CREDIT EDA STUDY

-BY ABHAY TOMER

INTRODUCTION

This assignment aims to give you an idea of applying EDA in a real business scenario. In this assignment, apart from applying the techniques that you have learnt in the EDA module, you will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

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.

To develop your understanding of the domain, you are advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough).

DATASET

This dataset has 3 files as explained below:

1. *'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.

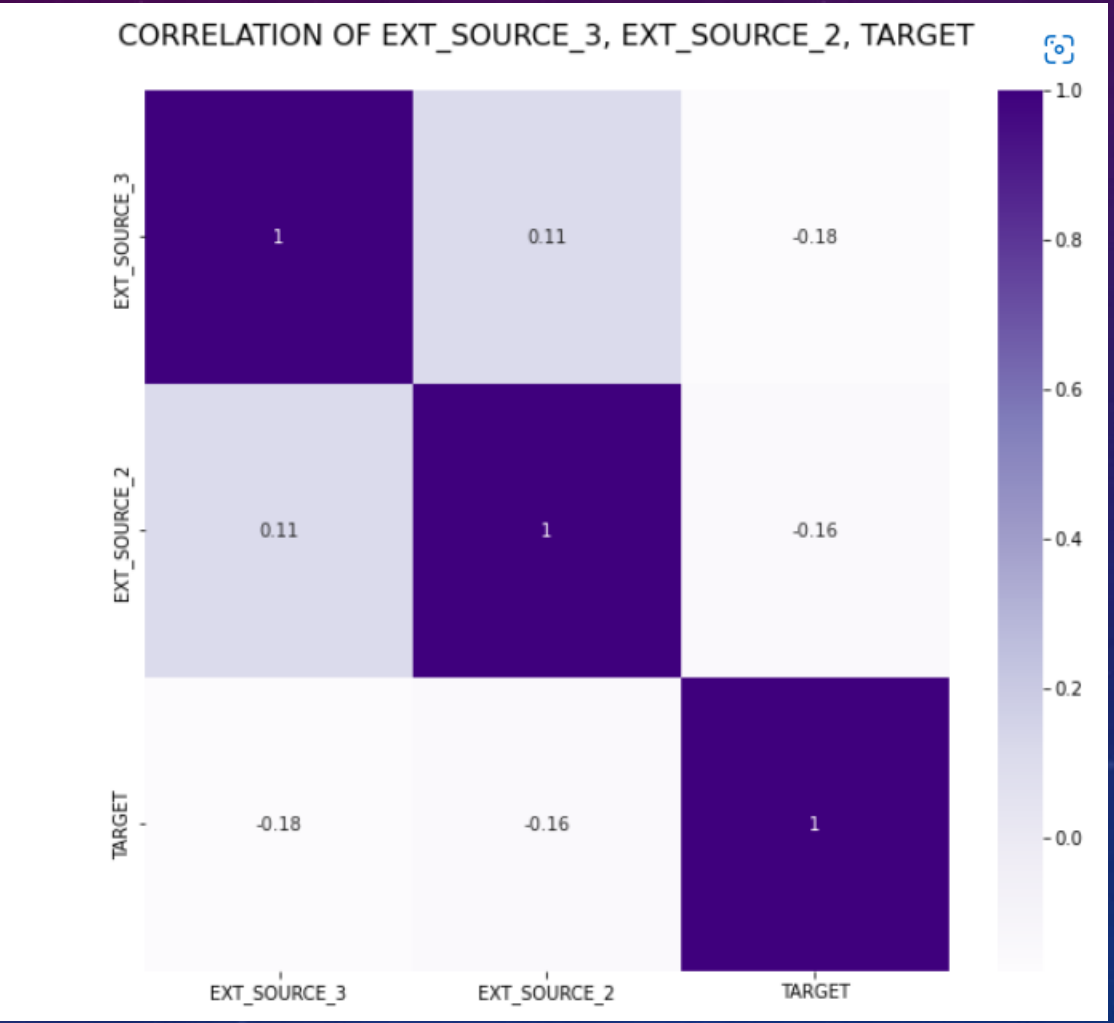
2. *'previous_application.csv'* contains information about the client's previous loan data. It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer.

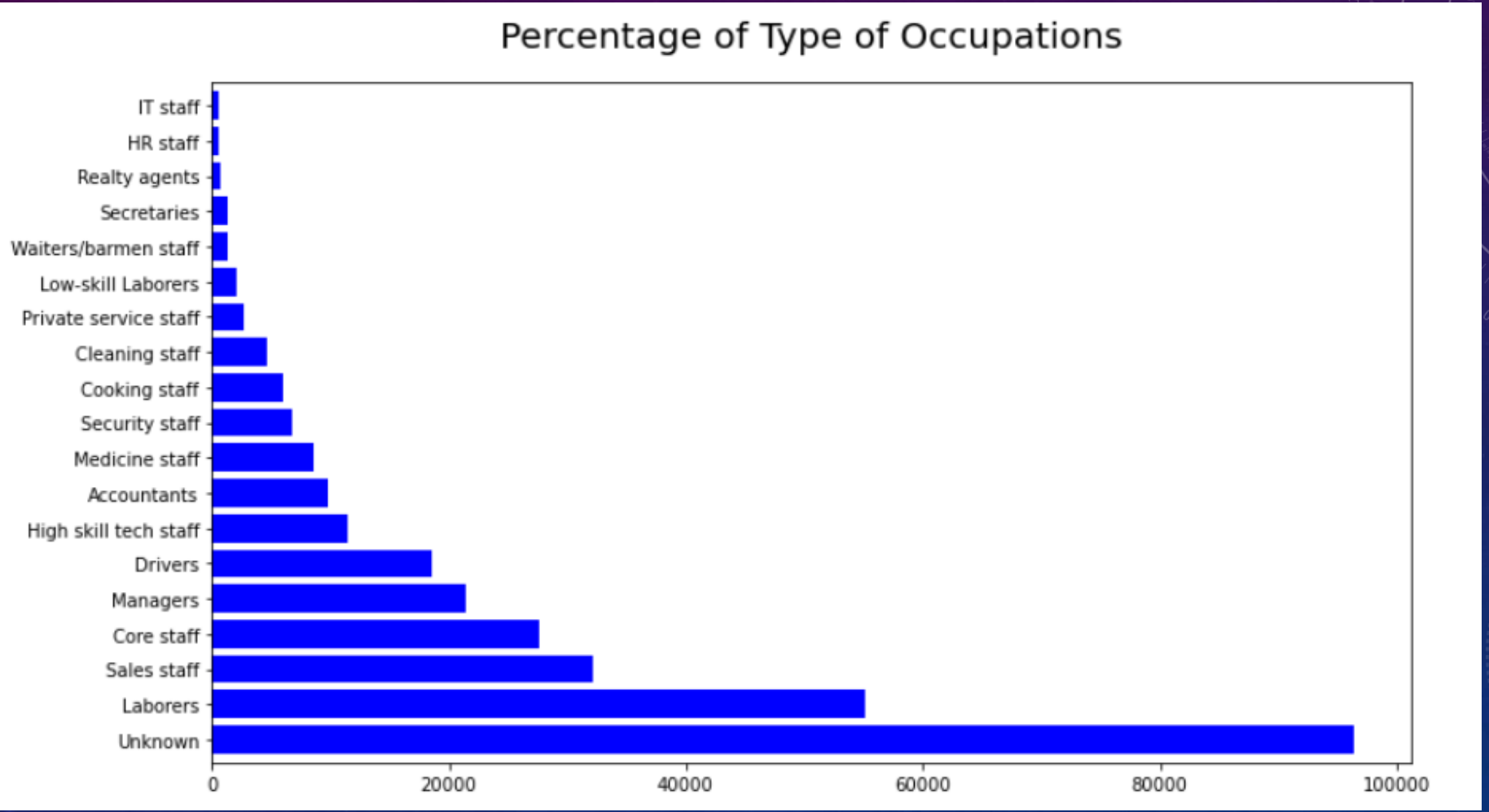
3. *'columns_description.csv'* is data dictionary which describes the meaning of the variables.

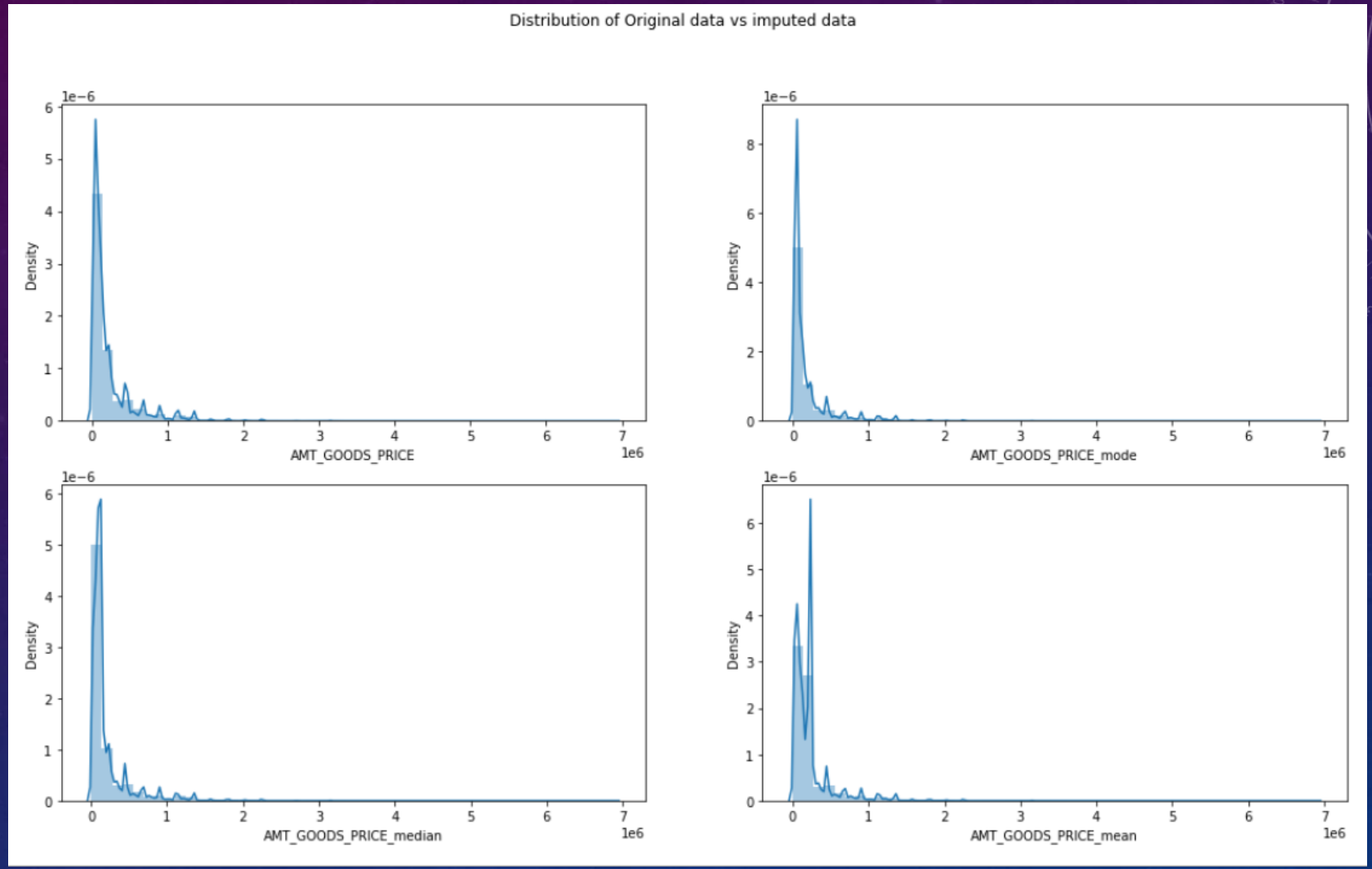
Report an error

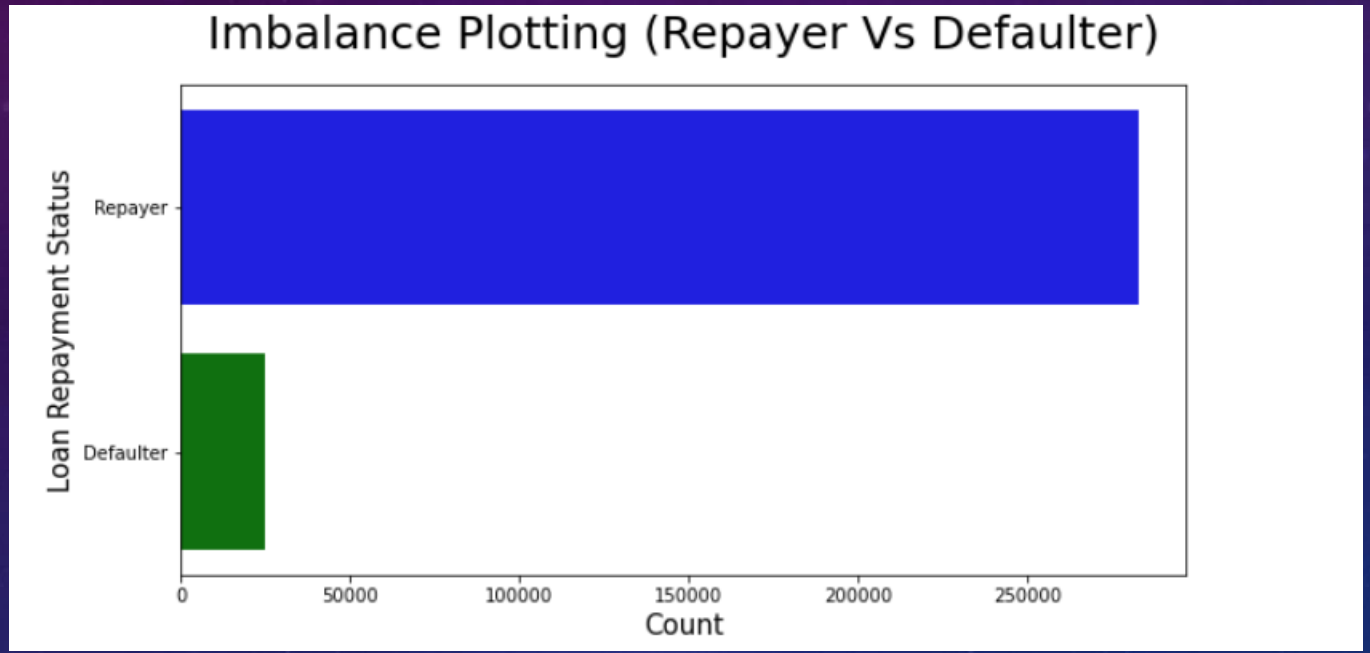
SOME OF THE STEPS

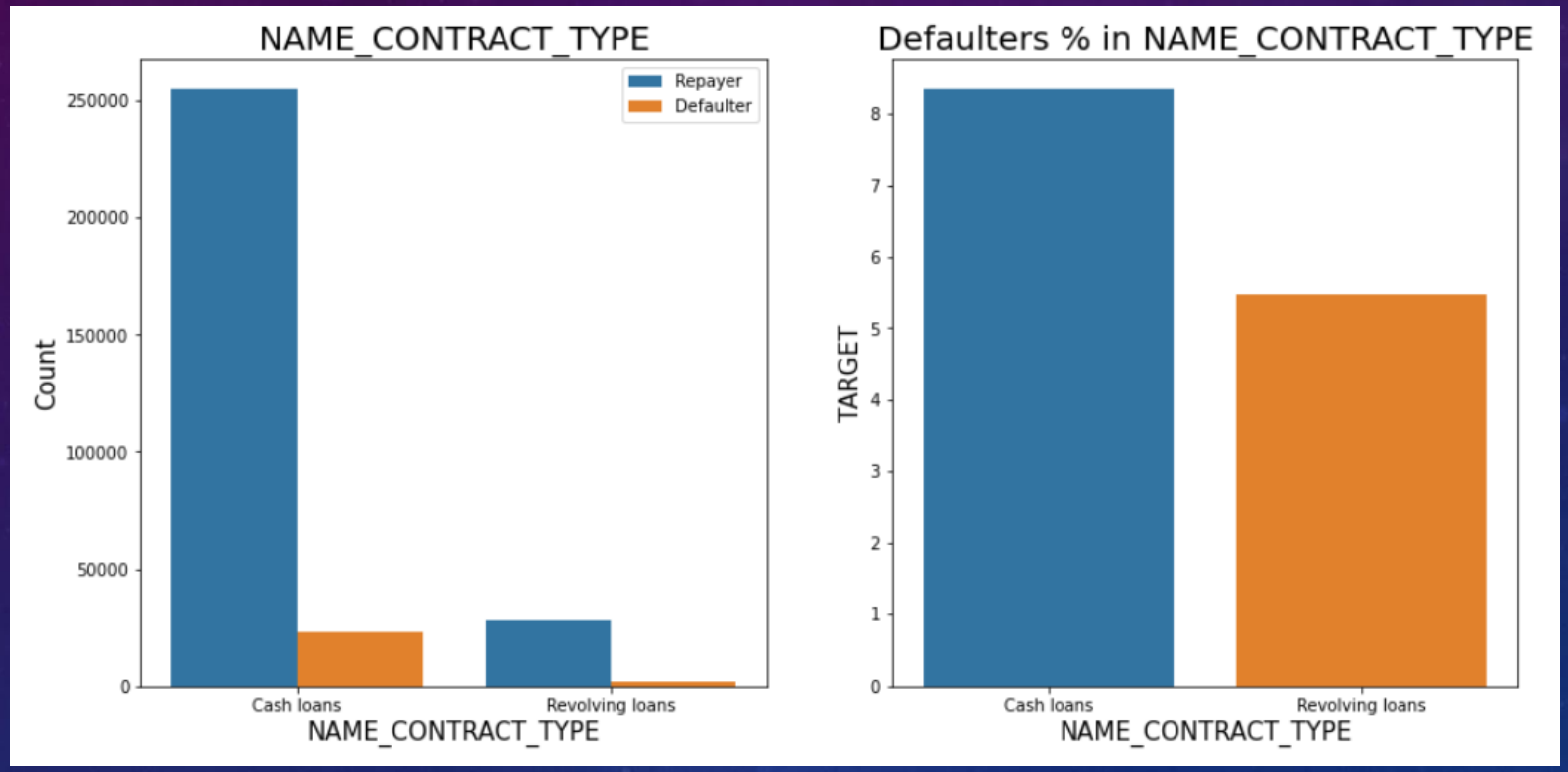
- **Data Cleaning**
- **Fixing the Rows and Columns**
- **Impute/Remove Missing values**
- **Standardize value**
- **Derived Metrics**
- **Univariate Analysis**
- **Bivariate Analysis**

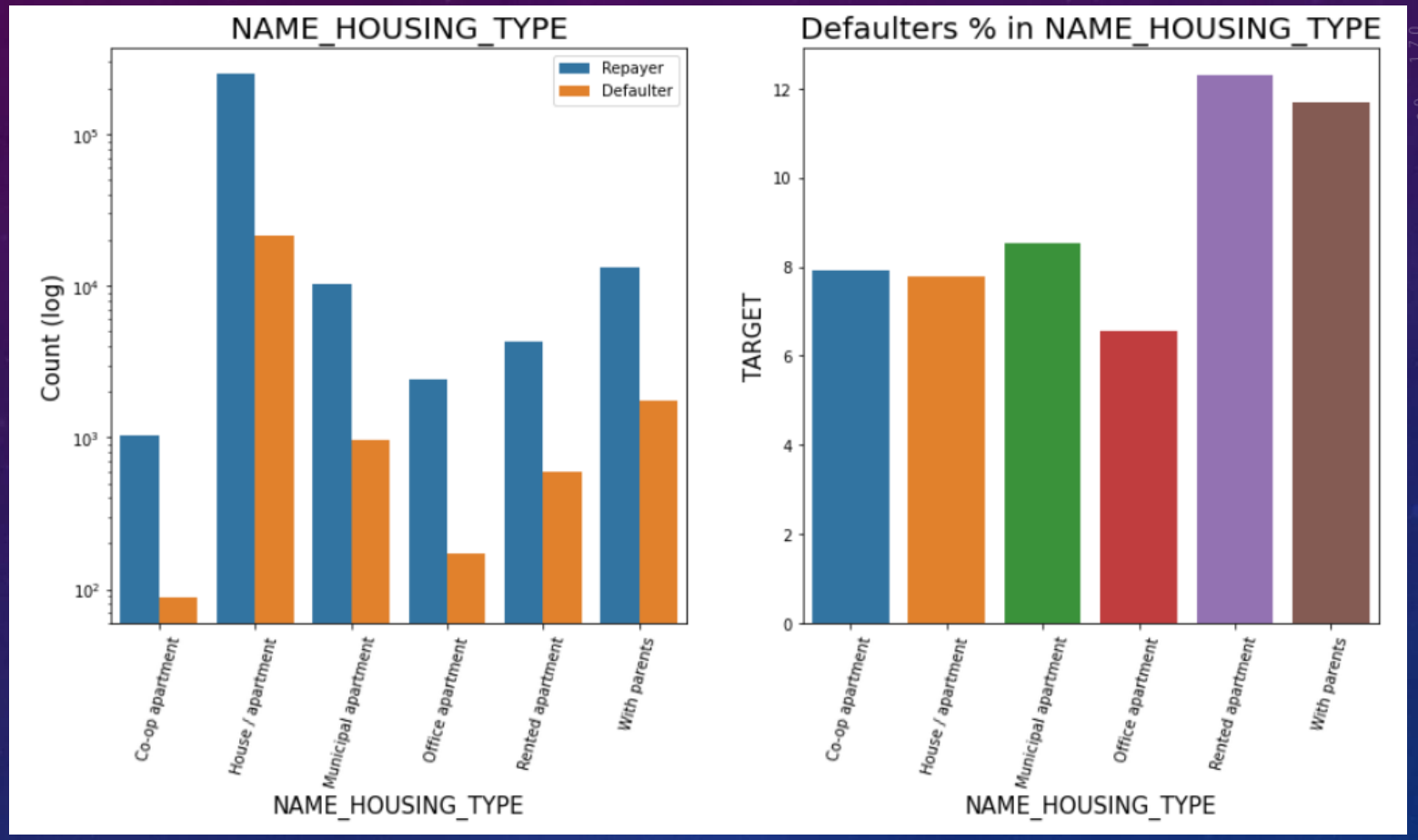

CORRELATION OF EXT_SOURCE_3, EXT_SOURCE_2, TARGET
EXT_SOURCE_3
EXT_SOURCE_2
TARGET
1
0.11
-0.18
0.11
1
-0.16
-0.18
-0.16
1
EXT_SOURCE_3
EXT_SOURCE_2
TARGET
1.0
0.8
0.6
0.4
0.2
0.0

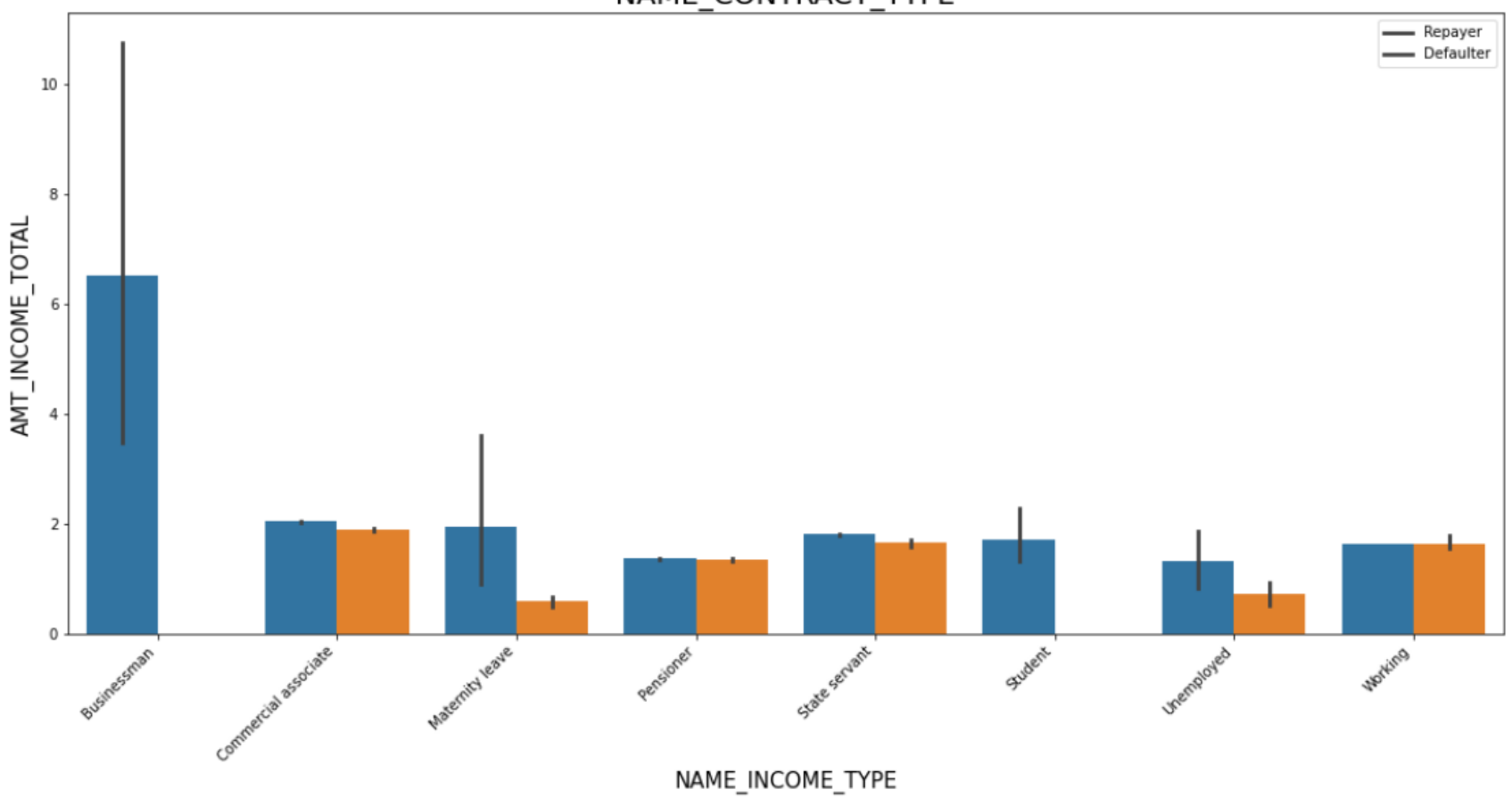

Percentage of Type of Occupations
IT staff
HR staff
Realty agents
Secretaries
Waiters/barmen staff
Low-skill Laborers
Private service staff
Cleaning staff
Cooking staff
Security staff
Medicine staff
Accountants
High skill tech staff
Drivers
Managers
Core staff
Sales staff
Laborers
Unknown
0
20000
40000
60000
80000
100000


Distribution of Original data vs imputed data
1e−6
Density
AMT_GOODS_PRICE
1e6
AMT_GOODS_PRICE_mode
AMT_GOODS_PRICE_median
AMT_GOODS_PRICE_mean

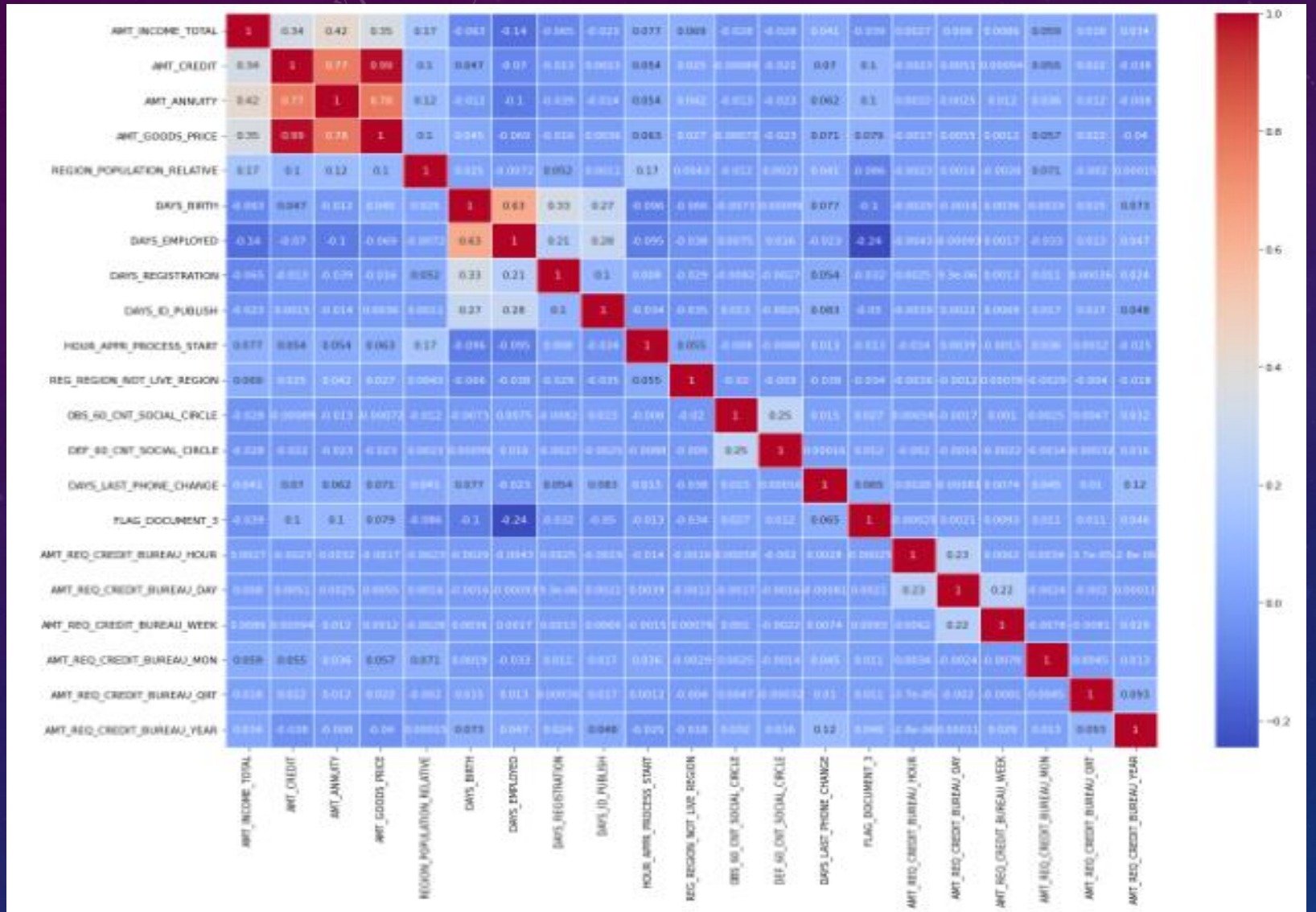

Imbalance Plotting (Repayer Vs Defaulter)
Loan Repayment Status
Repayer
Defaulter
0
50000
100000
150000
200000
250000
Count


NAME_CONTRACT_TYPE
Repayer
Defaulter
250000
200000
150000
100000
50000
0
Count
Cash loans
Revolving loans
NAME_CONTRACT_TYPE
Defaulters % in NAME_CONTRACT_TYPE
8
7
6
5
4
3
2
1
0
TARGET
Cash loans
Revolving loans
NAME_CONTRACT_TYPE


NAME_HOUSING_TYPE
Repayer
Defaulter
Count (log)
10^5
10^4
10^3
10^2
Co-op apartment
House / apartment
Municipal apartment
Office apartment
Rented apartment
With parents
NAME_HOUSING_TYPE
Defaulters % in NAME_HOUSING_TYPE
TARGET
12
10
8
6
4
2
0
Co-op apartment
House / apartment
Municipal apartment
Office apartment
Rented apartment
With parents
NAME_HOUSING_TYPE



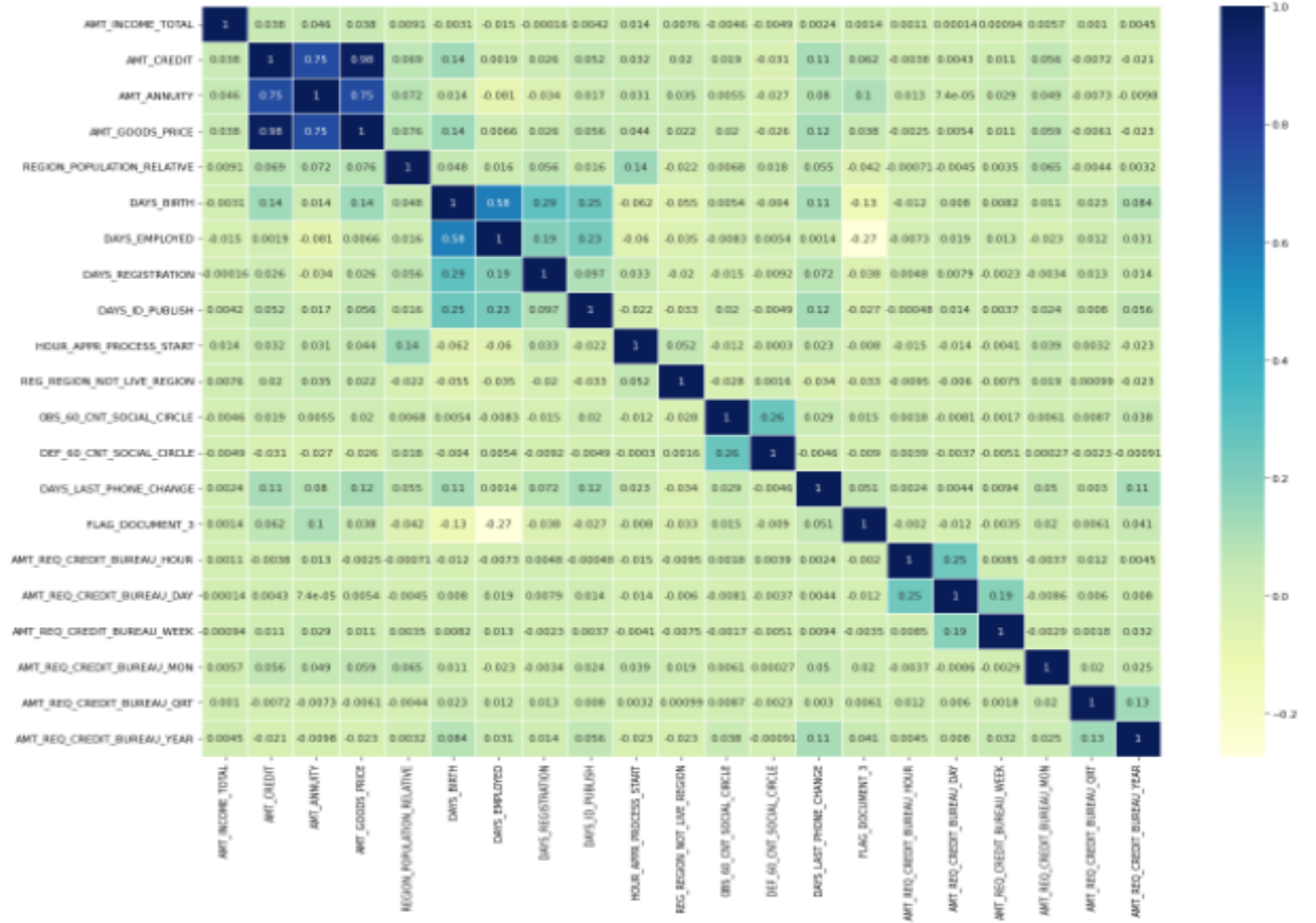
- **It can be seen that Businessman income is the highest and the estimated range with default 95% confidence level seem to indicate that the income of a Businessman could be in the range of slightly close to 4 lakhs and slightly above 10 lakhs**



Correlating factors amongst repayers

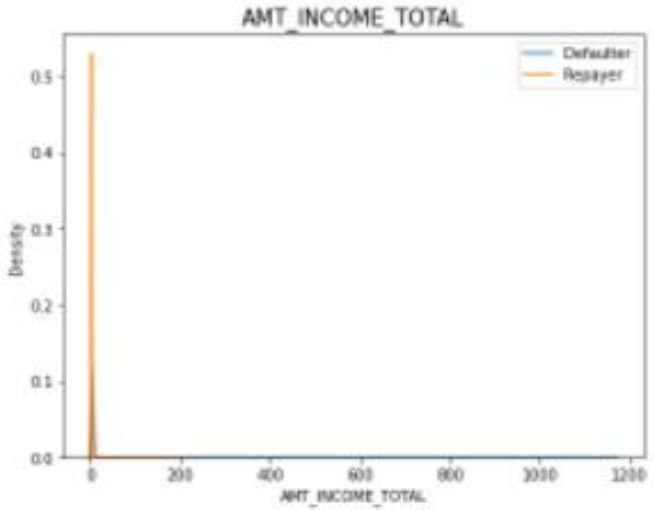
Credit amount is highly correlated with:

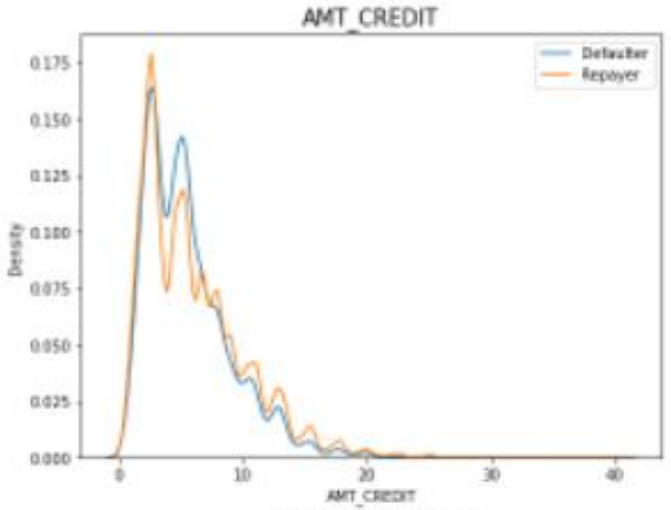
- Goods Price Amount
- Loan Annuity
- Total Income

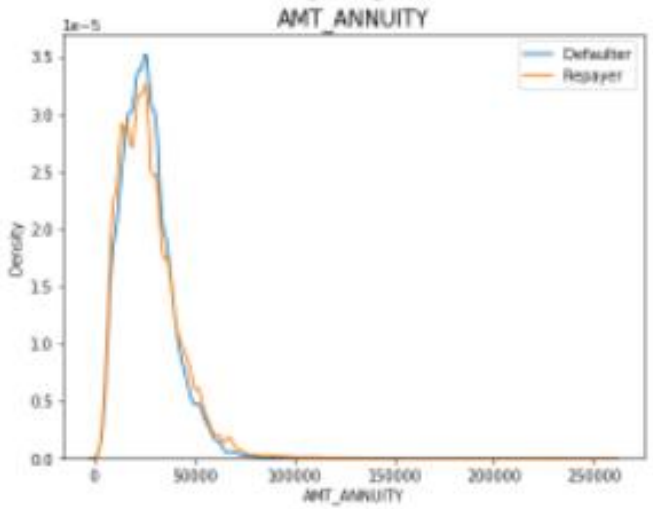


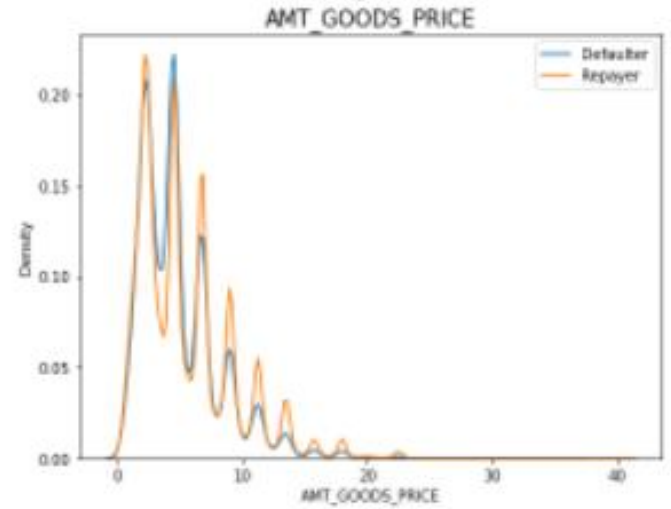
Correlating factors amongst repayers

- Credit amount is highly correlated with good price amount which is same as repayers.
- Loan annuity correlation with credit amount has slightly reduced in defaulters(0.75) when compared to repayers(0.77)
- We can also see that repayers have high correlation in number of days employed(0.62) when compared to defaulters(0.58).
- There is a severe drop in the correlation between total income of the client and the credit amount(0.038) amongst defaulters whereas it is 0.342 among repayers.
- Days_birth and number of children correlation has reduced to 0.259 in defaulters when compared to 0.337 in repayers.
- There is a slight increase in defaulted to observed count in social circle among defaulters(0.264) when compared to repayers(0.254)

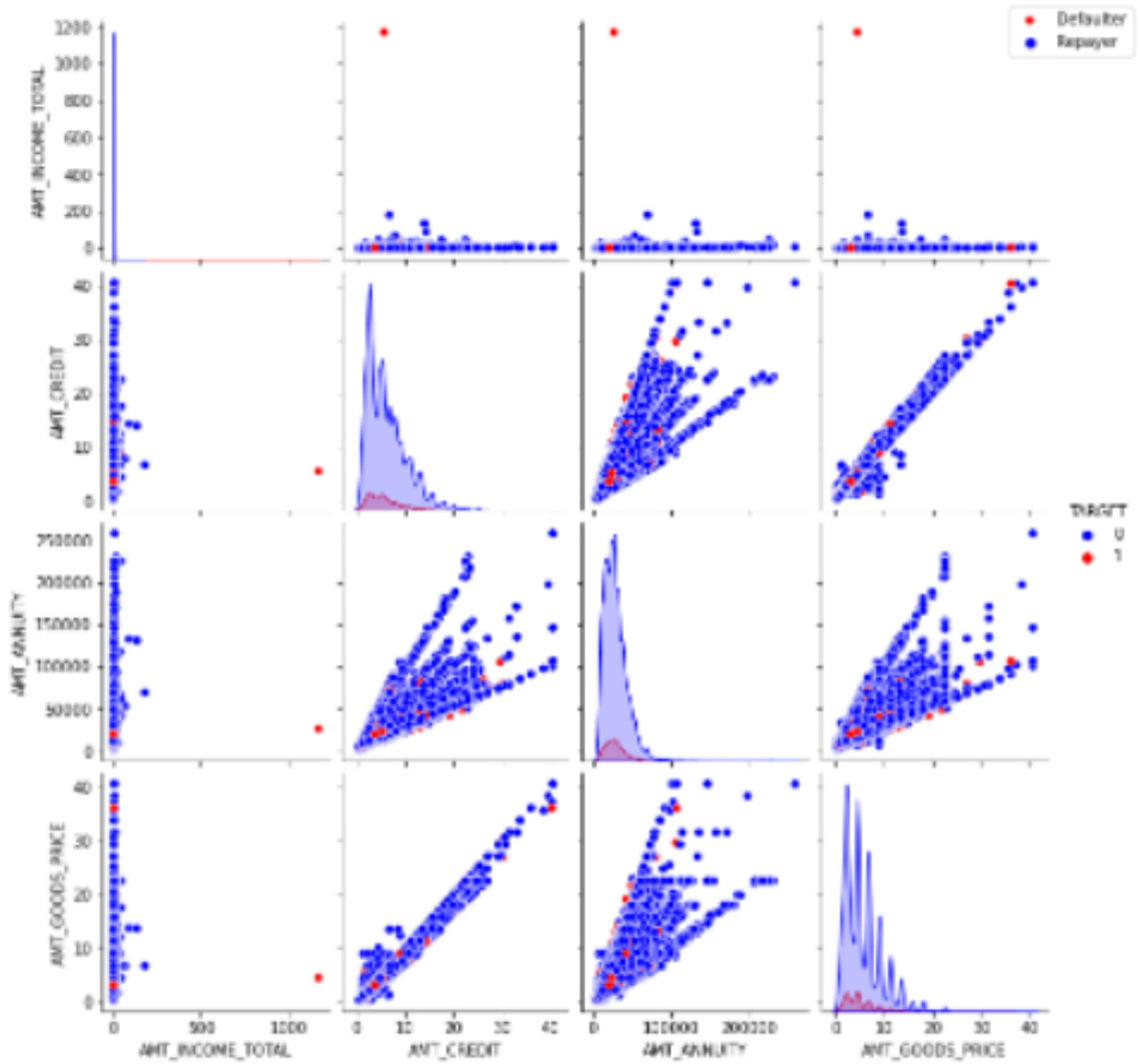




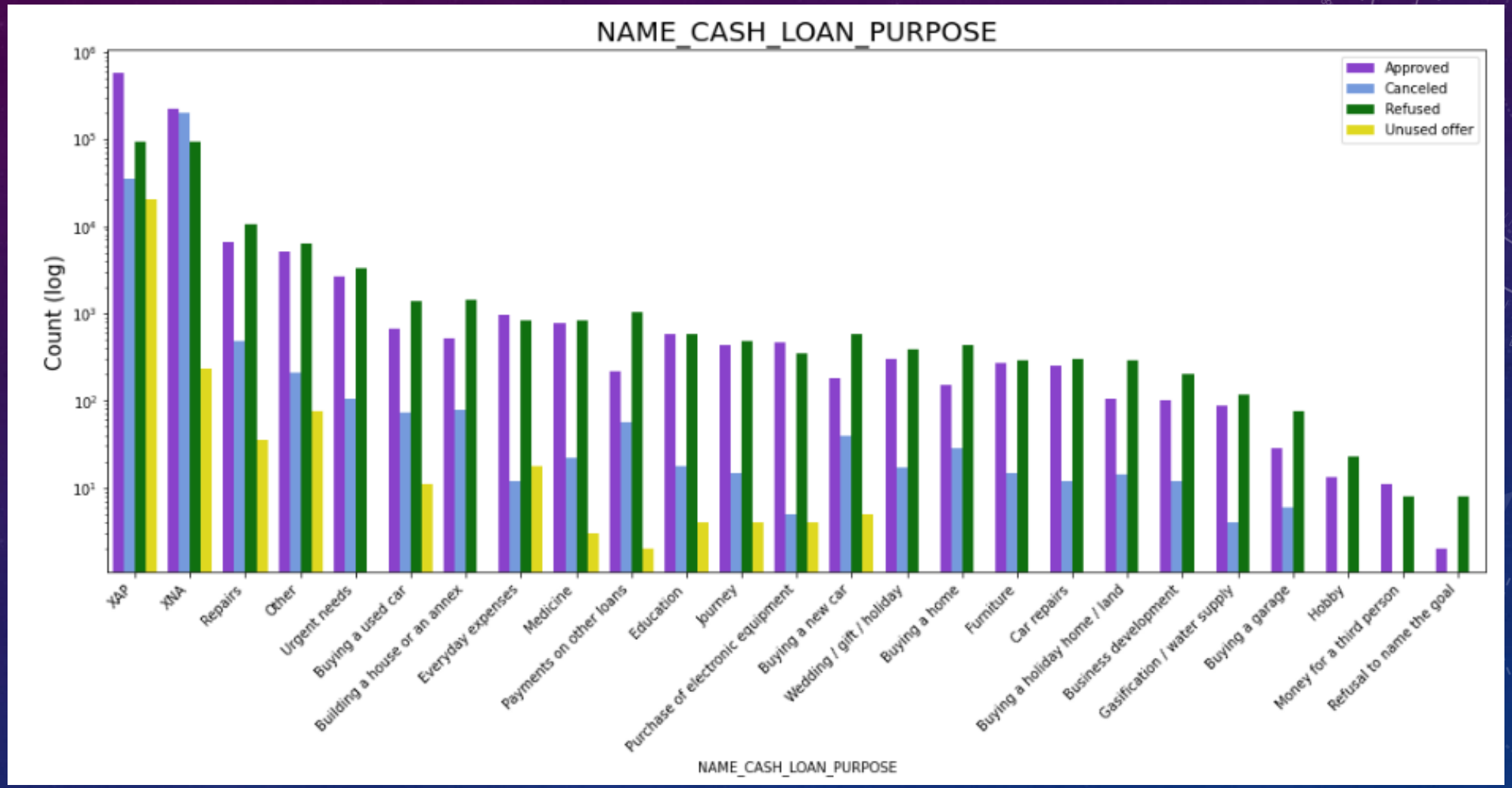


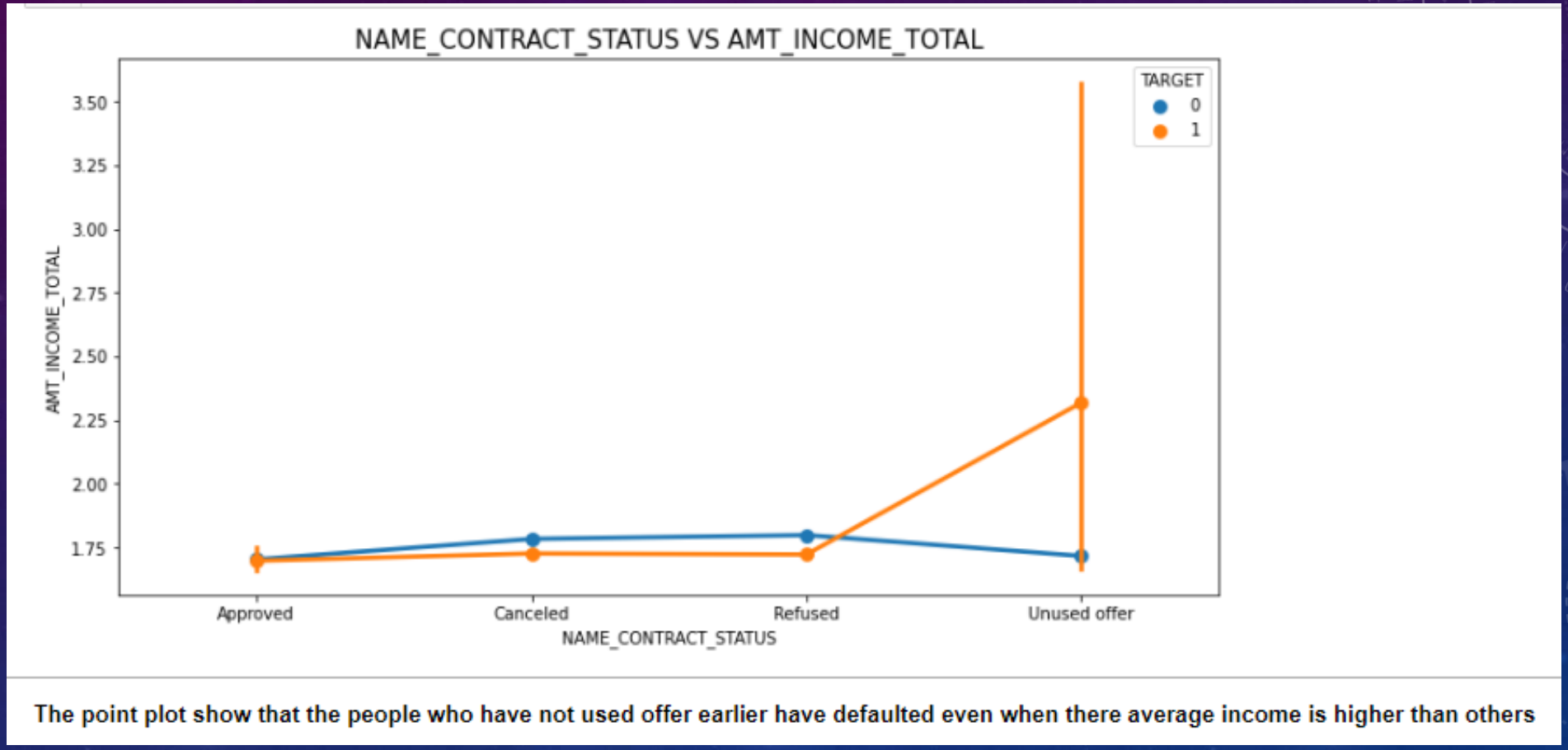


- Most number of loans are given for goods price below 10 lakhs
- Most people pay annuity below 50K for the credit loan
- Credit amount of the loan is mostly less then 10 lakhs
- The repayers and defaulters distribution overlap in all the plots and hence we cannot use any of these variables in isolation to make a decision



- When Annuity Amount > 15K and Good Price Amount > 20 Lakhs, there is a lesser chance of defaulters
- Loan Amount(AMT_CREDIT) and Goods price(AMT_GOODS_PRICE) are highly correlated as based on the scatterplot where most of the data are consolidated in form of a line
- There are very less defaulters for AMT_CREDIT >20 Lakhs


NAME_CASH_LOAN_PURPOSE
Count (log)
Approved
Canceled
Refused
Unused offer
XAP
XNA
Repairs
Other
Urgent needs
Buying a used car
Building a house or an annex
Everyday expenses
Medicine
Payments on other loans
Education
Journey
Purchase of electronic equipment
Buying a new car
Wedding / gift / holiday
Buying a home
Furniture
Car repairs
Buying a holiday home / land
Business development
Gasification / water supply
Buying a garage
Hobby
Money for a third person
Refusal to name the goal
NAME_CASH_LOAN_PURPOSE



The point plot show that the people who have not used offer earlier have defaulted even when there average income is higher than others

CONCLUSION

After analyzing the datasets, there are few attributes of a client with which the bank would be able to identify if they will repay the loan or not. The analysis is consisted as below with the contributing factors and categorization:

- A. Decisive Factor whether an applicant will be Repayer:
- NAME_EDUCATION_TYPE: Academic degree has less defaults.
- NAME_INCOME_TYPE: Student and Businessmen have no defaults.
- REGION_RATING_CLIENT: RATING 1 is safer.
- ORGANIZATION_TYPE: Clients with Trade Type 4 and 5 and Industry type 8 have defaulted less than 3%
- DAYS_BIRTH: People above age of 50 have low probability of defaulting
- DAYS_EMPLOYED: Clients with 40+ year experience having less than 1% default rate
- AMT_INCOME_TOTAL: Applicant with Income more than 700,000 are less likely to default
- NAME_CASH_LOAN_PURPOSE: Loans bought for Hobby, Buying garage are being repayed mostly.
- CNT_CHILDREN: People with zero to two children tend to repay the loans.

B. Decisive Factor whether an applicant will be Defaulter:

- CODE_GENDER: Men are at relatively higher default rate
- NAME_FAMILY_STATUS : People who have civil marriage or who are single default a lot.
- NAME_EDUCATION_TYPE: People with Lower Secondary & Secondary education
- NAME_INCOME_TYPE: Clients who are either at Maternity leave OR Unemployed default a lot.
- REGION_RATING_CLIENT: People who live in Rating 3 has highest defaults.
- OCCUPATION_TYPE: Avoid Low-skill Laborers, Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff as their default rate is huge.
- ORGANIZATION_TYPE: Organizations with highest percent of loans not repaid are Transport: type 3 (16%), Industry: type 13 (13.5%), Industry: type 8 (12.5%) and Restaurant (less than 12%). Self-employed people have relative high defaulting rate, and thus should be avoided to be approved for loan or provide loan with higher interest rate to mitigate the risk of defaulting.
- DAYS_BIRTH: Avoid young people who are in age group of 20-40 as they have higher probability of defaulting
- DAYS_EMPLOYED: People who have less than 5 years of employment have high default rate.
- CNT_CHILDREN & CNT_FAM_MEMBERS: Client who have children equal to or more than 9 default 100% and hence their applications are to be rejected.
- AMT_GOODS_PRICE: When the credit amount goes beyond 3lakhs, there is an increase in defaulters.

C. Factors that Loan can be given on Condition of High Interest rate to mitigate any default risk leading to business loss:

- NAME_HOUSING_TYPE: High number of loan applications are from the category of people who live in Rented apartments & living with parents and hence offering the loan would mitigate the loss if any of those default.
- AMT_CREDIT: People who get loan for 3-6 Lakhs tend to default more than others and hence having higher interest specifically for this credit range would be ideal.
- AMT_INCOME: Since 90% of the applications have Income total less than 3Lakhs and they have high probability of defaulting, they could be offered loan with higher interest compared to other income category.
- CNT_CHILDREN & CNT_FAM_MEMBERS: Clients who have 4 to 8 children has a very high default rate and hence higher interest should be imposed on their loans.
- NAME_CASH_LOAN_PURPOSE: Loan taken for the purpose of Repairs seems to have highest default rate. A very high number applications have been rejected by bank or refused by client in previous applications as well which has purpose as repair or other. This shows that purpose repair is taken as high risk by bank and either they are rejected, or bank offers very high loan interest rate which is not feasible by the clients, thus they refuse the loan. The same approach could be followed in future as well.

SUGGESTIONS:

- **90% of the previously cancelled client have actually repayed the loan. Record the reason for cancellation which might help the bank to determine and negotiate terms with these repaying customers in future for increase business opportunity.**
- **88% of the clients who were refused by bank for loan earlier have now turned into a repaying client. Hence documenting the reason for rejection could mitigate the business loss and these clients could be contacted for further loans.**