CREDIT EDA ASSIGNMENT

RISK ANALYTICS IN BANKING AND FINANCIAL SERVICES

PROBLEM STATEMENT

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. Suppose you work for a consumer finance company which specialises in lending various types of loans to urban customers. You have to use EDA to analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

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.

ASSUMPTIONS

- Entire analysis and visualizations are done in Python
- Features unwanted for analysis are dropped
- ‘XNA’ values are treated as NA

OVERALL APPROACH

- Datasets are loaded into the Jupiter Ipython notebook.
- Understanding the datatypes present in the datasets mainly Continuous/Numeric, Categorical/Object, Ordinal, Time..
- Data Cleaning

1. **Fixing rows and columns**

Delete unwanted Columns

Rename Columns if required

Create new Columns

2. **Impute/Remove missing values**

Identify the missing values % for all columns in the dataset

Set a threshold value for % and drop all the columns containing nulls beyond the threshold value

Identify the continuous variables with missing values and replace them with the median value.

Identify the categorical variables with missing values and replace them with mode vales and in some cases create a new category

3. **Outlier Treatment**

Identify the columns having outliers using box plot.

For continuous columns with outliers ,treat the outliers capping with upper bound values and flooring with lower bound values.

Binning the variables wherever required

Use imputation methods for getting rid of outliers.

OVERALL APPROACH

4. **Standardizing values**

Identify the columns that need a change in datatype and change accordingly.

- Univariate Analysis

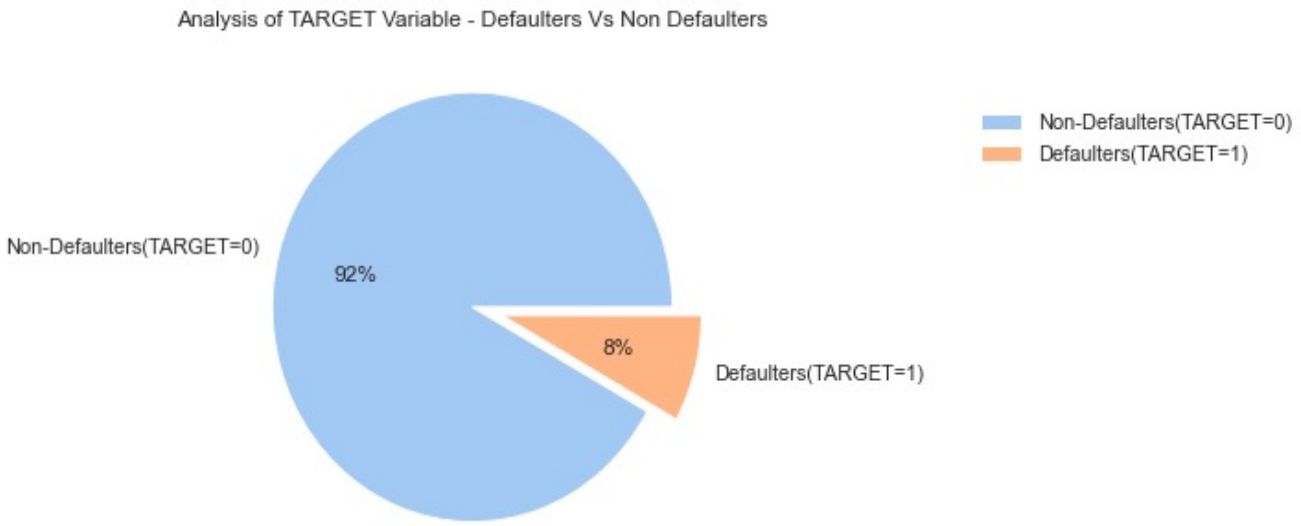
Perform univariate analysis to understand the distribution of variables across the datasets

1. Categorical unordered univariate analysis
2. Categorical ordered univariate analysis
3. Numeric variable univariate analysis

4. Segmented univariate analysis splitting the dataset based on Target variable.

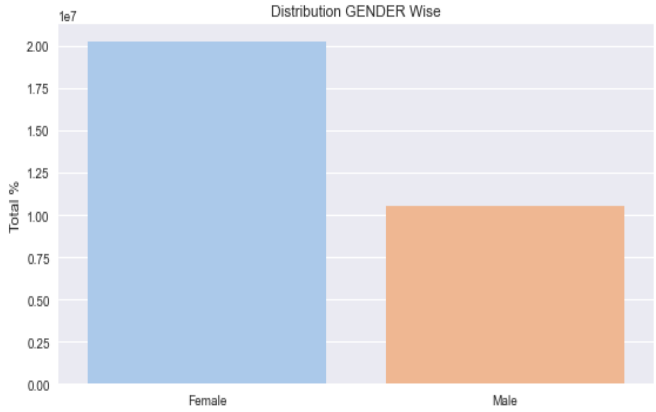
- Getting the top10 correlation between variables on segmented datasets
- Bivariate Analysis
- Multivariate Analysis
- Merging the datasets new application and previous application datasets
- Understanding the distribution based on previous application status
- Deriving the insights

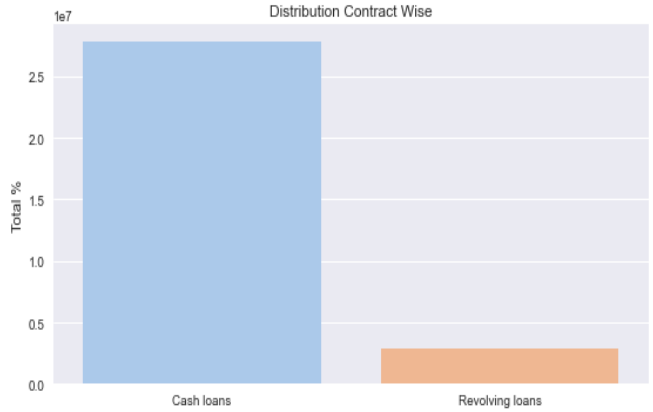
Target Variable - Imbalance

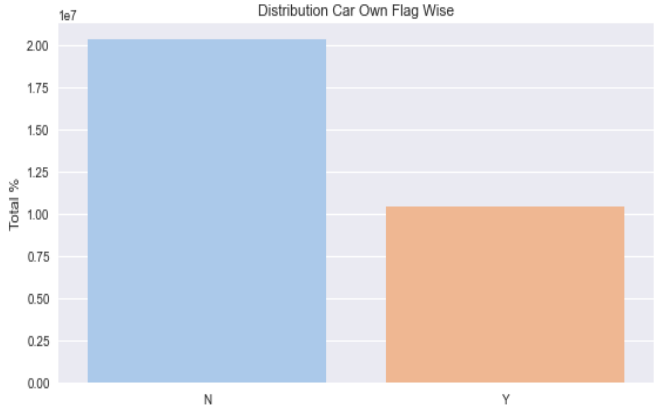


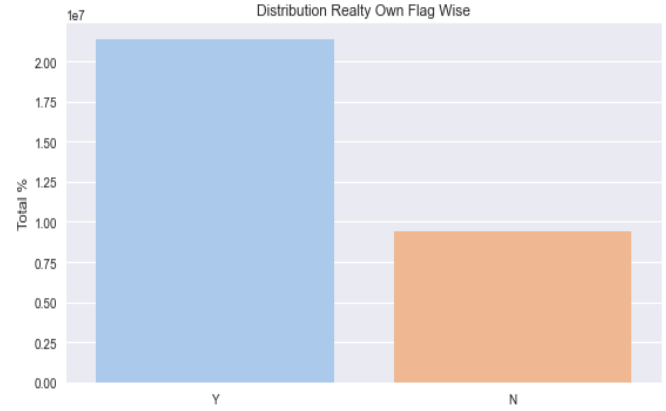
There is a clear Imbalance in the data % of Defaulters is 8% whereas for Non-Defaulters is 92%

Categorical un-ordered Univariate analysis





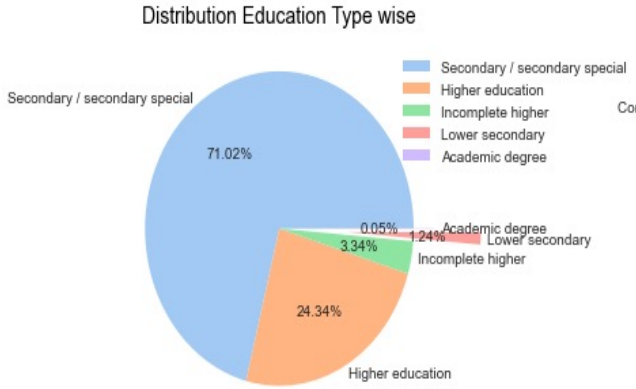


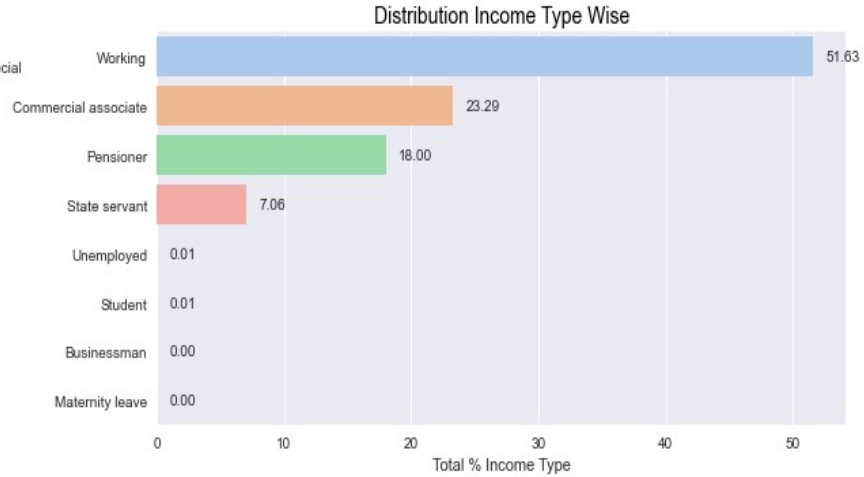


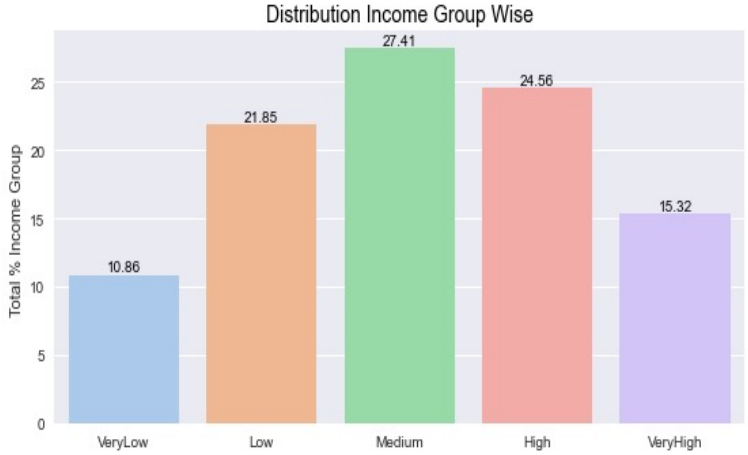
Insights:

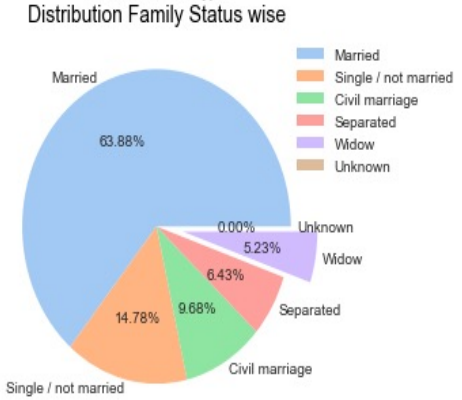
1.In the new applications,% of female applicants is higher compared to male
2.Cash loans are most preferred type of loans
3.Majority of applicants don’t own a car
4.Majority applicants own a Realty.

Categorical ordered Univariate analysis





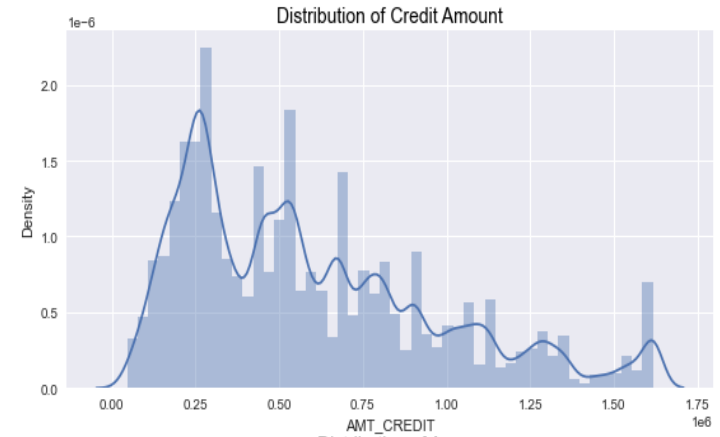


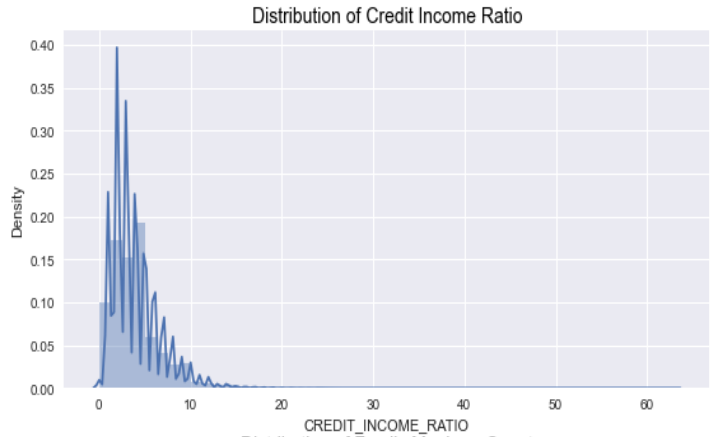


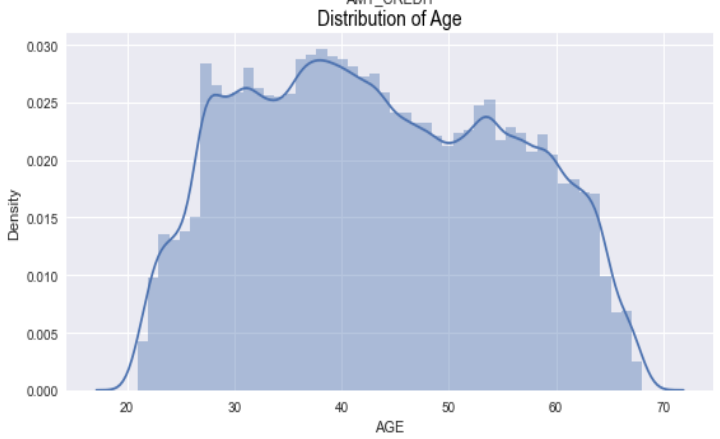
Insights:

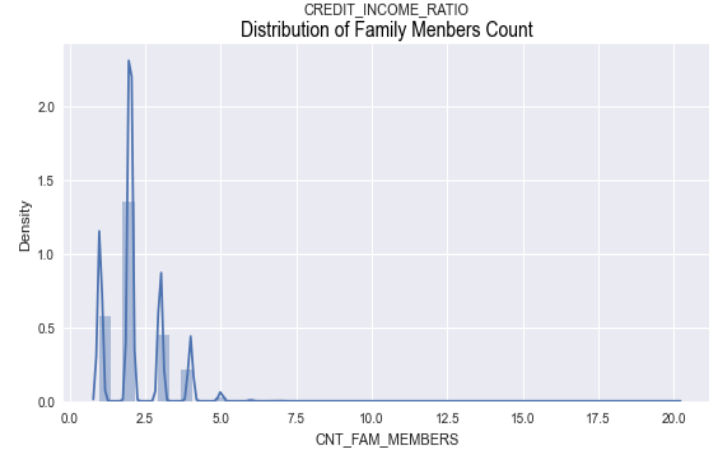
1.Large number of applicants belong to Secondary Education Category.
2.Working Professionals are opting more for loans compared to other categories.
3.Large number of applicants live in House/apartment they own.
4.Married people are the highest among the applicants.

Numerical variable Univariate analysis





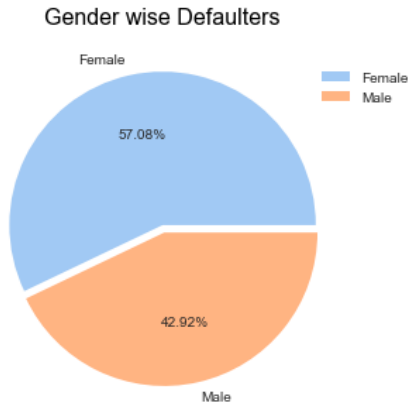


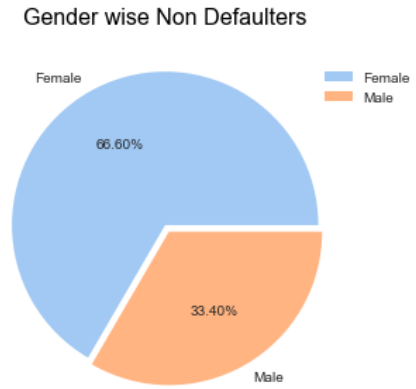


Insights:

1.Density of applicants with low Credit Amount was high,this shows most applicants were given low credit.
2.There were very high number of applicants whose Credit to Income Ratio was on lower side.
3.Max number of applicants were of age between 25-40.
4.Small families with count <=3 were among the highest applicants.

Segmented Univariate analysis

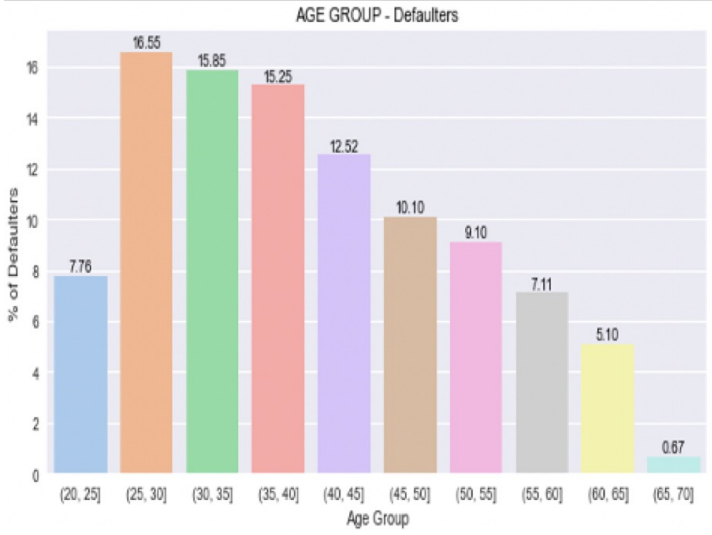


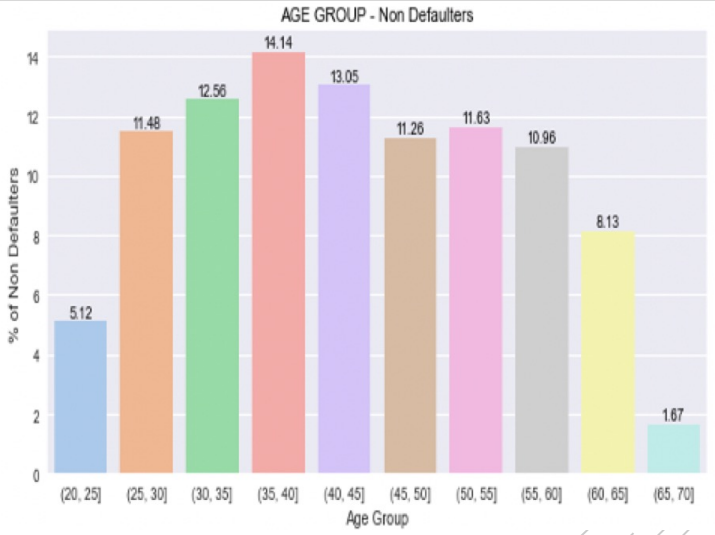


Insights:

-> It is obvious that females constitutes to higher percentage 57.08% with in defaulter segment taking into consideration the total applicants.

-> Rate of defaulters in Female are much lower compared to Males as the % of Female Defaulters is lower than % of Female Non-Defaulters in contradictory to Males.

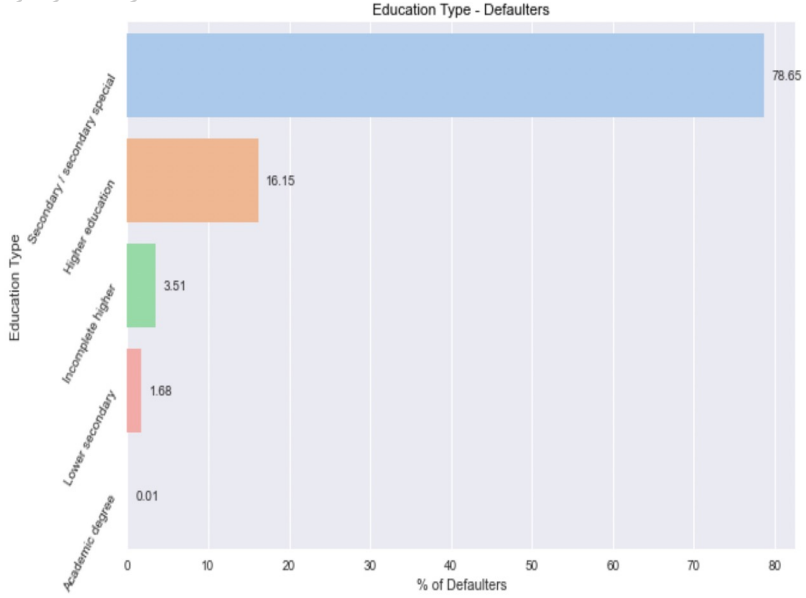


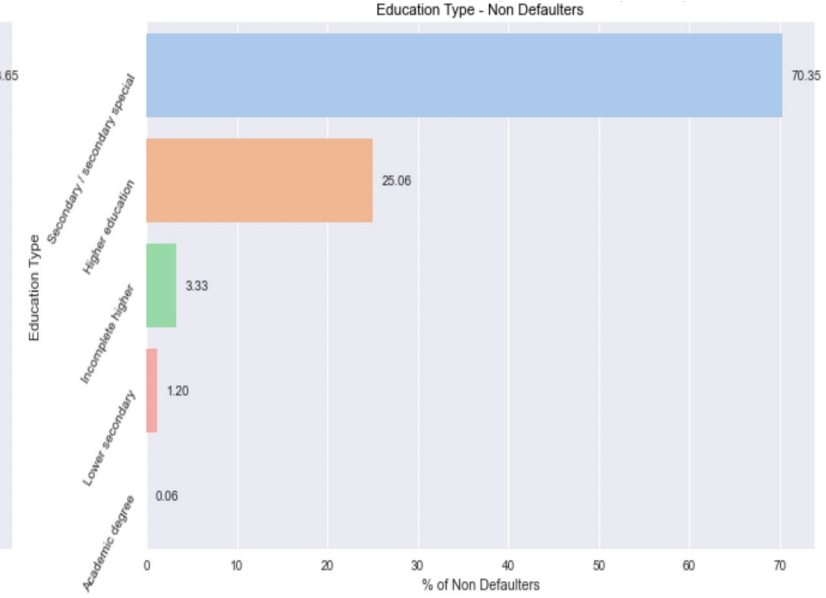


Insights

-> Applicants falling into Age group 20 to 40 are at high risk of being defaulters.

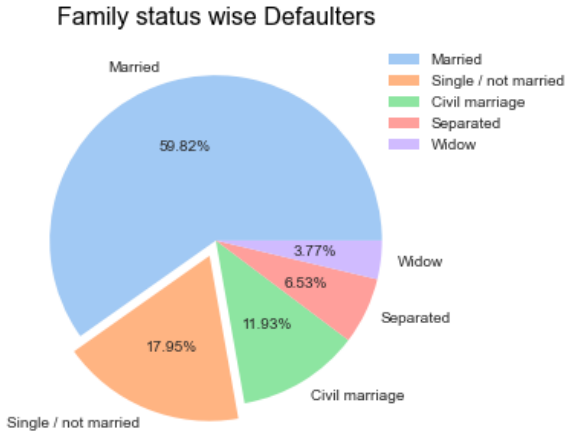
Segmented Univariate analysis

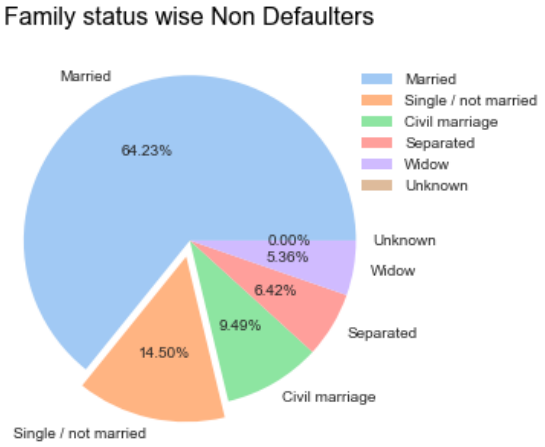




Insights

-> Rate of defaulters is high among people with Secondary or Lower Secondary Education. Higher the education, lower the risk of default.

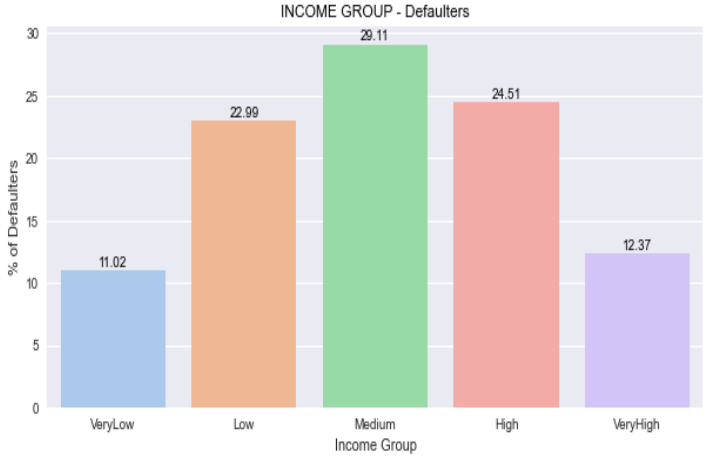


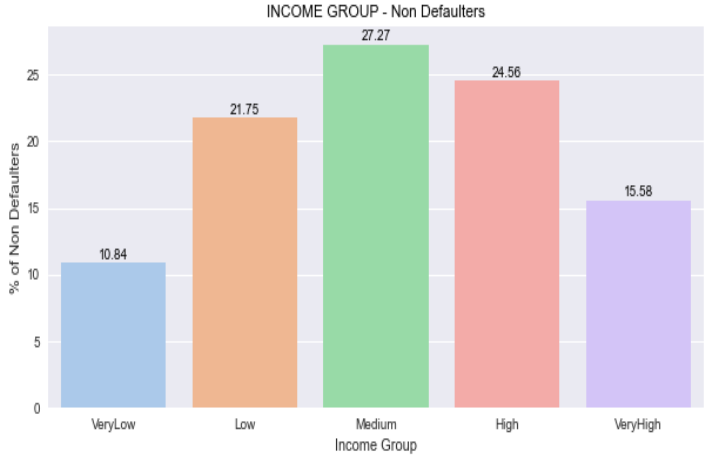


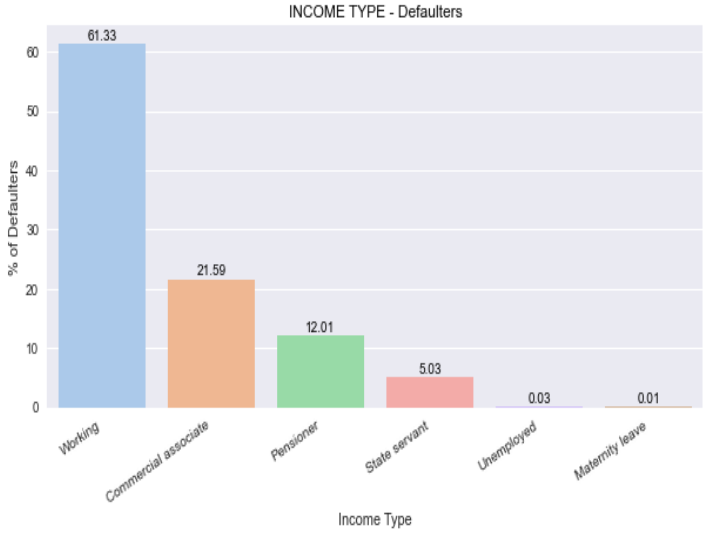
Insights

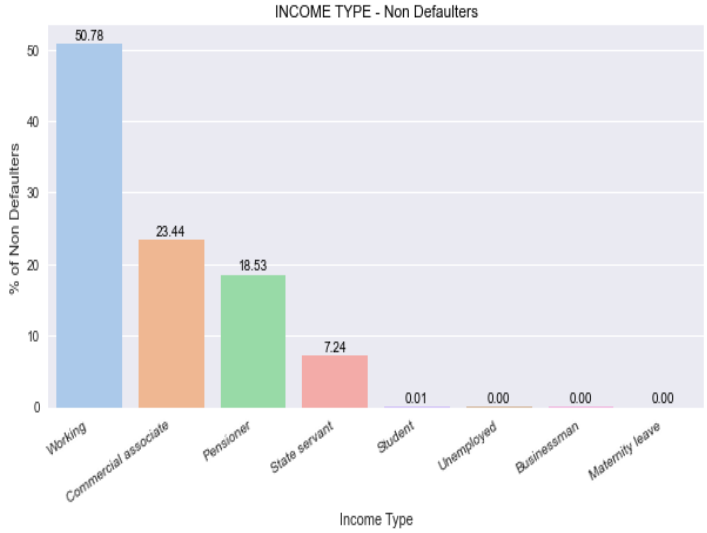
Single, Civil marriage and Separated people have the risk of default.

Segmented Univariate analysis









Insights

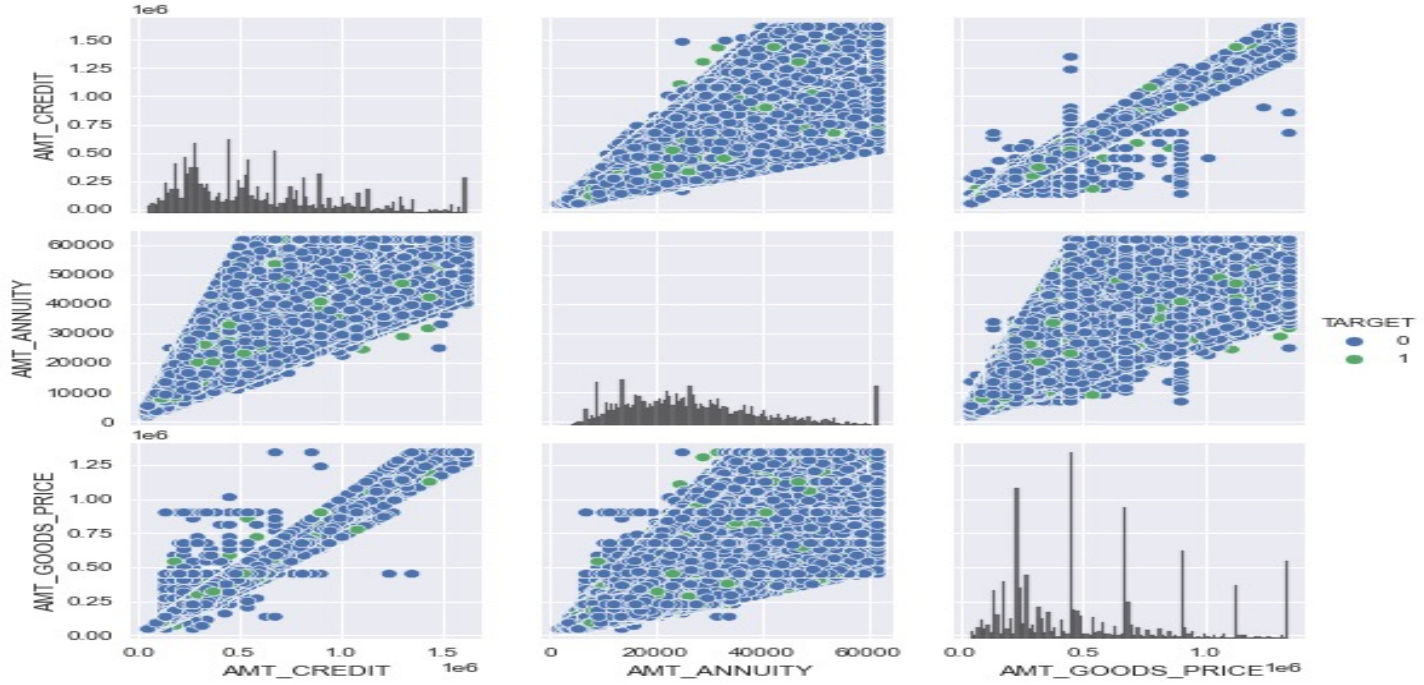
Working professionals falling into Very Low, Low and Medium Income levels fall into high rate of default category. Higher the Income of applicant, lower is the chance of default.

Bi Variate analysis

Insights

Its clear that the AMT_GOODS_PRICE and AMT_CREDIT vary linearly. Higher the goods price for the consumer loans, higher was the Credit for both Defaulters and Non-Defaulters.

We can observe the density of defaulters is decreases as Credit amount increases. Applicants taking lower credits are most likely to be the defaulters.

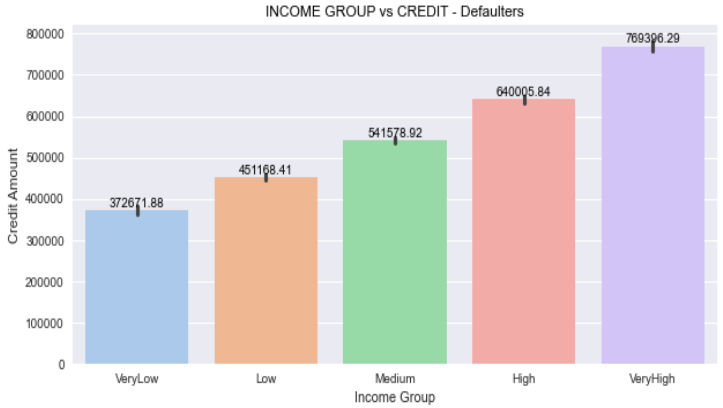


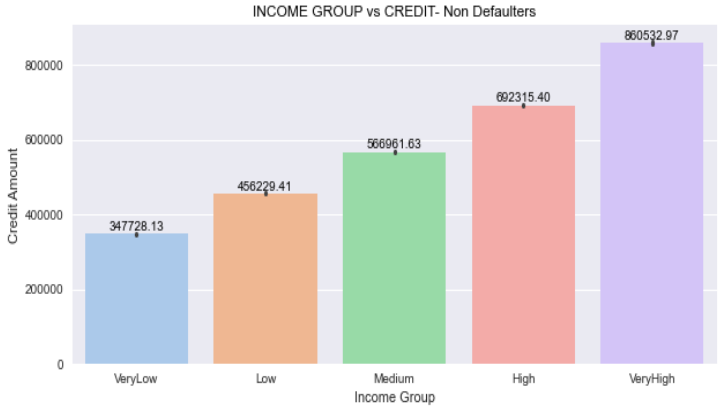
Insight

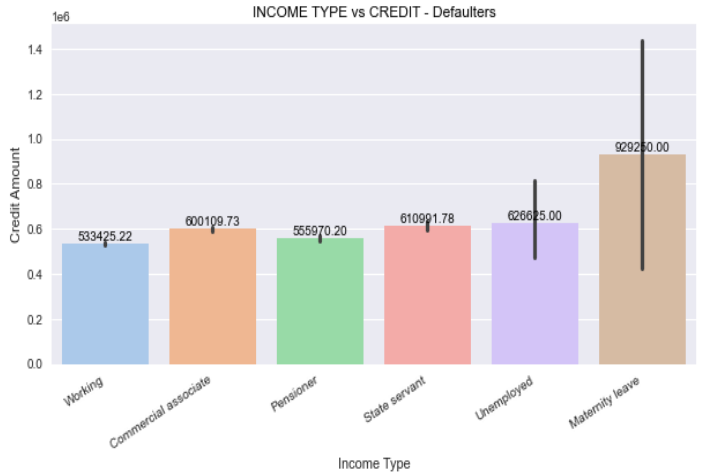
AMT_CREDIT,AMT_ANNUITY and AMT_GOODS_PRICE are linearly related to each other, if one increases other follows the same.

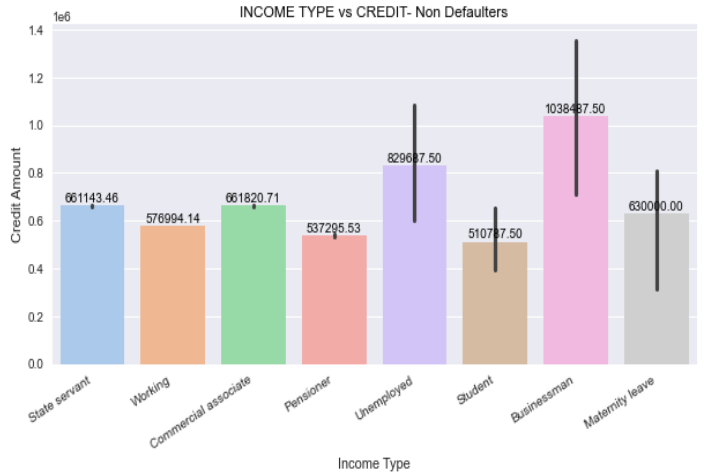
Density in all 3 features seems to be more at lower levels than higher.

Bi Variate analysis







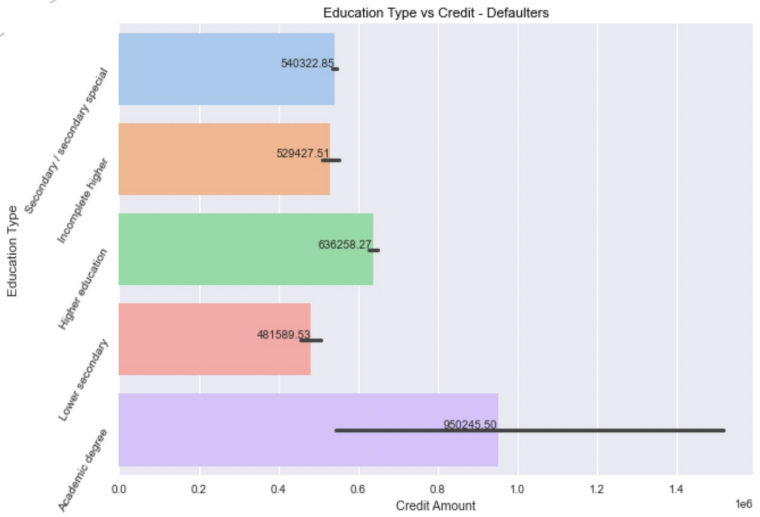


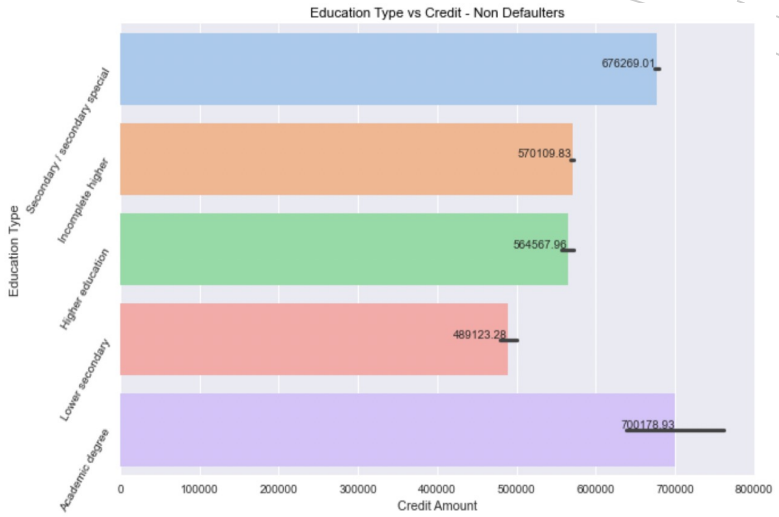
Insights

Average Credit amount is higher for Defaulters than Non-Defaulters for Very low-income group. More Credit is getting loss from this group than getting payee back.

Pensioners and Maternity leave applicants average credit amount is very high when they default its a loss to the organization.

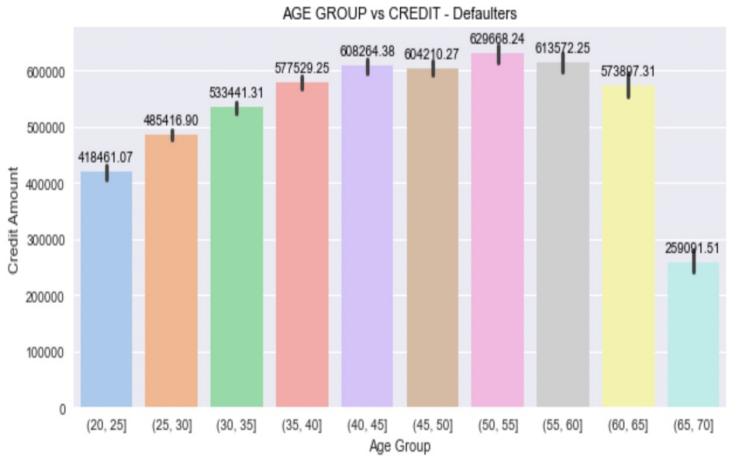
Bivariate analysis

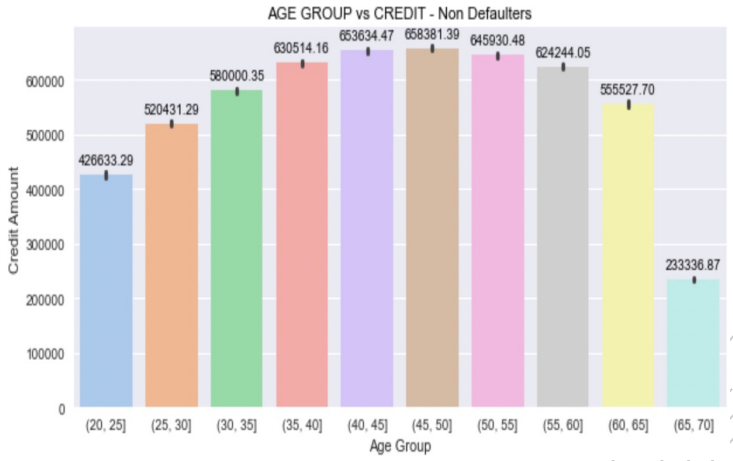




Insight

Though the less educated people constitute to defaulters more, higher education category is likely to default more credit than pay back



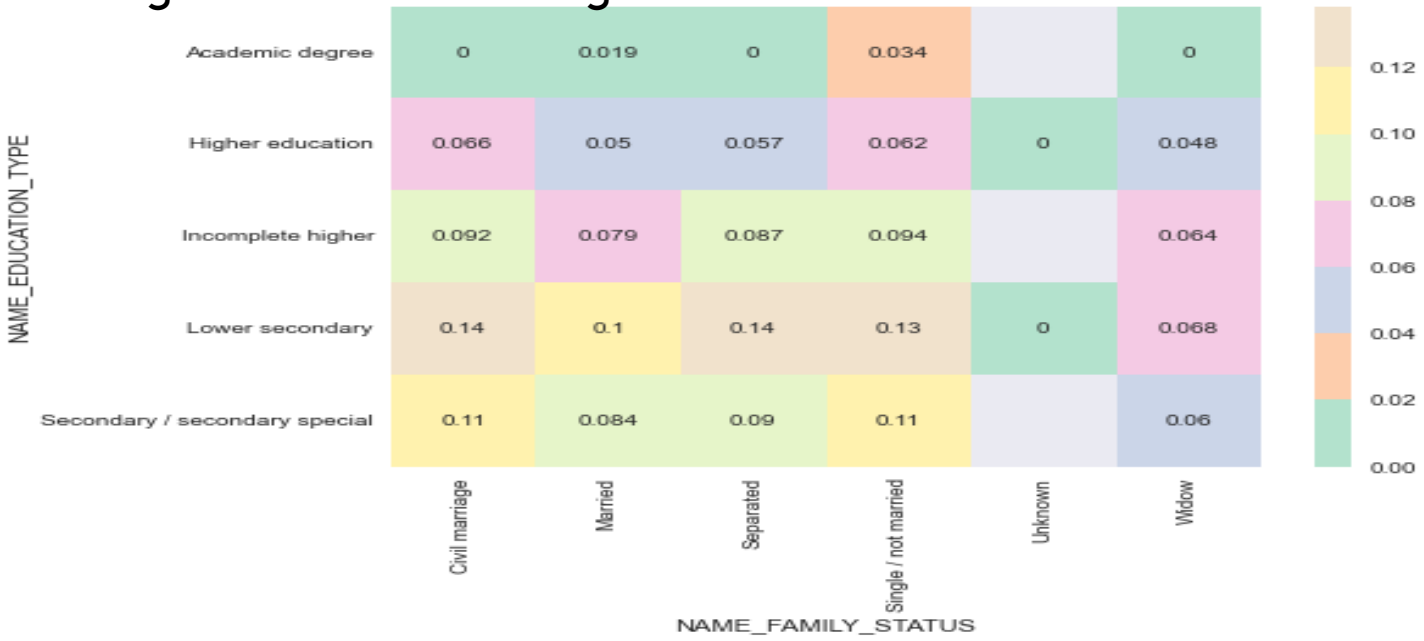


Insight

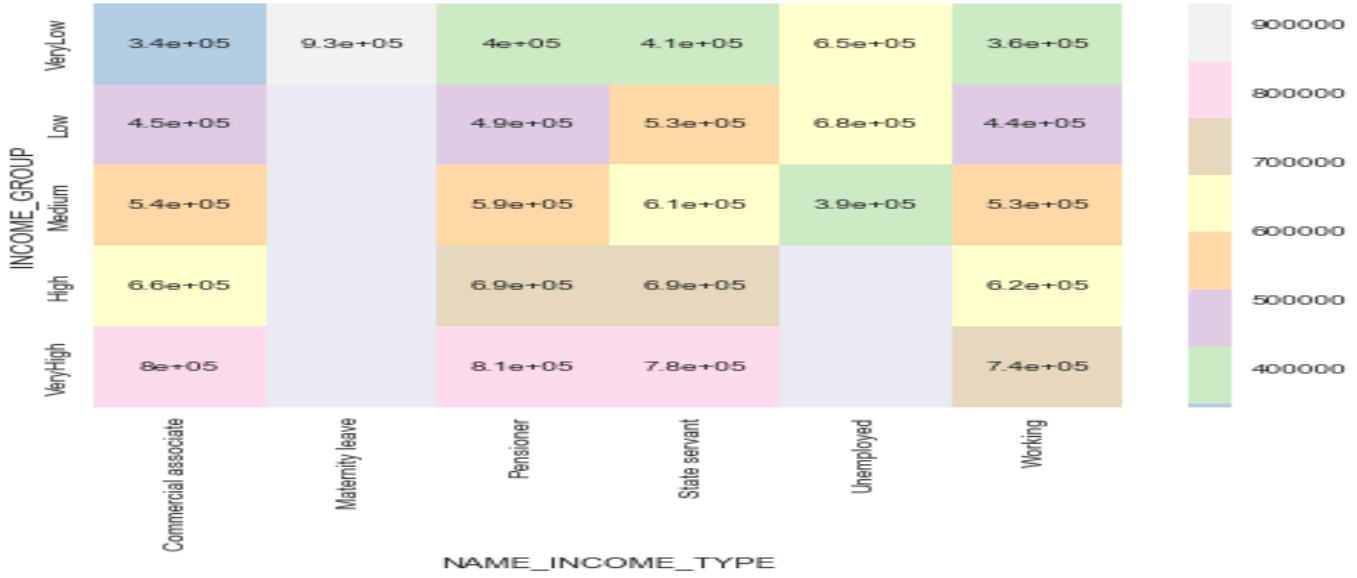
Age group >60 has higher default of Credit compared to repay.

Multivariate analysis

Highest defaulter categories

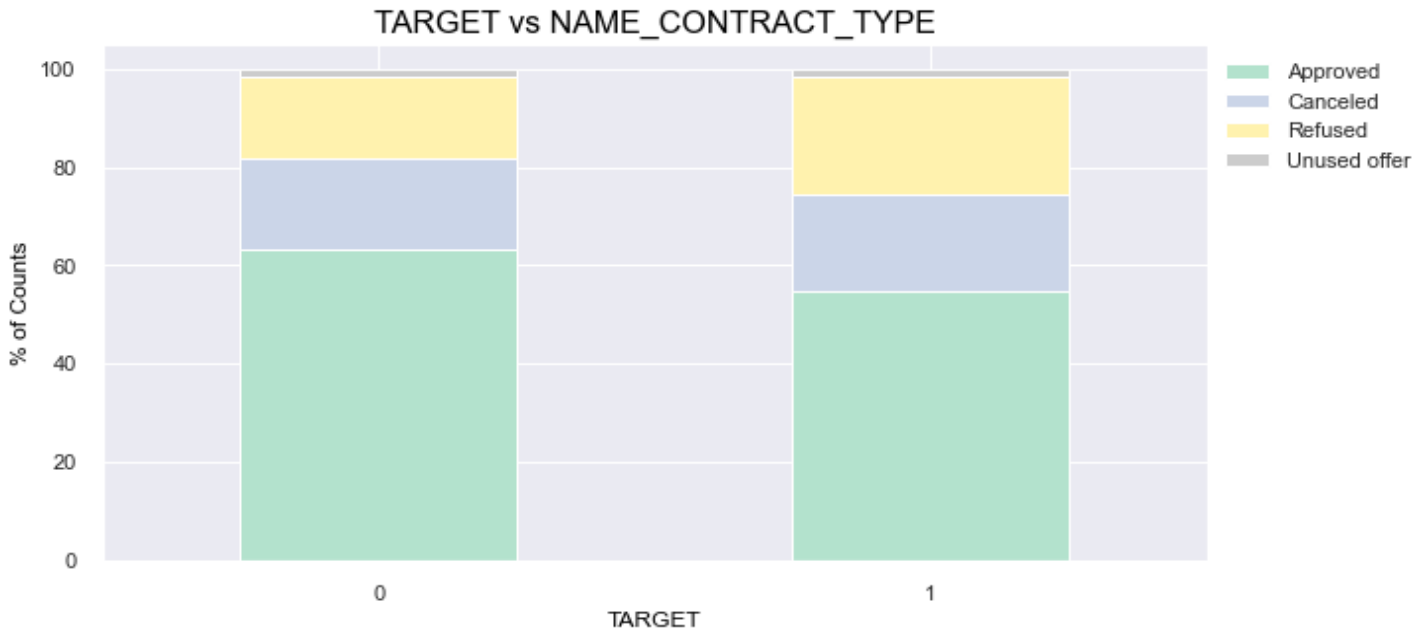


Applicants Lower secondary and Secondary education background belonging to Civil Marriage ,Single and Separated Family status are more among defaulters



Applicants belonging to Very high-income range and are Commercial associates, Pensioners and State Servants tend to contribute to more loss of credit due to default

Analysis with Previous applications

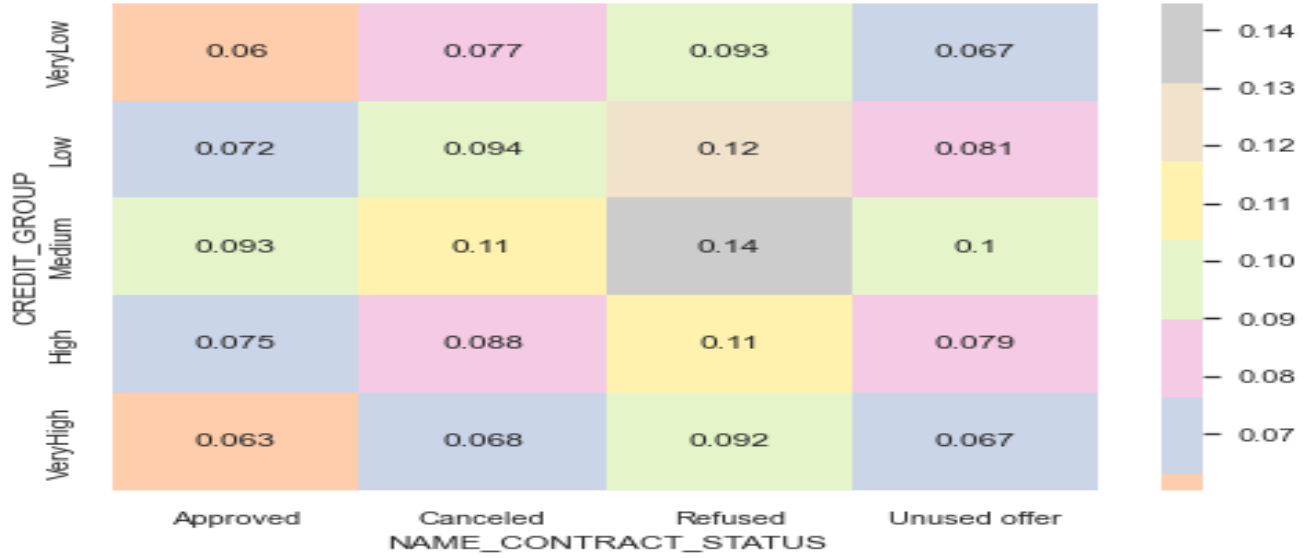


Insights

1.Defaulters are less when their previous application got Approved.

2.In contradiction of above applicants whose previous application was Refused defaulted more.

Applicants who were refused loan previously and belong to medium credit range default more



Conclusion

- Applicants opted for cash loans than revolving loans.
- Highest number of applicants belonged to Female Gender, Family status married, Education Secondary and Higher, Working professionals with Income range between low to Medium
- The Credit of loan increased with the price of goods for which loan is given
- Below are found to be high risk categories wrt defaulters

  Gender: Male

  Age Groups: 20 – 40

  Education : Lower Secondary and Secondary

  Income Groups: Very low ,Low and Medium

  Income Type: Working

  Family Status: Single, Civil Marriage

  Occupation Type: Laborer, Sales Staff and Drivers

  Region Rating Client : 3
- Below categories are causing high loss of credit due to default as the average credit is very high.

  It is suggested to avoid high Credit Amount for these.

  Age Group: >60

  Work Exp:>30 years

  Education : Higher Education

  Income Type: Pensioner

  Family Status: Widow, Civil Marriage
- The applicants whose Previous application status was Reject, are becoming highest defaulters.