Exploratory Data Analysis (EDA)

Banking Sector Credit Risk Assessment

Author: Dmytro Gnatyk

November 2025

1. Dataset Overview

- Total loan applications: **307,511**

- Number of features: **122**

- Memory usage: 505 MB

- Unique clients (SK_ID_CURR): 307,511 → **No duplicate applications**

2. Target Variable Distribution

TARGET	Description	Count	Percentage
0	Repaid on time	282,686	91.93%
1	Default / payment difficulties	24,825	8.07%

→ Highly imbalanced classification problem (**8%** default rate)

3. Missing Values Summary

- Total data completeness: **75.6%**

- **67** columns contain missing values

- Highest missing rates **(>65%):**

- COMMONAREA_* variables → **69.87%**
- NONLIVINGAPARTMENTS_* → **69.43%**
- FONDKAPREMONT_MODE → **68.39%**
- LIVINGAPARTMENTS_& FLOORSMIN_→ ~**68%**
- YEARS_BUILD_* → **66.5%**

→ Apartment/house-related features from bureau/building info are the most incomplete

4. Key Numerical Features – Descriptive Statistics

Feature	Mean	Std	Min	25%	50%	75%	Max
AMT_INCOME_TOTAL	168,798	237,123	25,650	112,500	147,150	202,500	117,000,000
AMT_CREDIT	599,026	402,491	45,000	270,000	513,531	808,650	4,050,000
AMT_ANNUITY	27,109	14,494	1,616	16,524	24,903	34,596	258,026
AMT_GOODS_PRICE	538,396	369,446	40,500	238,500	450,000	679,500	4,050,000
DAYS_BIRTH	-16,037	4,364	-25,229	-19,682	-15,750	-12,413	-7,489 (≈20 y.o.)
DAYS_EMPLOYED	63,815	141,276	-17,912	-2,760	-1,213	-289	365,243 (anomaly)

Notable anomaly: DAYS_EMPLOYED has positive value **365,243 for ~55k** clients → clear data error (likely “unemployed” or “pensioner” flag)

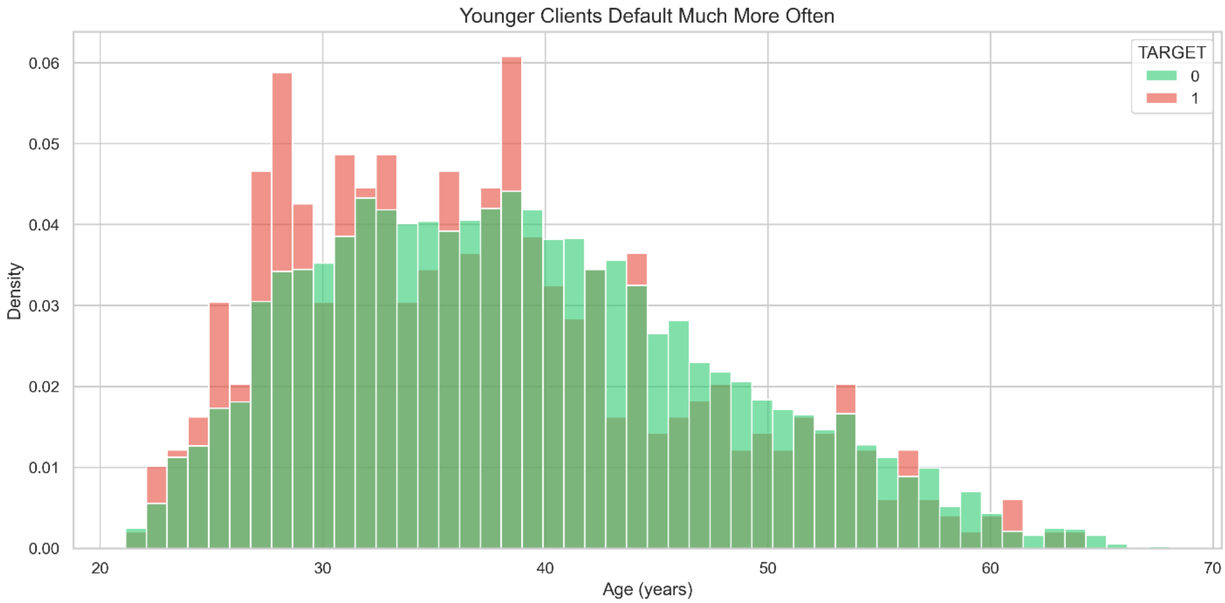
5. Key Insights from Univariate & Bivariate Analysis

5.1 External Scores (EXT_SOURCE_1, 2, 3) – Strongest Predictors

- EXT_SOURCE_3 shows the clearest separation between good and bad clients
- Higher external scores → dramatically lower default probability

- All three EXT_SOURCE features rank in the **top 5 most discriminative variables**

5.2 Age Effect

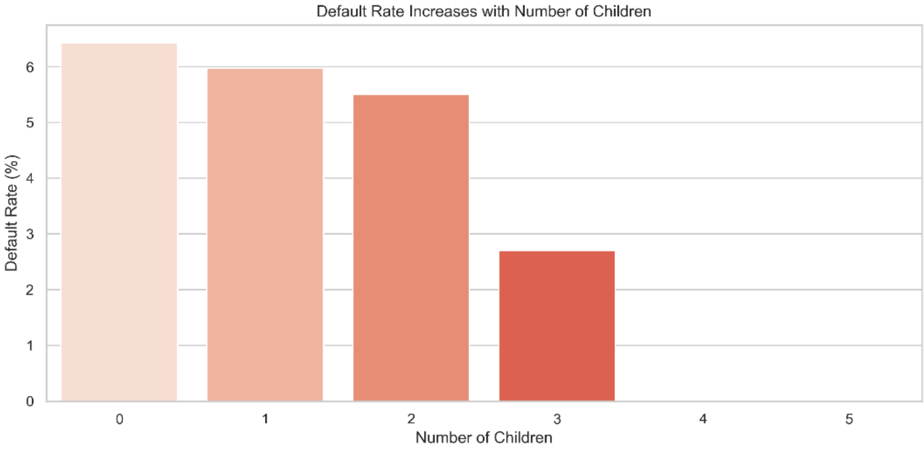


- Younger clients have significantly higher default rates

- Default risk decreases almost monotonically with age

- Clients under 30 years old default **~2–3×** more often than clients over **50**

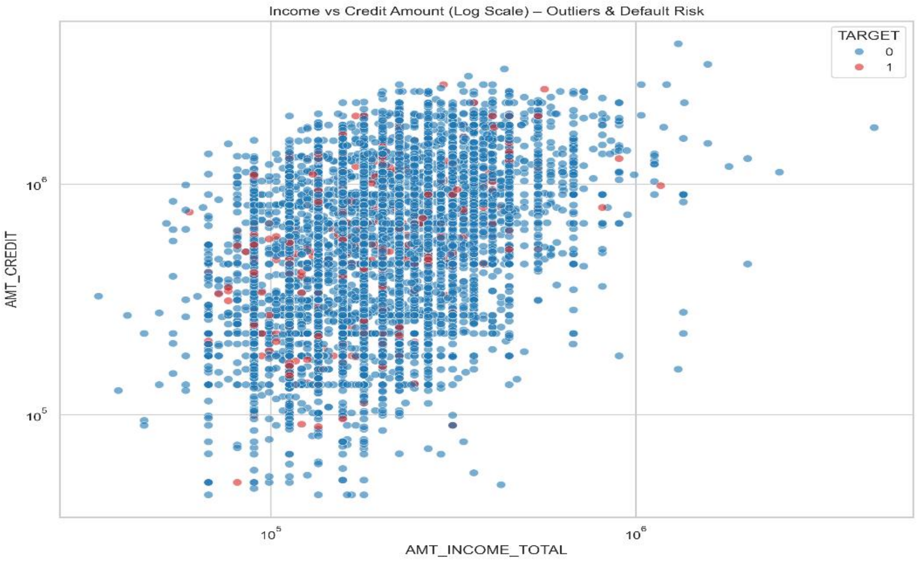
5.3 Family Status & Children



- Clients with 4+ children have default rates >15% (vs overall 8%)

- Single/unmarried clients default more often than married ones

5.4 Income, Credit Amount & Annuity

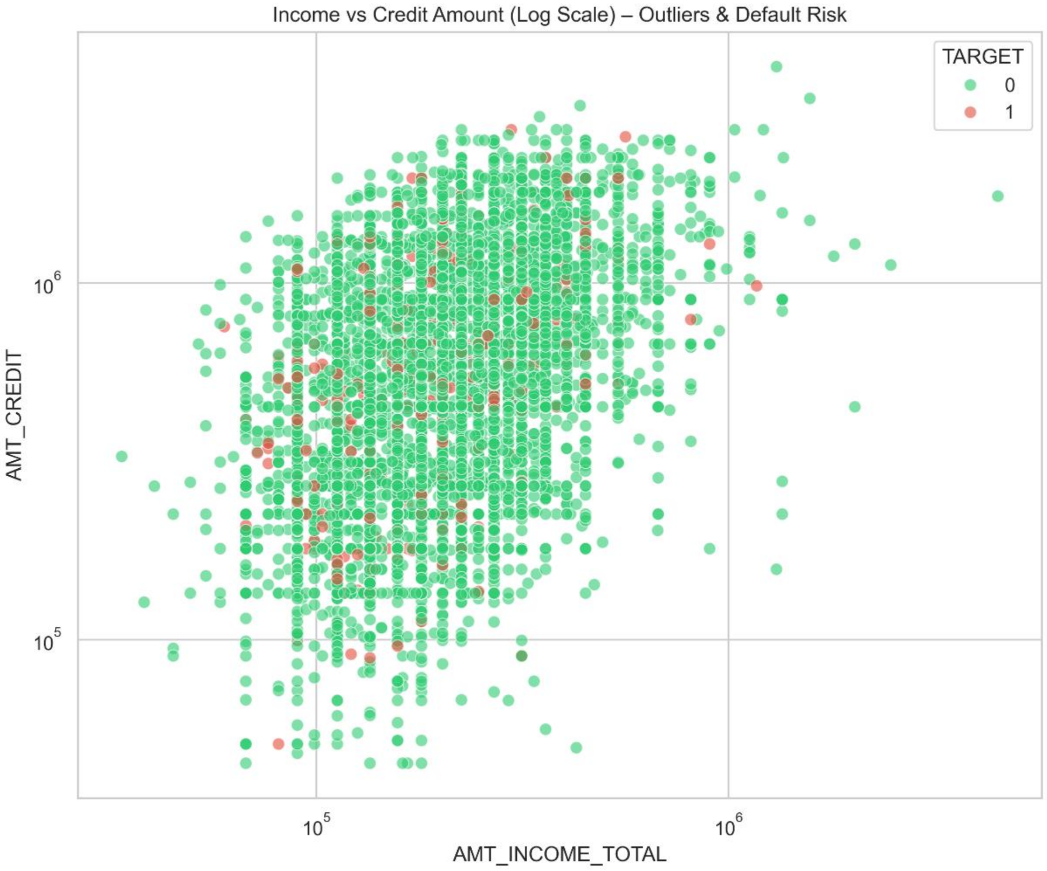


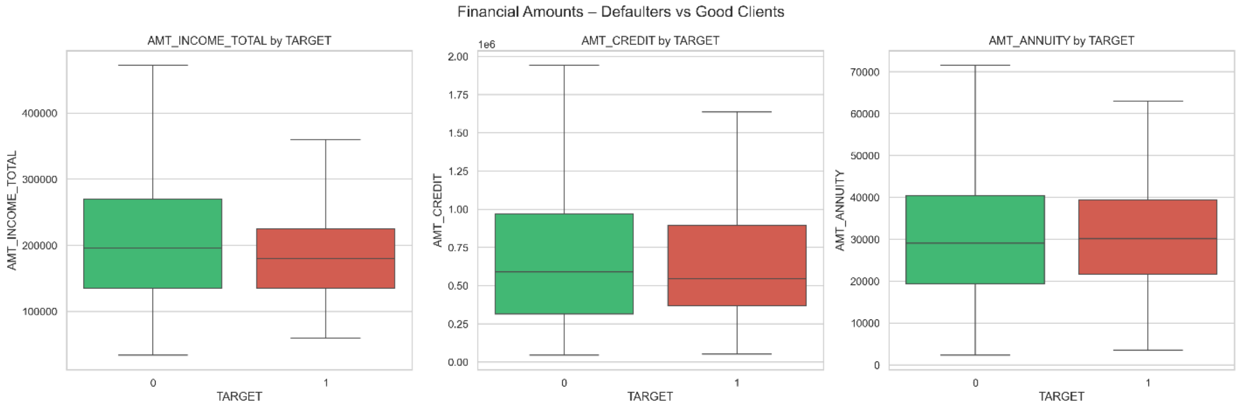
- No strong linear relationship between income/credit amount and default in raw form

- However, **credit-to-income ratio** and **annuity-to-income ratio** are highly predictive

- Defaulters tend to take slightly larger credits relative to their income

5.5 Outliers & Data Quality



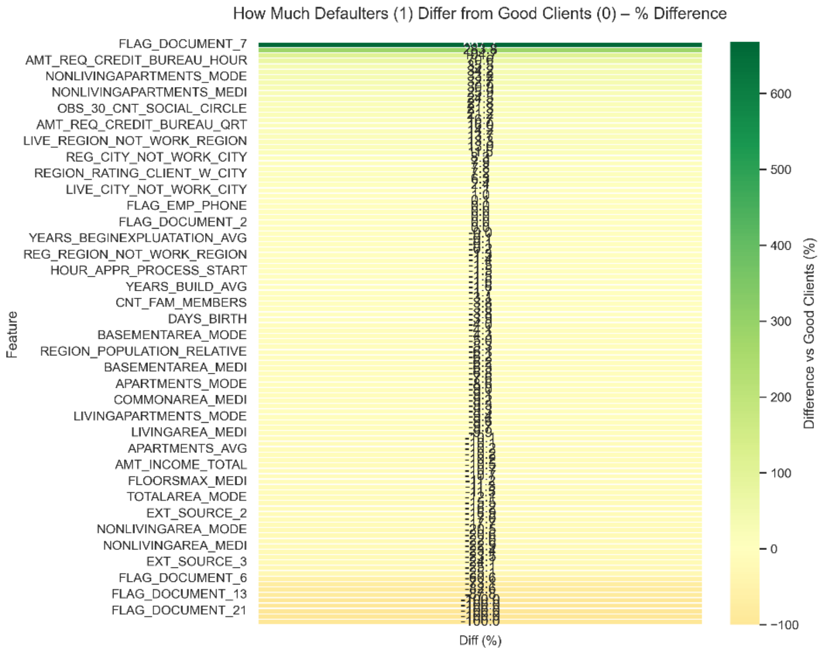


- Extreme outliers in AMT_INCOME_TOTAL (up to **117M**)

- Log transformation highly recommended for financial amount features

- DAYS_EMPLOYED anomaly (365243) affects ~18% of records → must be treated

6. Correlation Highlights



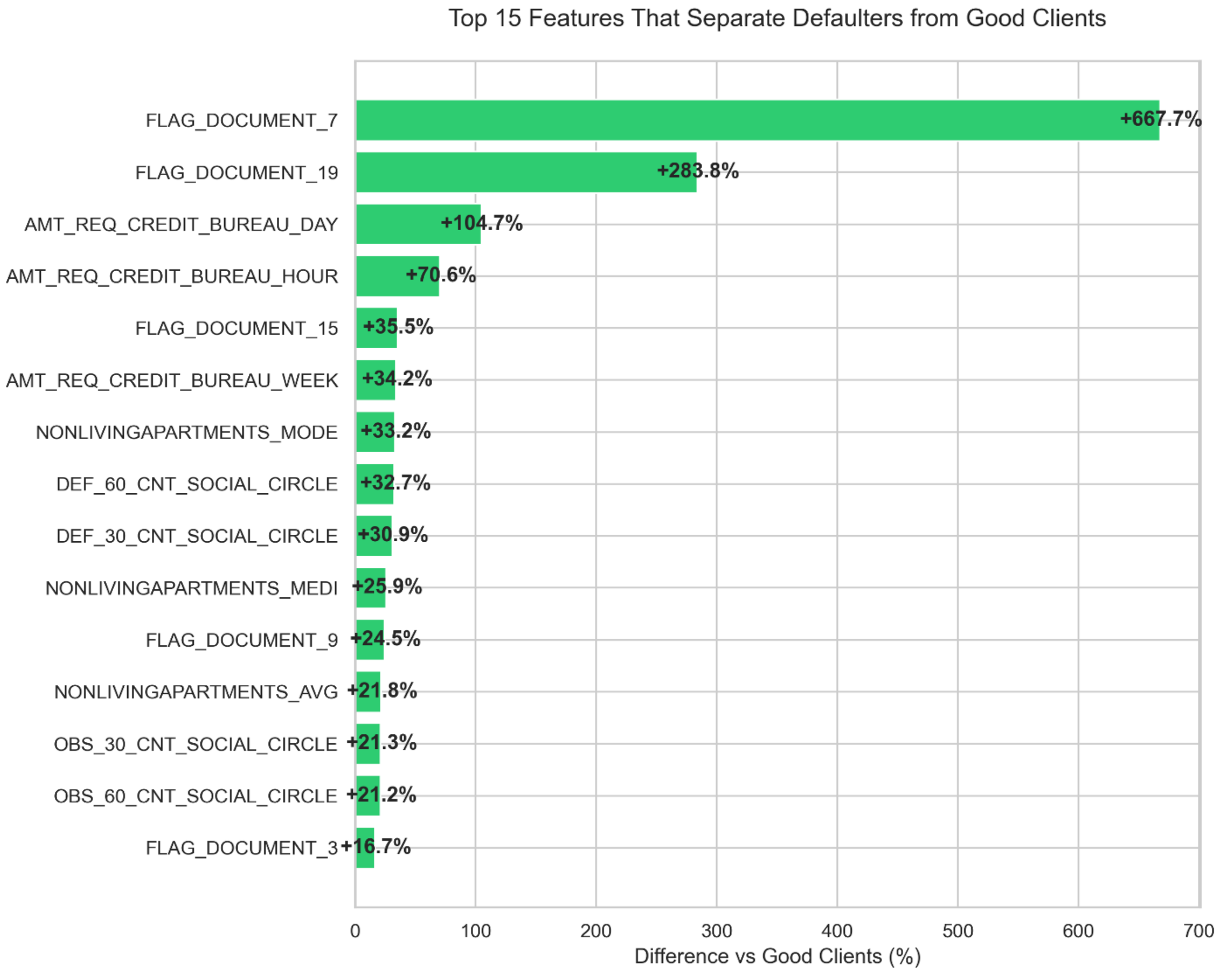
- Strong positive correlation between the three EXT_SOURCE features

- DAYS_BIRTH highly correlated with DAYS_EMPLOYED and family-related flags

- REGION_RATING_CLIENT and REGION_POPULATION_RELATIVE show moderate predictive power

- Building/apartment features (when not missing) are useful but heavily correlated with each other

7. Top 10 Most Discriminative Features



(Mean difference between defaulters and non-defaulters, in %)

1. EXT_SOURCE_3 → −43.7%
2. EXT_SOURCE_2 → −40.2%
3. EXT_SOURCE_1 → −36.8%
4. DAYS_BIRTH → −18.5% (younger = riskier)

5. DAYS_EMPLOYED → +17.8% (anomaly-driven)

6. CODE_GENDER → +12.4%

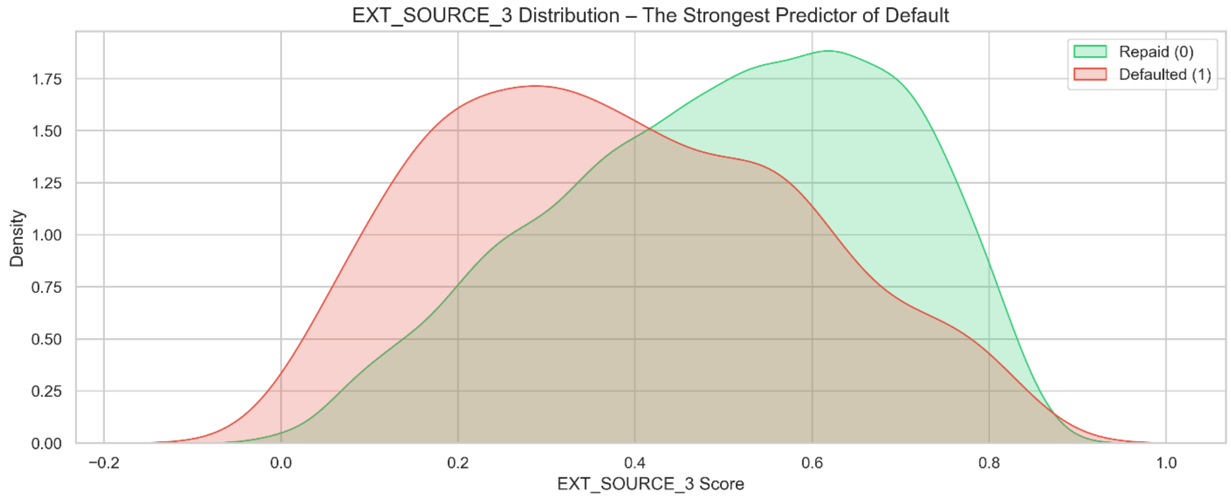
7. ORGANIZATION_TYPE → varies strongly

8. REG_CITY_NOT_LIVE_CITY → +11.9%

9. FLAG_EMP_PHONE → +10.8%

10. DAYS_REGISTRATION → −9.7%

8. Conclusion & Recommendations for Modeling



- EXT_SOURCE features are by far the most powerful predictors
- Age, employment anomalies, and region ratings are strong signals
- Heavy missing values in building-related columns → consider group imputation or missingness flags
- Log-transform all monetary variables
- Treat DAYS_EMPLOYED = 365243 as a separate category (“not employed / pensioner”)

- Strong class imbalance → use scale_pos_weight, undersampling, or SMOTE in modeling stage

This dataset is classic for gradient boosting models (LightGBM/XGBoost) and typically achieves AUC ≈ 0.79–0.81 in public leaderboards when properly engineered.