CREDIT RISK ANALYSIS

EDA CASE STUDY

By Kratika Jain

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.

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.

Credit risk analysis will help the company to make a decision for loan approval based on the applicant's profile. Which controls loss of business to the company and avoid financial loss for the company.

Steps to be taken

1.Data understanding and sourcing

2.check for Data quality issues and Binning

3.check for Data imbalance and univariate,segmented univariate & Bivariate analysis.

4.Merging of application data with previous application data.

5.Data analysis by univariate, segmented univariate , Bivariate analysis and correlation.

6.Conclusion

Data Understanding

There are two data sets present which are :

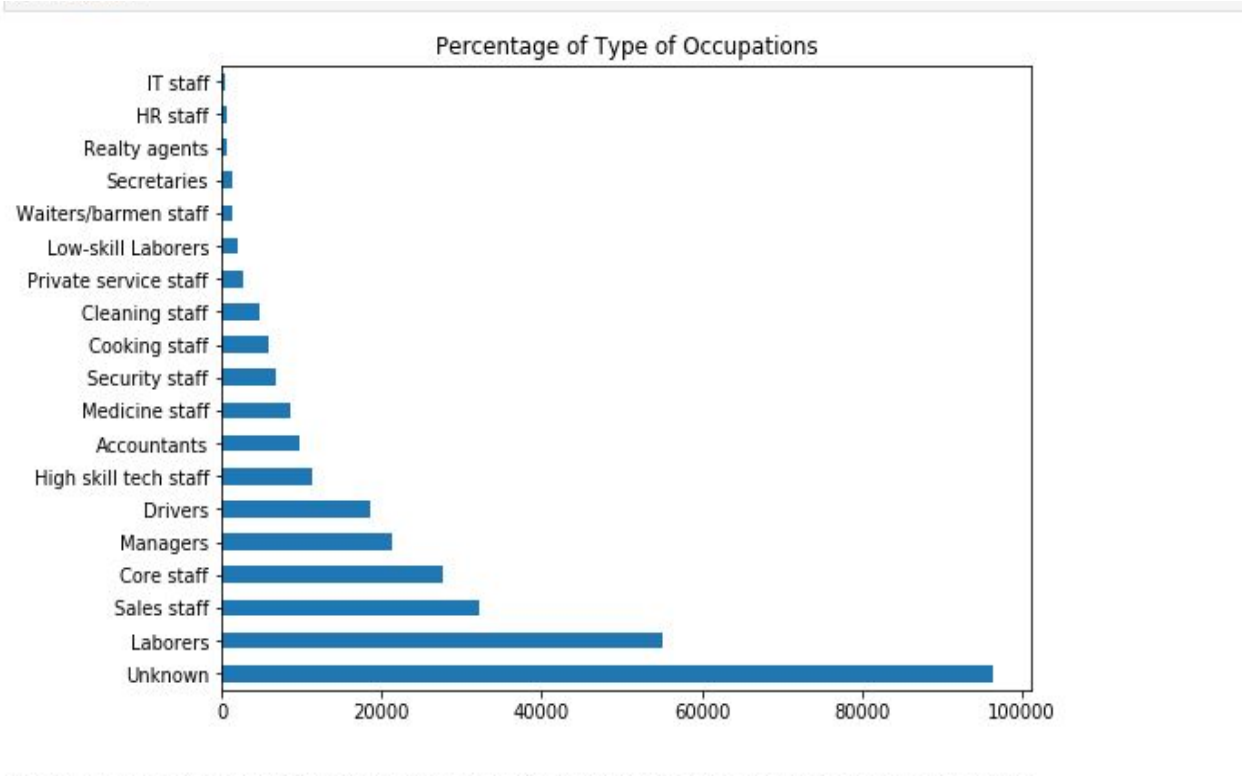
1)Application Data

2)Previous Application data

Application data have 307511 rows and 122 columns. And have 41 columns which have more than 50% null values.We will drop those columns

Previous Application data have 1670214rows and 37 columns. And have 4 columns which have more than 50% null values.We will drop those columns

This graph presents the percentage of type of occupation.It gives us an insight that we have unknown values at first and Laborers as second for occupation

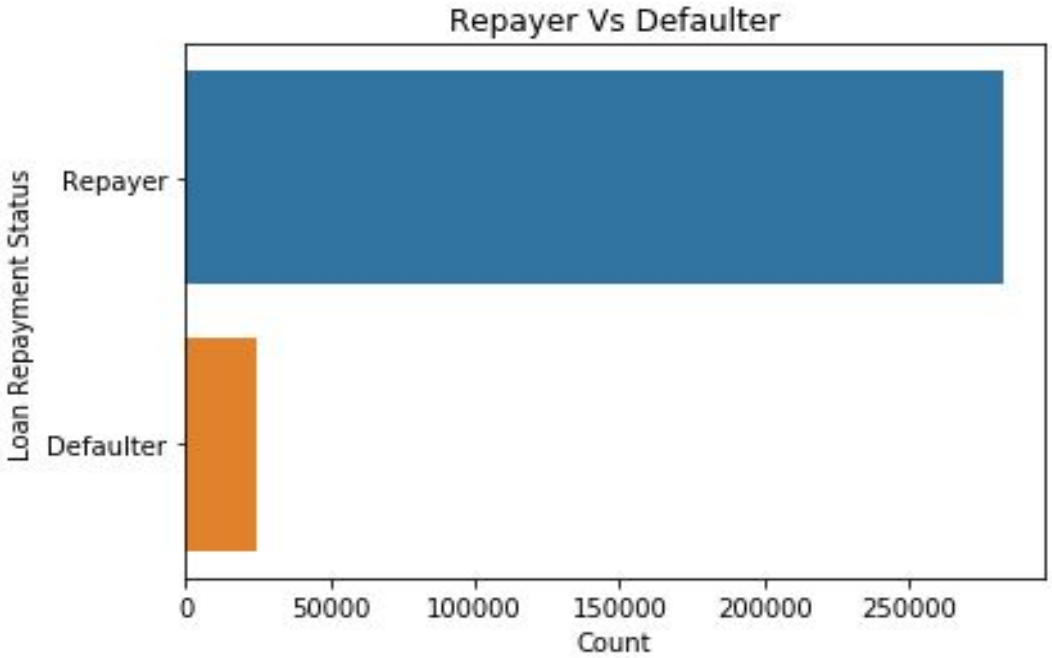


Imbalance Data

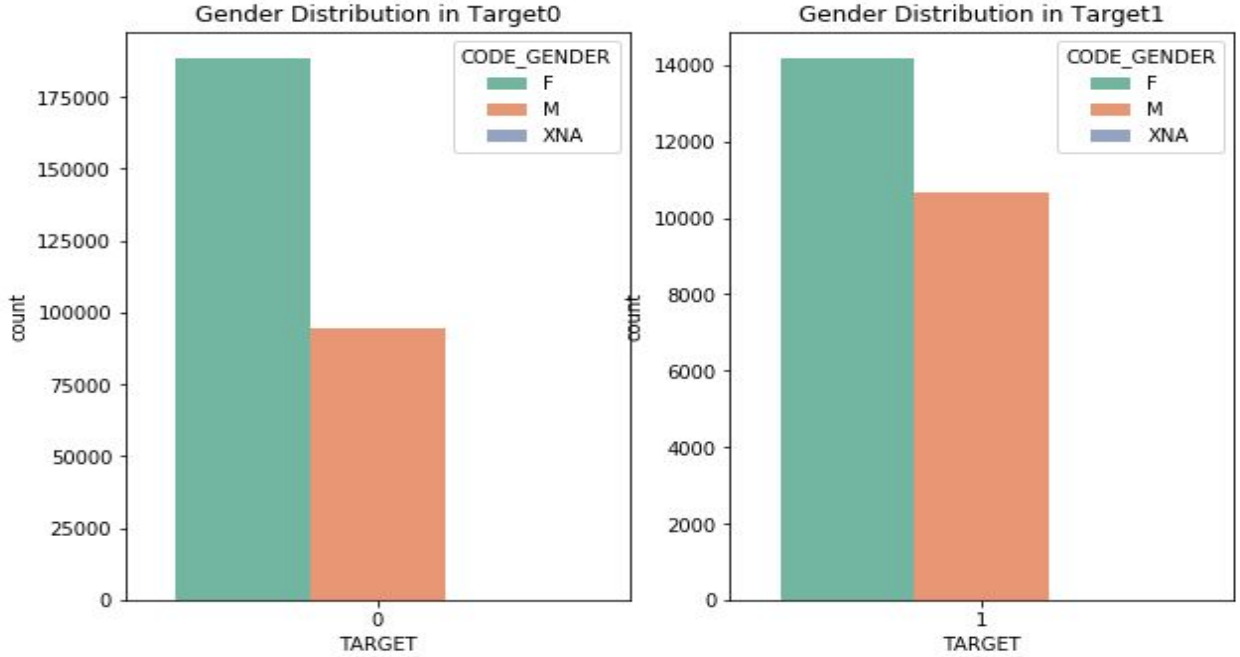
The Repayer percentage is 91.93%

The Defaulter percentage is 8.07%

The imbalance ratio is 11.39



Univariate Analysis

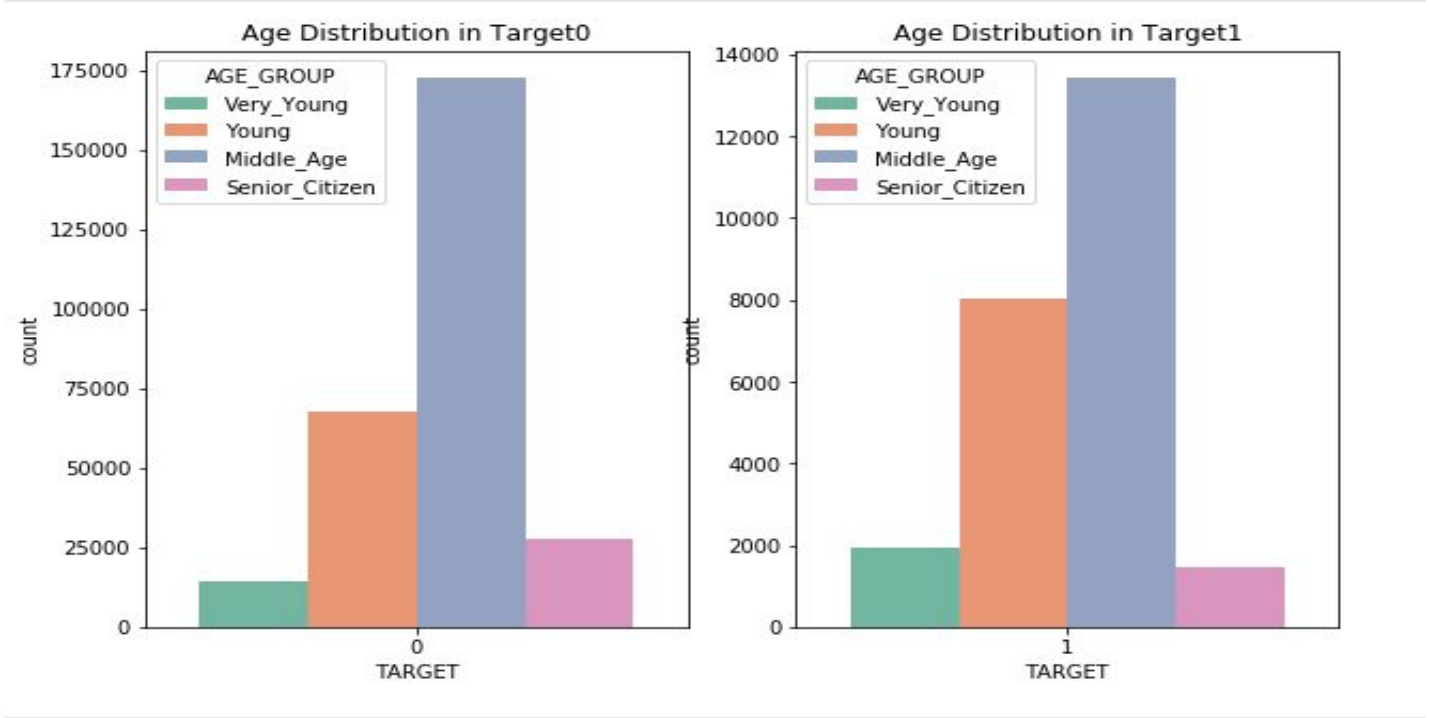


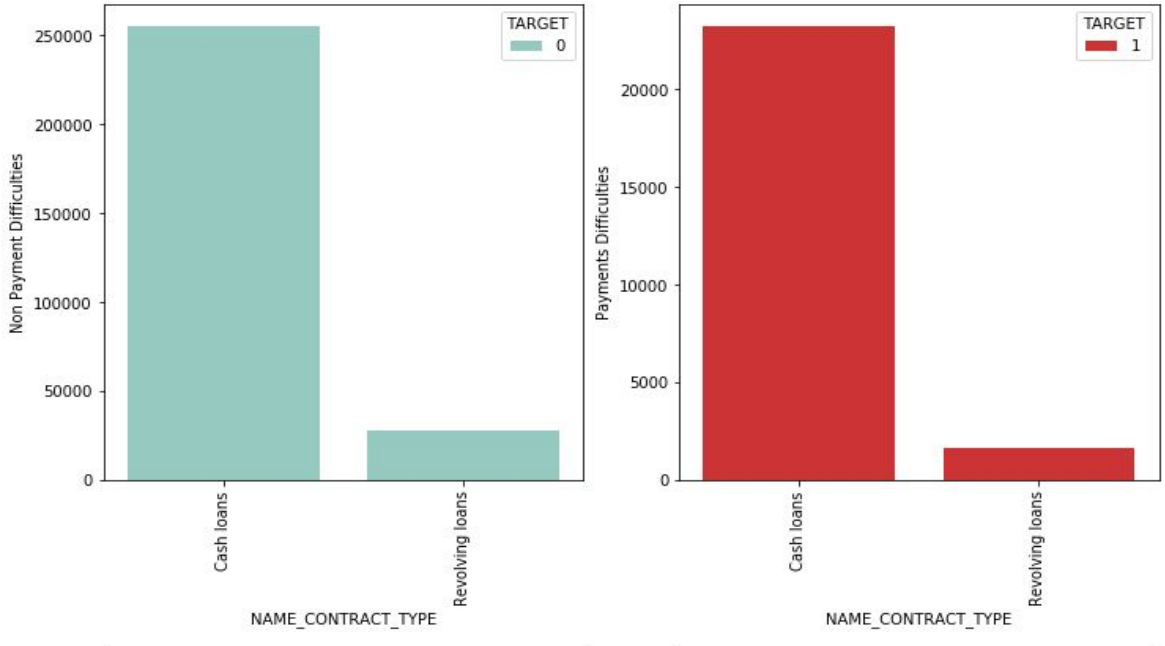
We can see that Female clients applied higher than male clients for loan.

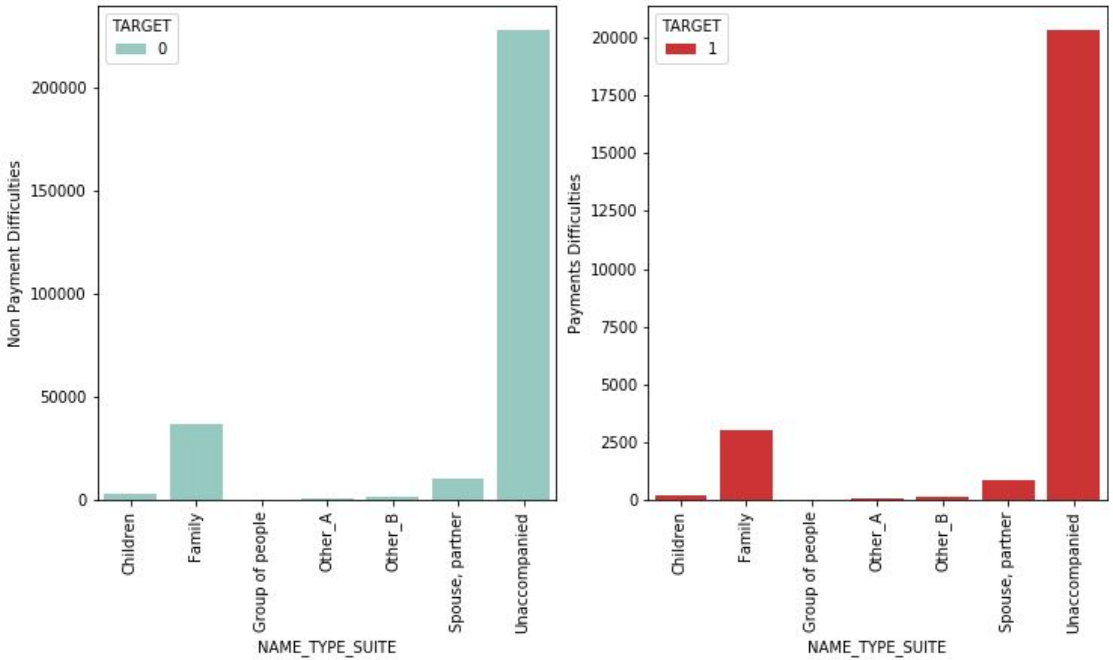
Age Distribution

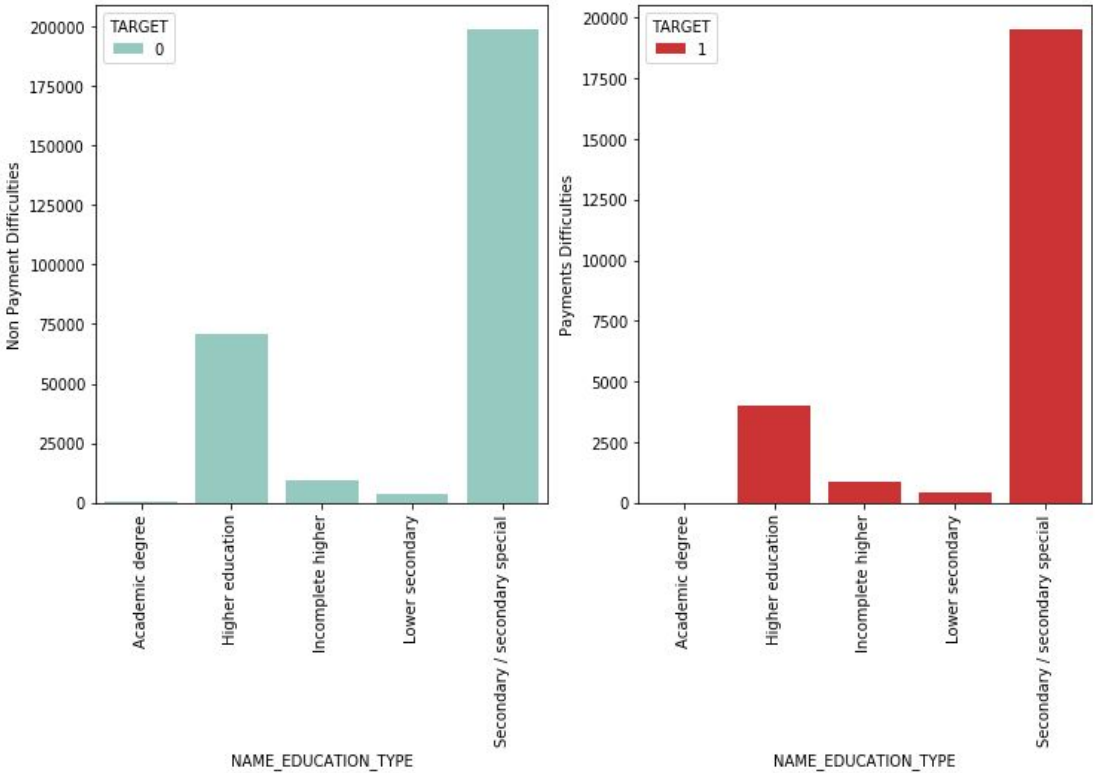
Middle Age(35-60) the group have applied higher than any other age group for loans in the case of Defaulters as well as Non-defaulters.

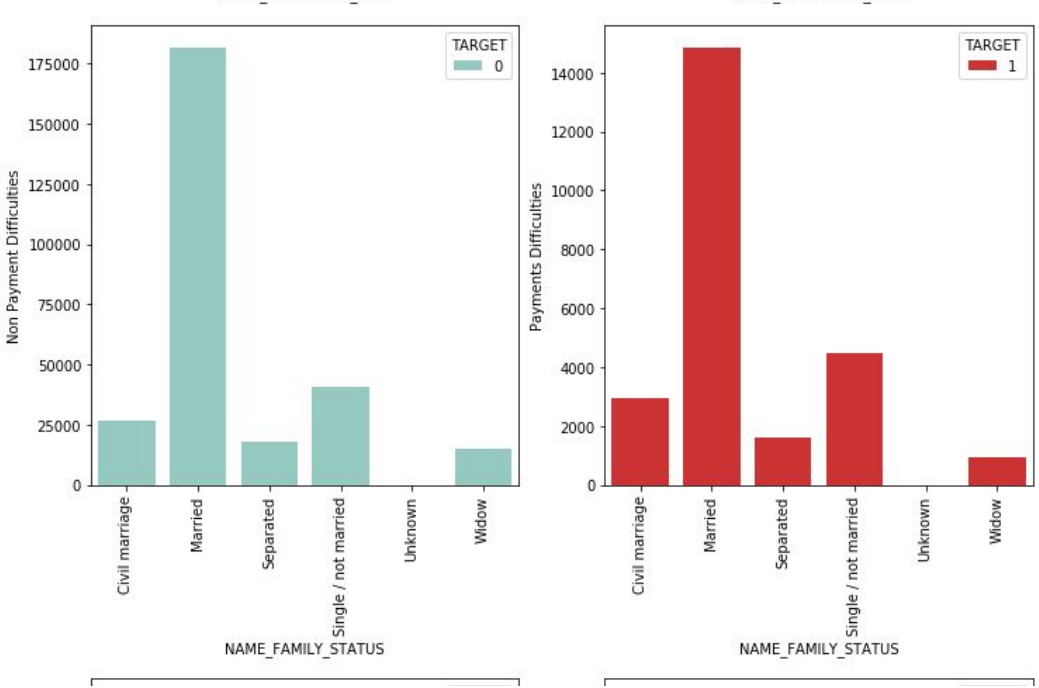
Very Young and Senior citizens have applied very less for loan.




TARGET
0
Non Payment Difficulties
250000
200000
150000
100000
50000
0
Cash loans
Revolving loans
NAME_CONTRACT_TYPE
TARGET
1
Payments Difficulties
20000
15000
10000
5000
0
Cash loans
Revolving loans
NAME_CONTRACT_TYPE


TARGET
0
Non Payment Difficulties
200000
150000
100000
50000
0
Children
Family
Group of people
Other_A
Other_B
Spouse, partner
Unaccompanied
NAME_TYPE_SUITE
TARGET
1
Payments Difficulties
20000
17500
15000
12500
10000
7500
5000
2500
0
Children
Family
Group of people
Other_A
Other_B
Spouse, partner
Unaccompanied
NAME_TYPE_SUITE


TARGET
0
Non Payment Difficulties
200000
175000
150000
125000
100000
75000
50000
25000
0
Academic degree
Higher education
Incomplete higher
Lower secondary
Secondary / secondary special
NAME_EDUCATION_TYPE
TARGET
1
Payments Difficulties
20000
17500
15000
12500
10000
7500
5000
2500
0
Academic degree
Higher education
Incomplete higher
Lower secondary
Secondary / secondary special
NAME_EDUCATION_TYPE


TARGET
0
Non Payment Difficulties
175000
150000
125000
100000
75000
50000
25000
0
Civil marriage
Married
Separated
Single / not married
Unknown
Widow
NAME_FAMILY_STATUS
TARGET
1
Payments Difficulties
14000
12000
10000
8000
6000
4000
2000
0
Civil marriage
Married
Separated
Single / not married
Unknown
Widow
NAME_FAMILY_STATUS

By all this univariate analysis we have some insights :

1)Most of the clients have applied for Cash Loan while very less have applied for Revolving loan for both Defaulters as well as Non-defaulters.

2)AMT_GOODS_PRICE with medium value have highest applied for loan for Non defaulters and for defaulters both low and medium values applied for loan.

3) Most of the clients were unaccompanied while applying for the loan and with few clients a family member was accompanying for both Defaulters and Non-Defaulters. But who was accompanying client while applying for the loan doesn't impact on the default.Also both the populations have same proportions.

4)Working,Commercial associate and Pensioner are more likely to apply for the loan, highest being the Working class category. Businessman, students and Unemployed Less likely to apply for loan . Working category have high risk to default. State Servant is at Minimal risk to default.

5)Very few clients have an academic degree. Secondary or Secondary Special education are more likey to apply for the loan and they are at higher risk to default.People with Academic degree are least likely to default.

6)Married person seems to be applied most for the loan compared to others for both Defaulters and Non-Defaulters. In case of Defaulters,people having single relationship or widows are less risky.

7)Most of the clients own a house or living in a apartment for both Defaulters and Non-Defaulters.

8)There is no noticable difference in days for both Defaulters and Non-defaulters.

9)Medium salary range people are more likely to apply for the loan for both Defaulters and Non-defaulters. people having low and medium income are at high risk to default.

10)people applied for Medium Credit Amount of the loan for both Defaulters and Non-defaulters.

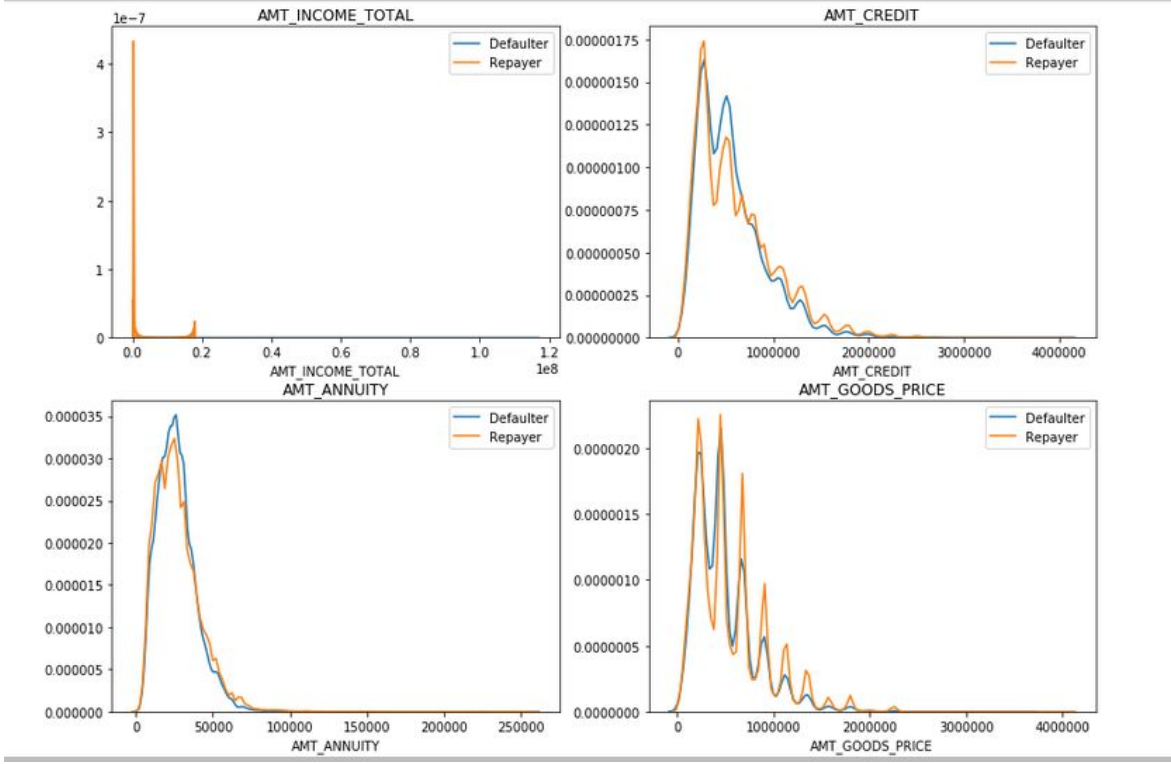
Numerical Univariate Analysis

1)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.

2)Credit amount of the loan is mostly less then 10 lakhs.

3)Most people pay annuity below 50K for the credit loan.

4)Most no of loans are given for goods price below 10 lakhs.

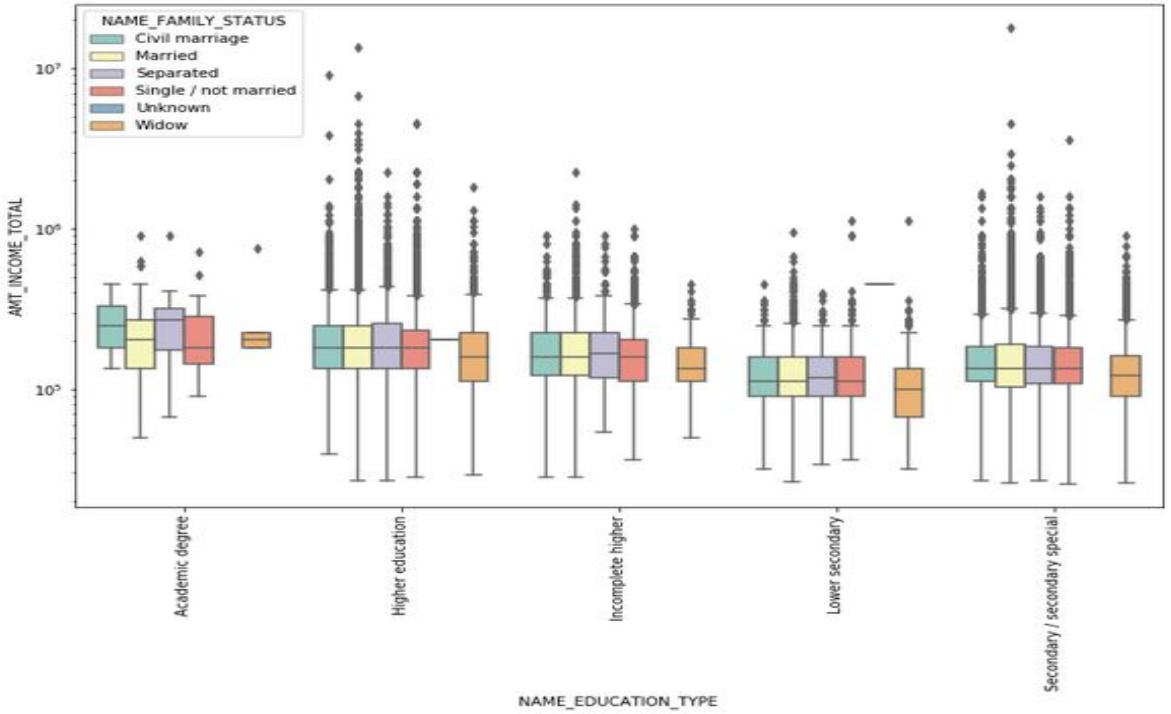


Bivariate Analysis

1)Widow with Academic degree have very few outliers and all people having academic degree have very less outliers.

2)People having Higher Education, Incomplete Higher Education, Lower Secondary Education and Secondary/Secondary Special have a higher number of outliers with Higher education being the largest.

3)Some of the people who haven't completed their Higher Education have higher income and some of the people having Secondary/Secondary Special Education have higher incomes



Some analysis which we did from correlation of target0 and target1

From this correlation heatmap of target0 we can see that:

1)Credit amount is highly correlated with Goods Price Amount and Loan Annuity

2)AMT_CREDIT is inversely proportional to the CNT_CHILDREN, means the Credit amount is higher for fewer children count have and vice-versa.

3)CNT_CHILDREN is also inversely proportional to Days_BIRTH.

From this correlation heatmap of target0 we can see that:

1) Credit amount is highly correlated with good price amount which is same as target0.

2) Loan annuity correlation with credit amount has slightly reduced in target1(0.81) when compared to target0(0.83)

3) People permanent address does not match the contact address are having fewer children.

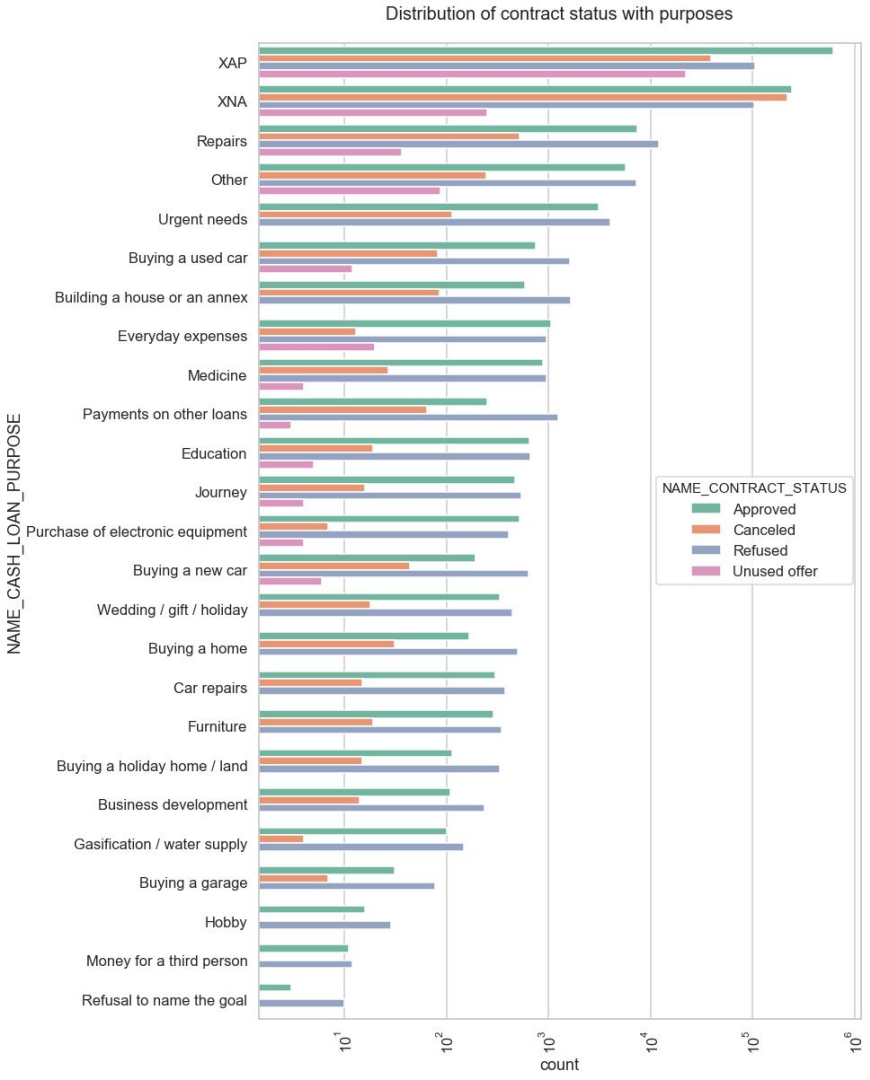
4) Days_birth and number of children correlation has reduced target1 when compared to target0.

Distribution of contract status with purpose

1)For repairs there are most of the rejection for loan.

2)Loan purpose has high number of unknown values (XAP, XNA).

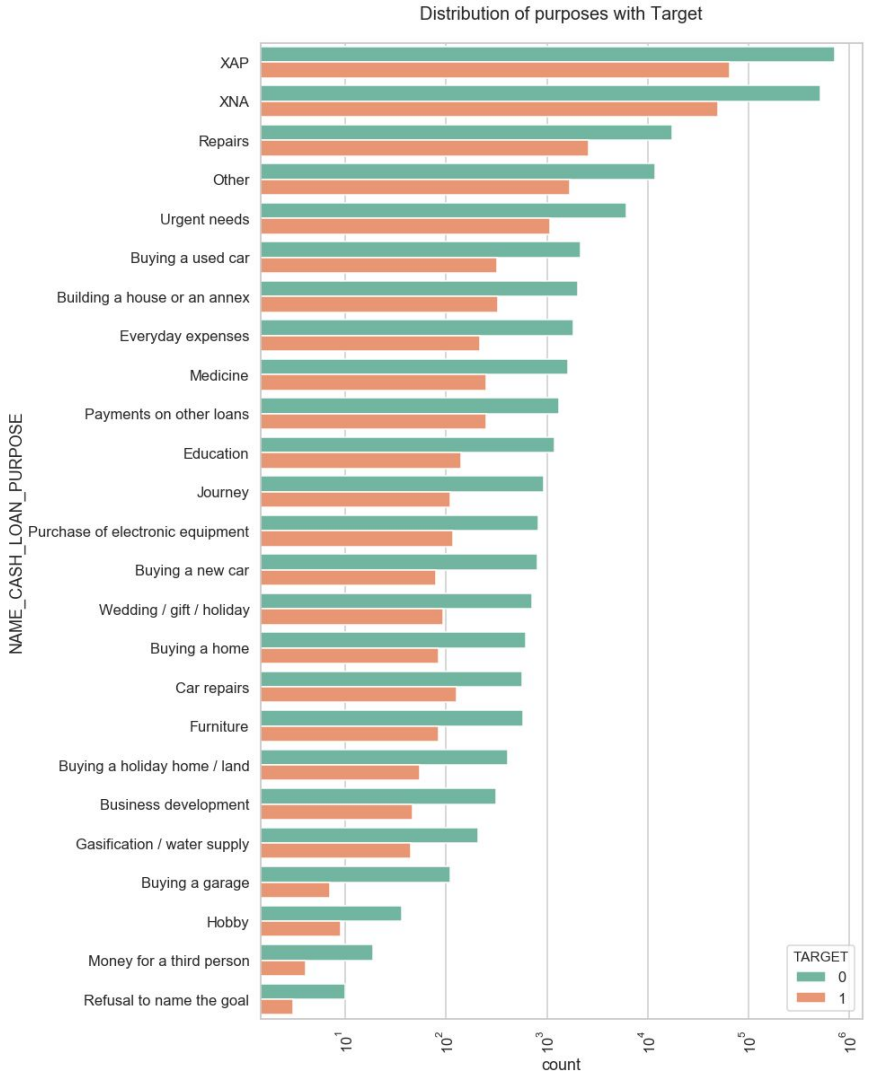
3)For Education purpose we have equal number of approval and rejection.¶



Distribution of purpose with target

1)Loan purposes with 'Repairs' are facing more difficulites in payment on time.

2)In Buying a garage', 'Business developemt', 'Buying land','Buying a new car' and 'Education' loan payment is significant higher than facing difficulties.



Conclusion

After analysing the data we found that:

Academic degree has less defaults.

Student and Businessmen have no defaults.

RATING 1 is safer.

People with Trade Type 4 and 5 and Industry type 8 have defaulted less than 3%

People above age of 50 have low probability of defaulting

People with 40+ year experience having less than 1% default rate

Applicant with Income more than 700,000 are less likely to default

Loans bought for Hobby, Buying garage are being repayed mostly.

People with zero to two children tend to repay the loans.