OCTOBER 2025

Loan Default Prediction

EDA Solution by Rakesh Mahakur

Prepared by
Rakesh Mahakur

Designed by
Rakesh Mahakur

Team
Data Insights Team

Impact

Missed defaulters hurt portfolios

False alarms block good customers

$1.5 Million

Estimated losses from defaults annually

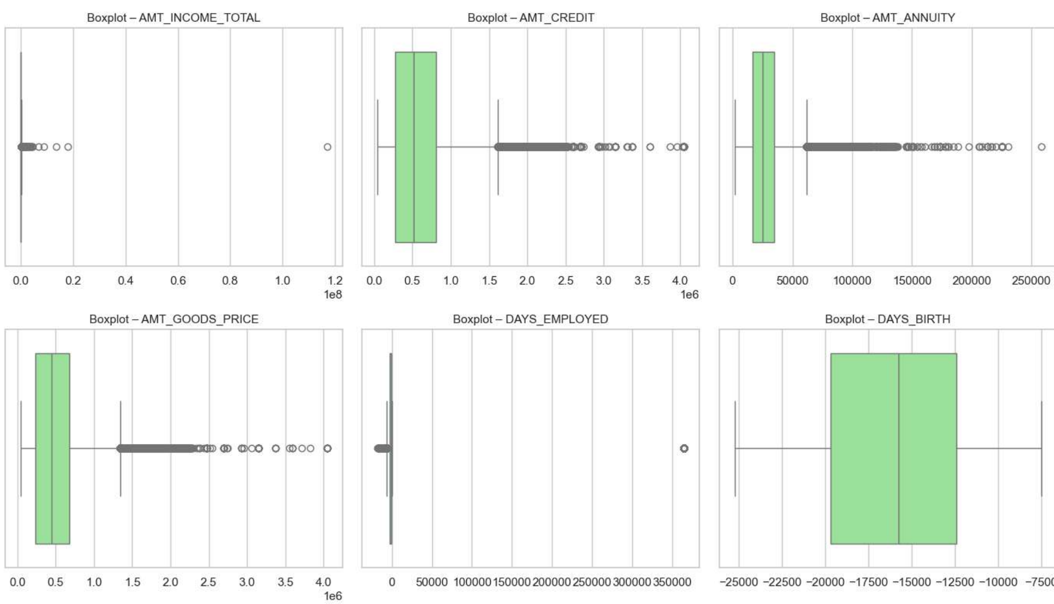


Figure: Visualization imported from your IPYNB (Python output)

Data Sources, Assumptions, and Governance for EDA

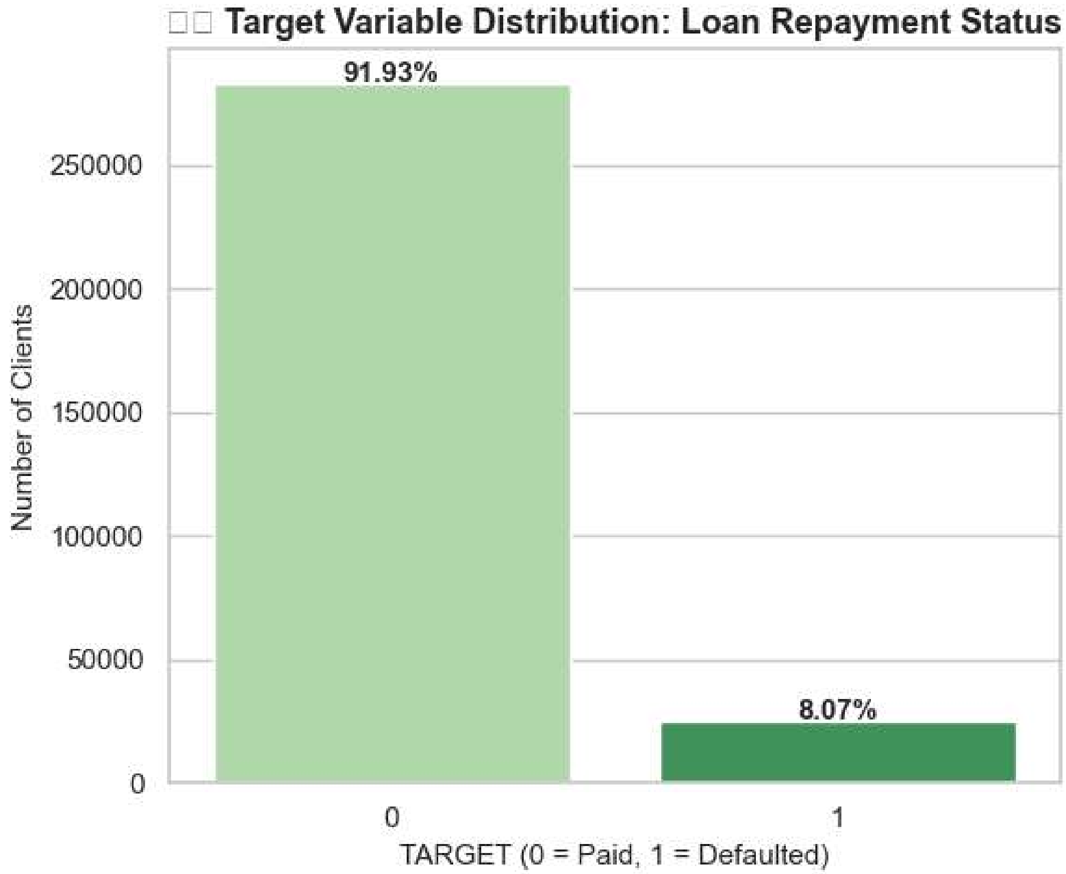
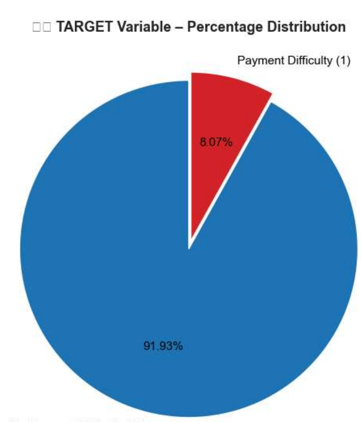


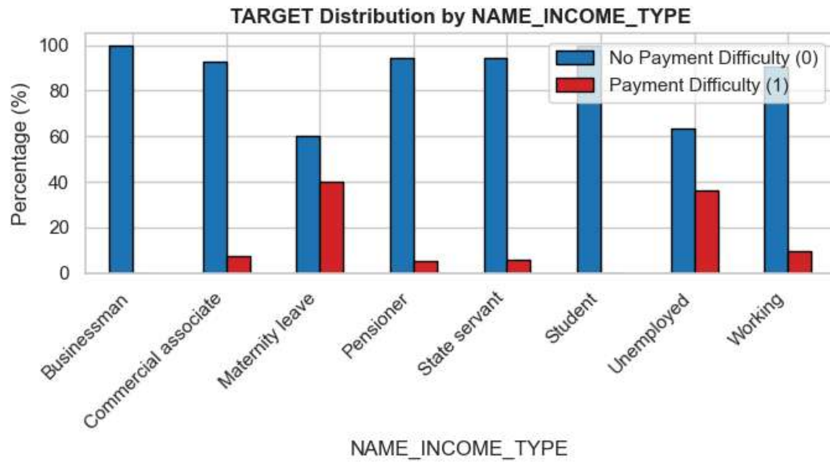
Figure: Visualization imported from my IPYNB (Python output)

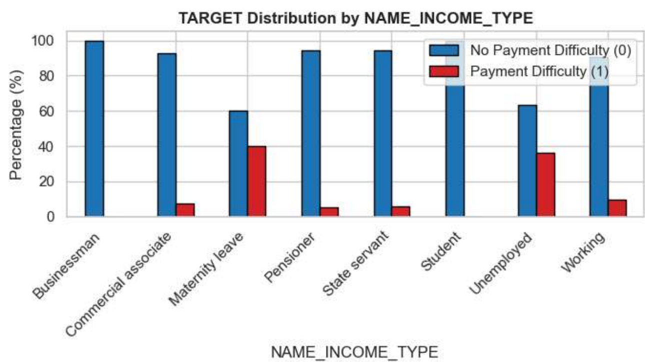
TARGET Variable – Percentage Distribution (Pie Chart)

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







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

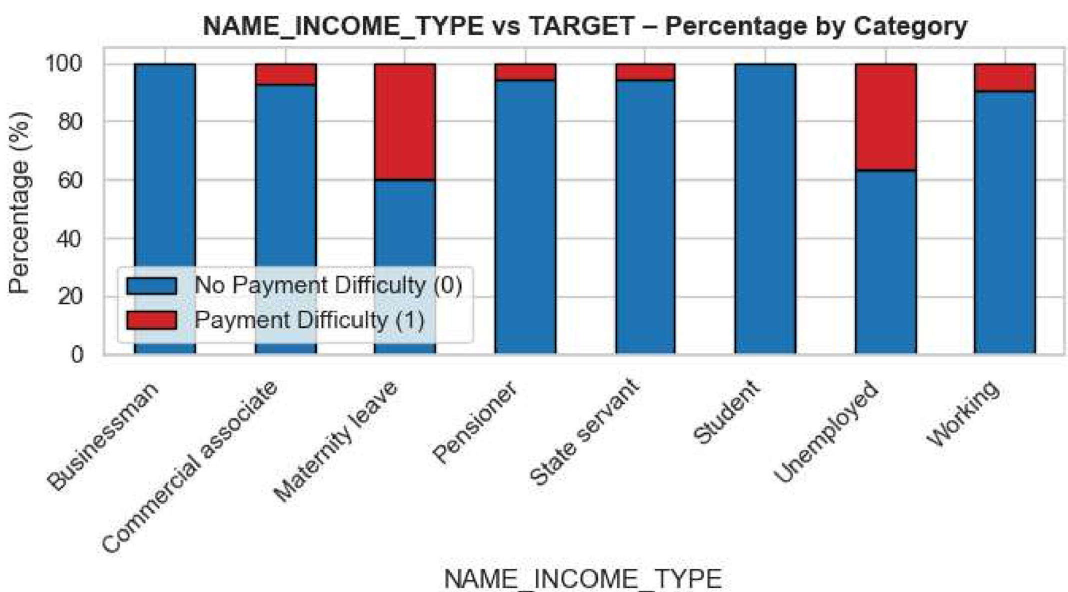
Key Results Summary

Model Findings

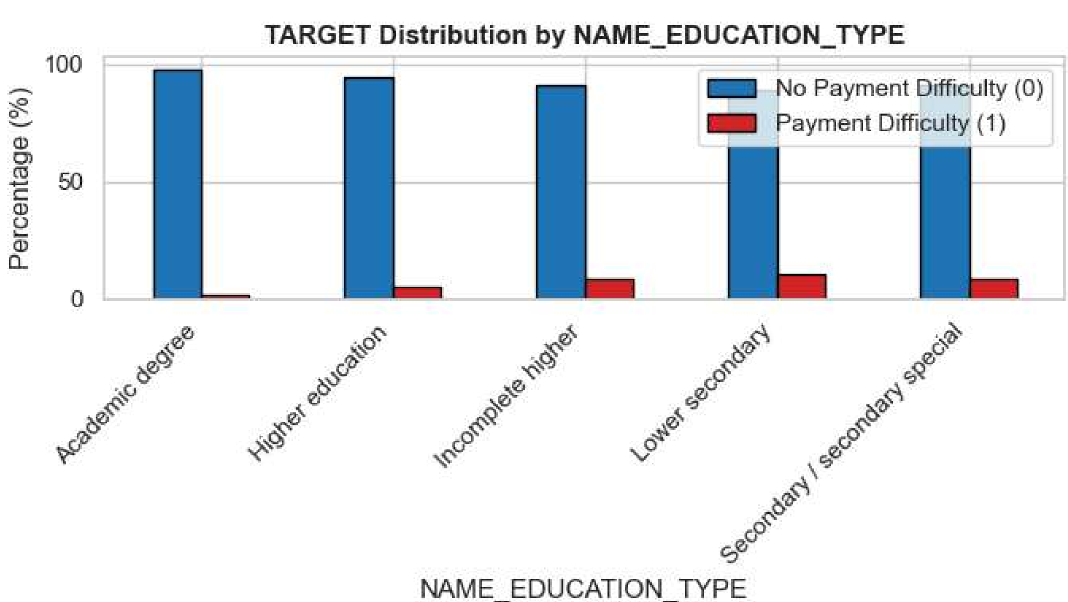
Top model selected for deployment

85% AUC

Measure of model accuracy



Action Plan Timeline



Why this timeline matters

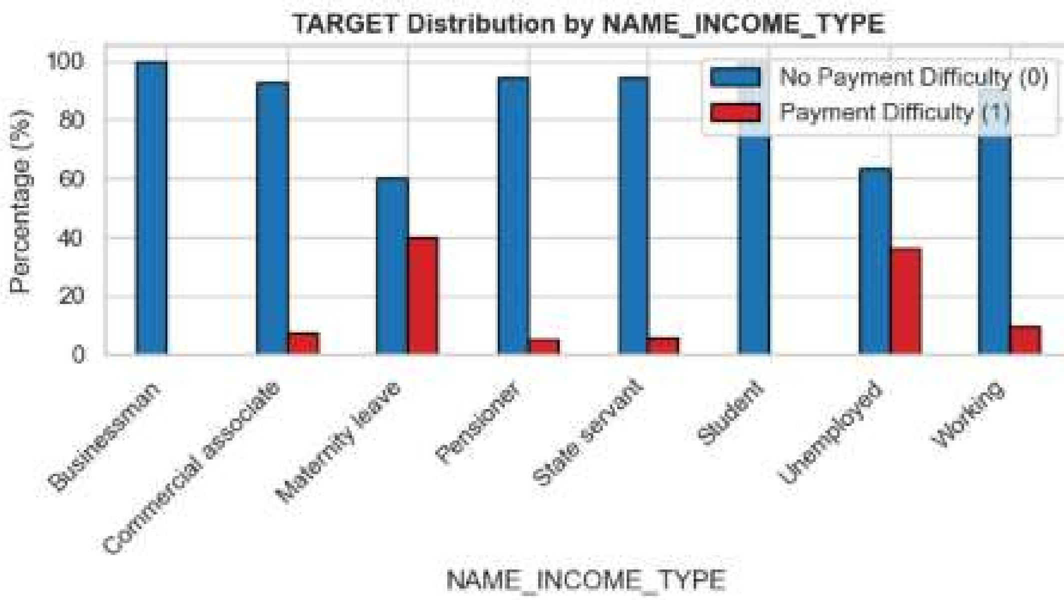
The action plan outlines the essential steps to implement and steward the model. Threshold tuning at deployment aligns decisions with business risk appetite, while A/B testing validates improvements against current processes. Regular monitoring ensures accuracy and effectiveness.

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.



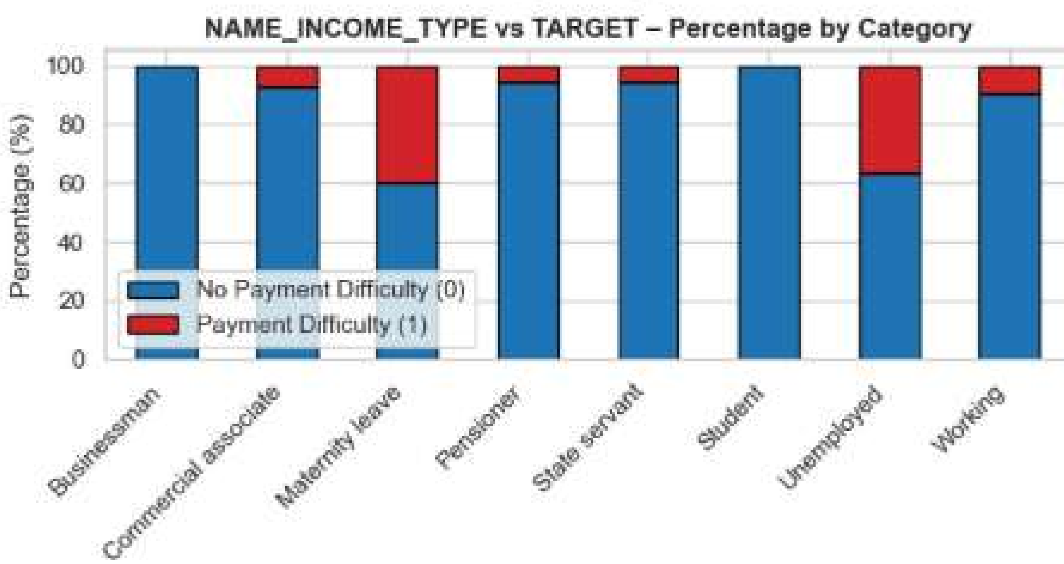
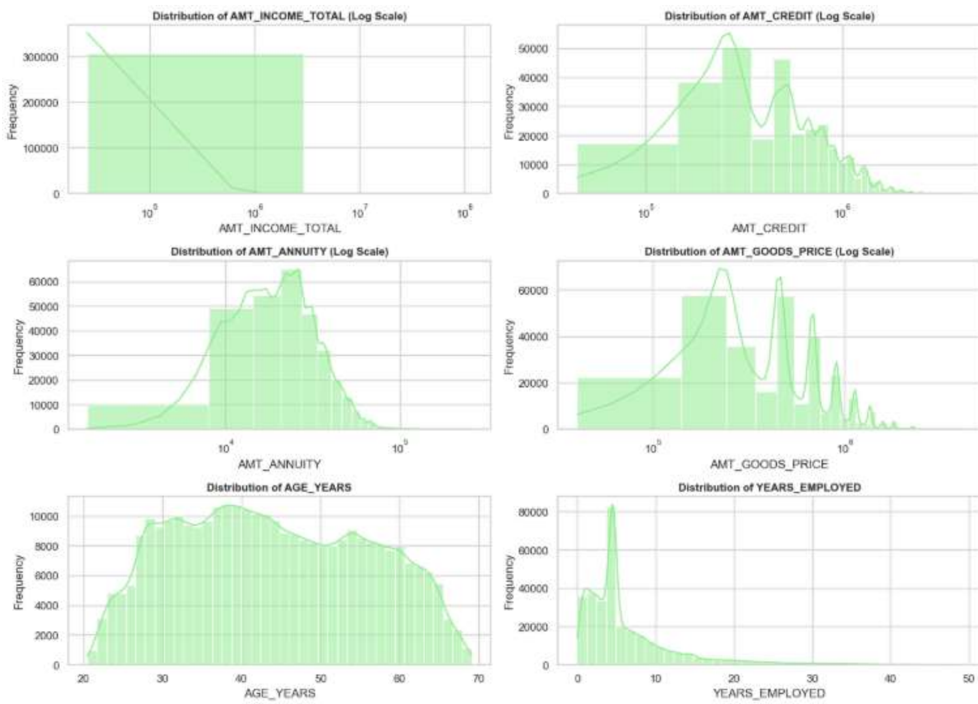


Figure 4a: TARGET % by NAME_INCOME_TYPE

Bivariate Analysis – Numeric Features vs TARGET (Brief)

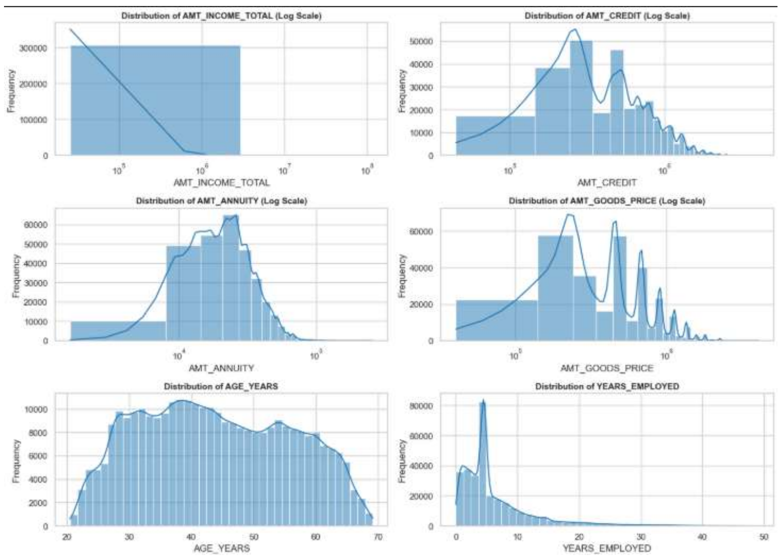


Validation & Threshold Tuning

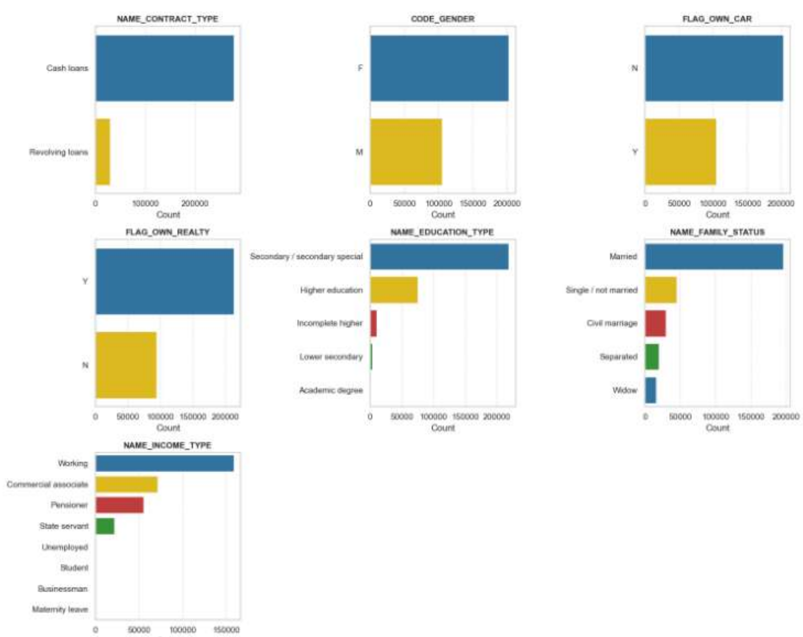
Validation focuses on confusion matrix, recall, precision, and ROC-AUC. Threshold tuning balances catching defaulters (recall) against avoiding false alarms (precision). This allows us to align decisions with risk appetite and operational needs. Continuous monitoring ensures the model adapts to changing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

Univariate Analysis

Continuous features (log where needed) and categorical frequencies.



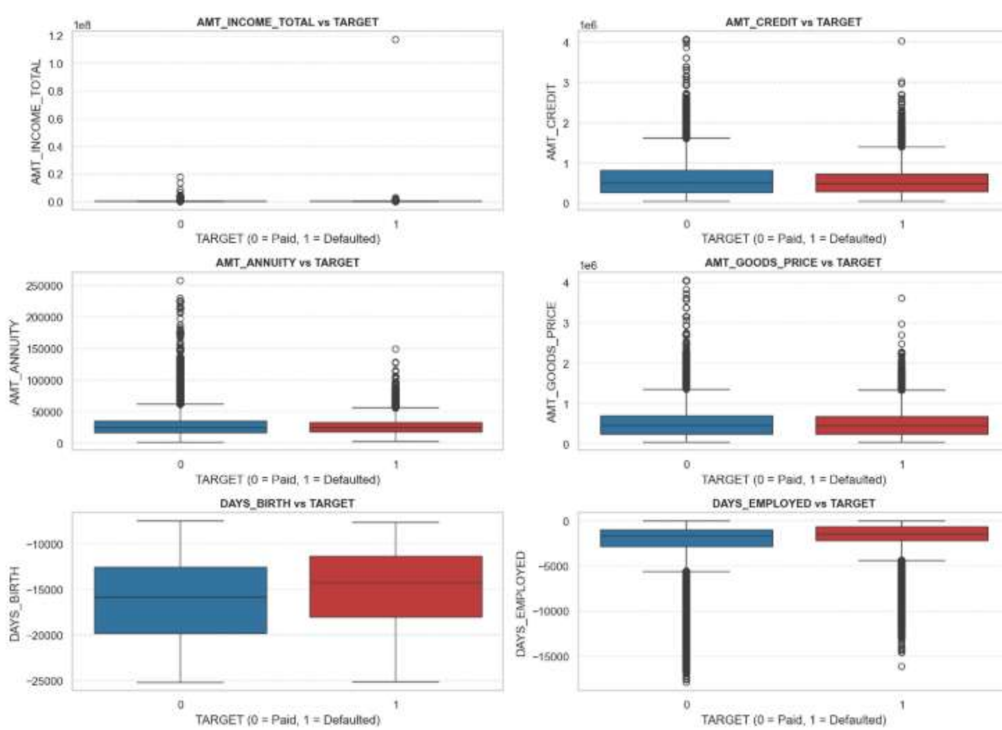
Visualizing Categorical Feature Distributions



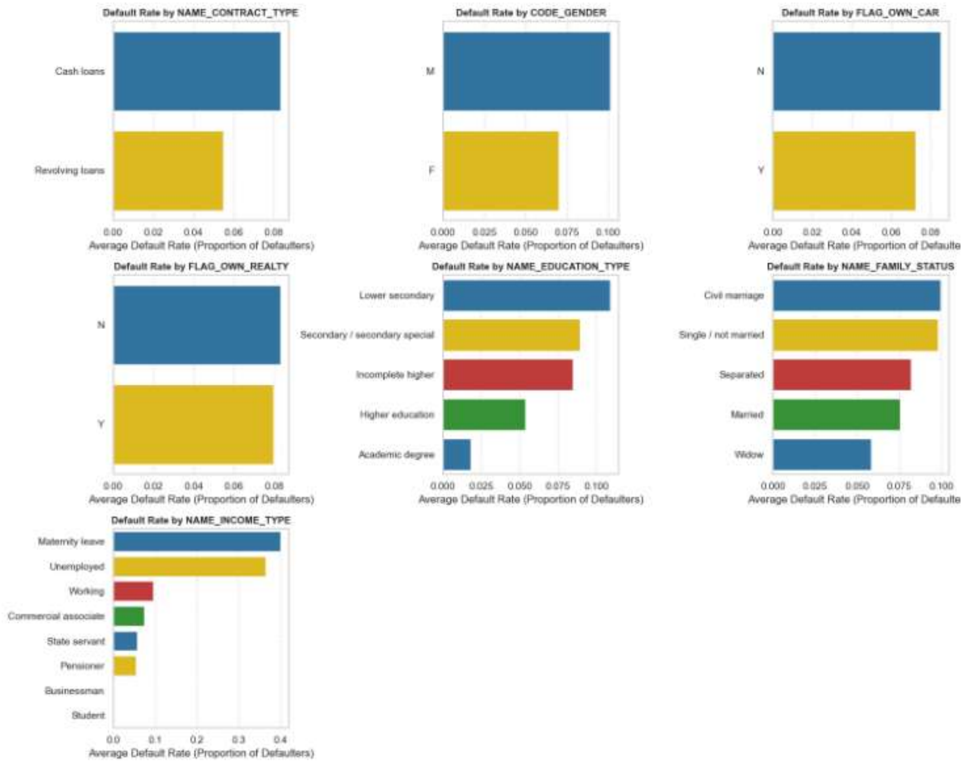
Comparing Numeric Variables by TARGET (Quick View)

Comparing Numeric Variables by TARGET (Quick View)

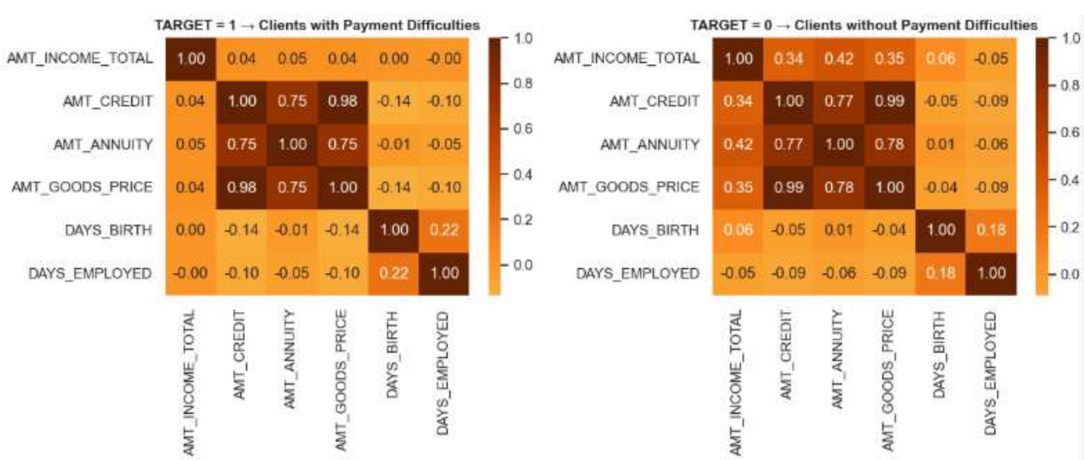
- **TARGET 0:** repaid; **TARGET 1:** defaulted.
- Boxplots show **median**, **IQR (spread)**, and **outliers** per group.
- Check if defaulters have **lower income** (`AMT_INCOME_TOTAL`).
- Examine **higher credit/price** (`AMT_CREDIT`, `AMT_GOODS_PRICE`) among defaulters.
- Assess **payment load**: 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.**



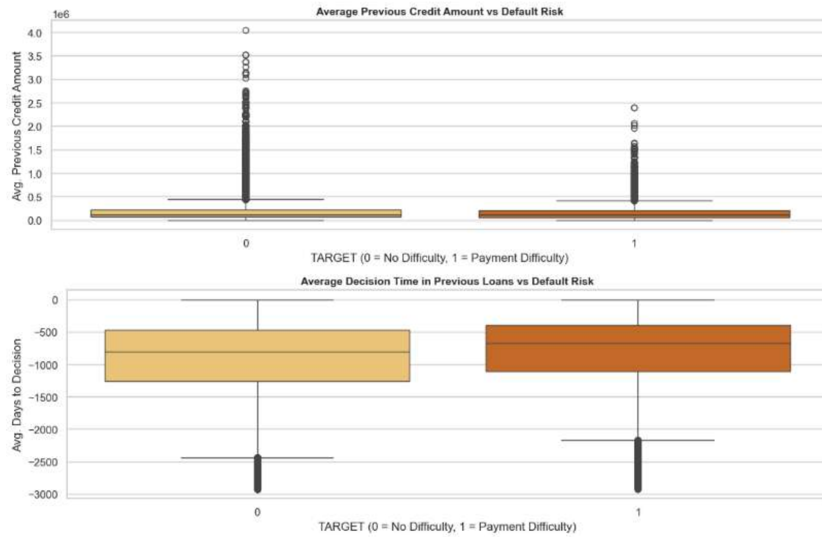
Analyzing Default Rates Across Categorical Variables



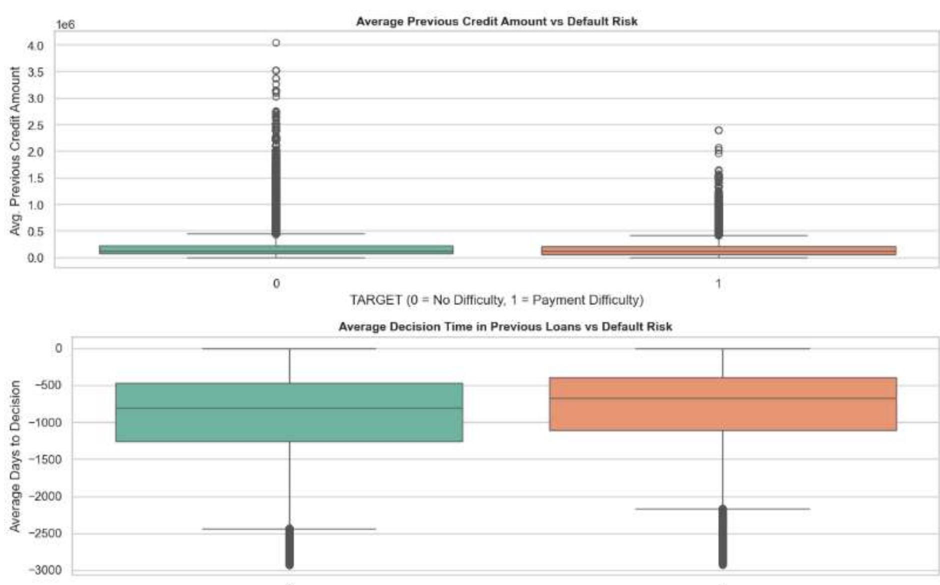
Correlation Among Key Financial Variables

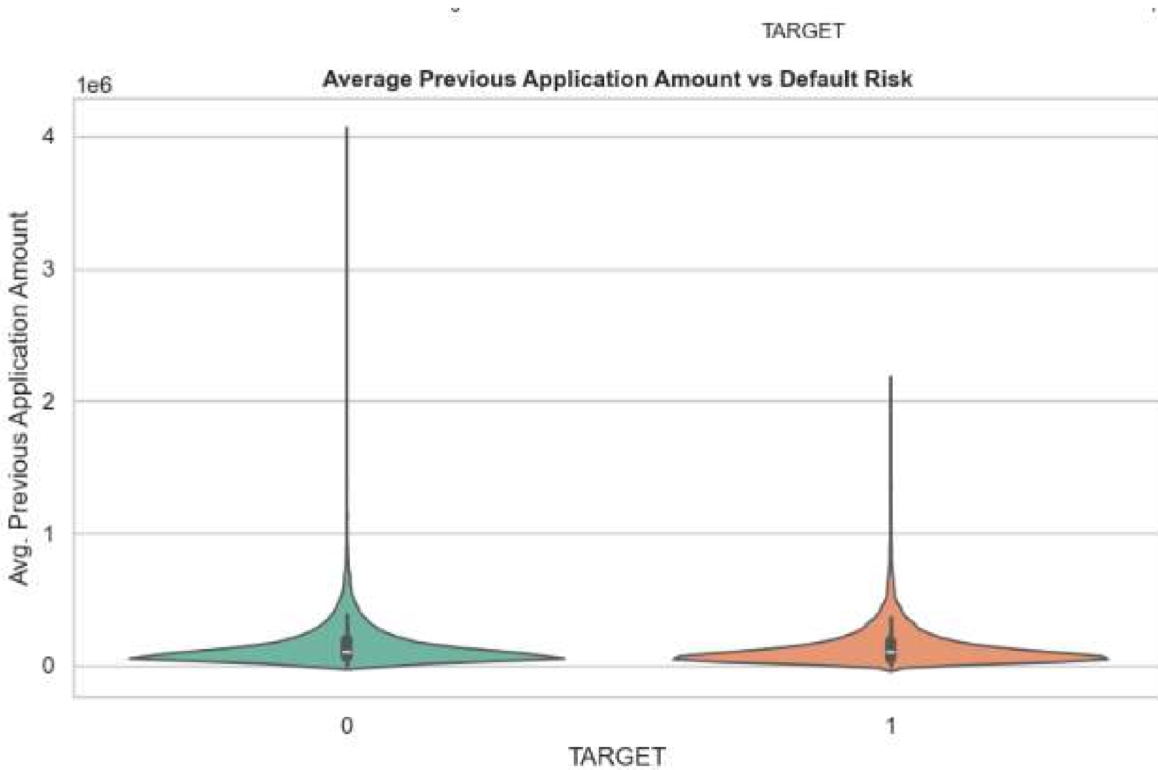


Previous Application Analysis – Brief Summary



Previous Loan Behavior vs 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.