OGTIP Internship Project: Python

EDA - Exploratory Data Analysis for Loan Default Risk Prediction

Ramya Panchatcharam
15/09/2024

Table of contents

Introduction

Project Focus : Apply Exploratory Data Analysis (EDA) to identify patterns predicting loan repayment difficulties.

Objective: Minimize lending risk by identifying factors indicating loan default likelihood.

Outcome: Assist in loan approval decisions based on repayment likelihood.

Objective

EDA Techniques: Identify key factors behind loan defaults.

Goal: Enhance risk assessment to approve applicants with higher repayment potential.

Project Approach

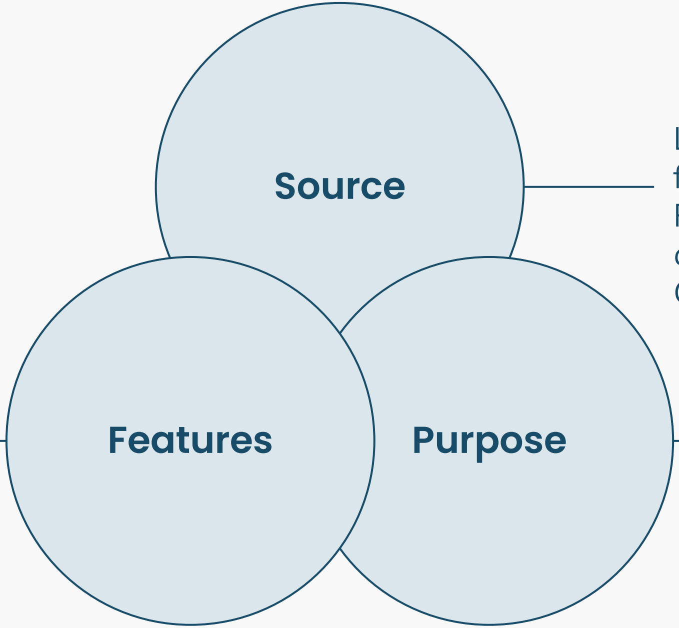
Identify Key Variables: Predictive of loan defaults.

Data Cleaning: Ensure reliable and accurate analysis.

Explore Relationships: Find indicators of default risk.

Summarize Findings: Provide actionable insights for risk management.

About the Dataset



Source: Loan records from a financial institution. Main File: Previous_application, application_data, and Column Description

Features: Client profiles, payment history, loan amounts, interest rates, etc.,

Purpose: Understand borrower characteristics and create effective risk mitigation strategies.

Key Concepts and Challenges

Exploratory Data Analysis (EDA):

Discover patterns in data.

Risk Analytics:

Assess variables affecting loan default risk.

Data Preprocessing

Clean and transform data for accurate results.

Decision-Making

Use data-driven insights to inform lending decisions.

Financial Risk Management:

Use analysis to minimize default risk.

EDA - Exploratory Data Analysis

Steps for EDA – Overview

• Import Libraries	• Multivariate Analysis
• Load Data	• Outlier Detection
• Initial Data Exploration	• Data Distribution
• Handle Missing Values	• Relationship Between Days-Based Variables and Risk
• Univariate Analysis	• Data Quality Checks
• Bivariate Analysis	• Check Target Value for Imbalance
• Time-Based Analysis	• Risk Analytics
• Target Variable Analysis	• Merge Datasets

Steps for EDA

Import Libraries

Import key libraries: pandas, numpy, matplotlib, seaborn, etc

```
# load the previous data -
previous_application=pd.read_csv('previous_application.csv')
```

```
# load the application data -
application_data=pd.read_csv('application_data.csv')
```

Load Data

Load datasets into Pandas DataFrame for analysis.

```
# Shape of previous_application data
previous_application.shape
```

```
(1670214, 37)
```

```
# Shape of application data
application_data.shape
```

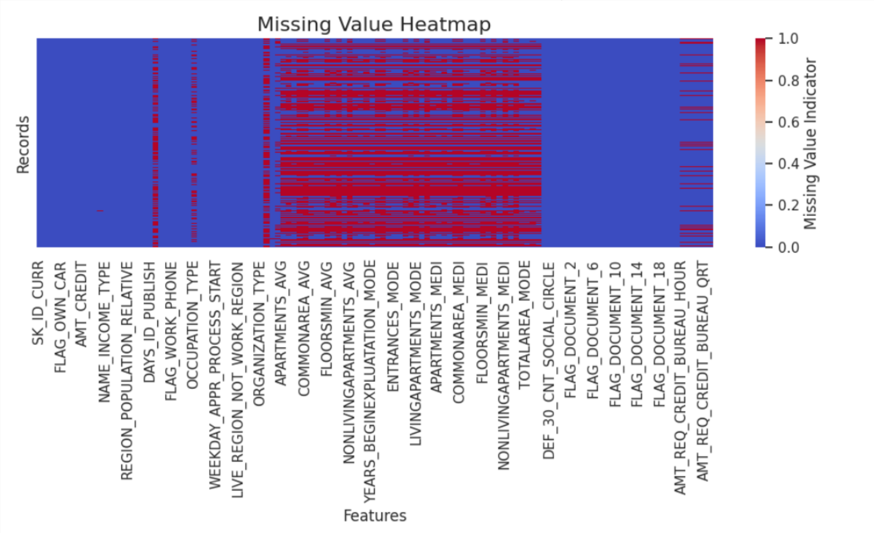
```
(307511, 122)
```

Shape of Data

Size of the data find using the shape method

Steps for EDA – Handling Missing Values

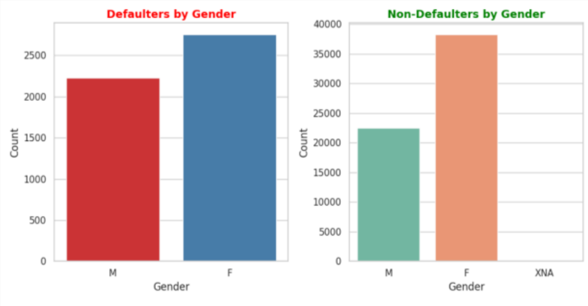
- Identify missing data using isnull().sum().
- Treat missing values with techniques like fillna(), dropna(), or imputation.



Use a seaborn heatmap to visualize missing values in application_data, with red indicating high missingness and blue showing complete columns.

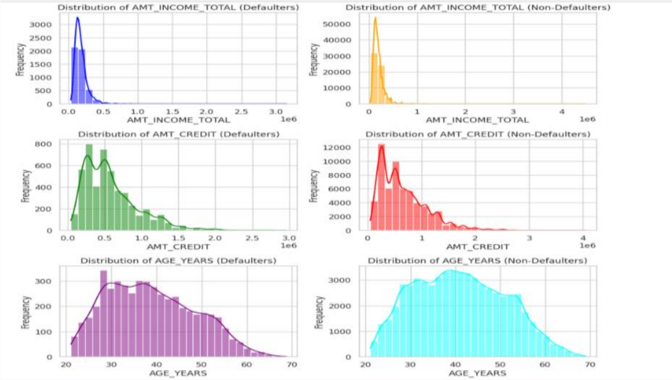
Steps for EDA – Univariate Analysis

1. Analyzing how loan repayment is affected by Gender: Defaulters vs. Non-Defaulters



- Females have a higher default rate but also more non-defaulters.
- The "XNA" category for missing gender data is significant, indicating incomplete information.

2. Visualizing the Distribution of Numerical Features

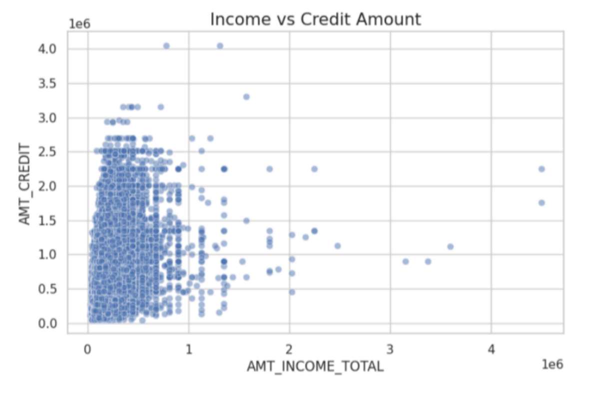


- Income level and loan amount are key factors in predicting loan default, but age shows no clear distinction.
- Further analysis is needed to understand age's interaction with income and credit history.

For a deeper analysis, refer to the notebook.

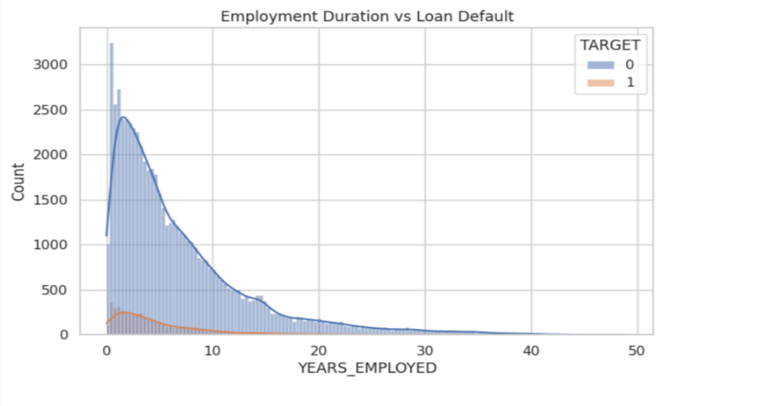
Steps for EDA – Bivariate Analysis

1. Visualizing Relationships Between variables in application_data file



There is a positive trend, other factors likely affect the relationship between income and credit amount.

2. Relationship Between DAYS_EMPLOYED and Loan Default (TARGET)

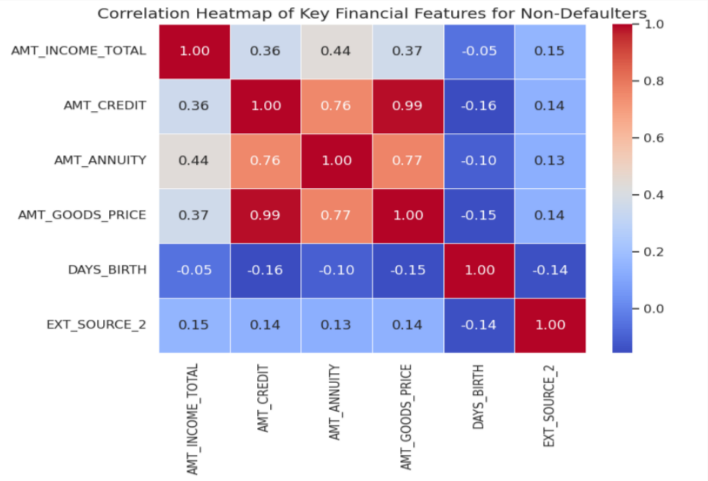


longer employment durations seem to reduce default risk, but further analysis is needed for deeper insights.

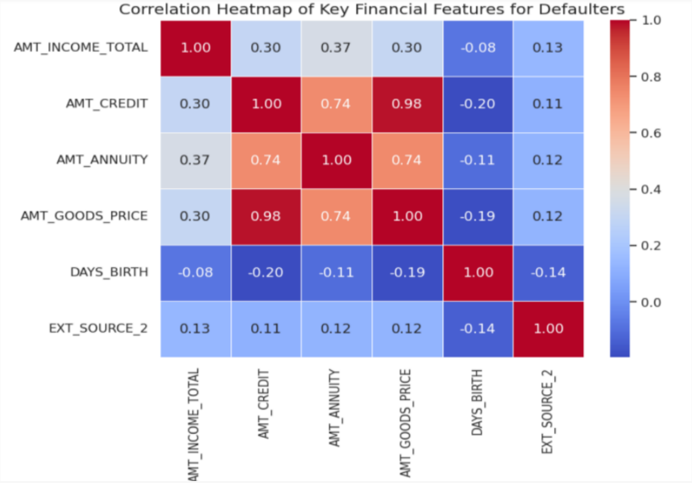
For a deeper analysis, refer to the notebook.

Steps for EDA – Bivariate Analysis

3. Correlation Heatmap of Key Financial Features for Non-Defaulters



4. Correlation Heatmap of Key Financial Features for Defaulters



The heatmaps shows strong links between AMT_CREDIT, AMT_ANNUITY, and AMT_GOODS_PRICE. AMT_INCOME_TOTAL is moderately correlated with AMT_CREDIT. DAYS_BIRTH and EXT_SOURCE_2 show weak relationships with financial features in both (Non-defaulters and defaulters)

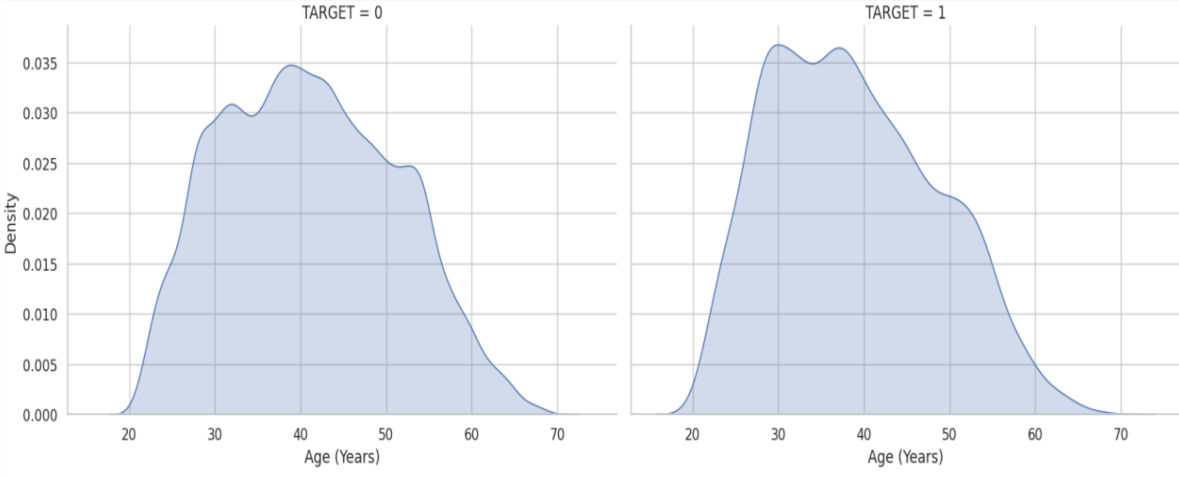
Steps for EDA – Bivariate Analysis

5. Violin Plot for AMT_CREDIT by NAME_INCOME_TYPE in application_data

Income type is a key factor in credit amount distribution, influencing credit profiles and risk assessment.

Steps for EDA - Time-Based Analysis

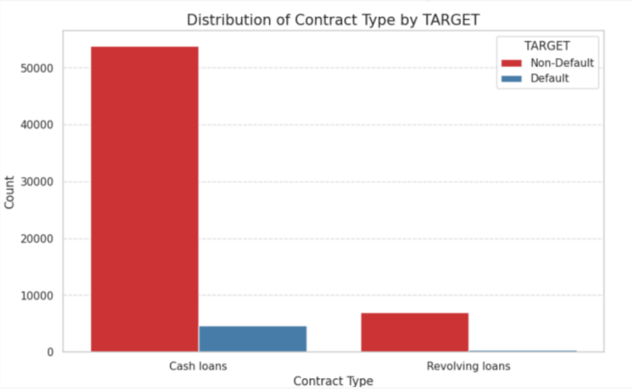
1. Compare the distribution of AGE_YEARS and YEARS_EMPLOYED for defaulters vs. non-defaulters.



This visualization provides a comparative view of how age varies between the two target groups, helping to identify any potential age-related patterns in loan default behavior.

Steps for EDA – Target Variable Analysis

1. Distribution of Contract Type by TARGET in application_data



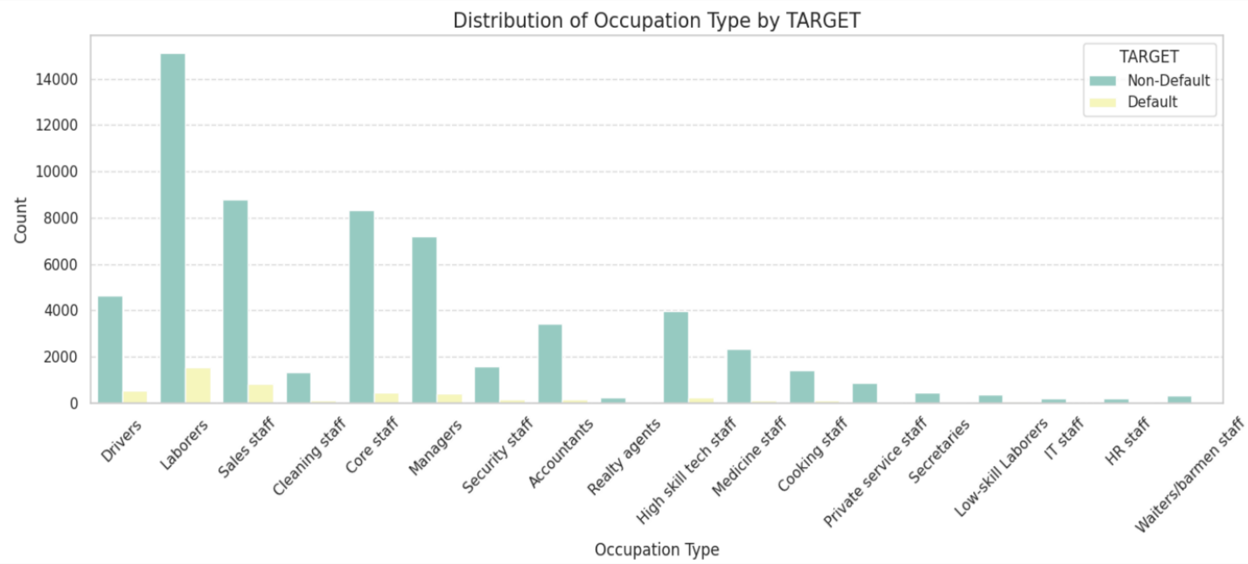
The chart shows cash loans are the most common contract type at TARGET, with most individuals holding cash and revolving loans not classified as "Default."

2. Distribution of Gender by TARGET in application_data

The chart shows that TARGET has more female employees than male employees.

Steps for EDA – Target Variable Analysis

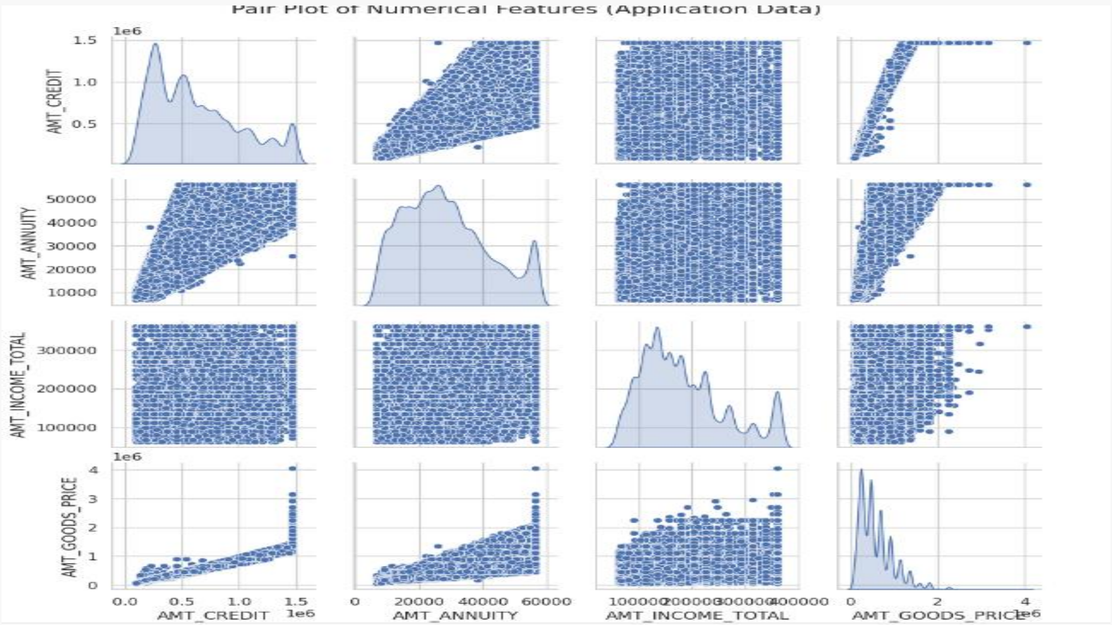
3. Distribution of Occupation Type by TARGET in application data



- The chart shows the distribution of occupation types by TARGET, which is likely a company or organization.
- It shows that drivers, laborers, and sales staff are the most common occupations.

Steps for EDA – Multivariate Analysis

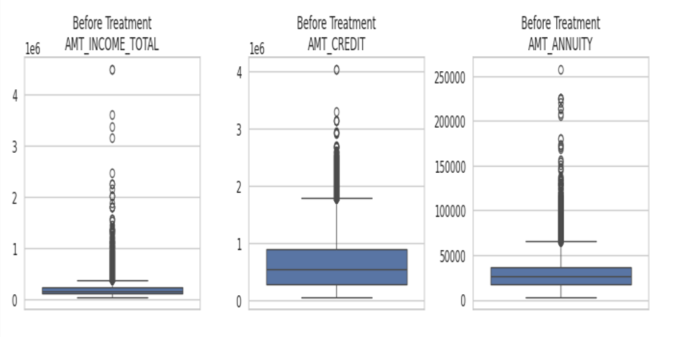
1. **Distribution of Occupation Type by TARGET in application data**

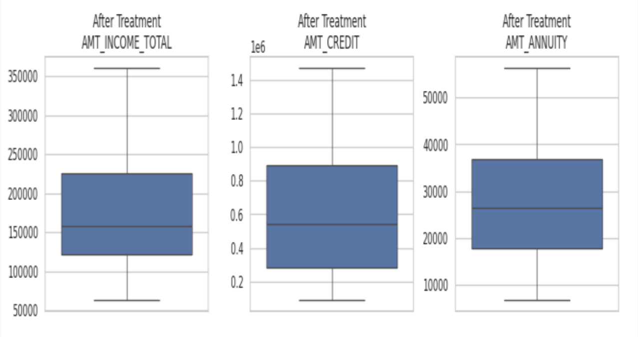


- The chart shows the distribution of occupation types by TARGET, which is likely a company or organization.
- The pair plot reveals significant correlations, distributions, and outliers among the features.
- Distributions: Right-skewed distributions for income and credit amounts.

Steps for EDA – Outlier Detection

1. To Detect Outliers in the Application_data and Decide on Their Treatment

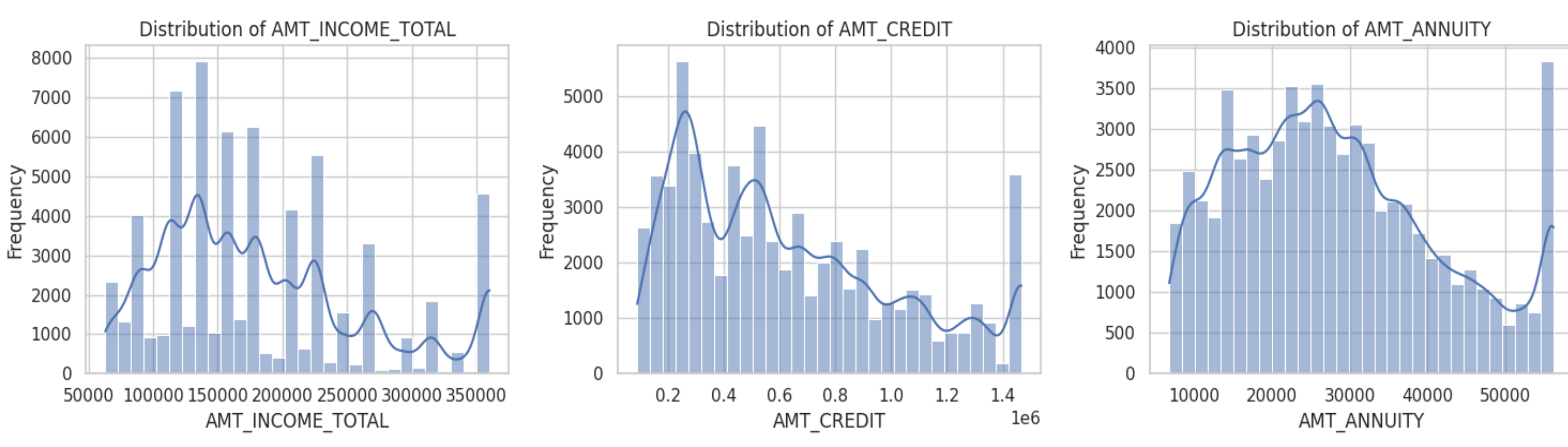




- Using Caps outliers in a DataFrame using a specified percentile-based approach. It then visualizes the distribution of the data after outlier treatment using box plots.

Steps for EDA – Data Distribution

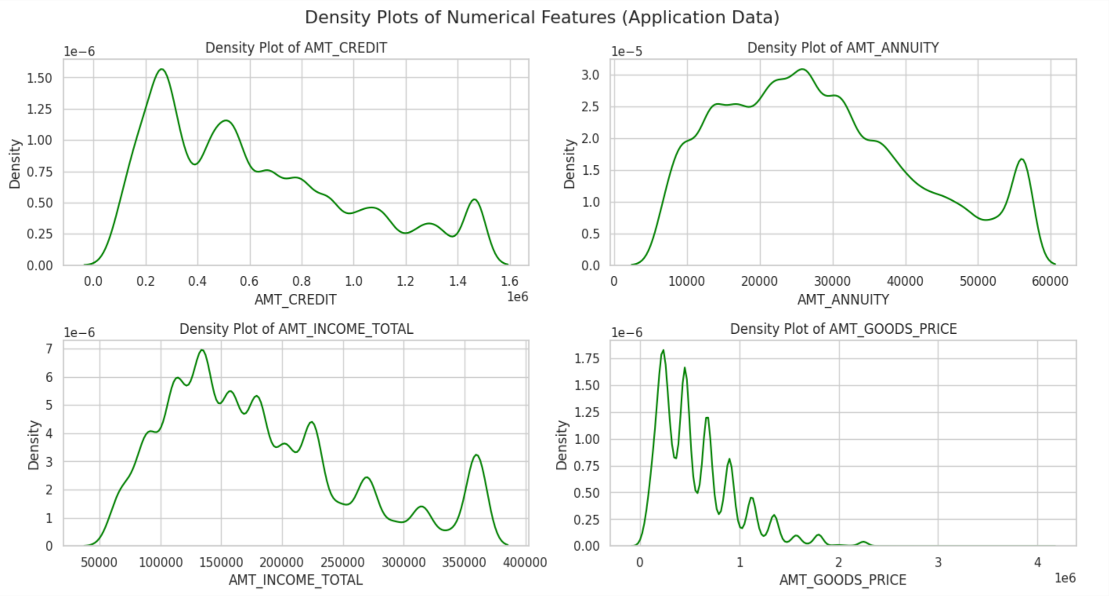
1. Distribution Analysis of Key Numerical Features



- The distributions of all three variables exhibit right-skewness, indicating that a majority of individuals have lower values for these financial metrics, while a smaller number have higher values.
- This suggests that there is a significant disparity in income, credit, and annuity payments among the individuals in the dataset.

Steps for EDA – Data Distribution

2. Density Plots of Numerical Features in Application Data

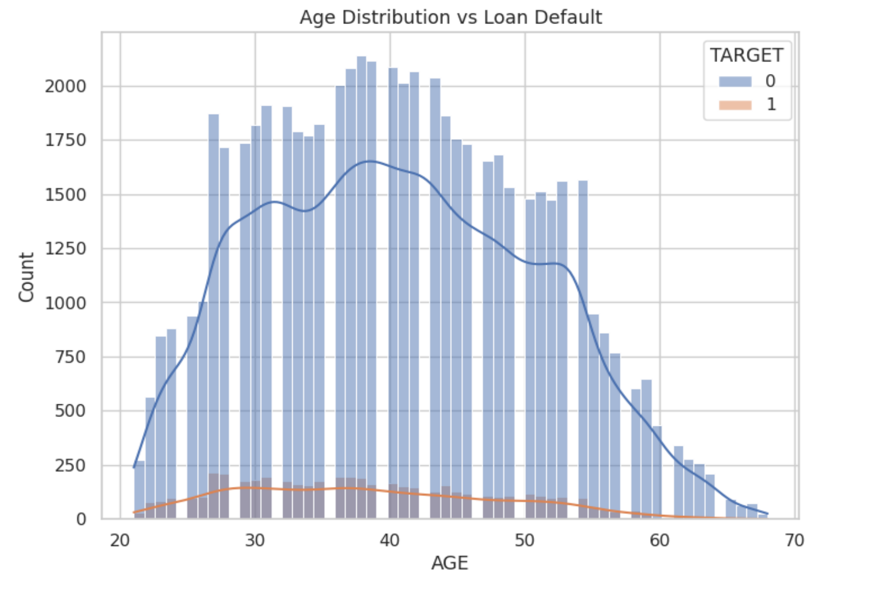


- The density plots provide valuable insights into the distribution of the numerical features in the application data. They help identify the shape, skewness, and potential outliers in the data.

Steps for EDA – Detecting Relationships

1. Detecting Relationships between Days-Based Variables and Risk

- Relationship between age and loan default, with younger individuals being more likely to default on loans compared to older individuals.
- However, it's important to note that other factors may also influence loan default, and further analysis would be needed to draw definitive conclusions.



Steps for EDA – Imbalance Check

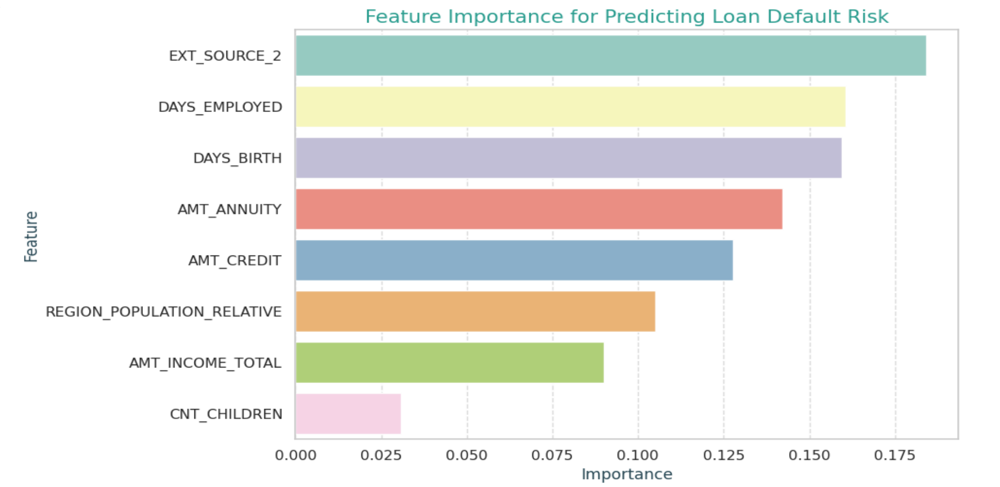
1. Check the Target Value if it is imbalanced

```
Original Distribution of TARGET:
TARGET
0    42629
1     3421
Name: count, dtype: int64

Distribution of TARGET After SMOTE:
TARGET
0    42629
1    42629
Name: count, dtype: int64
```

- Using the SMOTE method to address the imbalance in the target value, the distribution is adjusted to 0: 47,626 and 1: 42,629. Before applying SMOTE, the distribution was 0: 42,629 and 1: 3,421.

Steps for EDA – Risk Analytics

1. Feature importance using a RandomForeestClassifier model




```
Feature Importances:
                      Feature  Importance
4                EXT_SOURCE_2    0.184279
5               DAYS_EMPLOYED    0.160583
3                  DAYS_BIRTH    0.159359
2                 AMT_ANNUITY    0.142076
1                  AMT_CREDIT    0.127829
7  REGION_POPULATION_RELATIVE    0.105113
0            AMT_INCOME_TOTAL    0.089906
6                CNT_CHILDREN    0.030855
```

- EXT_SOURCE_2 and DAYS_EMPLOYED are the most influential features in predicting loan defaults.
- AMT_ANNUITY and AMT_CREDIT also significantly impact default risk.
- CNT_CHILDREN has the least impact among the top features.

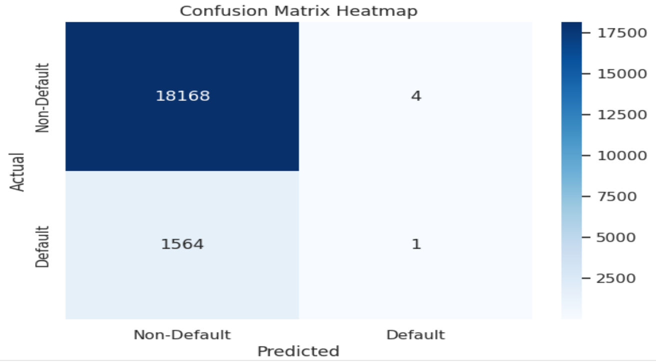
These insights help prioritize which features to focus on for risk management and model improvement.

Steps for EDA – Decision Making

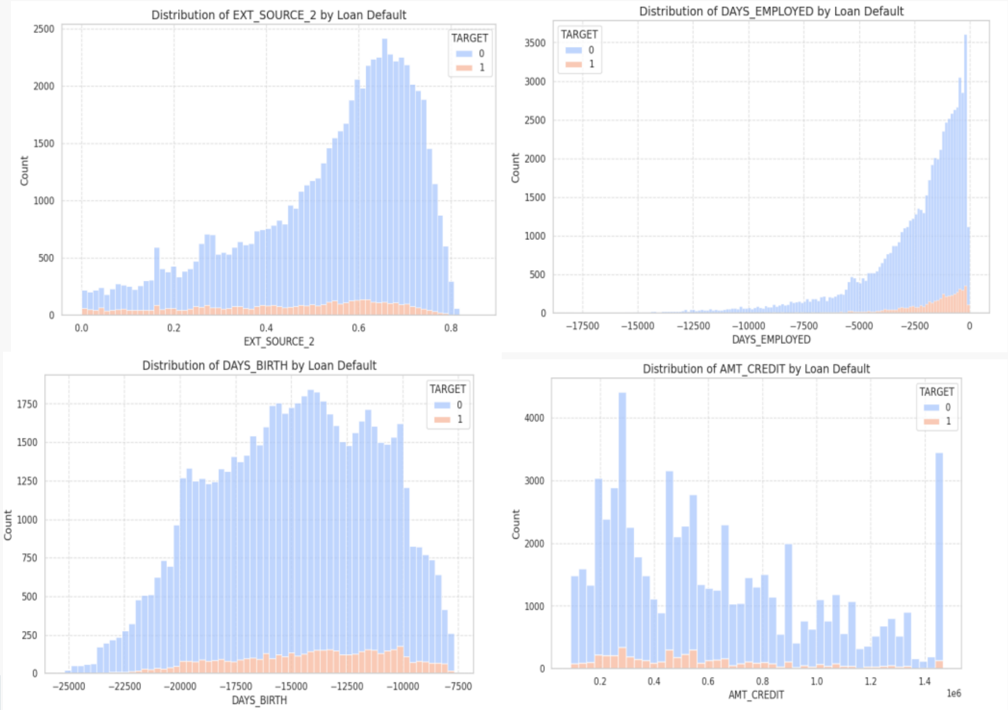
- The results suggest that the Random Forest model is performing well in predicting the "Non-Default" class, with high precision and recall.

- However, it struggles to predict the "Default" class, as indicated by the low precision and recall for this class.

```
RandomForest Classification Report:
              precision    recall  f1-score   support

           0       0.92      1.00      0.96     18172
           1       0.20      0.00      0.00      1565

    accuracy                           0.92     19737
   macro avg       0.56      0.50      0.48     19737
weighted avg       0.86      0.92      0.88     19737

RandomForest AUC Score: 0.6439

RandomForest Confusion Matrix:
[[18168     4]
 [ 1564     1]]

RandomForest Loan Decisions (First 10):
['Approve', 'Approve', 'Approve', 'Approve', 'Approve', 'Approve', 'Approve', 'Approve', 'Approve', 'Approve']
```



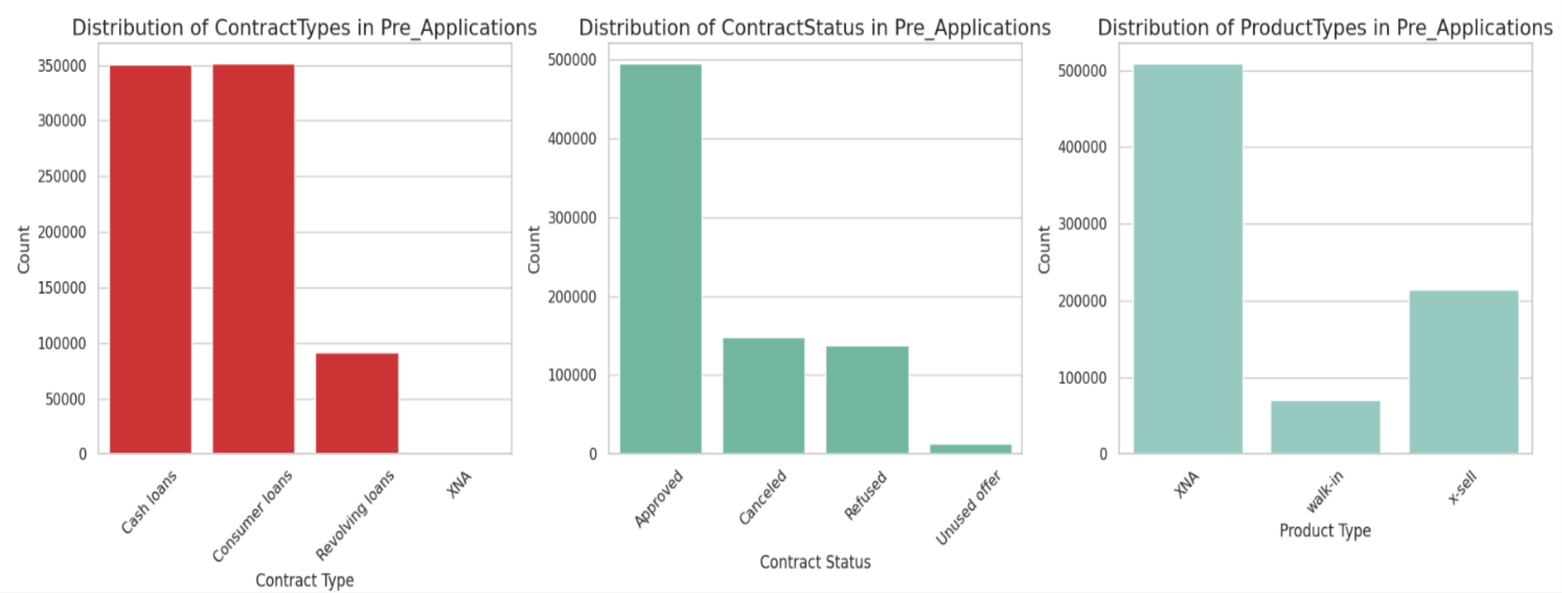
Steps for EDA – Financial Risk Management




```
Risk Profile Summary:
Risk Category
Very Low     17965
Low           1080
Moderate        48
High             0
Very High        0
Name: count, dtype: int64
```

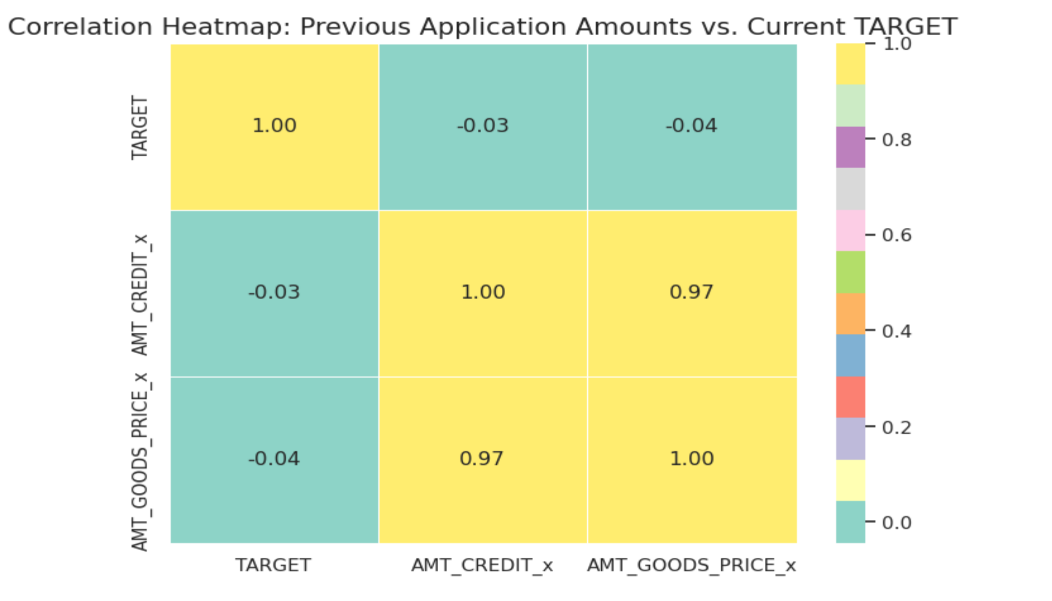
Steps for EDA – Merge Two dataset

```
# Merge the datasets on SK_ID_CURR
merged_data = pd.merge(application_data, previous_application, on='SK_ID_CURR', how='left')
merged_data.head()
```



Steps for EDA – Risk Analytics

Heatmap of Previous Application Amounts vs. Current TARGET



- The heatmap reveals a strong positive relationship between TARGET and AMT_CREDIT_X, while the relationship between TARGET and AMT_GOODS_PRICE_X is weak and negative.
- Additionally, AMT_CREDIT_X and AMT_GOODS_PRICE_X are strongly positively correlated.

Present Findings

- **Correlation Between Features:** Strong links between AMT_CREDIT, AMT_ANNUITY, and AMT_GOODS_PRICE.
- **Age vs Loan Default:** Younger applicants tend to default more, with additional factors influencing risk.
- **Numerical Distributions:** Income, credit, and annuity levels show right-skewed distributions.
- **Target Imbalance:** SMOTE balanced target values from (Non-Default: 42,629, Default: 3,421) to equal numbers.
- **Risk Analytics:** EXT_SOURCE_2 and DAYS_EMPLOYED are key predictors, though the model is more accurate with non-defaults (AUC: 0.6439).
- **Financial Risk:** Most applicants are categorized as very low to low risk.

For a deeper analysis, refer to the notebook.

Conclusion

The EDA revealed key factors influencing loan default, including strong correlations between financial variables and a higher default likelihood among younger individuals.
Imbalance in the dataset was corrected using SMOTE, and risk analytics highlighted EXT_SOURCE_2 and DAYS_EMPLOYED as key predictors. While the model performed well for non-defaults, it requires improvement in detecting defaults. These insights will aid in better loan approval decisions and risk management.

For a deeper analysis, refer to the notebook.

Thanks!

Any questions?

samramya@gmail.com