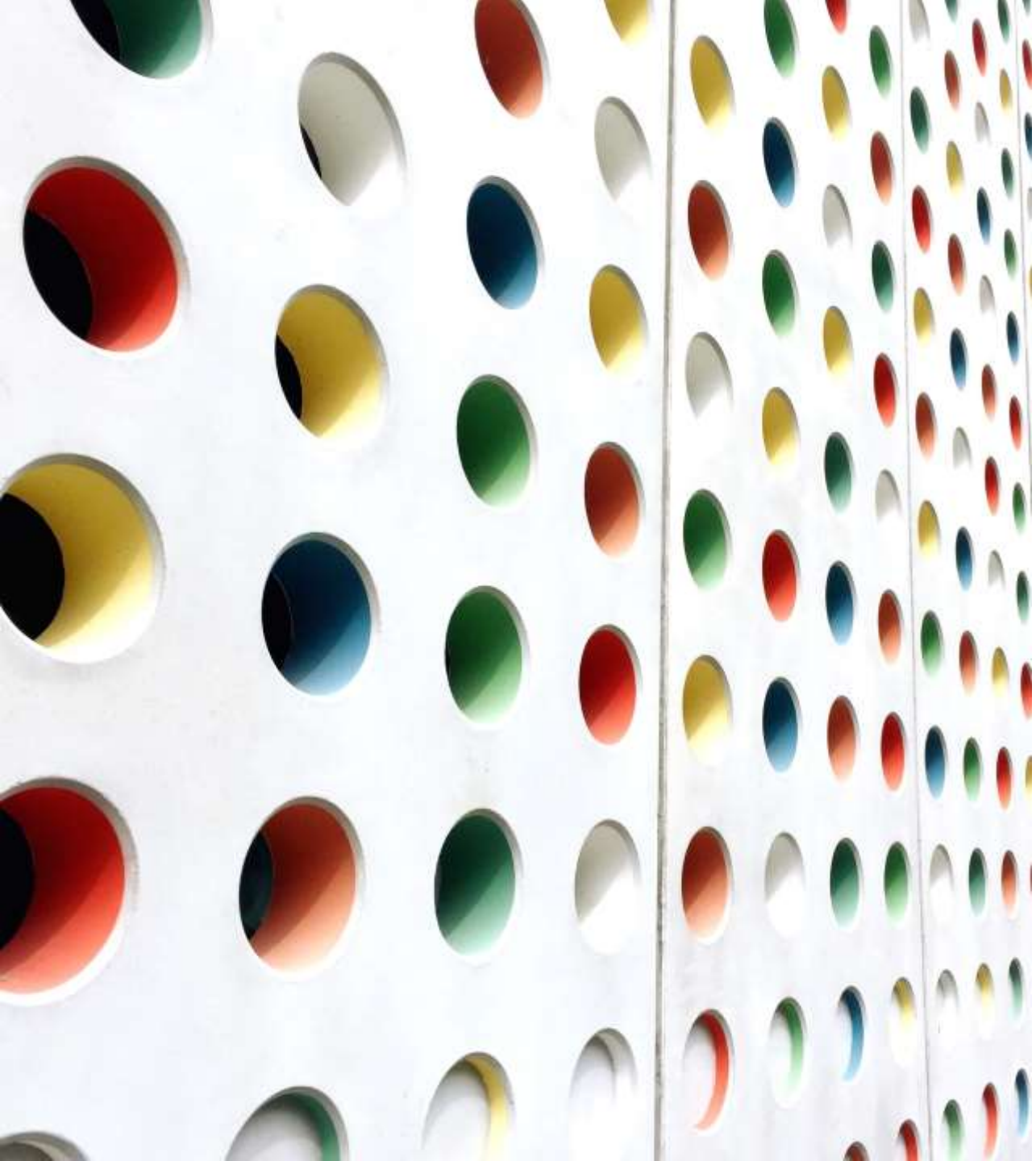


Credit EDA Assignment

AMLAN ANSUMAN ROUT

Business Understanding:

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.

The data given below contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios: The client with payment difficulties: he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample, All other cases: All other cases when the payment is paid on time.

When a client applies for a loan, there are four types of decisions that could be taken by the client/company):

Approved: The Company has approved loan Application Cancelled: The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client, he received worse pricing which he did not want. Refused: The company had rejected the loan (because the client does not meet their requirements etc.). Unused offer: Loan has been cancelled by the client but at different stages of the process.

In this case study, you will use EDA to understand how consumer attributes and loan attributes influence the tendency to default.

Business Objectives / Problem Statement:

This case study aims to identify patterns which indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

To develop your understanding of the domain, you are advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough.

Expected Results

Present the overall approach of the analysis in a presentation. Mention the problem statement and the analysis approach briefly. Identify the missing data and use appropriate method to deal with it. (Remove columns/or replace it with an appropriate value) Hint: Note that in EDA, since it is not necessary to replace the missing value, but if you have to replace the missing value, what should be the approach. Clearly mention the approach. Identify if there are outliers in the dataset. Also, mention why do you think it is an outlier. Again, remember that for this exercise, it is not necessary to remove any data points. Identify if there is data imbalance in the data. Find the ratio of data imbalance. Hint: How will you analyse the data in case of data imbalance? You can plot more than one type of plot to analyse the different aspects due to data imbalance. For example, you can choose your own scale for the graphs, i.e. one can plot in terms of percentage or absolute value. Do this analysis for the 'Target variable' in the dataset (clients with payment difficulties and all other cases). Use a mix of univariate and bivariate analysis etc. Hint: Since there are a lot of columns, you can run your analysis in loops for the appropriate columns and find the insights. Explain the results of univariate, segmented univariate, bivariate analysis, etc. in business terms. Find the top 10 correlation for the Client with payment difficulties and all other cases (Target variable). Note that you have to find the top correlation by segmenting the data frame w.r.t to the target variable and then find the top correlation for each of the segmented data and find if any insight is there. Say, there are 5+1(target) variables in a dataset: Var1, Var2, Var3, Var4, Var5, Target. And if you have to find top 3 correlation, it can be: Var1 & Var2, Var2 & Var3, Var1 & Var3. Target variable will not feature in this correlation as it is a categorical variable and not a continuous variable which is increasing or decreasing.
Include visualisations and summarise the most important results in the presentation. You are free to choose the graphs which explain the numerical/categorical variables. Insights should explain why the variable is important for differentiating the clients with payment difficulties with all other cases. You need to submit one/two Ipython notebook which clearly explains the thought process behind your analysis (either in comments of markdown text), code and relevant plots. The presentation file needs to be in PDF format and should contain the points discussed above with the necessary visualisations. Also, all the visualisations and plots must be done in Python(should be present in the Ipython notebook), though they may be recreated in Tableau for better aesthetics in the PPT file.

EDA Approach/Steps:

1. Understanding business scenario and objective of performing EDA

2. Data collection and inspection for current application dataframe

3. Data cleaning:
    a. Fixing Rows and columns
        - Incorrect or extra rows
        - Summary rows
        - Missing or inconsistent or misalligned column names
    b. Handling missing values:
        - Dropping columns with > 35% missing values
        - Imputing remaining missing values appropriately using mean, mode, median, Unknown type
        - Handling negative values in numeric columns where appropriate. e.g. Days columns
        - Handling any disguised missing values like XNA, XAP, etc
        - Standardising text/numbers as required. e.g. converting flags to 1 or 0, etc
        - Identifying and dropping additional columns not relevant for this analysis
        - Creating derived columns as appropriate. e.g.
            - Summing all Credit Bureau Request to single column
            - Converting Days birth to Age, Days employed to Years employed

    c. Oulier detection and treatment:
        - Removing outliers
        - Binning ordered categorical columns to segments. e.g. Income amount, Credit amount, External source
    d. Identifying data imbalance and ratio

4. Analysis and gaining insights from data:
    a. Univariate analysis on categorical unordered columns
    b. Univariate analysis on categorical ordered columns
    b. Univariate analysis on numerical columns
    c. Univariate analysis on segmented ordered categorical columns
    d. Bivariate and multivariate analysis on:
        - numeric-numeric columns
        - numeric-categorical columns
        - categorical-categorical columns
        - segmented categorical columns

5. Data collection, inspection, cleaning and analysis for previous application dataframe

6. Merging current and previous application dataframes using SK_ID_CURR (Loan ID)

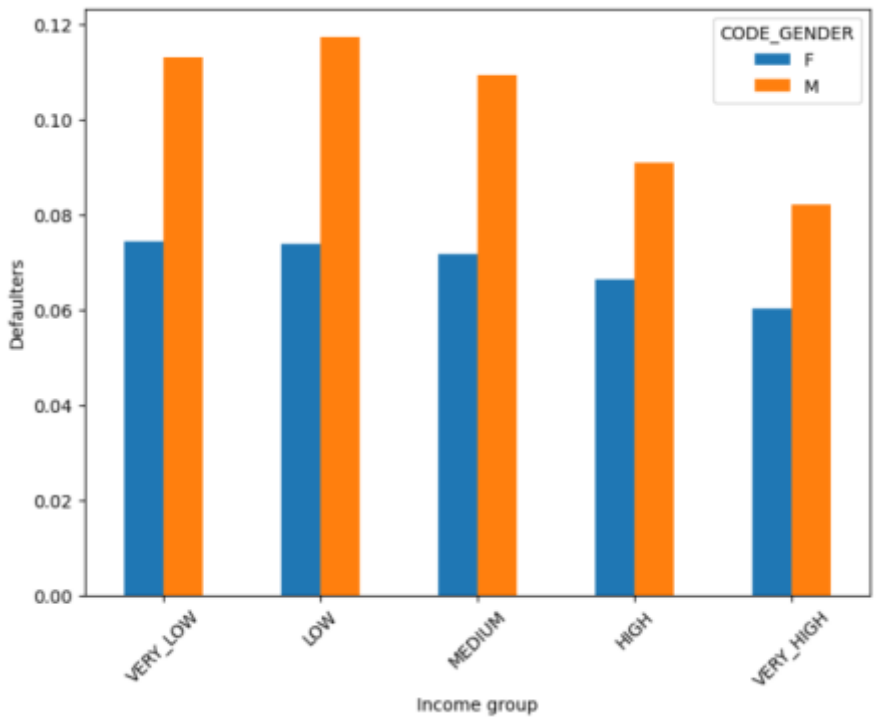
7. Cleaning merged dataframe (handling missing values, standardising, etc)

8. Performing univariate, bivariate and multivariate analysis on merged dataframe

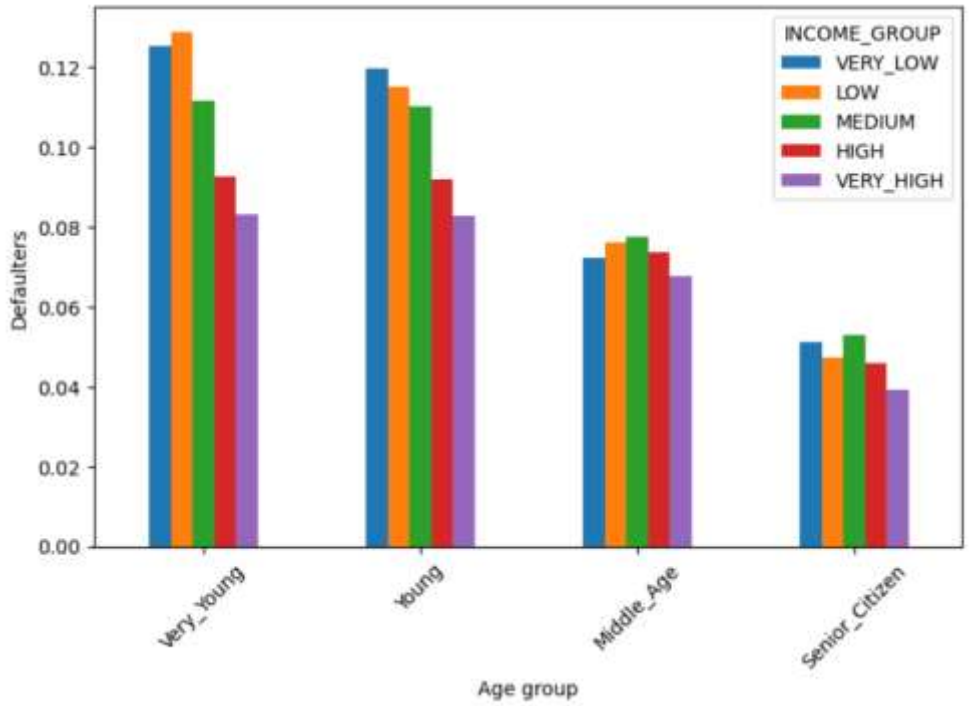
9. Providing insights and recommendations

Current applications

Income groups & gender



Income groups & age groups



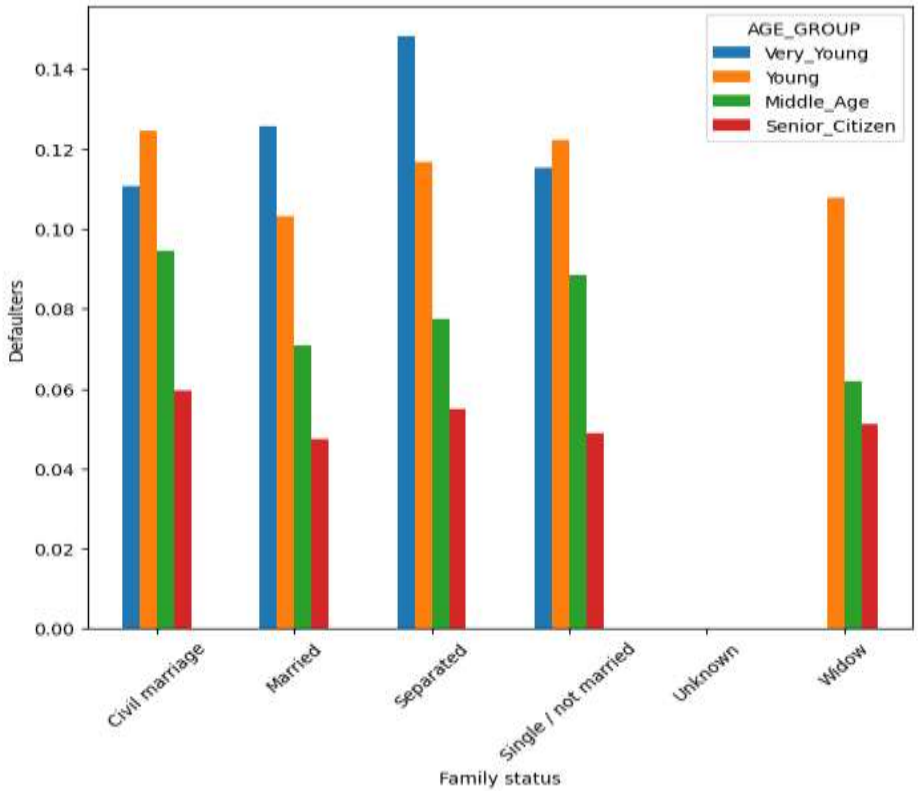
Insights:-

1. High income groups are less defaulter as compared to lower income groups.
2. Mid age and senior citizens with all income groups are less likely to default.
3. Males are higher in default as compared to females

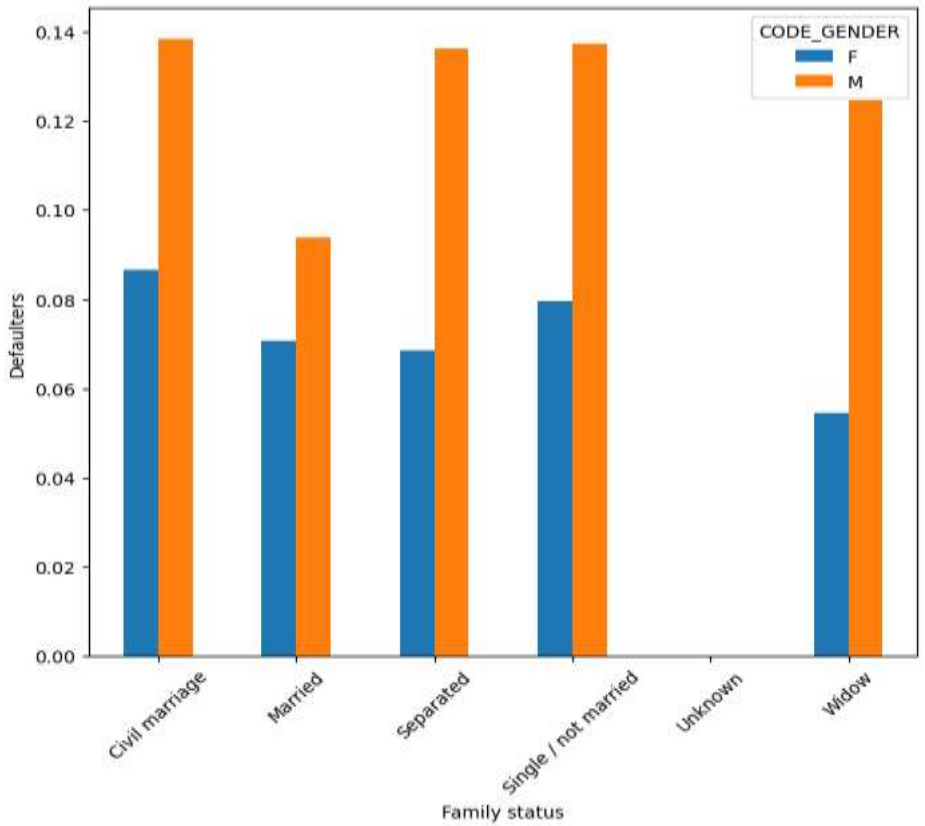
Recommendation:-

1. Safer to grant loan for mid age and senior citizen clients with higher income.
2. Risky to grant loans for very young and young people with low income groups.

Family status & age group



Family status & gender



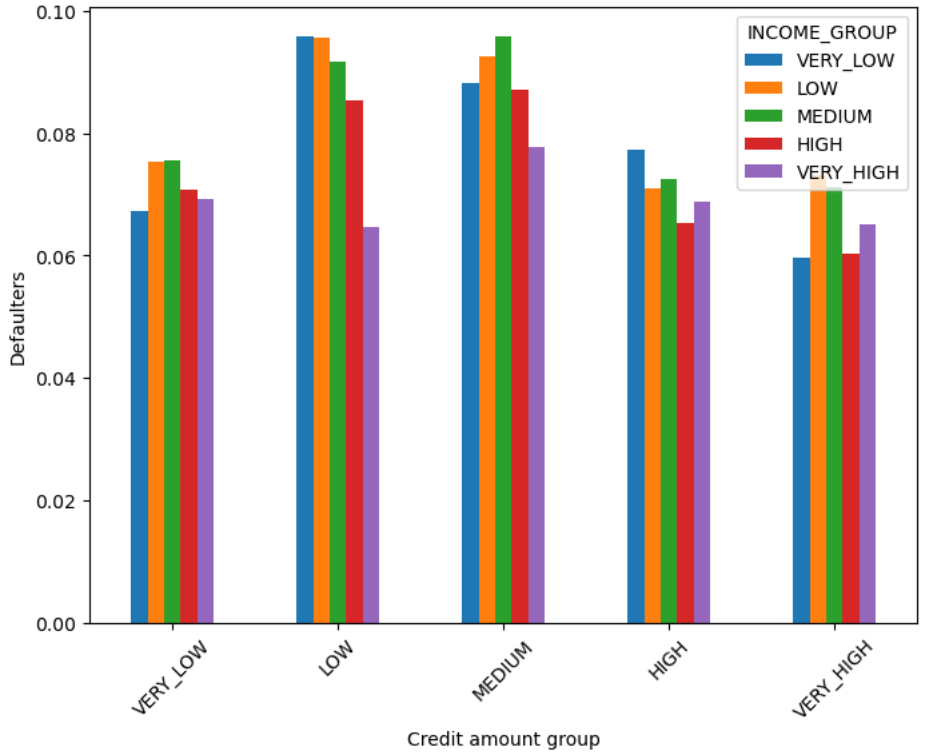
Insights:-

1. Senior people irrespective of family status are less likely to be defaulted.
2. Vey Young and Young people are more likely to be defaulted in all family status.
3. Males are more like to be defaulted than females.

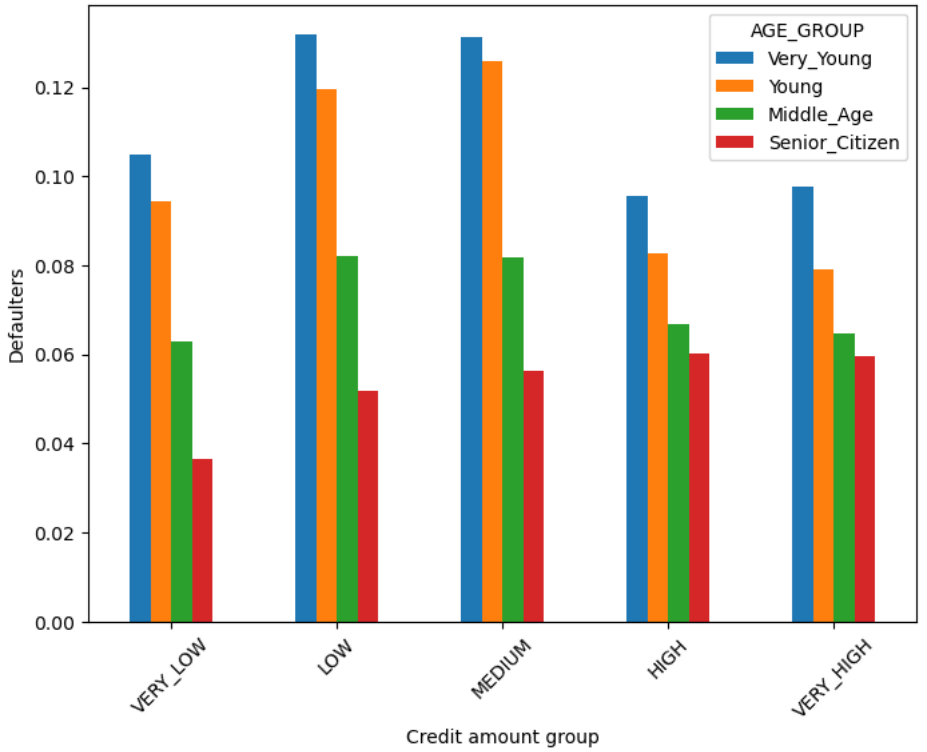
Recommendations:-

1. Safer to grant loan for senior citizen of all family status.
2. It is risky to grant loan for single, separated and civil marriage young men.

Credit amount group & income group



Credit amount group & age group



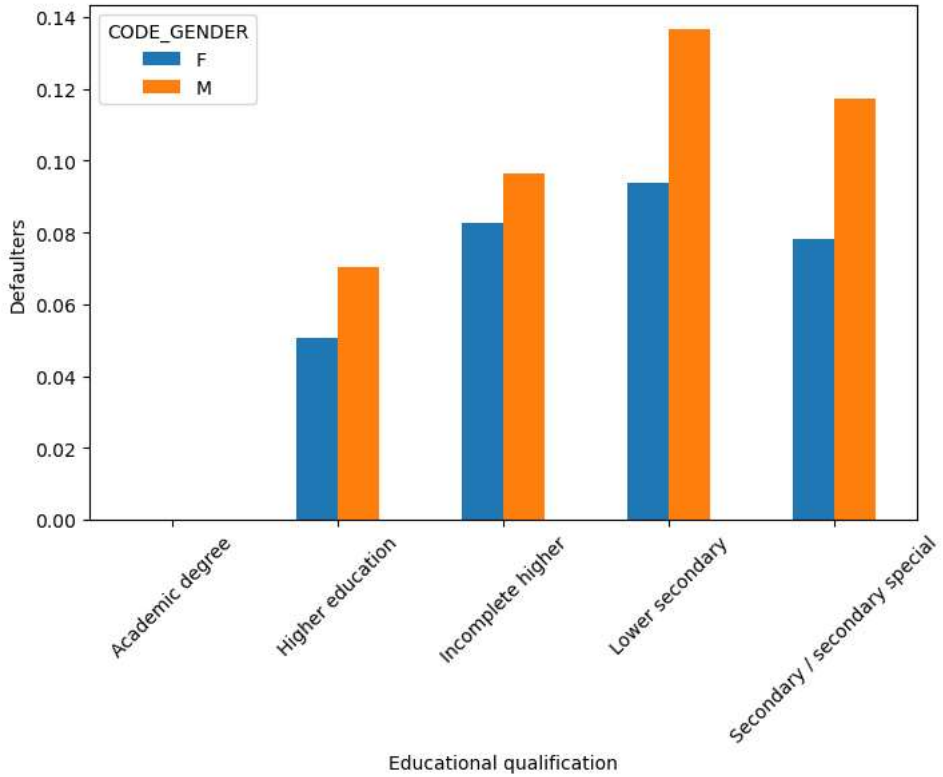
Insights:-

1. Across all income groups clients with medium credited amount are highly defaulted followed by low and high credit amount.
2. Young clients with medium and low credit amount are most likely defaulted.

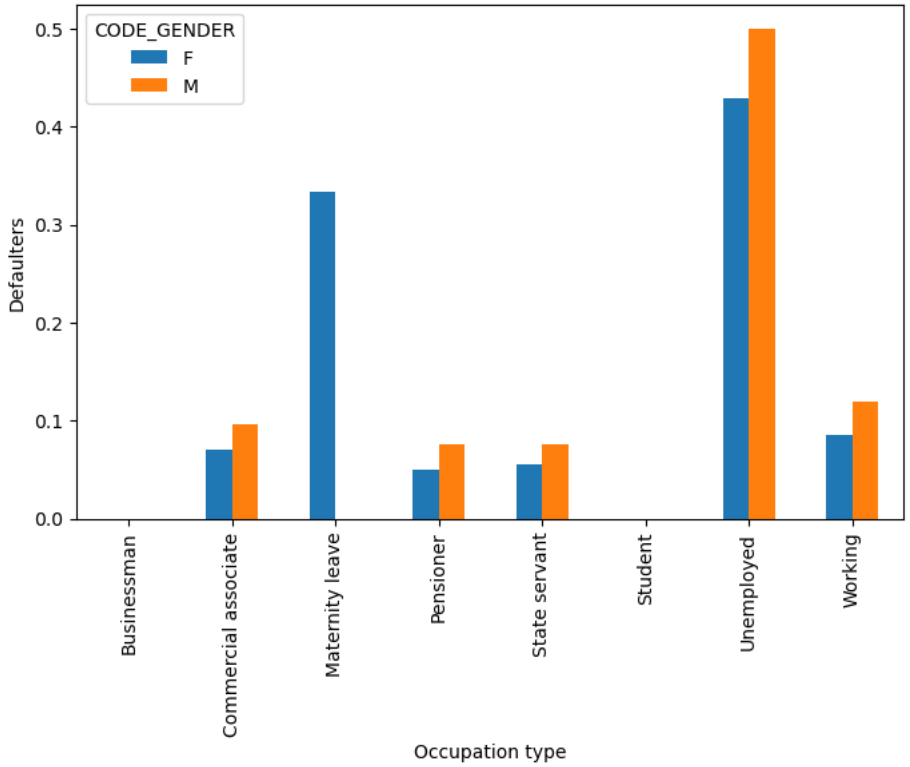
Recommendations:-

1. Recommended to grant slightly higher amount of loan to all income groups.
2. It is very risky to grant medium and low credit amount of loan to young clients.

Educational qualification & gender



Occupation type & gender



Insights:-

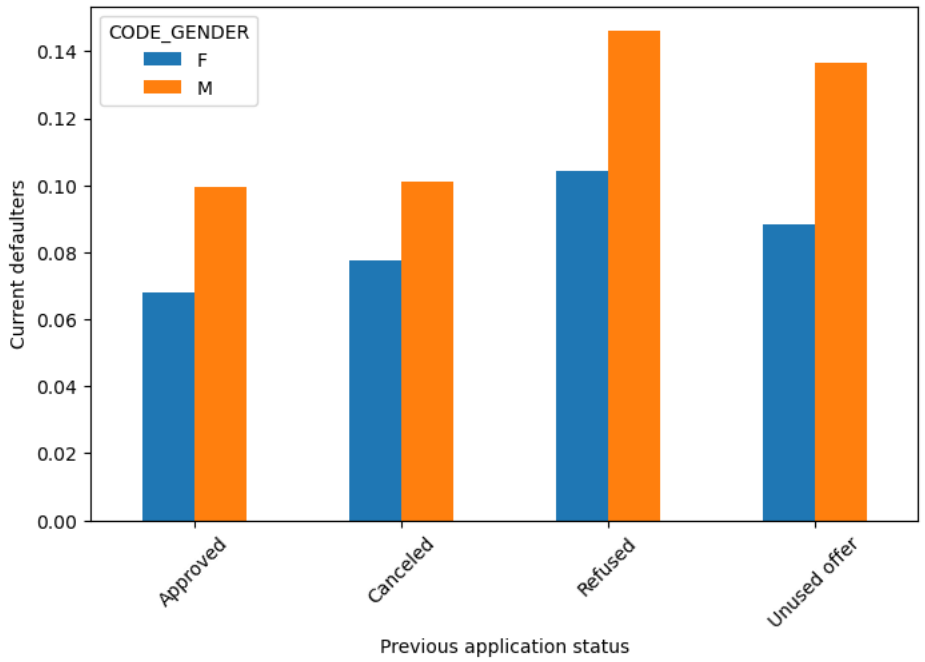
1. Higher educated people are less defaulters and lower secondary educated people are more.
2. Unemployed clients along with clients with maternity leave are high in default.

Recommendations:-

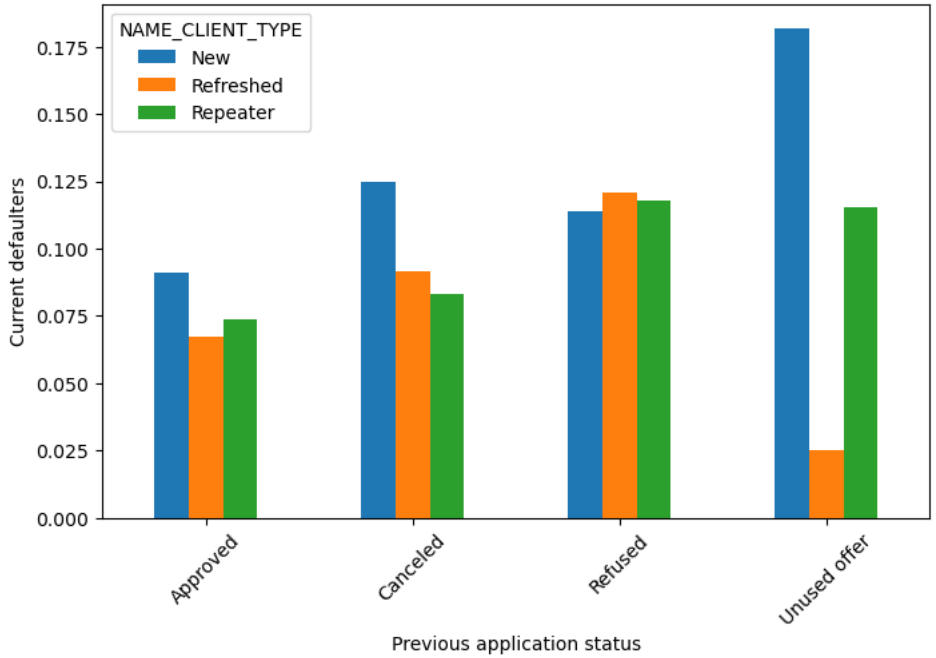
1. Safe to grant loans to higher educated clients across all occupations except unemployed and women on maternity leave.

Previous application status Vs Current application defaulters

Previous application status & gender



Previous application status & client type



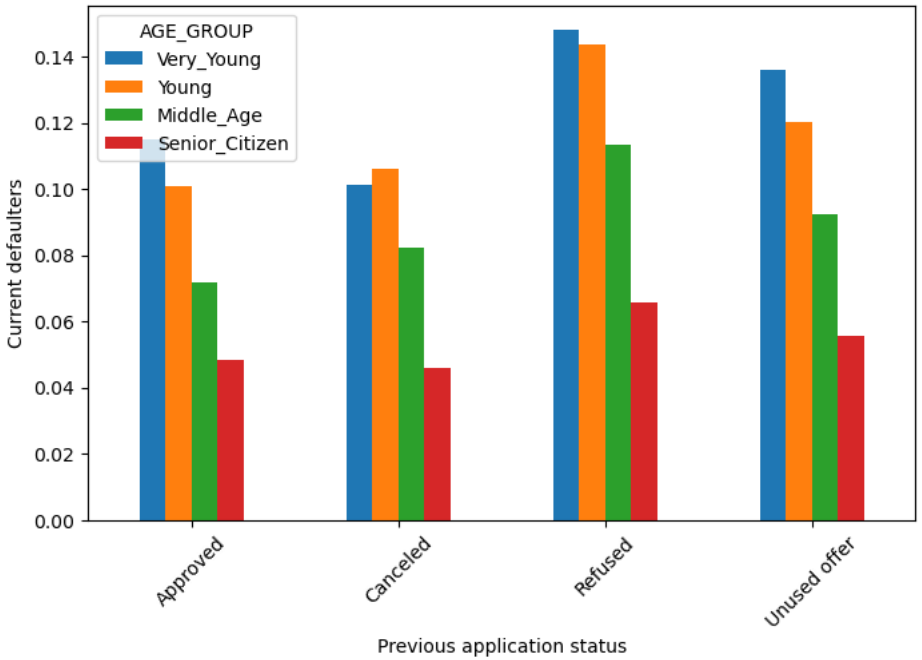
Insights:-

1. Previously refused and unused offer applications were more defaulted in male.
2. New clients with previously unused offer are more defaulted.

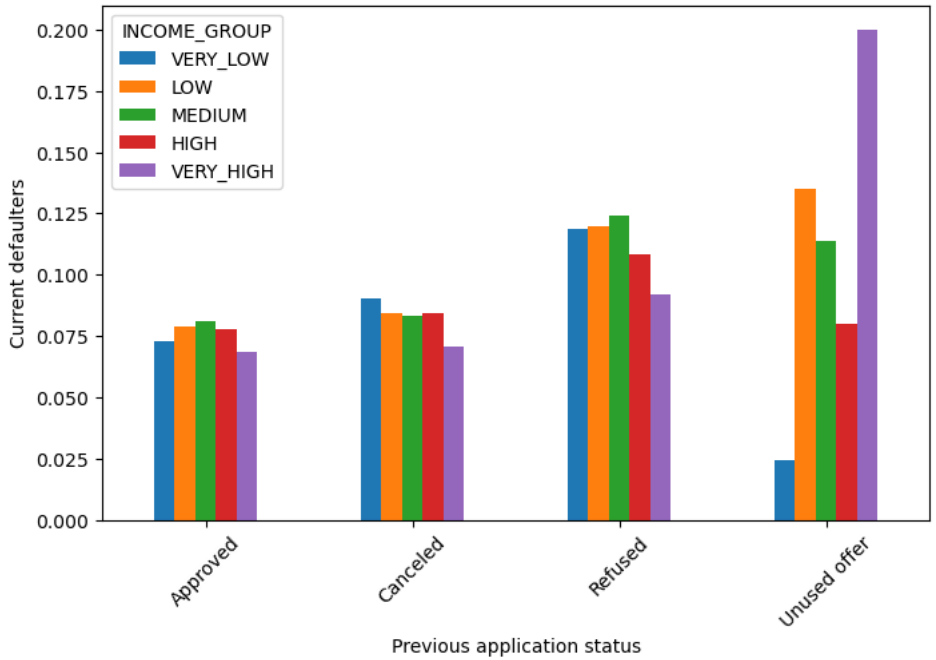
Recommendations:-

1. It is recommended to provide loans to previously approved females.
2. There is a risk to grant loans for clients, whose applications were refused or unused previously.

Age group & previous application status



Income group & previous application status



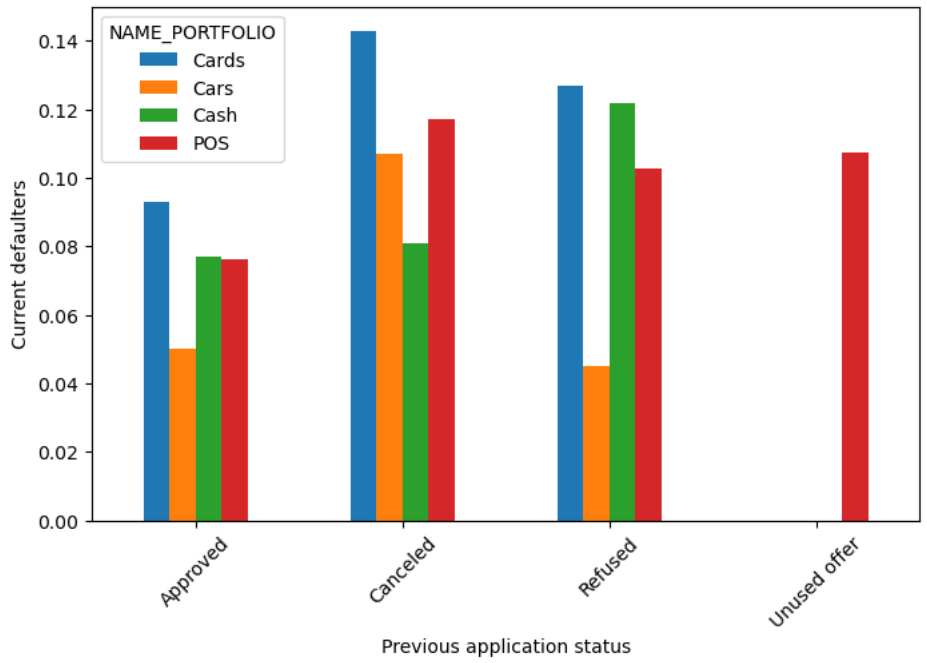
Insights:-

1. Young people, who were previously refused are mostly defaulted.
2. The senior citizens are less defaulted irrespective of their previous application status.
3. In all income groups previously refused applicants are more defaulted.

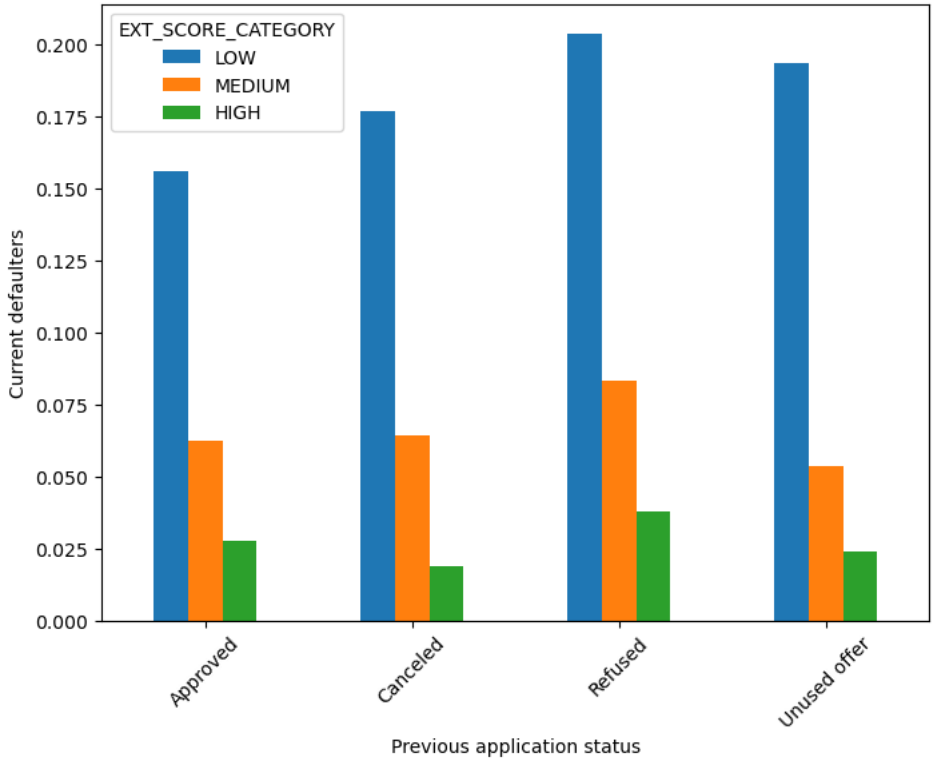
Recommendations:-

1. Safer to grant loans for senior citizen.
2. Lesser risk to grant loans for approved applicants to all income groups.

Portfolio & previous application status



External source score & previous application status



Insights:-

1. The previous applications for portfolio Cards and POS are mostly defaulted .
2. Previously refused applications for Cash are also higher in defaults.
3. Low external scorer are highly defaulted irrespective of their previous application status.

Recommendations:-

1. It is safer to grant loans for any portfolio for previously approved applicants.
2. It is high risk to grant loans for applicants, who have poor external score especially whose applications were previously refused, unused or cancelled.

Insights & Recommendations

Highly Recommended client groups:

1. **Approved clients in previous application.**
2. **Senior citizens in all categories.**
3. **Clients with higher education and high income.**
4. **Client with high External Score.**
5. **Clients with married marital status.**
6. **Female clients as compared to males.**

High Risk client groups:

1. **Clients with previously refused, unused offers or cancelled applications.**
2. **Young clients are more likely to default as compared to middle aged and senior citizens.**
3. **Lower secondary and secondary educated clients.**
4. **Clients with poor external score.**
5. **Clients with marital status - Single, separated, widow and civil partnership**
6. Low income clients with previously refused application status.