Credit EDA Assignment

(Bank Defaulters Data)

-by Devendra Sirsath

Problem Statement

- Whenever a customer applies for loan, bank can either lend or reject it. Some of the clients, to whom bank has given loan, tend to default whereas some of the clients who were capable or eligible for loan but bank had refused to lend them loan because of some reason.
- In both the cases there is business loss for bank as non-potential customer has been given loan and potential customer has been rejected.
- In order to mitigate the above situation, bank needs some strong parameters or factors which would help in taking the crucial decision of approving or rejecting loan.
- To identify these parameters or factors we are solving this assignment so that bank can take better decisions.

Problem Statement

- In this Assignment, we have to find patterns which would clearly indicate that the client has difficulties in re-payment so that same knowledge could be used to predict the possibility of defaulting for the future clients.
- This is Known as Risk Analysis and currently majority of financial institutions as well as other sectors are using it.

Assumptions

- “XNA”, ”XAP” are considered as Null values.
- In few cases, where “XNA”, ”XAP” value count is very high, they are considered as complete different category.
- In Application data set, “XNA” value in “ORGANIZATION_TYPE” is replaced by “Pensioner“ because of systematic pattern in the missing values.

Approach

- Steps:
  1) Understanding the data –

     imported the data and understood the variables with the help of column description data.

  2) Basic Sanity checks –

     using info, shape, description commands to check the if the data is practically correct or not. Eg. If any column related to time/day is negative then it is to be corrected to positive, values exceeding the maximum permissible limit should be handled.

Approach

3) Identifying and Imputing/Removing missing values OR removing entire column –

checked the missing values with the help of “isna”, ”isnull” & sum or mean commands.

- If more than 40% data is missing from a variable then dropped that variable
- If less than 5% data is missing then the rows can be dropped but decided to ignore that values as data was quite large
- For missing data between 5% to 40%, if variable is categorical then imputed the data with mode and for numerical variable imputed with appropriate measure of central tendency. Note that we can even ignore the missing data in this category instead of imputing with the incorrect one.

4) Dropping unnecessary/irrelevant variables -

After observing the dataset and understanding the variables, we can drop the variables which does not have much relevance with the objective of case study. For this we can use our domain knowledge along with data visualization.

Approach

5) Identifying & Handling Outliers –

Extreme values in a variable are called as outliers and it can be roughly observed with the help of describe command and further confirmed by plotting distribution plot or box plot.

Some outliers are due to data error (e.g. employment years = 1000) and some are practical values. Depending on the variable and outliers we can either impute or remove the outliers, bin them, cap the data to certain percentile.

6) Standardizing Values –

Using info command we can observe the datatypes. Ensure that numerical columns have int or float datatype and categorical columns have categorical datatype. Changed the datatype where ever required.

Approach

7) Checking Data Imbalance –

Data imbalance is checking the percentage positive & negative outcome of the target variable in this case defaulters-1 & repayers-0.

8) Univariate, Categorical/Numerical Univariate Analysis –

Understanding each variable separately and wrt some category.

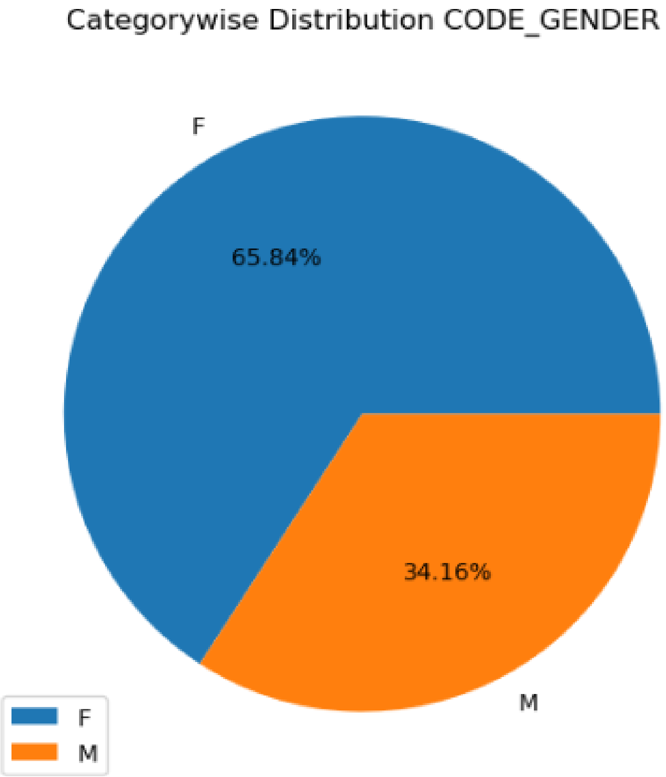
In this case, we found out the count of customers in a specific category wrt TARGET variable. We used pie chart, count plot, bar plot, distribution plot etc.

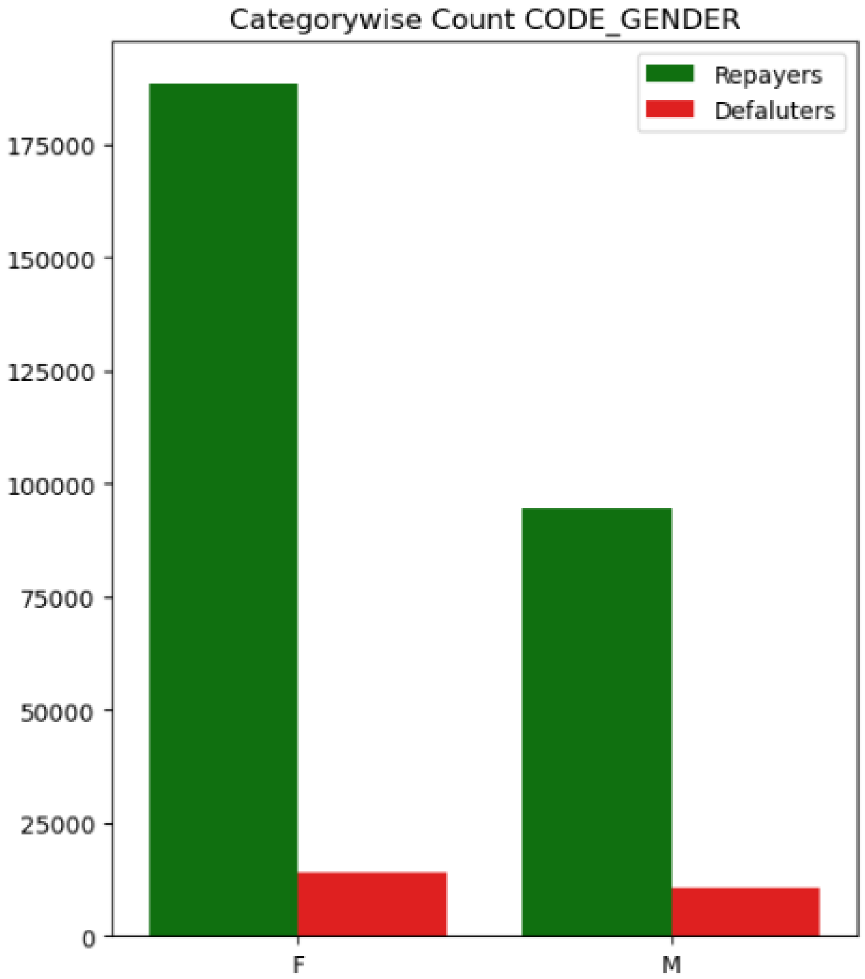
9) Bivariate & Multivariate Analysis –

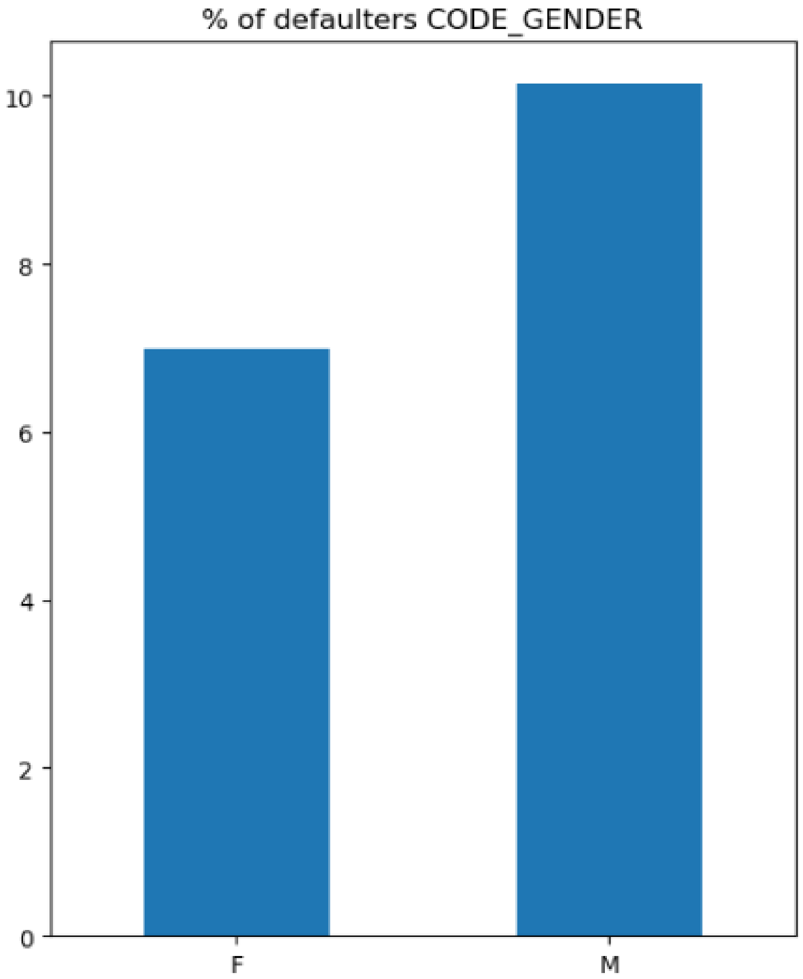
Analyzing two or more variable at a time wrt target variable. The variables can be categorical or numerical or mix of both.

We used pairplot, boxplot, heatmaps, factorplot, bar plots etc.

Results of Analysis

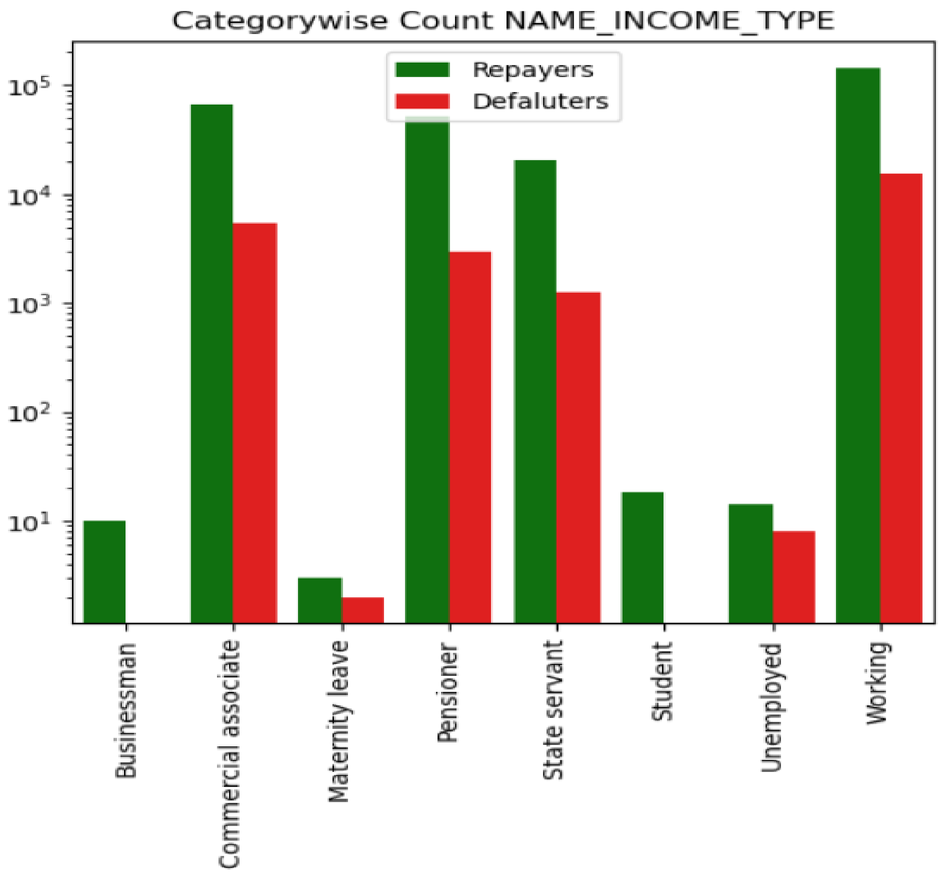

Categorywise Distribution CODE_GENDER

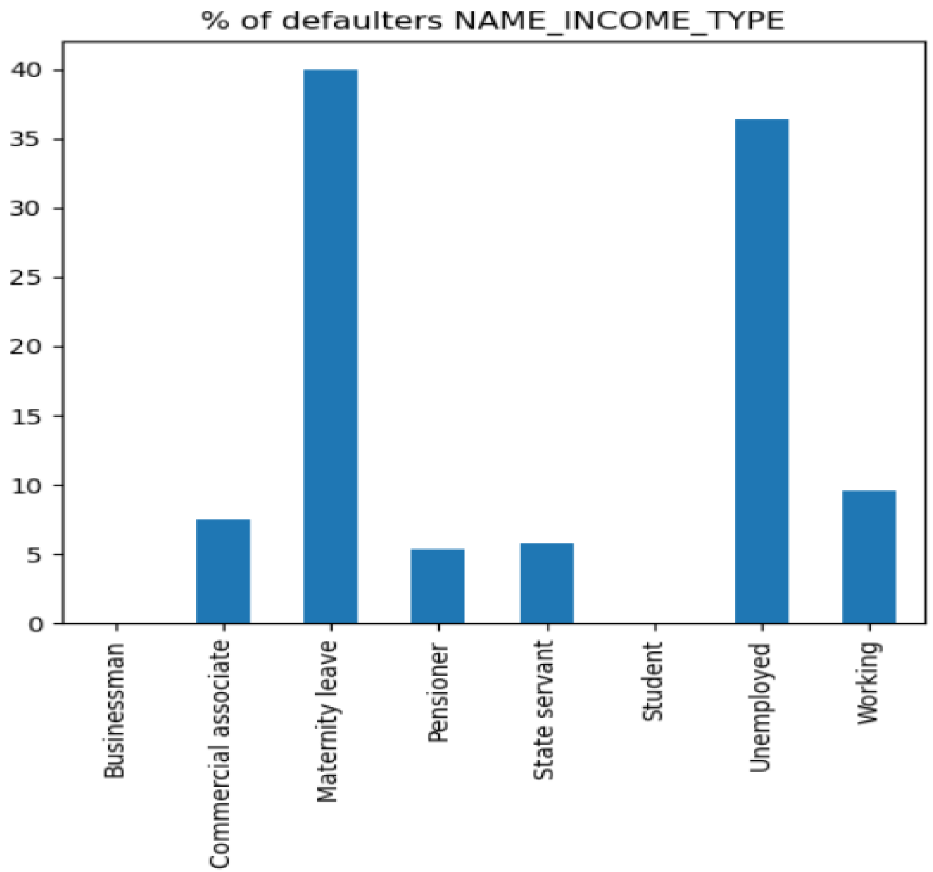

Categorywise Count CODE_GENDER


% of defaulters CODE_GENDER

The number of female customers is almost double than number of male customers still the chances of male defaulting on a loan is higher than that of female.

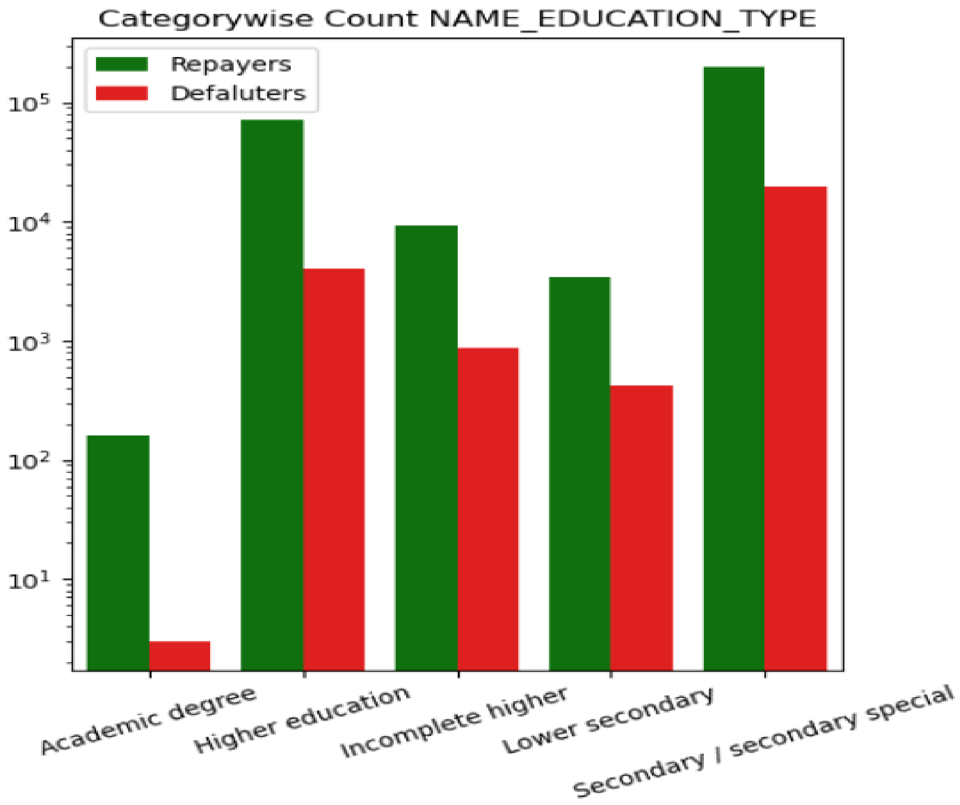
Results of Analysis

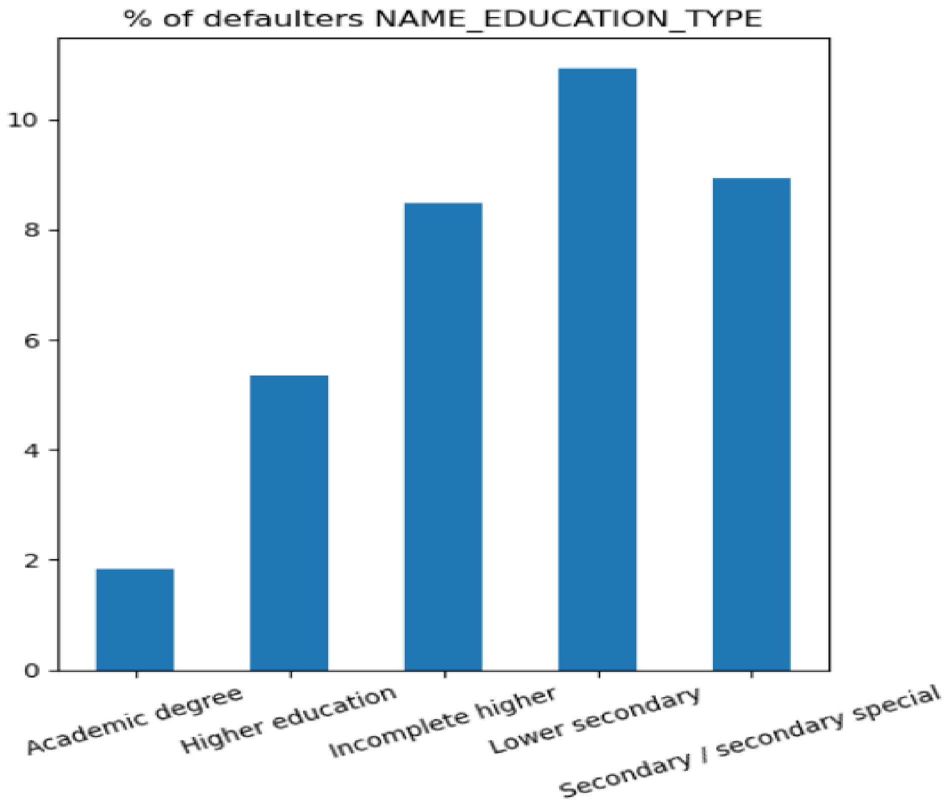




- It is least risky for the bank to give loan to businessman and students(0 default rate).
- Commercial associates, pensioner, state servant & working category fall under average rate of defaulting.
- the customers who are on Maternity leave have the highest possibility of defaulting. This is obvious because the expenses greatly increases in maternity period. Although, they have very low count as compared to other categories
- The people who are unemployed have the second highest default rate

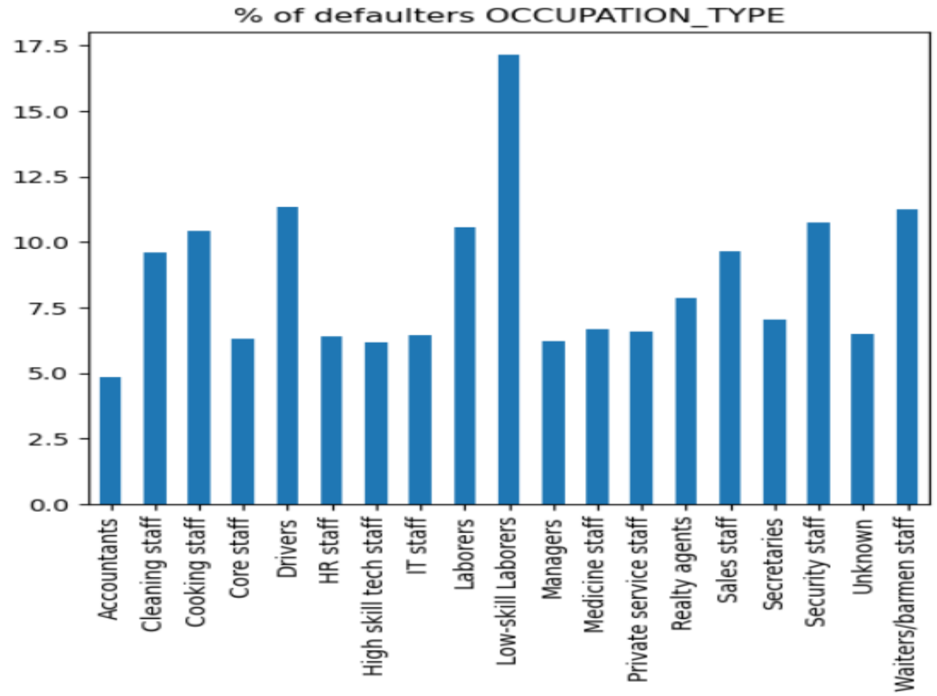
Results of Analysis

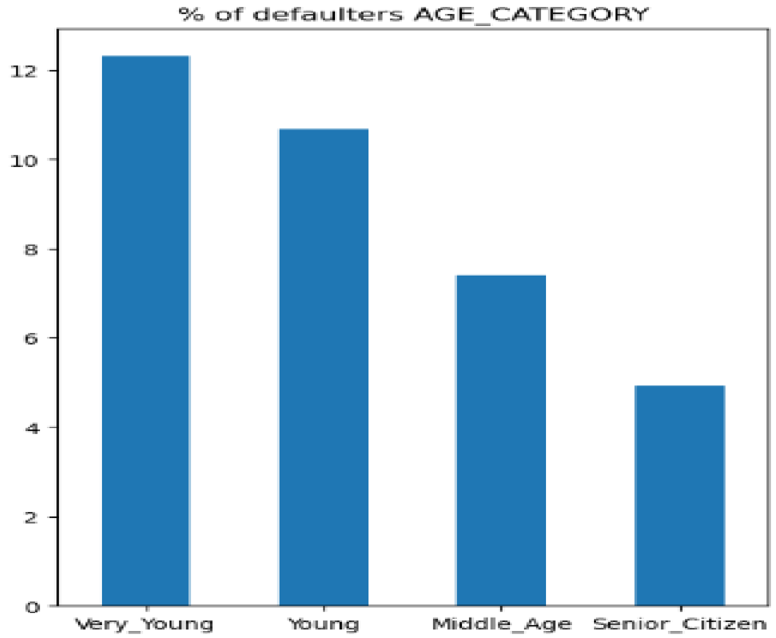




- Customers with academic degree have lowest rate of defaulting but the count is quite less as comparted to other categories.
- Lower secondary category has highest default rate (around 11%).
- secondary/secondary special have second highest default rate. they also contribute to the highest count.
- To reduce the default rate, bank should focus on Incomplete higher, Lower secondary & secondary/secondary special category customers as they have defaulting rate above avg(approx 8 %)

Results of Analysis



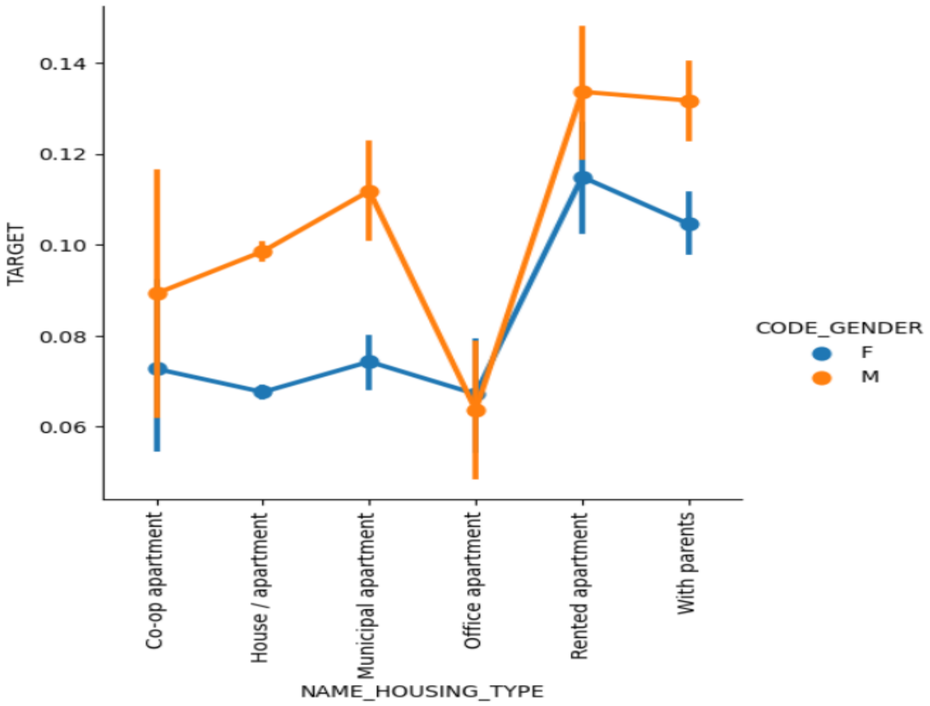


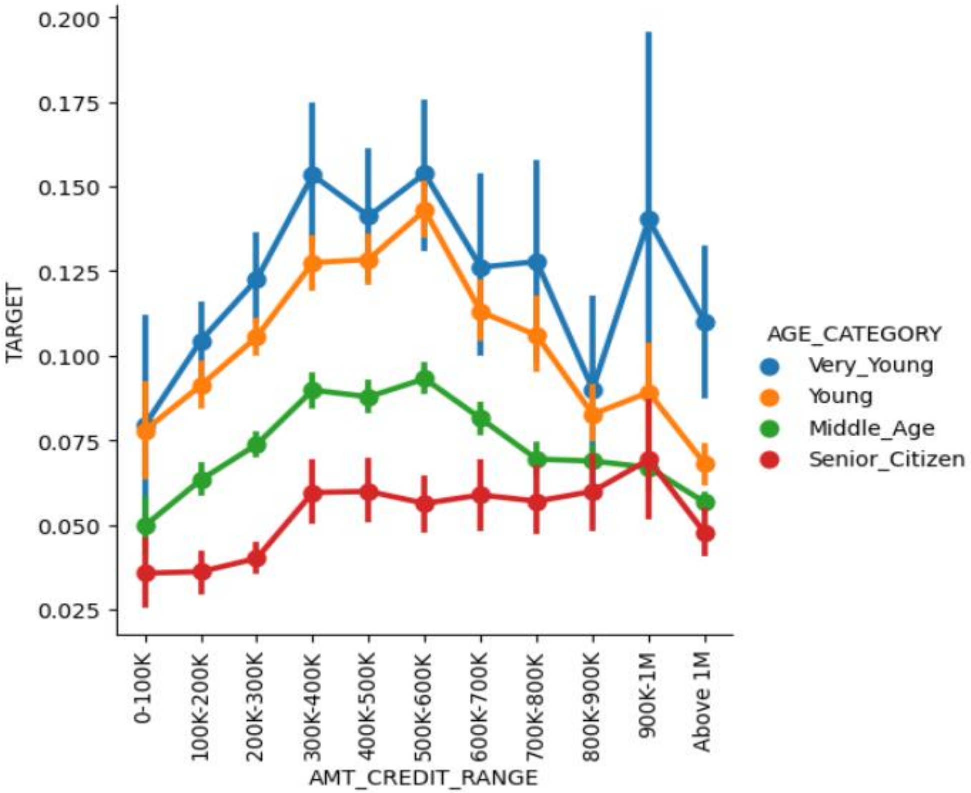
•Low-skilled Laborers have the highest defaulting rate (approx 17%) Accountants have lowest defaulting rate(Approx 5%).

•Laborers contribute to highest count of customers with approx 10% defaulting rate

- Almost 61% customers belong to middle age category(35-60 years) who have avg defaulters rate approx 8 %.
- Although very young customers(20-25 years) does not contribute much to the total count still they have the highest possibility of defaulting (around 12%)
- Senior citizens are have lowest defaulting rate approx 5%.

Results of Analysis

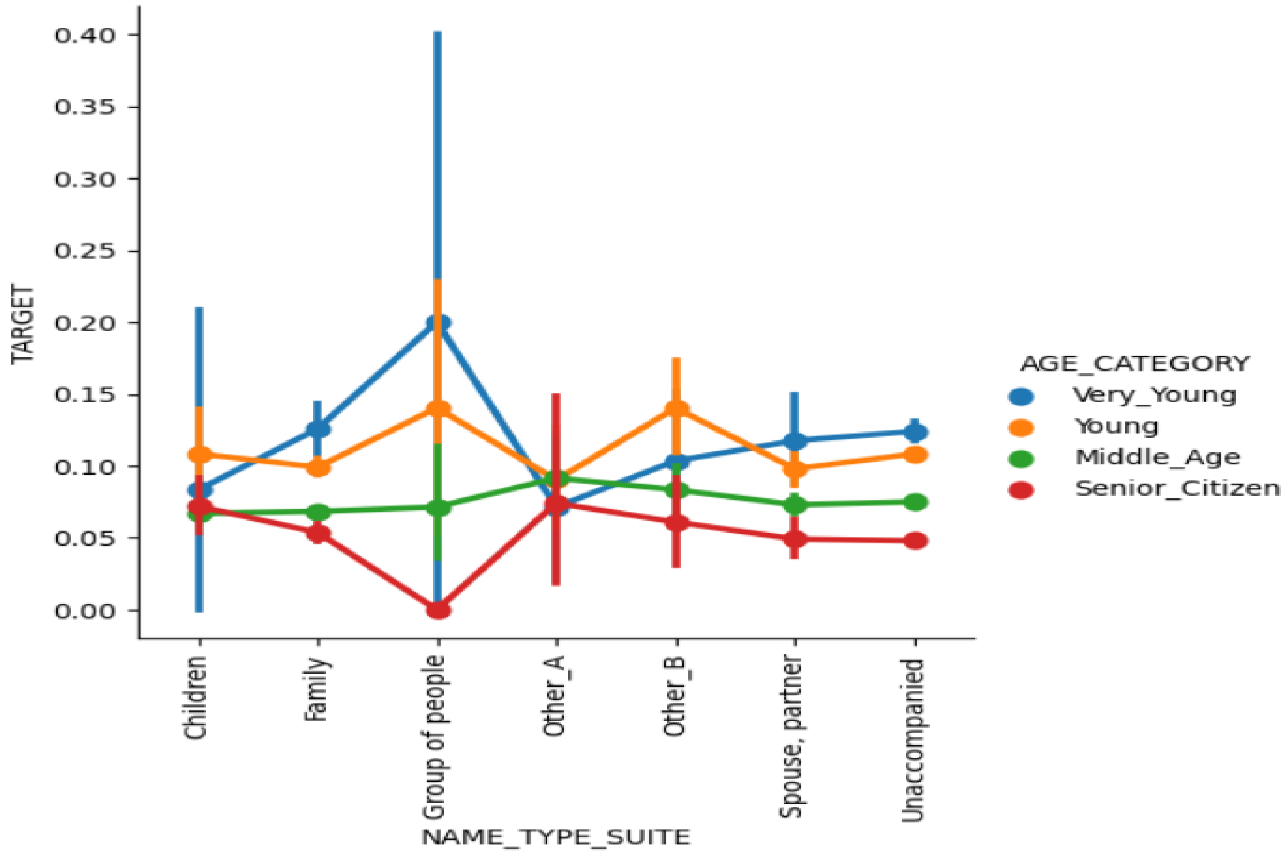




Only in Office apartment category males are less likely to default than females.

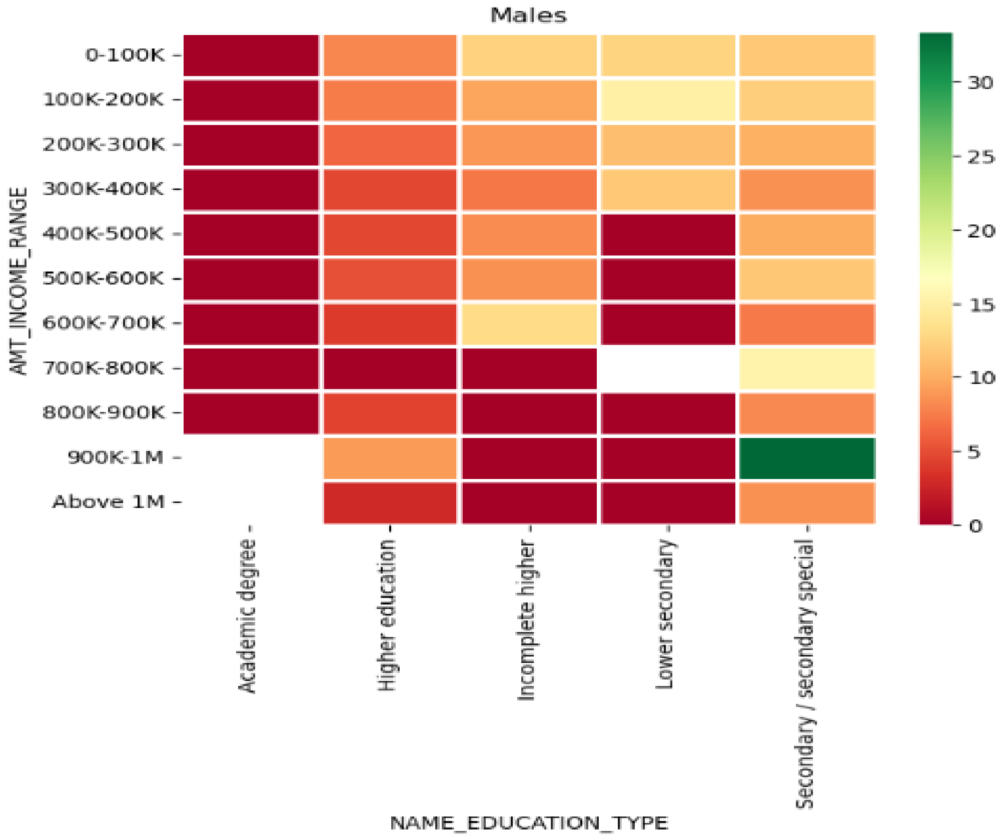
Although Very young customers have higher default rate than other age group customers they are least likely to default in below 100K loan amount.

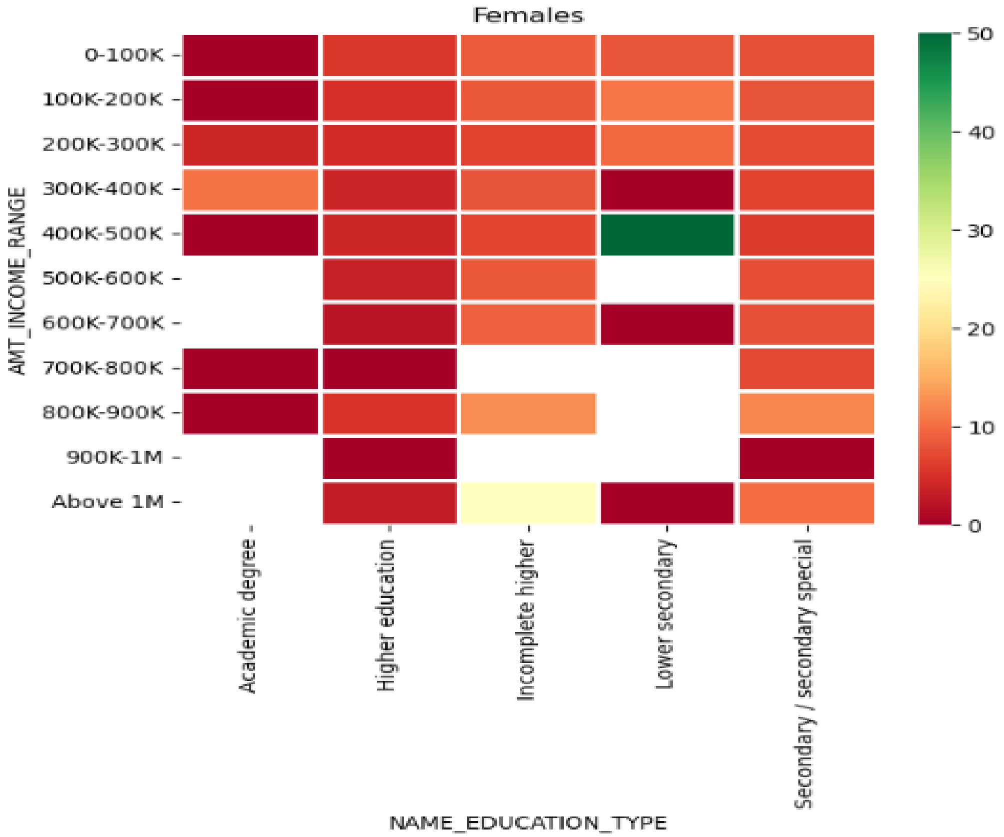
Results of Analysis



Senior citizen(above 60 years) who come with group of people are least likely to default(0%) and exactly opposite is the case with Very Young customers(~20%).

Results of Analysis

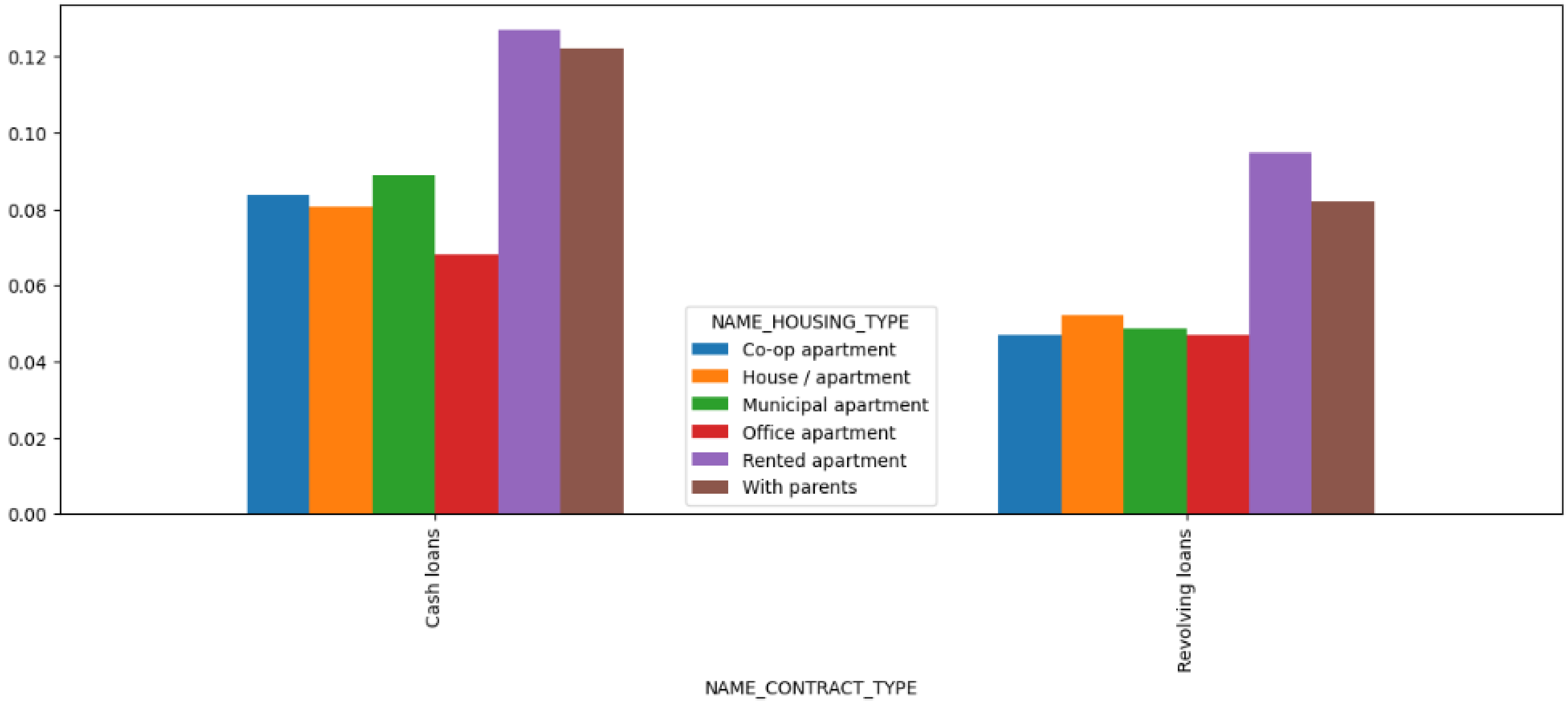




Note:

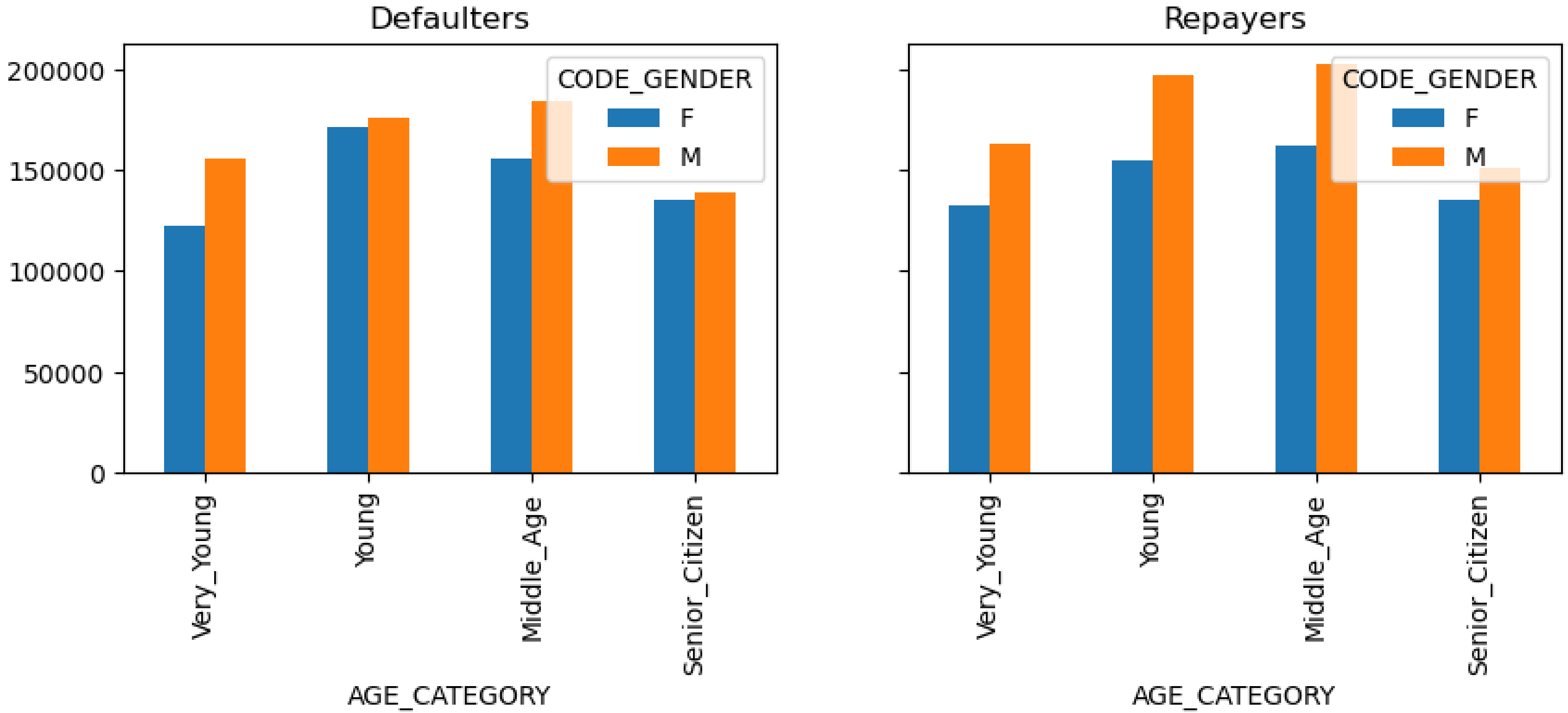
- We can say that Males with Salary range between 900K-1M with Secondary/Secaondary special education have highest possibility of defaulting
- Also Females with salary range between 400-500K with Lower secondary education have highest possibility of defaulting

Results of Analysis



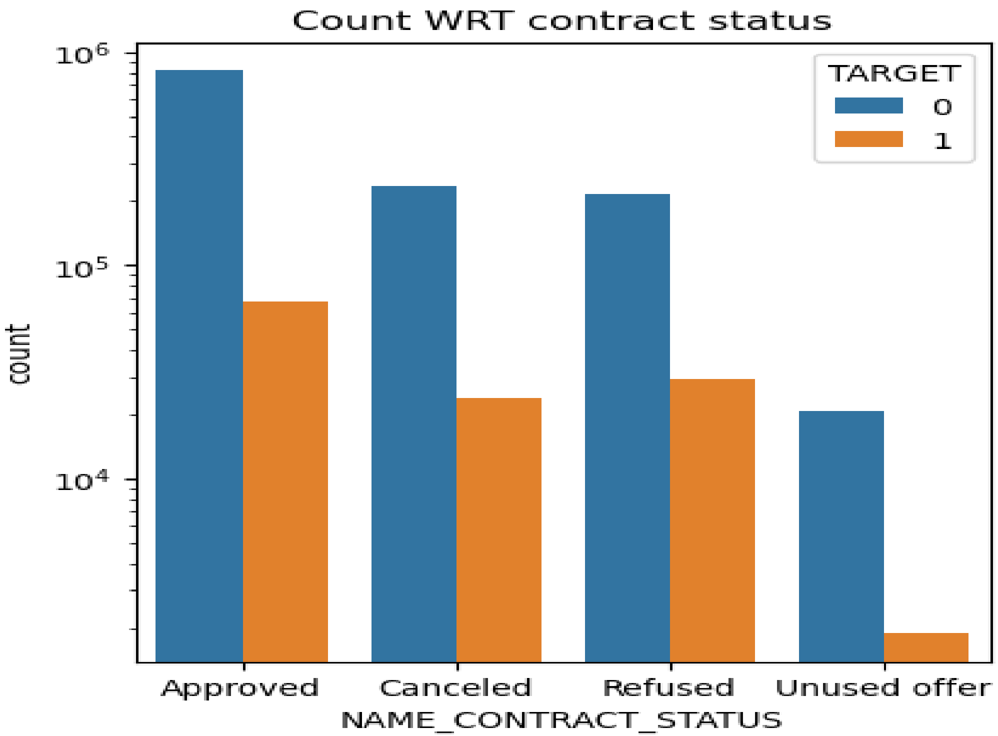
Excluding the office apartment customers, the customers who possess house/apartment have lesser chances of defaulting in cash loans than revolving loans as compared to other housing types

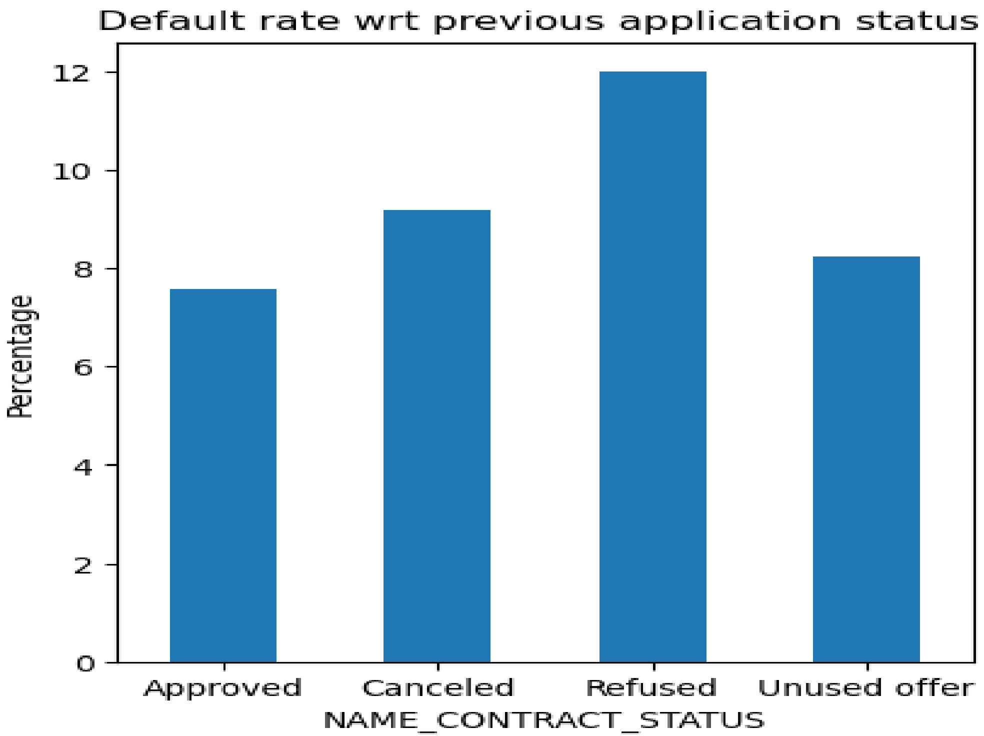
Results of Analysis



- In both the cases, default and repayment, males have higher mean salary than females, still we had seen that the possibility of default is higher in male category.
- Also it is interesting to see that females in Young category(25 to 40 years) who are defaulting have higher mean salary than females in Young category who are repaying.
- In male category, the repayers have slightly higher mean salary than that of defaulters.

Results of Analysis(Merged Dataset)



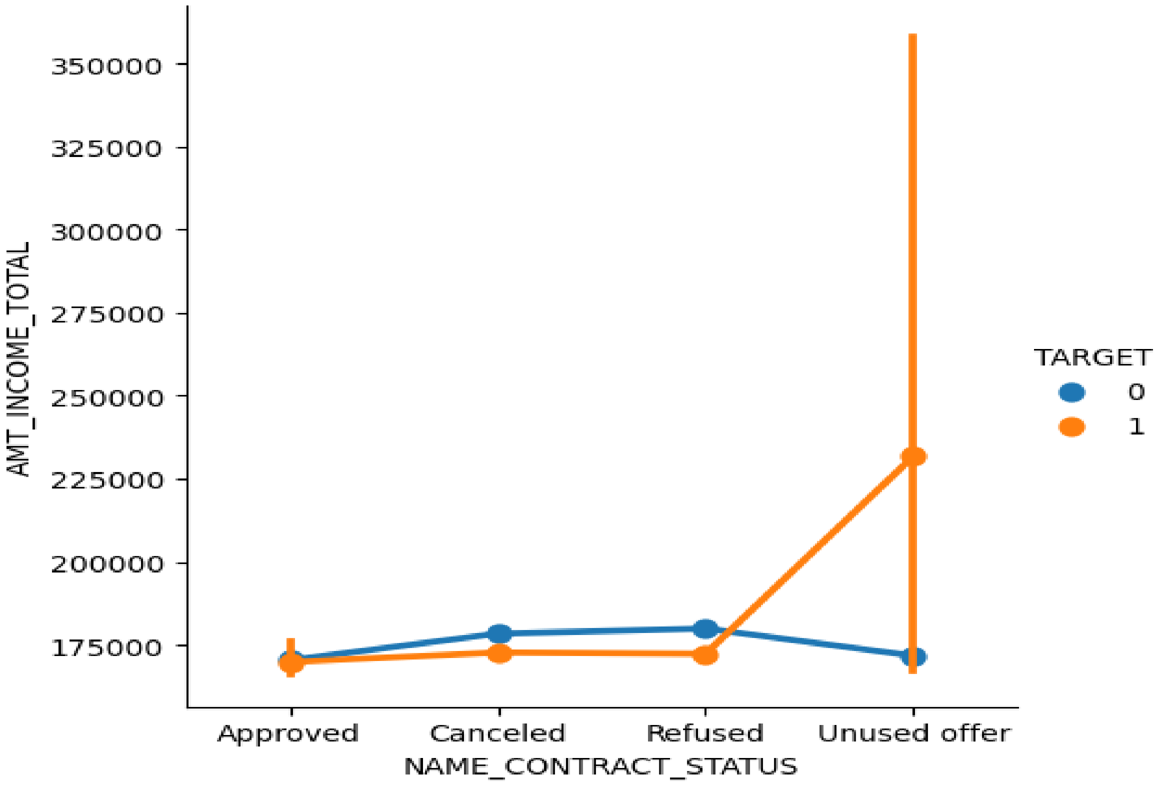


•From previously approved loan, it is evident that the customers are least likely to default i.e ~7.5%. Although this category have highest number of customers.

•The customers whose loan was previously refused have defaulted the most in currently approved loan(~12%) but on the brighter side, 88% are repayers. If we decide to reject on the basis of previous status will lead to 88% business opportunity loss.

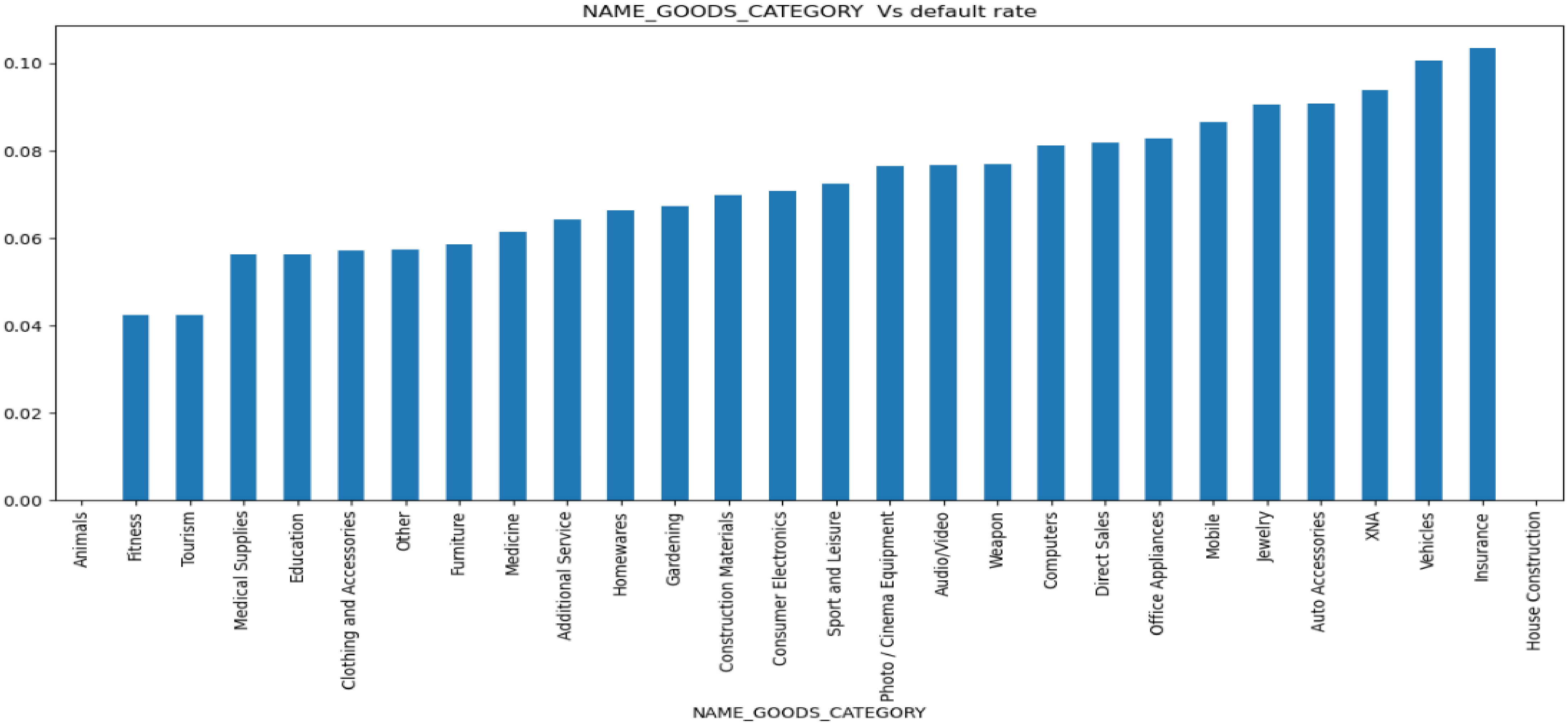
•The customers who were previously eligible for loan but did not take it have around 8% default rate which is fine. But if we want to further reduce the default rate we should always recalculate the eligibility considering the current situation instead of directly passing on the offer in this case.

Results of Analysis(Merged Dataset)



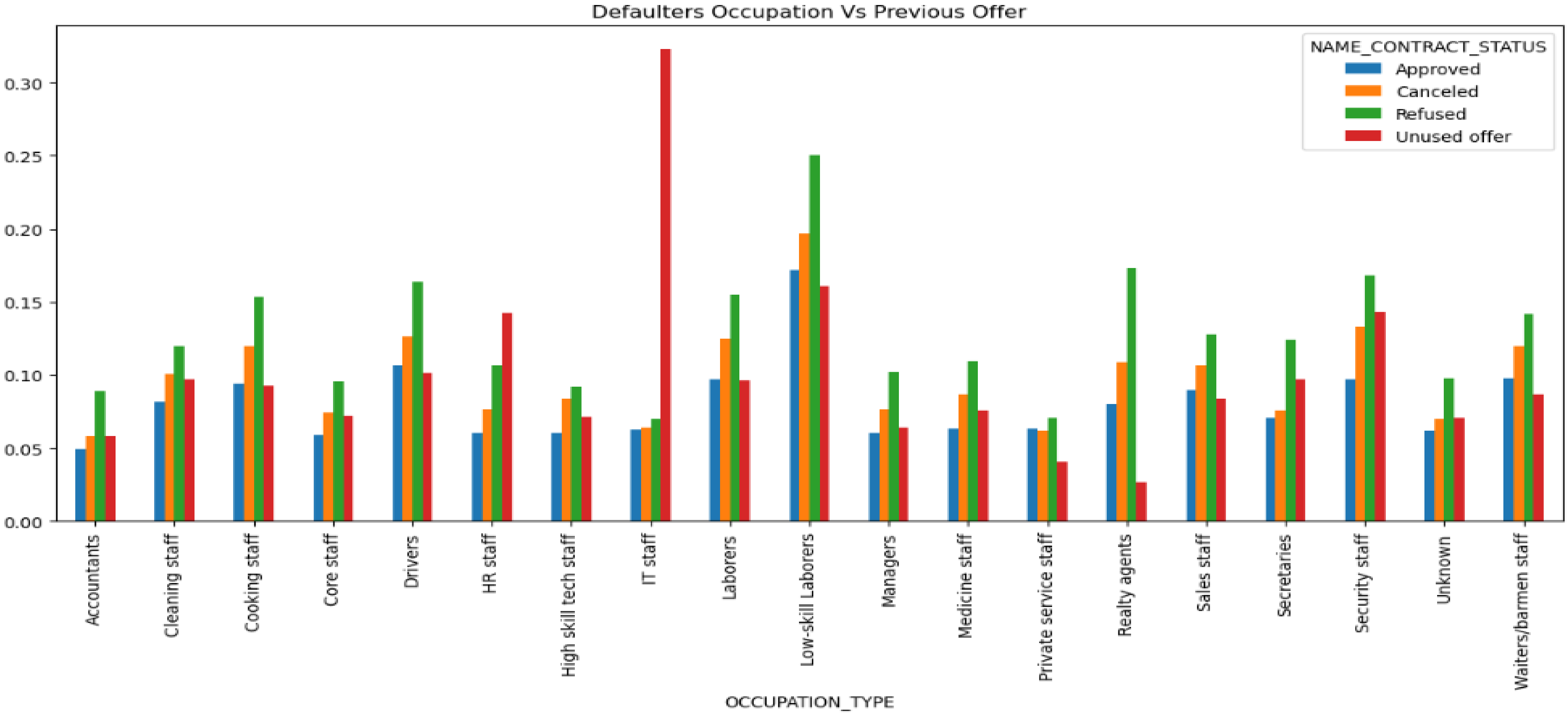
The customers who had loan offer but did not take it previously and currently defaulted the loan have higher avg salary than that of repayers. There is definitely some catch in this case, there is a possibility that bank has offered them the same previous loan offer amount without considering the current situation.

Results of Analysis(Merged Dataset)



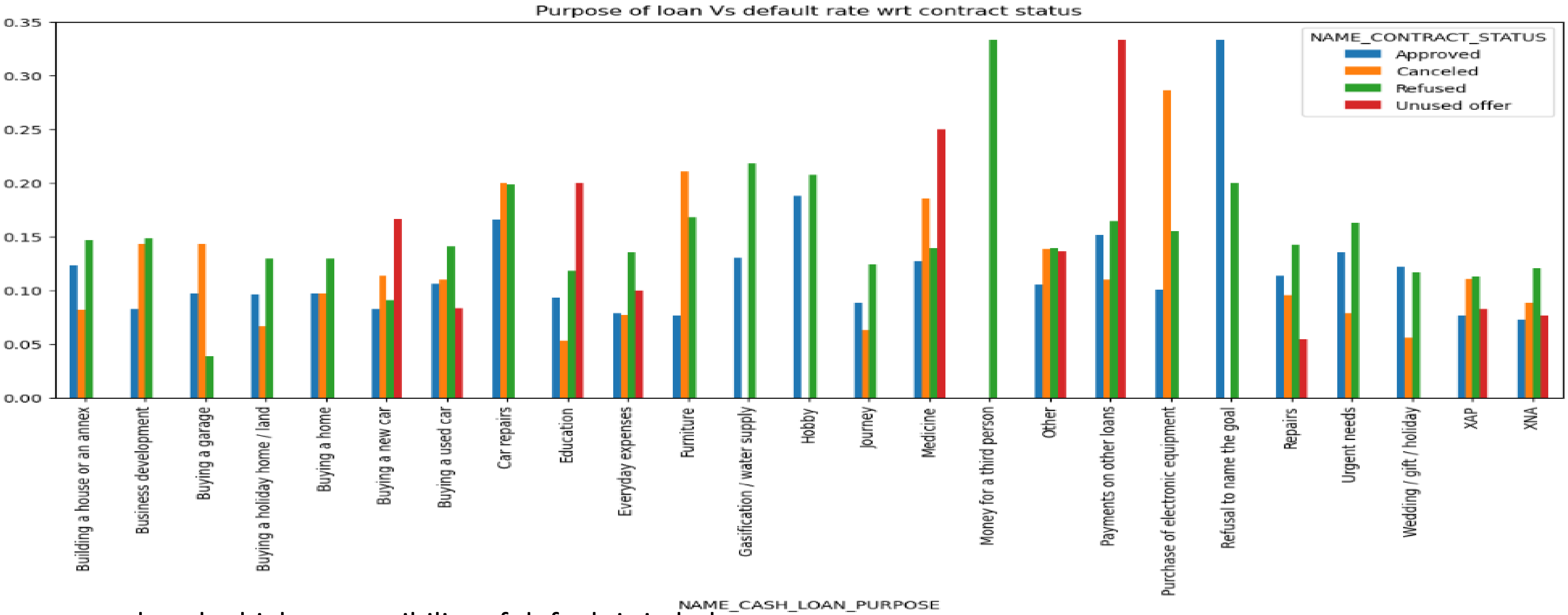
- Customers who have taken loan for fitness or Tourism are least likely to default around 4%
- Customers who have taken loan for Vehicle or Insurance are most likely to default around 10%

Results of Analysis(Merged Dataset)



- IT staff have the highest possibility of defaulting if they have taken unused offer loan which is more than 30%.

Results of Analysis(Merged Dataset)



- We can see that the highest possibility of default is in below purpose:

1) Money for a third person previously refused loan
2) Payment on the other loans previously given offer
3) Refusal to name the goal previously approved loan.
4) Purchase of Electronic equipment previously cancelled loan.

- We can clearly see that first 3 are that kind of cases where loan offer should be rejected because the default rate is almost 34%.
- Even in fourth case, we should further drill down to the reason of cancelling the loan and then take decision accordingly.

Conclusion

- <u>Factors which indicates if customer will default or not are:</u>

1. NAME_FAMILY_STATUS : Single people or who have civil marriage have higher default rate.
2. NAME_INCOME_TYPE: Customers who are on Maternity leave or unemployed have higher default rate.
3. NAME_EDUCATION_TYPE: Customers with Lower Secondary & Secondary education have higher default rate.
4. OCCUPATION_TYPE: Low-skill Laborers, Cooking staff , Drivers, Laborers, Security staff, Waiters/barmen staff are more likely to default (default rate more than 10%)
5. ORGANIZATION_TYPE: People from Transport: type 3 , Industry: type 13, Industry: type 8 have higher default rate (more than 12%) be approved for loan or provide loan with higher interest rate to mitigate the risk of defaulting.
6. DAYS_BIRTH: Very young customers (age group of 20-25) have higher probability of defaulting
7. DAYS_EMPLOYED: Customers with lesser work experience have high default rate.
8. CNT_CHILDREN & CNT_FAM_MEMBERS: More number of family members(similar to more number of children) are more likely to default
9. AMT_GOODS_PRICE: When the credit amount goes beyond 3lakhs, there is an increase in defaulters.
10. NAME_GOODS_CATEGORY: Customers who have taken loan for Vehicle or Insurance are most likely to default around 10%
11. CASH_LOAN_PURPOSE : Loan taken for buying a garage is most likely to be repaid

Conclusion

- Factors which indicates if customer will repay or not are:

1. CODE_GENDER: Female customers have lesser default rate than male customers (almost 3% lesser)
2. NAME_EDUCATION_TYPE: Customers with Academic degree are less likely to defaults.
3. NAME_INCOME_TYPE: Student and Businessmen have no defaults.
4. ORGANIZATION_TYPE: Customers with Trade Type 4, Industry Type:12 and Transport : Type 1 have defaulted less than 5%
5. OCCUPATION_TYPE: Accountants are less likely to default
6. DAYS_BIRTH: People above age of 60 have less probability of defaulting
7. NAME_HOUSING_TYPE : Customers with office apartment are least likely to default
8. DAYS_EMPLOYED: Customers with 25+ year experience(Expert category) are least likely to default
9. AMT_INCOME_TOTAL: Customers with Income more than 700,000 are less likely to default
10. AMT_CREDIT_RANGE : Customers with less than 100K or more than 1M loan amount are less likely to default
11. CNT_FAM_MEMBERS : Customer with 4 or less family members are less likely to default.
12. NAME_GOODS_CATEGORY: Customers who have taken loan for fitness or Tourism are least likely to default around 4%.
13. CASH_LOAN_PURPOSE : The customers who refuse to name the purpose of loan are most likely to default.

Recommendation

- Its better to take extra precaution while approving loans to people with occupation Low-skill Laborers, Cooking staff , Drivers, Laborers, Security staff, Waiters/barmen staff
- People from organization type Transport: type 3 , Industry: type 13, Industry: type 8 have higher default rate (more than 12%). Need to consider more factor while approving loan or provide loan with higher interest rate to mitigate the risk of defaulting.
- In each given contract status, below purpose of loans should be avoided or given with higher interest rate
  - Refused loan- Money for a third person
  - Unused offer- Payment on the other loans
  - Approved loan- Refusal to name the goal previously.
  - Cancelled loan - Purchase of Electronic equipment.

- If any customer had unused offer, then instead of lending the loan with the same offer, some reworking to be done in the loan offer after a certain period. IT staff is 35% likely to default if they opt for unused offer loan.
- It’s better to either avoid lending loan or lend loan at higher interest rate for below categories:
  - Males with Salary range between 900K-1M with Secondary/Secondary special education(~30% default)
  - Females with salary range between 400-500K with Lower secondary education(~50% default)