Data Bootcamp

Credit Risk Exploratory Data Analysis

Group 12 | February 2026

Risk Assessment
Data-Driven Insights

Skye Xi
Jonathan Cain Sab
Rajesh Krishnan

02 ▪ Key Variables

Risk Assessment Framework

T

TARGET

Primary Outcome Variable

0	1
On-time Payments	Payment Difficulties

Variable Categories

Employment

DAYS_EMPLOYED, AMT_INCOME_TOTAL

Demographics

YEARS_BIRTH_CATEGORY, CODE_GENDER

Socioeconomic

NAME_EDUCATION_TYPE, NAME_INCOME_TYPE

Family Status

NAME_FAMILY_STATUS

Financial

AMT_CREDIT, AMT_GOODS_PRICE

Historical

NAME_CLIENT_TYPE, NAME_CONTRACT_STATUS

Risk Factor Hierarchy

Structural Stability — High Impact

Employment tenure, education level, age

Financial Magnitudes — Medium Impact

Credit amount, goods price, income

Interaction Effects — High Impact

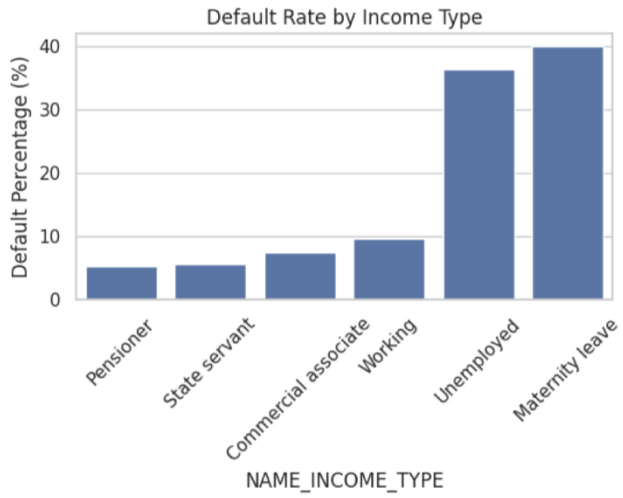
Loan-to-income, tenure-income relationships

Key Insight

Stability indicators provide stronger risk separation than raw financial values alone.

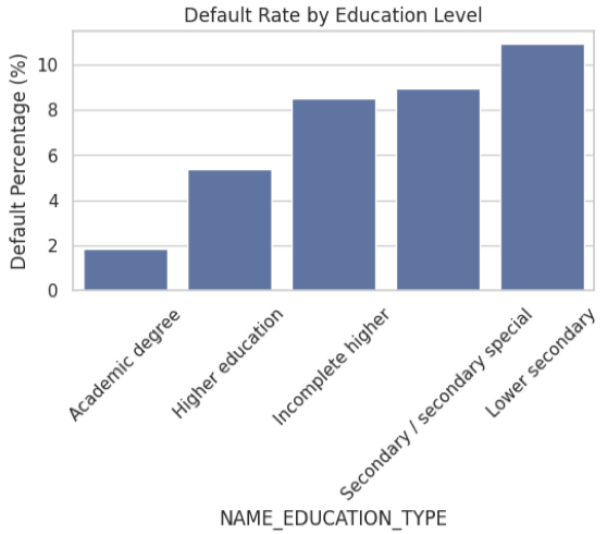
Categorical Variables

Income Type



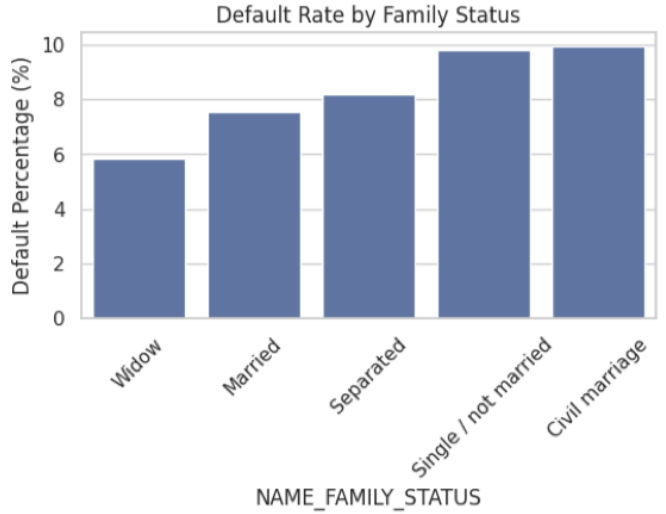
Income Type: Stable income groups (pensioners, state servants) show the lowest default risk (~5–6%), while unemployed or maternity-leave borrowers exhibit the highest (~36–40%), though these estimates are less reliable due to small samples. Core workforce segments fall in the moderate range (~7–10%).

Education Level



Education Level: A clear inverse relationship exists between education and default risk. Borrowers with academic degrees have the lowest default rates (~2%), increasing steadily as education decreases, peaking near ~11% for the lowest education category. Education is therefore a strong risk proxy.

Family Status

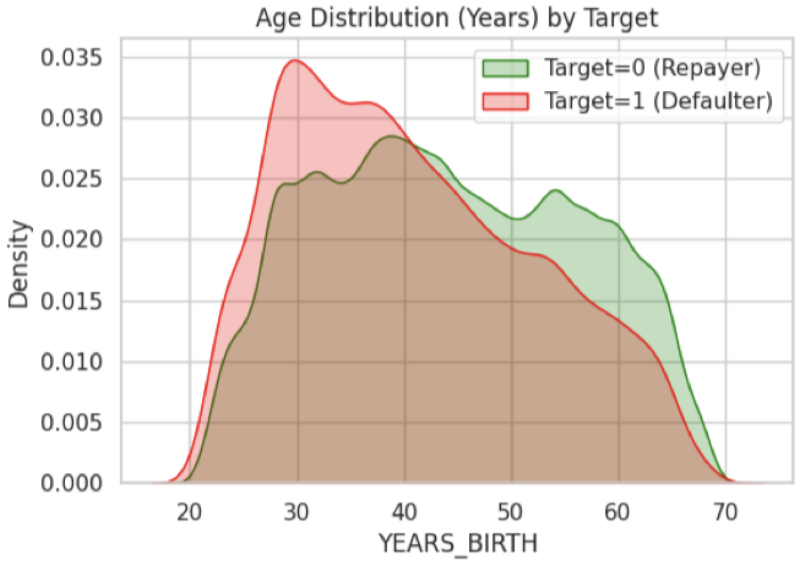


Family Status: Differences are present but less pronounced. Widowed borrowers have the lowest defaults (~6%), while single or civil-marriage borrowers approach ~10%, suggesting household stability modestly affects risk.

Key Finding: Borrowers with stronger financial and structural stability—reflected in stable income sources, higher education levels, and more stable household structures—demonstrate significantly lower loan default risk, making these factors the most powerful predictors of repayment performance.

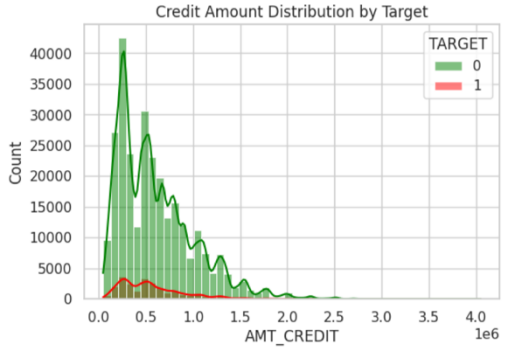
Numerical Variables

Age Distribution by Repayment Status



Key Insight: Defaulters (TARGET=1) concentrated in **25-35 age range**, while repayers show broader middle-age distribution. Younger applicants are overrepresented among defaulters.

Credit Amount Distribution



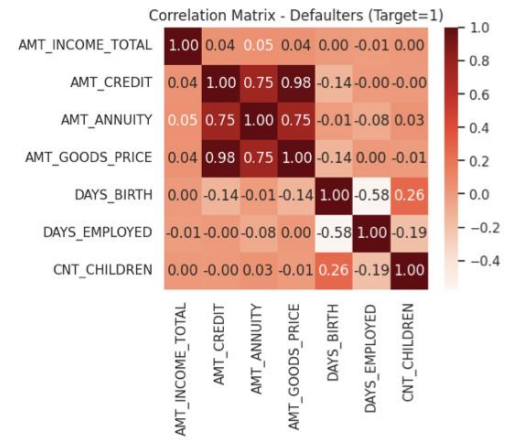
Goods Price Distribution

Loan & Goods Amounts:
Both variables are heavily right-skewed and show similar distributions across defaulters and non-defaulters, suggesting that absolute loan size alone is not a strong discriminator of risk; relative metrics (e.g., loan-to-income ratios) are likely more predictive.

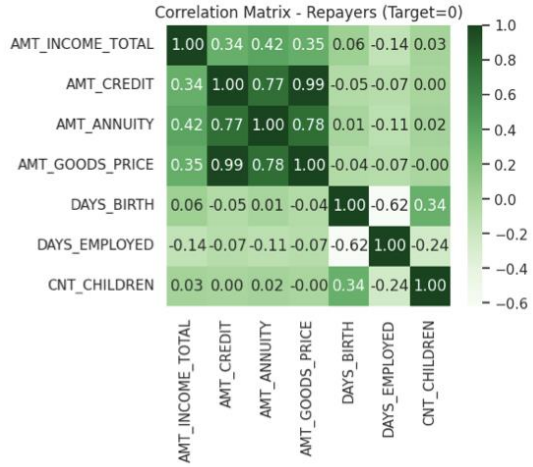
Critical Finding: Age is a strong risk-associated factor, while raw financial magnitudes show limited discriminatory power. Loan-to-income relationships may be more predictive than absolute amounts.

Correlation Structure

Defaulters (TARGET=1)



Repayers (TARGET=0)



Repayers Pattern

Income ↔ Credit: 0.34

Income ↔ Annuity: 0.42

Income-aligned loan sizing observed

Defaulters Pattern

Income ↔ Credit: 0.04

Income ↔ Annuity: 0.05

Weak income association

Income–Loan Alignment:
Among on-time payers, income is moderately correlated with credit amount (0.34) and annuity (0.42), indicating loan sizing is aligned with repayment capacity. In contrast, correlations for defaulters are near zero (0.04–0.05), suggesting weak income-based underwriting within this group.

TARGET	mean	std	min	max
0	169077.722266	110476.268524	25650.0	18000090.0
1	165611.760906	746676.959440	25650.0	117000000.0

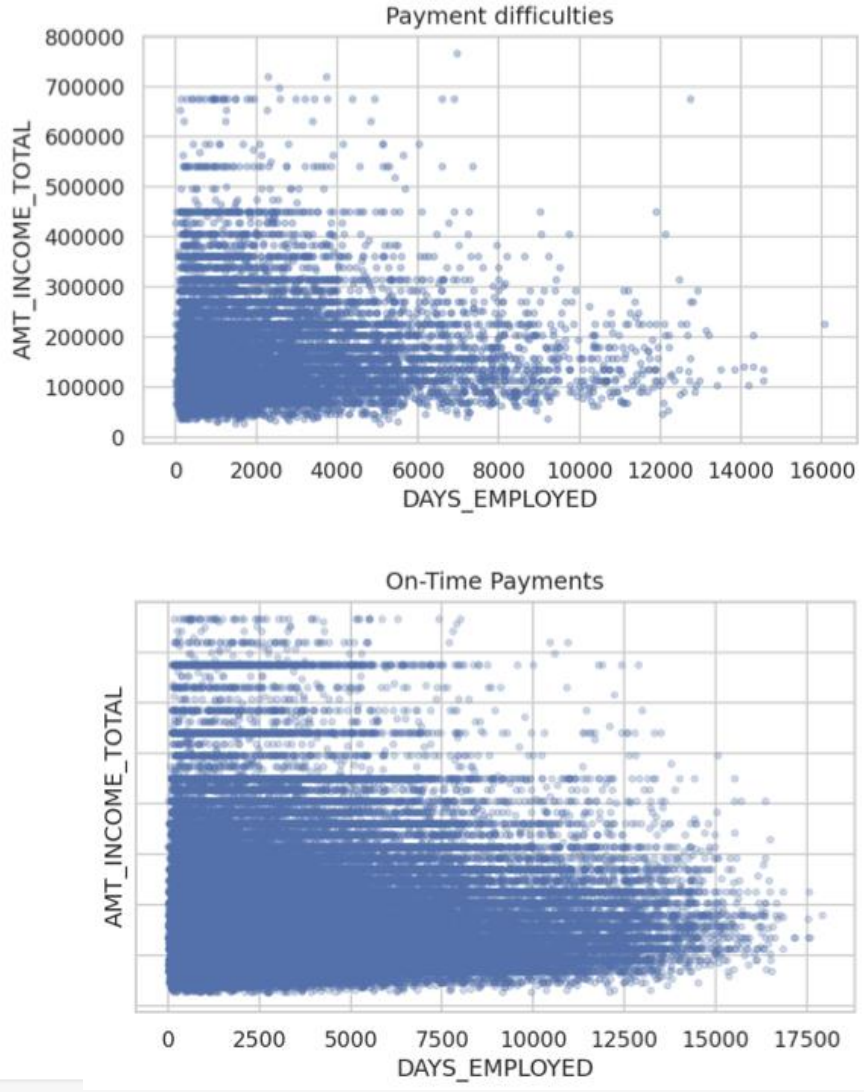
Variance & Heterogeneity:
Both groups have similar income means and ranges, but defaulters exhibit substantially higher variance. This implies greater income instability or heterogeneity among defaulters rather than differences driven by income level alone.

Key Insight

Default risk is driven less by absolute income level and more by how well loan obligations are calibrated to a borrower's repayment capacity. The sharp divergence in correlation patterns indicates that defaulters often received loans misaligned with their income, pointing to underwriting decisions that insufficiently accounted for repayment ability.

Numeric Variable Relationships: Employment Duration vs. Income Analysis

Employment Duration vs. Total Income by Repayment Status



Employment Tenure

Long employment histories are more prevalent among on-time payers, indicating job stability as a key predictor.

Median Tenure:
Repayers: ~8.5 years
Defaulters: ~4.2 years

Income Overlap

Income ranges overlap strongly between repayment groups, suggesting income alone is insufficient for risk assessment.

Income Variance:
Defaulters show higher variance, indicating greater heterogeneity.

Key Conclusion

Tenure appears to be a more informative stability proxy than income alone.

Categorical Variables vs Repayment Outcome

Previous application outcomes by age bucket.

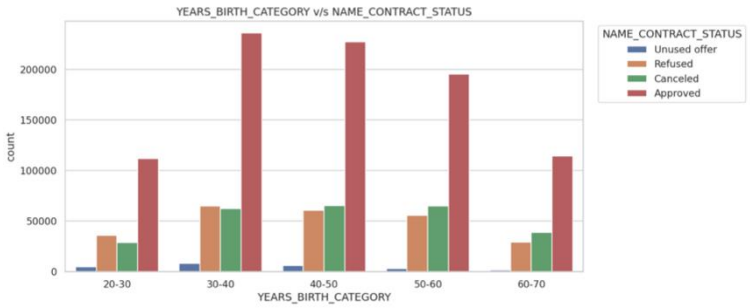


Figure 7: Previous application outcomes by age bucket.

Conclusion: Approval counts are highest for ages 30–50 and decline after age 50; approvals dominate outcomes in all age buckets.

Family Status vs. Previous Application Outcomes

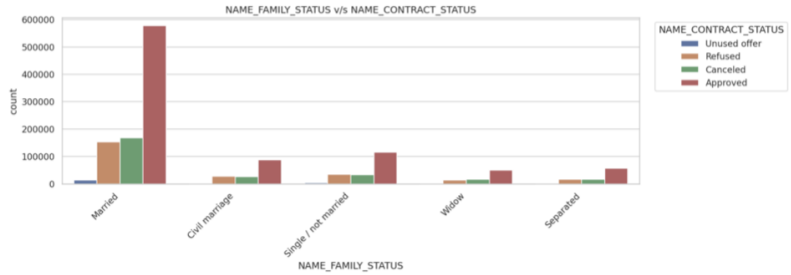


Figure 8: Previous application outcomes by family status.

Conclusion: Married clients account for the largest volume of approvals (count-based), but approvals remain the most frequent outcome across all family-status categories.

Client Type vs. Previous Application Outcomes

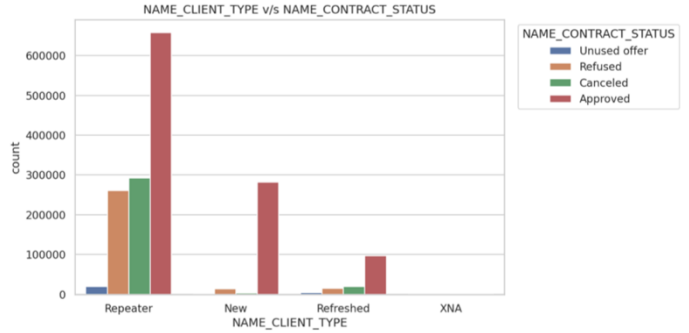


Figure 9: Previous application outcomes by client type.

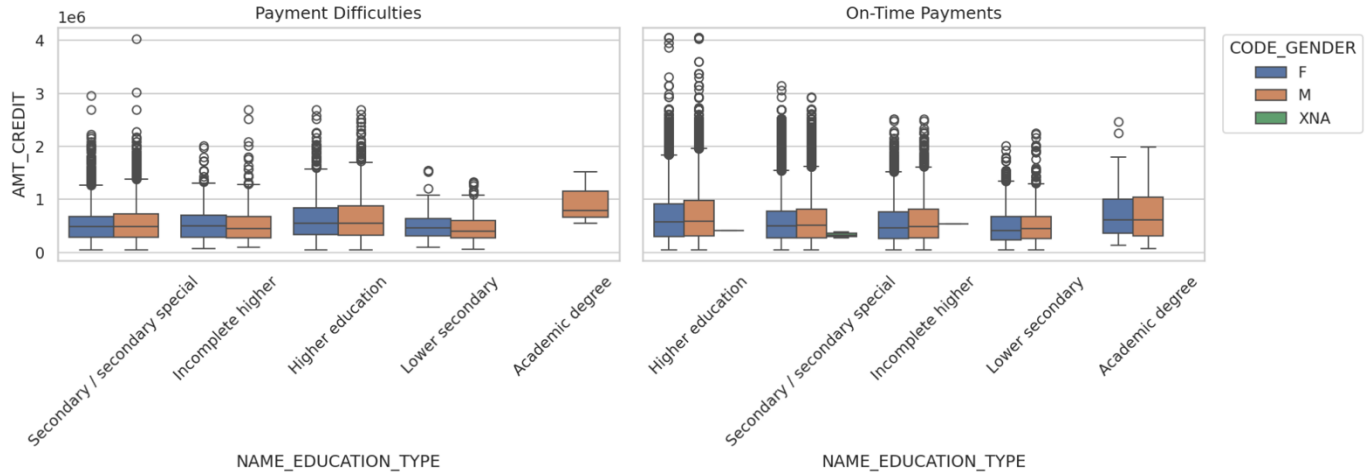
Conclusion: Repeat clients dominate approval counts; however, this is count-based and should be complemented with approval-rate comparisons for inference.

Key Insight

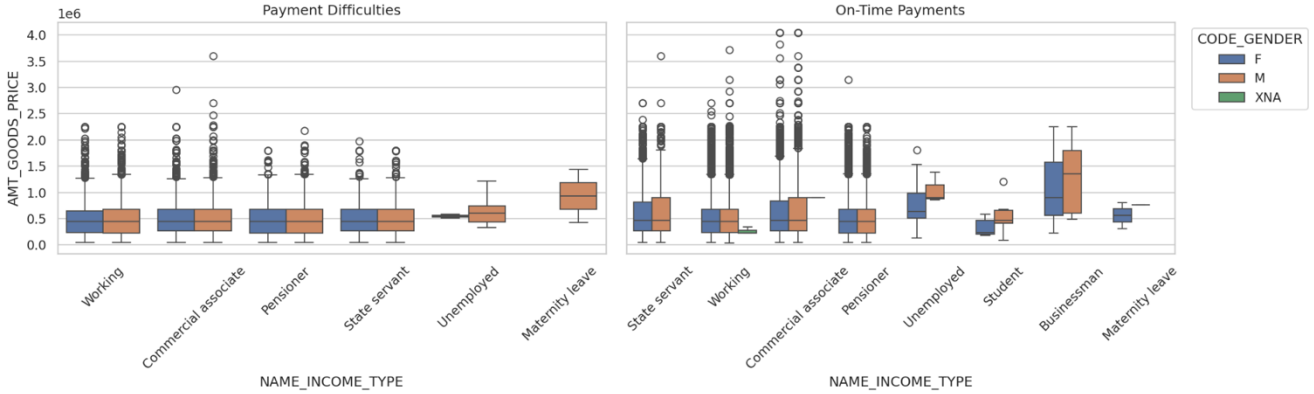
1. Approval counts are highest for ages 30–50 and decline after age 50; approvals dominate outcomes in all age buckets.
2. Married clients account for the largest volume of approvals (count-based), but approvals remain the most frequent outcome across all family-status categories.
3. Repeat clients dominate approval counts; however, this is count-based and should be complemented with approval-rate comparisons for inference.

Socioeconomic Factors vs Financial Magnitudes

Education vs. Credit Amount



Income Type vs. Goods Price



Minor data-quality issue: Small number of unknown-gender (“XNA”) records.

Gender

Higher education levels are associated with larger typical credit amounts; gender differences are **minor** relative to education differences.

Income Overlap

Income source differentiates financed goods price more than gender; business/commercial categories show higher typical goods price, while student/unemployed categories are lower.

Key Conclusion

Income type appears to explain more variation in financed goods price than gender differences.

Actionable Recommendations

1 Risk Segmentation Signals

Stability-related attributes (employment tenure, education level, age) provide stronger separation between repayment groups than raw financial magnitudes. These variables should be prioritized in early screening rules.

✓ **Action:** Implement education-based risk tiers and tenure thresholds

2 Underwriting Alignment

Loan amounts appear income-aligned among reliable borrowers but not among defaulters. The near-zero correlations for defaulters suggest **underwriting misalignment**.

✓ **Action:** Enforce income-consistency checks and debt-to-income limits

3 Feature Prioritization

Relative metrics (e.g., loan-to-income ratio, tenure-income interactions) are likely more predictive than absolute values. Raw credit amounts showed limited discriminatory power.

✓ **Action:** Engineer ratio-based features for scoring models

4 Operational Considerations

Segments with extreme percentages but **very small sample sizes** (e.g., Unemployed at 36-40% default rate, n=22) should not drive policy decisions. Count-dominant groups should be evaluated using rates, not raw volume.

✓ **Action:** Use approval/default rates for count-dominant segments

Strategic Impact: Incorporating structural stability indicators and interaction-based measures can help lenders more effectively distinguish reliable applicants from high-risk ones, directly supporting the business objective of optimizing approval decisions.

Key Takeaways

Data-Driven Risk Optimization

Stability Over Magnitude

Structural stability indicators (employment tenure, education, age) provide stronger risk separation than raw financial values.

Alignment Matters

Income-aligned loan sizing among repayers vs. misalignment among defaulters highlights the importance of underwriting consistency.

Relative Metrics Win

Loan-to-income ratios and interaction-based measures are more predictive than absolute loan amounts or income values.

Core Insight

Incorporating structural stability indicators and interaction-based measures enables lenders to **more effectively distinguish reliable applicants from high-risk ones**, directly supporting the optimization of approval decisions and minimizing both opportunity loss and credit risk.

Data Bootcamp Group 12
Skye Xi ▪ Jonathan Cain ▪ Sab Rajesh Krishnan

February 2026
Credit Risk EDA Report