Credit Risk Analysis

Exploratory Data Analysis for Loan Default Prediction

Problem Statement

- Loan companies face challenges lending to clients with limited credit history.
- Clients may take advantage by defaulting.

- **Objective**: Use EDA to identify attributes that indicate a client's likelihood of defaulting.
- Help avoid denying credit to good customers and approving risky ones.
- Aim to protect good customers and reduce risky approvals.

Dataset Overview

application_data.csv	Loan application info at submission
previous_application.csv	History of prior loan attempts and results
columns_description.csv	Data dictionary for feature understanding
TARGET	Label : 1 = payment difficulty, 0 = all others

Handling Missing Data

- Missing values were analyzed column-wise.
- Columns with more than 50% missing data were removed as they carry limited predictive power and can introduce noise.
- For columns with less than 50% missing data

- **Numerical values** : replaced with median.
- **Categorical values** : replaced with mode.
- Ensured consistency and preserved distribution shape.

Outlier Detection

- Used boxplots and IQR method to detect outliers.
- Found extreme values in AMT_INCOME_TOTAL and AMT_CREDIT.
- Outliers retained to preserve data distribution.
- Detected Outliers In The Given DataSets.

Class Imbalance (Target)

- TARGET = 0 : 91.9% (Non-defaulters)
- TARGET = 1 : 8.1% (Defaulters)
- Imbalance visualized using bar and pie charts.
- Important to handle during modeling.

Univariate & Segmented Analysis

- Used KDE plots to compare distributions across target classes.
- Defaulters tend to have lower income.
- External scores like EXT_SOURCE_2 are strong differentiators.
- Automating Segment Analysis (with Loops)
- This approach enabled efficient exploration of trends for both defaulters and non-defaulters.

Bivariate Analysis

- Explored relationships : AMT_CREDIT vs AMT_ANNUITY, EXT_SOURCE scores.
- Strong correlation between credit amount and annuity.
- Defaulters have lower EXT_SOURCE_2 and EXT_SOURCE_3 scores.
- Compared Age vs Employment days

Top 10 Correlation Analysis

- Non-defaulters: EXT_SOURCE_2, EXT_SOURCE_3, AMT_CREDIT.
- Defaulters: EXT_SOURCE_1, EXT_SOURCE_2, DAYS_BIRTH.
- Used segmented correlation analysis by TARGET 0 & 1.
- Plotted Graph for both TARGET 0 & 1 For Easier Understandings.

Previous Applications Insights

- Refused previous loans → **2.3x** more likely to default.
- Loan type and credit channel influenced risk.
- Canceled or unused offers generally safer segments.
- To better understand the influence of variables, the “columns_description.csv” file was used.
- Plotted Countplot For Better Understandings.

Business Implications

- Clients with payment issues have **40%** lower median income.
- Top correlation for reliable clients: `AMT_CREDIT` vs `AMT_ANNUITY` (**0.95**).
- Clients with previous refused loans are **2.3x** more likely to default.
- External source (`EXT_SOURCE_2/3`) are critical for risk assessment.
- Use risk-based pricing and targeted rejection.
- Recommend data-driven credit scoring.

Conclusion

- EDA revealed patterns between client attributes and default risk.
- Strong predictors: EXT_SOURCE_x, DAYS_BIRTH, AMT_CREDIT.
- **Next** :
- Build predictive models and apply insights.
- Implement stricter checks for high-risk groups.