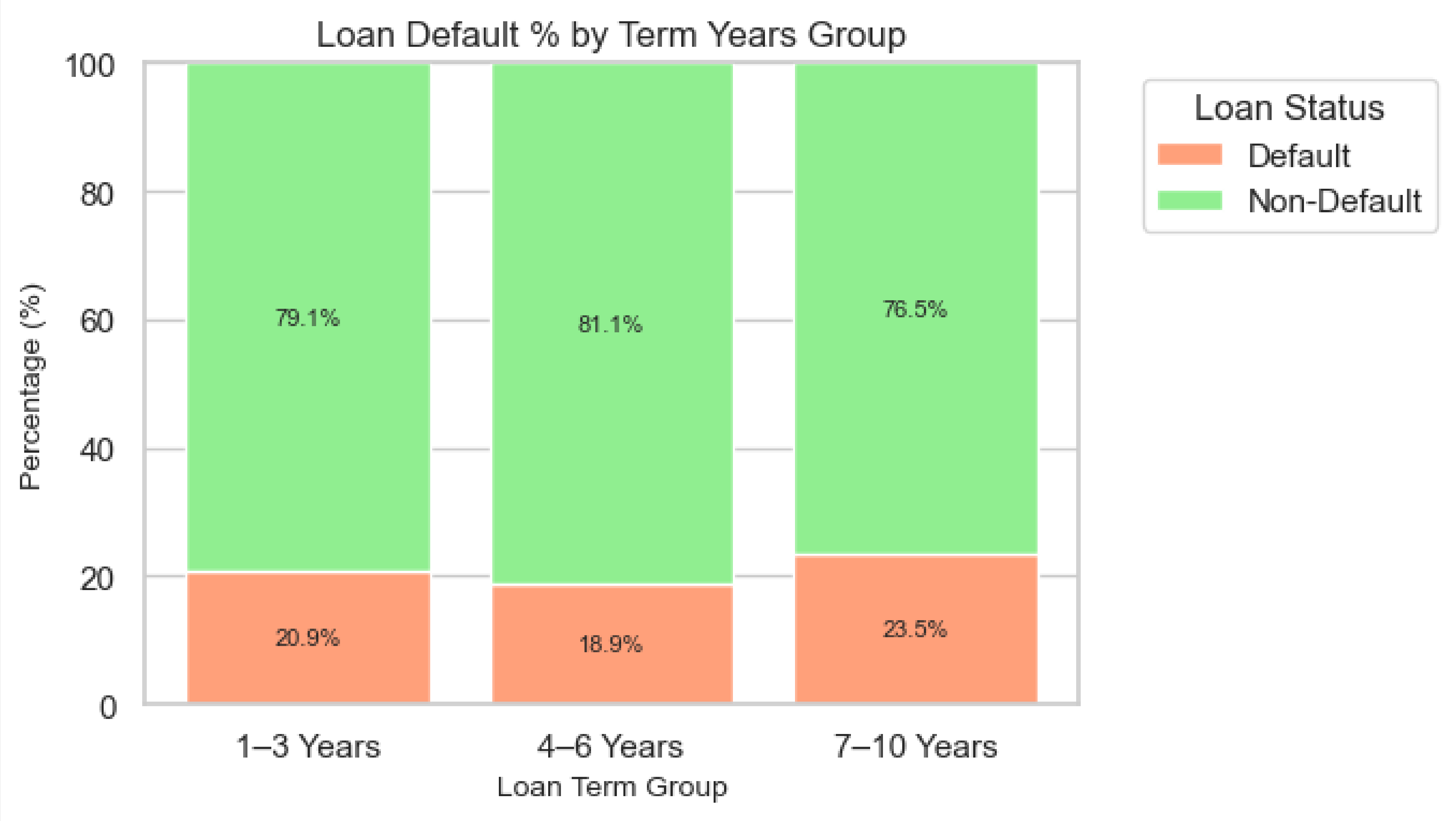


Insight:

- Own & Mortgage holders show the lowest default rates — stable living situations help repayment.

- Renters & Others face higher risk — financial uncertainty may impact loan reliability.

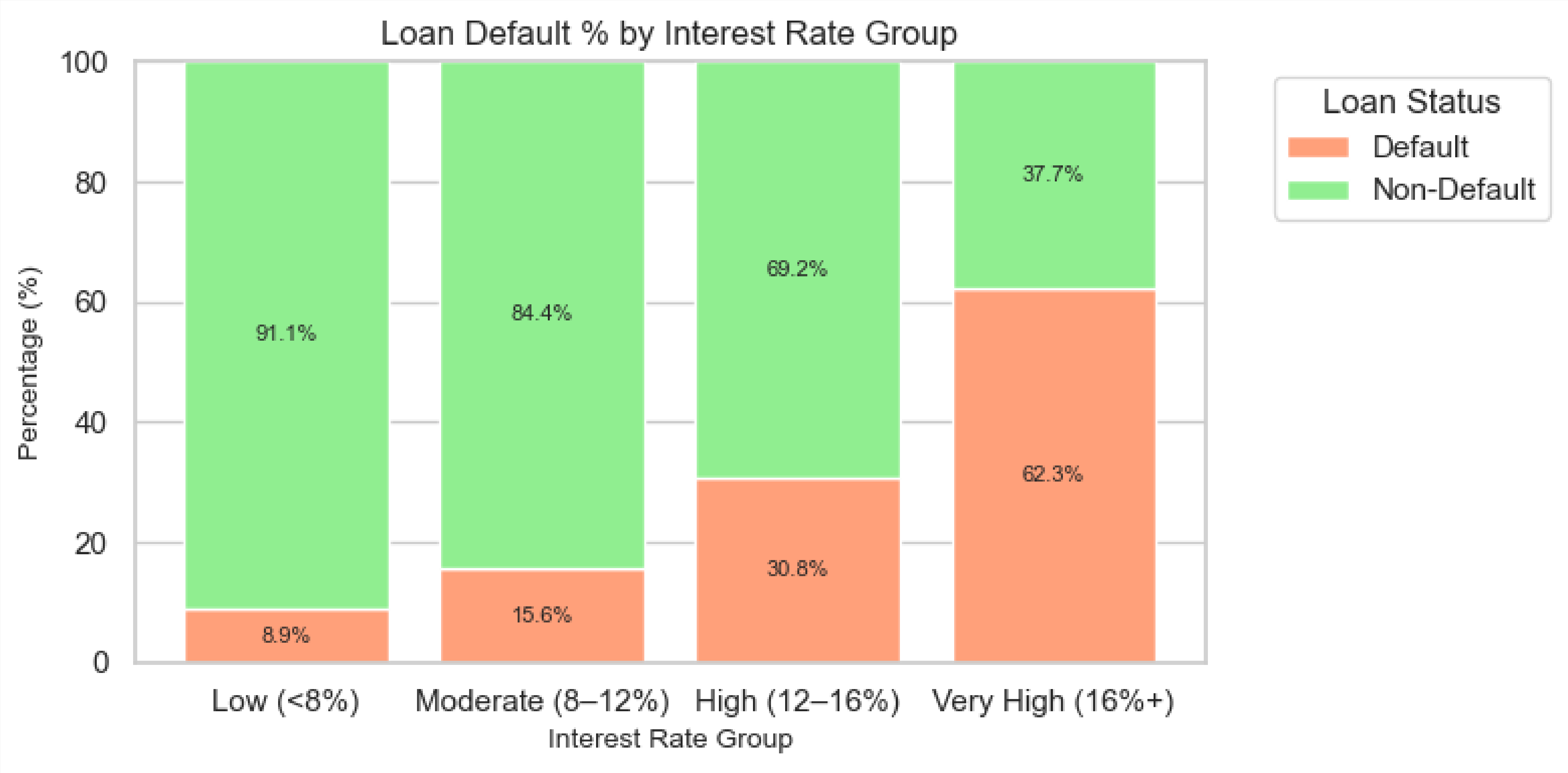
Term Years Group vs Loan Status



Insight:

- 4–6 year loans have the lowest default rate — repayment feels more balanced.
- Short (1–3 yrs) and long (7–10 yrs) terms show higher defaults — either too rushed or stretched too long.

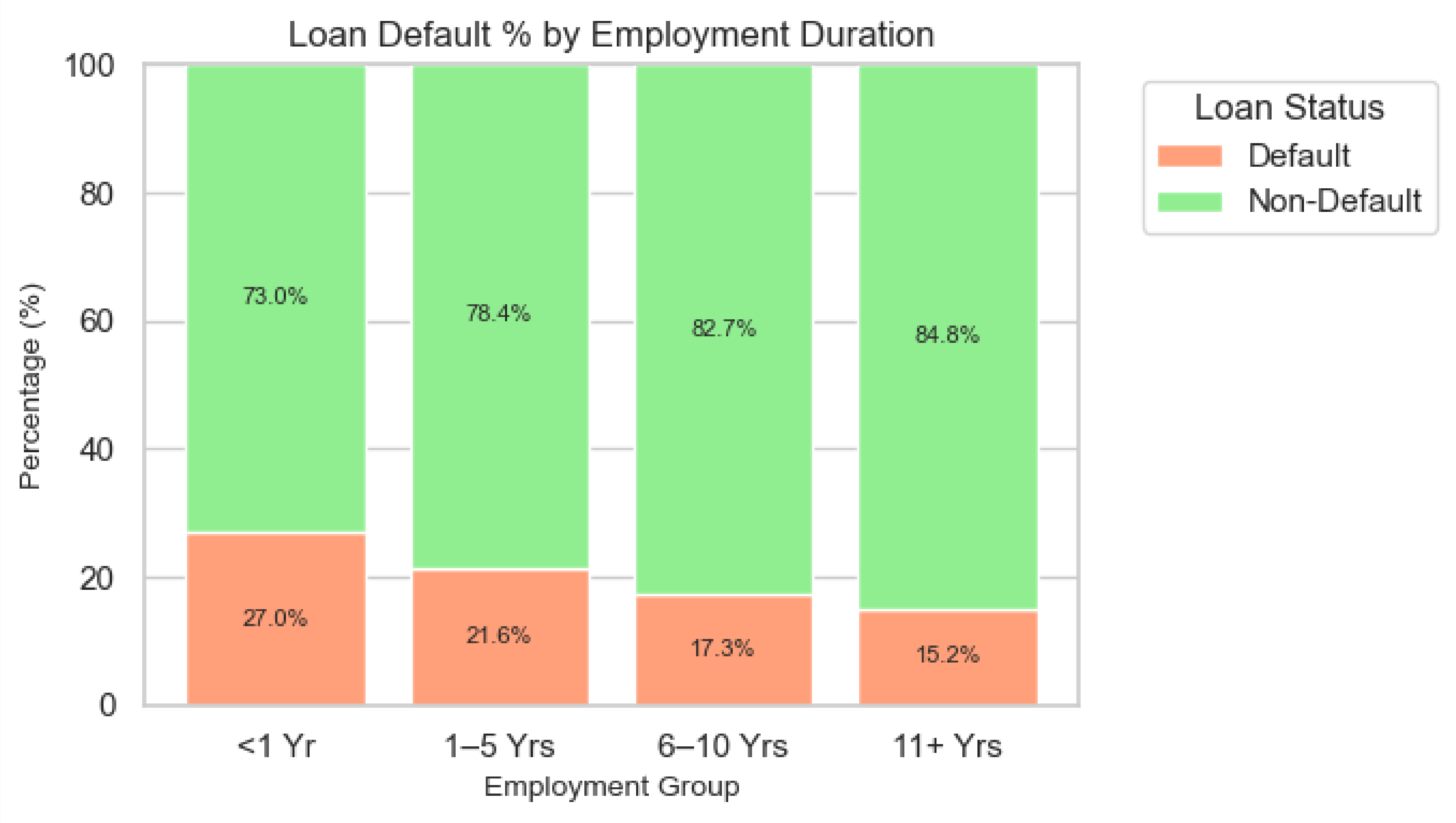
Interest Rate Group vs Loan Status



Insight:

- Borrowers with Very High interest rates are most likely to default.
- Suggests that higher rates may burden borrowers, especially those already flagged as risky

Employment Group vs Loan Status

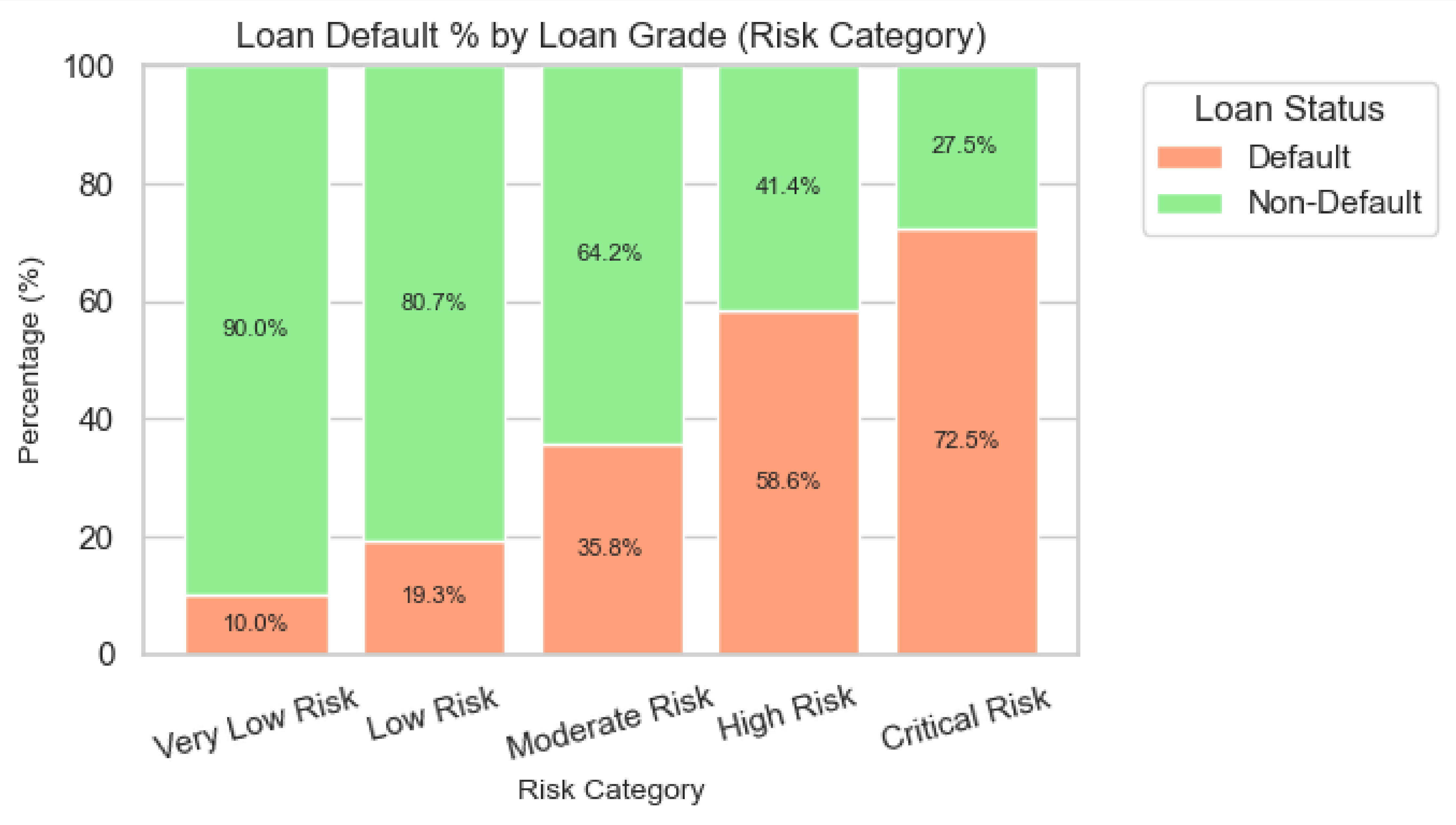


Insight:

- Default rate decrease as job duration increases — people with longer work history repay more reliably.

- Shortest employment group (<1 year) shows highest risk — stability matters in financial commitments

Loan Grade Named vs Loan Status



Insight:

- Default risk rises sharply from Very Low to Critical grades — high-risk loans see over 70% defaults.
- Safer lending happens in Very Low and Low grades — borrowers repay reliably.

Chi-square Test

🔍 **Purpose:**

To check whether two categorical variables are independent or associated.

Basic Idea:

It compares the observed frequencies with the expected frequencies in a contingency table.

- If p-value < 0.05 → Reject Null Hypothesis → Relationship exists
- If p-value ≥ 0.05 → Fail to reject Null → No relationship

Cramer’s V Test

🔍 **Purpose:**

After Chi-Square confirms the relationship, Cramér’s V tells how strong that relationship is (magnitude).

Basic Idea:

It uses the Chi-Square statistic and adjusts it to give a value between 0 and 1.

- 0.00 → No association
- 0.01 – 0.10 → Weak association
- 0.10 – 0.30 → Moderate association
- 0.3–0.50 → Strong association
- >0.5 → Very strong association

Chi-Square Test & Cramér’s V Test Results

Feature	p-value	Cramér’s V	Association with Loan Status	Interpretation
Age Group	< 0.0001	0.032	Weak Association	Age has minimal impact on loan default; not a strong predictor.
Income Group	< 0.0001	0.293	Moderate Association	Borrower income is moderately associated with default behavior.
Loan Amount Group	< 0.0001	0.094	Weak Association	Loan amount has weak influence on default probability.
Loan Intent Group	< 0.0001	0.142	Moderate Association	Loan purpose moderately affects chances of default.
Loan Grade Named	< 0.0001	0.373	Strong Association	Loan grade is a strong predictor of loan default risk.
Home Ownership	< 0.0001	0.251	Moderate Association	Home ownership status shows strong relation to repayment behavior.
Employment Group	< 0.0001	0.093	Weak Association	Employment duration shows limited effect on default.
Term (Years) Group	< 0.0001	0.044	Weak Association	Loan term length has a weak association with default status.
Loan Interest Rate Group	< 0.0001	0.323	Strong Association	Higher interest rates strongly relate to increased default risk.

Loan Prediction

Model Selection

Two models used for classification:

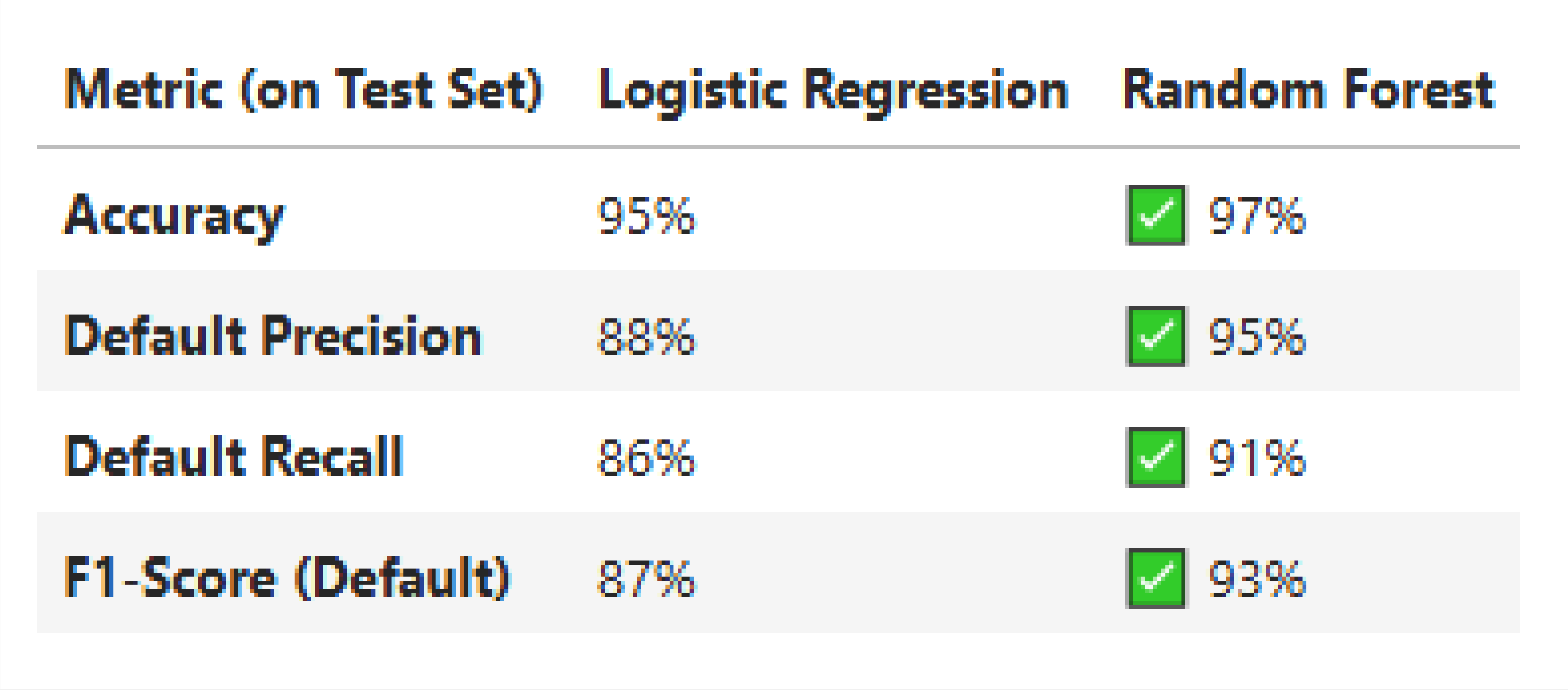
- 🔹 Logistic Regression (baseline)
- 🔹 Random Forest (advanced, tree-based)

Both trained on:

- Cleaned & engineered features
- OneHot + Scaled inputs
- 80/20 train-test split (32K+ records)

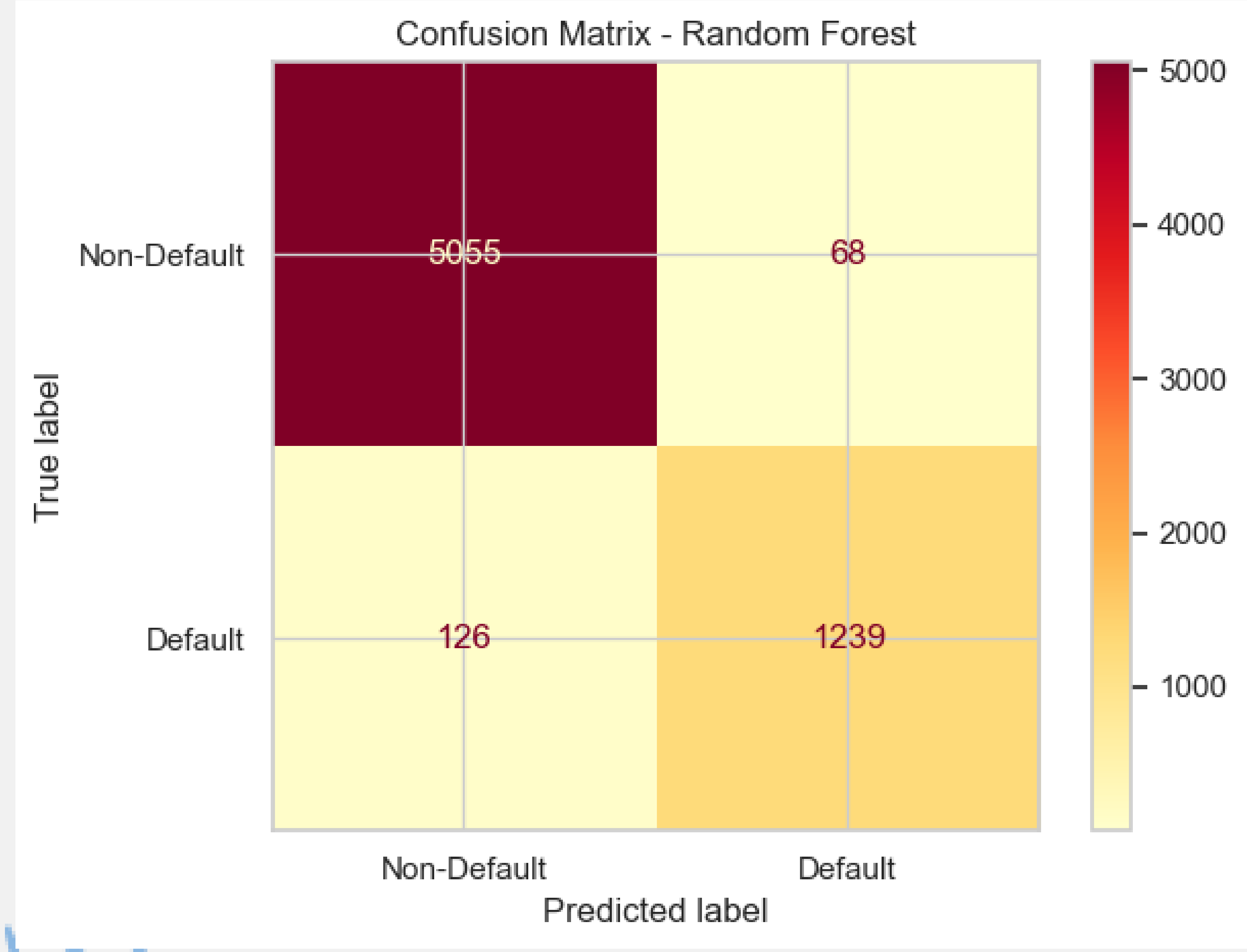
Model Evaluation Metrics & Performance comparison

- Accuracy → Overall correct predictions
- Precision → Out of predicted defaulters, how many were correct?
- Recall → Out of all actual defaulters, how many were caught?
- F1-Score → Balance between precision & recall (risk control)



Metric (on Test Set)	Logistic Regression	Random Forest
Accuracy	95%	✅ 97%
Default Precision	88%	✅ 95%
Default Recall	86%	✅ 91%
F1-Score (Default)	87%	✅ 93%

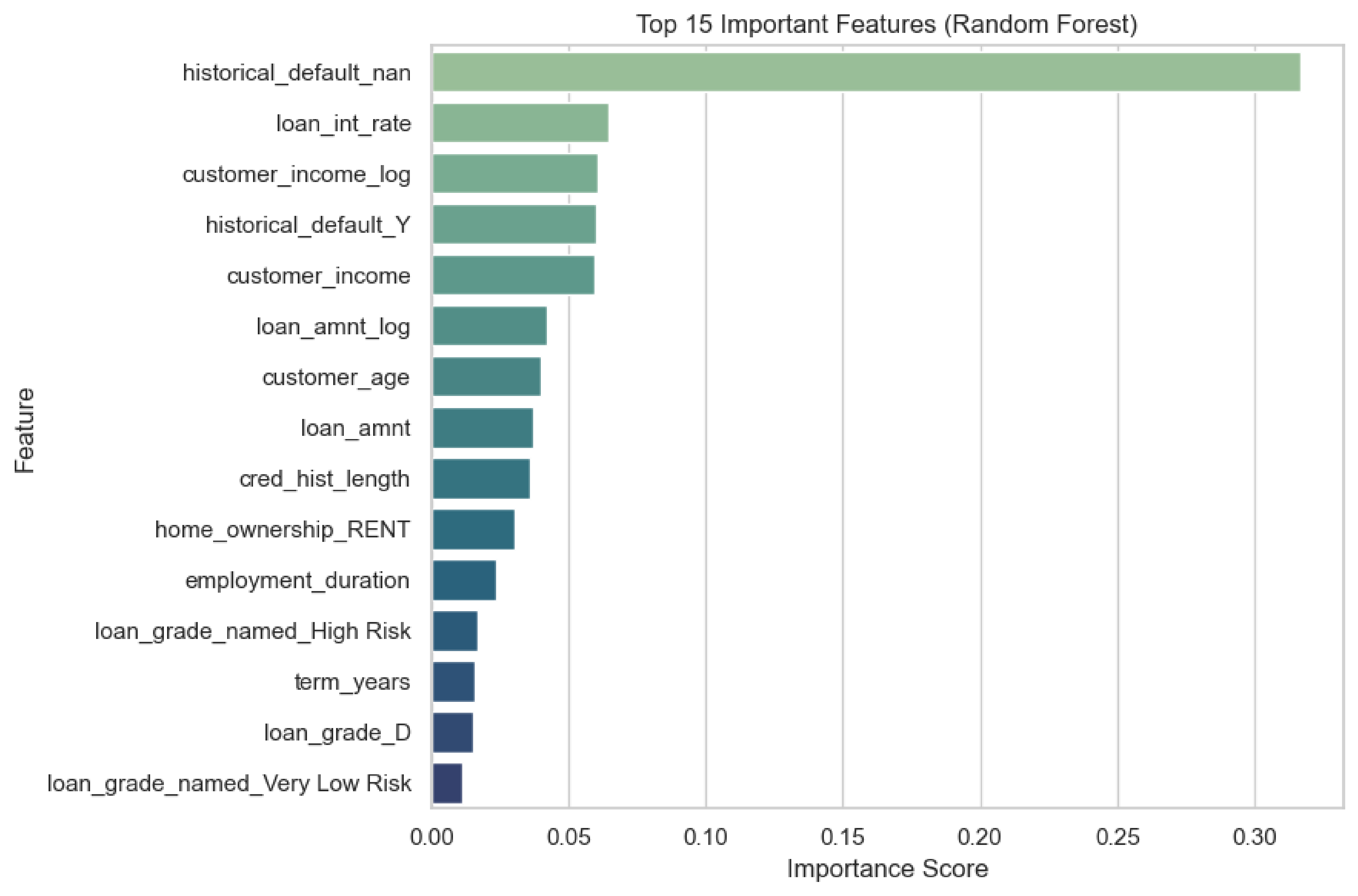
Confusion Matrix – Random Forest Model



Insights:

- Correct predictions: 6294 cases.
- Wrong predictions: 194 cases.

Top Predictive Features – Random Forest Model



Insights:

- Most important: historical_default_nan stands out as the key driver, holding nearly 30% of total importance.
- Next top group: Features like loan_int_rate, customer_income_log, and historical_default_Y each contribute around 10% importance.

Conclusion

📊 EDA and Visualization revealed key risk groups, including low-income borrowers, young applicants, and larger loan amounts

⍰ Statistical tests confirmed strong associations between default risk and features like loan grade, home ownership, and interest rate

⚙ Random Forest model delivered 97% accuracy, correctly identifying 91% of actual defaulters with strong reliability

🚀 The model is ready for real-world use to enhance loan approval processes and reduce financial risk exposure

Thank you

Any Question?