PRESENTATION ON CREDIT EDA CASE STUDY

Done by. Sathish Kumar E

PURPOSE

To analyse the data set and help the organization make a better decision in loan approval based on applicants profile.

This will help the organization avoiding the financial loss to the company.

STEPS

- Understand the dataset
- Check on the quality of data (missing values, outliers, etc) and address it accordingly.
- Check for data imbalance and perform univariate and bivariate analysis.
- Merging of application data with the previous application data.
- Recommendations and Risks.

Loaded the ‘application_data.csv’

Initial shape of the dataset -> (307511, 122)

There were 49 columns with 33% and more null values. Those columns were dropped from the dataset, shape of the updated dataset is (307511, 73)

Still 19 more columns contains null values ranging from 1 to 96391.

Columns having a very few null values was replaced with the most occurring value in the column, like in the Gender column, there were only 4 missing values, it was replaced with ‘F’ which was the most appearing in that column.

But, in some column, the count of null values, was too big, and replacing it with the most appearing value would change the entire look of dataset. For example, OCCUPATION_TYPE had 96,391 null values, it was not possible to replace it with most appearing value and this column cannot be dropped as well, because loan applicants occupation is an important factor for analysis. So, the null values in such columns was replaced by NaN.

'DAYS_BIRTH’, 'DAYS_EMPLOYED’, 'DAYS_REGISTRATION’, 'DAYS_ID_PUBLISH',

'DAYS_LAST_PHONE_CHANGE’

This 5 columns had values on days count, like 730, 1000, 2000, etc. It was divided by 365 and was converted to years.

Added 3 new columns,

1) RANGE_AMT_INCOME
2) RANGE_AMT_CREDIT
3) BINS_DAYS_BIRTH

This columns was used to categorize the income into high, low, medium, very high, very low. Same followed for credit as well. For birthdays, classified into Young, Senior Citizen, Middle Age, Very Young.

Some of the important columns in the dataset were,

AMT_INCOME_TOTAL (Income of the client)

AMT_CREDIT (Credit amount of the loan)

AMT_ANNUITY (Loan annuity)

CODE_GENDER (Gender of the client)

TARGET (Target variable)

OCCUPATION_TYPE (What kind of occupation did the client have)

ORGANIZATION_TYPE (Type of organization where client works)

NAME_EDUCATION_TYPE (Level of highest education client achieved)

And few more columns were also used for analysis.

In the next step we are dividing the dataset into two different datasets for analysing based on the client difficult on loan payment.

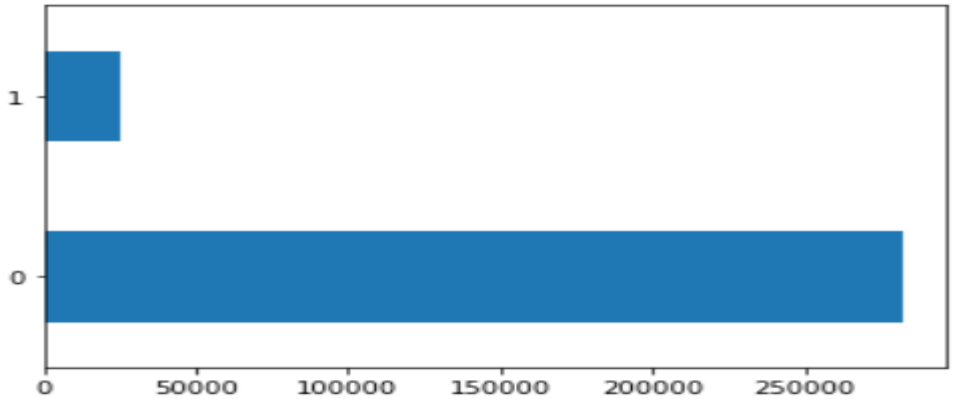
Dividing our dataset into 2 different datasets based on clients loan payment difficulties and all other case.

In TARGET column, there are two values, 0 & 1.

‘0’ -> all other cases (non loan payment difficulties)

‘1’ -> client with loan payment difficulties

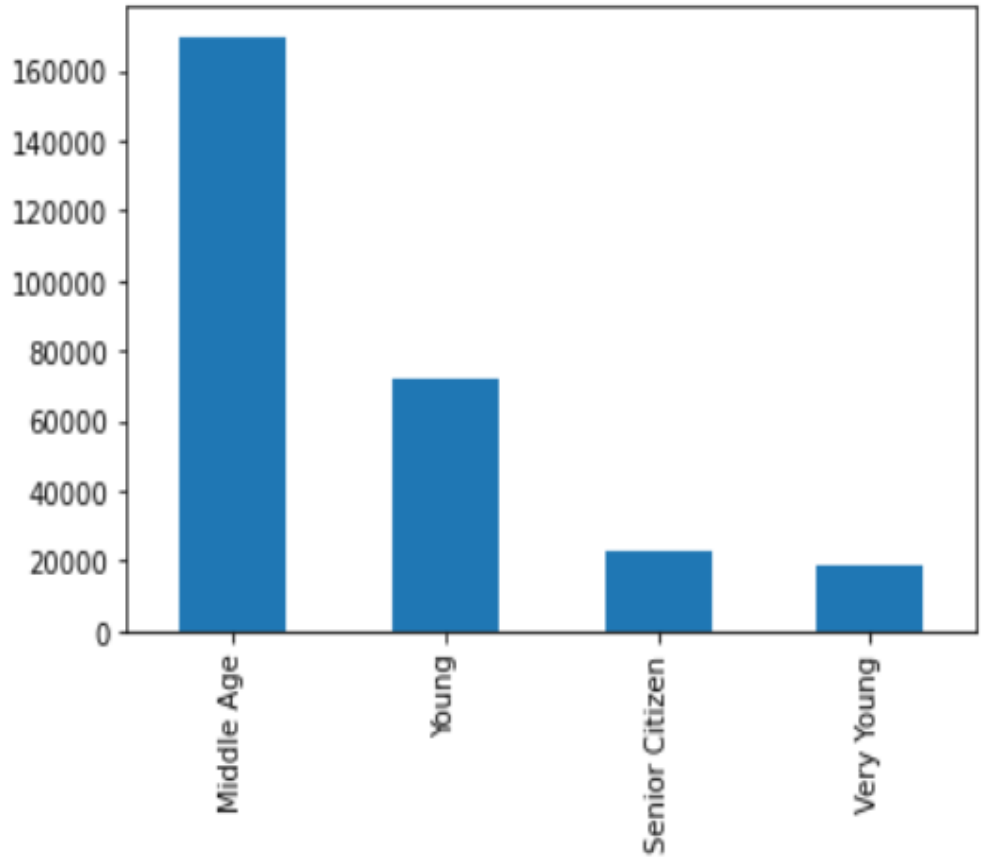
0	282686
1	24825



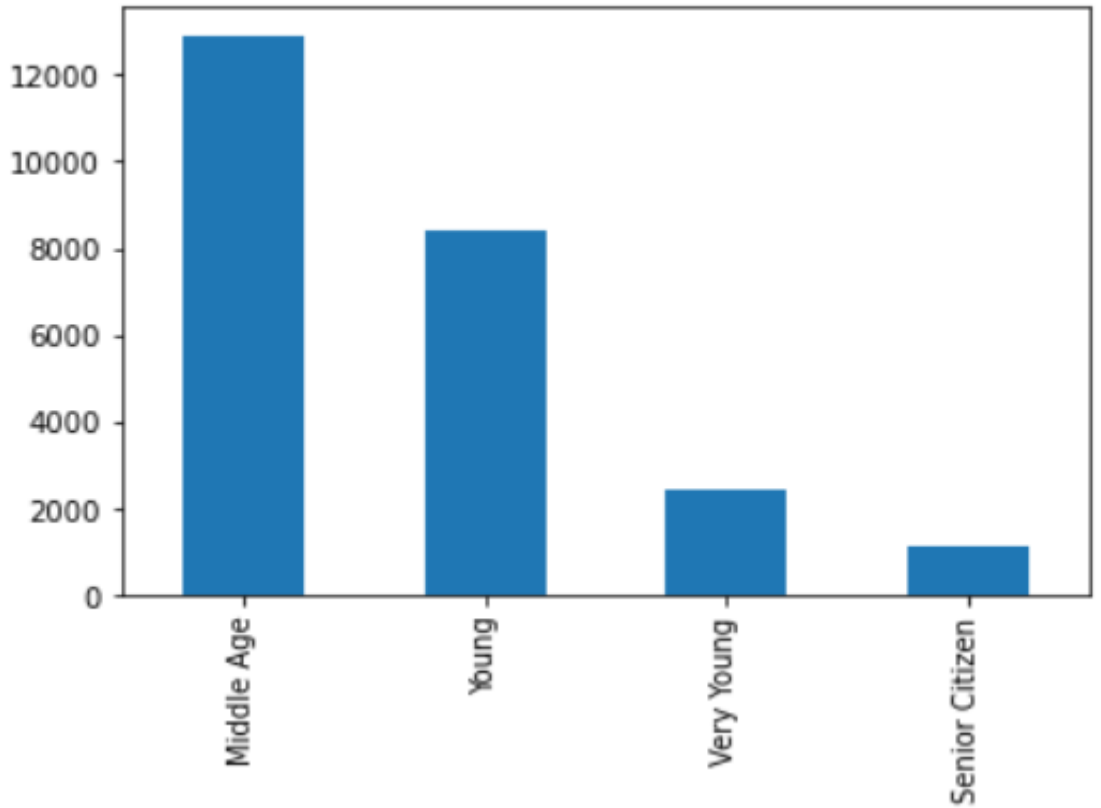
Now, dataset corresponding to loan payment difficulties will contain 24825 records.
And the other dataset corresponding to non loan payment difficulties will contain 282686 record.

DISTRIBUTION ON LOAN APPLICATS AGE

Distribution of Age of Non Loan Payemnt Difficulties

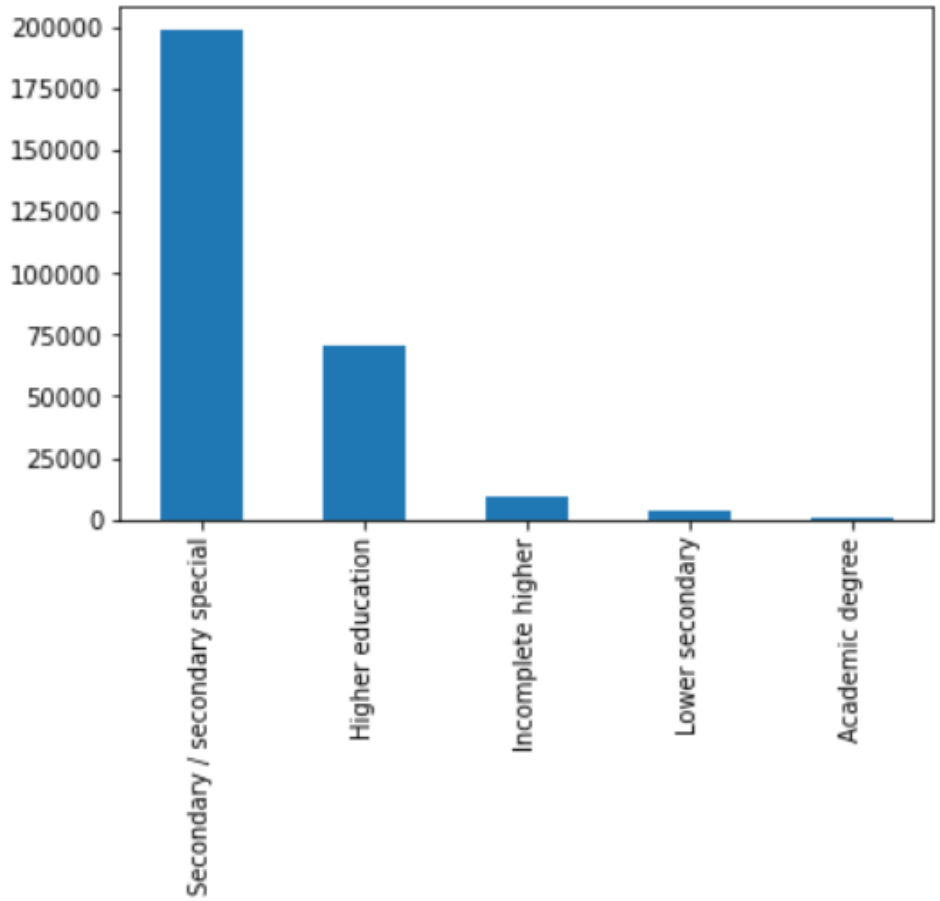


Distribution of Age of Loan Payemnt Difficulties

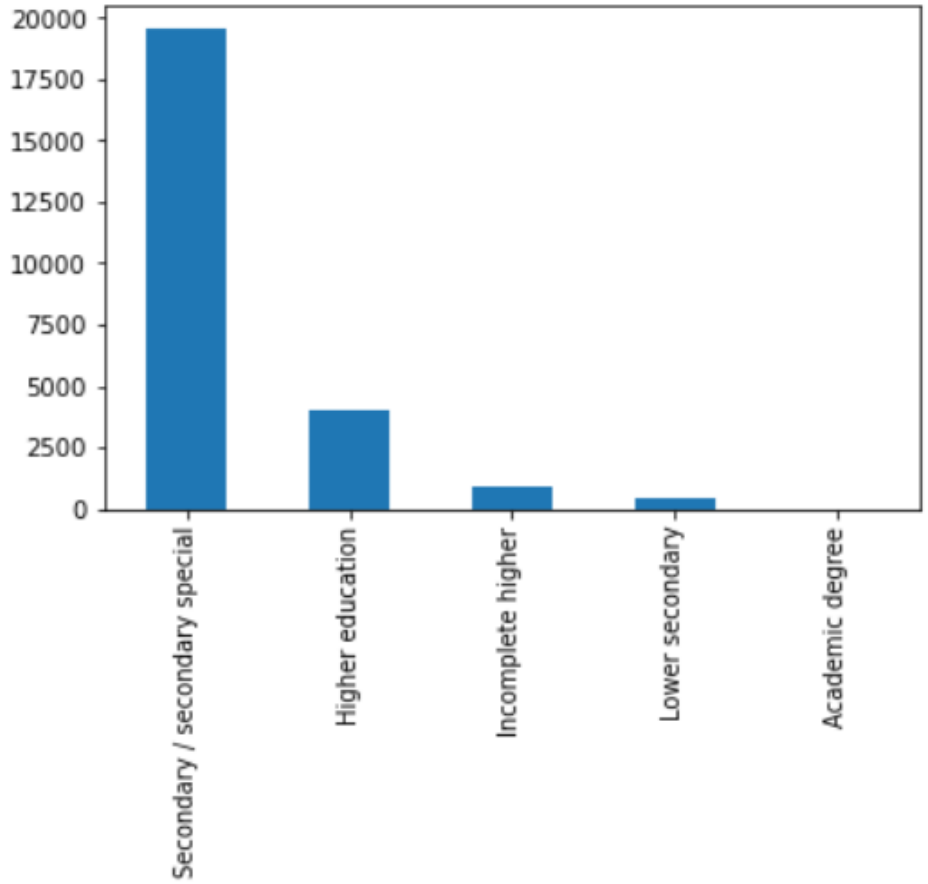


DISTRIBUTION ON LOAN APPLICANTS EDUCATION

Educational Qualification of Non Loan Payemnt Difficulties

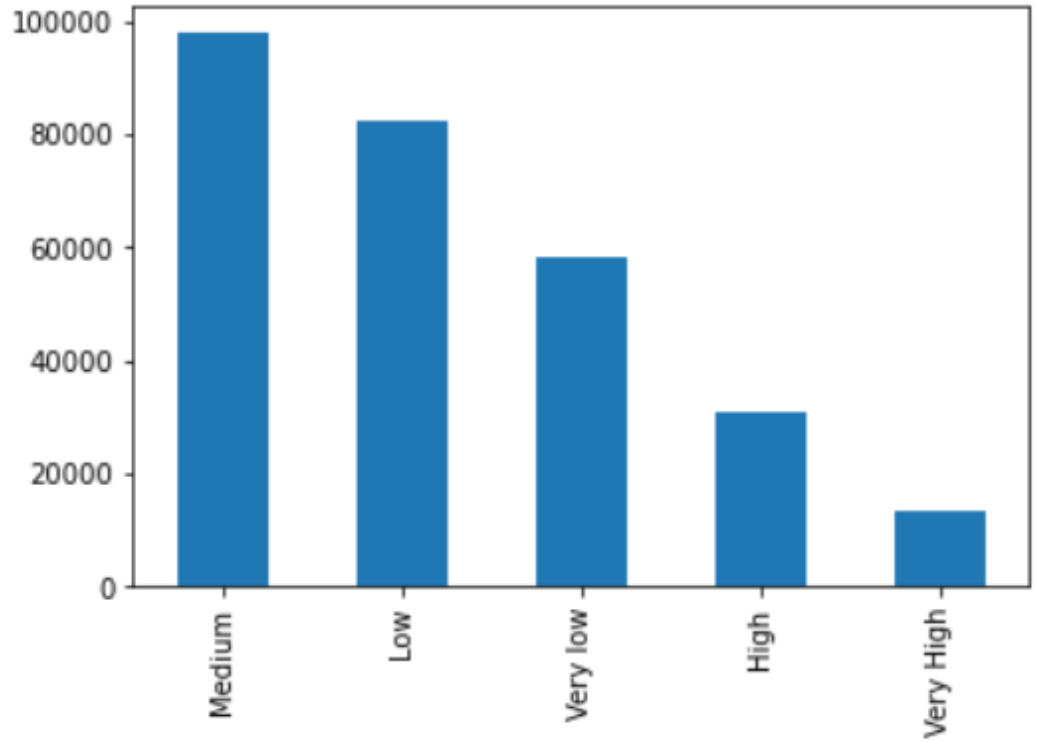


Educational Qualification of Loan Payemnt Difficulties

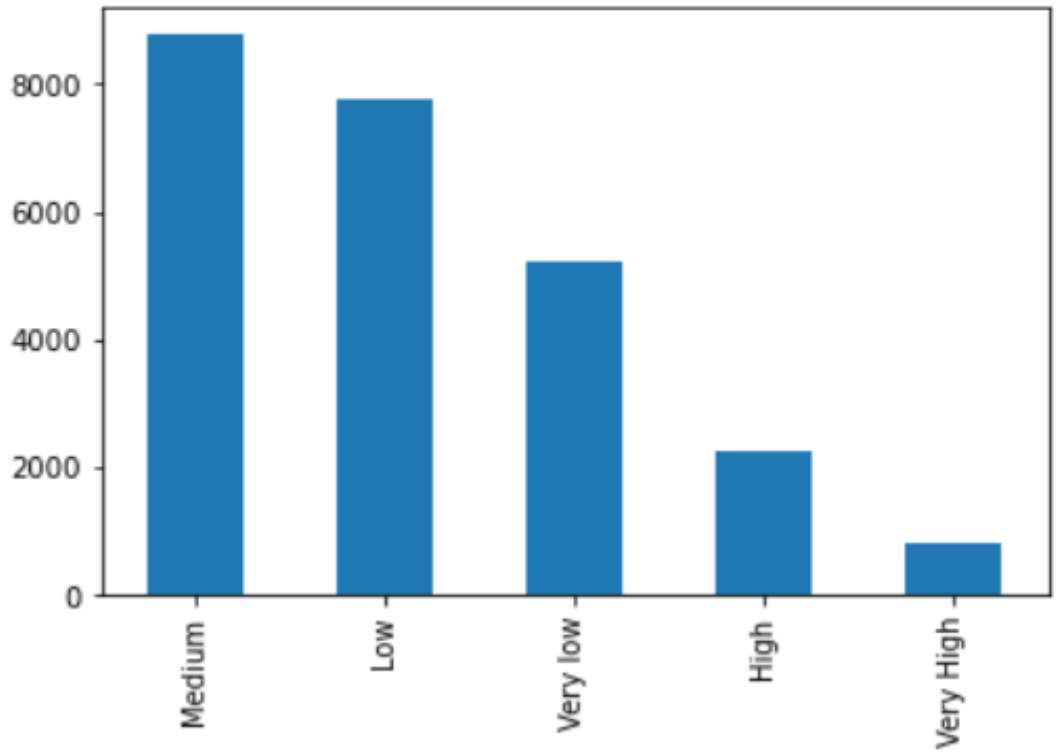


INCOME DISTRIBUTION OF LOAN APPLICANTS

Income distribution of Non Loan Payemnt Difficulties

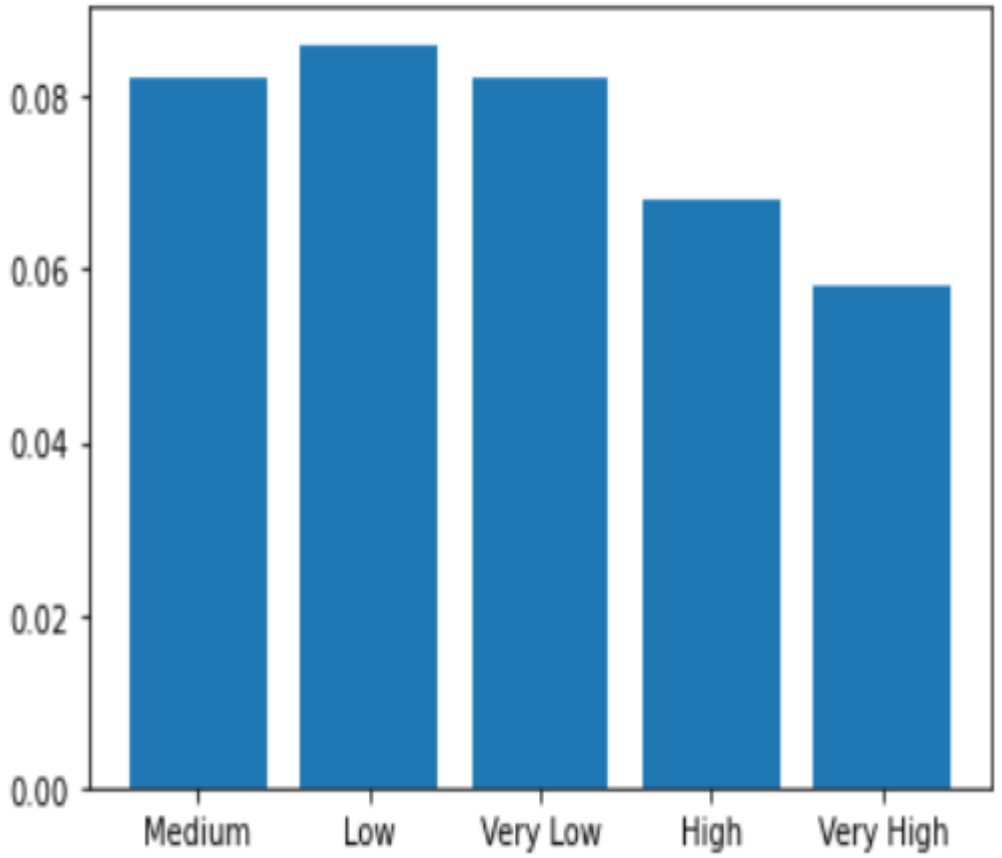


Income distribution of Loan Payemnt Difficulties



PERCENTAGE OF PEOPLE FAILING TO PAY LOAN UNDER INCOME CATEGORY

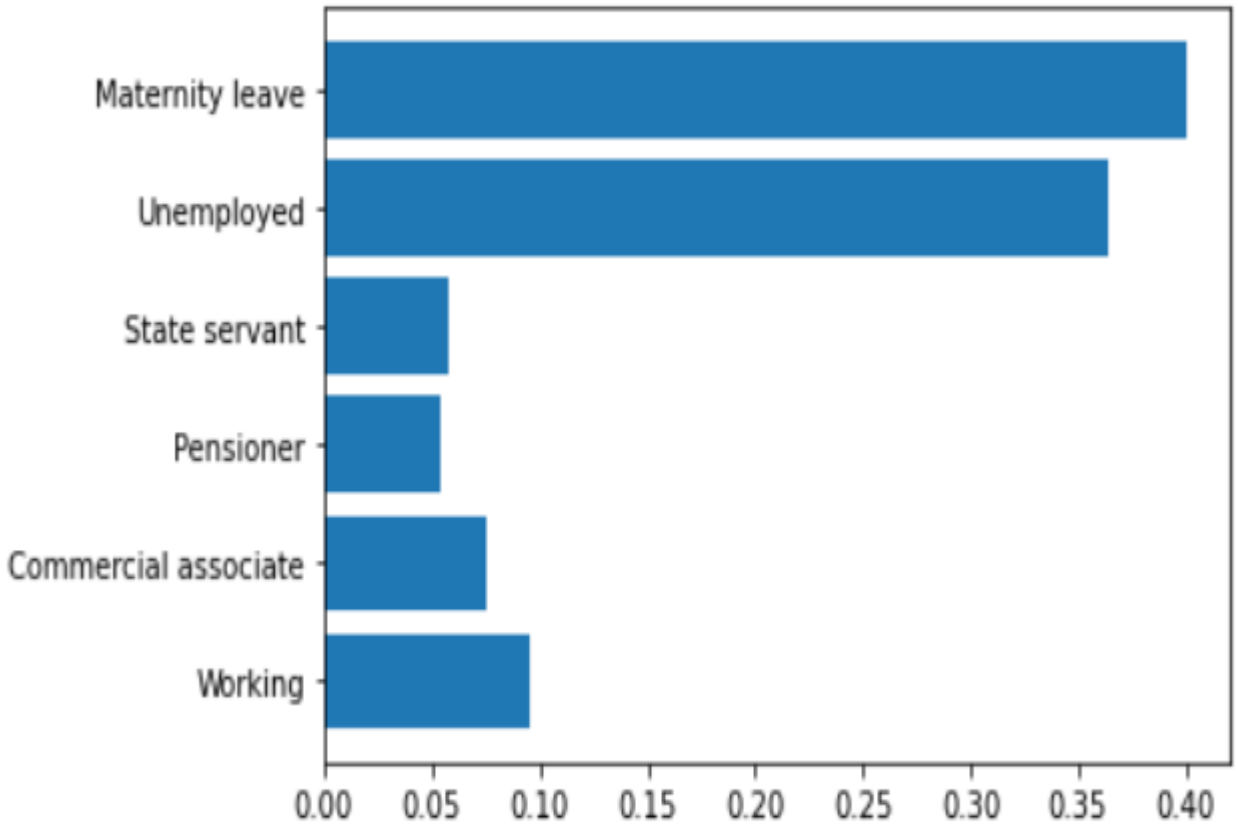
From this bar chart, we can observe, people belonging to high and Very High category in RANGE_AMT_INCOME has not faced much problem in loan payment.



PERCENTAGE OF PEOPLE FAILING TO PAY THE LOAN UNDER EACH INCOME CATEGORY

Clients who are on Maternity leave and Unemployed have been facing difficulty in paying the loan amount.

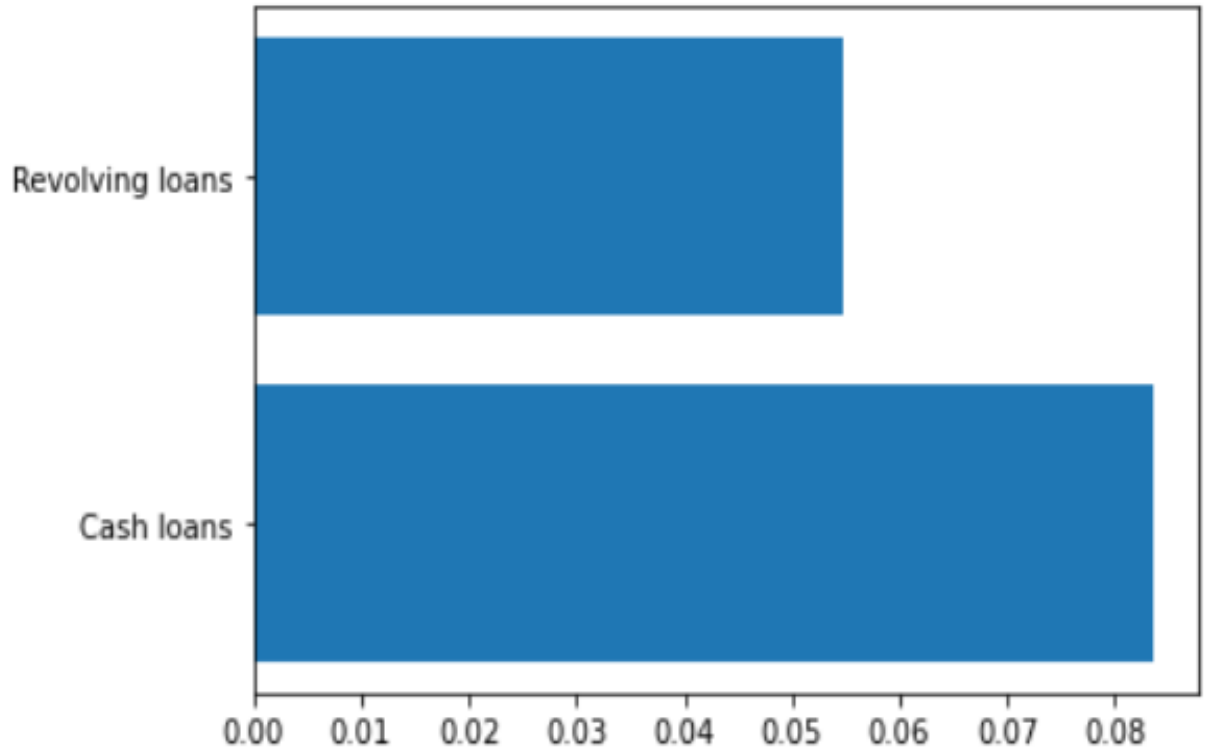
But the people who are working have not faced much problem in paying the loan.



PERCENTAGE OF PEOPLE FAILING TO PAY LOAN UNDER THE FOLLOWING TWO CATEGORIES IN CONTRACT

Clients who have taken revolving loans are not facing much challenge in loan payment.

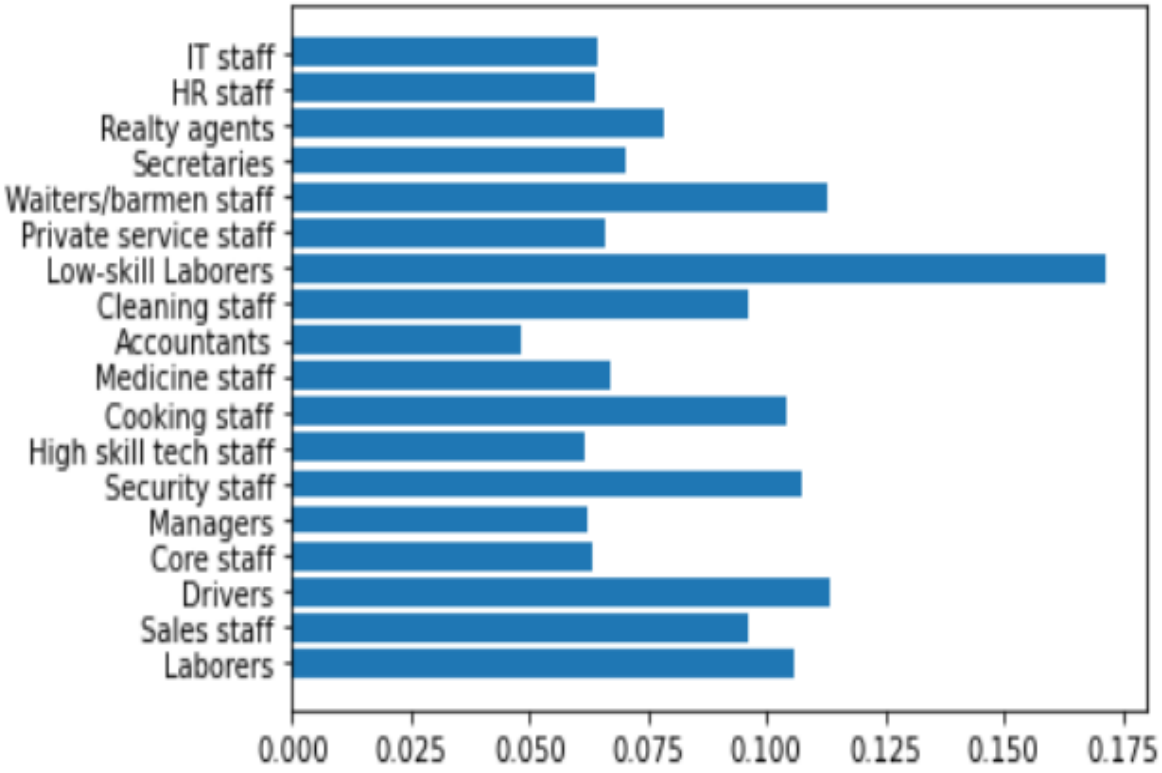
Whereas, on other hand, clients who have taken Cash loans, have been facing difficulty in loan payment.



PERCENTAGE OF PEOPLE FACING DIFICULTY TO PAY THE LOAN UNDER VARIOUS OCCUPATION CATEGORY

Low-skilled workers are not able to pay the loan amount regularly. Also, clients occupied in the category waiters are also facing a challenge in loan paying.

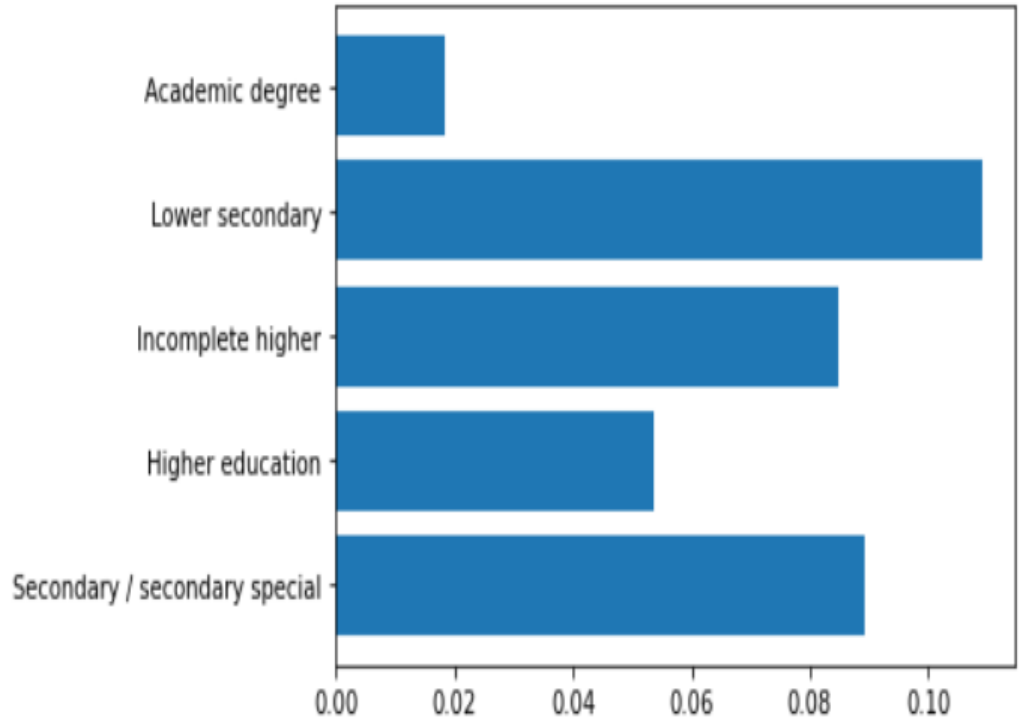
But, the people who are working as IT staff, Accountants, Managers are not facing much problem in paying the loan amount.



PERCENTAGE OF PEOPLE FACING DIFFICULTY TO PAY THE LOAN BASED ON THEIR EDUCATIONAL QUALIFICATION

People who have completed academic degree and higher education are not facing much problem in loan paying.

But the people who are in the category of Lower Secondary, Secondary / secondary special and Incomplete higher are facing a challenge to pay they loan amount.



PIVOT TABLE, with index as ‘CODE_GENDER’ and ‘RANGE_AMT_INCOME’ with columns as ‘NAME_HOUSING_TYPE’

	NAME_HOUSING_TYPE	Co-op apartment	House / apartment	Municipal apartment	Office apartment	Rented apartment	With parents
CODE_GENDER	RANGE_AMT_INCOME						
F	Very low	0.077381	0.071219	0.078868	0.064607	0.130316	0.116444
	Low	0.041860	0.071107	0.079445	0.084233	0.118478	0.101940
	Medium	0.078704	0.066224	0.068685	0.061896	0.115044	0.103129
	High	0.137931	0.057038	0.065621	0.043478	0.072398	0.083821
	Very High	0.117647	0.049137	0.071130	0.050847	0.041667	0.073864
M	Very low	0.071429	0.109391	0.154047	0.085366	0.106796	0.172869
	Low	0.124031	0.111245	0.139923	0.099010	0.145105	0.142665
	Medium	0.080214	0.100895	0.100681	0.079070	0.146779	0.123111
	High	0.072464	0.078455	0.096552	0.021186	0.104167	0.092348
	Very High	0.047619	0.063762	0.052846	0.022388	0.096296	0.109848

PIVOT TABLE with index as ‘CODE_GENDER’ and ‘RANGE_AMT_INCOME’ with column as ‘NAME_EDUCATION_TYPE’

CODE_GENDER	NAME_EDUCATION_TYPE / RANGE_AMT_INCOME	Academic degree	Higher education	Incomplete higher	Lower secondary	Secondary / secondary special
F	Very low	0.000000	0.056068	0.086399	0.080193	0.076778
	Low	0.000000	0.049022	0.080075	0.113889	0.079523
	Medium	0.000000	0.050254	0.078431	0.096983	0.075692
	High	0.105263	0.041516	0.074313	0.038961	0.070736
	Very High	0.076923	0.037289	0.082251	0.066667	0.065930
M	Very low	0.000000	0.080411	0.123967	0.125000	0.118066
	Low	0.000000	0.073305	0.097778	0.142857	0.123693
	Medium	0.000000	0.070086	0.095130	0.150515	0.113466
	High	0.000000	0.055911	0.074627	0.081633	0.093484
	Very High	0.000000	0.044080	0.077586	0.064516	0.089939

RECOMMENDED GROUP

- Belonging to the income category High and Very High
- Under income category Working and Commercial associate
- Who are willing to take revolving loans
- Are in occupation category IT staff, Accountants and Managers
- Have completed academic degree and higher education

RISKY GROUP

- Female applicants who are on maternity leave
- Who are unemployed
- Clients who are taking cash loans
- Low skilled workers
- Applicants who are not educates