Credit EDA Assignment

Submitted by - Arpan Das

Table Of Contents

Problem Statement

A financial institution specializing in loan provision has supplied two comprehensive datasets detailing customer information, loan-related info, and additional characteristics recorded by the institution. The data is organized into two distinct datasets:

1. **Application_data**: This dataset contains information pertaining to the current status of active loans, including a **TARGET** variable indicating whether a customer has defaulted on loan payments.
2. **Previous_application**: This dataset encompasses information related to past loan application.

The objective of the analysis is to conduct exploratory data analysis (EDA) to identify patterns or driver variables that serve as indicators of loan default risk. The insights derived from this analysis will enable the institution to formulate actions to better identify potential high-risk customers and issues if any.

High Level Approach (Detailed Approach mentioned in below slides)

1. Import the data frames and do variable analysis
2. Remove the columns which don't make sense , too many nulls, duplicates, so you don't have clean up unnecessary cols
3. Impute the rest of the columns with appropriate values
4. Rename/Merge-create a new columns if required
5. Univariate analysis to remove outliers
6. Multi variate Analysis – in the order of (a) Numerical – numerical (a) Categorical – numerical (b) Categorical – Categorical
7. To the same analysis of other data frame
8. Merge two data frames to check if new insights are available with additional columns

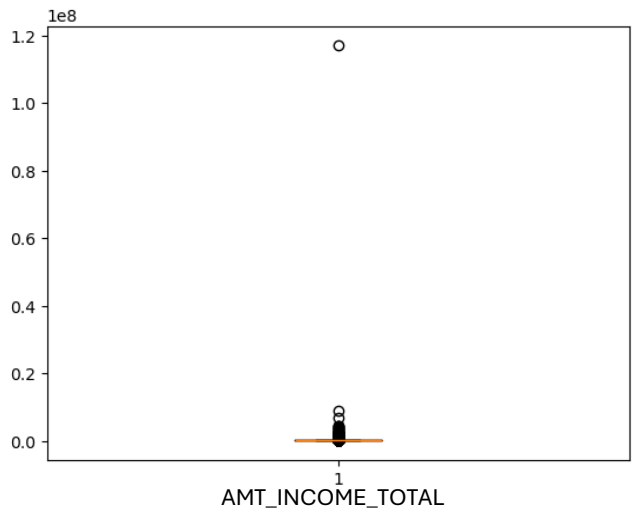
Detailed Approach and Analysis (Important Graphs attached)

1. **Application_data dataset Analysis**
   a) Import the csv into a data frame
   b) Analyse the columns using df.head()/tail() , info, describe - get the insights into columns make assumptions

```
days birth,days_employed,days_registration, id_publish,Days_last_phone_change have negative values- may need conversion to date field
Flag_***_**** - fields if the customer did provide info may be useful to identify customer's interest
Occupationtype,OrganizationType,region rating, OBS_30/60,DEF_30/60 ,AMT_REQ_CREDIT_BUREAU_* may be useful columns for default rate. Check
REG_CITY_NOT_LIVE_CITY,Days_last_phone_change like cols - may point to fraudulent behavior should be checked
Columns that are bank's operational columns and shouldn't have any impacts on the default rate - Weekday_appr_process_start,
HOUR_APPR_PROCESS_START,EXT_SOURCE_1,EXT_SOURCE_2,EXT_SOURCE_3
The columns related to Apartment specification seem redundant, also seem like have lot of missing values, check and drop
Flag_document_# - document type details not available, check and drop the columns that have lot of null values
```

   c. Check for null value columns- as there are many columns with lot of null values. I checked which columns have more than 50% null and which columns are 40% null. Based on the column relevance on default rate decided it's better to drop columns with more than 40% null, instead of imputing the values and distorting the data

c. Removed columns that are operational columns and have no bearing on default rate - WEEKDAY_APPR_PROCESS_START,HOUR_APPR_PROCESS_STAR
d. Removed some duplicate columns (identified definition e.g.- EG_CITY_NOT_LIVE_CITY)
e. There were lot of credit bureau columns (categorical) which seem important but had null. Post (index) check identified that all of these credit bureau columns suffer for null values in same rows. So deleted those rows, as it will be wrong to impute with most used or anything else
f. Removed some more columns which had null values and were not clearly defined or seemed irrelevant - NAME_TYPE_SUITE, EXT_SOURCE_3
g. Imputed the rest of columns with null values with median
h. Converted the DAYS_BIRTH column in Years Age of customer as its important column to check
i. Univariate Analysis :Checked for outliers and removed one big outlier in AMT_INCOME_TOTAL. Rest of the important numeric columns have outliers but spread across, so decided not to remove



c. Multivariate Analysis

i. Checked the correlation matrix to identify highly correlated columns any insights new info on column relationship

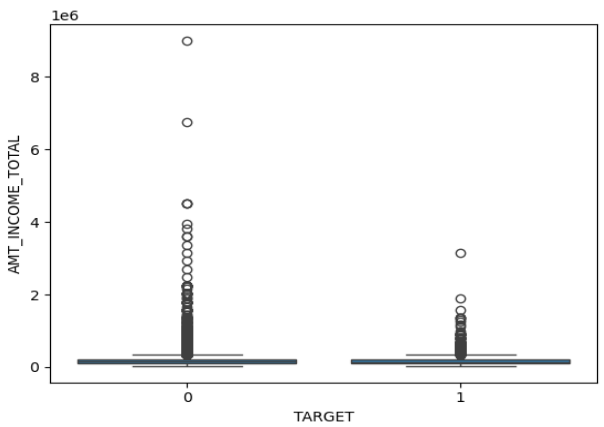
high corelated columns - check if any additional relationship can be figured out

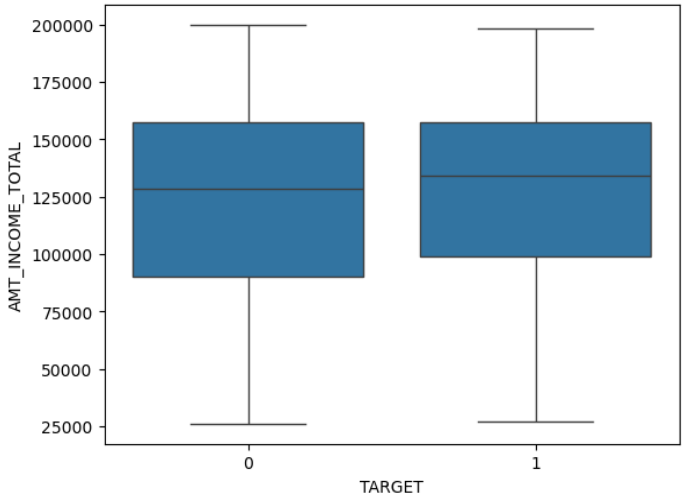
- CNT_CHILDREN is highly correlated with CNT_FAM_MEMBERS - expected
- AMT_CREDIT is highly correlated with AMT_GOODS_PRICE - expected loan amount dependent on underlying goods
- FLAG_EMP_PHONE is high negative correlation with DAYS_EMPLOYED - expected
- REGION_RATING_CLIENT_W_CITY highly correlated with REGION_RATING_CLIENT - expected nearly same
- LIVE_REGION_NOT_WORK_REGION highly correlated with REG_REGION_NOT_WORK_REGION - expected
- OBS_30_CNT_SOCIAL_CIRCLE highly correlated with OBS_60_CNT_SOCIAL_CIRCLE
- DEF_60_CNT_SOCIAL_CIRCLE highly correlated with DEF_30_CNT_SOCIAL_CIRCLE

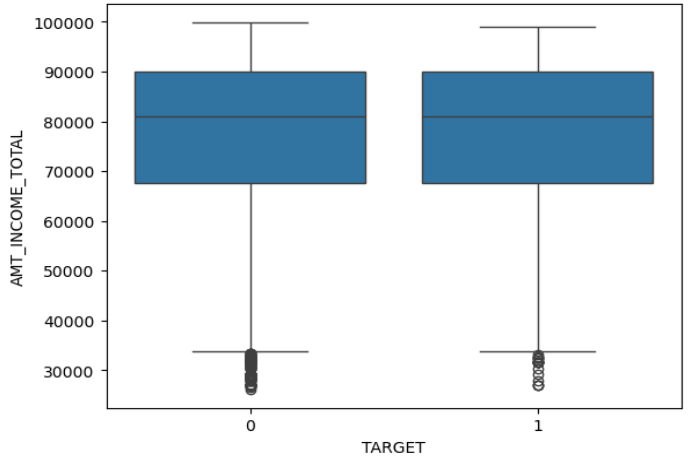
ii. Insights from multivariate analysis and plotting –

Started with categorical (TARGET) to numerical variable analysis . All fields checked (through describe and graph) few important one mentioned below

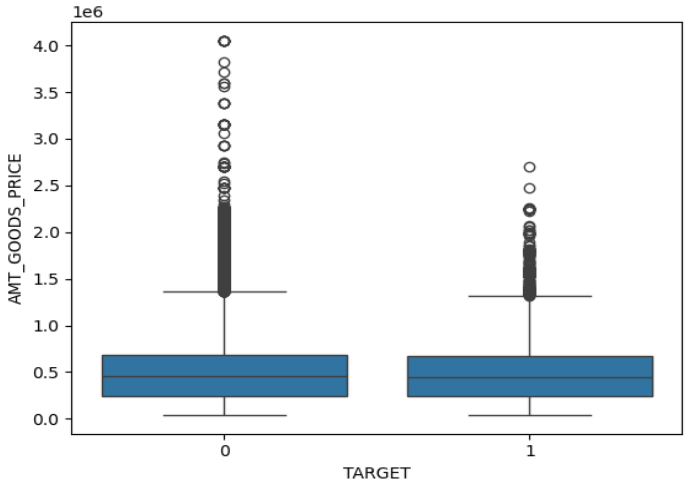
- AMT INCOME TOTAL doesn't impact the Default Rate i.e. TARGET

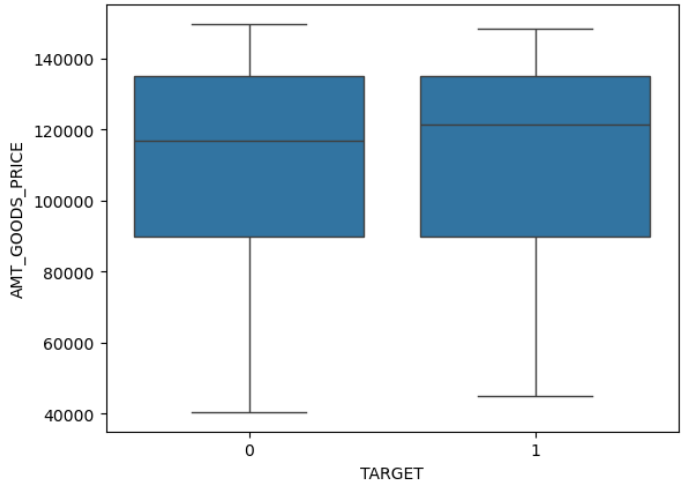


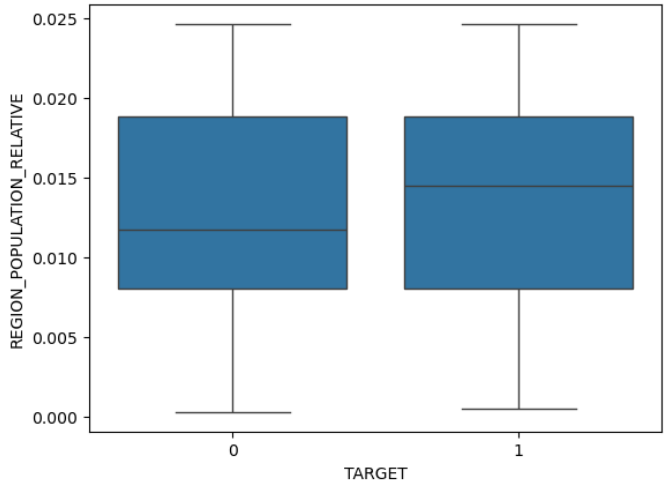




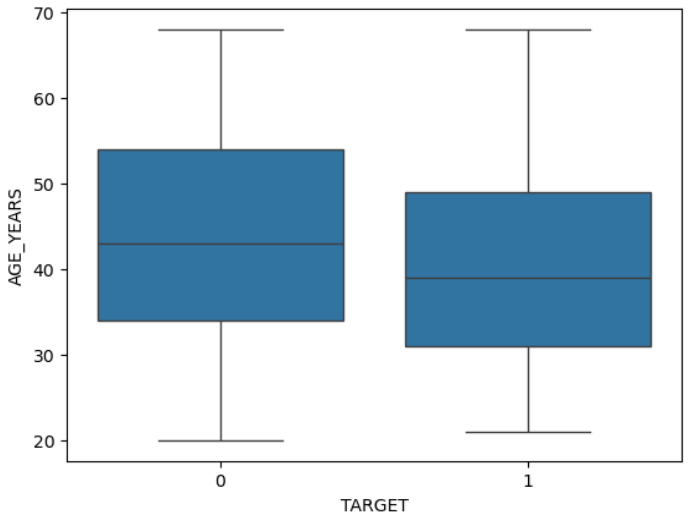
- AMT_GOODS_PRICE and REGION_POPULATION_RELATIVE doesn't impact the Default Rate i.e. TARGET

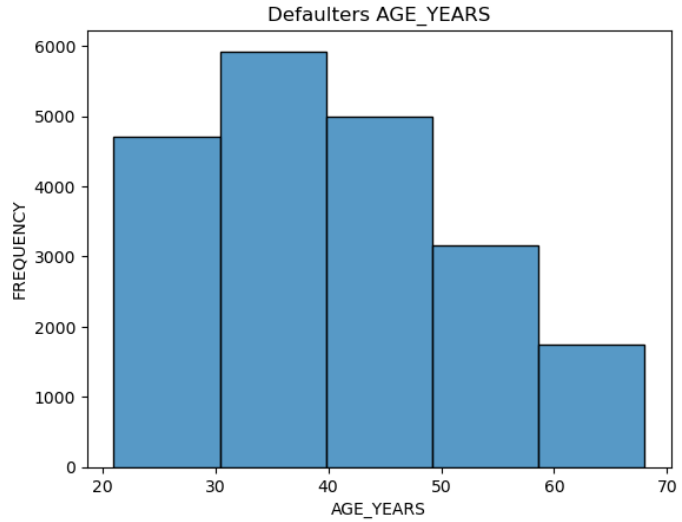


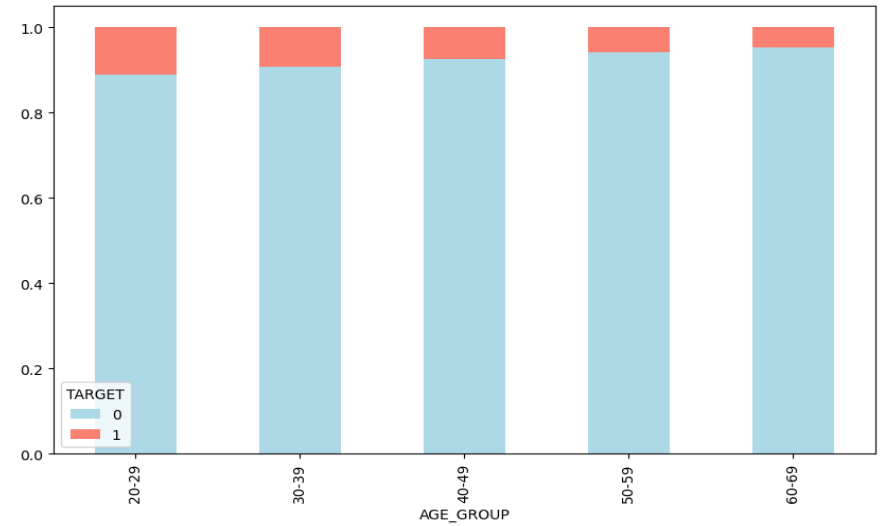




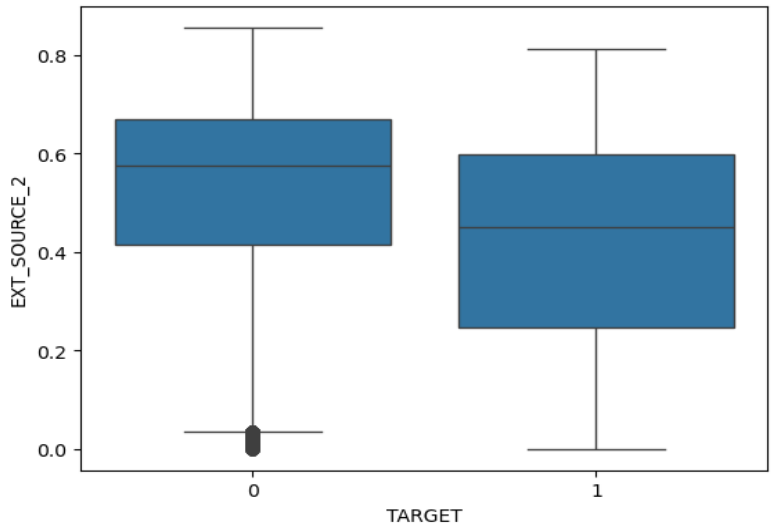
- AGEYEARS seem to hold some differences when checked. It seemed 30-40 years default more, but they may maximum number for applicants also , so checked through stacked bar to confirm default rate is high is lower age groups





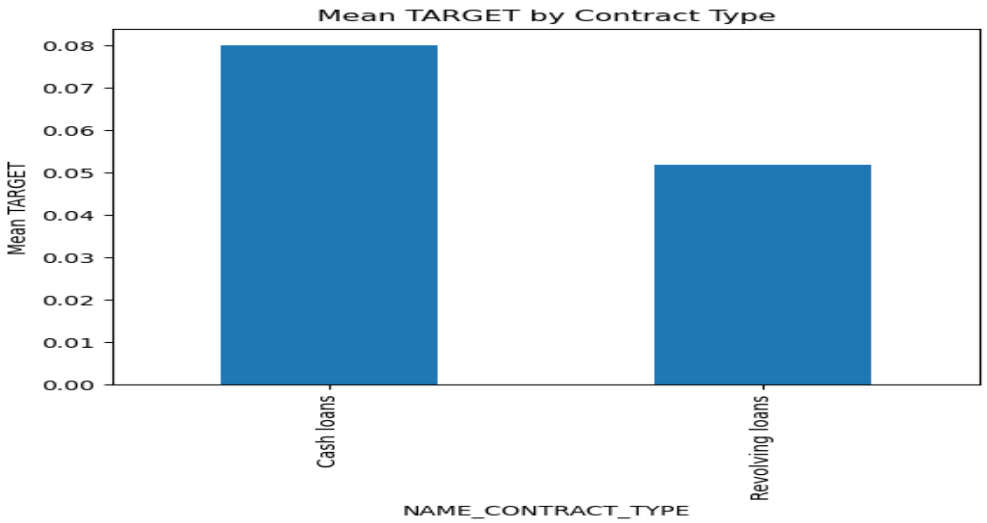


- EXT_SOURCE_2 seem to have a lot of impact on the TARGET, lower values of EXT_SOURCE_2 have less default

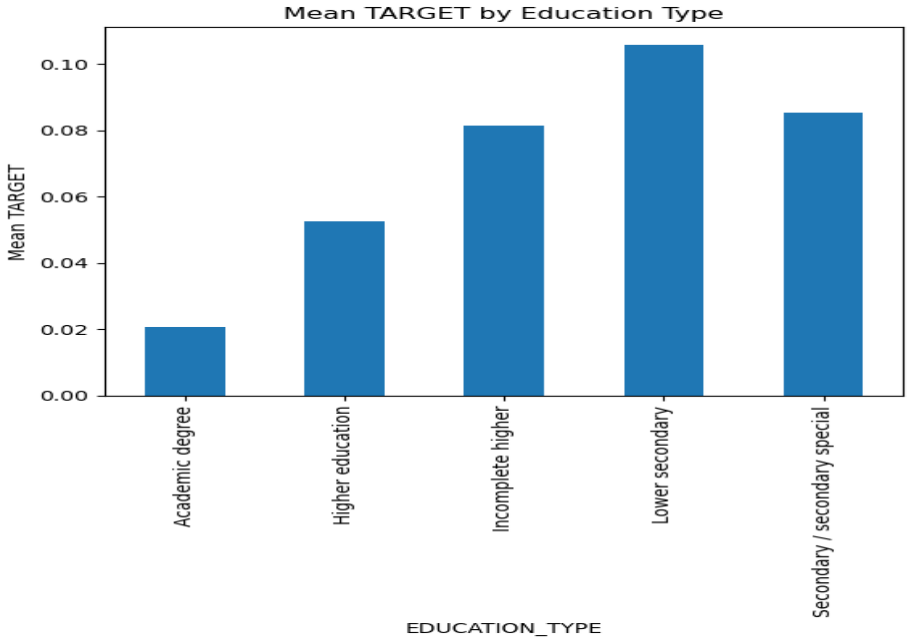


After Categorical to numerical variable analysis did –Categorical - Categorical Analysis Multivariate Analysis (Few important ones where some insights were found mentioned below)

- Cash loans are more risky than revolving credit

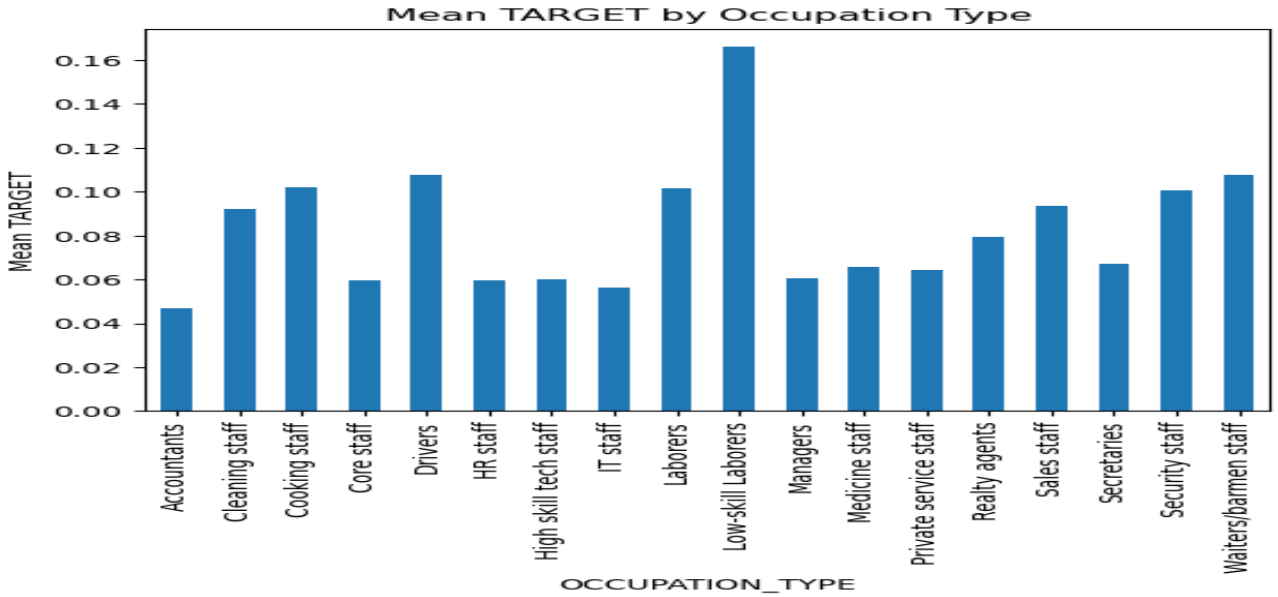


- Better the education of the customer lesser the default rate

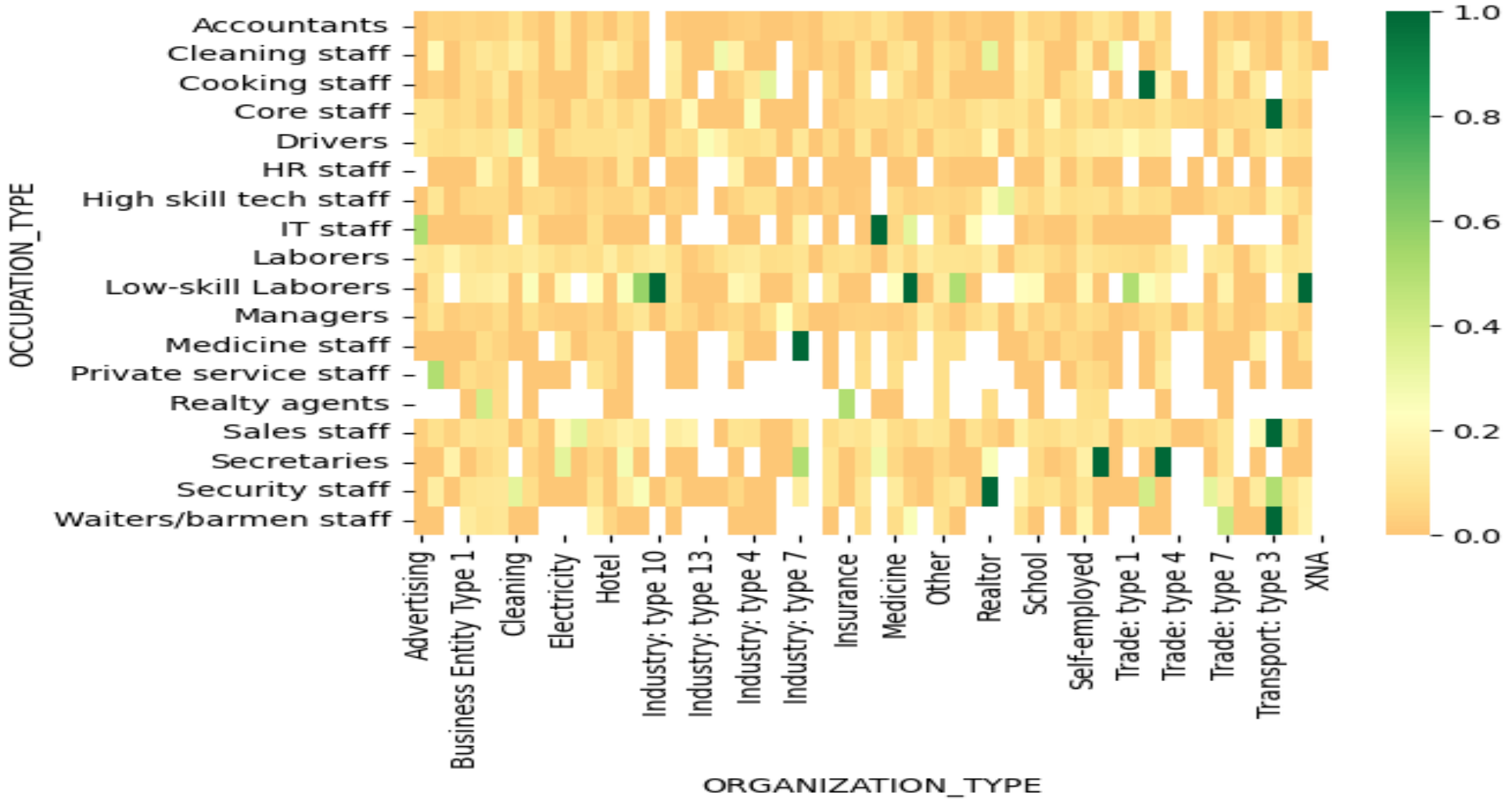


- NAME_FAMILY_STATUS and NAME_HOUSING_TYPE didn't have much to contribute towards the default rate

- Low Skilled had higher default rate



- Wanted to check if all low skill workers have higher propensity to default or based on some organization type, but no such pointers



2. Analysis of Previous_Application_Dataset

First step was to clean up and analyze the previous application dataset and then to merge it application dataset

a) Post import analyze the columns using head and tail

```
Analysis of the columns from head and tail
NAME_CONTRACT_TYPE, AMT_ANNUITY,
RATE_DOWN_PAYMENT,RATE_INTEREST_PRIMARY,RATE_INTEREST_PRIVILEGED,NAME_TYPE_SUITE - lot of NaN
NAME_CASH_LOAN_PURPOSE  NAME_CONTRACT_STATUS,CODE_REJECT_REASON,CHANNEL_TYPE,NAME_SELLER_INDUSTRY,NAME_YIELD_GROUP,NAME_SELLE
R_INDUSTRY check if the categorical values have any impact on the default rate of new app
PRODUCT_COMBINATION - is combination of multiple columns
DAYS_FIRST_DRAWING,DAYS_FIRST_DUE,DAYS_FIRST_DUE_1ST_VERSION,DAYS_LAST_DUE,DAYS_TERMINATION - are operational columns
NFLAG_INSURED_ON_APPROVAL - seems imp ,check if it can be used as a predictor variable
```

b) Checked null values , decided to delete the columns with more than 40% null values

c) Some bank's operational columns (such as WEEKDAY_APPR_PROCESS_START,HOUR_APPR_PROCESS_START) were deleted

d) Some irrelevant columns from business perspective (ANNUITY) and columns which had many null and will be wrong to impute (CNT_PAYMENT) were deleted

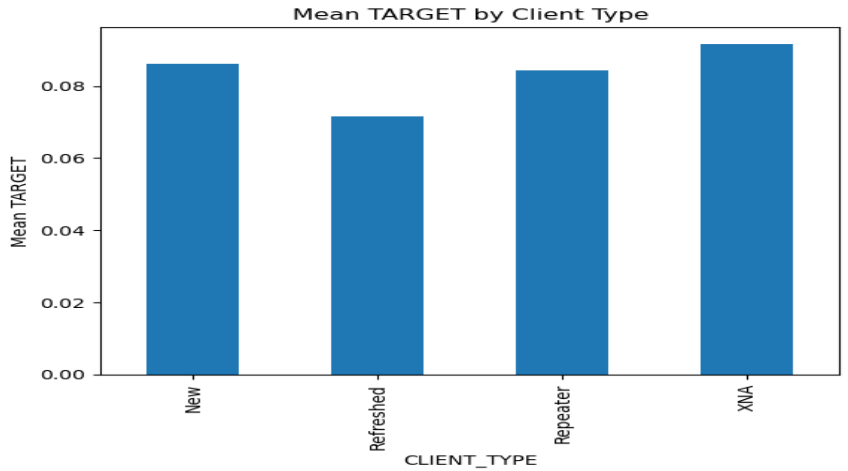
e) Amts_goods_price was imputed with mean

f) Checked the correlation matrix of other numeric columns - only AMT_GOODS_PRICE is highly correlated to AMT_CREDIT is expected

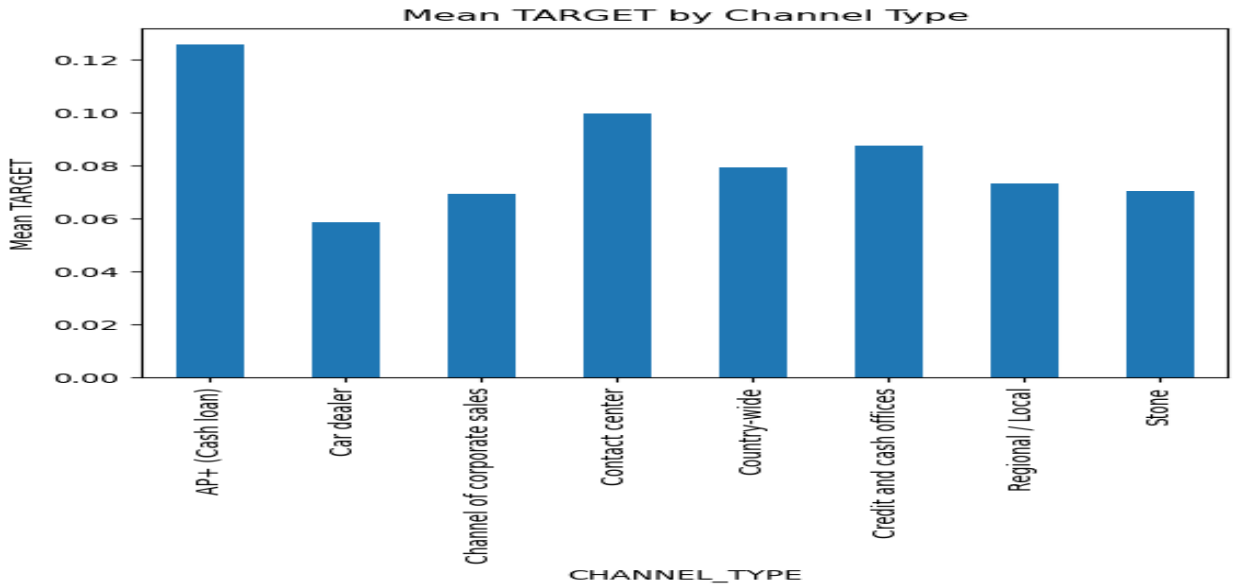
g) Merged the previous application data frames into application data frame on SK_ID_CURR

3. Multivariate analysis using Merged table - new columns - few key ones mentioned below

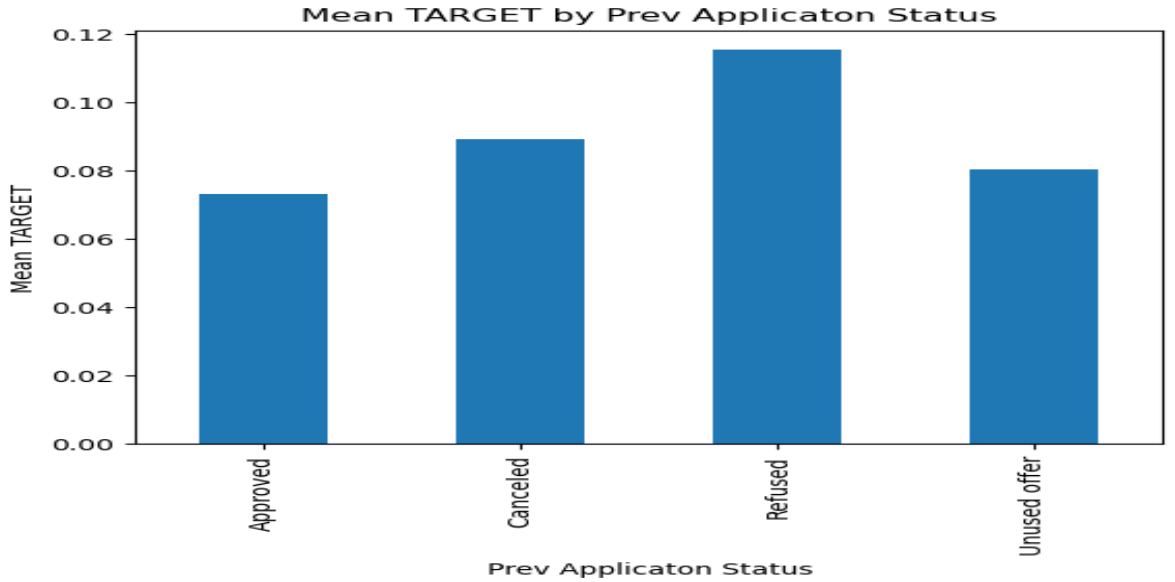
- Not impact difference in TARGET based on CLIENT TYPE



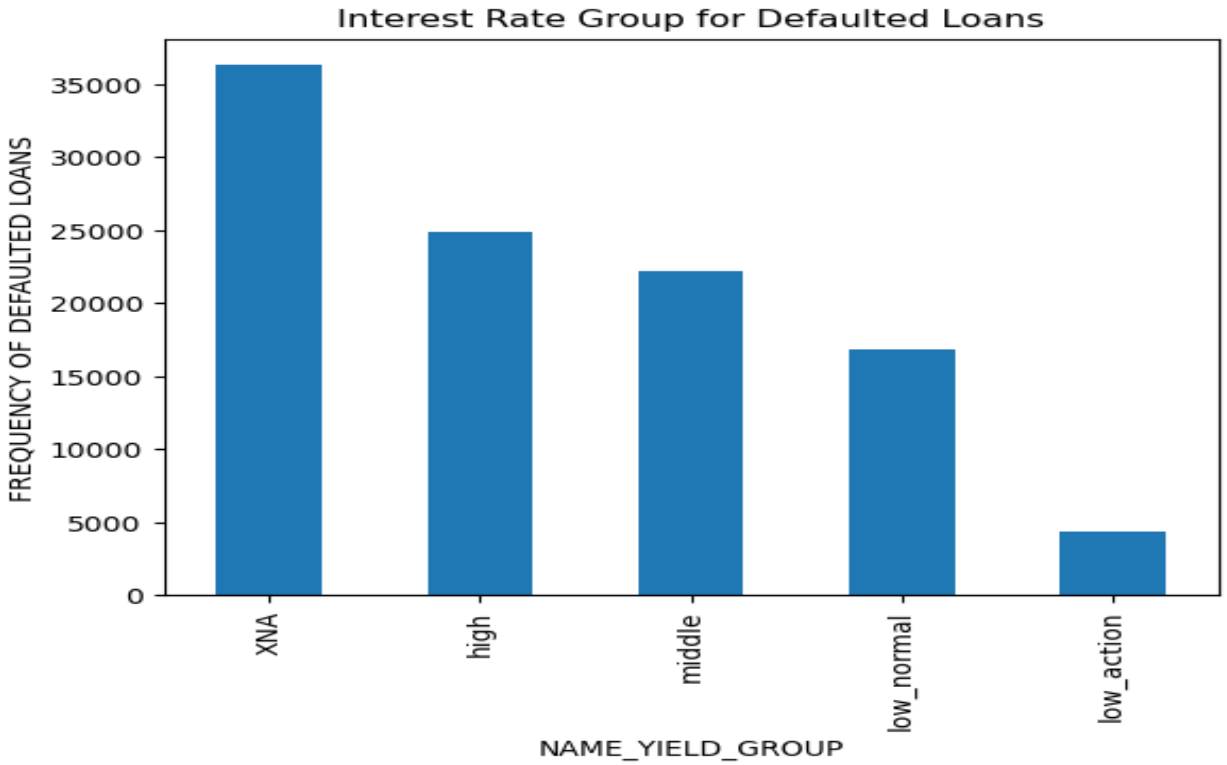
- One Channel Type has AP+ has higher default rate over others, bank should be inspect why the channel is getting wrong customer



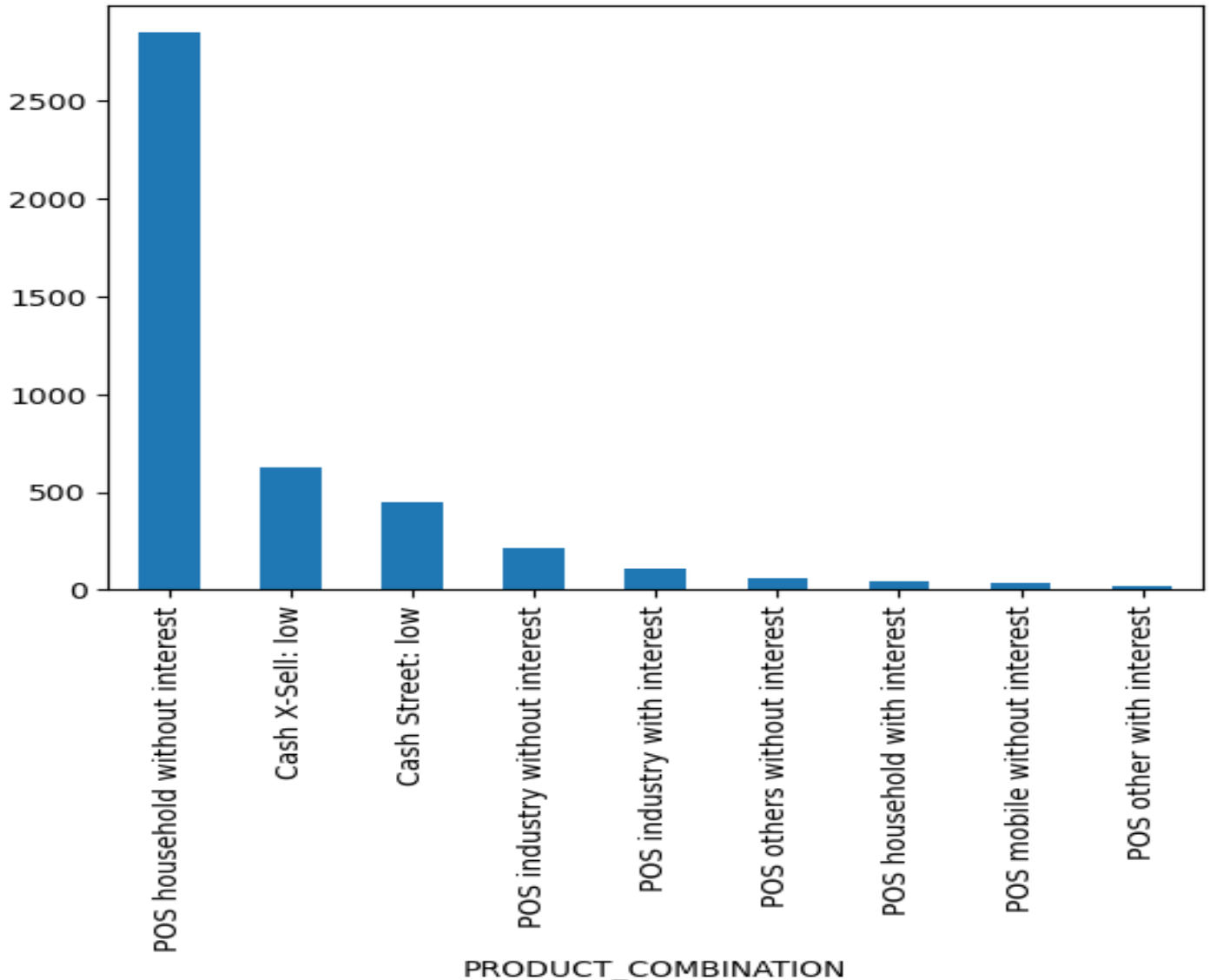
- Bank has provided loans to many applicants whom it has refused earlier, and they have higher default rate compared to others. May be tapping cancelled and unused clients will be better than going ahead with refused



- Wanted to check if Bank is getting higher return for risk, as the merged data frame has column related to interest rate charged by the bank. Seemed like bank risk return trade-off is sound, but not conclusive to data abstraction(not sure what XNA group stands for, rest seems good)



- To figure which product combination bank is getting risk return trade off wrong assuming low_action is lowest yield I potted the product vs yield. Seems like Seems like should reevaluate interest rate for group- *POS household without interest*



THANK YOU