Presentation on Credit EDA Case Study

Varun Vaidhiya A

Purpose of this Case Study

- Credit Risk Analysis is Used by Banks and other Loan Providing Organizations to identify the risk involved in providing loan to a particular person
- In this case study we have analyzed the data of huge number of applicants and got some interesting insights by doing data Analysis

Steps

- **1. Importing Libraries**
- **2. Importing and checking application_data.csv and previous_application.csv**
- **3. Data Cleaning - "application_data.csv"**
- **4. Data Cleaning - "previous_application.csv"**
- **5. Data Analysis**

Processes used to achieve above steps

1.Data understanding and sourcing

2.check for Data quality issues and Binning

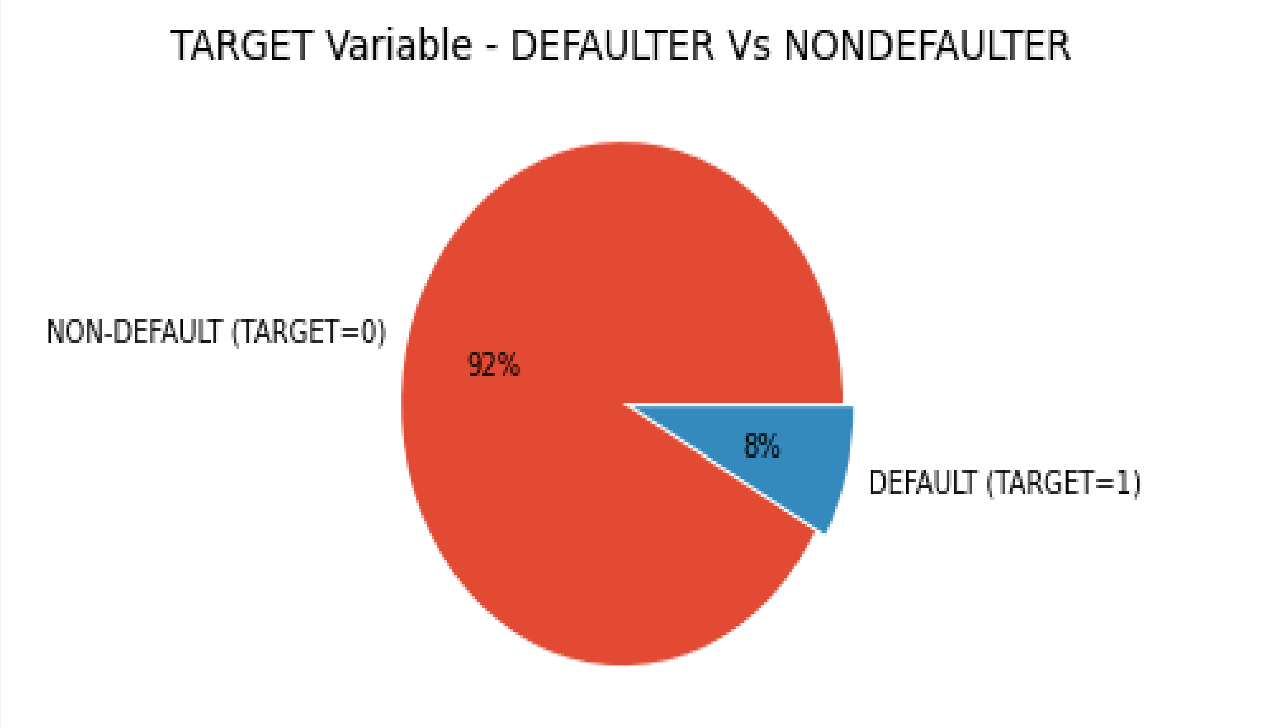
3.check for Data imbalance and univariate, segmented univariate & Bivariate analysis, correlation.

4.Merging of application data with previous applicationdata

5.Data analysis by univariate, segmented univariate ,Bivariate analysis and correlation

6.Recommendations and Risks

Checking for Data Imbalance

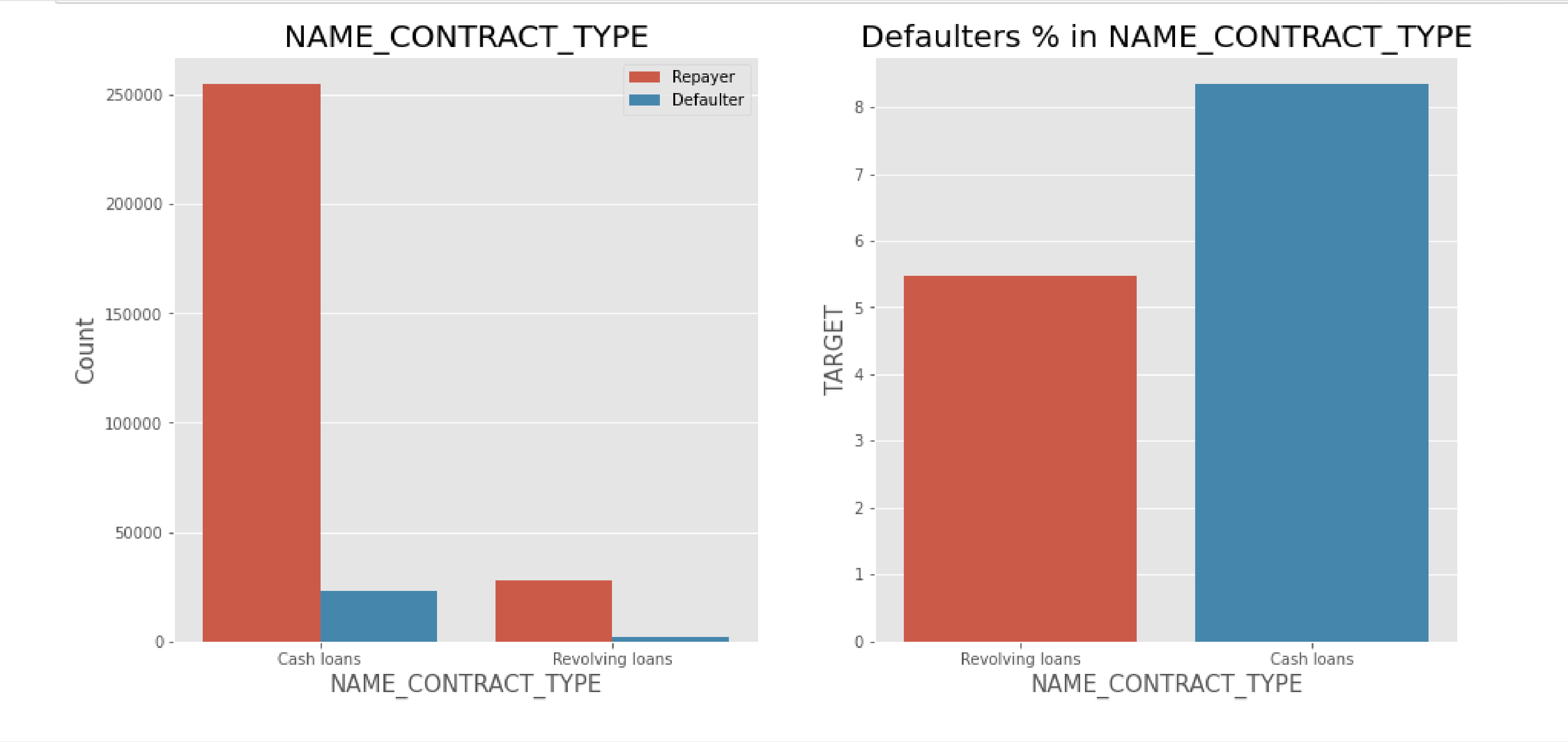


Inference

- Repayer Percentage is 91.93%
- Defaulter Percentage is 8.07%
- Imbalance Ratio with respect to Repayer and Defaulter is given: 11.39/1 (approx)

Categorical Variable Analysis

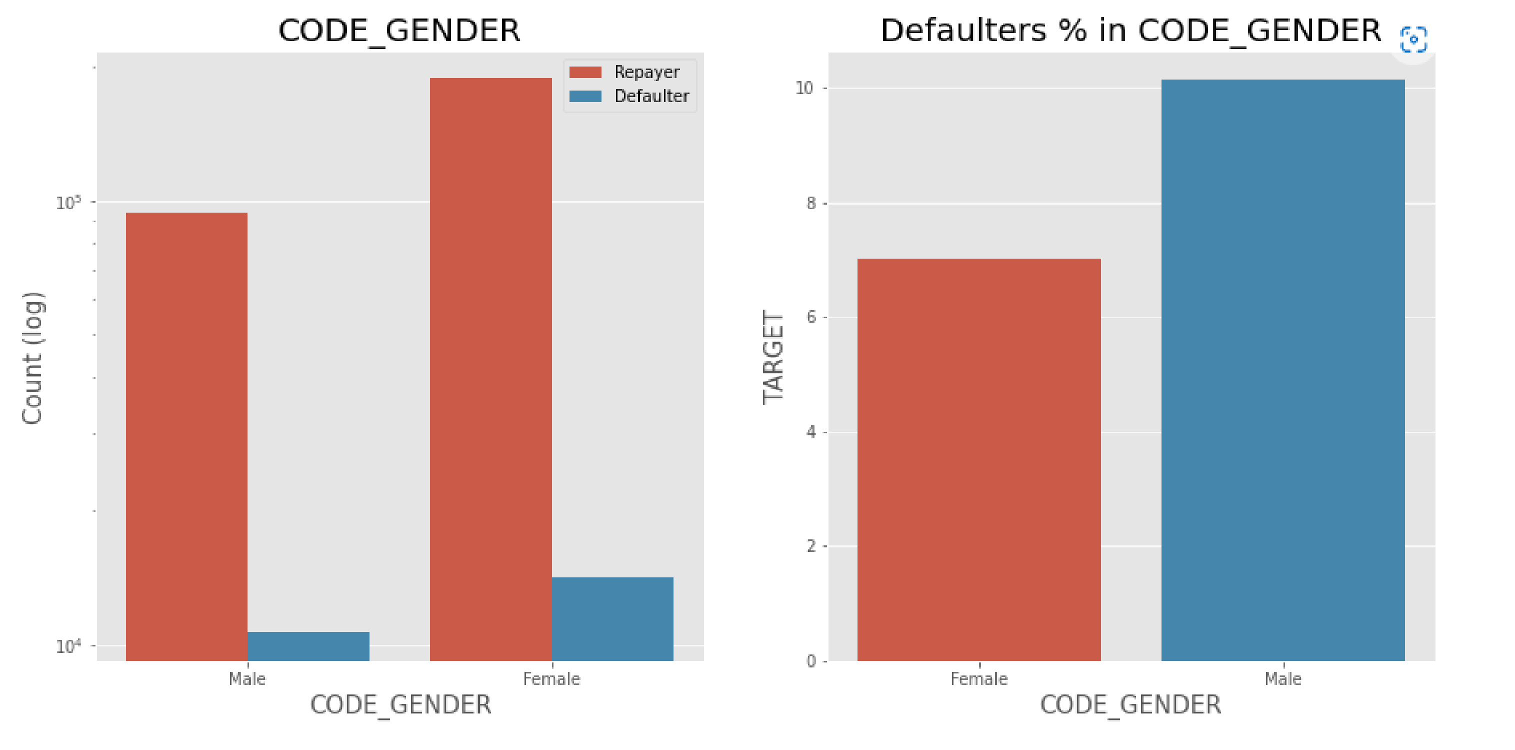
Checking the contract type based on loan repayment status



Inference-Contract type

- Revolving loans are just a small fraction (10%) from the total number of loans
- Around 8-9% Cash loan applicants and 5-6% Revolving loan applicant are in defaulters

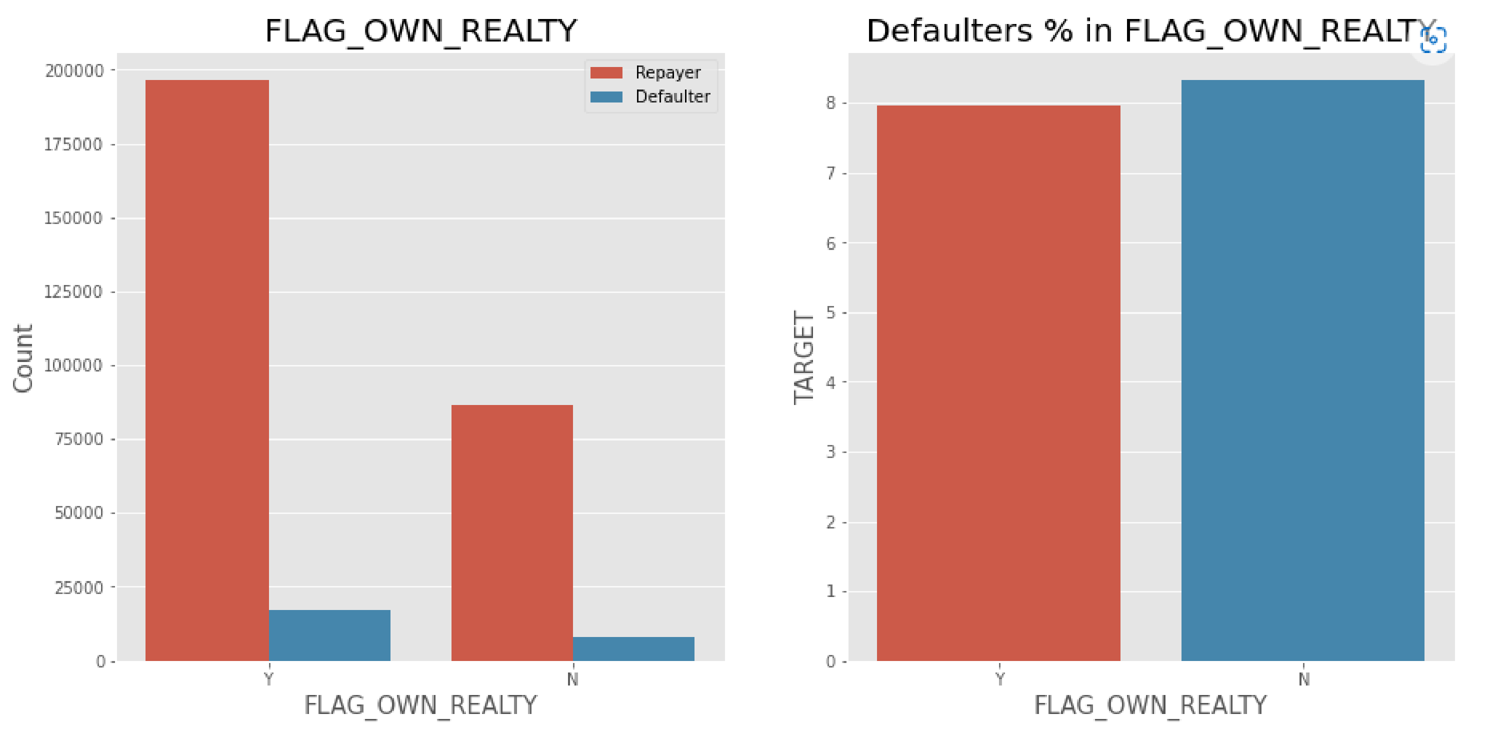
Checking the type of Gender on loan repayment status



Inferences: Gender Type

- The number of female clients is almost double the number of male clients.
- Based on the percentage of defaulted credits, males have a higher chance of not returning their loans about 10%, comparing with women about 7%

Checking if owning a real estate is related to loan repayment status

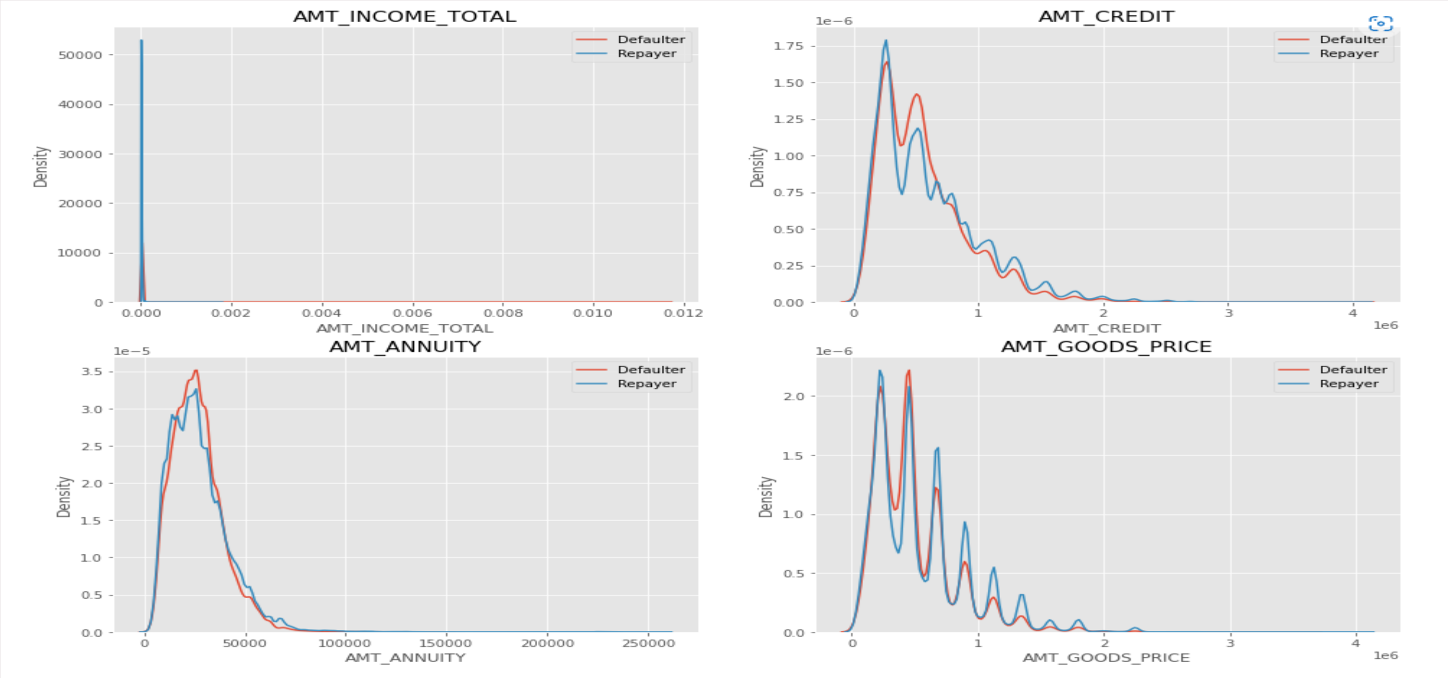


Inferences:

- The clients who own real estate are more than double of the ones that don't own.
- The defaulting rate of both categories are around the same (~8%). Thus we can infer that there is no correlation between owning a reality and defaulting the loan.

Numerical Univariate Analysis

- Plotting the numerical columns related to amount as distribution plot to see density

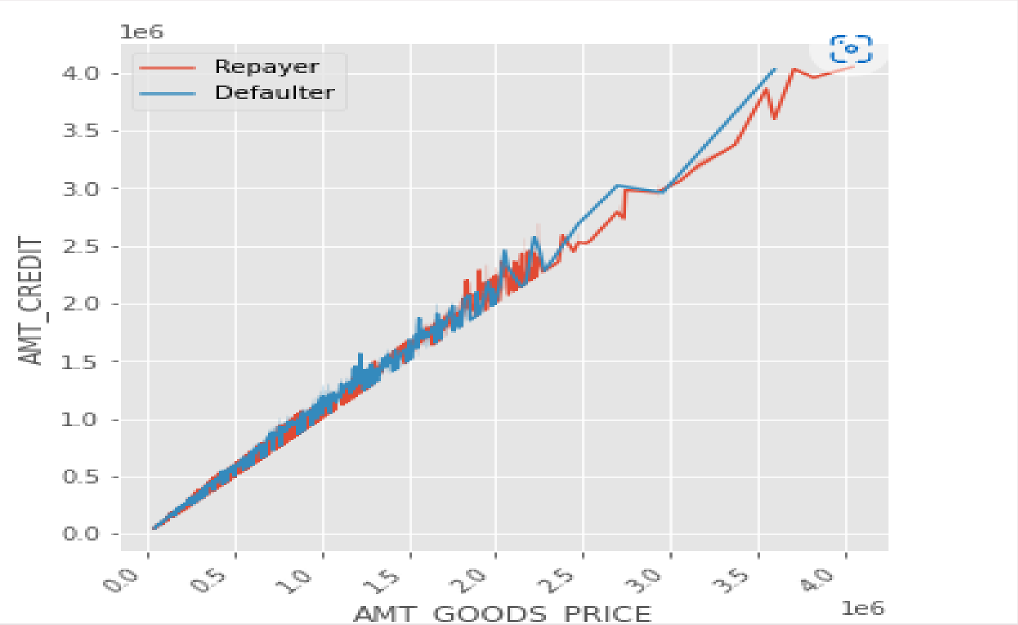


Inference

- Most no of loans are given for goods price below 10 lakhs
- Most people pay annuity below 50K for the credit loan
- Credit amount of the loan is mostly less then 10 lakhs
- The repayers and defaulters distribution overlap in all the plots and hence we cannot use any of these variables in isolation to make a decision

Numerical Bivariate Analysis

- Checking the relationship between Goods price and credit and comparing with loan repayment staus



Inference

- When the credit amount goes beyond 30 Lakhs, there is an increase in defaulters.

Conclusion

- From the above Data Analysis we can infer that the groups that is less likely to default is :
  - Clients who are working as a state servant.
  - Old people of any income group.
  - Client with high income category.
  - Old female client.
  - Client with higher education (female).
  - Any client whose previous loan was approved.
- And groups that are more likely to default is
  - Lower secondary educated clients are the most in
  - number to be defaulted when their previous loans were cancelled or refused
  - Male clients with civil marriage.
  - Previously refused loan status group