

# Exploratory Data Analysis

(Credit EDA case study)





## 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 as their advantage by becoming a defaulter. Here we make use of 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

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):

- 
1. **Approved:** The Company has approved loan Application
  2. **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.
  3. **Refused:** The company had rejected the loan (because the client does not meet their requirements etc.).
  4. **Unused offer:** Loan has been cancelled by the client but on different stages of the process.

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



## Approach of the analysis

The steps followed for doing this analysis is as follows:

1. Data understanding
2. Data cleaning (Missing value handling and outlier handling)
3. Univariate, Bivariate and Multivariate analysis
4. Data Visualisation
5. Drawing and summarising Insights

First, we take a look at the 2 datasets in hand, check the rows and columns, shape of the datasets, the various data types, etc and get a broad understanding of the data at hand.




Second, we clean our data. We check for duplicates and missing values and then decide on the best way to handle them. If there are more than 40% of data missing from a given column, then we drop those columns.

If there an insignificant amount of missing values in a column for e.g. less than 10% or 15%, we then drop those records from our analysis.

If the amount of missing values is between 15 % and 40 %, then we examine the variable and take decision on either imputing, creating missing category or dropping the records depending on the situation and variable in question.

After dealing with missing values, we check for outliers in each dataset by using boxplots. If there are outlying points in the boxplots, we examine whether those data points are truly outliers or not (whether they may have occurred due to mistake or whether they are valid values or not). If a point is decided upon to be an outlier, we either subset them out or we cap the outliers depending on the variable at hand.



Third, we perform univariate, bivariate or multivariate analysis on the variables available in relation to our Target variable (which is either 1 for defaulters/people with payment difficulties or 0 for all other cases).

The goal is to find patterns on how the Target variable is affected i.e. how patterns of default can be determined from other variables in the dataset.

Univariate analysis concentrates on 1 variable at a time, bivariate analysis uses 2 variables (in this case we used Target variable and any other variable that we thought was relevant) and finally multivariate analysis that can take more than 2 variables and study how each affects the other.

The common tools used for this were value\_counts function, groupby functions, mean, median, quantiles, pivot tables, barplots, box plots, pieplots, etc.

We also accounted for imbalance in our dataset (target variable) by splitting the dataset into 2 parts.



Fourth, we visualised the insights using bar graphs, pie plots, boxplots, and heatmaps as visualisation is much more effective in conveying insights than raw numbers

Finally, we wrote down and summarized our observations or insights after each relevant graph.






## Observations and Insights derived from Bivariate analysis

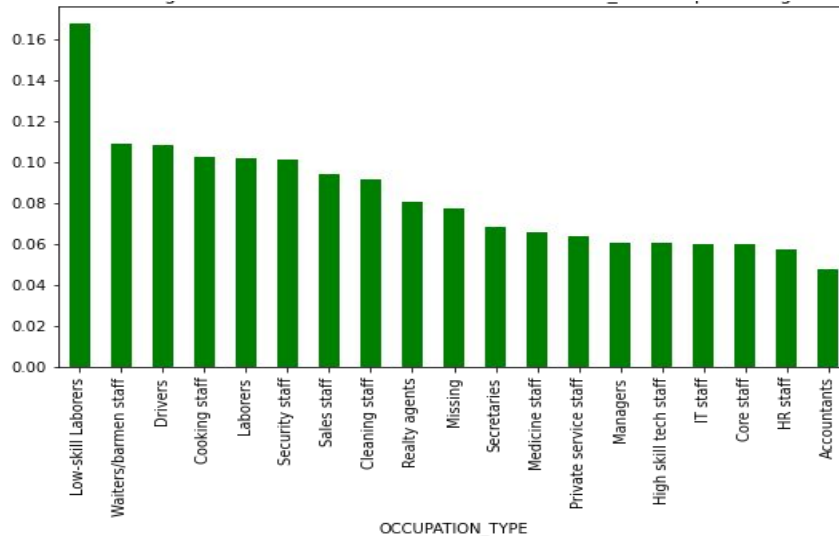
### Observations from Categorical variables:

- In education type column, we can see people with education upto lower secondary level and secondary/secondary special levels are more likely to have payment difficulties. This is probably due the fact people with less education tend to get low paying jobs and so their income is lower. On the other hand, people with Academic degrees and higher education are less likely to default.
- In the gender column, we can see that males are more likely to default then females. This is probably due to most females being housewives and are supported by their husbands so they don't need to request for loans as often.
- In the column Name\_type\_suite, we saw that people who are accompanied by individuals of 'Other' categories are more likely to have payment difficulties than those who are accompanied by children and family. This might be because people who face payment difficulties are given wrong advice regarding loans by third parties or maybe fraudsters.
- In the column organization type, we can see that people from transport, restaurant and construction type organizations are more likely to default. This is probably due to lower salaries. On the other hand, people from industry and trade are less likely to default
- In family status column, we can see that clients of civil marriage and those who are single/unmarried are more likely to have payment issues than clients who are widows or married.

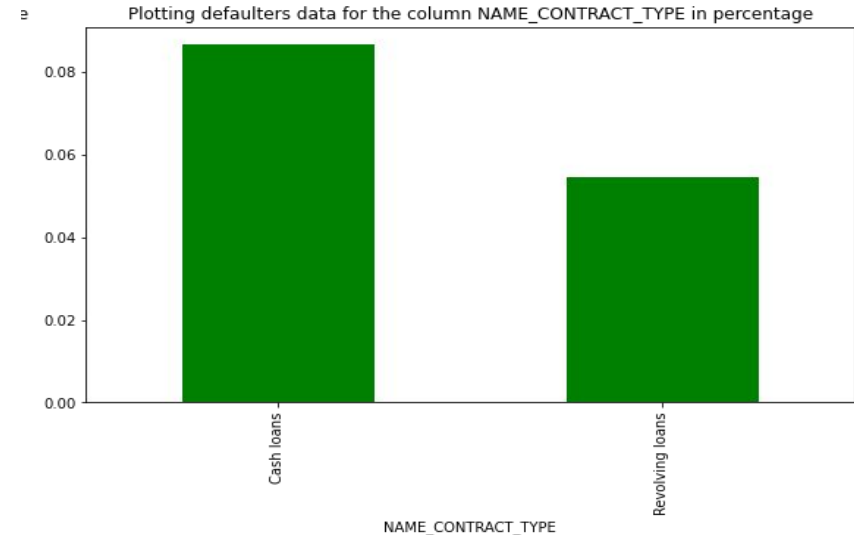
- 
- In columns Income type, we can see that working people and commercial associates are more likely to default as opposed to students and pensioners. This is probably because students are usually financially supported by their parents so they don't need to take loans often, and clients with pension are financially more secure than working people so they don't need to worry about loans either.
  - in column Name contract type, we found that cash loans usually face more default cases than revolving loans. This is probably due to the different nature and pay structure of these different types of loans.
  - In Occupation type, we found that low skill labourers, waiters/barmen, drivers, etc are more likely to face payment difficulties probably due to their lower income levels. However, accountants, HR staff, IT staff and managers are more likely capable of paying their due installments in time.
  - We also saw, people who do not own cars are more likely to have payment difficulties than car owners.
  - Finally, clients living in rented apartments or those who are living with their parents are more likely to face payment difficulties than those who are living in office apartments or co-operative apartments. People who rent apartments or those living with their parents may indicate lower income level or higher living expenses which makes them riskier to be approved loans.

Some graphs to support the insights we derived in the previous slides (Mean Target or average risk in y axis for all graphs)

Occupation type vs average risk

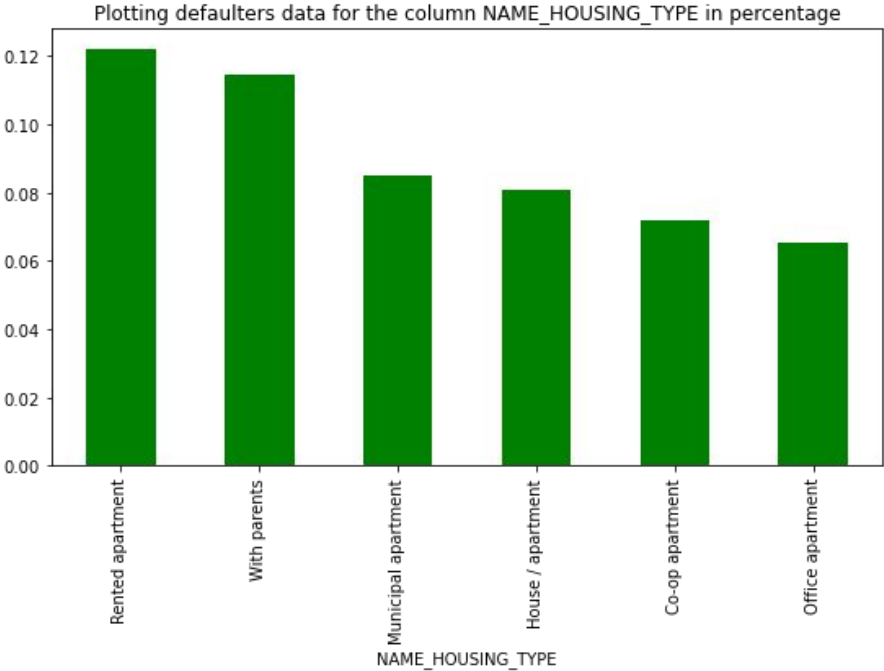


Contract type vs Average Risk

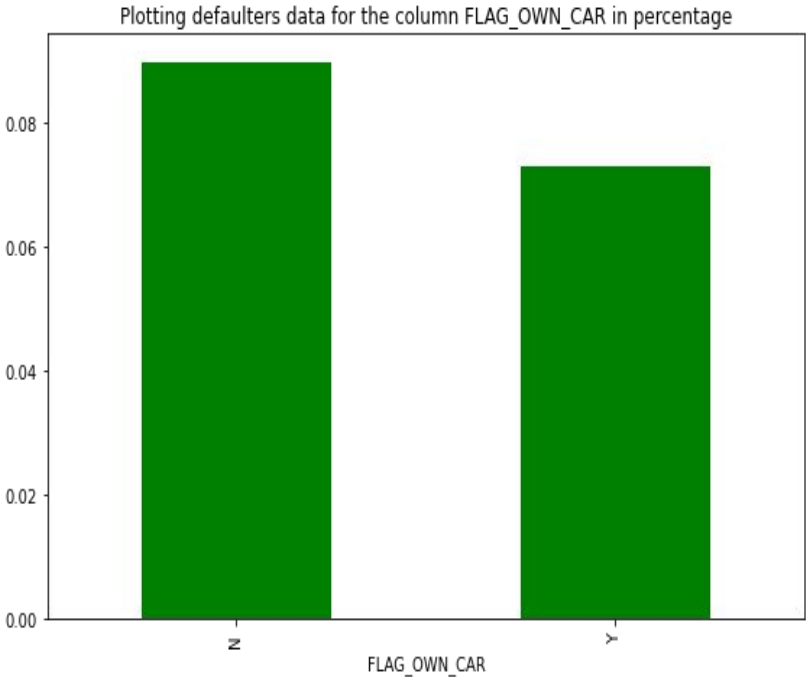




## Housing type vs Average risk




## Car owners/Non car owners vs Average risk





## Observations from Numerical variables:

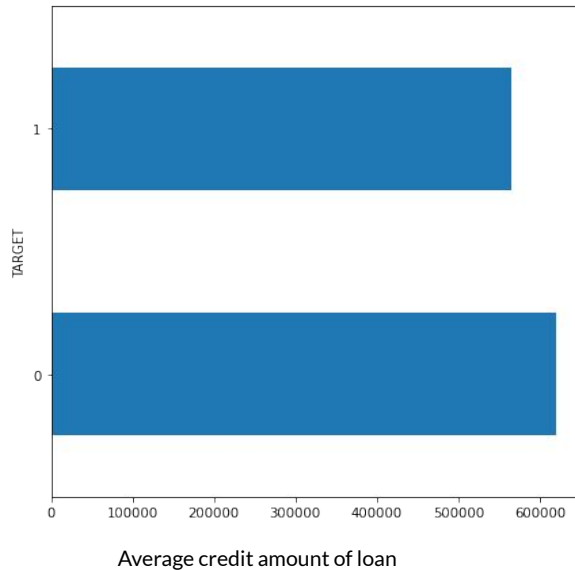
- **AMT\_REQ\_CREDIT\_BUREAU\_YEAR:** Higher average number of enquiries about clients to credit bureau one year before application is indicating higher risk
- **DEF\_60\_CNT\_SOCIAL\_CIRCLE:** Higher average observation of the clients with social surroundings that defaulted on 60 (days past due) DPD is indicating higher risk
- **REGION\_POPULATION\_RELATIVE:** There is lower risk with the clients from higher average relative population
- **AMT\_CREDIT:** There is higher risk with clients taking loans with lower credit amount on average
- **AMT\_GOODS\_PRICE:** There is lower risk with consumer loans having higher goods price on average
- **REGION\_RATING\_CLIENT:** Regions with higher rating are posing higher risk
- **CNT\_CHILDREN:** Clients with higher average number of children are posing higher risk.
- **DAYS\_ID\_PUBLISH:** Higher the average number of days before the application the clients changed their identity document with which they applied for the loan, higher is the risk
- **DAYS\_EMPLOYED:** There is lower risk for clients who have been employed at their current job for higher number of days before applying for the loan than those who have been employed for lesser duration on average.

- 
- DAYS\_BIRTH: There is lower risk with clients of higher age on average at the time of application
  - AMT\_ANNUITY: Clients with lower loan annuity on average pose higher risk
  - DAYS\_LAST\_PHONE\_CHANGE: There is higher risk posed by clients who changed their phone recently before making the application on average
  - DAYS\_REGISTRATION: There is higher risk posed by clients who changed their registration recently before making the application on average
  - AMT\_INCOME\_TOTAL: Clients with higher income level are posing less risk

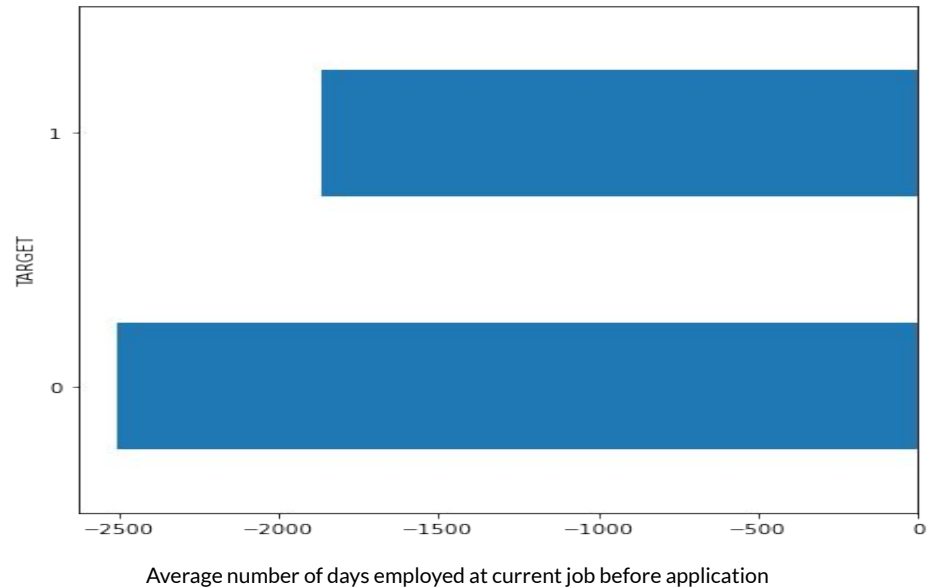


## Some graphs to support the insights we derived in the previous slides

Average credit amount of loan of vs Target

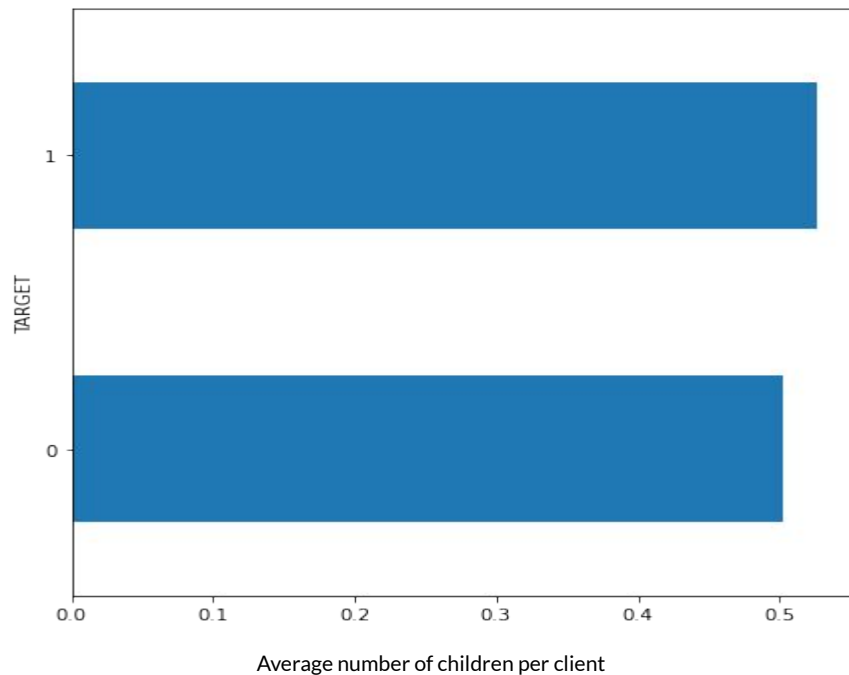


Average number of days employed at current job before application vs Target

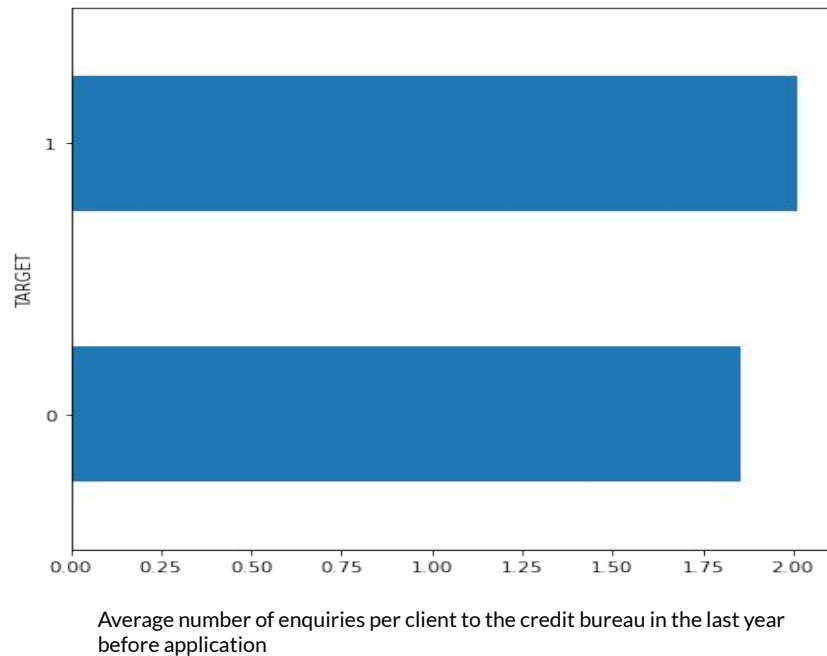




Average number of children per client vs Target



Average number of enquiries per client to the credit bureau in the last year before application vs Target



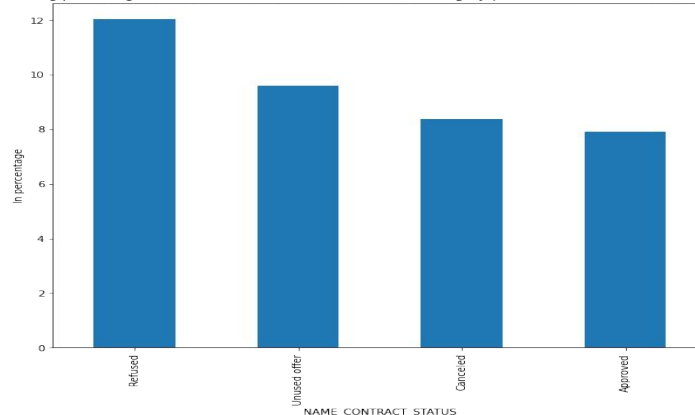


## Observations from previous application dataset(bivariate and multivariate analysis):

1. Clients who were refused the previous loan application are more likely to default than those who were approved
2. Clients of all types(Repeater, New, Refreshed or XNA) who had been refused the previous loan application are more likely to default than others. We can also see that new clients are likely to be more risky to give loans to than clients who are repeaters, refreshed or XNA.

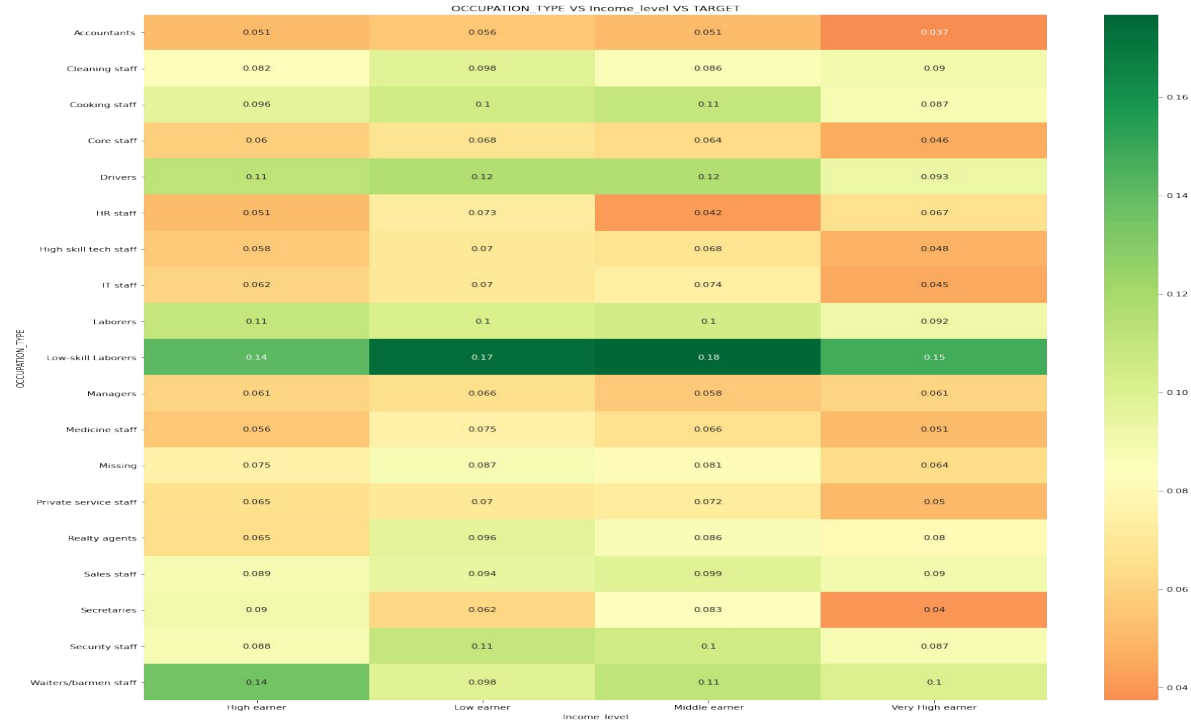
Contract status of previous application vs Target

Bar chart showing percentage contribution from each contract status category(previous loan) to the total number of defaulters



# Observations of Insights derived from Multivariate analysis

## OCCUPATION TYPE VS Income level VS TARGET:

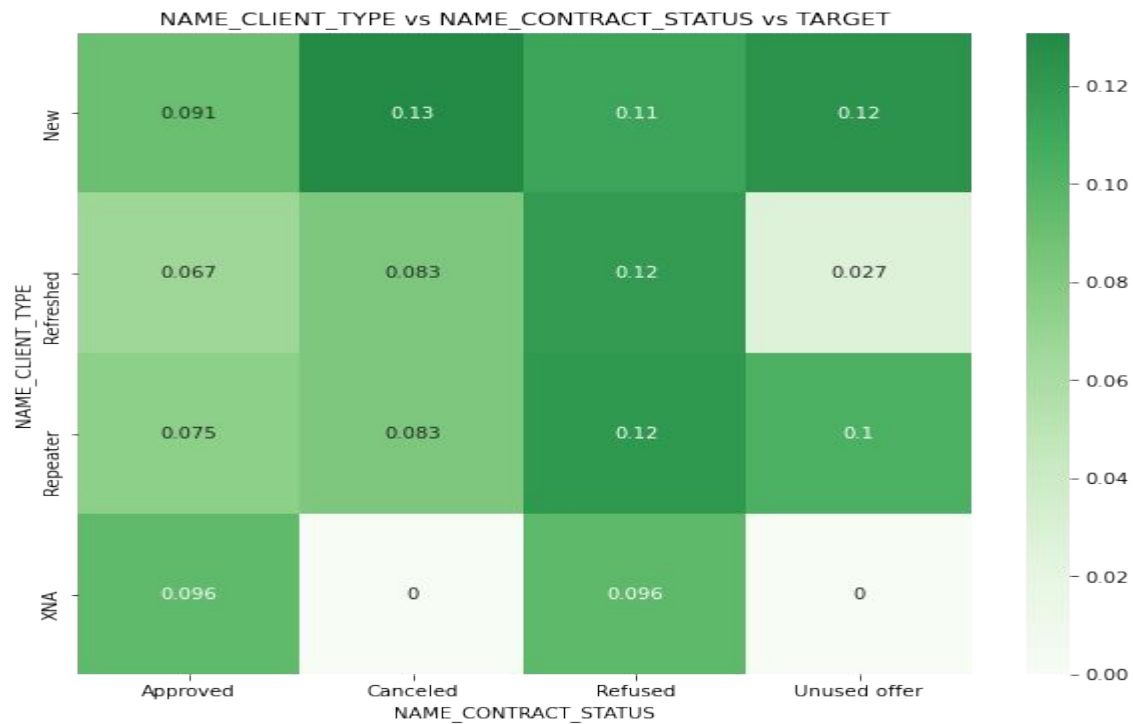




## **Observations:**

The above heatmap reinforces our previous observations firmly. We can clearly see that low skill labourers and waiters who earn less are high risk clients as opposed to Accountants, HR staff, core staff, high skill tech staff and secretaries who are at High or Very High income levels or middle income levels.

## CLIENT TYPE vs CONTRACT STATUS vs TARGET





## Observations:

As we can see from the above heatmap, clients of all types(Repeater, New, Refreshed or XNA) who had been refused the previous loan application are more likely to default than others. We can also see that new clients are likely to be more risky to give loans to than clients who are repeaters, refreshed or XNA.


In [ ]:



## Recommendations

Based on the entire analysis, we can make some recommendations in order to make better judge a client's credibility/their will default or not.

- The bank should categorise every new client into 2 categories: high risk or low risk based on on whether that person's probability of default.
- If the client is more likely to face payment difficulties then he/she should be categorised as a high risk client. On the other hand, if a client is less likely to face payment difficulties, then he/she should be categorised as low risk client.
- The decision on classifying a client into high risk or low risk category should be based his/her financial situation e.g. variables in the application dataset like income level, occupation, family status, etc.
- The bank must use the patterns or effects of these variables on default/risk that we learnt from this analysis in the classification decision. For e.g. labourers or barmen with low income level, clients who were refused previous loan application, new clients, unmarried people, people with higher number children, etc should be considered high risk.
- On the other hand, clients who are accountants or HR staff with high income, those who were approved on previous loan application, repeater or recurring clients, people who are widows or married, clients with lower number of children, etc should be considered low risk.

- 
- After classifying the client into to high risk or low risk, the bank can then decide whether to approve loan or not by assessing the level of risk posed..
  - In case of high risk clients, bank can either decide to reject their loan application or approve loan on some specific conditions.
  - If the risk of default is considerably high for a given client then the bank should reject the application.
  - However, if the bank does decide to give loan to a high risk client then such a loan should be given with lower principal/credit amount or lower period of tenure of the loan or higher rate of interest or a combination of all of these.
  - In case of low risk clients, the bank should assess the client's previous loan records and then decide to approve the loan. If the level of risk is low enough, then the bank can approve the loan to the client.