Credit EDA Assignment

- By Deepanjali Bhatt

Table of content

Problem Statement

Overview

- A consumer finance company which specialises in lending various types of loans to urban customers.
- When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:
  - If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.
  - If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

Objective

*To improve **Risk Assessment***

- Refine loan approval criteria to reduce risk.

*To minimize the **Credit Loss***

- Optimize portfolio quality to minimize financial losses.

*Employ targeted **Strategies** such as*:

- **Loan denial**: Refusing applications from high-risk borrowers.
- **Loan reduction**: Decreasing loan amounts for applicants with limited capacity.
- **Interest rate adjustment**: Charging higher rates to compensate for increased risk.
- **Term modification**: Altering loan durations to balance risk and borrower needs.

Summary

Defaulters' demography : All the below variables were established in analysis of Application dataframe as leading to default.

- Checked these against the Approved loans which have defaults, and it proves to be correct

    - Medium income

    - 25-35 years olds , followed by 35-45 years age group

    - Male

    - Unemployed

    - Labourers, Salesman, Drivers

    - Business type 3

    - Own House - No

- Other IMPORTANT Factors to be considered

    - Days last phone number changed - Lower figure points at concern

    - No of Bureau Hits in last week. Month etc – zero hits is good

    - Amount income not correspondingly equivalent to Good Bought – Income low and good value high is a concern

    - Previous applications with Refused, Cancelled, Unused loans also have default which is a matter of concern.   This indicates that the financial company had Refused/Cancelled previous application but has approved the current and is facing default on these.

Credible Applications refused

- Unused applications have lower loan amount. Is this the reason for no usage?

- Female applicants should be given extra weightage as defaults are lesser.

-

60% of defaulters are Working applicants. This does not mean working applicants must be refused. Proper scrutiny of other parameters needed

- Previous applications with Refused, Cancelled,Unused loans also have cases where payments are coming on time in current application. This indicates that possibly wrong decisions were done in those cases.

Approach and Methodology

Data Acquisition

- Data Sourcing
- Data Description

Data Preparation

- Data Cleaning
    - Dropping rows/columns that are completely null
    - Missing values treatment
    - Handling outliers
    - Filtering data

Exploratory Data Analysis(EDA)

- Univariate Analysis
    - Categorical
    - Numerical
- Bivariate Analysis
- Multivariate Analysis
- Inferences and Recommendations

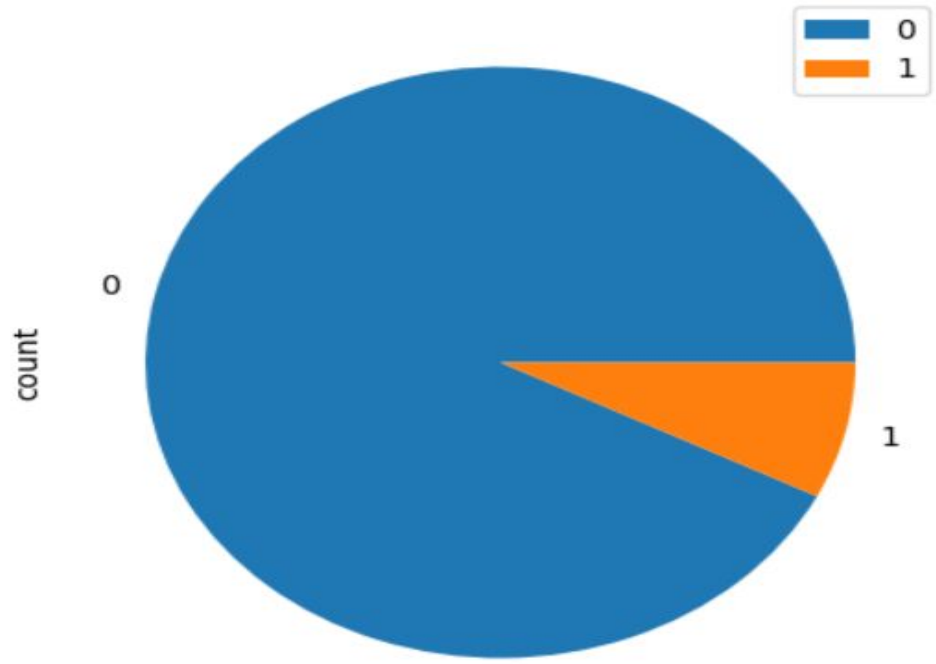
Note: Though we have analyzed all the relevant attributes in the loan data, this presentation includes analysis of only attributes that are most relevant to the business.

Plots and Insights

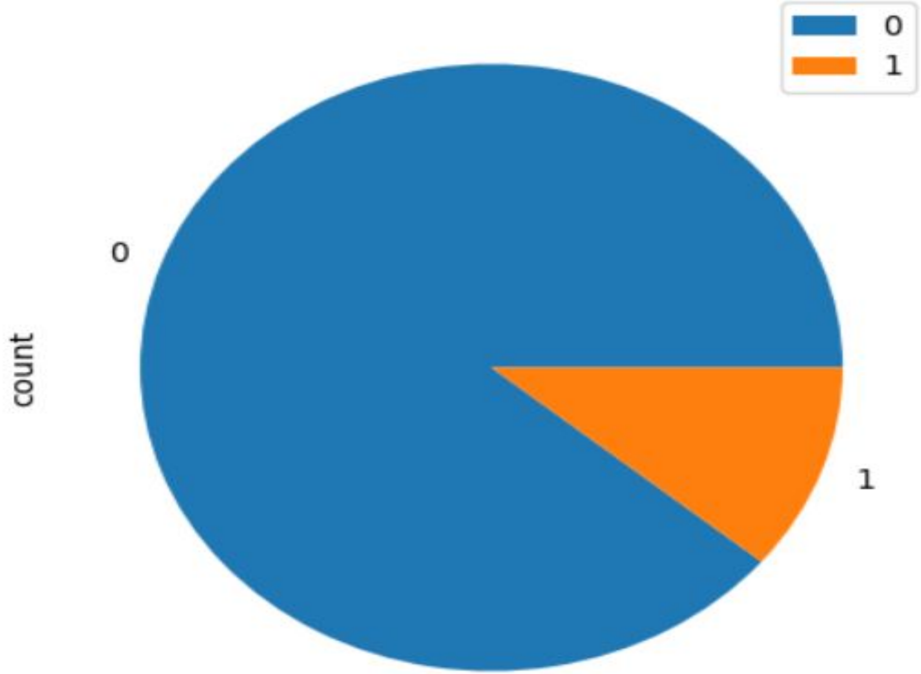
Summary of analysis

- Upon merging the datasets and conducting a comprehensive exploratory data analysis, several significant insights were identified.
- **Analysis 1:**
    - Analyzing Loan Status and Default Rates
    - To gain insights into the relationship between loan status and default rates, we analyzed the distribution of NAME_CONTRACT_STATUS for both target 0 (non-default) and target 1 (default) categories. By visualizing this data, we can identify patterns and trends that may influence the likelihood of default.

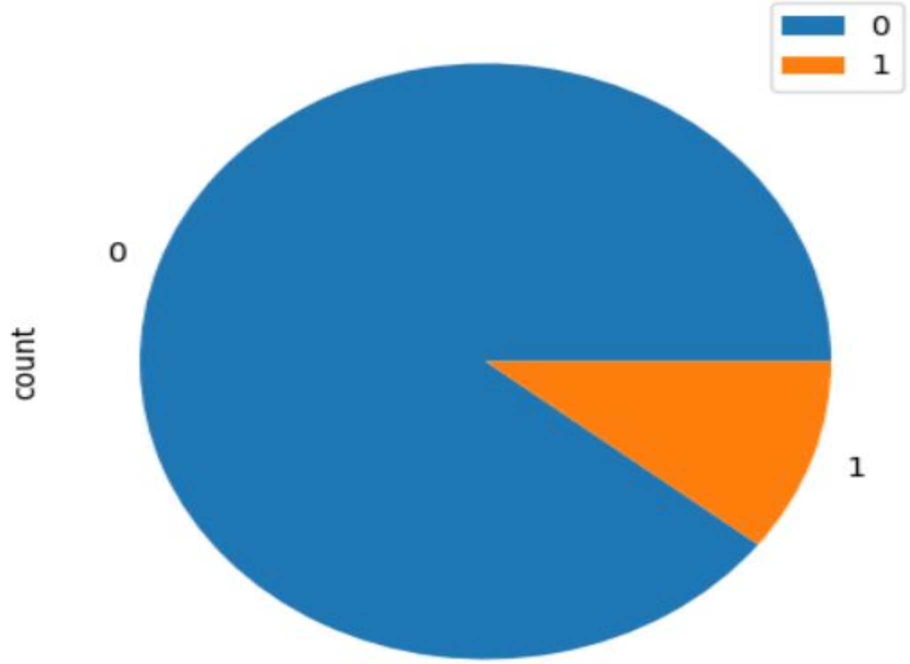
```
Target 0 and 1 for :  Approved
TARGET
0     0.925206
1     0.074794
Name: proportion, dtype: float64
```



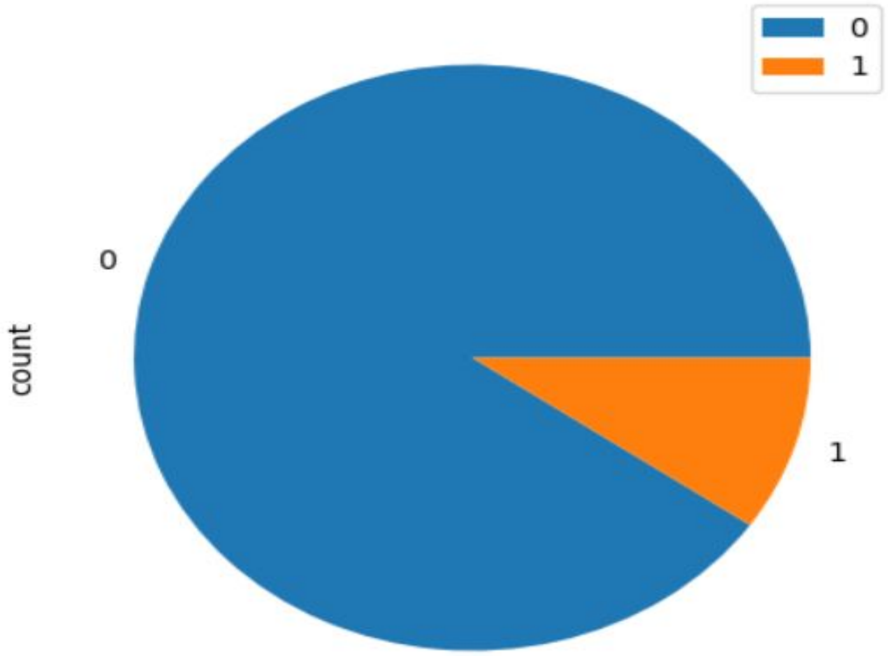
```
Target 0 and 1 for :  Refused
TARGET
0    0.889593
1    0.110407
Name: proportion, dtype: float64
```



```
Target 0 and 1 for :  Unused offer
TARGET
0    0.894602
1    0.105398
Name: proportion, dtype: float64
```

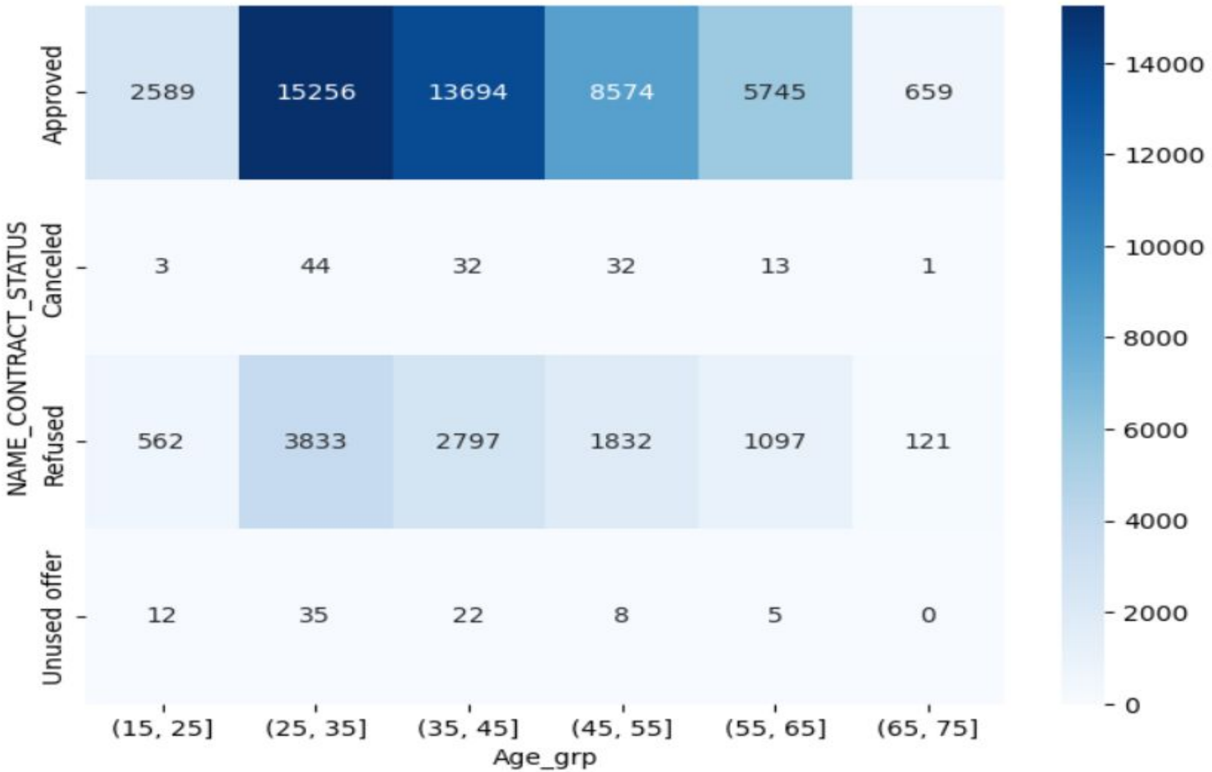


```
Target 0 and 1 for :  Canceled
TARGET
0    0.903251
1    0.096749
Name: proportion, dtype: float64
```



- Key insights on Analysis 1:
    - Approximately 7.5% of all approved loans result in default, indicating a certain level of credit risk.
    - Previous loan rejections are associated with higher default risk.
- **Analysis 2:**
    - Analyzing the impact of Contract Status and Age group on Default rates.
    - To understand the interplay between contract status, age group, and default rates, we have created a heatmap. This visualization allows us to identify patterns and trends that may influence the likelihood of default across different age groups and contract statuses.

NAME_CONTRACT_STATUS' vs 'AGE_GROUP



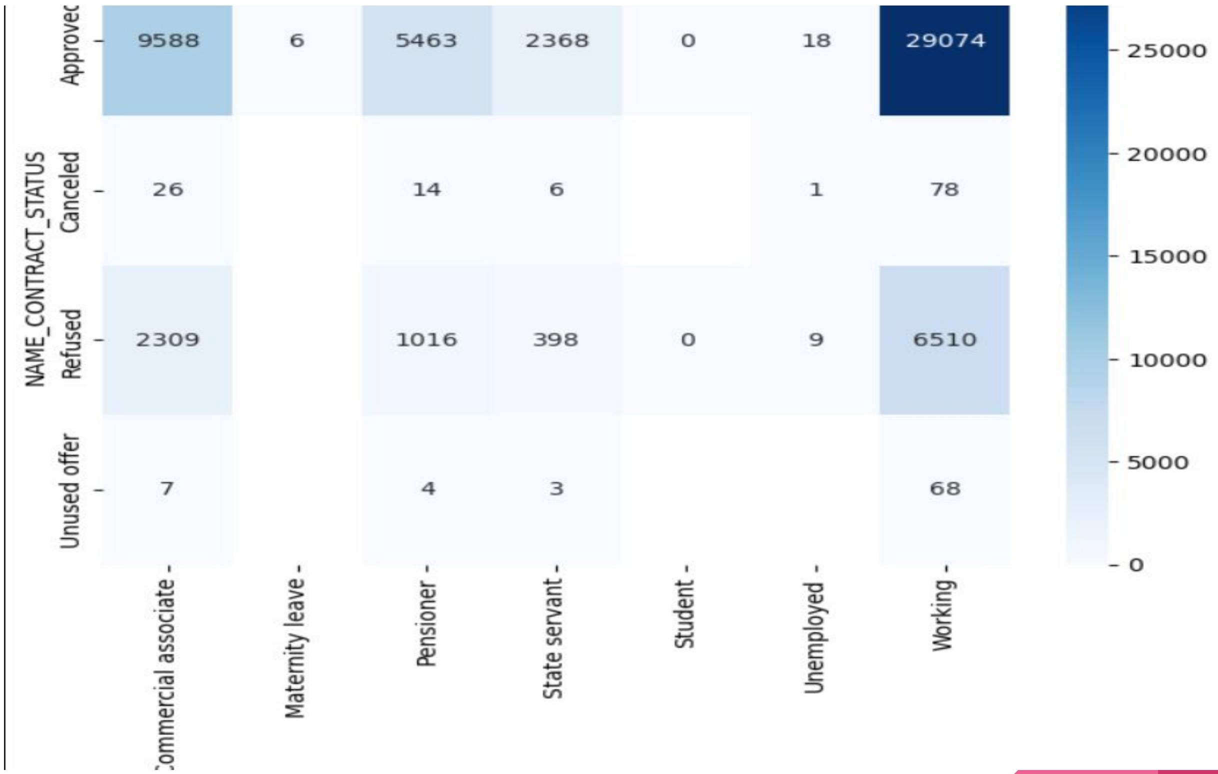
- Key insights on ‘NAME_CONTRACT_STATUS’ vs ‘AGE_GROUP’
    - Higher values in the matrix indicate a higher risk of default
    - Individuals aged 25-45, particularly those in the 35-45 age group, exhibit higher default rates among approved loans. This suggests that factors such as career instability, family commitments, or lifestyle choices may contribute to increased financial strain and a higher likelihood of default in this demographic.
    - Individuals with a history of rejected or cancelled loan applications are more likely to default on subsequent loans. This suggests that past credit behavior is a strong indicator of future risk.

- **Analysis 3:**

    - Analyzing the impact of contract status and Income type on default rates

    - To understand the interplay between contract status, income type, and default rates, we have created a heatmap. This visualization allows us to identify patterns and trends that may influence the likelihood of default.

NAME_CONTRACT_STATUS' vs NAME_INCOME_TYPE



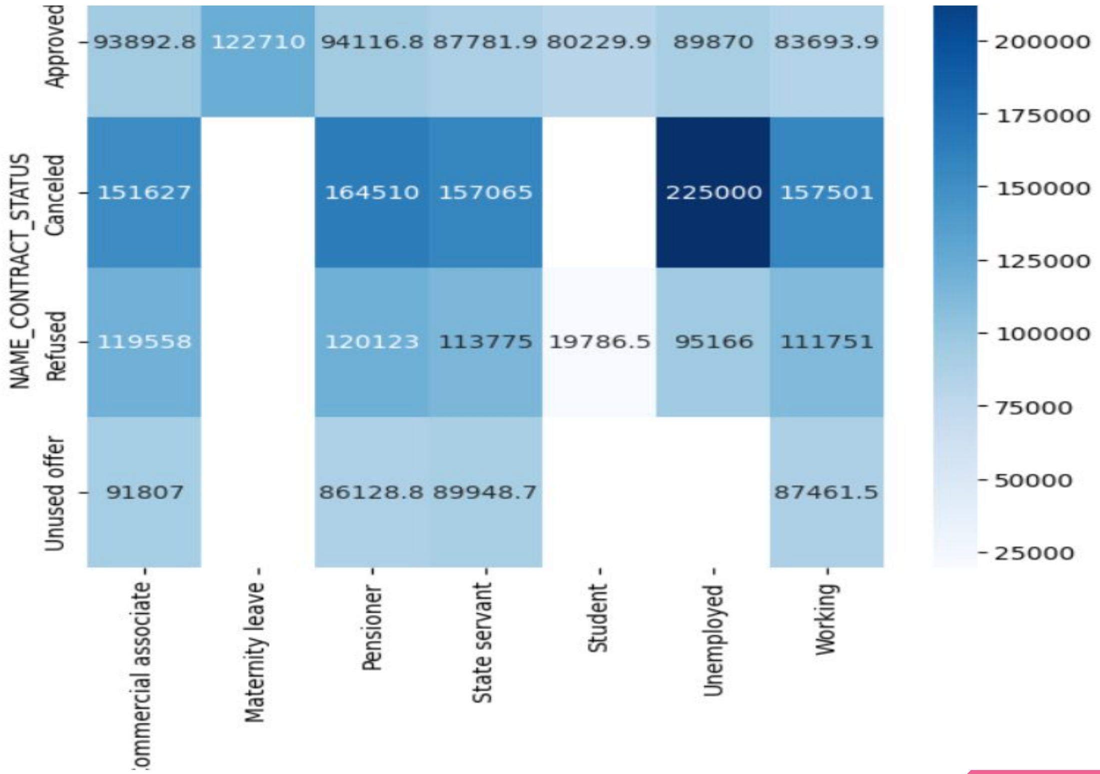
- Key Insights on NAME_CONTRACT_STATUS’ vs NAME_INCOME_TYPE
    - Higher values in the matrix indicate a higher risk of default
    - Employed applicants have the highest default rate.
    - The analysis reveals a correlation between previous loan rejections and subsequent defaults. This indicates that a thorough assessment of an applicant's credit history, including past loan applications and outcomes, is crucial for accurate risk assessment.
    - A significant number of working-class individuals who had previously been denied loans have subsequently defaulted on their current loans. This suggests that the bank may need to re-evaluate its credit assessment criteria for this demographic.

- **Analysis 4 :**

    - Analyzing impact of contract status and income type on loan amounts

    - To understand how contract status and income type influence loan amounts, we have created a heatmap. This visualization allows us to identify patterns and trends in the average loan amount across different combinations of contract status and income type.

NAME CONTRACT STATUS Vs NAME_INCOME_TYPE



Recommendations

The initial analysis provides a solid foundation for understanding the key demographic and financial characteristics of loan applicants. However, to further enhance the predictive capabilities of our models and improve decision-making, we recommend the following:

Data Quality:

- **Address Missing Values:** Implement strategies to handle missing values in critical columns like AMT_ANNUITY, AMT_GOODS_PRICE, and others. Consider imputation techniques or domain-specific knowledge to fill in missing data.
- **Outlier Detection and Treatment:** Identify and address outliers in continuous variables like AMT_INCOME_TOTAL and AMT_CREDIT to ensure data accuracy and model robustness.
- **Feature Engineering:** Create informative features from existing data, such as Debt-to-Income Ratio, Loan-to-Value Ratio, and time-based features to capture temporal trends.

Feature Engineering and Selection:

- **Feature Engineering:** Explore techniques like feature discretization, normalization, and interaction terms to extract valuable insights from the data.
- **Feature Selection:** Employ feature selection methods like correlation analysis, feature importance, or dimensionality reduction techniques to identify the most relevant features.

THANK YOU