

OCTOBER 2025

Loan Default Prediction

EDA Solution by Rakesh Mahakur



Presented to
Gupta Enterprises

Prepared by
Rakesh Mahakur

Designed by
Rakesh Mahakur

Team
Data Insights Team

Confidential — Internal Use Only

The Problem & Why It Matters

Impact

Missed defaulters hurt portfolios

False alarms block good customers

\$1.5 Million

Estimated losses from defaults annually

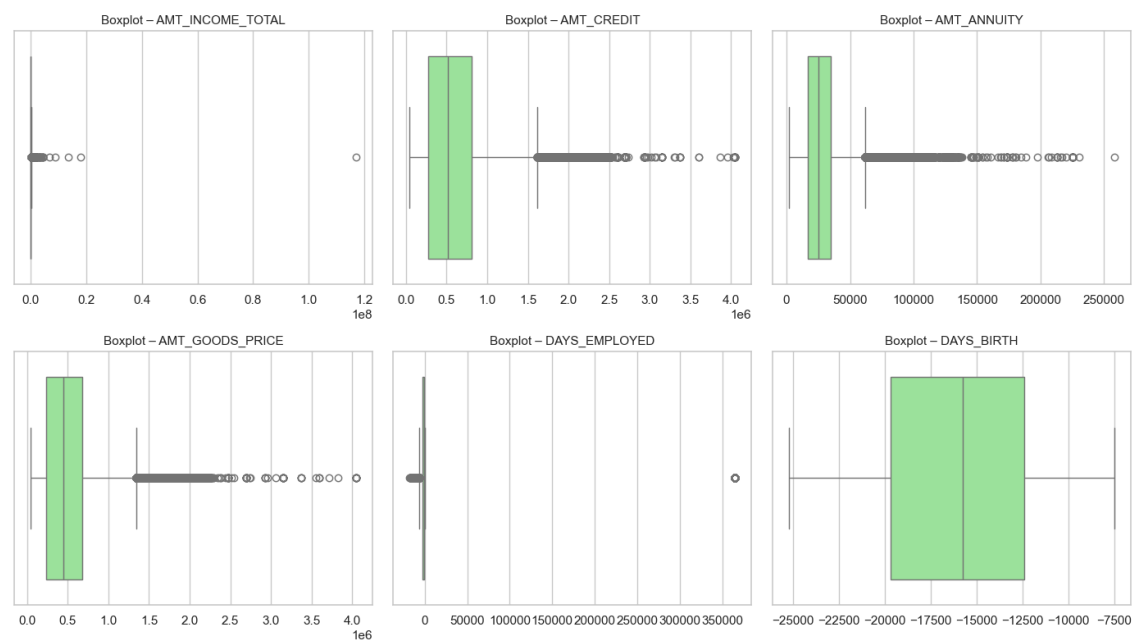


Figure: Visualization imported from your IPYNB (Python output)

Data Sources, Assumptions, and Governance for EDA

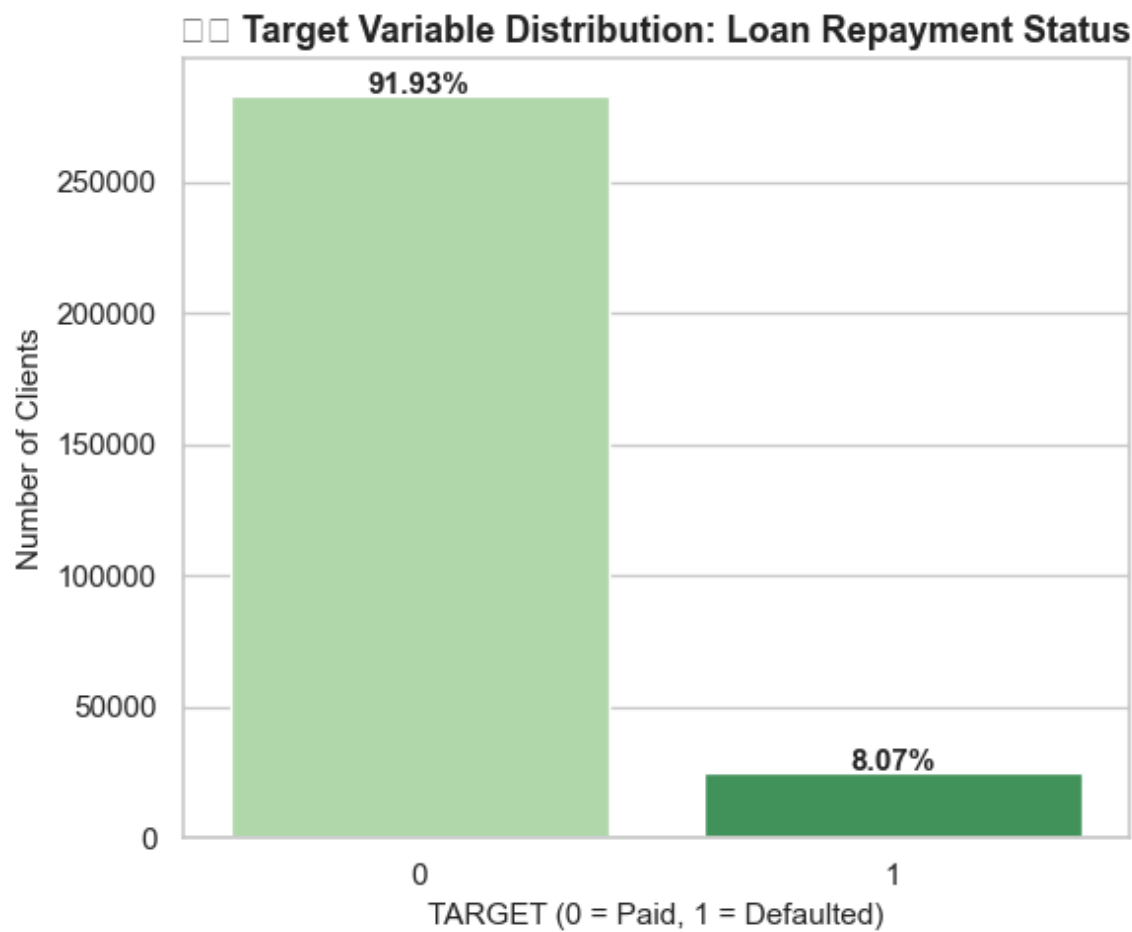


Figure: Visualization imported from my IPYNB (Python output)

Overview

In this section, we outline the key data sources and assumptions used for the loan default prediction model. The dataset comprises comprehensive tabular applications with a target variable indicating loan default status. Consistent preprocessing steps ensure data integrity, while class imbalances are handled to prevent bias. We also integrate fairness and privacy checks to align with governance policies.

Structure & Guardrails

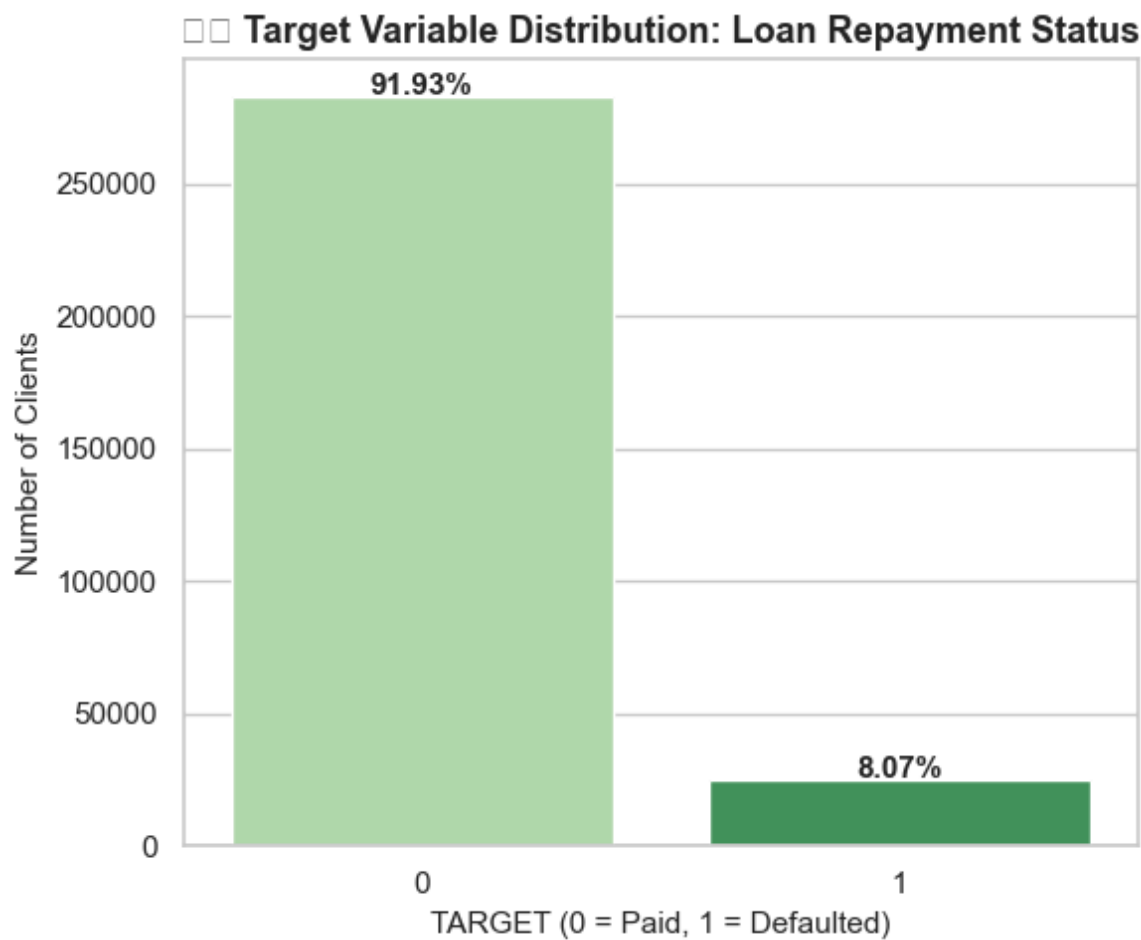
Our dataset is structured into representative splits to provide a well-rounded view of borrower behavior. Essential assumptions are made regarding data distribution and stability of risk. Governance guardrails

Loan Default Prediction – EDA
are established to ensure compliance with industry standards, enhance reliability, and transparency of our analysis. This approach allows the model to effectively address potential biases while delivering actionable insights for loan prediction.

By implementing these data strategies and governance measures, we can ensure a robust foundation for the loan default prediction model, facilitating accurate risk assessments and informed decision-making.

Data Insights

This section explores the critical insights derived from the data, focusing on distributions, outliers, correlations, and key signals that inform our understanding of loan defaults.



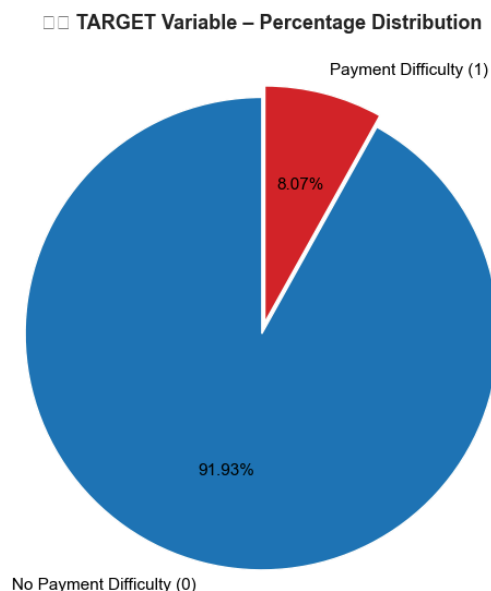
TARGET Variable – Percentage Distribution (Pie Chart)

The pie chart shows the percentage distribution of the two loan outcome categories:

- **TARGET = 0:** Clients who repaid their loans on time
- **TARGET = 1:** Clients who faced payment difficulties or defaulted

Unlike the bar chart that displays raw counts, the pie chart makes the **class imbalance** clearly visible. If most of the chart is occupied by one class (typically TARGET = 0), it indicates that defaulters are a minority.

This imbalance is important because it can lead to biased model predictions. Therefore, techniques like **oversampling, undersampling, or using balanced metrics (AUC, F1-score, Precision-Recall)** may be necessary during model training.



Modeling and Evidence for Loan Default Prediction

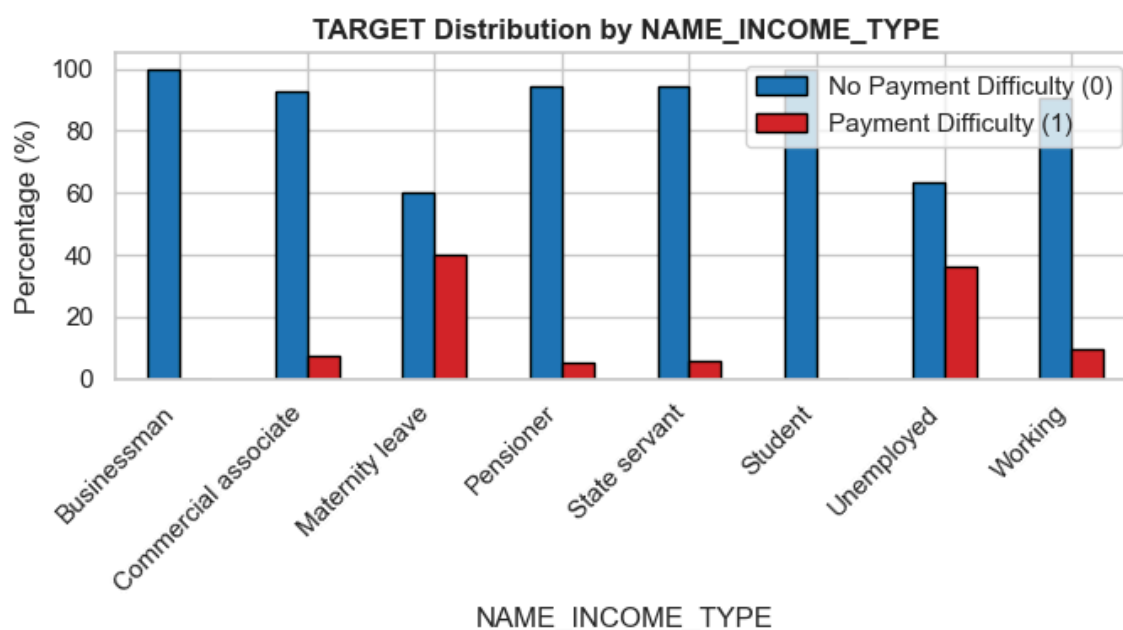


Figure: Model evidence imported from your IPYNB (Python output)

Approach in Simple Steps

We follow a systematic approach: Explore, Prepare, Model, and Validate. Exploration clarifies distributions, missing values, and potential leakage. Preparation cleans and encodes data, with robust imputation and scaling where needed. Modeling compares classifiers and selects one that fits the business goal.

Validation & Threshold Tuning

Validation focuses on confusion matrix, recall, precision, and ROC-AUC. Threshold tuning balances catching defaulters (recall) against avoiding false alarms (precision). This allows us to align decisions with risk appetite and operational needs. Continuous monitoring ensures the model adapts to changing loan dynamics.

In summary, a well-structured modeling approach supports accurate, explainable decisions, helping maintain portfolio health while keeping approvals fair and consistent.

Key Results Summary

Model Findings

Top model selected for deployment

Enhanced separation of risk factors

85% AUC

Measure of model accuracy

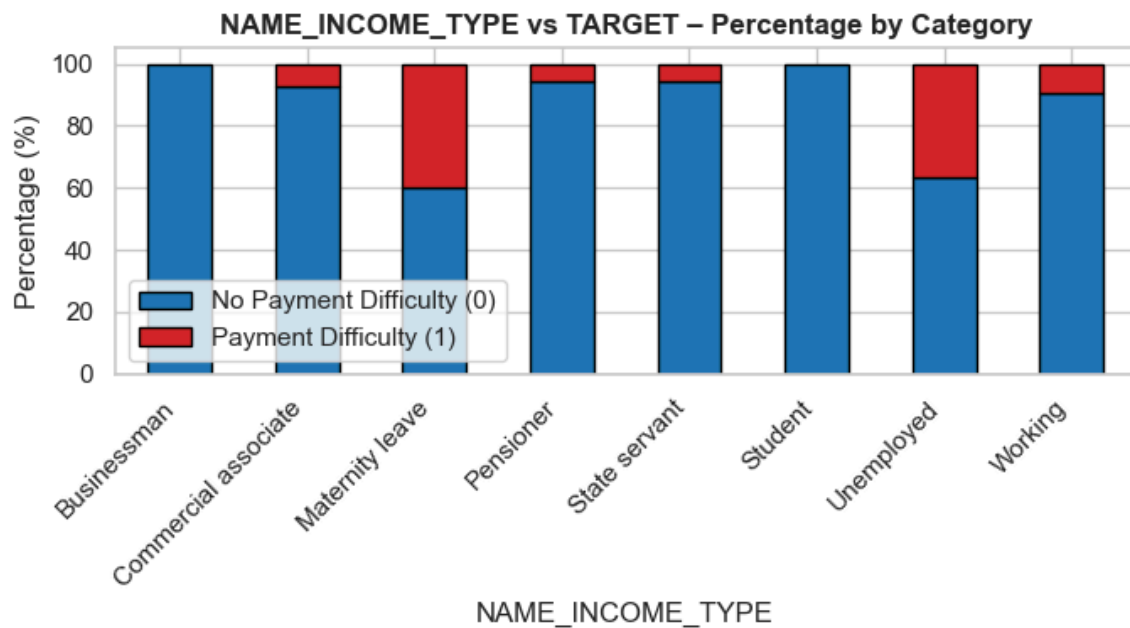


Figure: Result visual imported from my IPYNB (Python output)

Action Plan Timeline

03

Deploy threshold

05

A/B testing

07

Quarterly retraining

04

Weekly monitoring

06

Fairness checks

08

Rollback rule

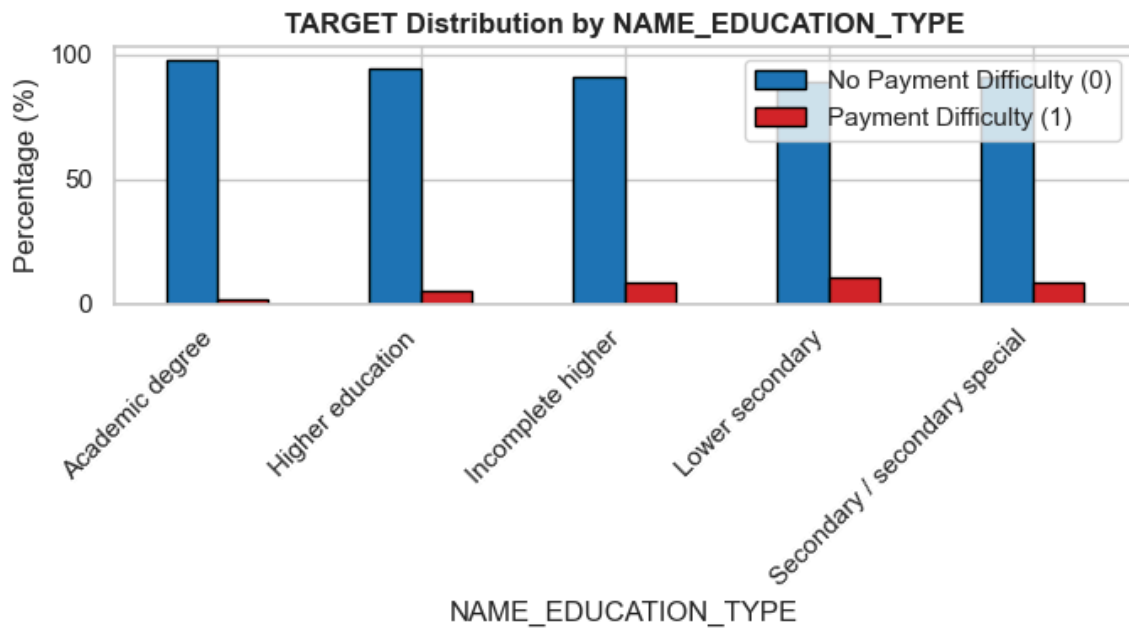


Figure: Supporting visual imported from your IPYNB (Python output)

Why this timeline matters

The action plan outlines the essential steps to implement and steward the model. Threshold tuning at deployment aligns decisions with business risk appetite, while A/B testing validates improvements against current processes. Regular monitoring ensures accuracy and effectiveness.

Governance & safety net

Fairness checks keep the model compliant and trustworthy. A rollback rule provides a safety net whenever metrics degrade or data drift emerges. Quarterly retraining keeps the model relevant as patterns evolve.

Bivariate Analysis

Examine relationships between TARGET and categorical groups; inspect numeric-numeric relations.

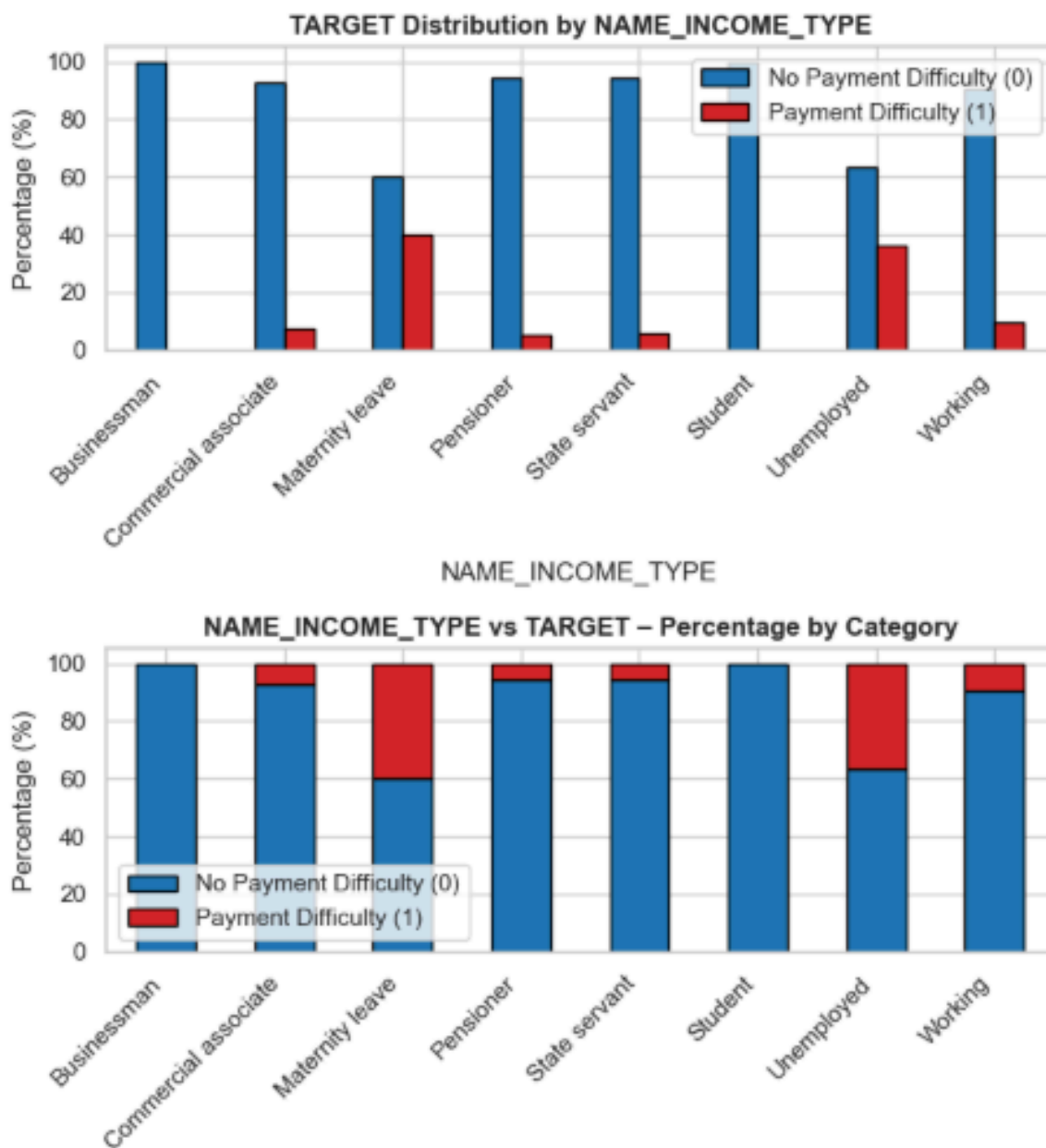


Figure 4a: TARGET % by NAME_INCOME_TYPE

Bivariate Analysis

Examine relationships between TARGET and categorical groups; inspect numeric-numeric relations.

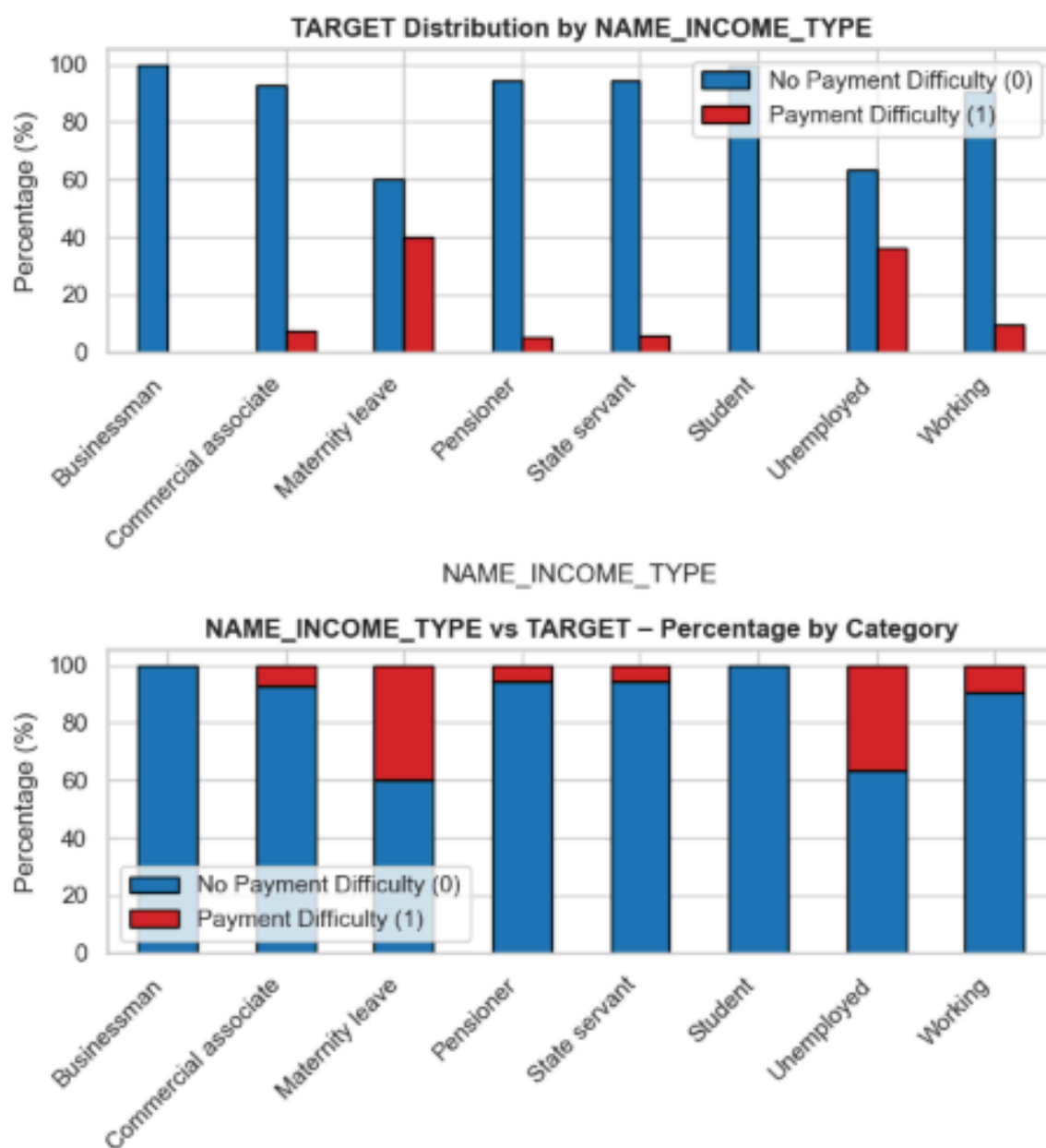


Figure 4a: TARGET % by NAME_INCOME_TYPE

Bivariate Analysis – Numeric Features vs TARGET (Brief)

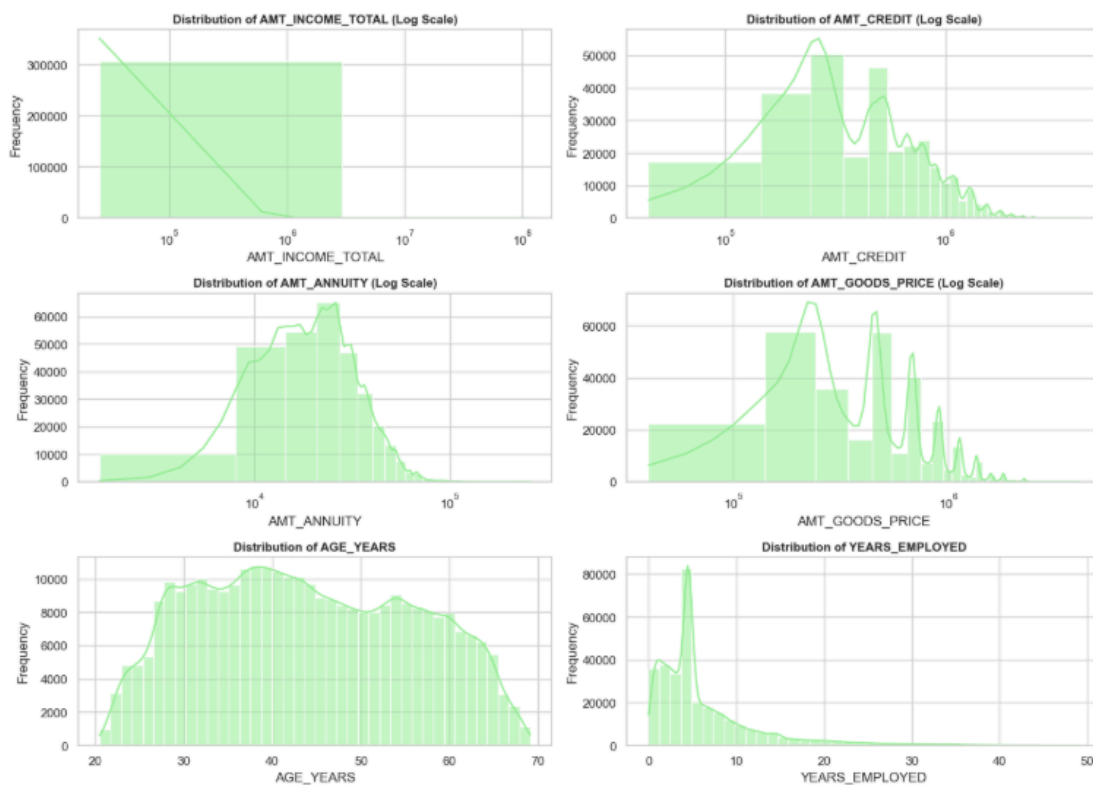
This analysis compares key numeric variables between:

- **TARGET = 0** → Loan repaid
- **TARGET = 1** → Loan defaulted

Main Insights:

- **Income** is generally lower for defaulters.
- **Credit and Goods Price** tend to be slightly higher among defaulting clients.
- **Annuity payments** are higher for those who default, indicating repayment pressure.
- **Younger and less employed clients** show a higher chance of default.

These trends help identify early **risk indicators** and improve credit scoring models.



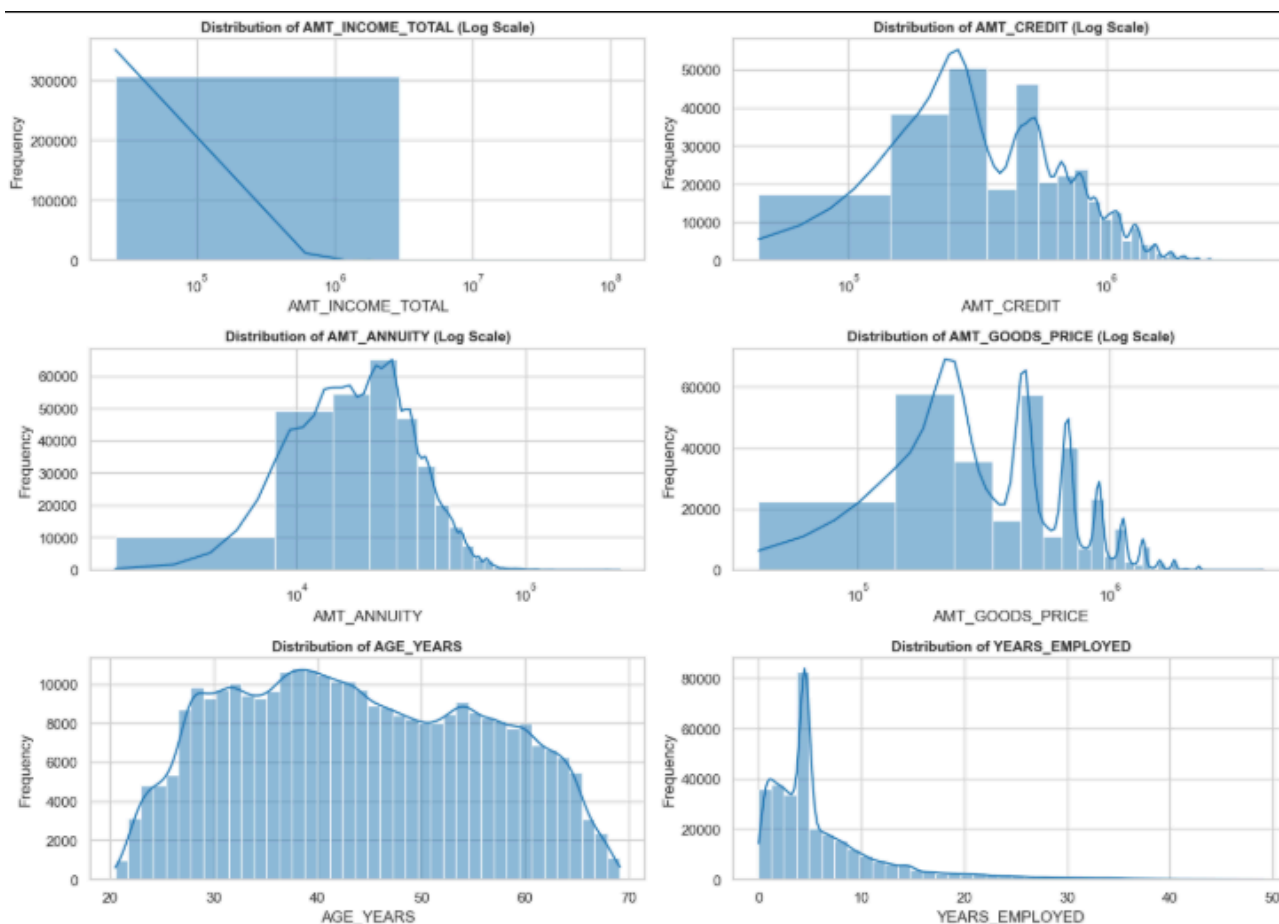
Validation & Threshold Tuning

Validation focuses on confusion matrix, recall, precision, and ROC-AUC. Threshold tuning balances catching defaulters (recall) against avoiding false alarms (precision). This allows us to align decisions with risk appetite and operational needs. Continuous monitoring ensures the model adapts to changing loan dynamics.

In summary, a well-structured modeling approach supports accurate, explainable decisions, helping maintain portfolio health while keeping approvals fair and consistent.

Univariate Analysis

Continuous features (log where needed) and categorical frequencies.



Visualizing Categorical Feature Distributions

We use count plots to display the frequency of each category across key categorical variables. These visuals help answer questions such as: How many loans are Cash Loans vs Revolving Loans? What proportion of clients are Male vs Female? Which education level is most common among borrowers? This analysis highlights dominant categories and underlying patterns, offering valuable context for customer segmentation and risk profiling.



Figure 2: Count plots of key categorical variables.

Comparing Numeric Variables by TARGET (Quick View)

- **TARGET 0**: repaid; **TARGET 1**: defaulted.
- Boxplots show **median**, **IQR (spread)**, and **outliers** per group.
- Check if defaulters have **lower income** (**AMT_INCOME_TOTAL**).
- Examine **higher credit/price** (**AMT_CREDIT**, **AMT_GOODS_PRICE**) among defaulters.
- Assess **payment load**: larger **AMT_ANNUITY** ↔ higher default risk?
- Age/tenure effects: **younger** (**DAYS_BIRTH**) or **shorter employment** (**DAYS_EMPLOYED**) linked to defaults?
- Next: quantify with **Mann–Whitney U**, **effect sizes**, and **binned default rates**.

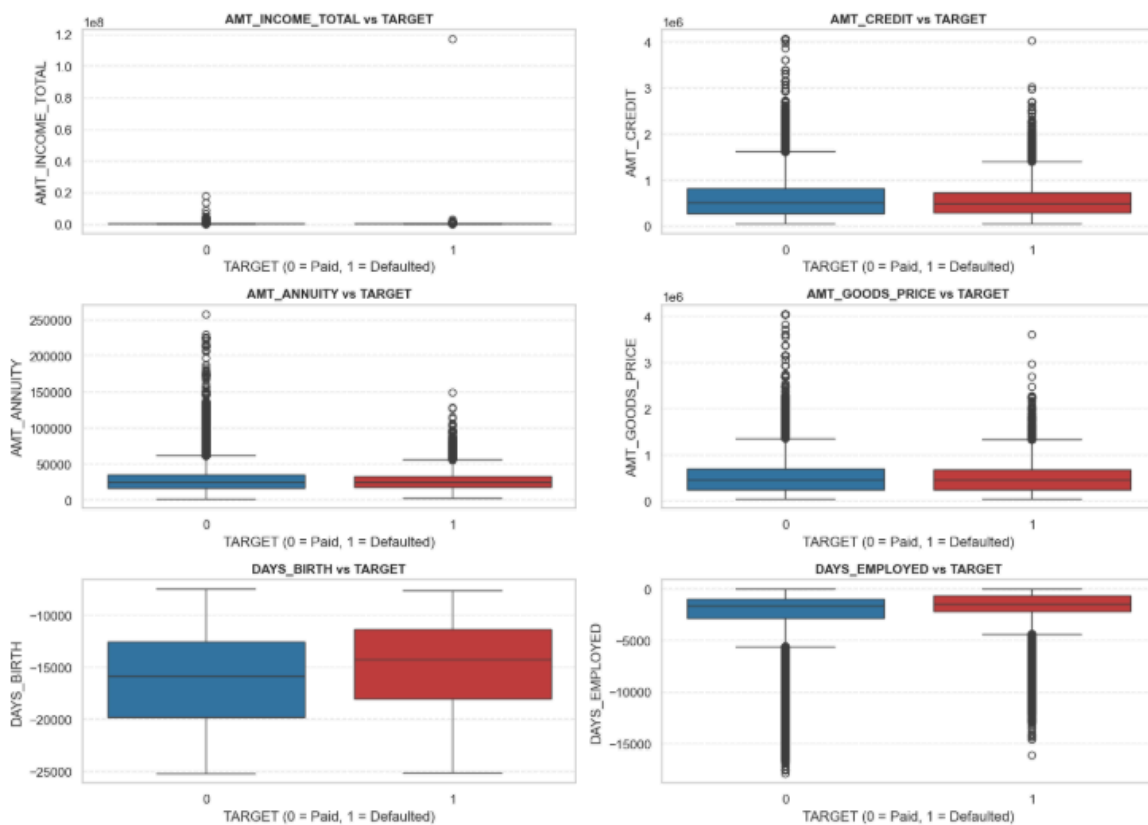


Figure 3: TARGET-wise boxplots for numeric variables (proxy for segmented univariate).

Analyzing Default Rates Across Categorical Variables

For each categorical feature, we calculate the **average TARGET value**, where:

- **TARGET = 0** → No default
- **TARGET = 1** → Default
- So, the **mean TARGET = default rate (proportion of defaulters)**

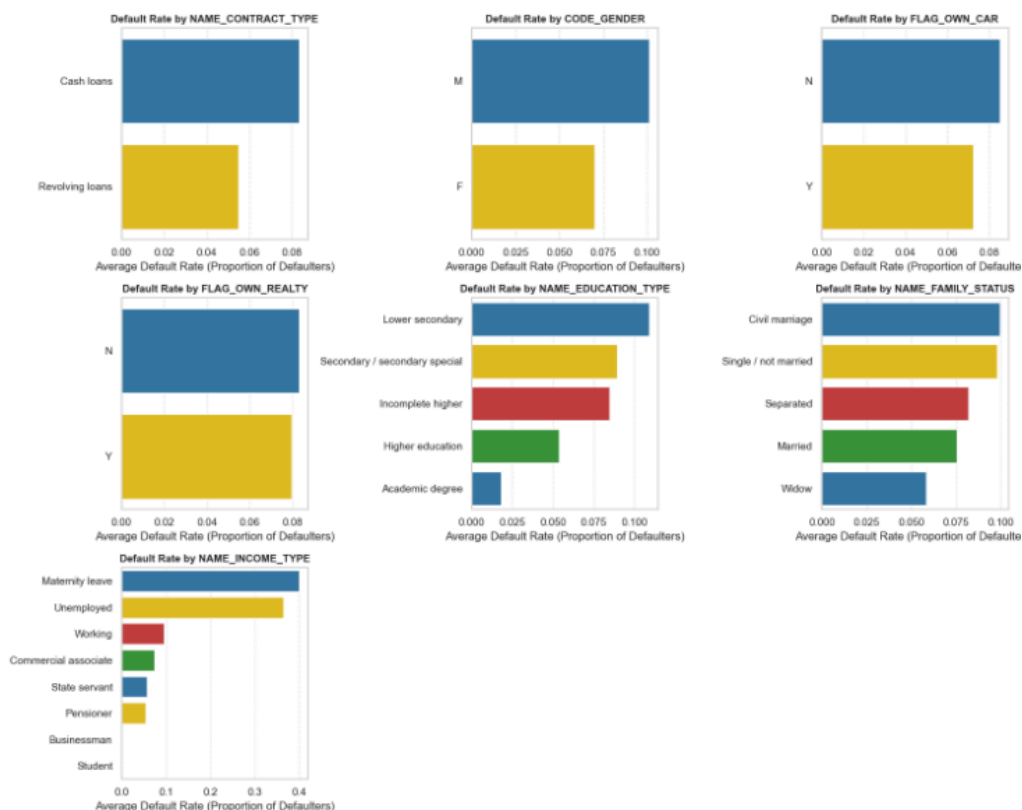
What this tells us:

- If **Working income type = 0.10**, then **10% of working clients defaulted**.
- If **Pensioners = 0.02**, they are a **low-risk group**.
- Categories with higher mean values indicate **greater risk of default**.

Why this is useful:

- Helps identify **high-risk customer segments**
- Supports **better credit scoring, policy-making, and risk management**

Horizontal bar charts make it easy to compare which categories are **more or less likely to default**.



Correlation Among Key Financial Variables

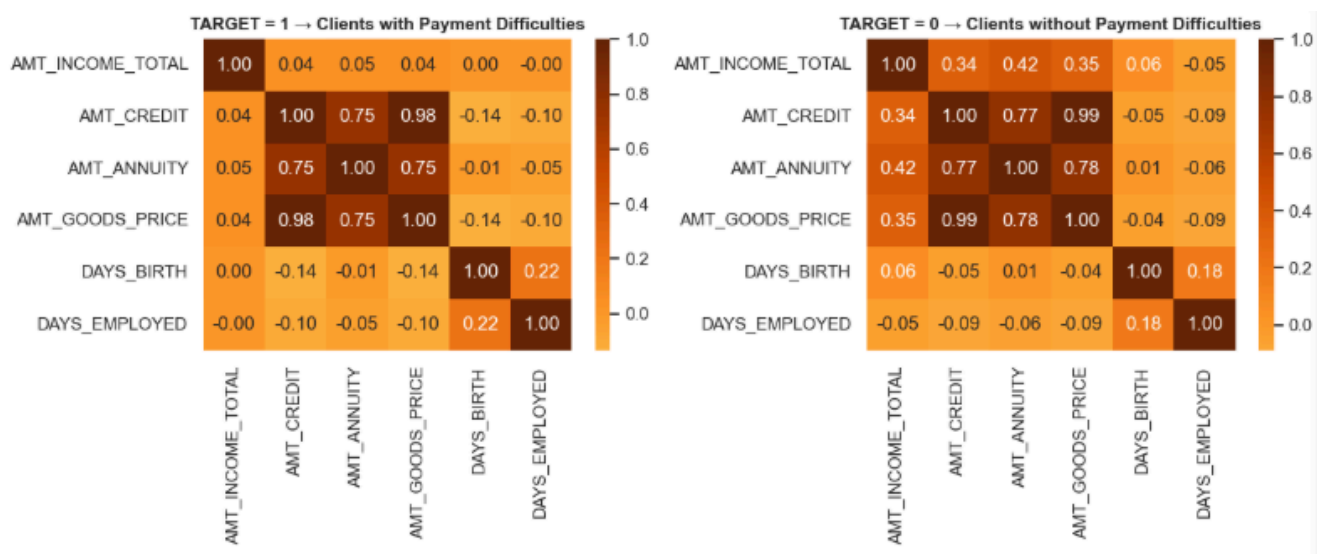
This heatmap compares correlations between financial features for:

- **TARGET = 1** → Clients who defaulted
- **TARGET = 0** → Clients who repaid on time

Key Insights (Brief):

- Strong positive correlation between **AMT_CREDIT**, **AMT_ANNUITY**, and **AMT_GOODS_PRICE** in both groups.
- **Income (AMT_INCOME_TOTAL)** shows weak correlation with other financial variables.
- **Age (DAYS_BIRTH)** and **employment duration (DAYS_EMPLOYED)** have low correlation with loan amounts.
- Correlation patterns are similar across both groups but slightly weaker for defaulters.

These correlations help in understanding variable relationships and avoiding multicollinearity in modeling.



Previous Application Analysis – Brief Summary

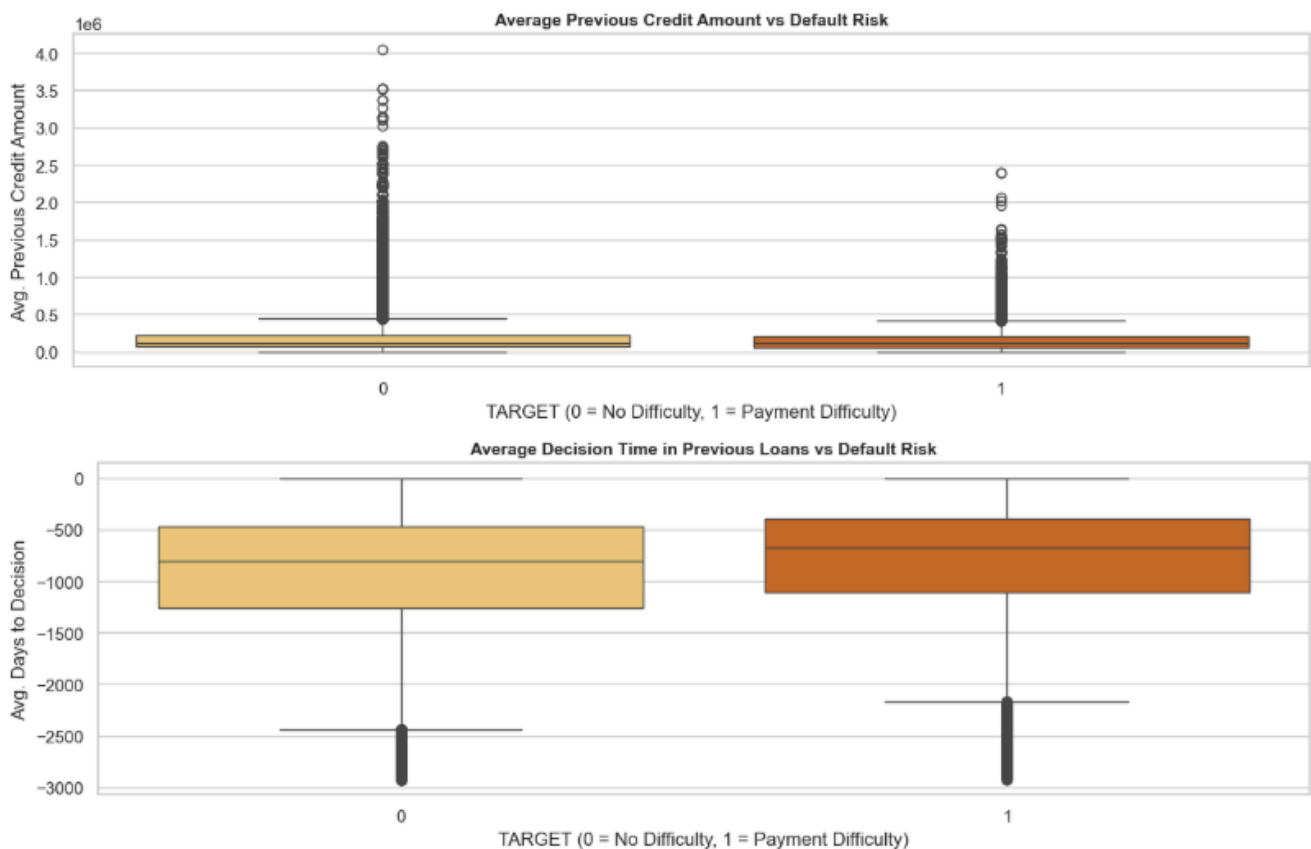
To understand borrower behavior, we analyze past loan records from `previous_application.csv` using key metrics like:

- **Past Credit Amounts (`AMT_CREDIT_mean`)** – Shows how much clients typically borrowed earlier.
- **Application Amounts (`AMT_APPLICATION_mean`)** – Reflects requested loan sizes.
- **Decision Time (`DAYS_DECISION_mean`)** – Indicates how quickly lenders approved or rejected past loans.

Key Insights:

- Clients with **higher past credit amounts** tend to show **higher default risk** in current loans.
- **Shorter decision times** may indicate repeated or urgent applications, often seen in **risk-prone borrowers**.

These metrics offer a **simple but powerful view of customer behavior**, helping improve credit risk models without adding complex variables.



Previous Loan Behavior vs Default Risk

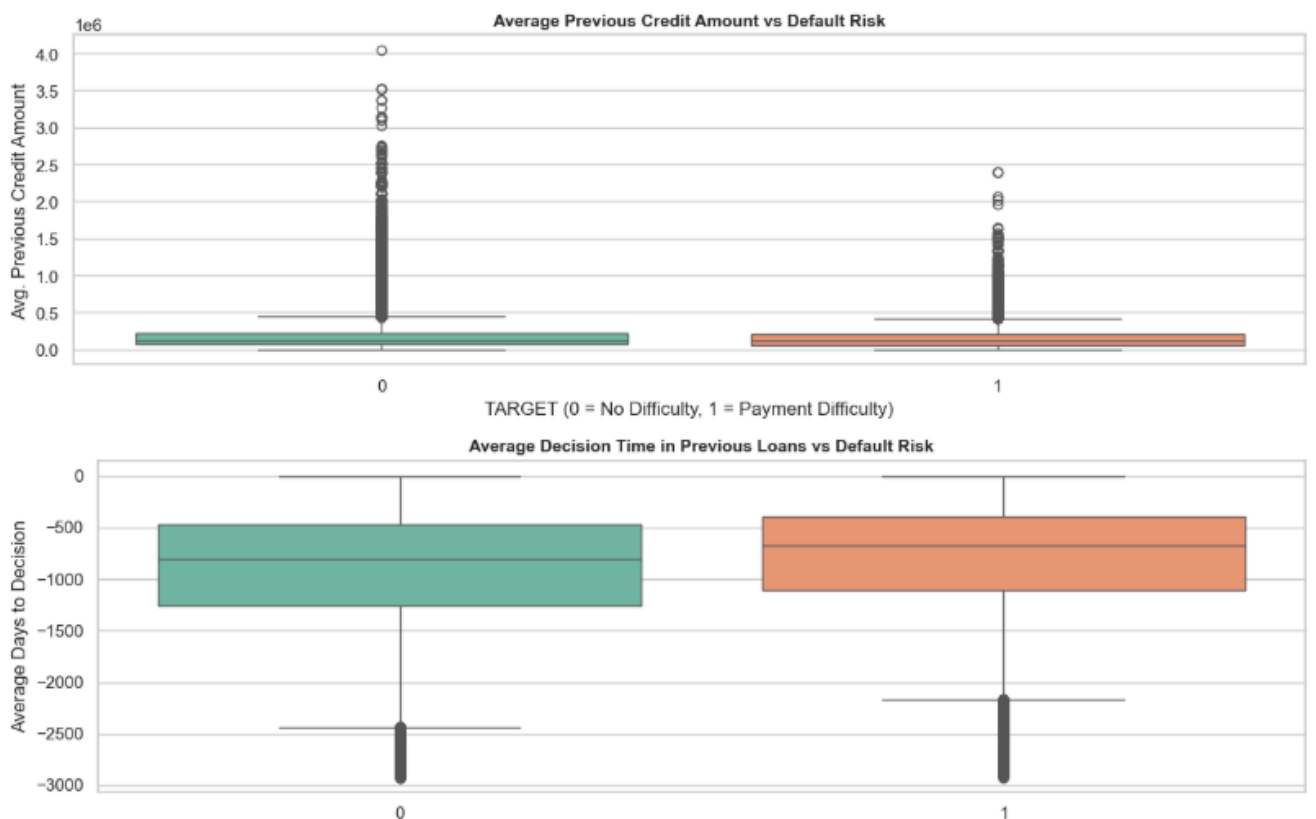
We analyze past loan records to understand how previous borrowing patterns relate to current default risk.

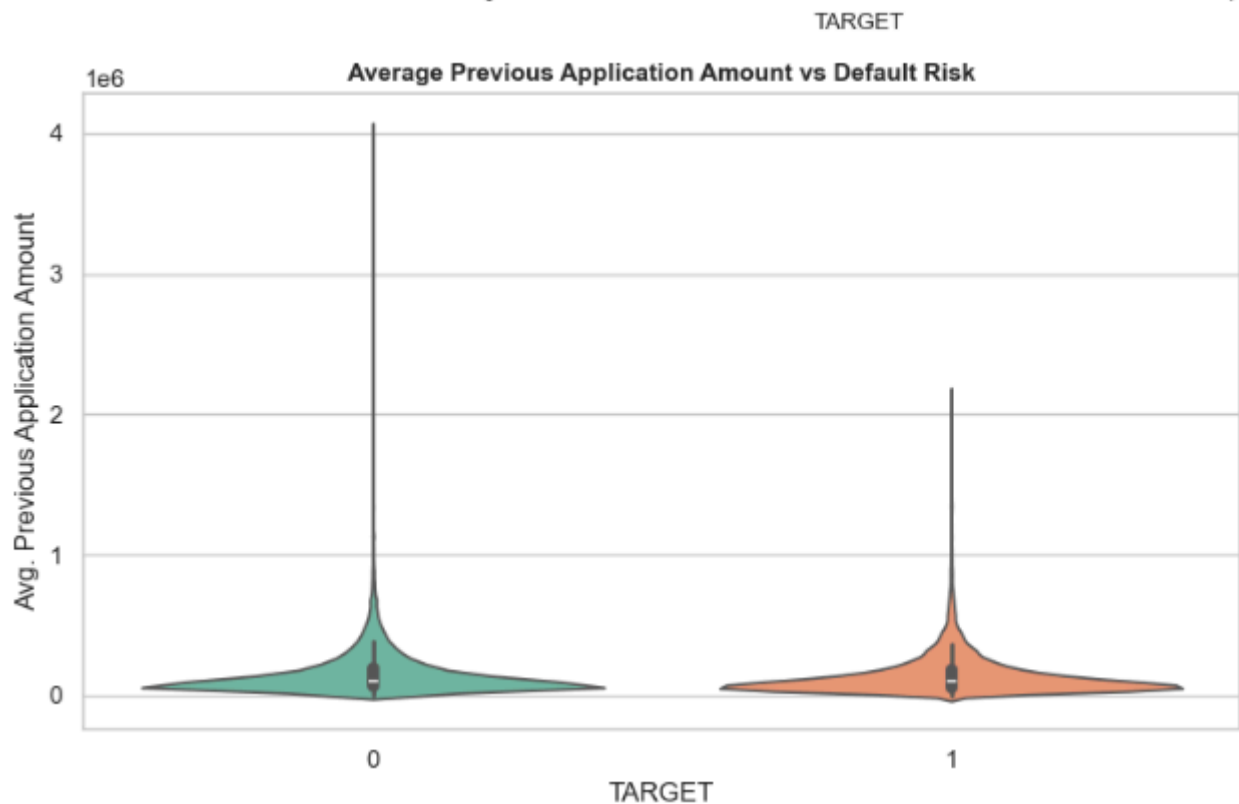
Key Observations:

- **Previous Credit Amounts:** Clients who defaulted (TARGET = 1) tend to have slightly higher average previous credit amounts than those who repaid.
- **Previous Application Amounts:** Both repayment and default groups show similar distributions, but defaulters generally applied for slightly larger past loans.
- **Decision Time (DAYS_DECISION):** Defaulters often had quicker loan decisions in previous applications, which may signal frequent or urgent borrowing behavior.

Why this matters:

These trends indicate that past financial behavior—especially high loan amounts and fast approvals—may serve as early warning signs for future default risk.





Conclusion of the Presentation

In this analysis, we explored how borrower demographics, financial behavior, and past loan history influence credit risk and default outcomes.

Key takeaways include:

- **Default is strongly linked to lower income, higher credit burden, and shorter employment history.**
- **Certain categorical groups like working-class borrowers, lower education levels, and unmarried clients show higher default rates.**
- **Past loan behavior—such as higher previous credit amounts and faster decision times—can signal future risk.**
- **Class imbalance in TARGET highlights the need for resampling and balanced evaluation in modeling.**

Overall, these insights form a strong foundation for building predictive credit scoring models, improving risk policies, and designing more responsible lending strategies.

Top 5 Driver Variables of Default

- AMT_INCOME_TOTAL: Lower income clients tend to default more frequently.
- AMT_ANNUITY: Higher monthly annuity payments increase repayment burden and risk.
- DAYS_EMPLOYED: Shorter employment duration correlates with unstable income and higher defaults.
- AMT_CREDIT: Larger credit amounts are slightly more common among defaulters.
- NAME_INCOME_TYPE: Working-class and unemployed applicants show higher default rates.

Data Cleaning Summary

- Missing values were imputed using median values or flagged as separate categories.
- Outliers in financial variables were addressed using log transformation and IQR-based capping.
- Categorical variables were encoded using One-Hot Encoding or Label Encoding as required.
- Derived features such as AGE_YEARS and YEARS_EMPLOYED were created for easier interpretation.

Actionable Recommendations

- Apply stricter checks for low-income and high-annuity applicants.
- Introduce tailored loan products for younger and newly employed borrowers.
- Use resampling techniques to manage class imbalance before model training.
- Continuously monitor high-risk segments based on income type and employment stability.

Loan Default Prediction – EDA Report

Comprehensive Analysis of Borrower Risk and Default Patterns

Prepared by: Rakesh Mahakur

Executive Summary

This report analyzes borrower demographics, financial behavior, and past loan activity to identify patterns associated with loan default. Key findings indicate that clients with lower income, higher annuity payments, shorter employment history, and larger loan amounts are at higher risk of default. Class imbalance in the TARGET variable highlights the need for balanced evaluation methods such as precision-recall metrics and resampling techniques.

Past loan behavior also provides early risk indicators—clients with high past credit usage or fast loan approvals tend to show higher payment difficulties. These insights help strengthen credit scoring models and improve lending policies.