Exploratory Data Analysis

Banking Dataset Case Study – Prepared by Sourav Das

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. Suppose you work for a consumer finance company which specializes in lending various types of loans to urban customers. You have to use EDA to analyze 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.

Business Objectives

This case study aims to identify patterns which indicate if a client has difficulty paying their installments 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.

Data Understanding

The Complete Dataset contains 3 files and below are the explanations of each files:

'application_data.csv' contains all the information of the client at the time of application. The data is about whether a client has payment difficulties along with many other information related to each client. This dataset having 122 columns and 307511 rows of information.

'previous_application.csv' contains information about the client’s previous loan data. It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer. This dataset having 37 columns and 1670214 rows of information.

'columns_description.csv' is data dictionary which describes the meaning of the variables. Basically this data to understand the business meaning of the data given in above two files.

Data Preparation

New application dataset is a big in terms of features. After analyzing the meaning of each features, many columns looks not revenant into our analysis hence all these need to be dropped from the dataframe. In total of 99 columns removed from New application dataset.

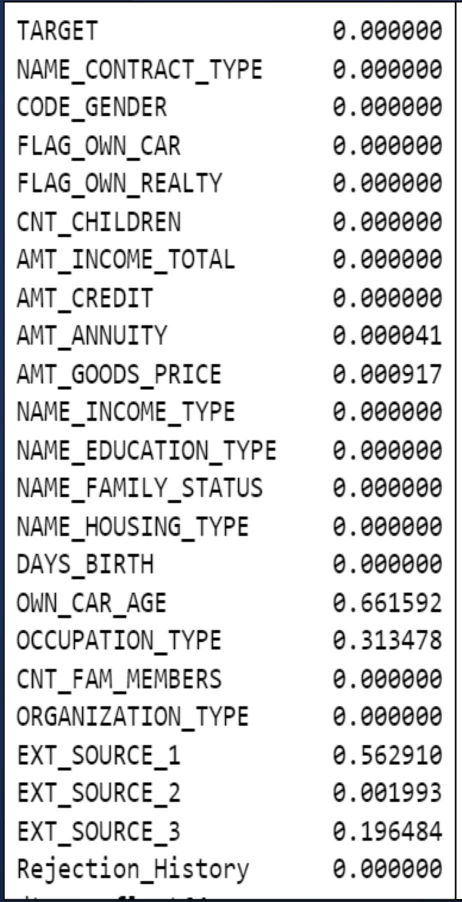
After understanding the meaning of Old application dataset features, decided to keep only 2 columns in that dataset 'SK_ID_CURR’ & 'NAME_CONTRACT_STATUS’. 'SK_ID_CURR’ is the current applicant ID, this is required for mapping with new application dataset. 'NAME_CONTRACT_STATUS’ is basically represents as mentioned below.

- 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 on different stages of the process.

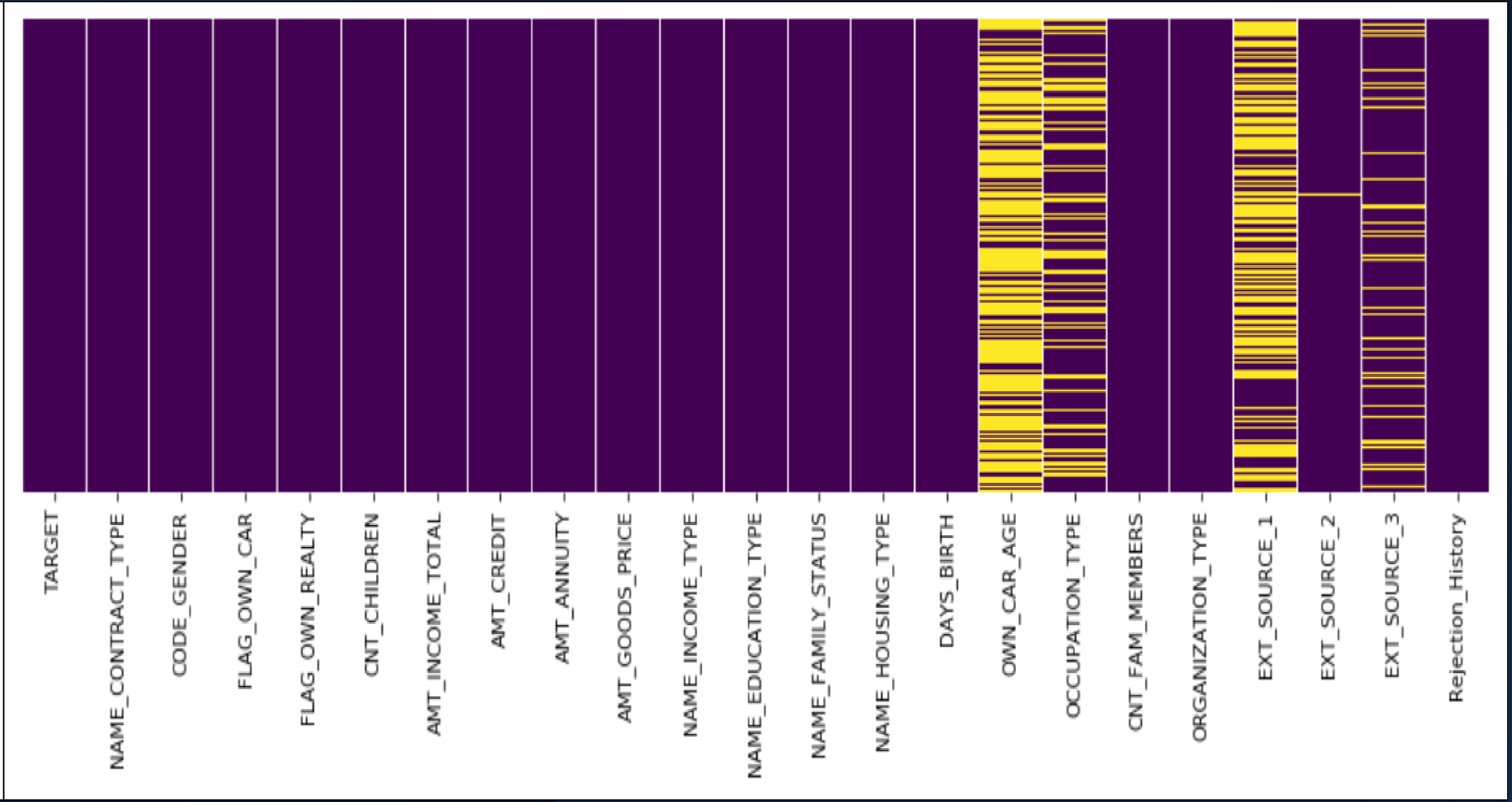
Counted how many times any applicant got Rejection before. Then merge these 2 above mentioned dataset into one, which is the final dataset for the analysis.

Missing value analysis

Below table shows % of missing values in each columns and attached below heatmap of missing values in entire dataset to understand randomness.



```
TARGET                 0.000000
NAME_CONTRACT_TYPE     0.000000
CODE_GENDER            0.000000
FLAG_OWN_CAR           0.000000
FLAG_OWN_REALTY        0.000000
CNT_CHILDREN           0.000000
AMT_INCOME_TOTAL       0.000000
AMT_CREDIT             0.000000
AMT_ANNUITY            0.000041
AMT_GOODS_PRICE        0.000917
NAME_INCOME_TYPE       0.000000
NAME_EDUCATION_TYPE    0.000000
NAME_FAMILY_STATUS     0.000000
NAME_HOUSING_TYPE      0.000000
DAYS_BIRTH             0.000000
OWN_CAR_AGE            0.661592
OCCUPATION_TYPE        0.313478
CNT_FAM_MEMBERS        0.000000
ORGANIZATION_TYPE      0.000000
EXT_SOURCE_1           0.562910
EXT_SOURCE_2           0.001993
EXT_SOURCE_3           0.196484
Rejection_History      0.000000
```



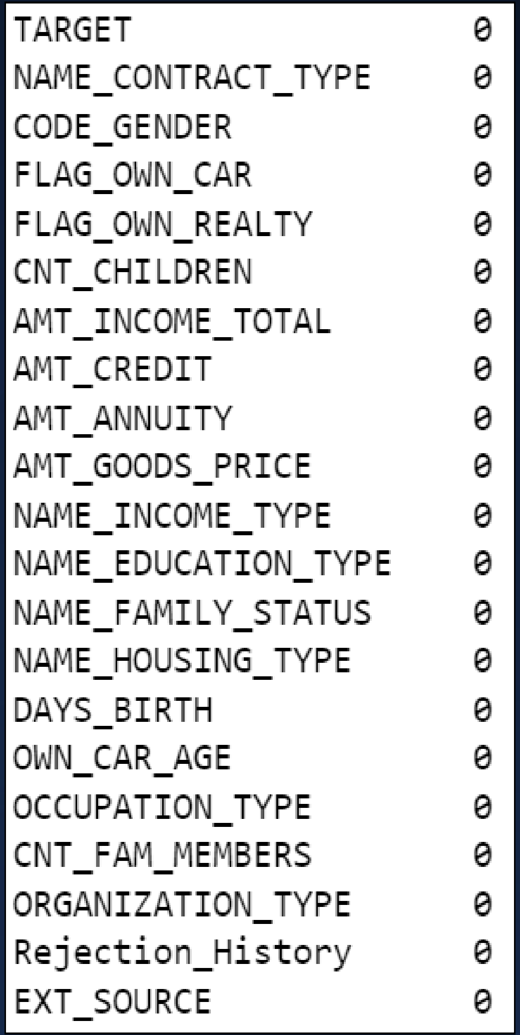
Impute/Remove Missing Values

Column "OWN_CAR_AGE" having very high missing values i.e 66%. But considering the fact if a person doesnt own car then it can be blank. Hence need to impute 0 values for "FLAG_OWN_CAR" No cases.

EXT_SOURCE_1, EXT_SOURCE_2 & EXT_SOURCE_3 are the score given by 3 different external companies. As we can see many entries are missing for EXT_SOURCE_1 & EXT_SOURCE_3. But EXT_SOURCE_2 having almost all the values. Hence instead of removing 2 columns we can make mean of all these 3 values and use it as single score.

One column named "OCCUPATION_TYPE" is having missing values of 31.35%. We cannot simply remove 31.35% of data hence either we can use some algorithm to predict missing values from other features in dataset or we can use it as new category as ‘Unknown’. Second opted here as first one required lot many time efforts and also it will enhance computation resource & time.

There are few missing values in column AMT_ANNUITY, AMT_GOODS_PRICE & OWN_CAR_AGE. As these all are numerical columns hence we can replace the missing values with median values. Right table shows count of missing after all these operations.



```
TARGET                 0
NAME_CONTRACT_TYPE     0
CODE_GENDER            0
FLAG_OWN_CAR           0
FLAG_OWN_REALTY        0
CNT_CHILDREN           0
AMT_INCOME_TOTAL       0
AMT_CREDIT             0
AMT_ANNUITY            0
AMT_GOODS_PRICE        0
NAME_INCOME_TYPE       0
NAME_EDUCATION_TYPE    0
NAME_FAMILY_STATUS     0
NAME_HOUSING_TYPE      0
DAYS_BIRTH             0
OWN_CAR_AGE            0
OCCUPATION_TYPE        0
CNT_FAM_MEMBERS        0
ORGANIZATION_TYPE      0
Rejection_History      0
EXT_SOURCE             0
```

Outlier analysis and Fixing columns

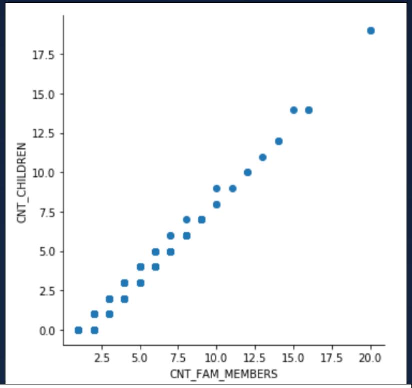
New Columns created as ‘Age’ from column ‘DAYS_BIRTH’ as ‘DAYS_BIRTH’ column values are in days and also in negative in numbers. ‘Age’ column is having clients Age in years.

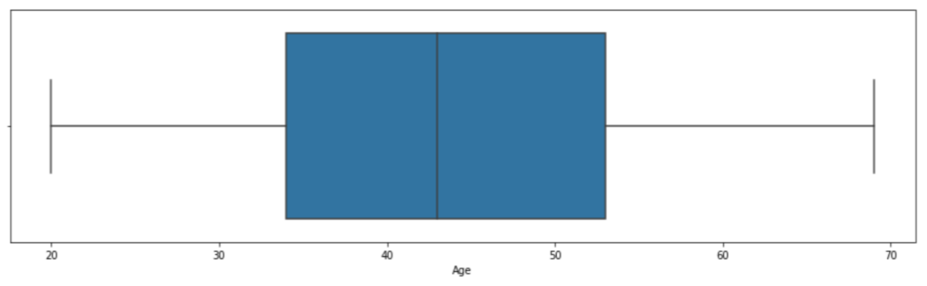
Categorical columns created like ‘Income_Group’, ‘Family_Members’ and ‘Age_Group’ for better analysis.

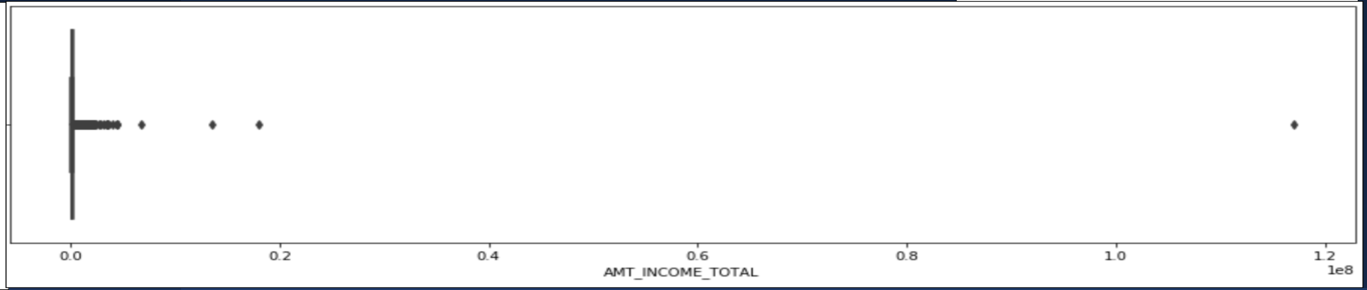
Data type conversion done for CNT_FAM_MEMBERS column as it cannot be in float hence needed to convert it to integer.

Outlier didn’t found in Age, Family_Members & CNT_Children.

Outliers found in Income column and above 99 percentile of this data removed from the dataset for better result in our analysis.





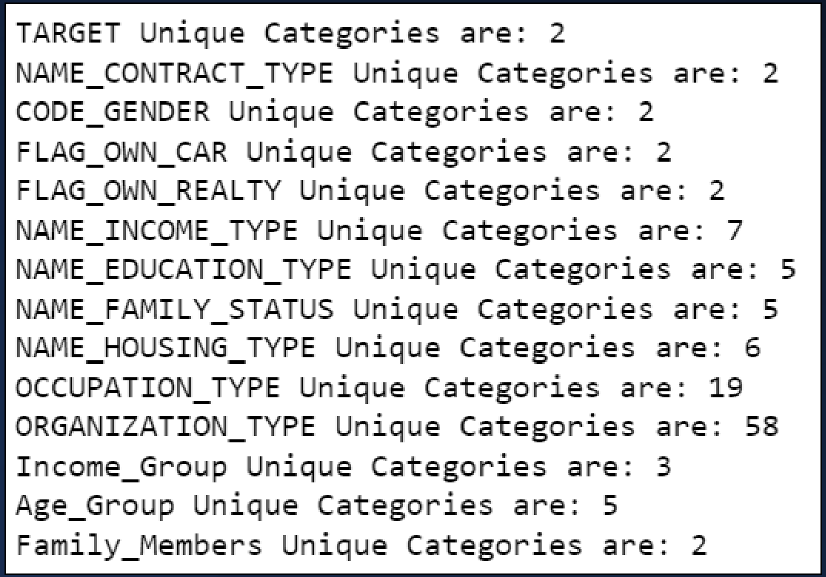


Cardinality and Fixing columns

OCCUPATION_TYPE & OCCUPATION_TYPE shows very high cardinality. Either we need to reduce cardinality or remove these columns to do easy analysis.

OCCUPATION_TYPE column cardinality reduced by making all types of staff categories into single category as ‘Staff’.

Lot many categories in ORGANIZATION_TYPE column. We can reduce of lesser types but its fine if we remove this column as it includes lot many NA data.

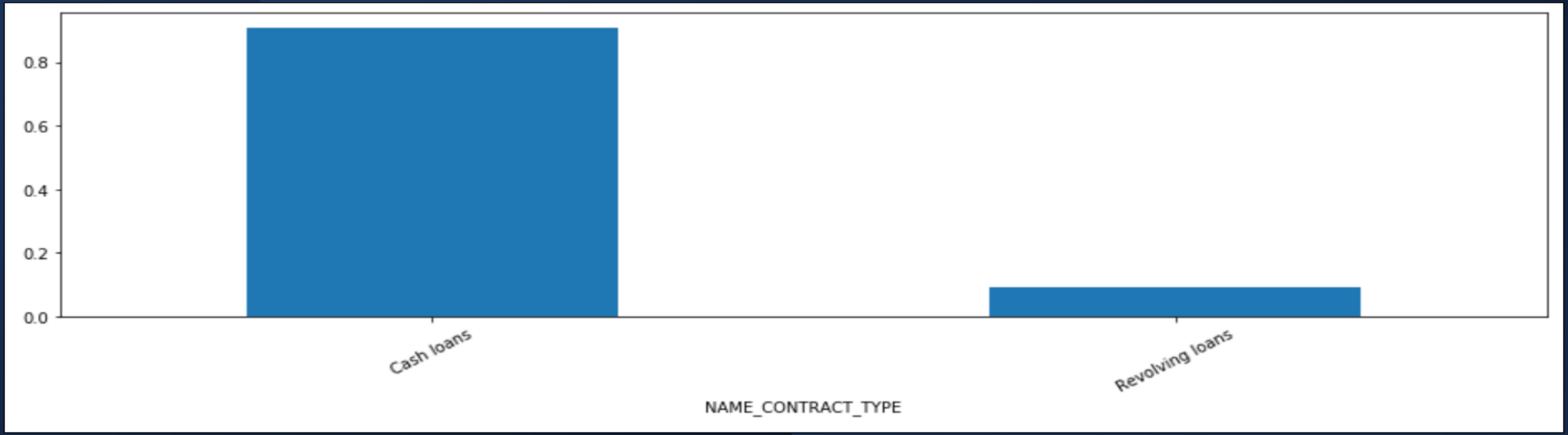


```
TARGET Unique Categories are: 2
NAME_CONTRACT_TYPE Unique Categories are: 2
CODE_GENDER Unique Categories are: 2
FLAG_OWN_CAR Unique Categories are: 2
FLAG_OWN_REALTY Unique Categories are: 2
NAME_INCOME_TYPE Unique Categories are: 7
NAME_EDUCATION_TYPE Unique Categories are: 5
NAME_FAMILY_STATUS Unique Categories are: 5
NAME_HOUSING_TYPE Unique Categories are: 6
OCCUPATION_TYPE Unique Categories are: 19
ORGANIZATION_TYPE Unique Categories are: 58
Income_Group Unique Categories are: 3
Age_Group Unique Categories are: 5
Family_Members Unique Categories are: 2
```

Data Imbalance

Checked the data imbalance in Contract type and found that nearly 90% of the data is for Cash Loans where as only 10% of data is for Revolving Loans.

So as data for Revolving Loans are very less hence recommended to have more data for this category of Loans for better prediction.



Univariate Analysis

After analyzing all different types of variables, below are the observations –

- Nearly 65% applicants are Female where 35% only for Male applicants.
- Nearly 70% of the applicants owns house.
- Majority of the applicants are working stage and nearly 20% are from commercial associate or Pensioner.
- Nearly 70% of the applicants education level is Secondary.
- Nearly 65% applicants are Married and 15% are Single.
- Nearly 85% applicants are living in House/Apartments.
- Most of the applicants are from age 30-40yrs followed by age 40-50yrs and 50-60yrs. Though applicants age 20-30 are also nearly 15% and above 60yrs are nearly about 12%.
- Nearly 70% of applicants having less than 3 family members.
- Client Incomes are positively skewed that means majority of applicants are with lower income level though there are many applicant exists whose income is very high.
- 5Lakhs is the median values of approved given credit. There are some cases where credit given is more than 20Lakhs as well.
- Score given by external sources are mostly below 0.5 for majority of the applicants.
- Majority are from age 34yrs to 53yrs and with mean/median 43yrs. Applicant max age is 69yrs and min age is 20yrs.

Bivariate Analysis

After analyzing all different types of variables with respect to TARGET variable, below are the observations –

- Higher % of risk in case of Male applicants.
- Higher % of risk in case applicants don't have car.
- Higher % of risk in case applicants don't have own house.
- Higher % of risk in case of applicants are Single or did Civil Marriage.
- Higher % of risk in case Medium and Low Income.
- Higher % of risk in case of applicants education level is Lower Secondary.
- Higher % of risk in case of applicants staying at Rented apartment or with parents.
- Higher % of risk in case of applicants are in Maternity leave or unemployed.
- Higher % of risk in case of Cash loans than Revolving loans.
- Higher % of risk in case of applicants belongs to Low-skill Laborers.
- Higher % of risk in case applicants age between 20-30 years.
- Higher % of risk in case of applicants family members are more than 3.
- Higher % of risk in case of lower aged applicants.
- Higher % of risk in case of lower income.
- Higher % of risk doesnt increase with increase in credit amount.
- Higher % of risk in case of lower score given by external companies.

Multivariate Analysis

After analyzing combinations of variables with respect to TARGET variable, below are the observations –

- Male applicant aged between 20-30 years having 13% probability of defaulting in loans.
- Applicant aged 20-30 years with Low income group having 13% probability of defaulting in loans.
- Male applicant who doesn't own car having 13% probability of defaulting in loans.
- Unemployed Male applicant having 70% probability of defaulting in loans.
- Unemployed Female applicant having 50% probability of defaulting in loans.
- Maternity leave applicant having 100% probability of defaulting in loans.
- Male applicant with education level lower secondary having 14% probability of defaulting in loans.
- Male applicant with education level secondary having 12% probability of defaulting in loans.
- Male applicant with marriage status other than 'Married' having ~14% probability of defaulting in loans.
- Low-skill laborers & Realty agents male applicant having ~18% probability of defaulting in loans.
- Low-skill laborers female applicant having 15% probability of defaulting in loans.
- Low-skill laborers with any income having ~17% probability of defaulting in loans.

Recommendation

Based on Exploratory data analysis below recommendation to improve business result –

1. Though the data is limited for Revolving loans but as per available data it seems recovery of this type of loan is more than Cash loan type. Hence to improve profit bank can provide more Revolving loans than cash loans.
2. Female clients loan recovery probability is higher than male clients. Hence bank can approve more female clients as its recovery rate is higher. It will improve bank profit.
3. Unemployed & Maternity leave clients are having highest risk of defaulting loans hence bank should not approve this type of applicants.
4. Male client with 20-30 years and with low income group having high risk of defaulting loans. Hence bank should reduce the approval of such applications.
5. Male client with lower secondary education having high risk of defaulting loans. Hence bank should reduce the approval of such applications.
6. Low skill laborers are high risk of defaulting loans even if client having higher income. Hence bank should reduce the approval of such applications.