Case Study on Credit Dataset

Aim: To find which metrics are relevant to making a decision of whether a loan applicant will repay the amount or whether he will have difficulties in repaying it

Vishnu Ram
Ripunjoy Goswami

Table of Content

1. Libraries Used and Approach taken
2. Handling Null values
3. Handling Outliers
4. Univariate Analysis
5. Segmented Univariate
6. Bivariate
7. Correlation Between Attributes
8. Holistic Recommendations to the company

1. Libraries Used and Approach taken

Libraries Used:

1. Pandas
2. Matplotlib
3. Seaborn

Approach taken:

1. In most of the cases we have used Percentage of likelihood to pay, to address the Imbalance (In TARGET Column ~90% Zeros are there ~10% Ones are there).
2. All our analysis is based on TARGET variable and the factors influencing it for all used columns from both the datasets
3. In some places Log scale has been used to address the same

2. Handling Null Values

- Dropped the columns that had null values of more than 35% as a thumb rule as it will affect the analysis a lot.
- Dropped rows which had null values of the columns which had less than 1% Null values
- In 6 AMT_REQ_CREDIT_BUREAU columns Imputed Median Values on null, as Ninety percentile value was nearly equal to Median.

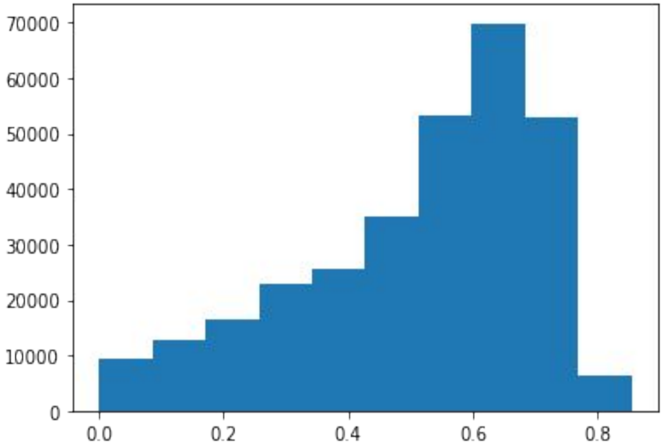
3. Handling Outliers

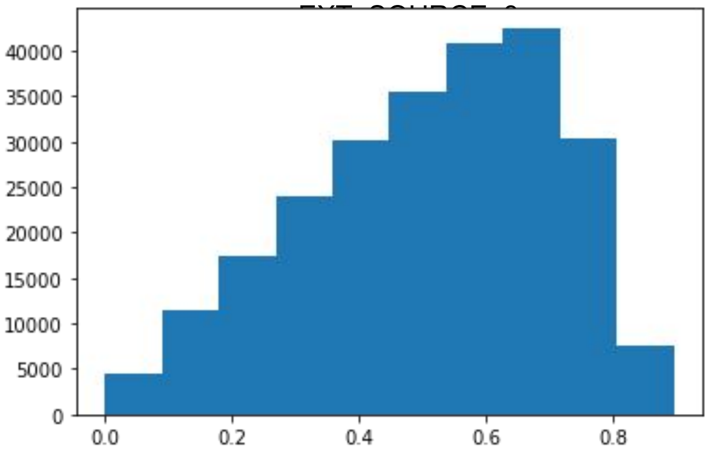
- In Gender Column 4 unusual value “XNA” were there. Those has been replaced with F as F was more than 50%
- In Org type, Replaced “XNA” with “Unknown” to present it better without loosing a good dimension
- Categorised list of columns into:
  - Categorical
  - Binary (0,1/Y,N)
  - Continuous Numerical
- Out of all, below are the selected column for Analysis
  - Dimensions: NAME_INCOME_TYPE, NAME_EDUCATION_TYPE, NAME_HOUSING_TYPE, OCCUPATION_TYPE, ORGANIZATION_TYPE, YEARS_BIRTH_RANGE, YEARS_EMPLOYED_RANGE, CNT_FAM_MEMBERS
  - Facts: AMT_INCOME_TOTAL, AMT_CREDIT, AMT_GOODS_PRICE_RANGE, DEF_60_CNT_SOCIAL_CIRCLE, EXT_SOURCE_2_RANGE

4. Univariate Analysis

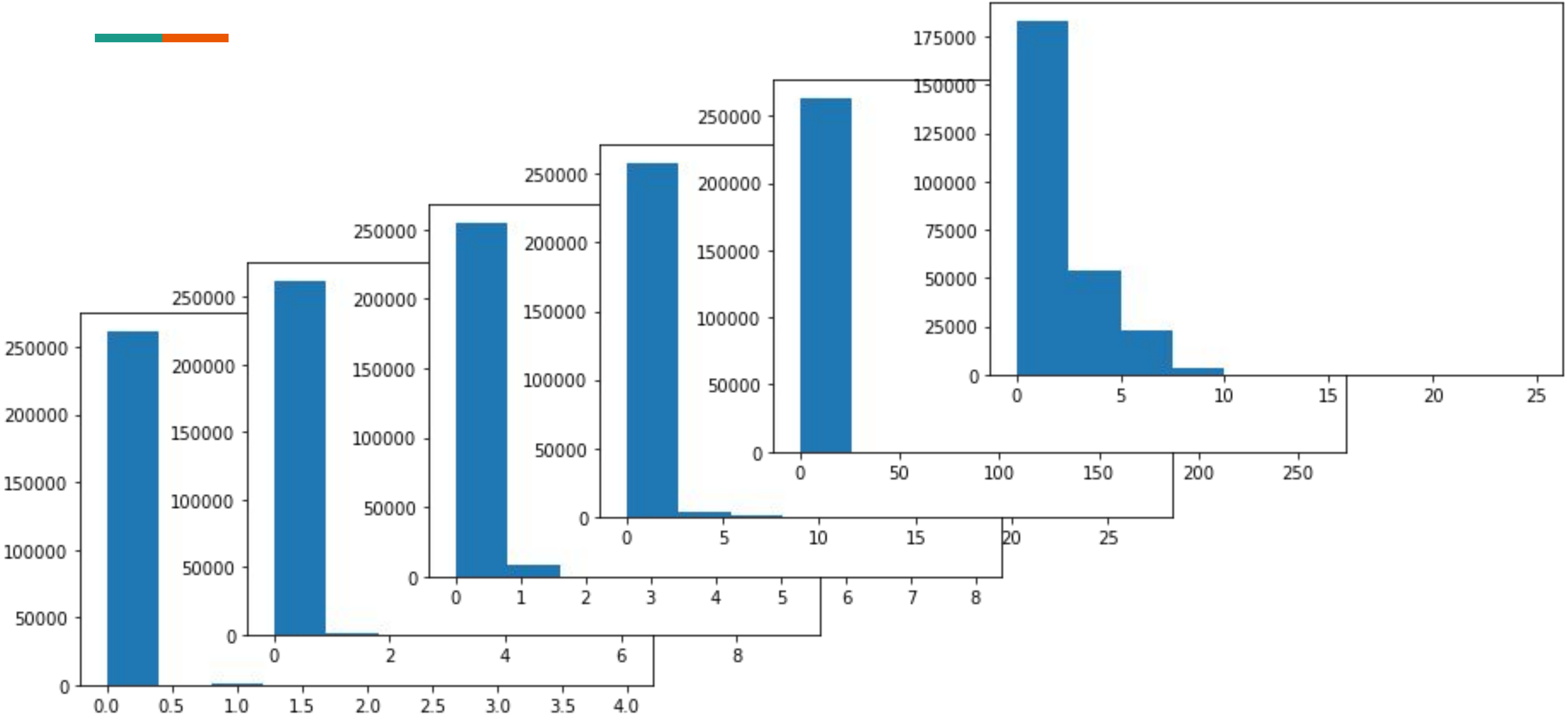
- Out of 3 External Source columns, Dropped the one which had more than 35% null Values, dropped another as the Histogram distribution is very similar
- In Occupation Type column, Imputed Value "Unknown" in place of Null as other category value of the columns are worth using.
- FLAG_MOBIL column has been removed as all the values had only the value 1 which won’t help our analysis

Removing 'EXT_SOURCE_3' Column, as it has more null values and it has a similar distribution as 'EXT_SOURCE_2'





Imputing median to AMT_REQ_CREDIT_BUREAU Columns

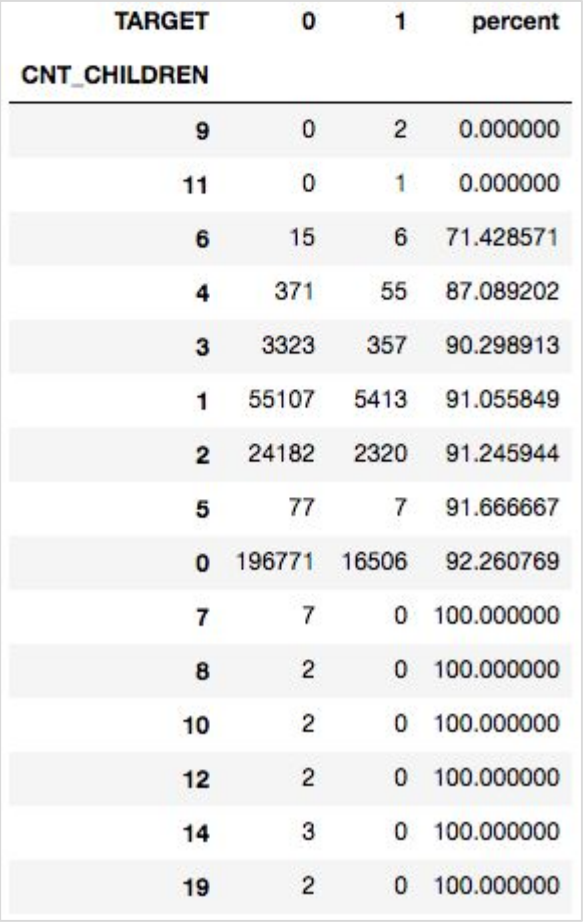


5. Segmented Univariate Analysis

- Taken “TARGET” column as Segmenting Element
- Using Pivot Table, Each Categorical column is taken in Index(Row) and Target is taken in Column for segmenting and count of target is taken in values.
- As a 3rd column percentage of people tend to pay (number of 0/count of target) has been taken
- The table has been sorted based on Percentage
- Selected 5 Columns for further analysis as given in the following slides

The Percentage Approach

The Percentage Approach to address the imbalance of target column. The image in right is an example of that

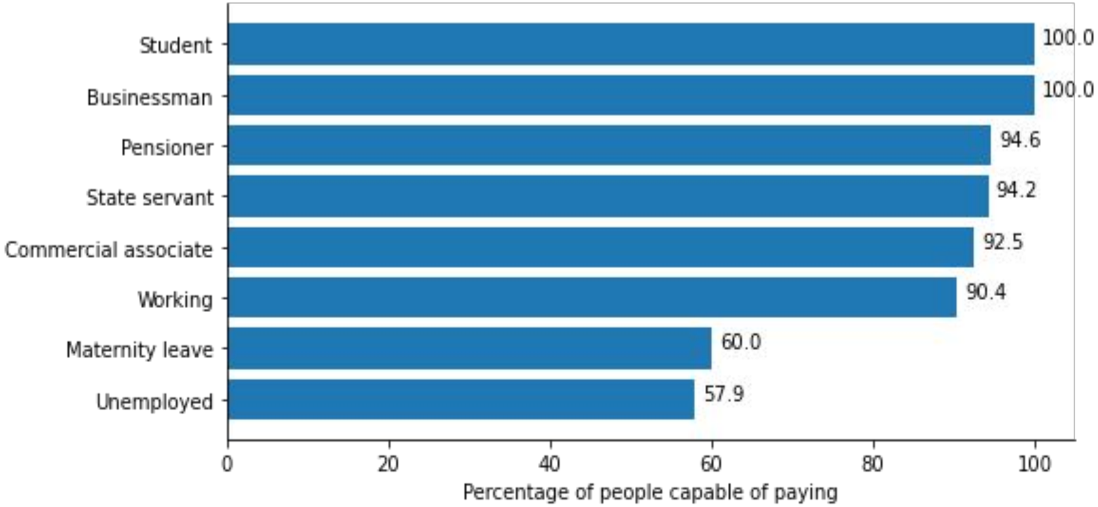


TARGET	0	1	percent
CNT_CHILDREN			
9	0	2	0.000000
11	0	1	0.000000
6	15	6	71.428571
4	371	55	87.089202
3	3323	357	90.298913
1	55107	5413	91.055849
2	24182	2320	91.245944
5	77	7	91.666667
0	196771	16506	92.260769
7	7	0	100.000000
8	2	0	100.000000
10	2	0	100.000000
12	2	0	100.000000
14	3	0	100.000000
19	2	0	100.000000

Repayment Likelihood based on INCOME_TYPE

Observation: People who are Unemployed and who are in Maternity Leave are less likely to Pay back the loan

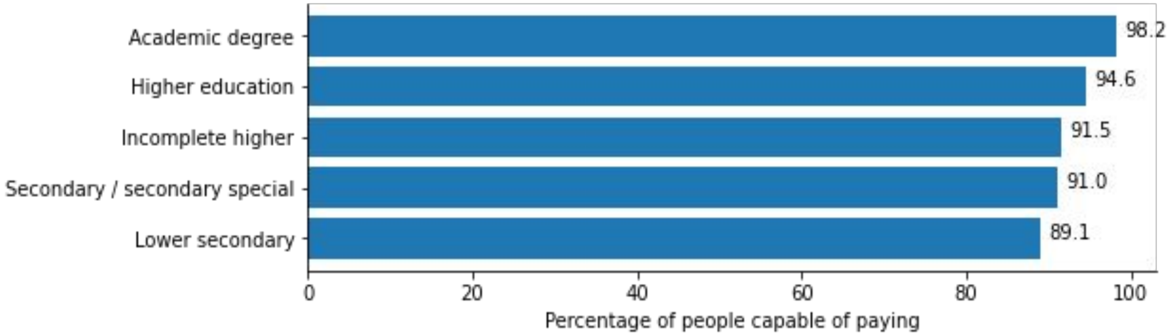
Recommendation: 1. Avoid Providing Loan to Maternity leave people and unemployed people.
2. Almost all Students and Businessmen tend to pay, don't lose them. For others increase interest rate



Repayment Likelihood based on EDUCATION_TYPE

Observation: 98.2 % of Academic Degree People are more likely to repay.

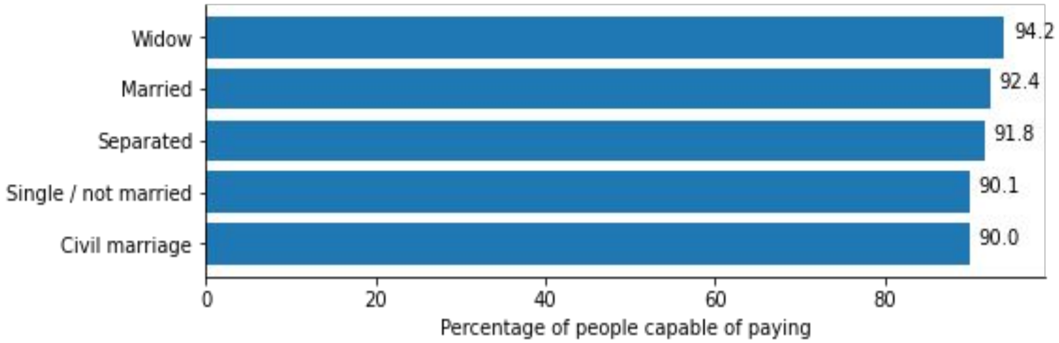
Recommendation: Should not lose an Academic degree candidate's loan application. Consider increasing the Rate of interest for the rest



Repayment Likelihood based on FAMILY_STATUS

Observation: 94.2 % of widow are more likely to repay.

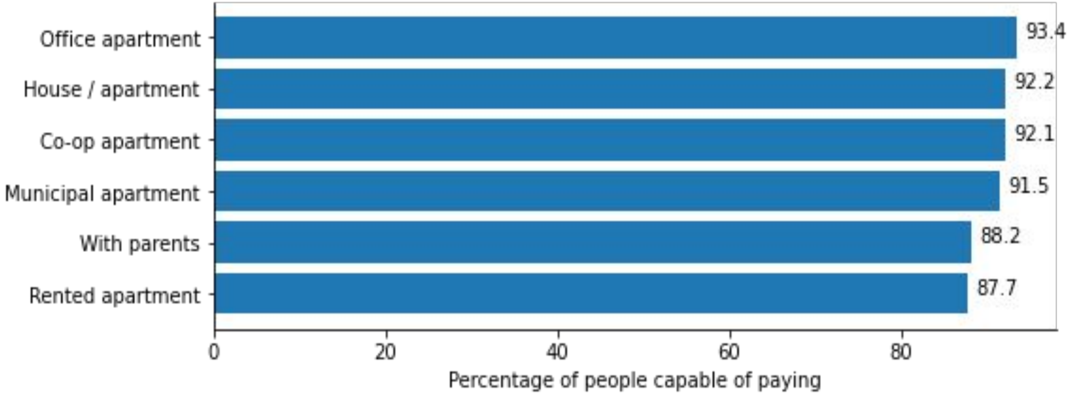
Recommendation: Should not lose an widow candidate's loan application. Consider increasing the Rate of interest for the rest



Repayment Likelihood based on HOUSING_TYPE

Observation: The Candidate who is staying in parent’s home and rented apartment are the one who have difficulty in paying the loan

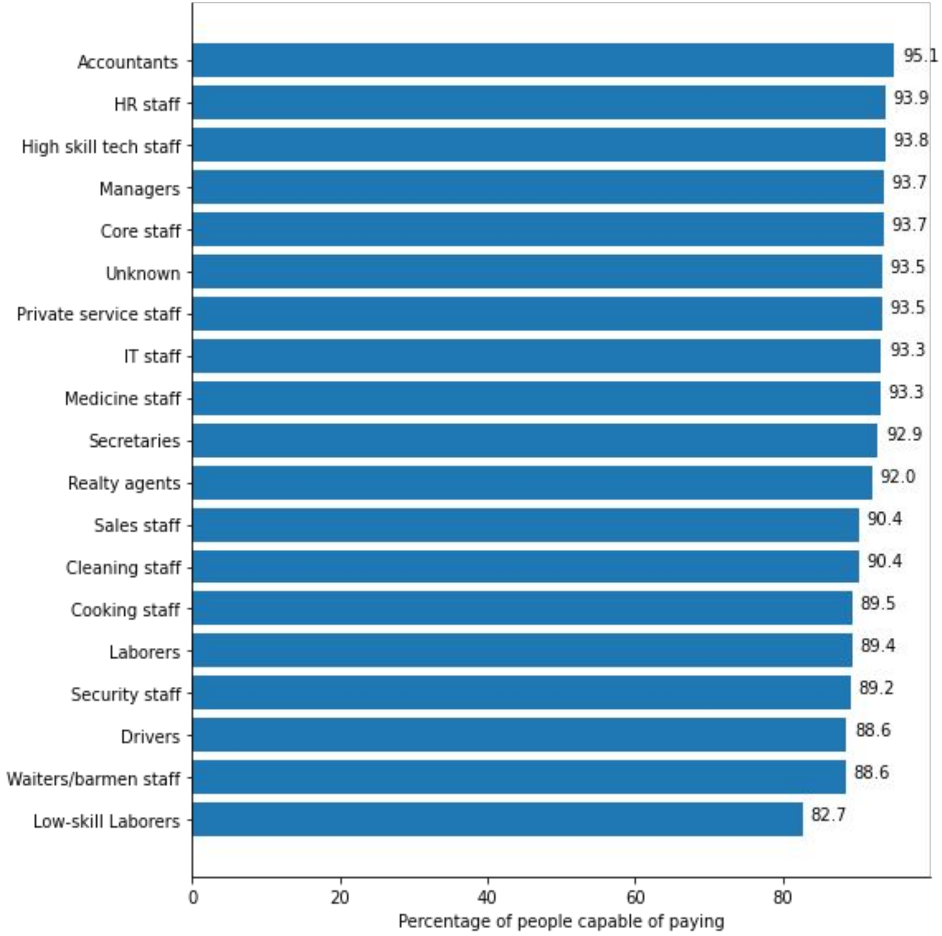
Recommendation: Avoid giving loan or increase rate of interest for those people



Repayment Likelihood based on OCCUPATION_TYPE

Observation: Sales Staff and non skill labour are less tend to pay the loan

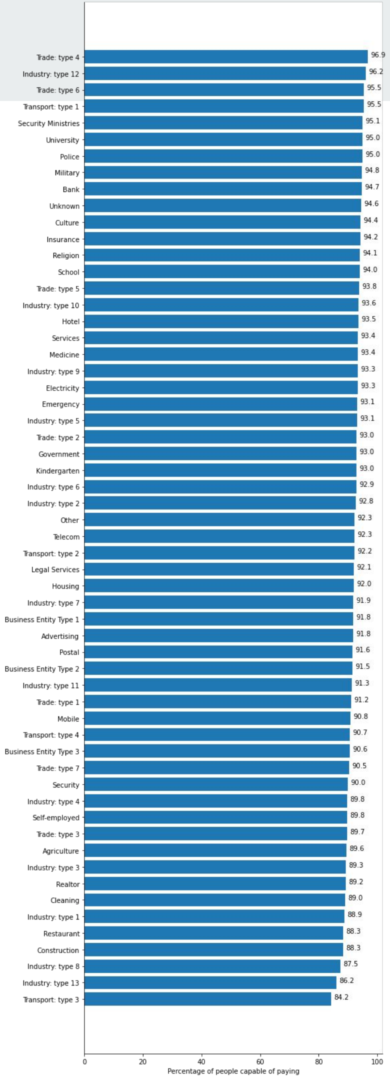
Recommendation: Avoid Approving loan or increase Interest Rate



Repayment Likelihood based on ORGANIZATION_TYPE

Observation: Industry type 8, Industry type 13, Transport type 3 people are less tend pay the loan

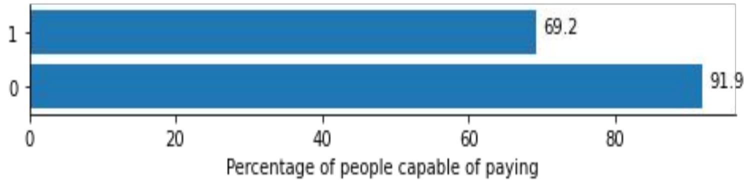
Recommendation: Avoid giving loan or increase rate of interest for those people



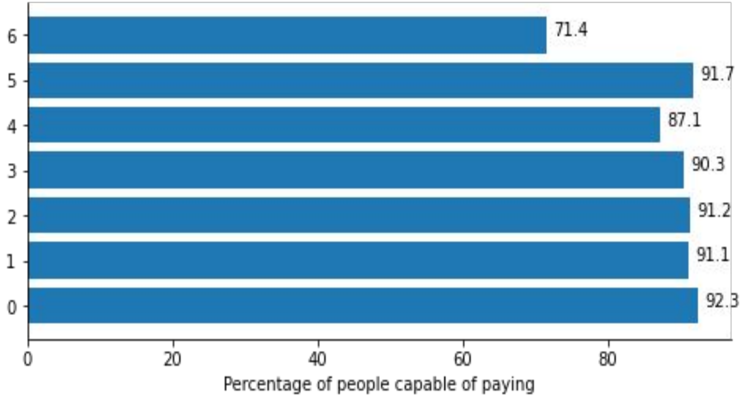
Repayment Likelihood based on FLAG_DOCUMENT_2

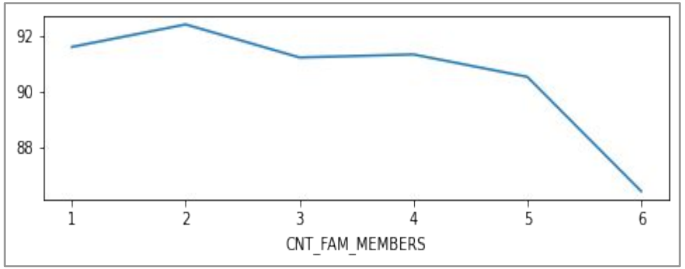
Observation: Out of all document types, Document type 2 had a significant impact. It could be Educational Document. Assuming that. People providing that document as security are less possible to pay back

Recommendation: Avoid giving loan or increase rate of interest for those people who give Document-2



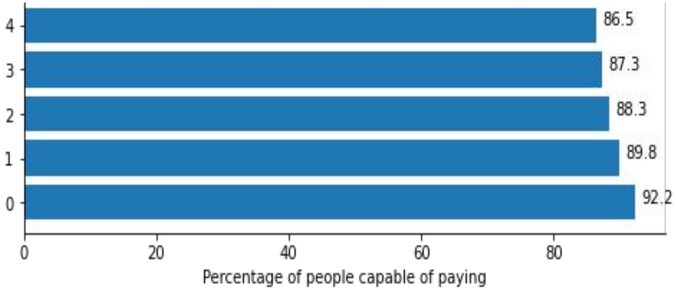
Repayment Likelihood based on CNT_CHILDREN & CNT_FAM_MEMBERS

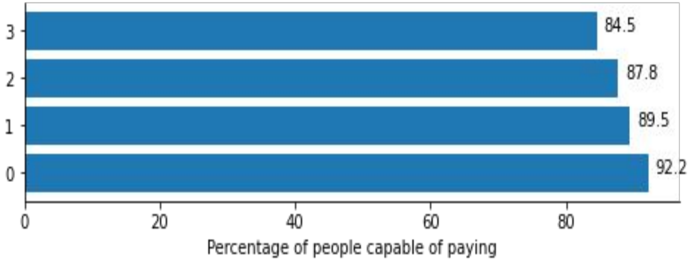




Observation: When the Number of children or number of family members increases the possibility of Paying back reduces.

Repayment Likelihood based on DEF_30_CNT_SOCIAL_CIRCLE & DEF_60_CNT_SOCIAL_CIRCLE





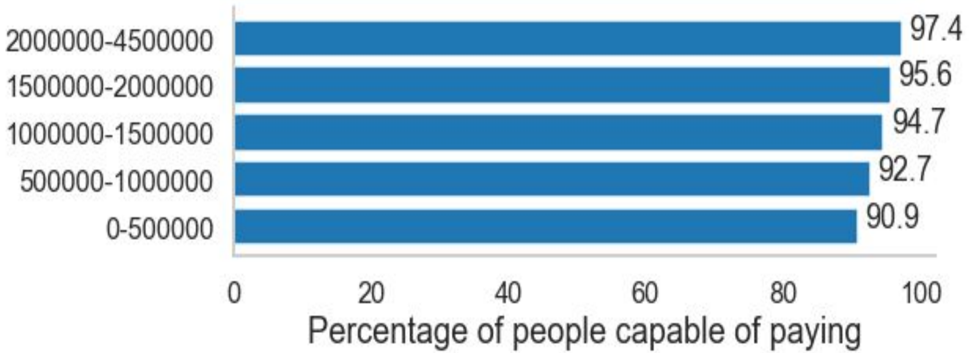
Observation: The observation of client's social surroundings defaulted on 30 DPD (days past due) and 60 DPD shows perfectly that if there are any defaulters in their surrounding their likelihood of paying decreases

Recommendation: Based on defaulter's area Interest rate could be factored.

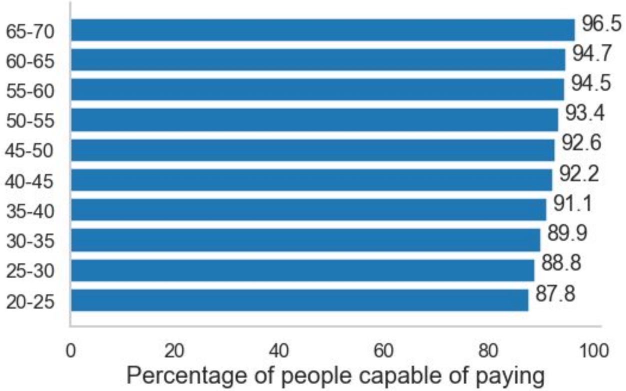
Repayment Likelihood based on GOODS_PRICE

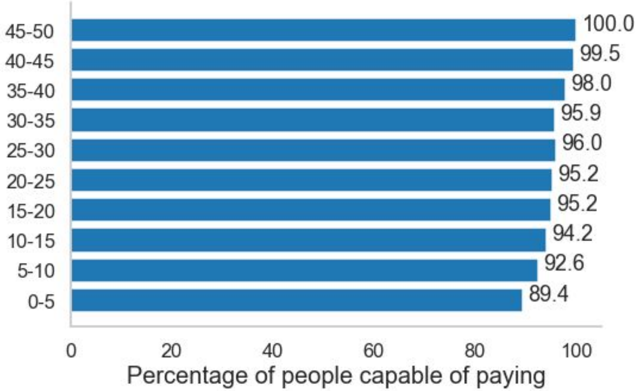
Observation: When Goods Price increases, the likelihood of paying back increases

Recommendation: Interest rate could be decreased for high goods price to encourage loan takers



Which Age group and Work Experience of Applicant are having good repayment likelihood

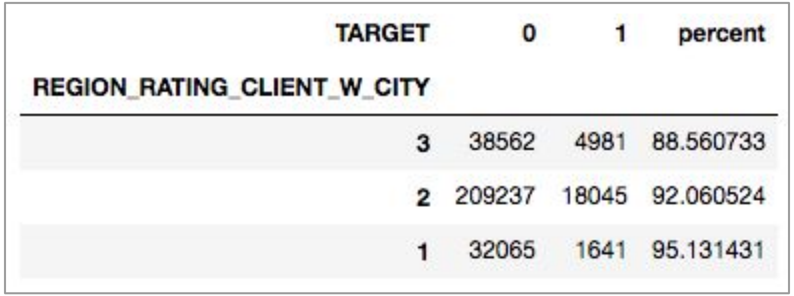




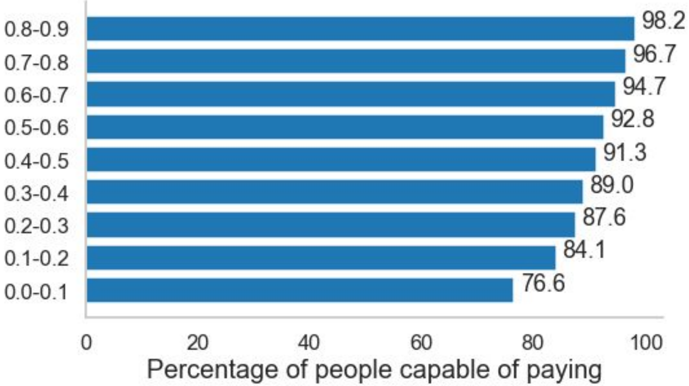
Observation: Age and work experience has a positive correlation with likelihood of paying back the loan.

Recommendation: Interest rate could be decreased for senior citizens.

Which Rating to Trust



TARGET	0	1	percent
REGION_RATING_CLIENT_W_CITY			
3	38562	4981	88.560733
2	209237	18045	92.060524
1	32065	1641	95.131431



Observation: REGION_RATING_CLIENT_W_CITY is not reliable as it is negatively correlating while, **EXT_SOURCE_2_RANGE** correlates Positively

Recommendation: Going ahead Ext Source Rating could be trusted more than Region Rating and region rating with respect to city of the client

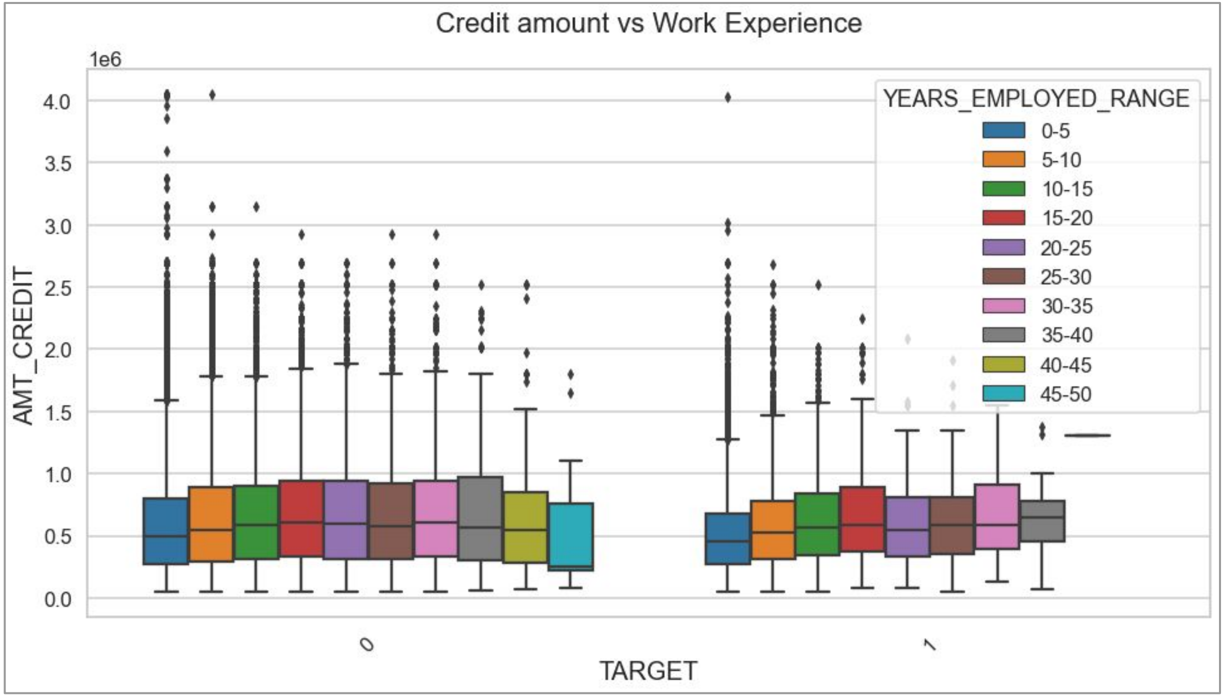
6. Bivariate Analysis

1. 5 Categorical variables (Dimensions) have been chosen and 10 Numerical variables (Facts) have been chosen from 122 columns.
2. 50 combinations of graphs have been studied and found out the upcoming insights

Credit Amount Vs Work Experience

Observations: Most of the applicants with more than 40 years tend to pay back the loan.

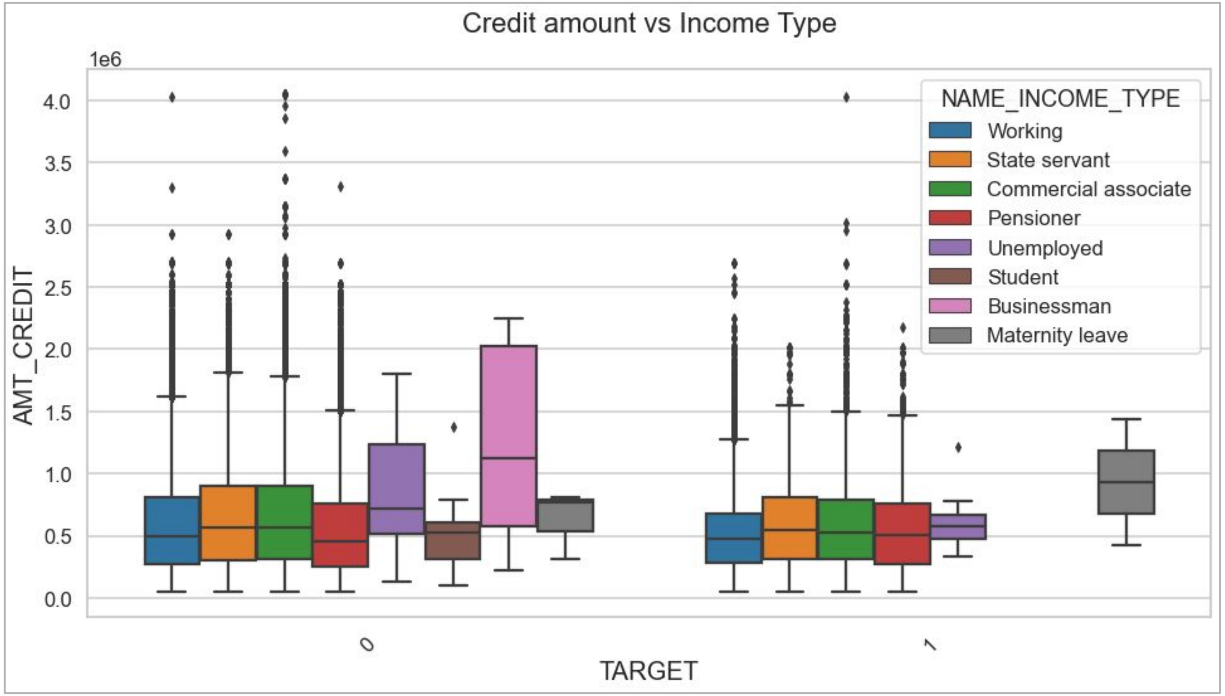
Recommendations: Interest Discounts could be provided to encourage them to take up loans



Credit Amount Vs Income Type

Observations: Nearly all Students and applicants tend to pay back the loan. High credit loans for maternity leave applicants have difficulty in paying the loan.

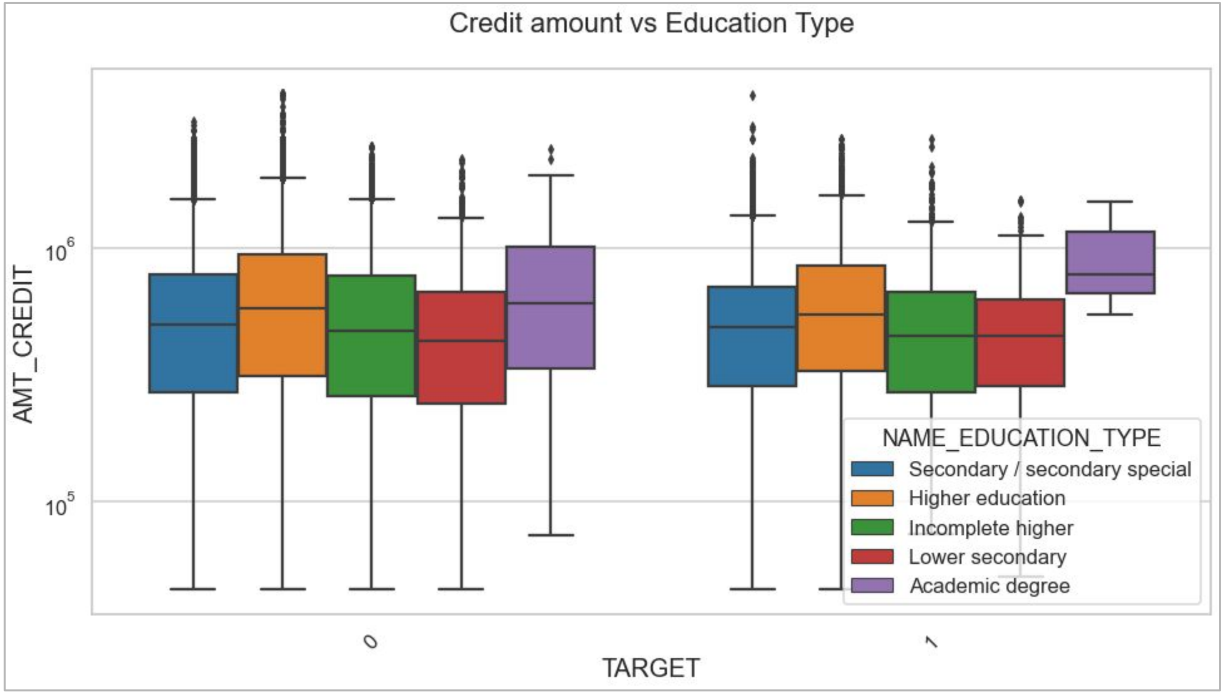
Recommendations: Interest Discounts could be provided to encourage Students and businessmen to take up loans. Maternity leave applicants should not be provided a loan greater than 10 lakh



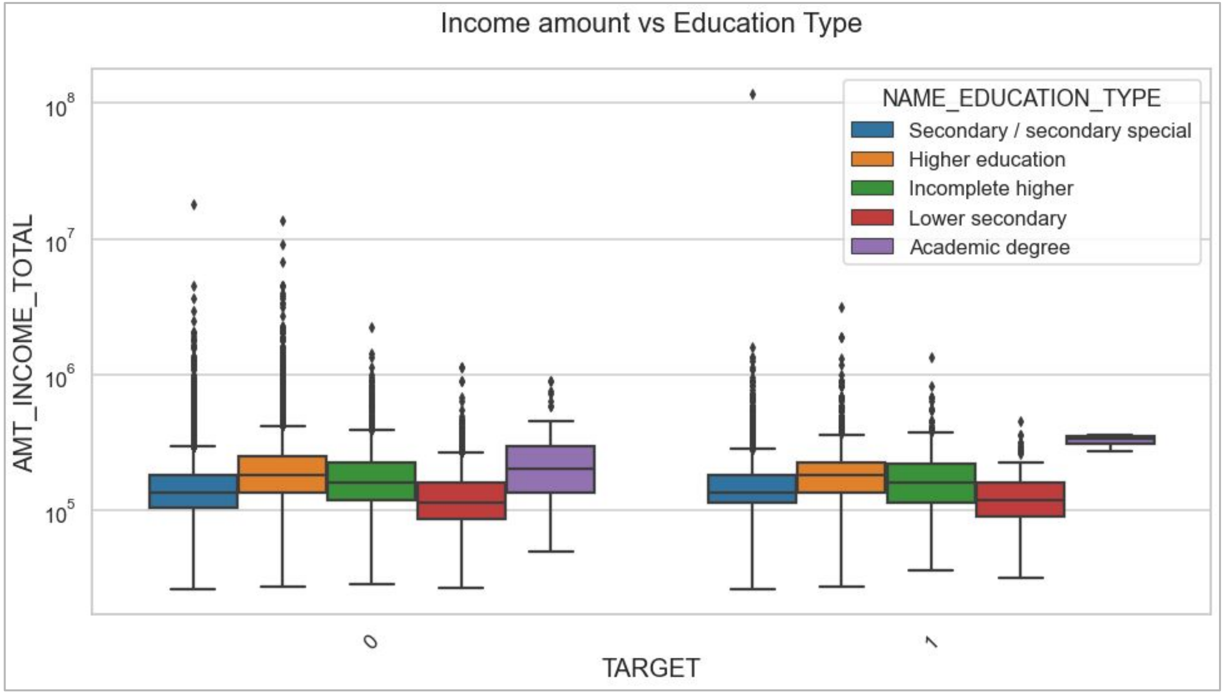
Credit Amount Vs Education Type

Observations: People holding Academic Degree are having difficulty in paying when credit amount is more than say 10 lakh

Recommendations: Interest could be increased for these applicants.



Income Amount Vs Education Type

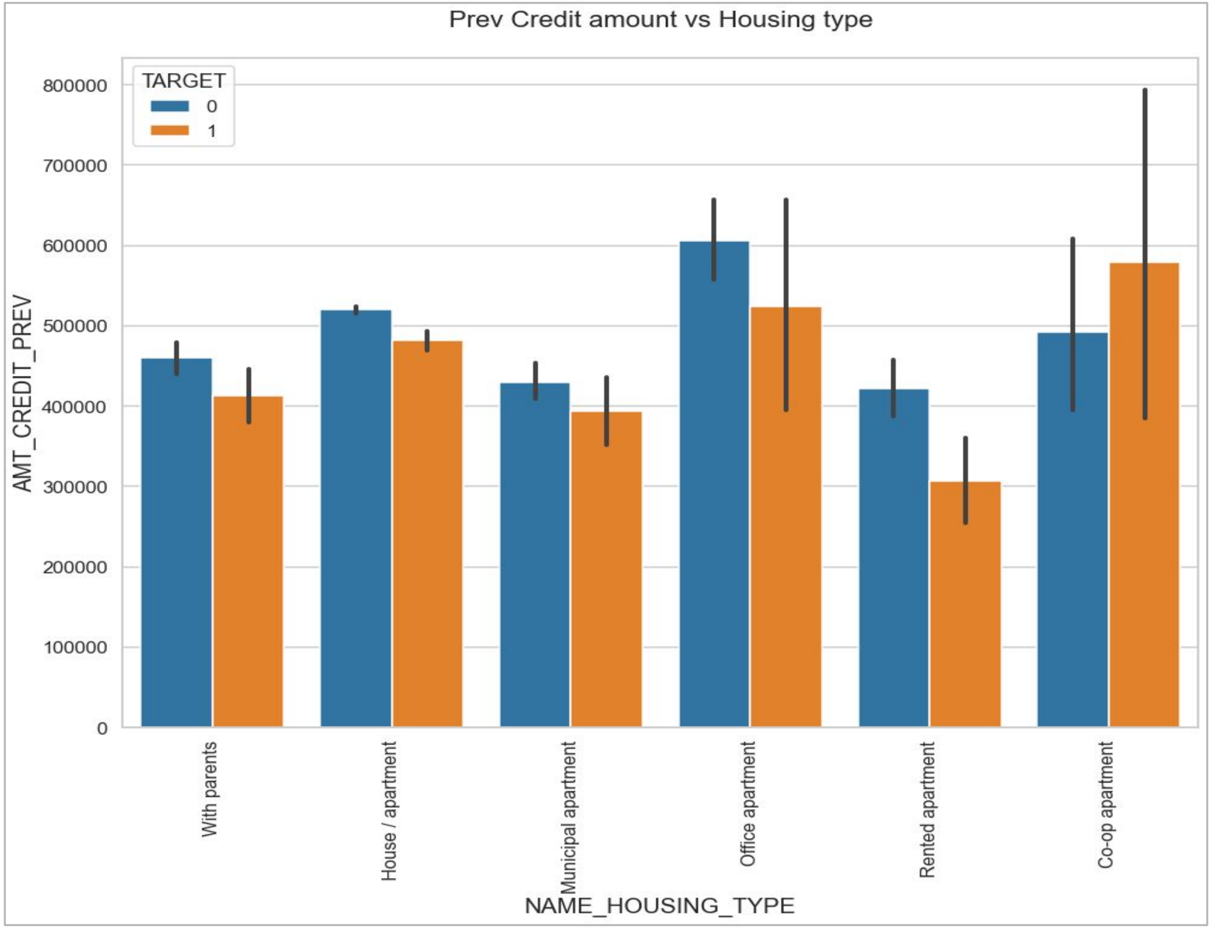


Observations: People holding Academic Degree and high salary are having difficulty in paying.

Recommendations: Interest could be increased for these applicants.

Previous Credit Amount Vs Housing

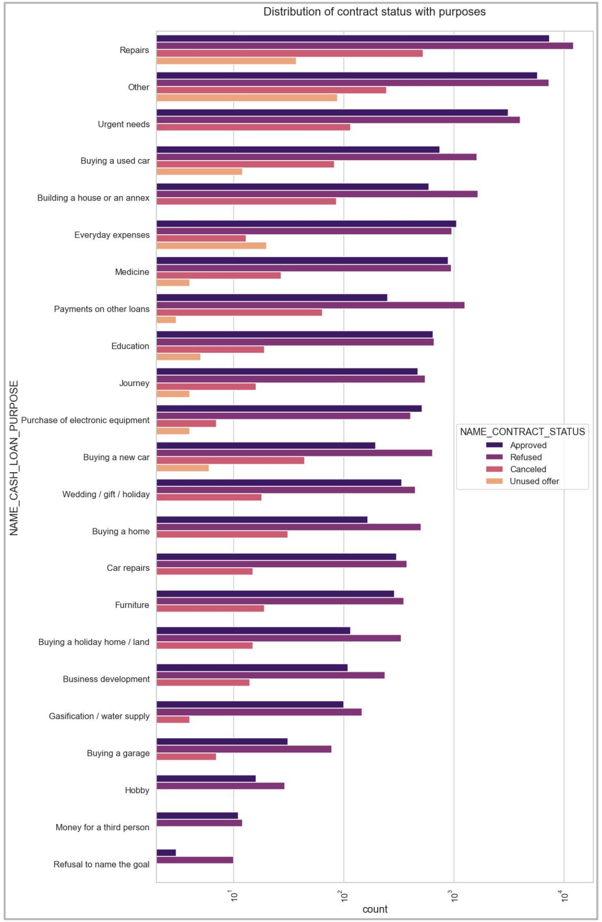
Here for Housing type,office apartment is having higher credit of target 0 and co-op apartment is having higher credit of target 1. So, we can conclude that bank should avoid giving loans to the housing type of co-op apartment as they are having difficulties in payment. Bank can focus mostly on housing type with parents or House\apartment or municipal apartment for successful payments.



Purpose of Loan Vs Contract Status

"Payments on other loans",
"Buying a home",
"Buying a holiday home/land",
"Buying a garage",
"Buying a new car"

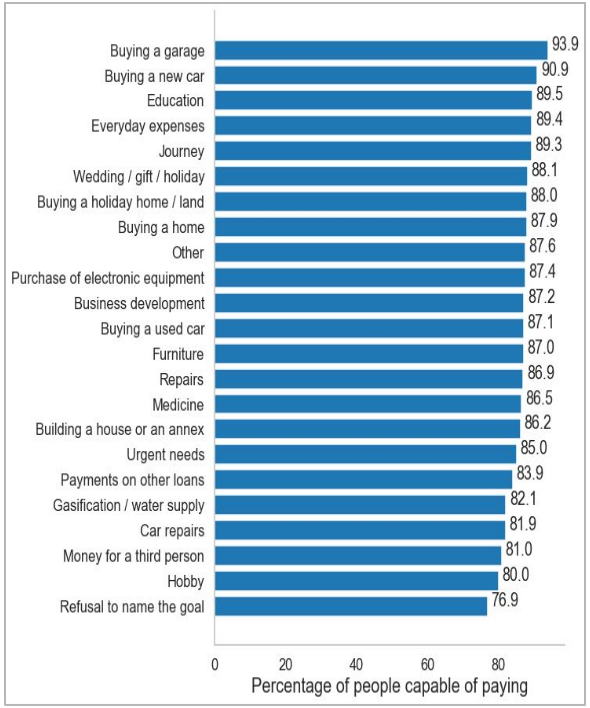
These categories loans is having significant higher rejection than approvals. While

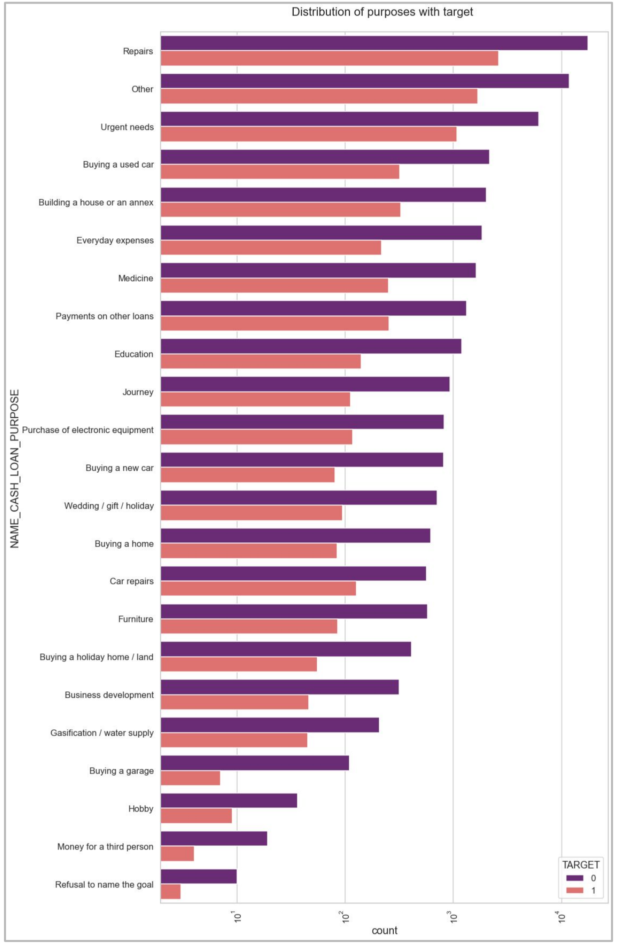


Good Paying Category applications are being Rejected

Observation: In Previous Slide we saw that, In previous applications,
"Buying a home",
"Buying a holiday home/land",
"Buying a garage",
"Buying a new car"
These columns were rejected more. But these are the segments that perform well.

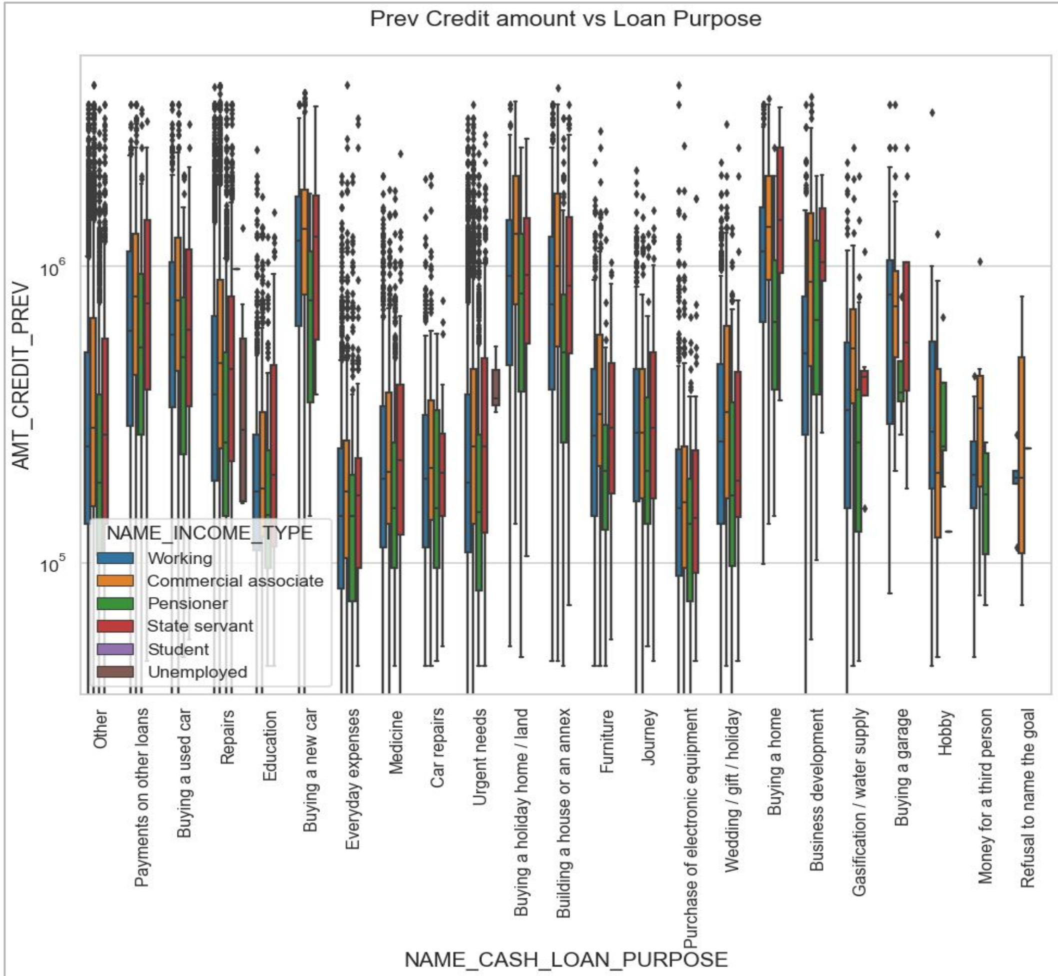
Recommendations: This category's approach should be relooked and approval should be encouraged.





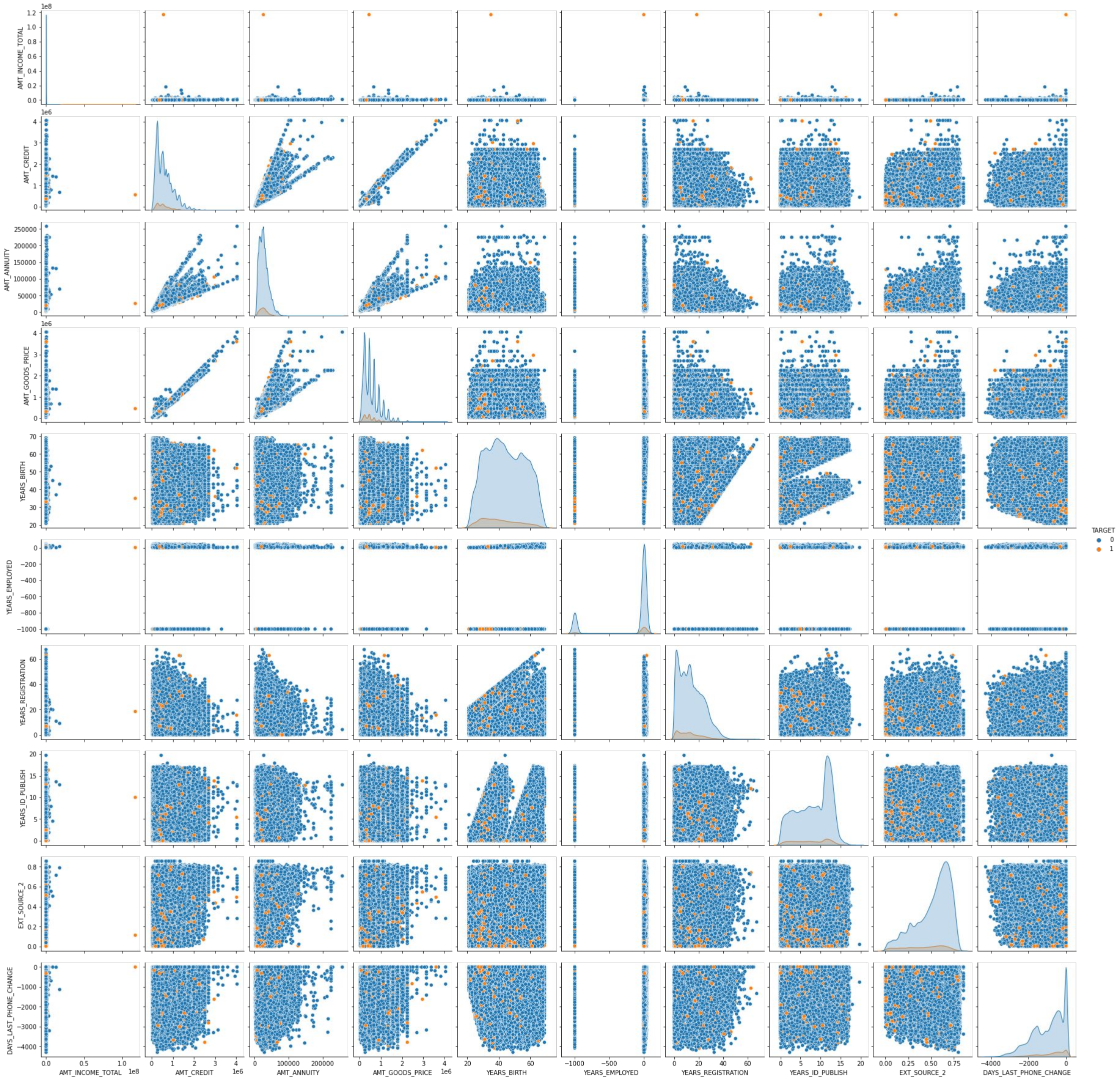
Bivariate for merged dataset

1. The credit amount of Loan purposes like 'Buying a home','Buying a land','Buying a new car' and 'Building a house' is higher.
2. Income type of state servants have a significant amount of credit applied
3. Money for third person or a Hobby is having less credits applied for.



7. Correlation Between Attributes

Pairplot has been made to understand whether the variables correlate with each other to understand the genuinity of the data.



Correlation of Selected columns

- Credit amount, Goods Price, Annuity are positively correlated with each other
- Low age people tend to default more
- Low External Rating score people tend to default more
- Lower age people has less External Rating Score

8. Holistic Recommendations to the company

The parameters recommended in the upcoming slide, highly influence the likelihood of the applicant of repaying.

The 10 major Attributes that helps us to judge the applicants are as below

1. Income Type
2. Education Type
3. Family Status
4. Housing Type
5. Occupation Type
6. Organisation Type
7. Count of Children (or) Count of Family Members
8. Number of observations of client's social surroundings defaulted on 30 DPD (days past due) (or) 60 DPD
9. Age Group and Work Experience
10. Purpose of Loan

Top 4 Continuous Numeric Value that helps us to judge the applicants are as below

1. External Rating Score
2. Credit Amount
3. Goods Price
4. Total Income

Thank you

- Vishnu Ram D
- Ripunjoy Goswami

Post Graduate Diploma in Data Science
International Institute of Information Technology - Bangalore
Upgrad