Credit Risk Analysis

EXPLORATORY DATA ANALYSIS

EDA and Presentation by :
Anirudh Saksena
1 January, 2026

TABLE OF CONTENTS

PROBLEM STATEMENT

Banks face financial loss when customers fail to repay loans on time.

The objective of this analysis is to identify key factors that influence loan default risk using historical loan application data.

By analyzing applicant demographics, income details, credit information, and past behaviour, we aim to:

- Understand patterns between defaulters and non-defaulters
- Identify variables that strongly indicate payment difficulties
- Provide insights that can help banks make better lending decisions and reduce default risk

DATASET OVERVIEW

Data Sources:

- Application Data: Client info at the time of application
- Previous Application Data: Client's past loan history
- Columns Description: Dictionary for understanding variables

General Overview:

- Number of Applications: 307,511
- Number of Features: 122 (after cleaning 73 kept for analysis)
- Target Variable: TARGET (0 = non-defaulter, 1 = defaulter)
- Class Imbalance: ~91.92% non-defaulters, 8.07% defaulters

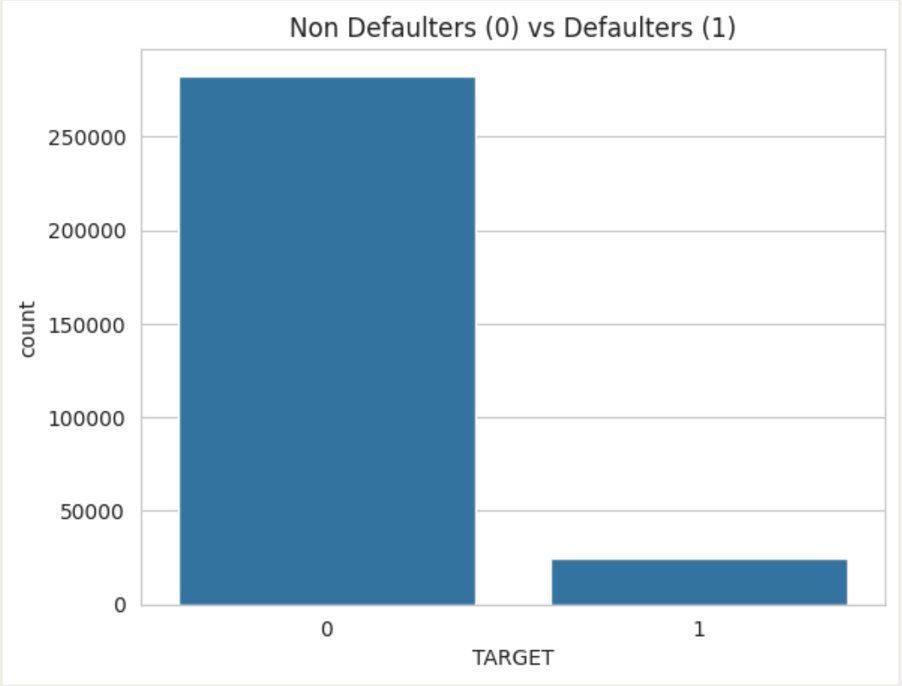
TARGET VARIABLE & DATA IMBALANCE

Target Variable & Data Imbalance

- Target variable: Loan Default (TARGET)
- ~92% non-defaulters, ~8% defaulters
- Data is highly imbalanced

Key Insight

- Default cases are rare but high-risk
- Analysis focuses on differences between defaulters and non-defaulters, not overall averages



HANDLING MISSING DATA

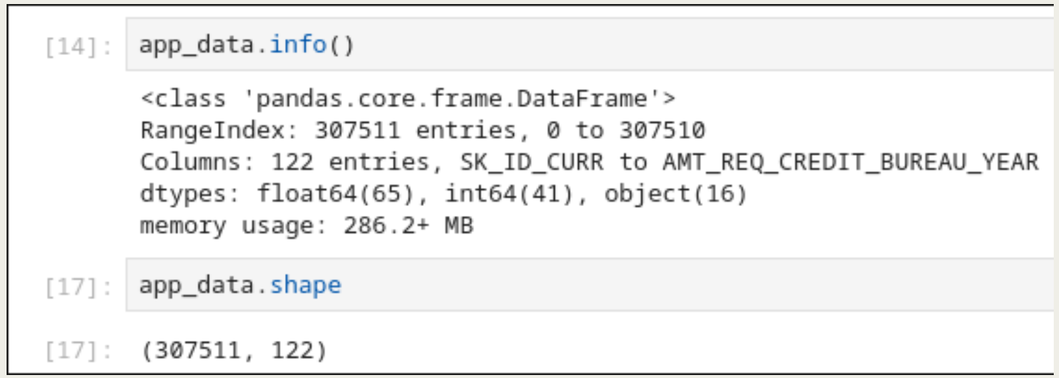
- Several features had very high missing values
- Features with >40% missing were removed
- Features with 5–40% missing were filled using simple methods
- Features with <5% missing were left unchanged

Key Insight

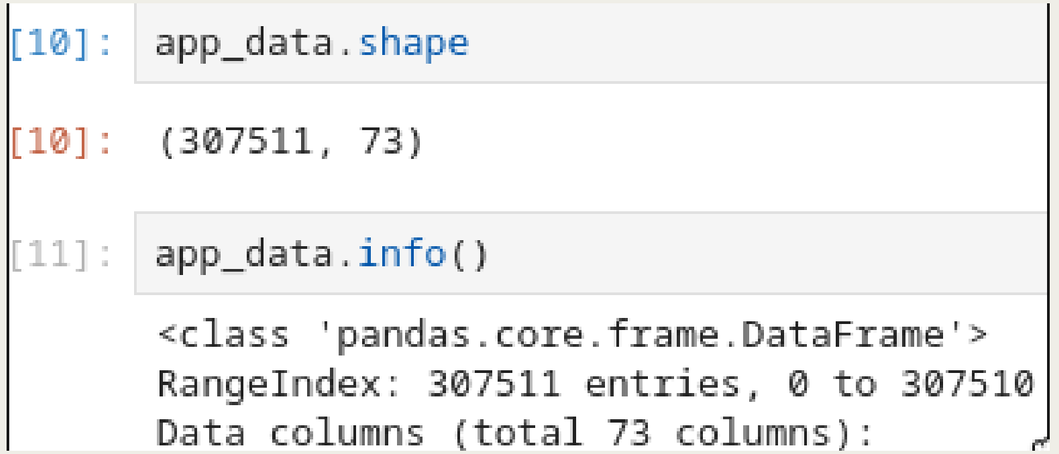
- High-missing features add noise, not value
- Cleaning reduced noise while keeping important risk signals

DATA CLEANING AND FEATURE ENGINEERING

- Columns with more than 40% missing values were dropped as they were not reliable for analysis
- Invalid and placeholder values were handled
- DAYS_EMPLOYED = 365243 was treated as missing
- Time-based columns (DAYS_) were converted into year-based features for easier interpretation
- New financial risk metrics were created:
  - CREDIT_INCOME_RATIO
  - ANNUITY_INCOME_RATIO
- Previous application records were aggregated per applicant to create features such as total applications, approval rate, and average credit amount. These were merged with the main dataset.



(ORIGINAL DATAFRAME)

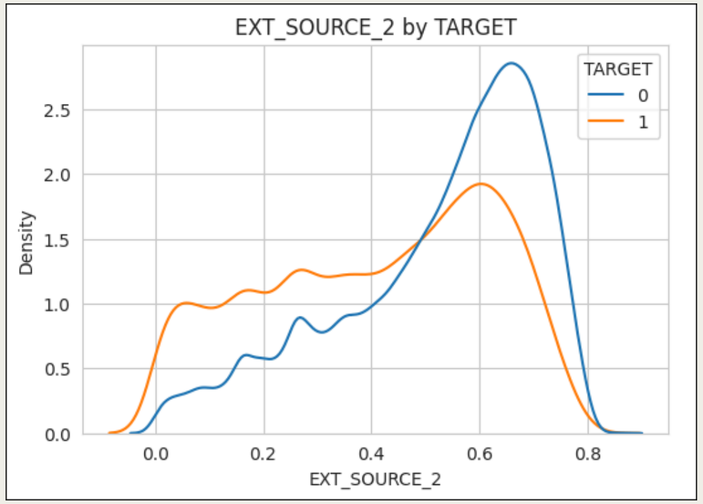


(CLEANED DATAFRAME)

UNIVARIATE ANALYSIS

External Source by Target

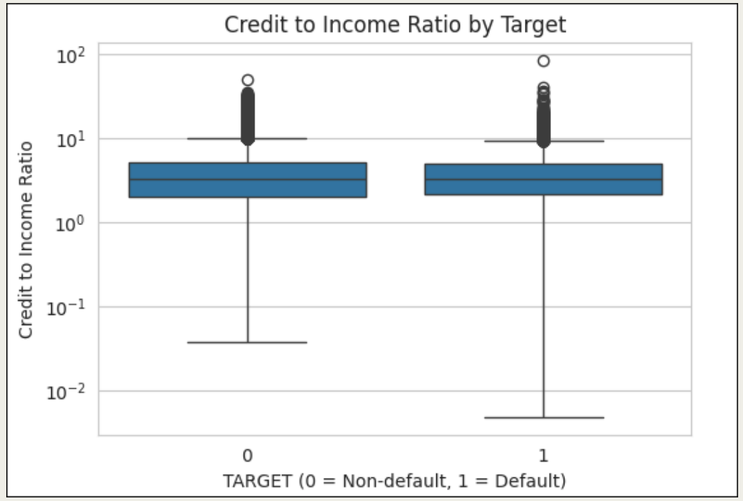
KDE Plot



Defaulters have much lower external credit scores than non-defaulters, showing strong separation.

Credit Income Ratio by Target

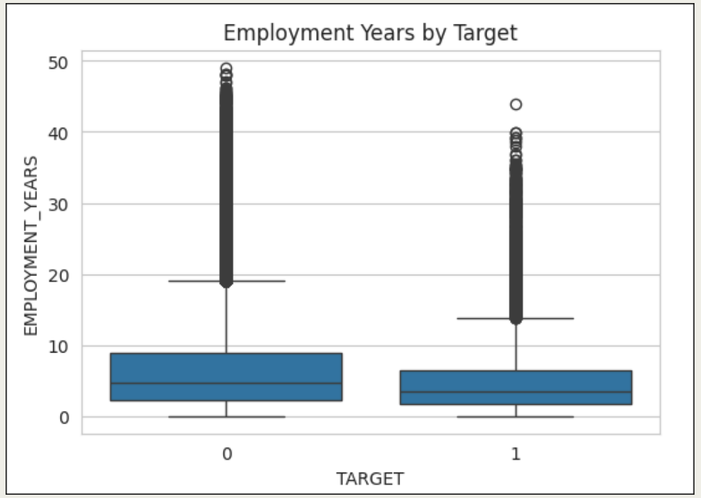
Boxplot



Higher credit-to-income ratios are more common among defaulters, indicating increased repayment burden and higher likelihood of payment difficulty.

Credit Income Ratio by Target

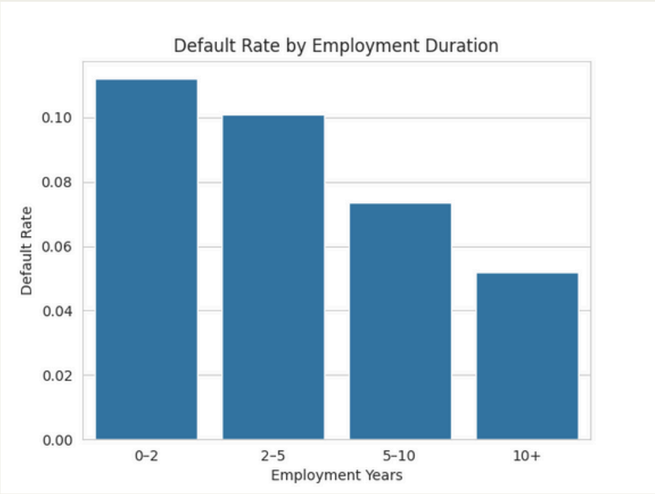
Boxplot



Applicants with shorter employment history default more often, while longer employment duration is associated with lower default risk and greater financial stability.

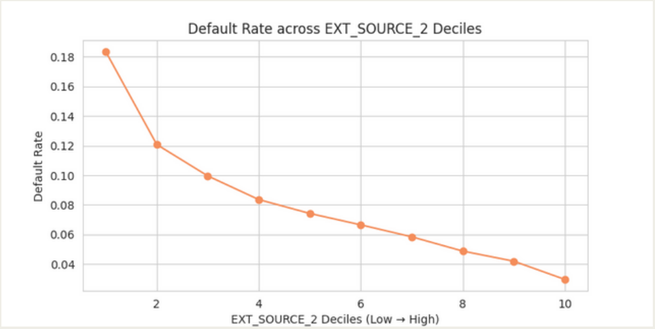
BIVARIATE ANALYSIS

Default Rate by Employment Duration
Barplot



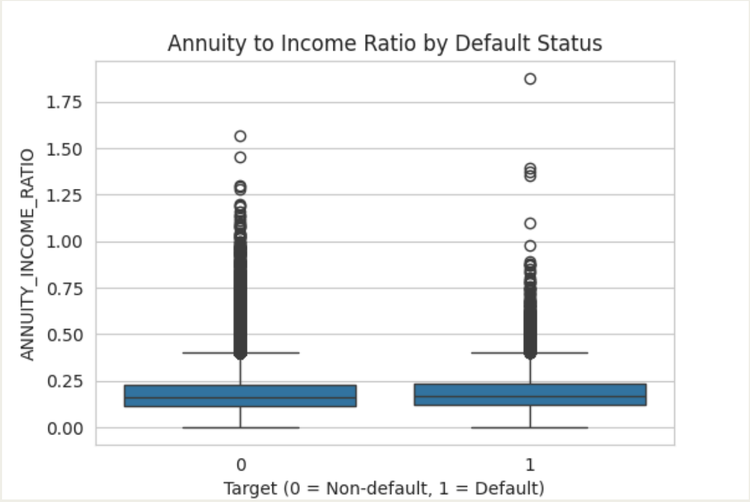
Applicants with **shorter employment** history show much **higher default rates**, especially when combined with high credit burden, highlighting employment stability as a key risk moderator.

Credit Income Ratio by Target
Boxplot



Default rate decreases sharply as external credit score increases, showing a strong non-linear relationship and confirming ***EXT_SOURCE_2*** as one of the most powerful predictors of default risk.

Credit Income Ratio by Target
Boxplot



Defaulters have consistently higher annuity-to-income ratios, indicating heavier monthly repayment pressure even at similar income levels, increasing the likelihood of payment difficulties.

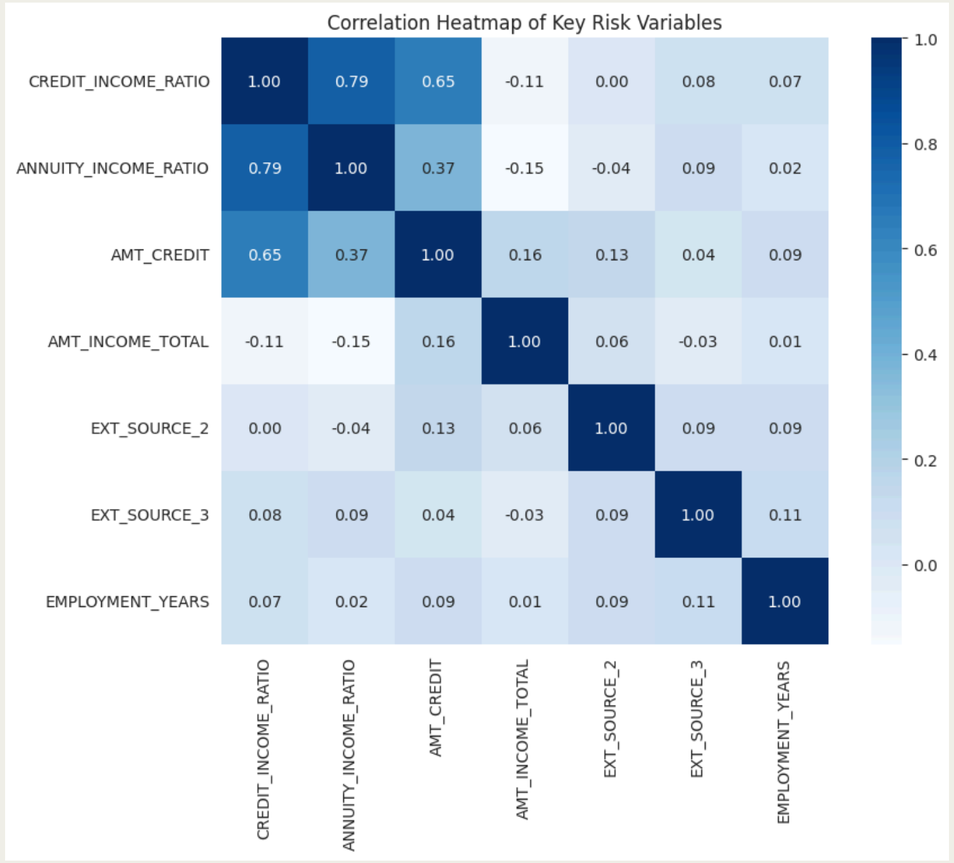
CORRELATION ANALYSIS

The heatmap includes only **key numeric variables** identified from earlier univariate and bivariate analysis.

These features were chosen because they directly represent income, credit burden, external risk scores, and employment stability, which are critical for default risk assessment.

Strong correlations are observed between credit amount, income, and repayment ratios, indicating higher financial stress when credit exposure increases.

External risk scores show consistent relationships with these financial variables, reinforcing their importance as reliable indicators of default risk.



KEY DRIVER VARIABLES

The following variables were identified as key drivers of default risk based on consistent patterns across univariate, bivariate, and correlation analysis:

- **CREDIT_INCOME_RATIO** – higher credit burden increases default risk
- **ANNUITY_INCOME_RATIO** – higher monthly repayment pressure leads to more defaults
- **EXT_SOURCE_2** – lower external scores are strongly linked to default
- **EXT_SOURCE_3** – consistently differentiates defaulters and non-defaulters
- **EMPLOYMENT_YEARS** – shorter employment history shows higher risk

BUSINESS RECOMMENDATIONS

- Apply stricter checks for applicants with high credit or annuity to income ratios
- Use external scores (EXT_SOURCE_2, EXT_SOURCE_3) as primary screening indicators
- Limit loan amount or increase interest rates for applicants with short employment history
- Combine income, credit burden, and employment stability for risk-based pricing

CONCLUSION

Key Takeaways:

- Loan default risk is strongly linked to credit burden and repayment pressure
- External risk scores clearly separate defaulters from non-defaulters
- Employment stability reduces the likelihood of default

Thank you!

ANIRUDH SAKSENA, 1ST JANUARY, 2026