

A decorative graphic on the left side of the slide consisting of two overlapping parallelograms. The front one is blue and the back one is a light green. They are positioned diagonally, with the blue one partially covering the green one.

Credit EDA Assignment

By **Ricky Chris Pinto**



Problem Statement

Main objective of this problem case study is to figure out the who among the applicants for a loan more likely to repay or default based on their past credit history as well as other factors.

Here we try to analyse the applicants with difficulty to repayment of loan based on their previous data and also the past applicants record, In order to analyse the risk of lending the loan and also to offer loan with higher interest rate to some potential candidates or to refuse the application of some risky candidates.



Data Understanding

There are mainly 3 datasets in our case study

- Application data set - which gives insights of the applicant whether he is defaulter or Repayer.
- Previous data ser - which tells us about previous application info of clients such as whether the application got Approved, Cancelled, Refused or Unused Offer.
- Column Description - which gives an overview of all the different columns present in our dataset.

Application data set :-

```
appData.head()
```


Out[47]:

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDI
0	100002	1	Cash loans	M	N	Y	0	202500.0	406597.
1	100003	0	Cash loans	F	N	N	0	270000.0	1293502.
2	100004	0	Revolving loans	M	Y	Y	0	67500.0	135000.
3	100006	0	Cash loans	F	N	Y	0	135000.0	312682.
4	100007	0	Cash loans	M	N	Y	0	121500.0	513000.



Application Data Set :-

- It consists of total of 307511 entries
- total 122 columns
- It consist of different types of data types.

- 
- Initially we check for null values in our data set.
 - We will remove the columns with more than 50% of missing values as they will affect our analysis.
 - We found that 41 columns which had null values more than 50% and hence dropped them.
 - We also found some 7 columns with more than 40% of missing values which had no meaning for our analysis so we dropped them



Handling Missing Values :

- There are over 30% missing values in occupation type so will replace them with a new category called '**Others**'.
- Number of enquiries to Credit Bureau about the client with different time period have around 13% missing values so we replace them with median of the particular column of our data set.
- EXT_SOURCE_3 which seems to have no correlation between our target variable so we can drop this

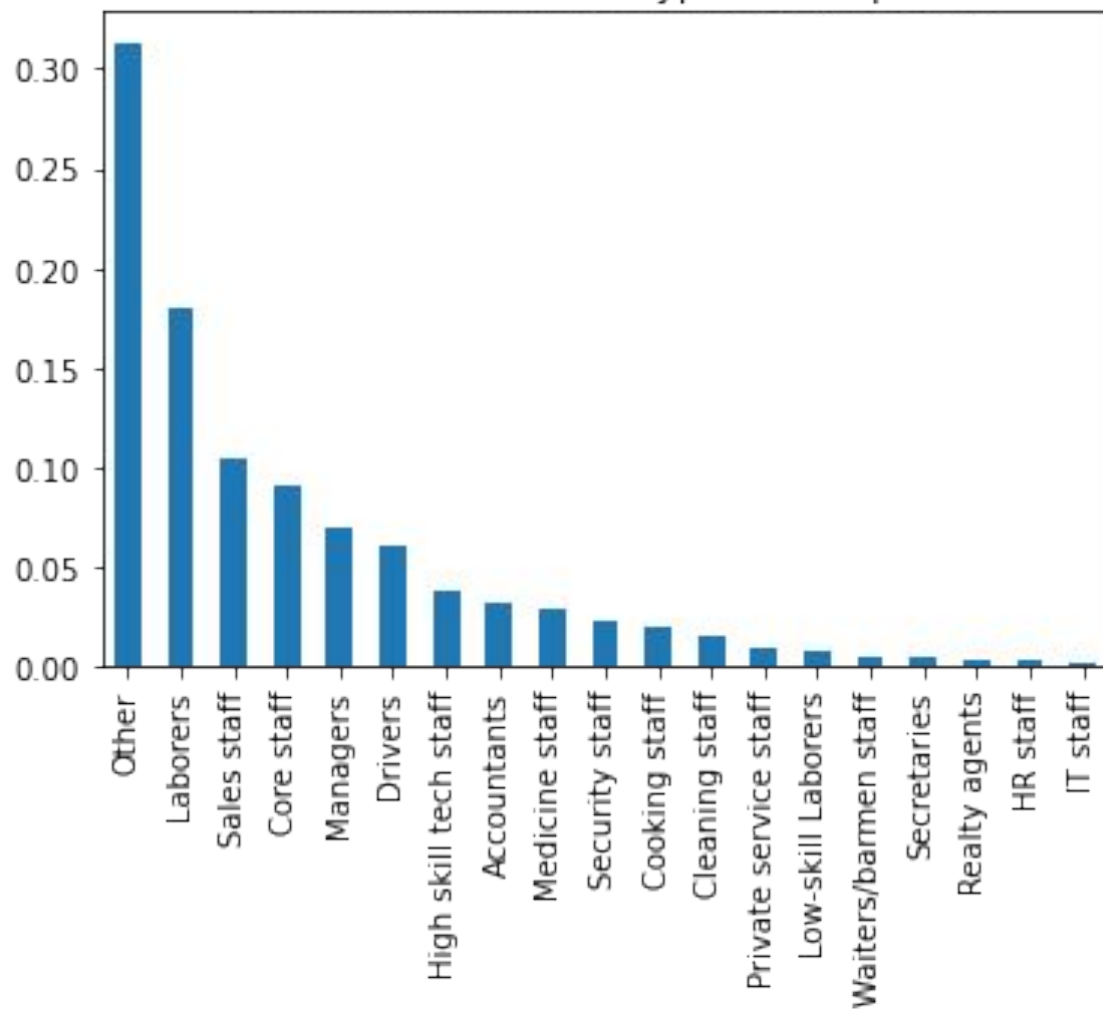
Correlation value = **-0.1789**



Value count of different type of Occupation

- As we observe in previous figure the category Others followed by Laborers have the highest number of applicants in our data set.
- Whereas in IT staff as well in HR staff has very low counts.

Value count of different type of Occupation





Handling improper values

Some of the date fields have negative values so we need to correct them,

- Applying the abs method on date so that to convert every value to a positive integer.
- Dividing ***DAYS_BIRTH*** by 365 to get a new column of age of the applicant.
- Then we group the age into different category to analyse it.

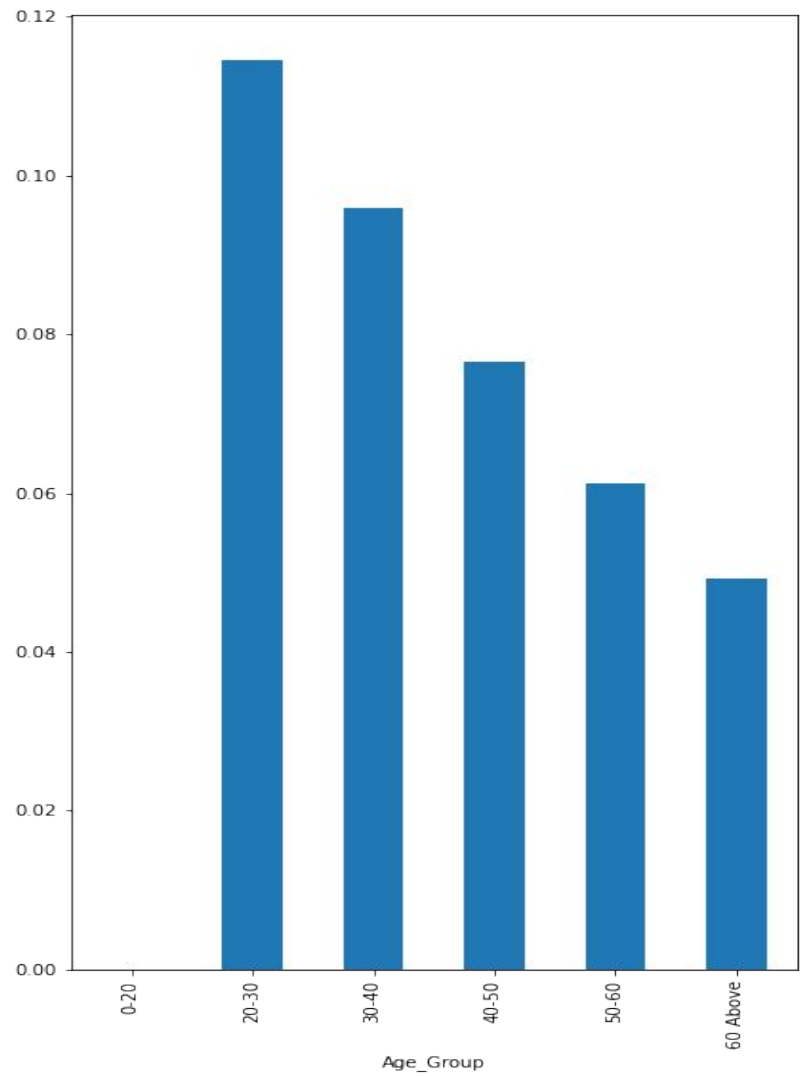
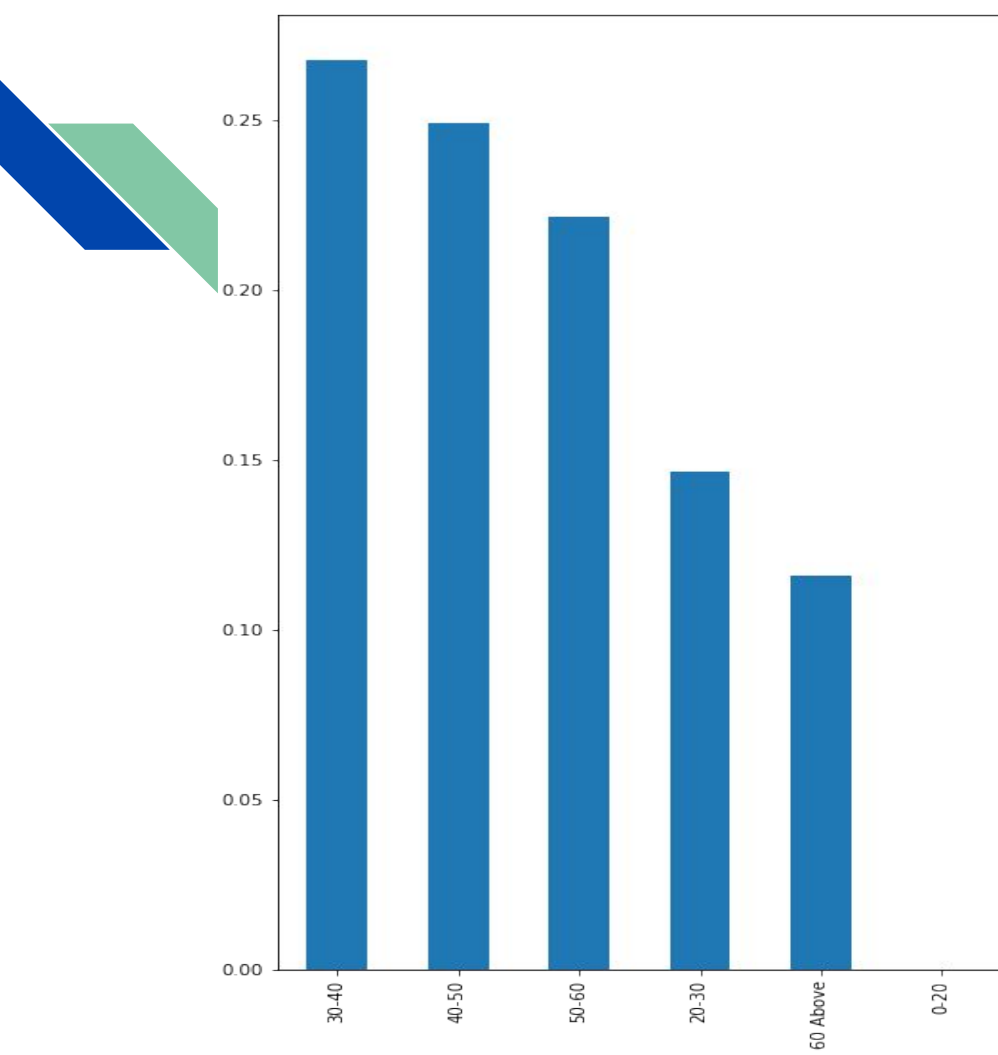


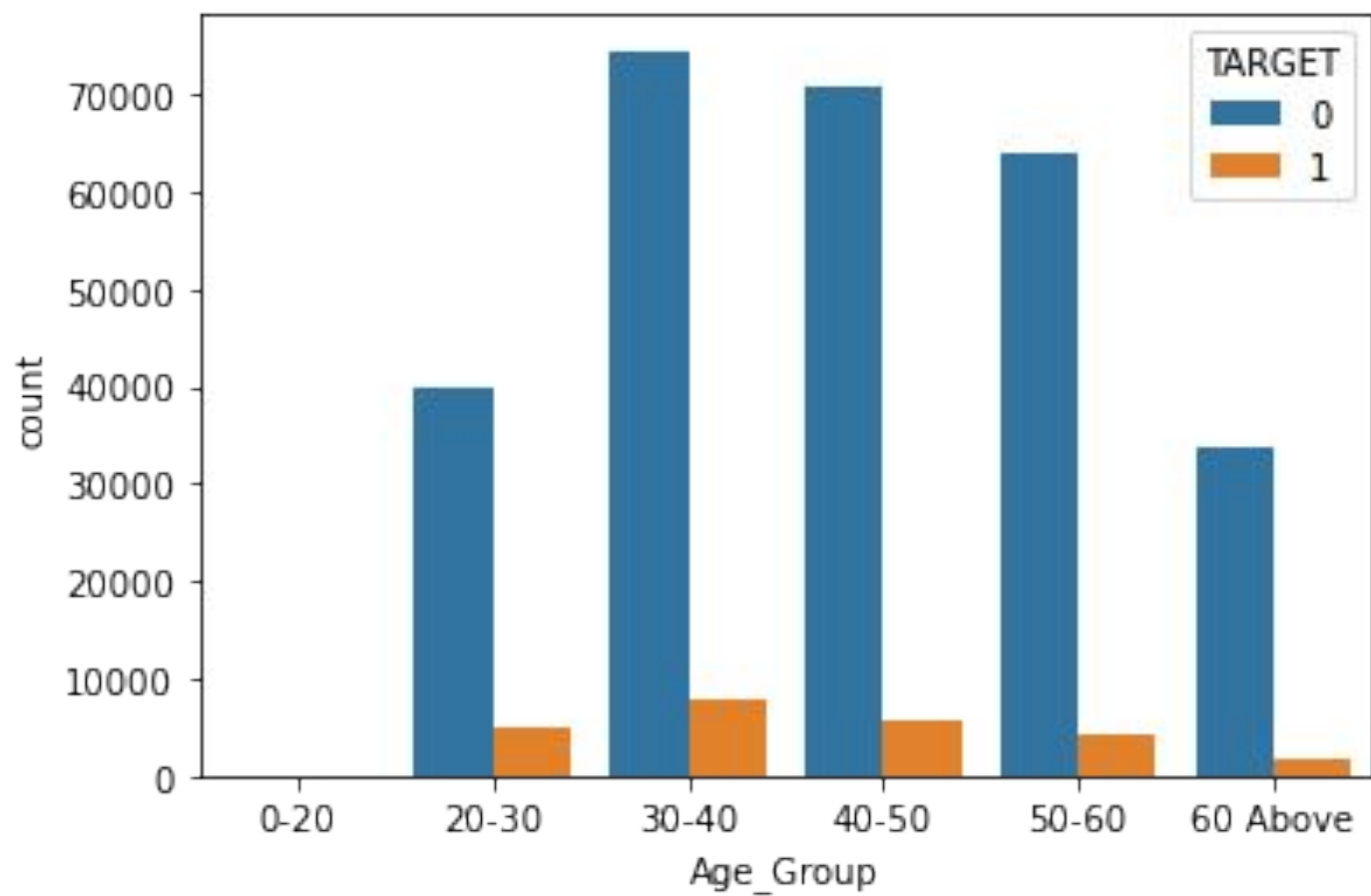
Age Vs Target

Figure on the left shows the count of applicants in various age groups and on right shows the percentage of defaulters or applicants having payment difficulties.

We can observe that

- 20-30 age group have more payment difficulties.
- Age group of 30-40 yr old are having the highest number of people







Education Type Vs Target

Figure shows the applicants in various education background who are having payment difficulties.

Observations :

- As we can observe that people with high education like an academic degree are less likely to default
- lower secondary has more rate of defaulting

NAME_EDUCATION_TYPE

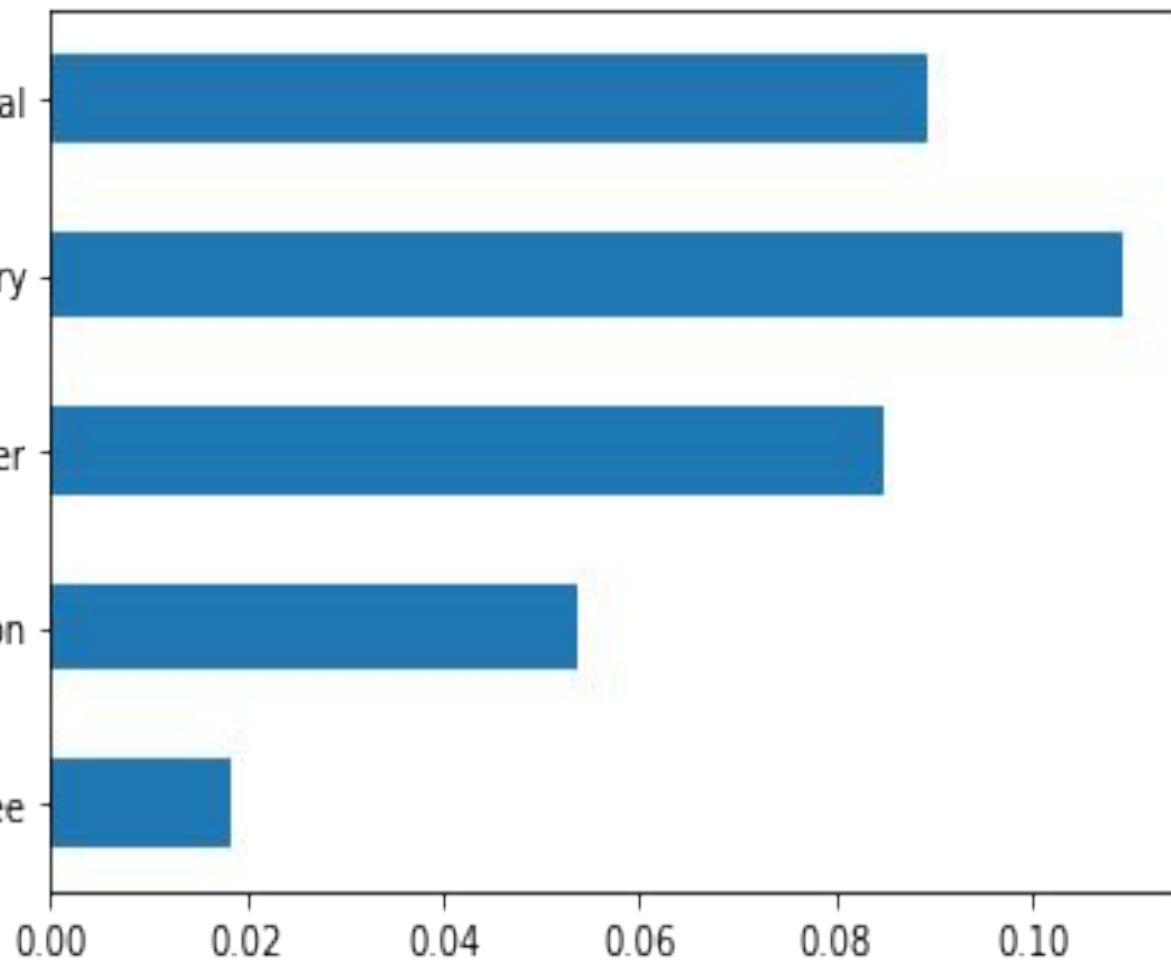
Secondary / secondary special

Lower secondary

Incomplete higher

Higher education

Academic degree





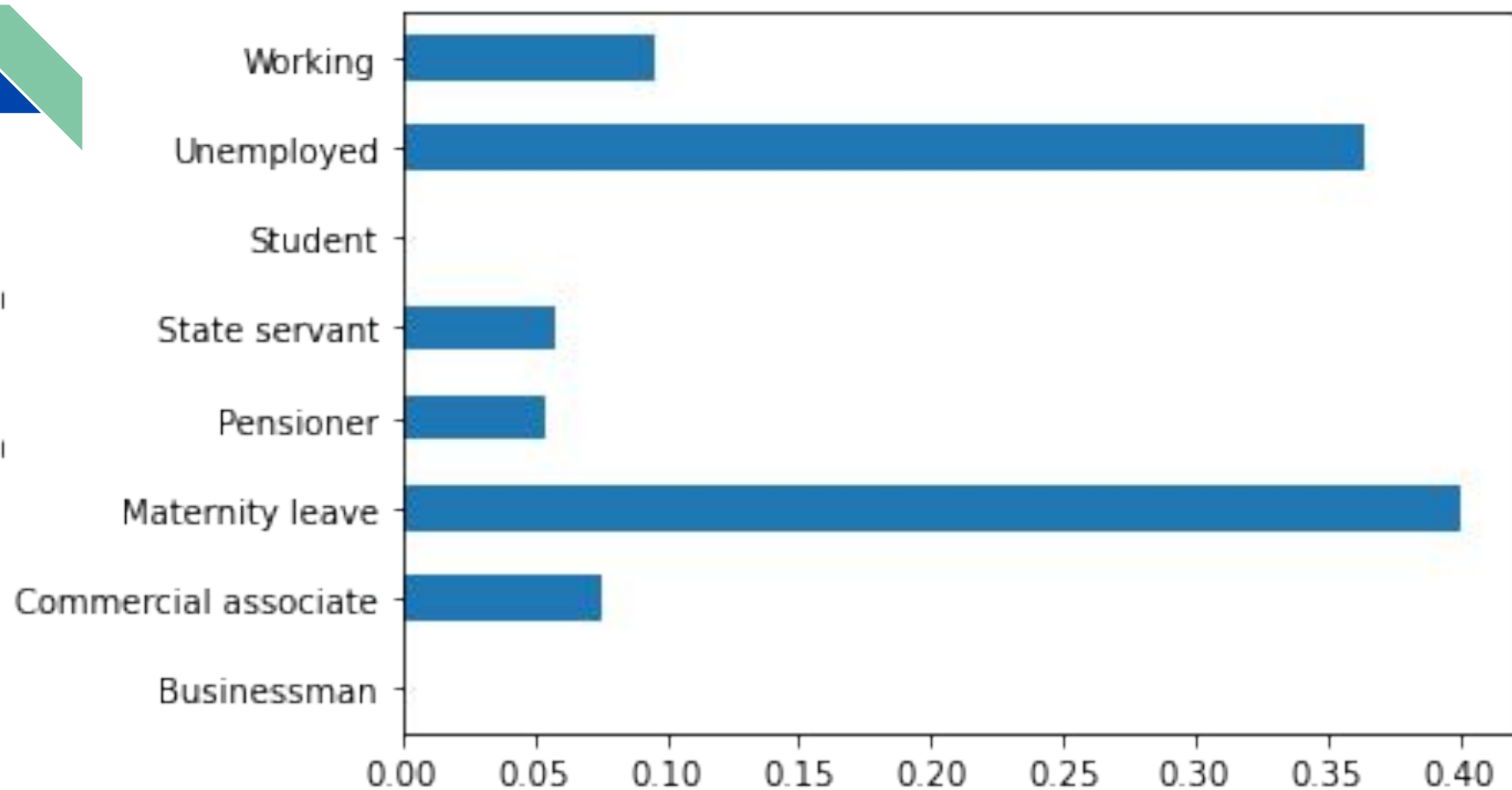
Income Type Vs Target

Figure shows the applicants in different income who are having payment difficulties.

Observations :

- Students and Businessmen don't have any defaulters in our data set
- mainly people who are not work are more likely to default

NAME_INCOME_TYPE



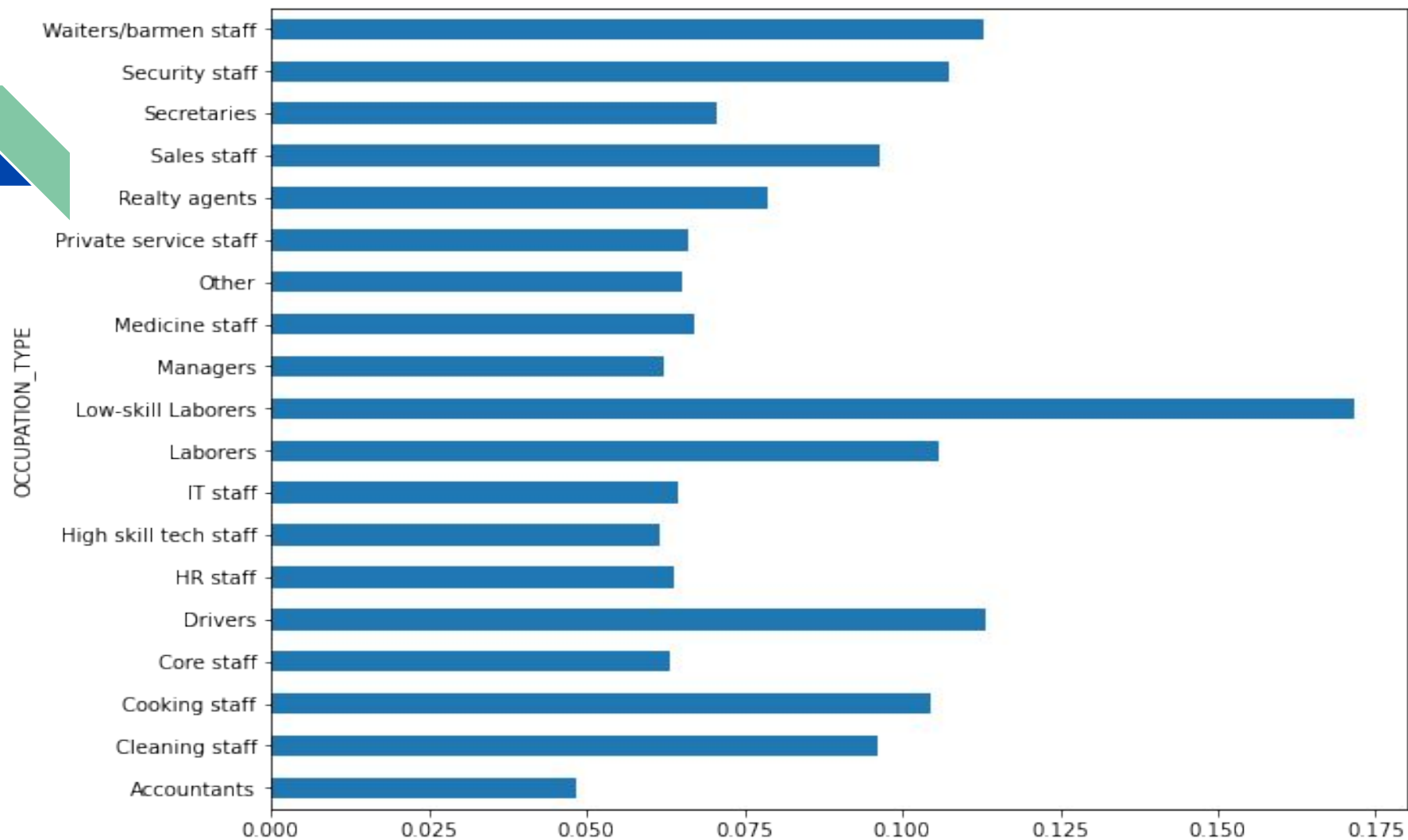


Occupation Type Vs Target

Figure shows the applicants with different Occupation who are having payment difficulties.

Observations :

- Accountants have a low rate of defaulting.
- Low-skill Laborers are majority of the defaulters.
- people with low education seems to be in the defaulter category.



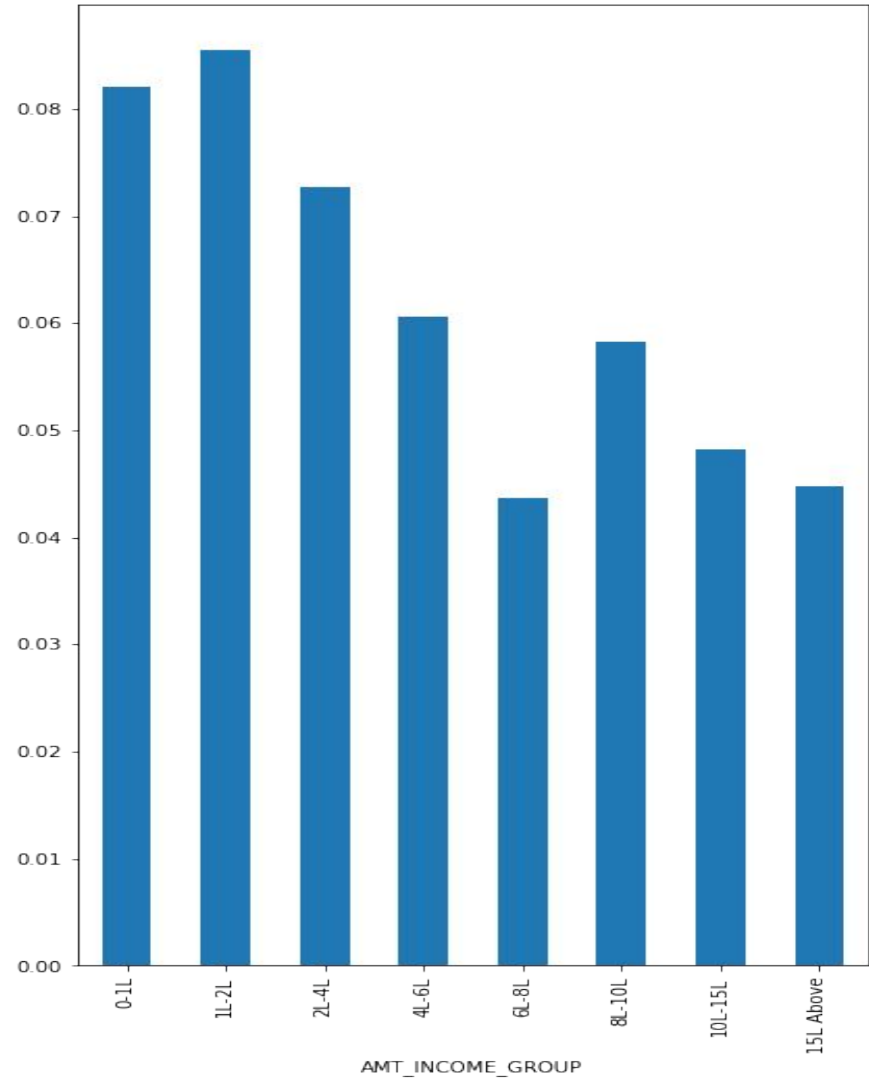
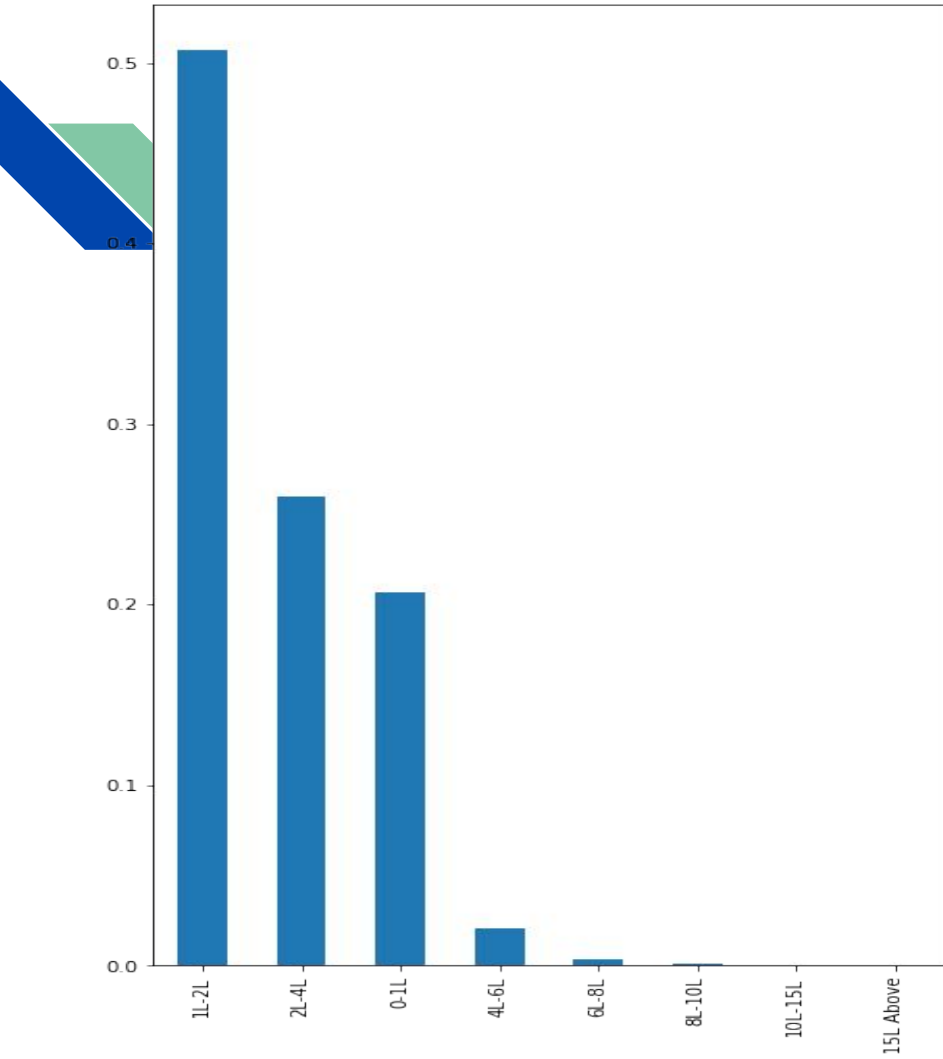


Income Vs Target

We will divide the income of applicants into various bins so we could easily group them and analyse.

Fig on the right shows Observations :

- People with low income (1-2Lk and 0-1LK) have high rate of default.
- People with more income more than 10 LKS very unlikely to default.

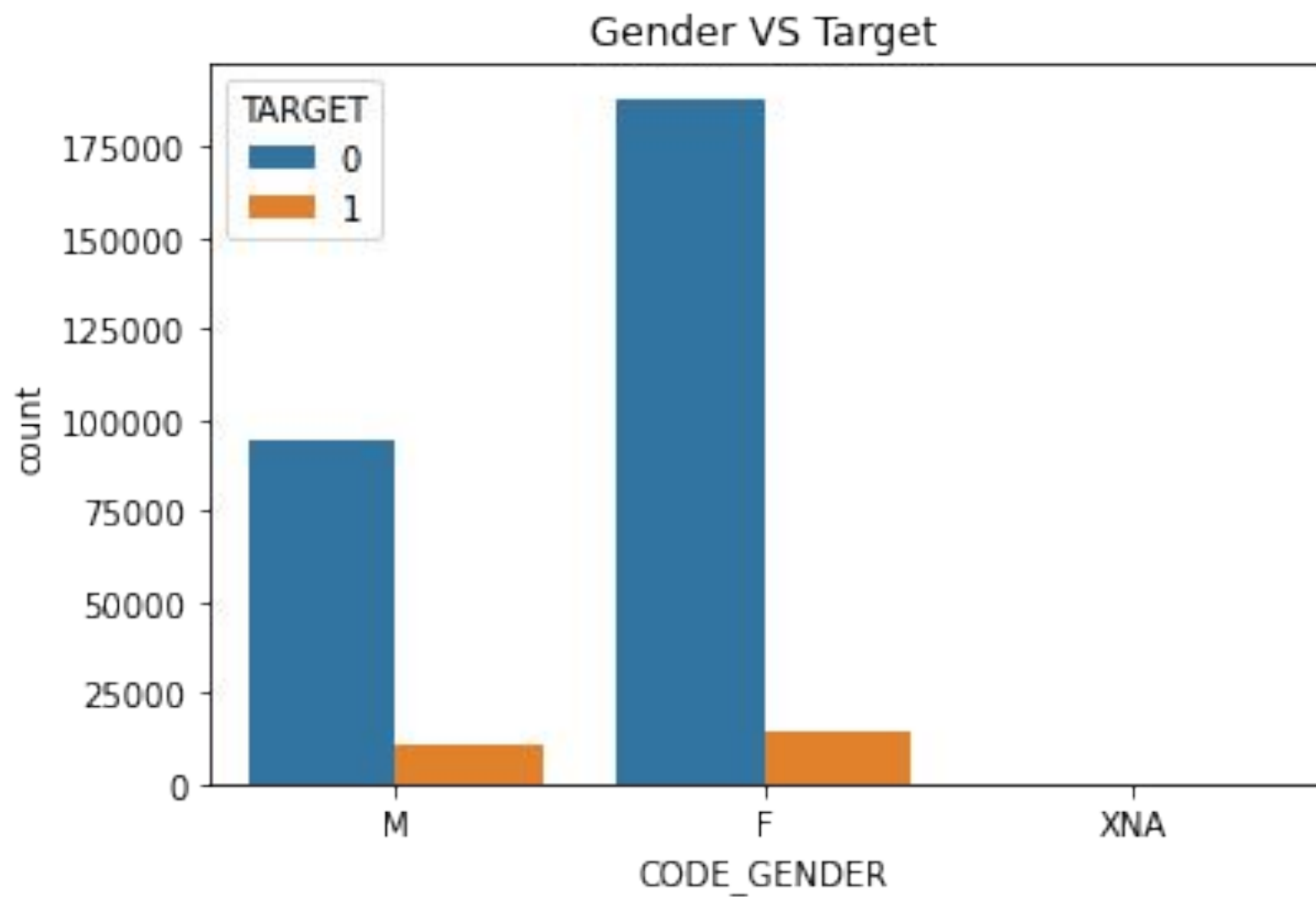




Gender comparison

Among the genders,

- Though the Female count is more Men are more likely to default



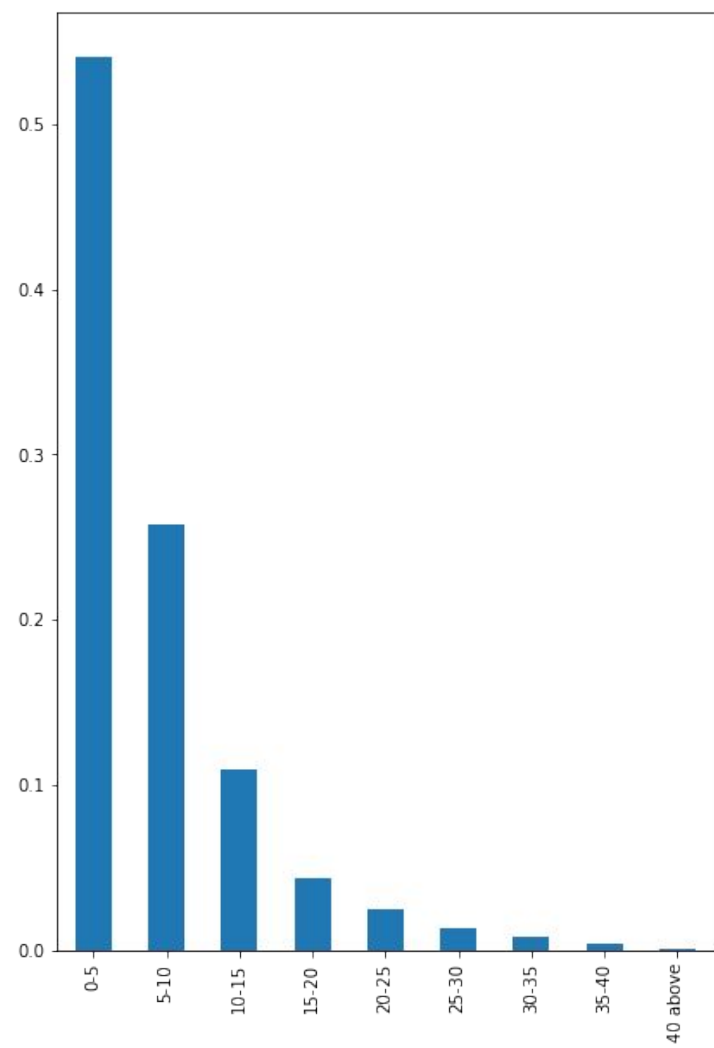
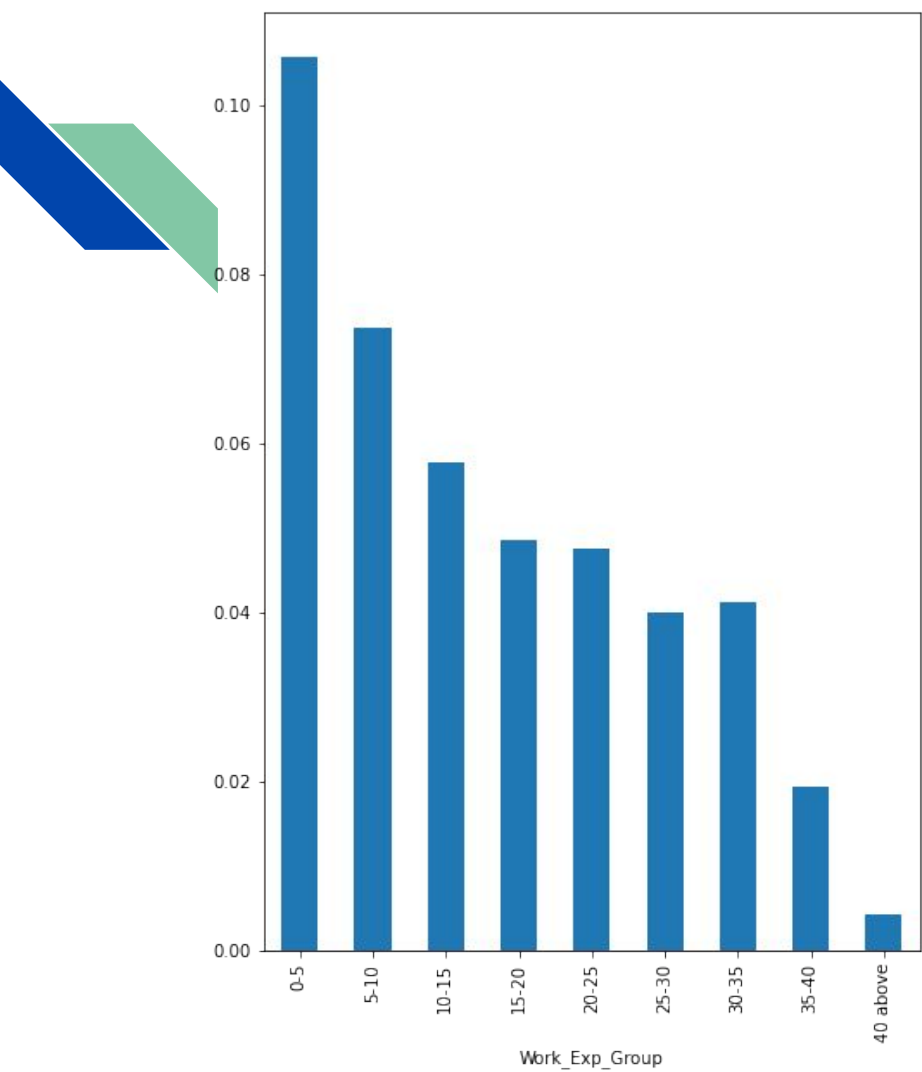


Work Experience vs default rate

Figure on the left shows the default rate of applicants across different groups of work experience that we have created for our analysis.

Observations:

- People with very low experience are more likely to default.
- 40+ yr of experience tend to have very less count as well as low default rate.





Previous application dataset

- It consists of 1670214 entries.
- Shape - (1670214, 37)
- Total of 37 columns
- Consists of data of various data types.



Outliers

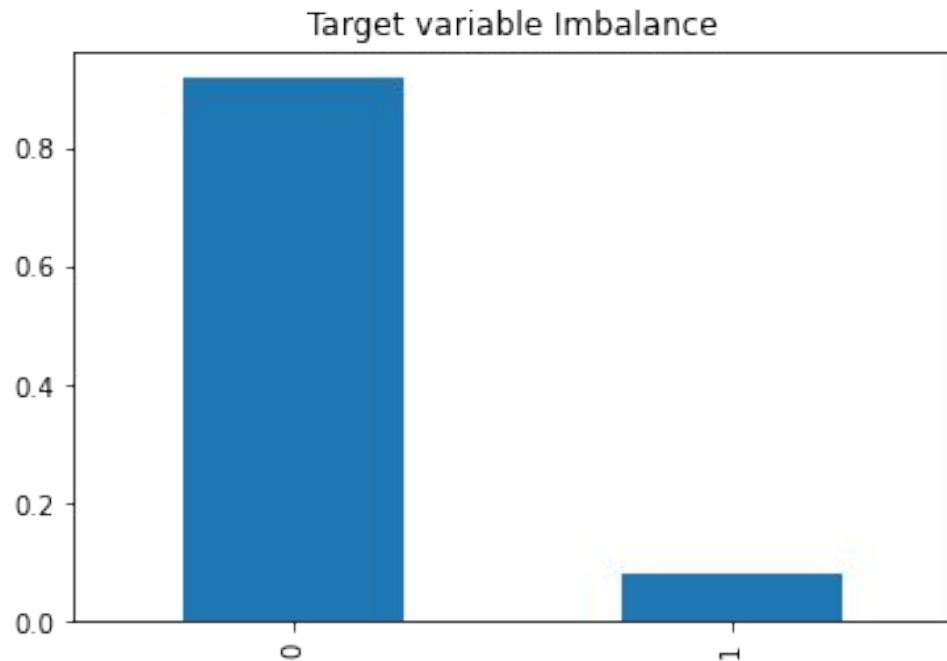
- Count children have some outliers as high as 18.
- days employed has one entry of 350000 which can be approx 959yrs which is an incorrect entry.
- days registration have many outliers
- income also have outliers but it can be a valid entry.
- credit also has outliers mainly because of the huge loans taken by business owners.



we can observe from the above figure that,

- Over 91% of the entries are of repayers
- and 8% are defaulters.

Imbalance ratio can be found by this i.e - **11.387**





Conclusion:

There are number of factors based on which bank can decide whether an applicant will repay the loan or default from it.

Applicants who are likely to repay the loan.

1. People with age more than 60 have less chance of defaulting.
2. People with high education like an academic degree are less likely to default
3. Students and Businessmen do not have any defaulters in our data set and most likely to repay the loan.
4. Accountants have aa low rate of defaulting.
5. Trade type-4 and industry type 12 have a low rate mostly business personal default very less.
6. People with more income more than 10 LKS very unlikely to default.
7. 40+ yr of experience tend to have very less count as well as low default rate.



People with payment difficulty (defaulters)

- 20-30 age group have more payment difficulties.
- Applicants with education such as lower secondary has more rate of defaulting.
- Mainly people who are don't work are more likely to default.
- Low-skill Laborers are majority of the defaulters.
- people with low education seems to be in the defaulter category.
- Transport type-3 has the highest rate of defaults.
- People with low income (1-2Lk and 0-1LK) have high rate of default.
- Though the Female count is more Men are more likely to default.
- Civil marriage and Singles have more rate of defaulting
- People with less experience are more likely to default.



Applicants who can be considered for a high interest rate loan as they can be considered risky

- People who have taken loan between the range 4-6L have most default rate.
Hence these applicants should be offered loan at high interest rates.
- Same can be done with the applicants whose income is low. Like the applicants with 2-4L of income can be considered for this category.