EXPLORATORY DATA ANALYSIS

Credit EDA Assignment – Module 9

PROBLEM STATEMENT

Using data to minimize the risk of losing money while lending to customers by understanding the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default and ensuring that the consumers capable of repaying the loan are not rejected.

ASSUMPTIONS

- Columns with missing values greater than 40% of missing data has been dropped in case of APPLICATION DATASET and columns with more than 35% of the missing data has been dropped in the category of PREVIOUS APPLICATION DATASET.

- COLUMNS related to GENDER, ANNUITY , LOAN TYPE, INCOME, EDUCATION, ORGANISATION, OCCUPATION, FAMILY COUNT, GOODS FOR LOANS, etc has been considered for analysis purposes as these factors are the driving factors to understand the capacity of the customers.

- Rest columns have been removed.

- For data with “XNA” and other non available values , requisite modifications have been made for different categories like imputing with median, mode and mean.

DATA UNDERSTANDING

- Our target Variable is "TARGET" column that tell us whether any applicant is bound to be defaulter or not based on certain factors.

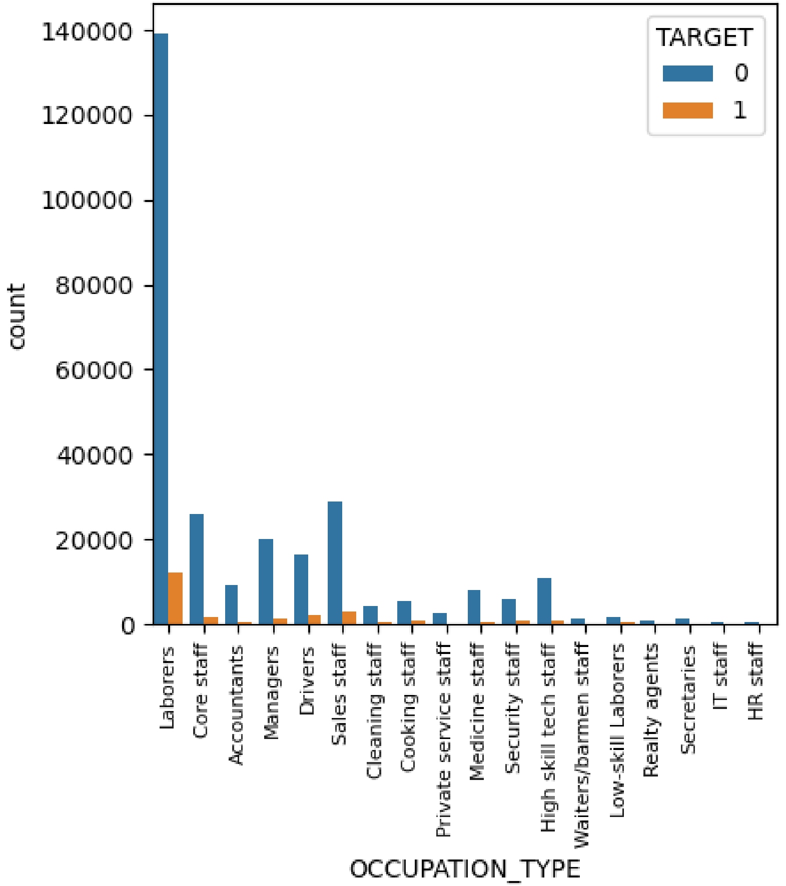
LISTED BELOW ARE THE COLUMNS TO BE FOUND SUITABLE FOR ANALYSIS REQUIRED TO PREDICT THE DEFAULTER CASES BASED ON FACTORS MENTIONED ALONG

- CODE_GENDER : Gender of the applicant
- CNT_CHILDREN : Number of children applicant has
- AMT_INCOME_TOTAL: Income of the applicant
- AMT_ANNUITY : Loan annuity
- AMT_GOODS_PRICE : Price of good for which loan is given
- AME_INCOME_TYPE : Clients income type

- NAME_EDUCATION_TYPE : Level of highest education
- OCCUPATION_TYPE : Type of occupation type of the client
- CNT_FAM_MEMBERS : Count of family members
- EXT_SOURCE_3 : Normalized score from external data source...

NOTE: Columns describing floor areas of the apartments, client type, dues if any, and other columns can be regarded as irrelevant as these do not support the capacity of the applicant or are either required by banks for application process. So, these can be ignored.

GRAPHICAL ANALYSIS – Application Data



OCCUPATION TYPE Vs DEFAULTERS/ NON DEFAULTERS

IN THE ABOVE GRAPH:

```
1. MAX number of defaulters fall under the category of labourers who have taken the loan.
2. There is no significant data for the HR staff.
3. SALES staff after labourers is the category with 2nd highest share of defaulters.
```

CONCLUSION :

Not much variation in TARGET audience with organisation type. HOWEVER, labourers record need to be looked upon while providing loans.

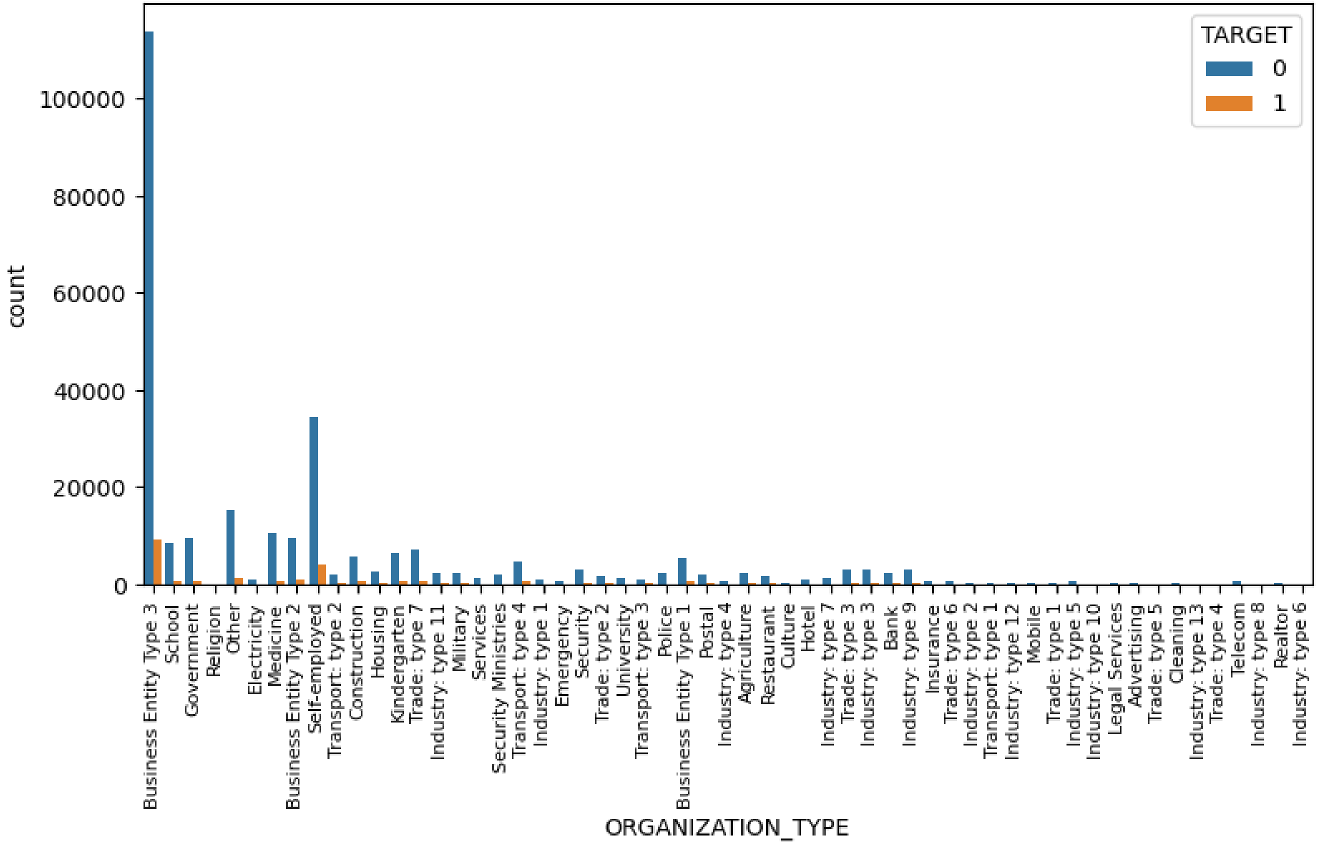
ORGANISATION TYPE Vs TARGET

-----

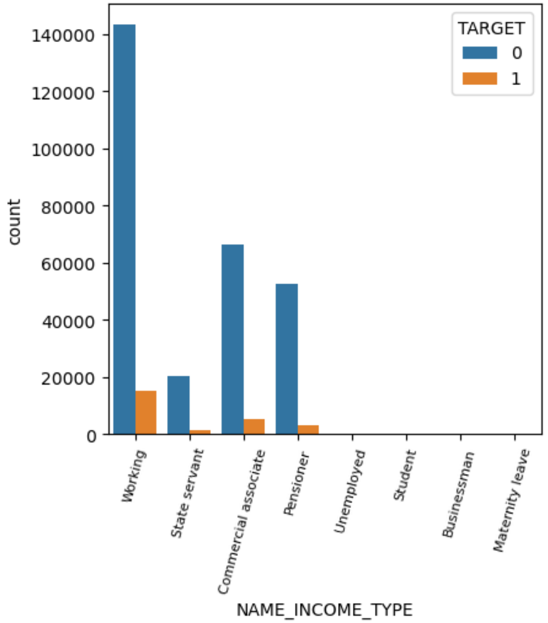
1. Business entity type 3 has highest number of defaulters as well as the one with the non defaulters.

2. The second position is ouccupied by self employed people

3. Other category falls under the 3rd place.



EDUCATION TYPE AND INCOME TYPE Vs TARGET VARIABLE

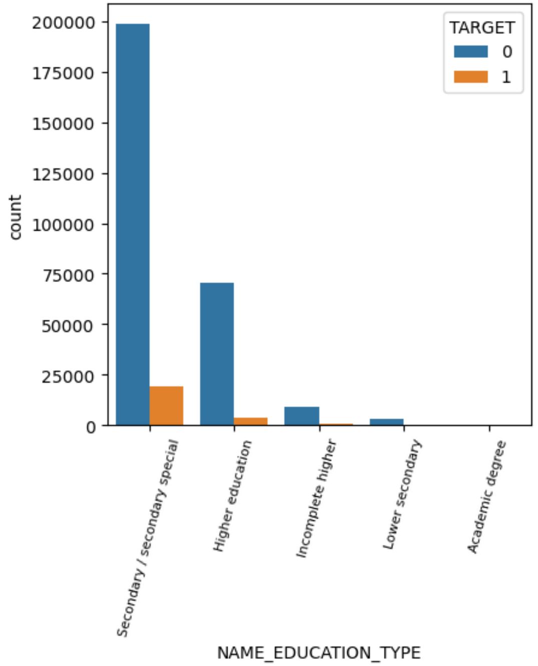


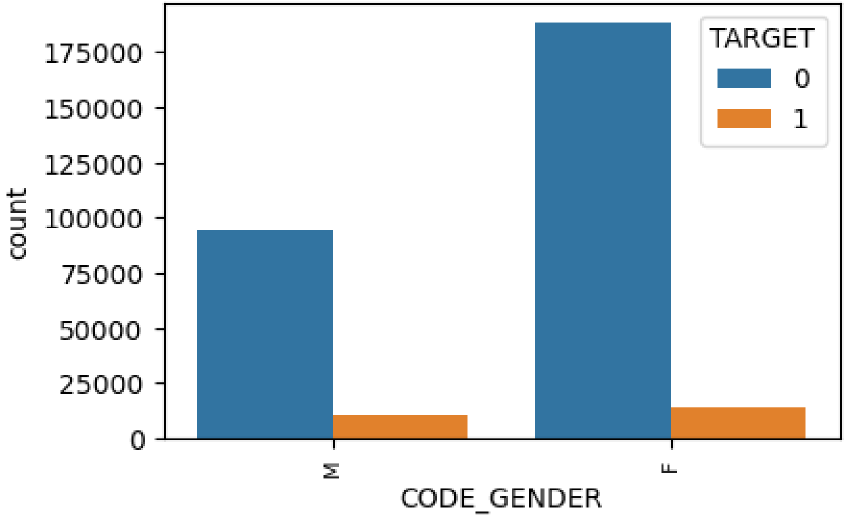
----

1. Secondary educated category has highest number of defaulters as well as non defaluters

2. People with higher education apply for loan and pay loan on time. This category has less number of defaulters.

3. Working category people apply for loan and fall under income category of people who might default.

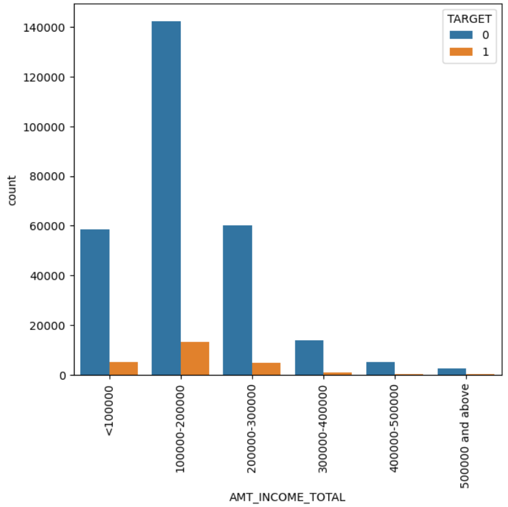


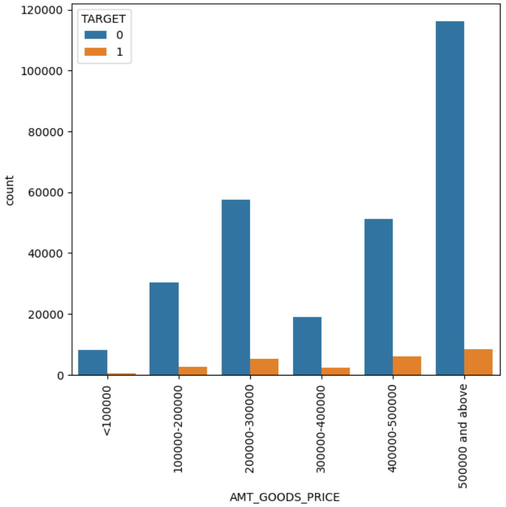


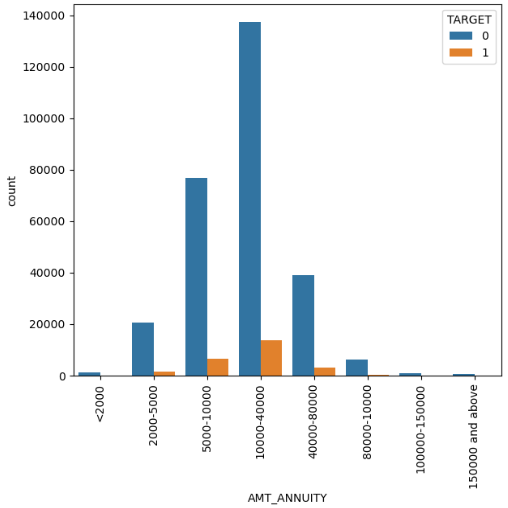
GENDER Vs TARGETS

There is not much significant difference between defaulters category in Terms of GENDER.

Females are much reliable category in terms of repayment of loans.







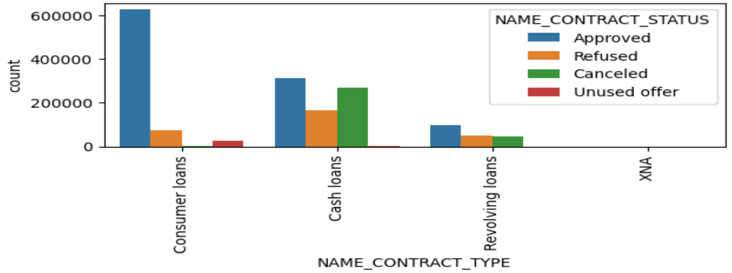
NUMERICAL VARIABLE Vs TARGET

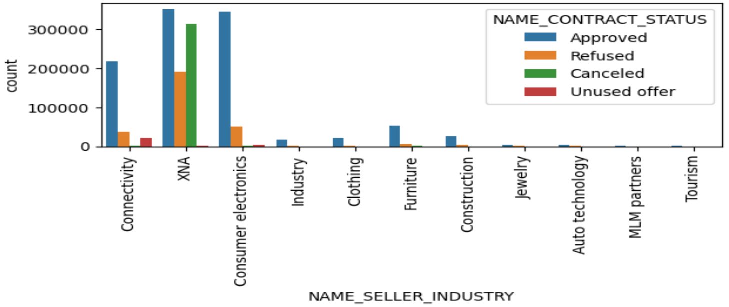
- People with salary under 100000 and under 200000 - 300000 range are the category have almost same ratio of defaulters and non defaulters

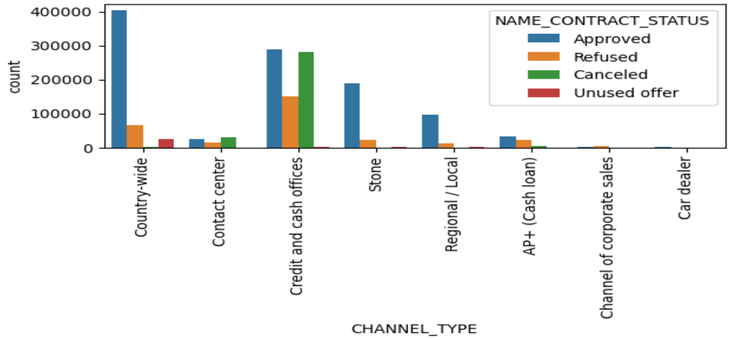
- People under 100000 - 200000 category are highest defaulters as well as non defaulters. These can be regarded as relaible category.

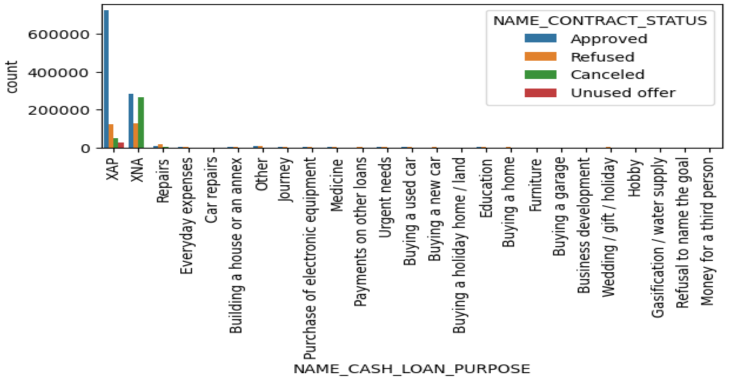
- People who applied for loan with goods price 200000 - 300000 have a significant share of non defaulters.

GRAPHICAL ANALYSIS - PREVIOUS APPLICATION DATASET









For PREVIOUS APPLICATION DATASET

----

1. CASH LOANS have been cancelled and refused in greater amount as compared to CONSUMER LOANS.

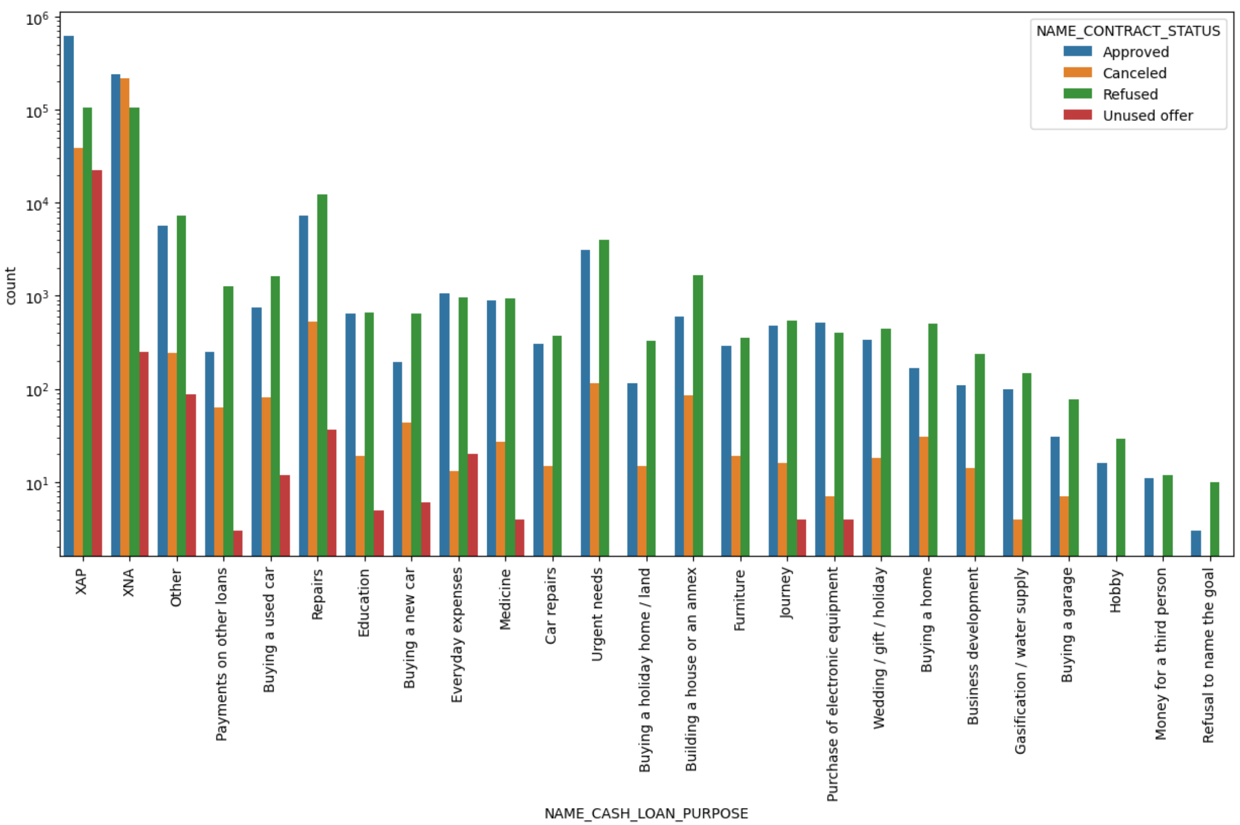
2. REVOLVING loans have a significant share of cancelled and refused to that of passed category.

3. For INDUSTRY category majority of data is unavailable.

4. CONSUMER CATEGORY loans got maximum approval for loans and after that connectivity category got the maximum approvals.

5. Repairs got maximum rejection in the CASH LOAN PURPOSE.

MERGED DATASET – ANALYSIS



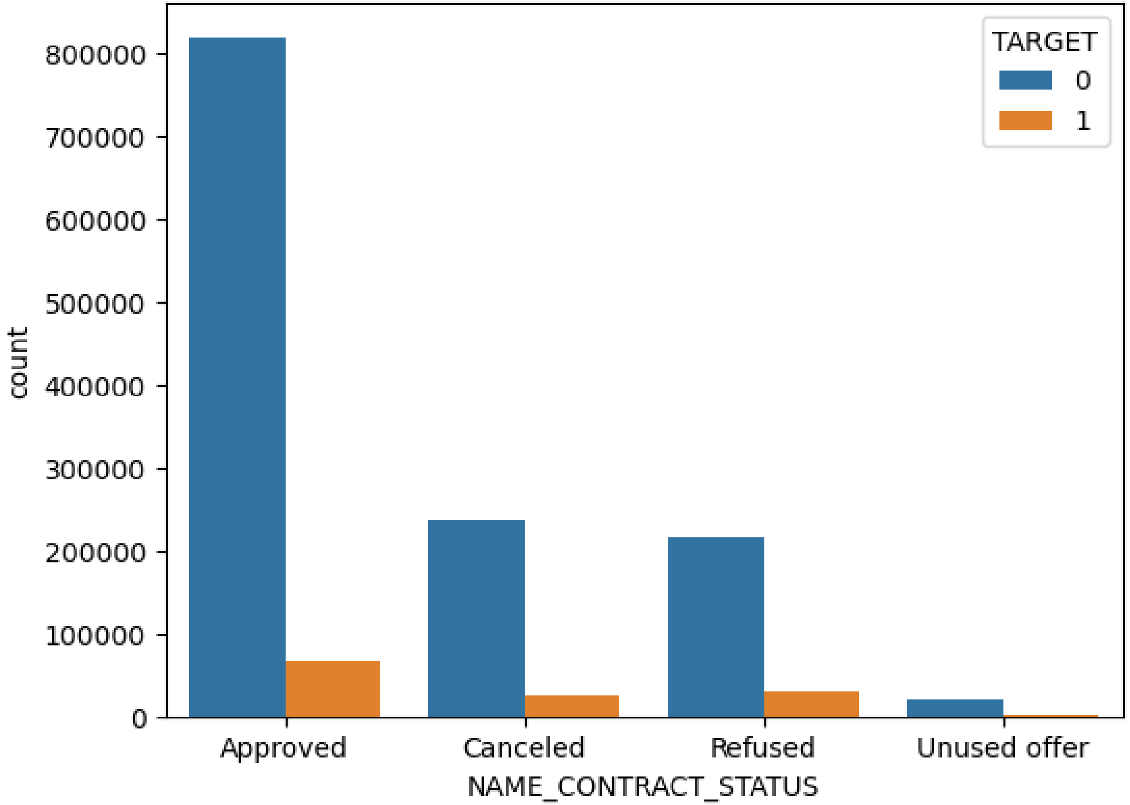
MERGED DATA ANALYSIS FOR VARIOUS CATEGORY

----

1. MAJORITY data is unavailable.

2. REPAIRS category got maximum refusal and cancelled.
3. URGENT NEEDS also fall under the category with almost equal refusal and cancelled category.

4. EVERYDAY expenses got maximum approval than refusal.



CONTRACT STATUS Vs TARGET VARIABLE

PEOPLE whose loan got refused or cancelled almost repayed the loans.

INSIGHTS AND CONCLUSION

OCCUPATION TYPE

- LOW SKILL labourers and drivers are highest defaulters.
- CORE STAFF , MANAGERS AND LABOURERS can be targeted with respect to the non deafulter and defaulter ratio .
- ACCOUNTANTS fall under less deafulters.

GENDER BASED

Females can be the target audience as they have high ratio of non defaulters to defaulters.

EDUCATION TYPE

WORKING PEOPLE AND HIGHER EDUCATION category can be targeted as they are less likely to default.

.

LOAN CATEGORY

CONSUMER LOANS can be targeted.

INCOME CATEGORY

PEOPLE income below 1 million and more than 1.5 million can be targeted.

ANNUITY AMOUNT

PEOPLE who can pay annuity of 100k are more likely to get loans.

MERGED DATA

1. MOST of the people who applied for loans have repayed.

2. Repairs category got the maximum cancellation. May be bank needs to check out for this segment before providing loans.

3. EVERYDAY expenses got maximum approval than refusal. So, this segment can be thought upon.

THANK YOU