Credit EDA Assignment

Submitted by: Praveen Rathi

upGrade & IITB | Data Science Program - January 2023

Scenario

Context

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it to their advantage by becoming a defaulter.

Risk

- Risk of application being approved who is likely to default in payment of loan
- Risk of application being rejected who is capable of repaying the loan

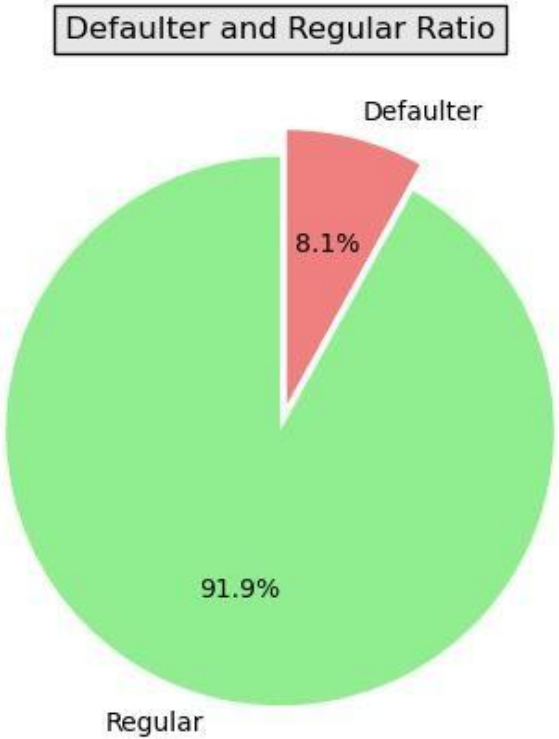
Approach

Through EDA, plot patterns, observe to gain insights of data to identify applicants with payment difficulties and good customers

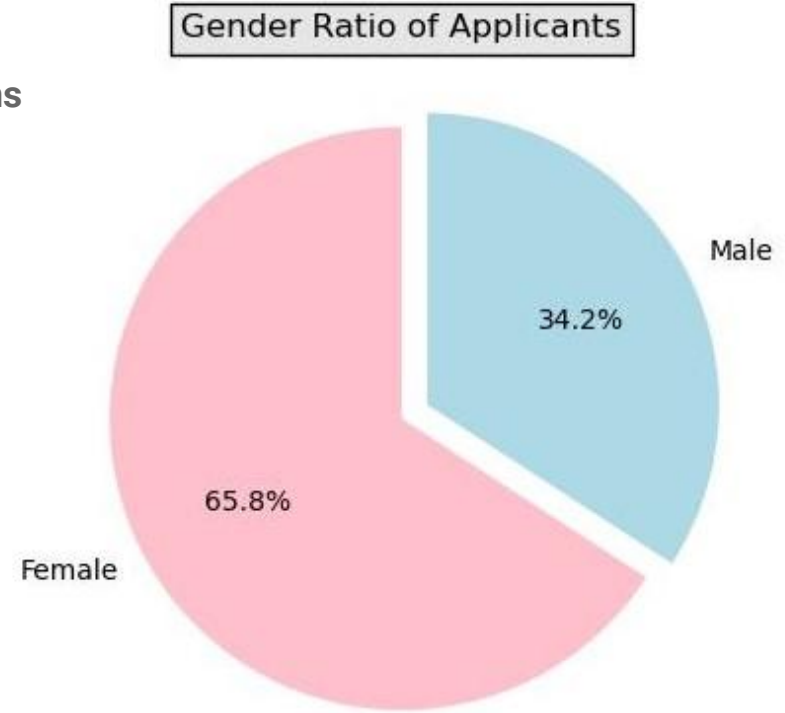
Approach step wise

- Observe overall dataset missing values and drop features having above 40% missing values
- Data standardization: negative days to positive, Target column categorise
- Univariate Analysis of categorical feature:
  - Imputing missing values with mode, drops rows having missing values
  - Plot pie chart
- Univariate Analysis of Numerical feature:
  - Plot box plot, describe to observe outliers
  - Handling outliers by binning the values for analysis
  - Plot histogram to observe distribution
- Segmented Univariate Analysis:
  - Segment dataset based on Target feature
  - Handle data imbalance: analyse the segmented dataset by converting values to percentage % and observe the pattern in each segmented dataset
- Bivariate analysis:
  - Pair plot - scatter plot of numeric features to observe the linear pattern if any
- Multivariate analysis:
  - Plot heatmap to observe the correlation among numeric features

Data Imbalance



Total Applications are 307511



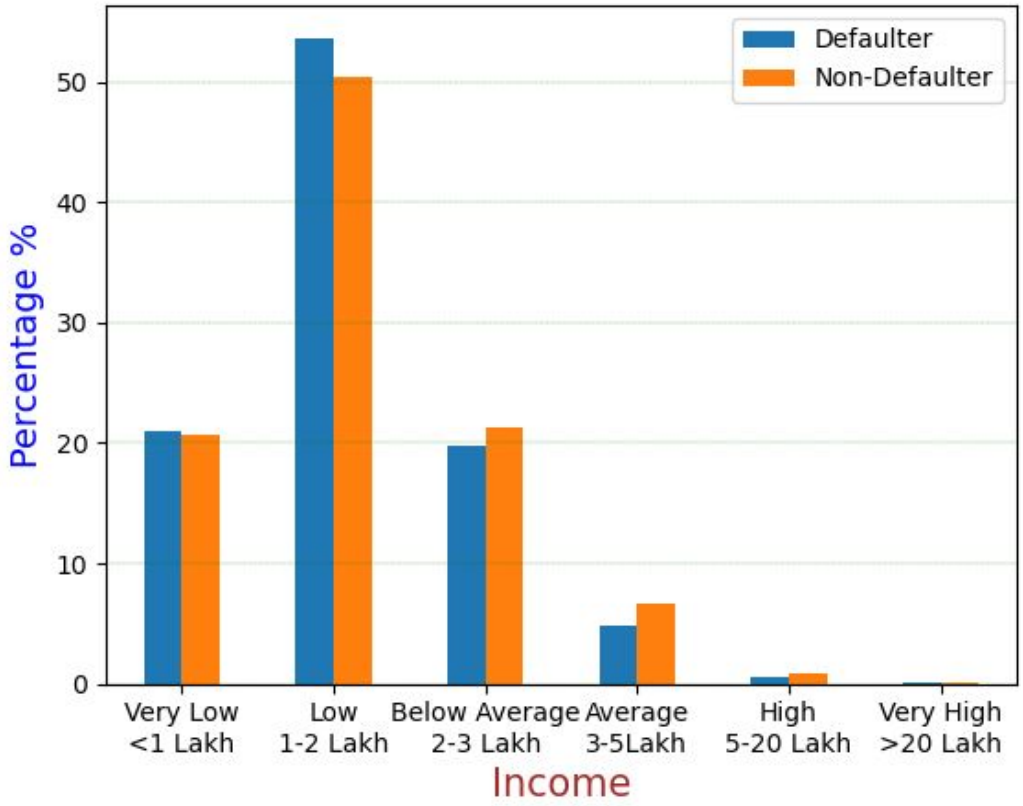
Income Analysis

Describe:

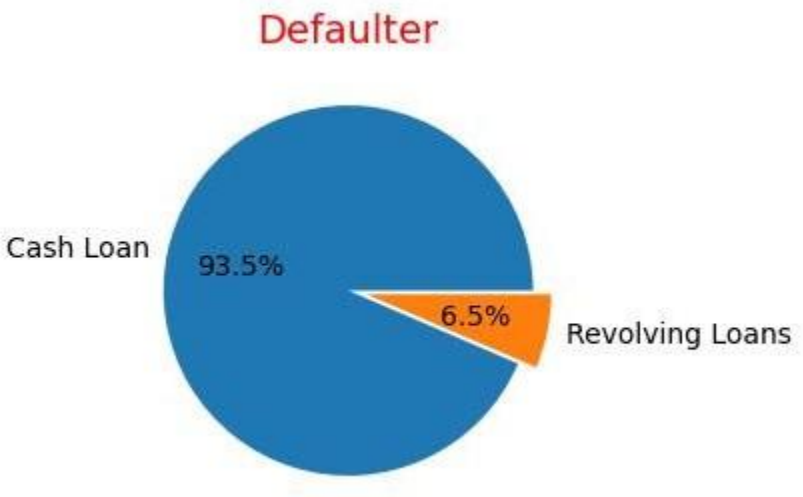
- We have received applications from 99.12% Applicants having income below 5 Lakh
- Average Income is 1.63 Lakh (Excluding applicants income more than 5 Lakh)
- Minimum Income is 0.26 Lakh only

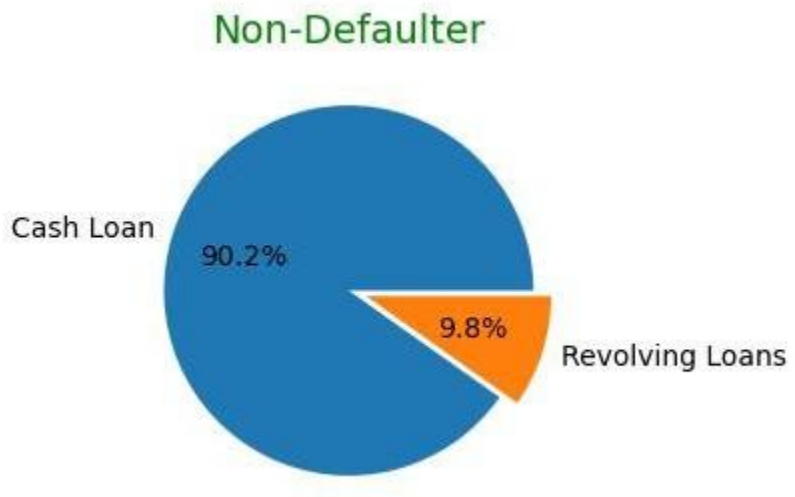
Insight:

- **Almost defaulters has income below 2 lakh**



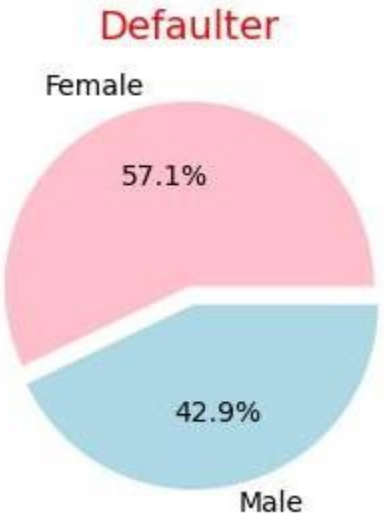
Loan Type analysis

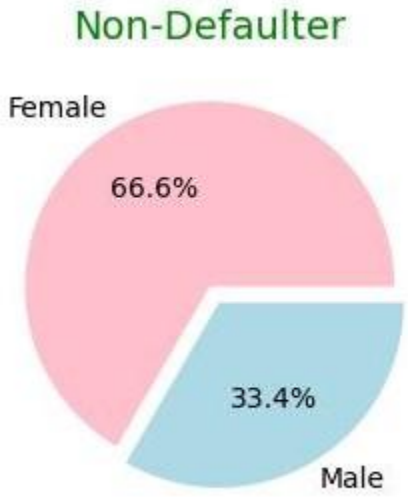




3% More chances of default in Cash Loan than Revolving Loans

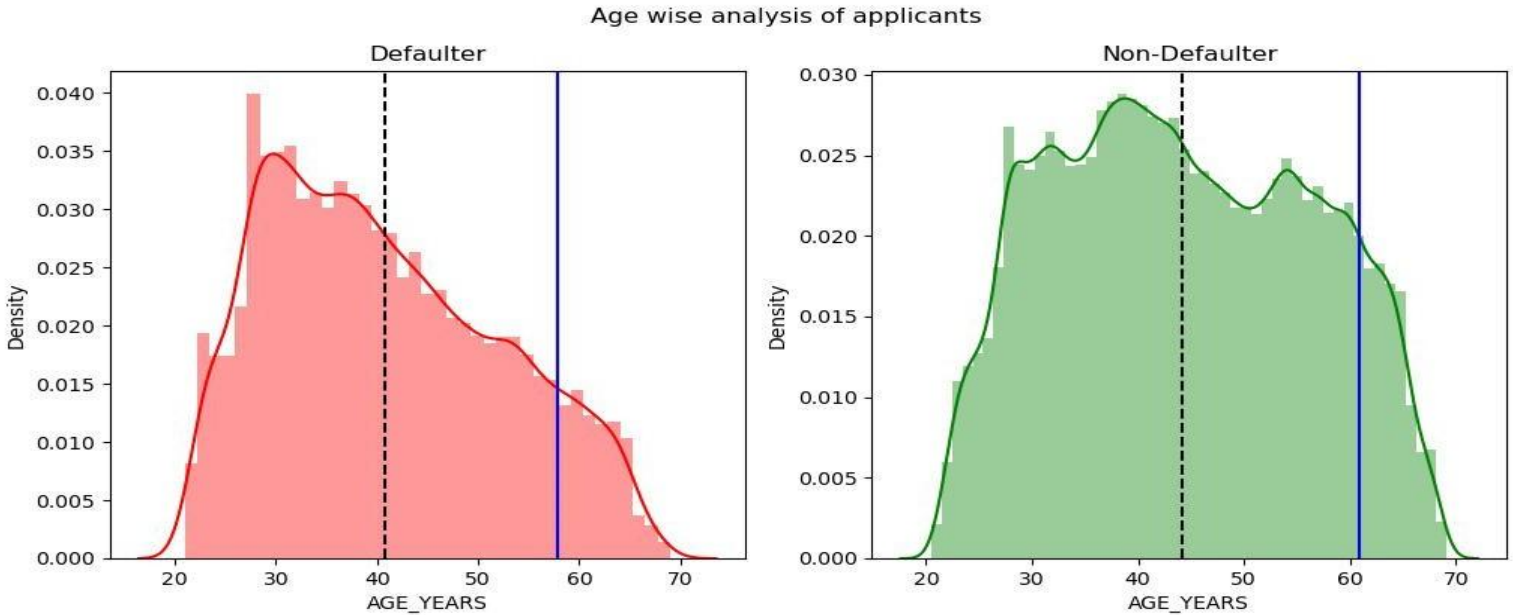
Gender wise analysis





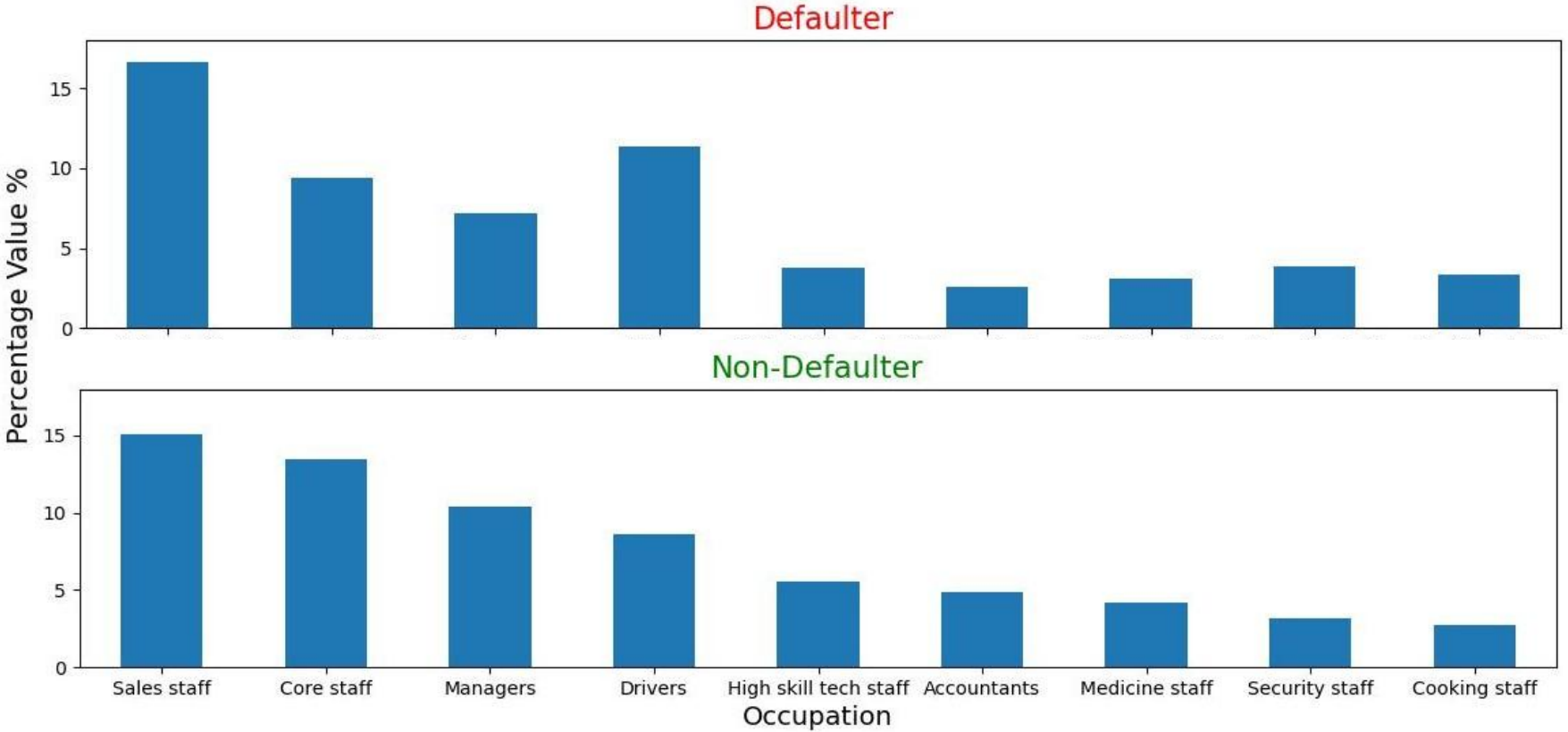
10% more chance that Males would default in payment than females

Analysis Age Group basis



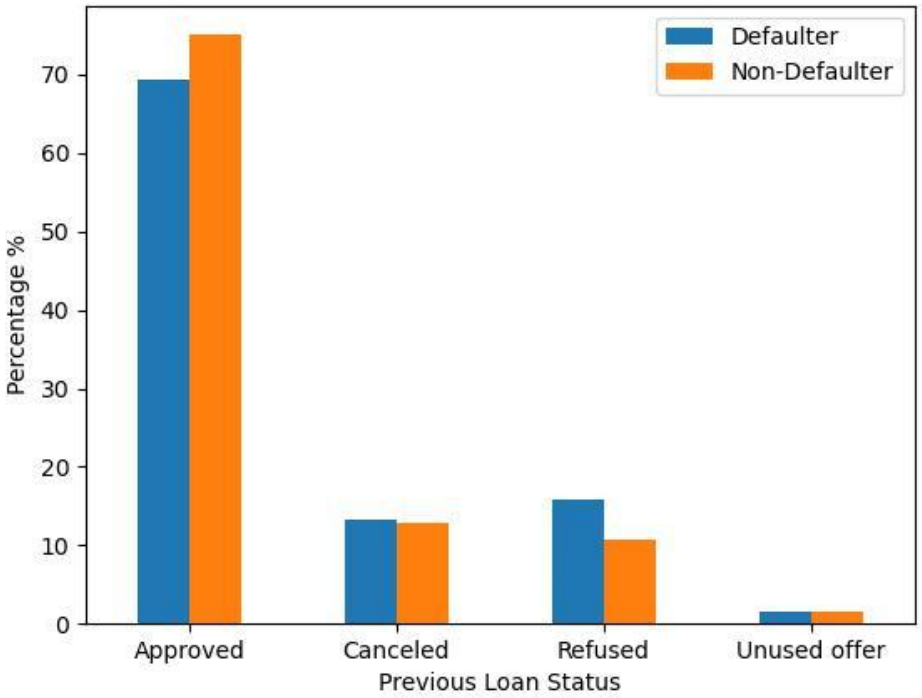
We see as the density of defaulter's age is more between 25 - 45 years whereas for non -defaulter it seems to be evenly distributed ranging from 25 - 65 years, hence this patterns shows as age is increasing defaulter counts are decreasing, we can think this as if older aged applicant is less likely to default than younger age

Analysis Occupation based



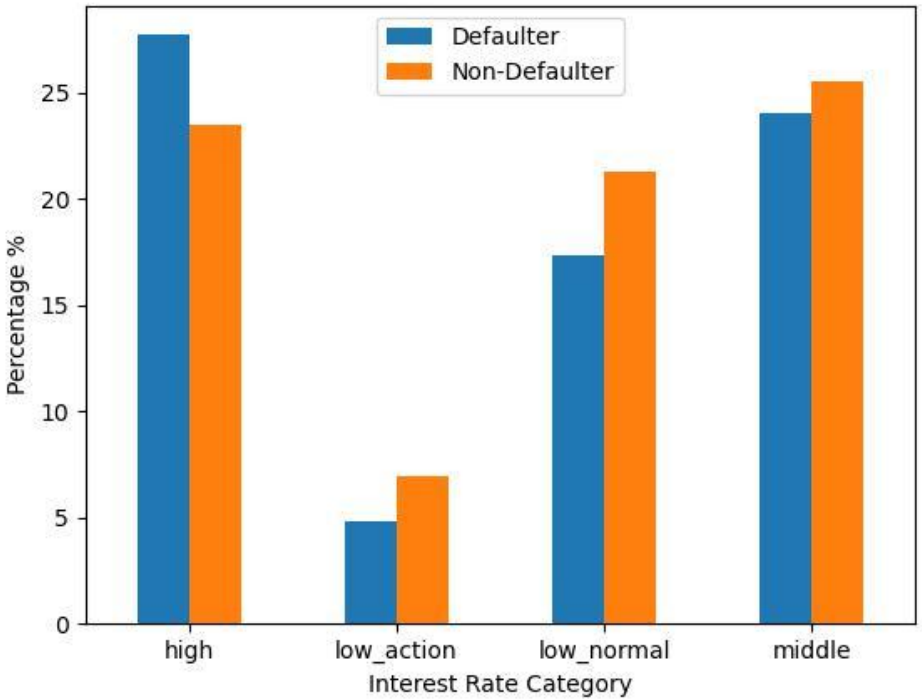
Sales Staff, Drivers are higher in defaulting than core staff and managers

Analysis Previous Loan Status



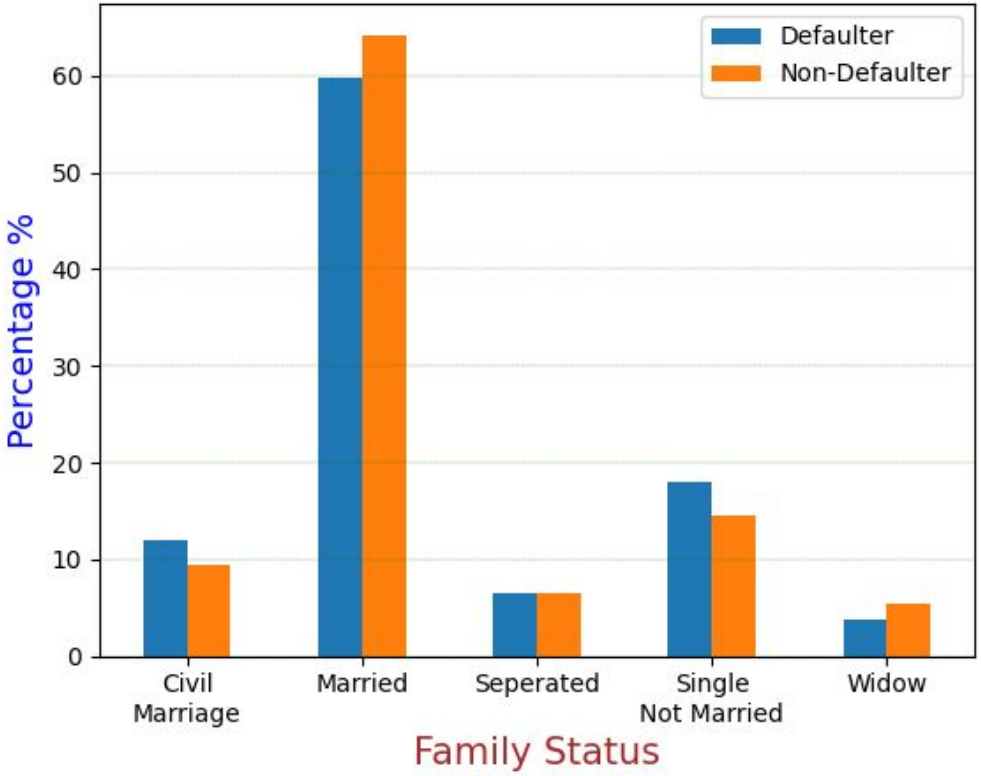
Applicants whose previous application was approved are 5% less likely to default and whose previous application was refused are 5% more likely to default

Analysis Interest Rate Category



High interest rate category applicants seems to default 5% more and others are lower side in making default

Analysis based on Family Status



Single and Civil Marriage applicants are higher in default ratio whereas Married and Widow are lower side in defaulting

Other Observations

- ➔ Observed linear relations through pair plot among credit amount, annuity amount, goods value and it's obvious that higher goods value would require higher credit amount resulting higher annuity value
- ➔ Most applicants lives in less populated region
- ➔ Applicants not owning car or reality are slightly higher side with payment difficulties than applicants owing car or reality
- ➔ There are 0.8% applicants found whose Annuity is higher than 50% of their Income. For an example where an applicant is earning Rs. 100 and spending more than Rs. 50 on Annuity, this means those applicants may face tight financial situation for paying annuity (However source of data of Income given in dataset is not known to us and it is assumed income and annuity are annual)