

EDA Bank Case Study

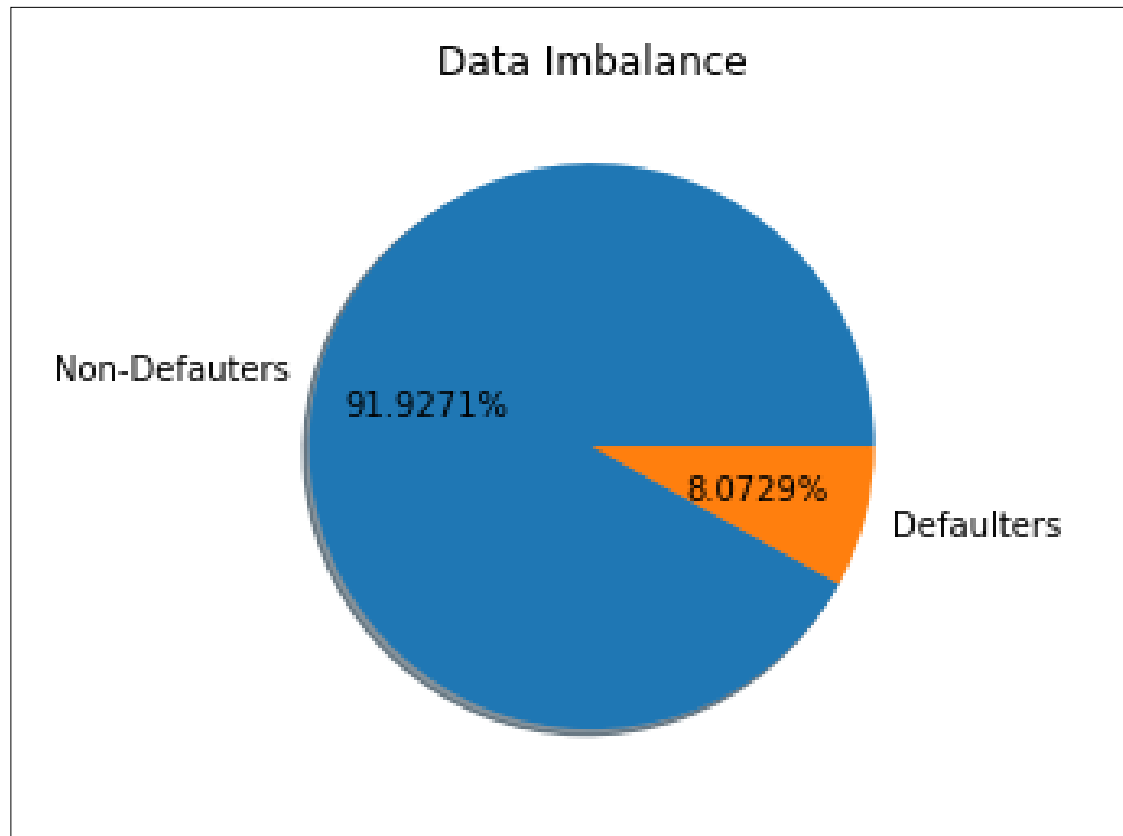
Submitted by : Shilpa Goel

Problem Statement and Approach

- Bank dataset is provided to identify defaulters in replaying loans, so that loan will not be provided to them in future.
- For this, we have to :
 - a. Clean the data
 - b. Bin columns
 - c. Perform univariate and bivariate analysis
 - d. Identify group of people with some common attributes who are defaulters.

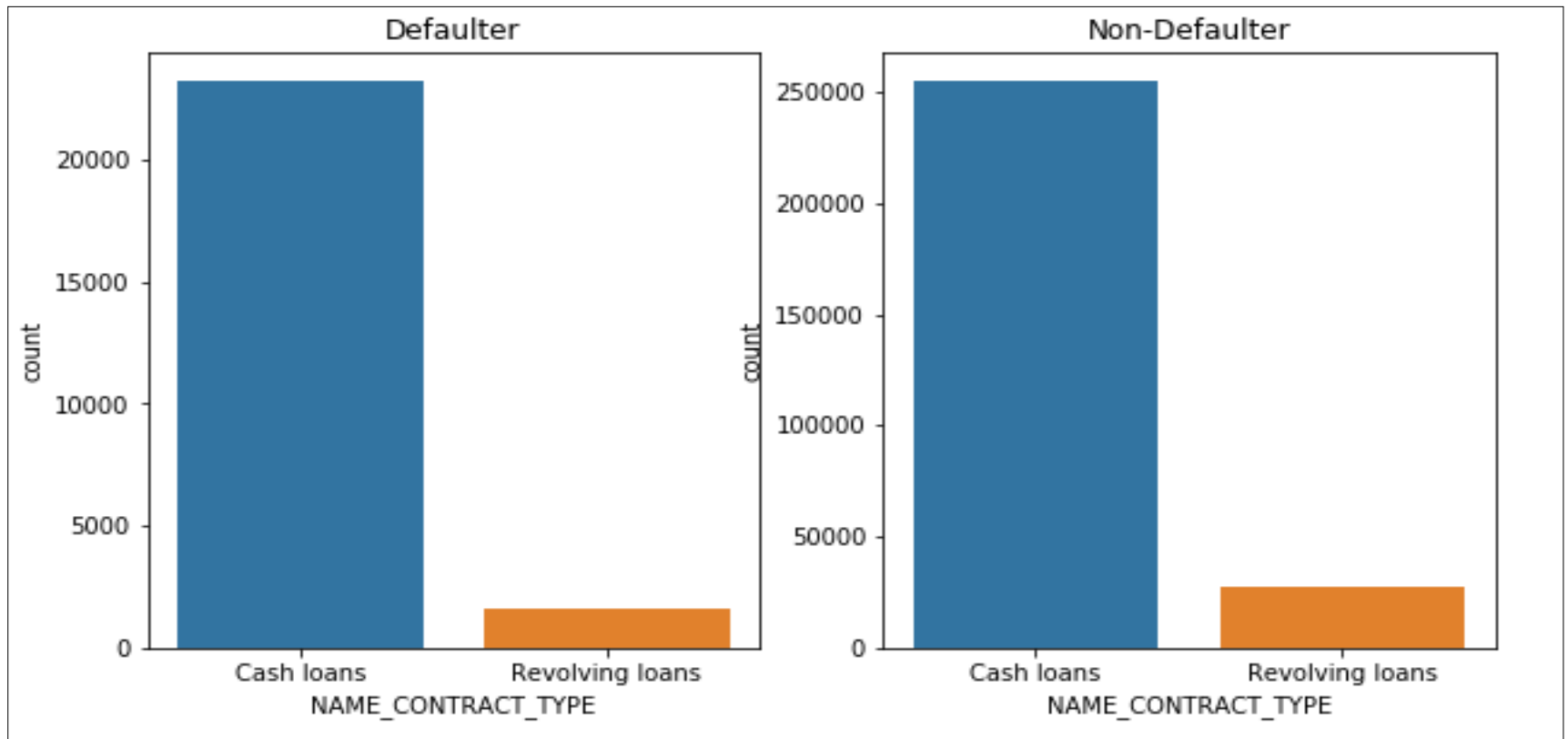
Data Imbalance

There is high data imbalance with 91.93% of Non-defaulters and 8.07% of defaulters.



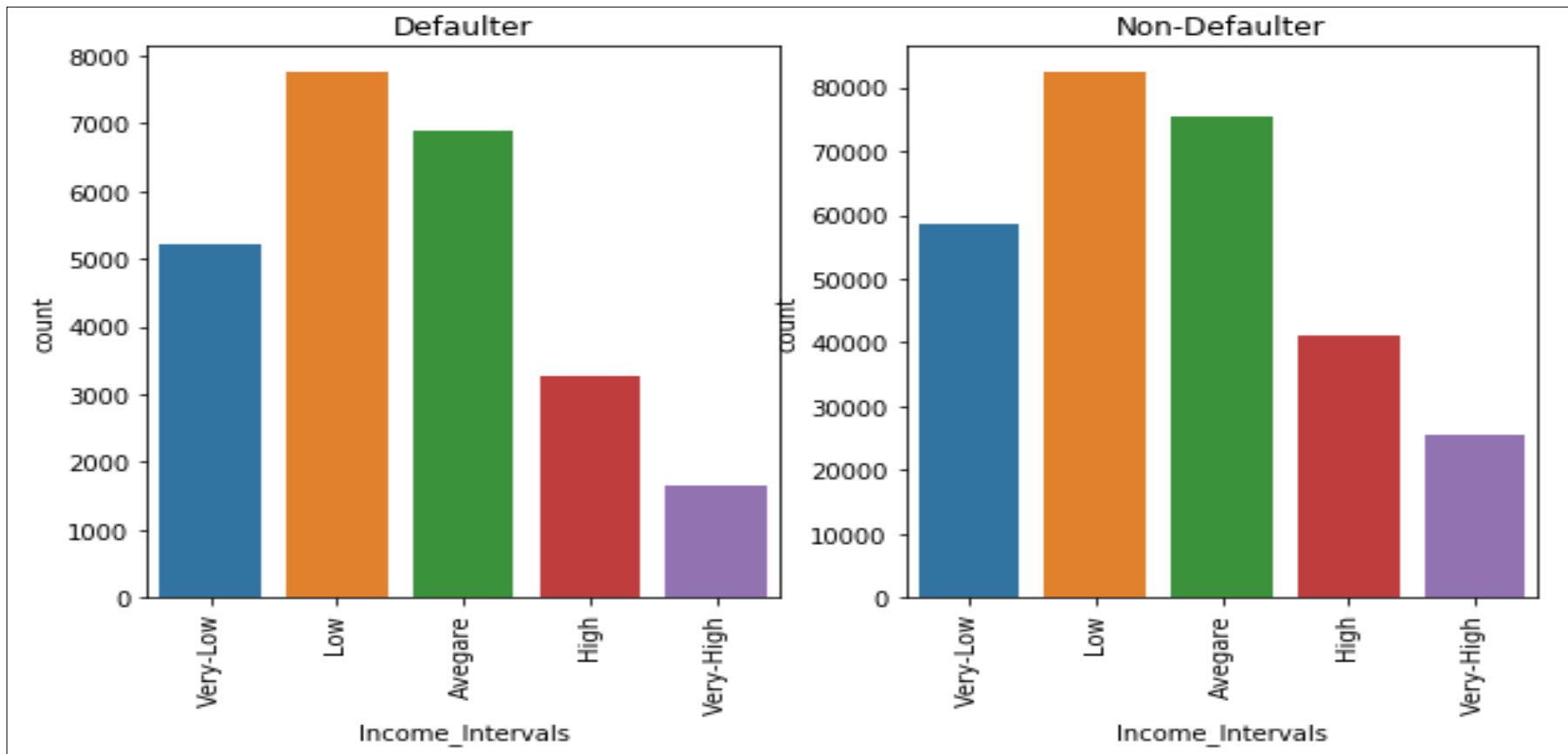
Unordered Categorical Variables

Defaulter count is comparatively high for Cash Loans type than Revolving Loans



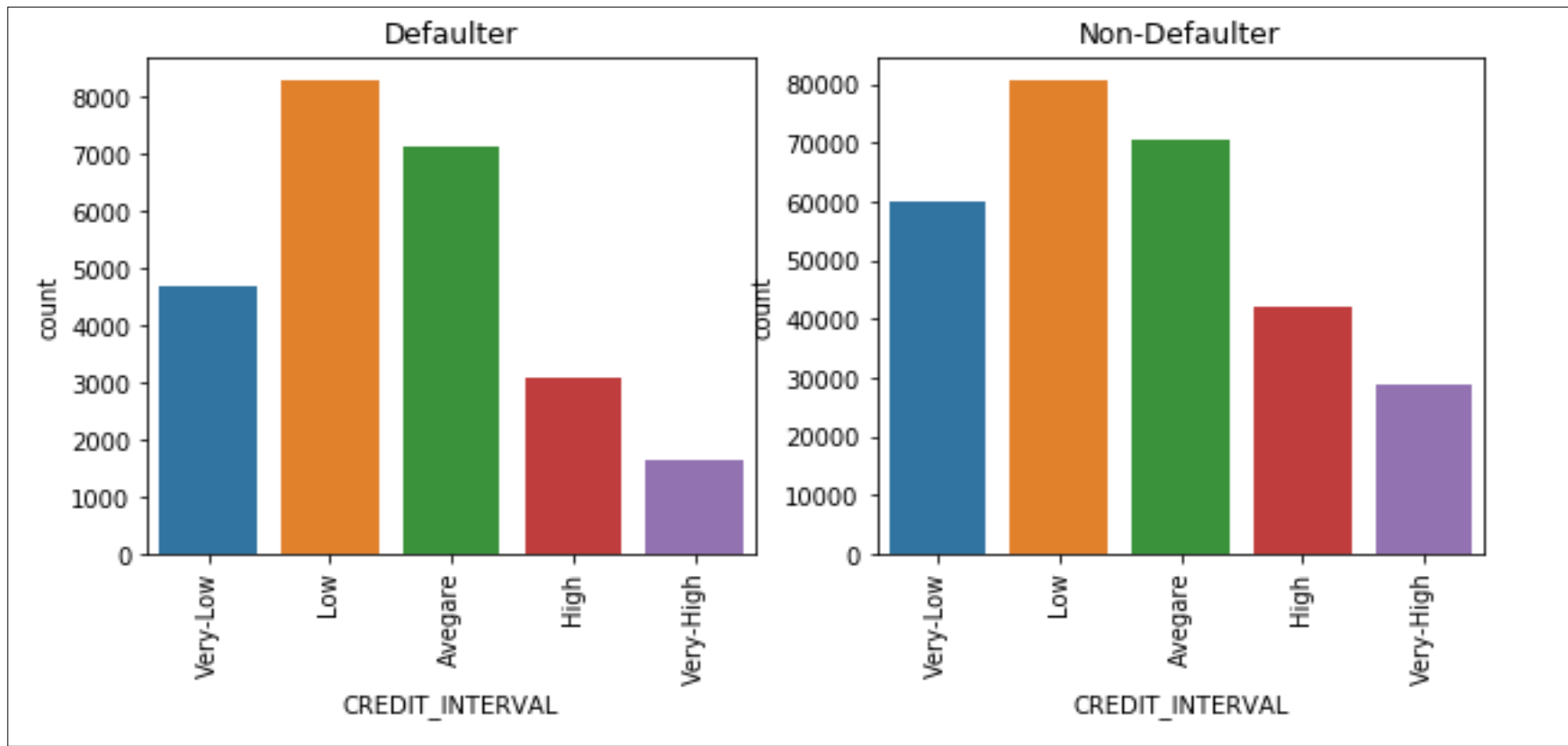
Ordered Categorical Variables

Percentage of people with higher income is low in defaulter that means person with higher salary has lesser chances of becoming a defaulter.



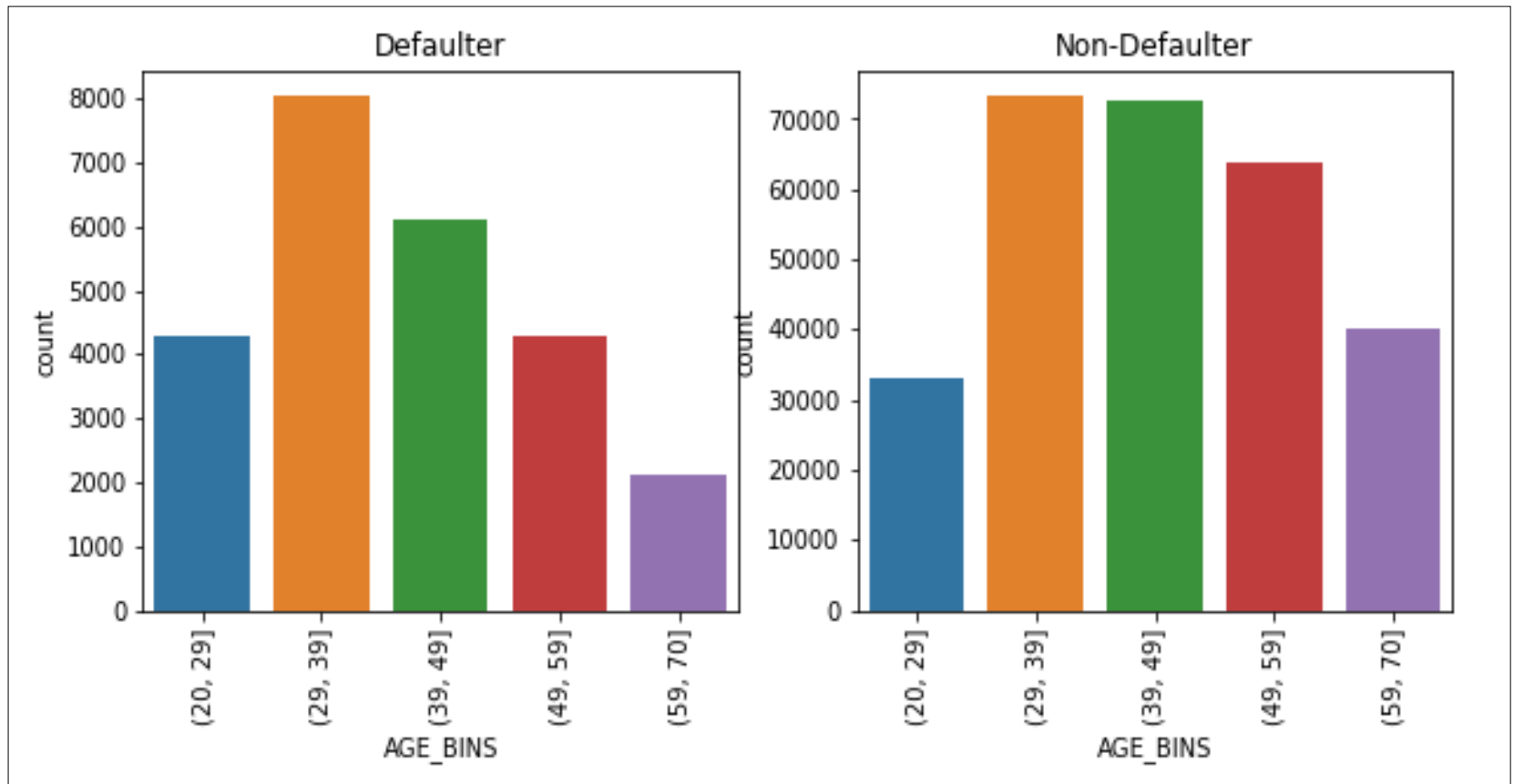
Ordered Categorical Variables

- Percentage of people with higher credit is low in defaulter that means person with higher credit has lesser chances of becoming a defaulter.
- But a defaulter person with high credit can make bigger loss to bank.



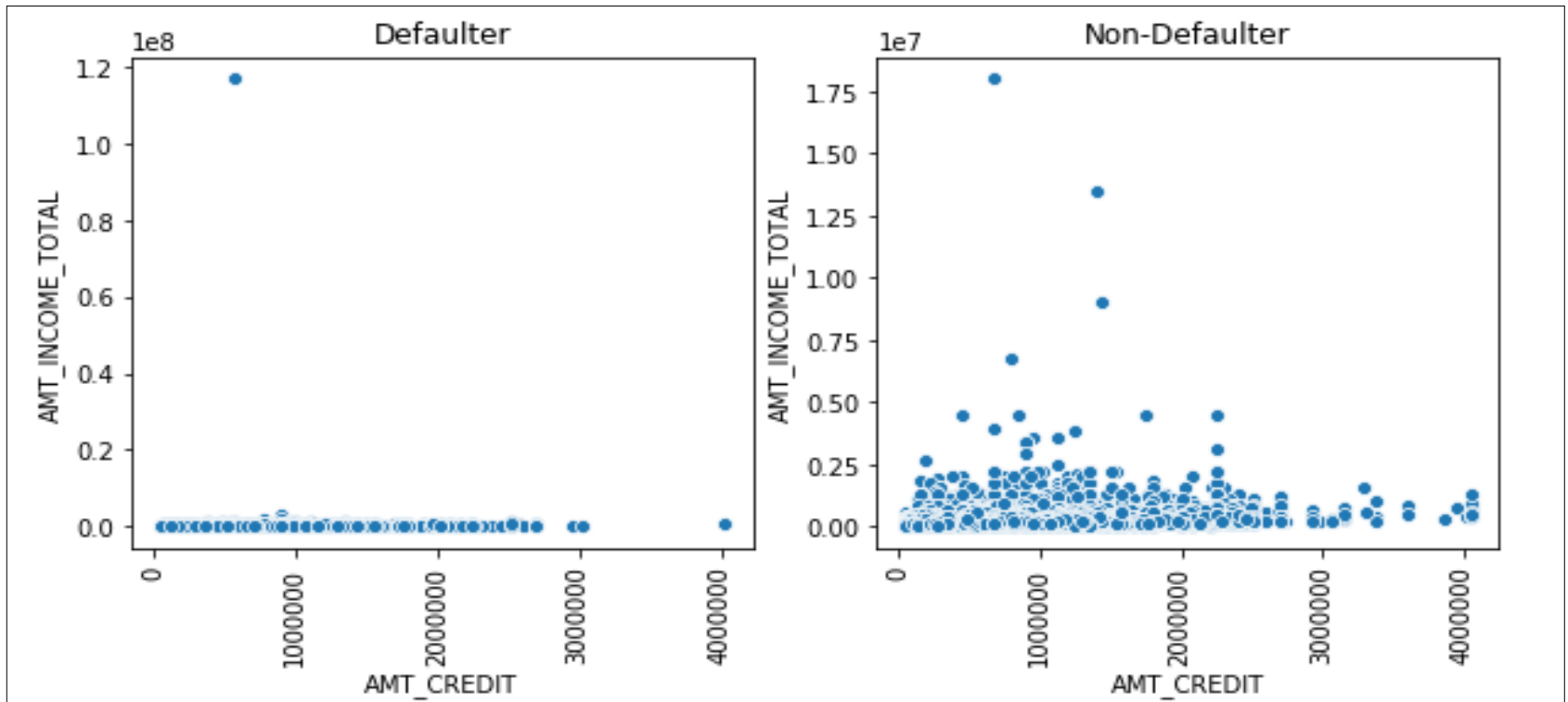
Ordered Categorical Variables

Percentage of people with older age is low in defaulter that means old people are more reliable than young in terms of repaying loans.

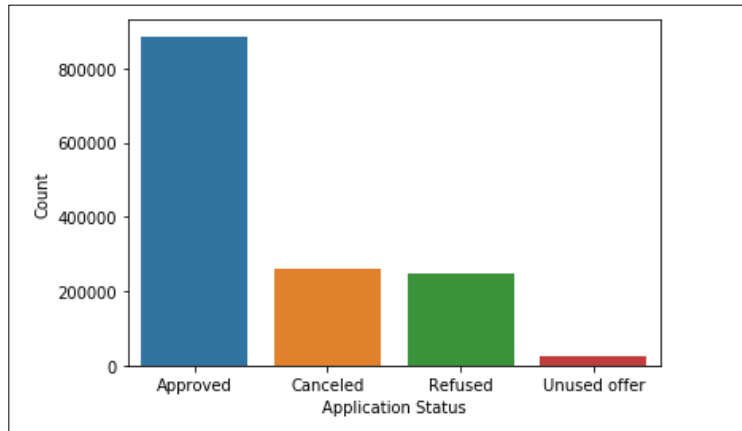


Bivariate Analysis

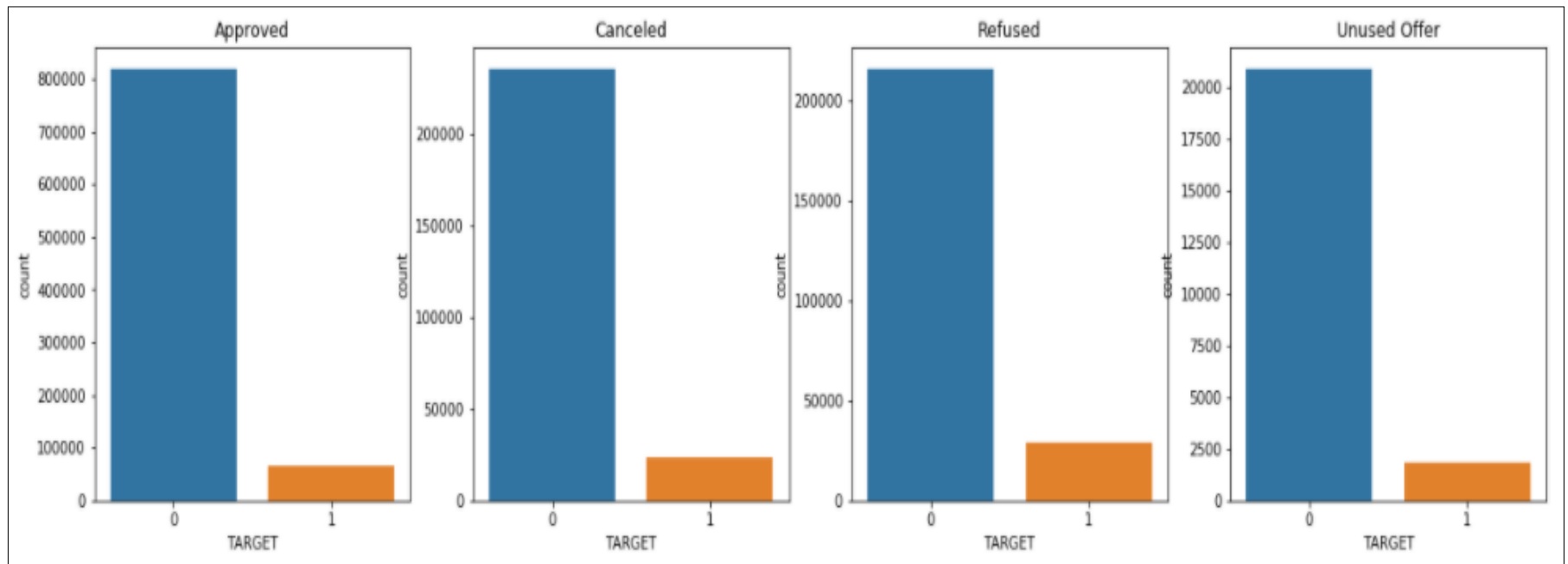
- Income distribution is more for Non-Defaulters for amount credit.
- Less defaulter credit amount greater than 30 Lakh



Merging Dataset

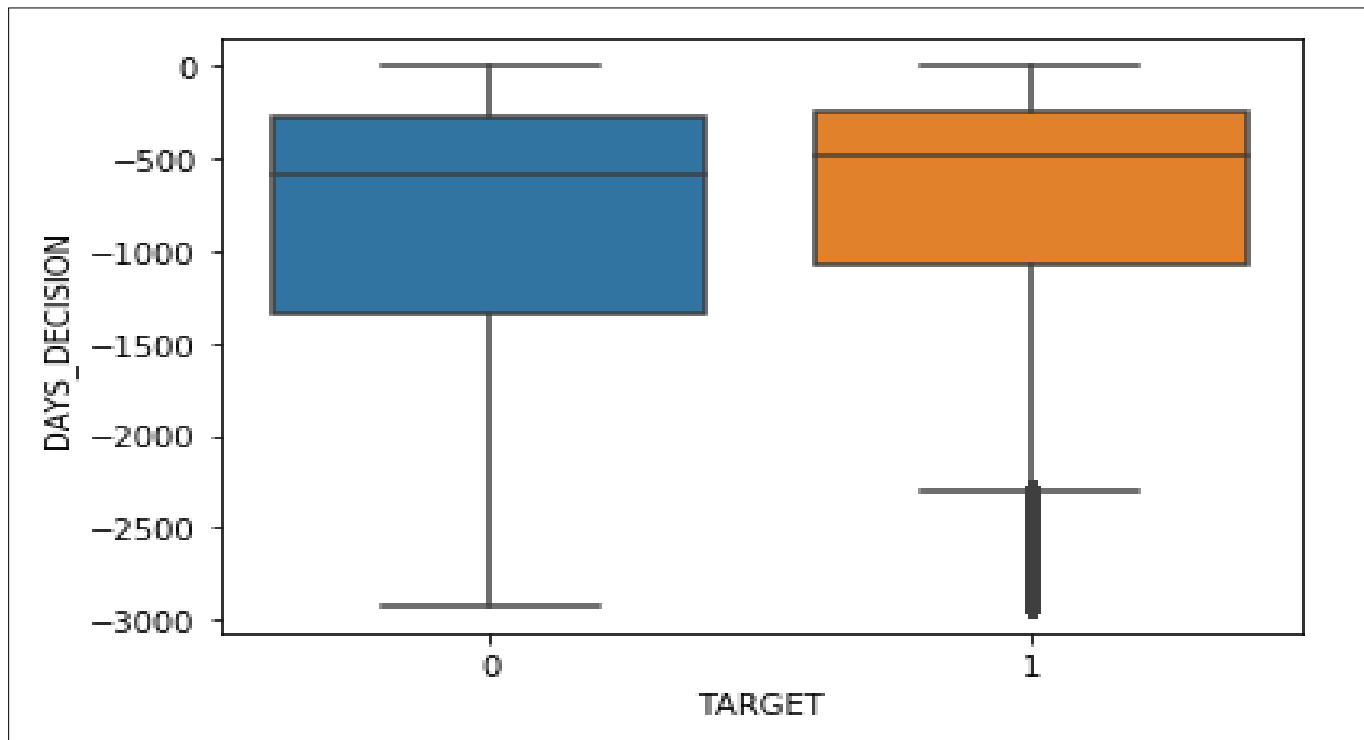


Divide dataset into 4 datasets
Refused offers has higher percentage of defaulters



Merged Dataset Analysis

As days increases in taking decision, defaulter case decreases



Conclusion

- Investigate more in case giving loan to person with low salary.
- Old age people are more reliable than young in terms of repaying loans.
- Person for whom loan was rejected earlier has a higher chance of becoming defaulter.
- More days should be taken in investigation to avoid defaulters.