Credit EDA Assignment

By Akshay P

Problem Statement

- A financial organization wants to know the driving factors behind loan defaulting so that they can utilise these variables to reduce the risk of loan default.
- Identifying such strong indicators will help them to reduce the loan amount/avoid approving loans to applicants with payment difficulties.
- It can also be used to minimize rejecting loan applications of people who are actually capable of repaying.

- Let's try to identify such factors

We have the following data:

1. 'application_data.csv' contains all the information of the client at the time of application. This 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.

Workflow

- Stored 'application_data.csv' file as inp0

1. Fixing rows and columns

- Removed columns containing more than 50% missing values, from the dataframe inp0.
- Rows with incorrect data were removed

Workflow

2. Handling missing values

- Handling missing values in numerical columns

  If the mean and median of the data were same or nearby(assumes that the data follows a normal distribution), the missing values were imputed with mean. Otherwise the missing values were imputed with median.

- Handling missing values in categorical columns

  Missing values were imputed with the mode of the categorical columns

  Missing values were kept as it is, where imputing may cause exaggeration of data.

3. Handling data types of the variables

- Dtypes of columns were converted into appropriate data types.

4 .Handling the outliers

- Outliers of variables were identified and were either imputed or capped.

5. Standardising the variables

6. Checking for data imbalance

7.Univariate analysis

8. Finding top 10 correlations

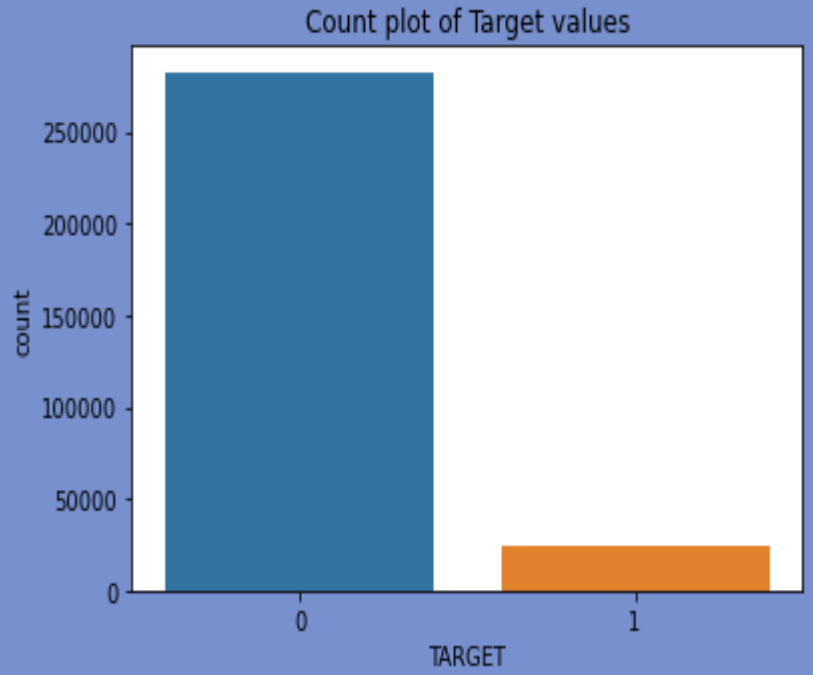
9. Bi/Multivariate analysis

- Correlation of Target values with other boolean columns(variables which had either 1 or 0 as values)were checked.
- Numeric- numeric analyses were done using scatter plot / correlation matrices
- Numeric-categorical variables were analyzed using bar charts and boxplots.
- Categorical-categorical analyses were done using bar charts

10. Reading previous_application.csv file as inp1

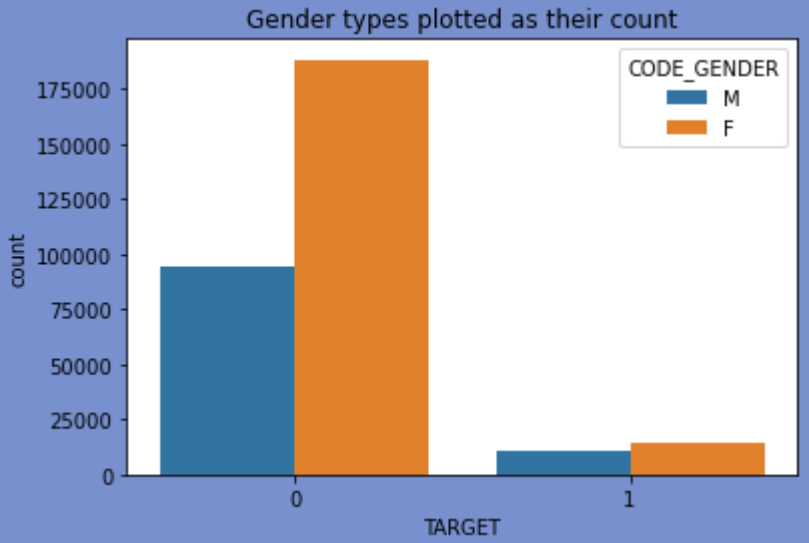
- Merged the two data frames inp0 and inp1 into final_df
- Few more analysis were performed on final_df.

Checking for data imbalance



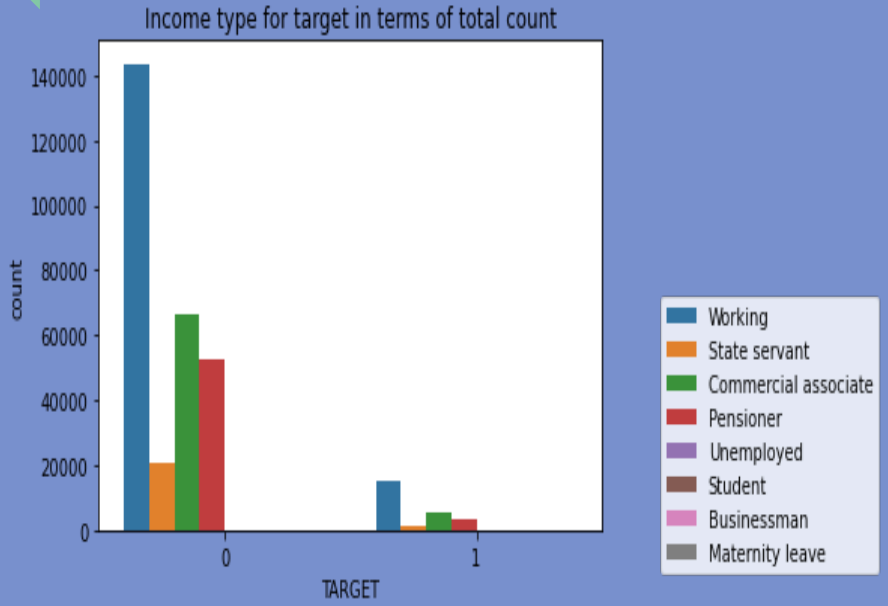
- ➔ The data is imbalanced
- ➔ Among the applicants 91.9% of them are non defaulters and 8.1% of them had payment difficulties during the previous applications
- ➔ Ratio of imbalance is 11.39 after cleaning the dataset.

Univariate analysis- Gender type



- Most of the applicants are females (~66%)
- As females apply for loan more than males, the number of females among defaulters is also high

Univariate analysis- Income type



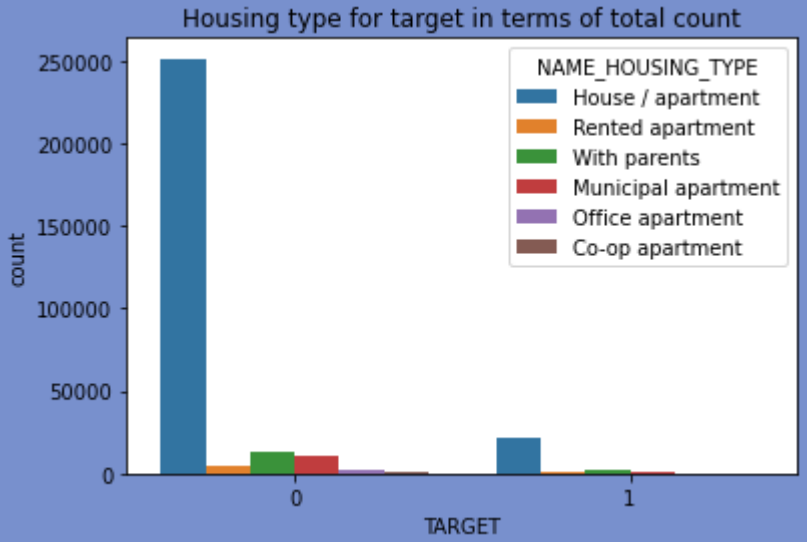
- Most of the applicants belong to working class or commercial associate class

- Among the applicants the number of people who are unemployed/student/businessman or on maternity leave are very small.

- As the number of applicants in income type categories increases, the number of defaulting also increases.

- Defaulting by students or businessmen are very less.

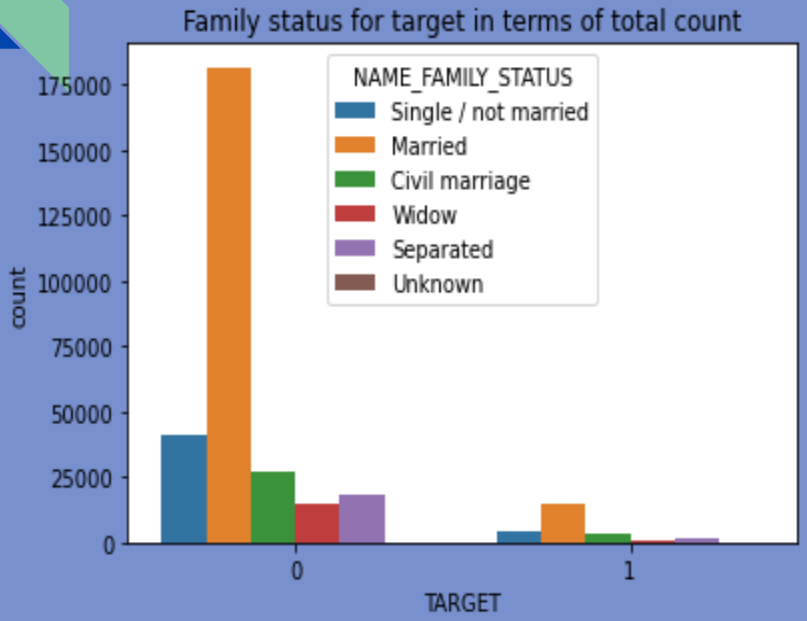
Univariate analysis- Housing type

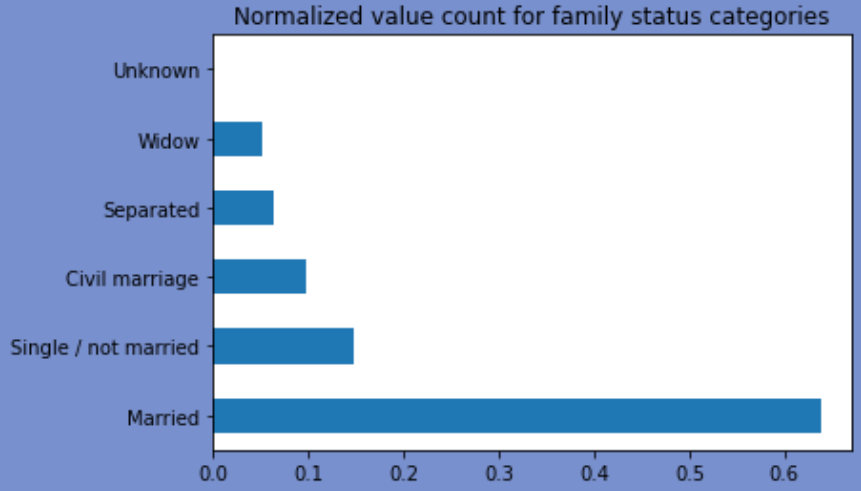


- Most of the applicants live in house/apartment and their number of defaulting is higher than applicants with other housing types

- Only 1.5% of applicants live in rented apartments.

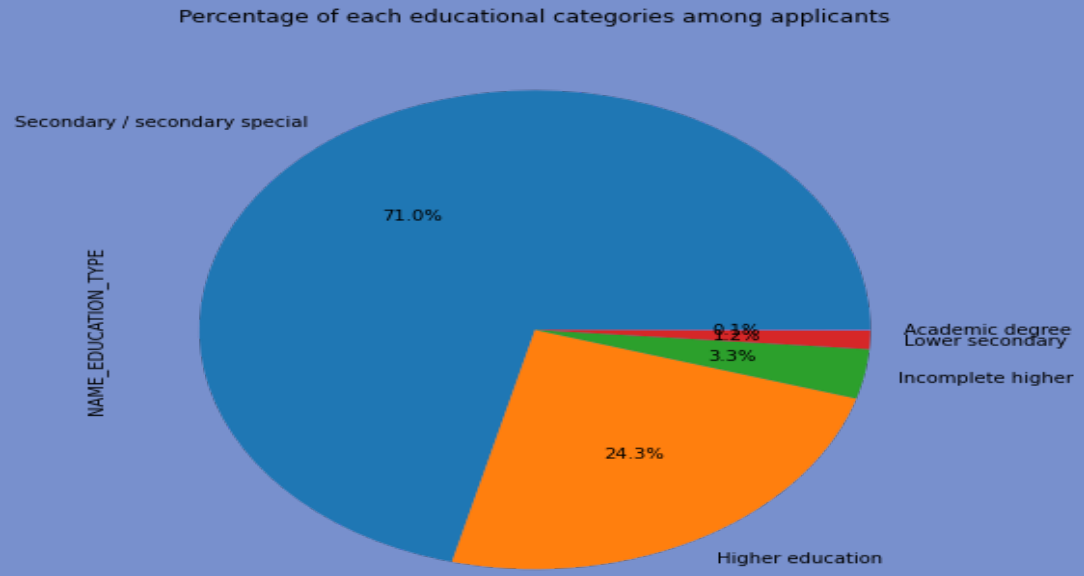
Univariate analysis- Family Status





- Married people are more likely to apply for loan.
- Around 64% of the applicants are married.

Univariate analysis- Education type



- People with secondary/secondary special level education have applied the most for the loan, and the least is by people with academic degree/lower secondary level education.

Top ten correlations- Among non defaulters

Top 10 correlation (in target0)

1. OBS_30_CNT_SOCIAL_CIRCLE~OBS_60_CNT_SOCIAL_CIRCLE 2. FLOORSMAX_AVG ~FLOORSMAX_MEDI

3. YEARS_BEGINEXPLUATATION_MEDI ~YEARS_BEGINEXPLUATATION_AVG 4. FLOORSMAX_MODE ~FLOORSMAX_MEDI

5. AMT_CREDIT ~AMT_GOODS_PRICE 6. FLOORSMAX_AVG ~FLOORSMAX_MODE

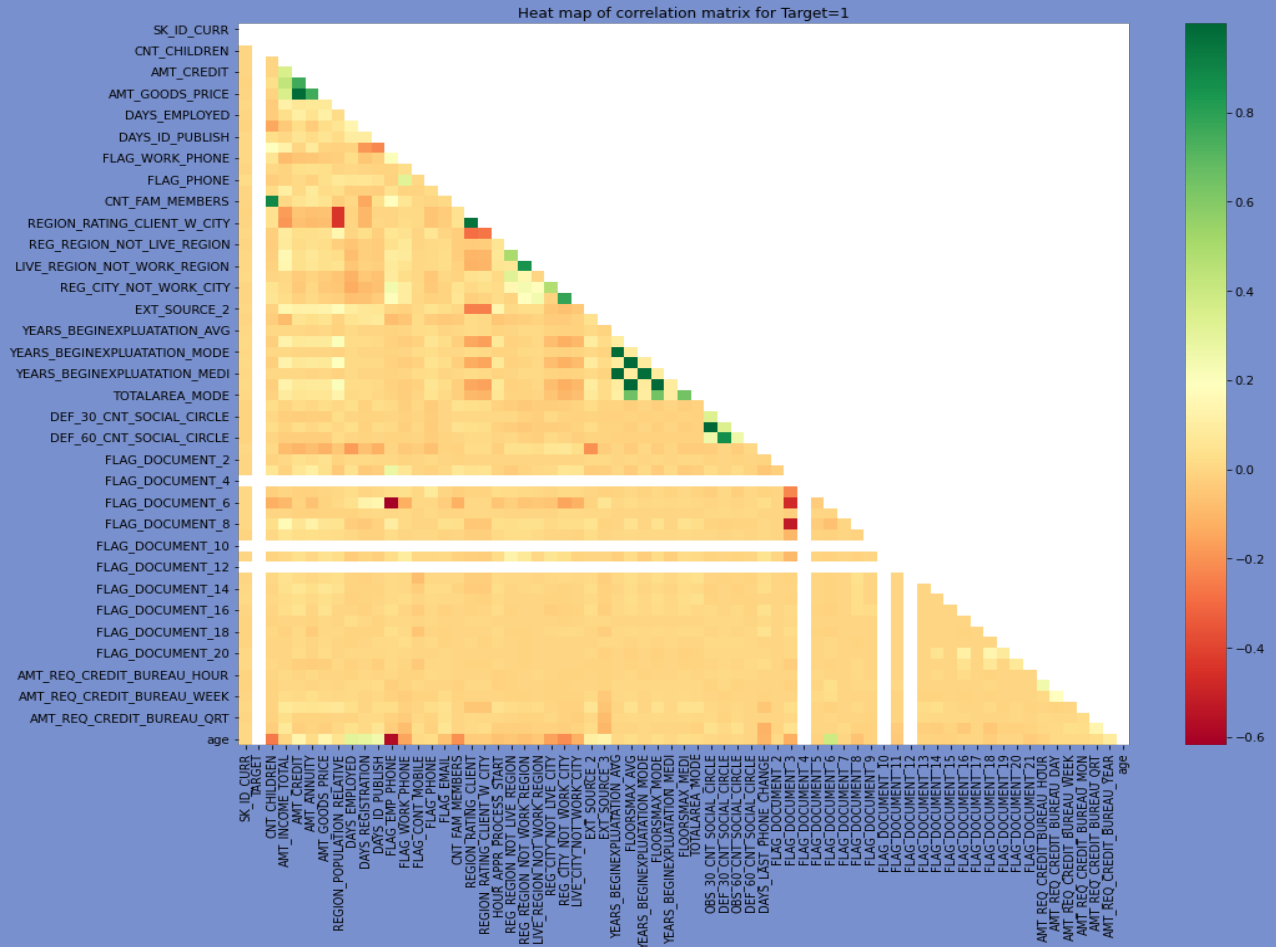
7. YEARS_BEGINEXPLUATATION_AVG ~YEARS_BEGINEXPLUATATION_MODE

8. YEARS_BEGINEXPLUATATION_MEDI ~YEARS_BEGINEXPLUATATION_MODE

9. REGION_RATING_CLIENT ~REGION_RATING_CLIENT_W_CITY

10. CNT_CHILDREN ~CNT_FAM_MEMBERS

Top ten correlations- Among defaulters



Top 10 correlation (in target1)

1. OBS_30_CNT_SOCIAL_CIRCLE ~OBS_60_CNT_SOCIAL_CIRCLE

2. FLOORSMAX_MEDI ~FLOORSMAX_AVG

3. YEARS_BEGINEXPLUATATION_MEDI ~YEARS_BEGINEXPLUATATION_AVG

4. FLOORSMAX_MEDI ~FLOORSMAX_MODE 5. FLOORSMAX_AVG ~FLOORSMAX_MODE

6. AMT_CREDIT ~AMT_GOODS_PRICE

7. YEARS_BEGINEXPLUATATION_AVG ~YEARS_BEGINEXPLUATATION_MODE

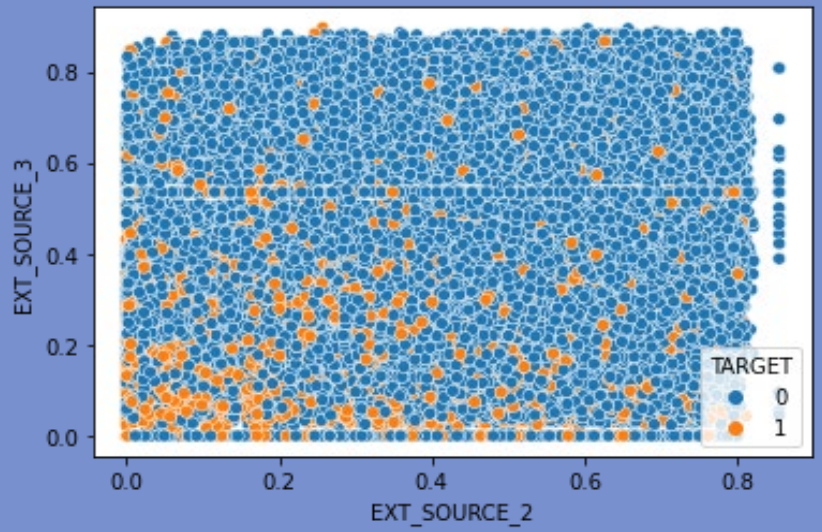
8. YEARS_BEGINEXPLUATATION_MEDI ~YEARS_BEGINEXPLUATATION_MODE

9. REGION_RATING_CLIENT ~REGION_RATING_CLIENT_W_CITY

10. CNT_CHILDREN ~CNT_FAM_MEMBERS

Bi/multi variate analysis

EXT_SOURCE_2 vs EXT_SOURCE_3

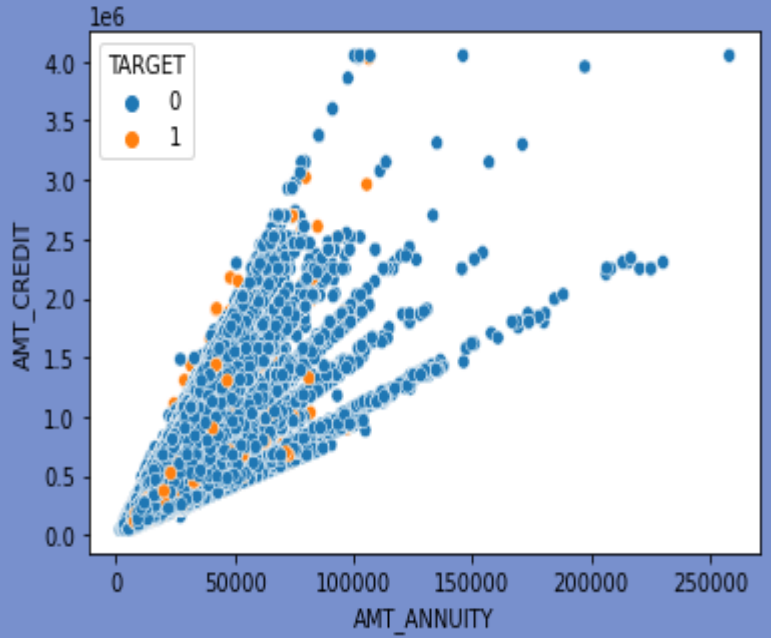


- The correlation between two external scores are not strong (correlation coefficient=0.094)

- However we could see that most of the people who had difficulty in repaying the amount have low external scores.

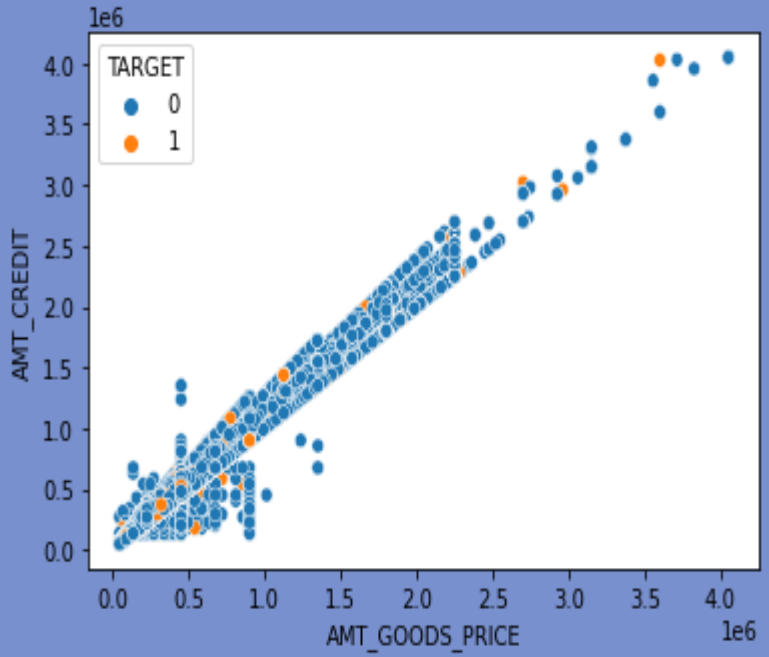
- So EXT_SOURCE_2 & EXT_SOURCE_3 are good indicators of defaulting.

AMT_ANNUITY vs AMT_CREDIT



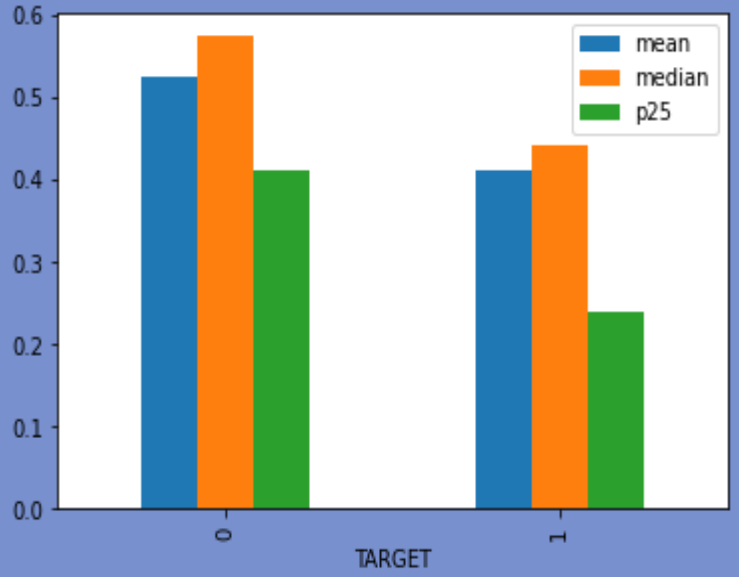
- As the loan annuity increases the credit amount also increases

AMT_CREDIT vs AMT_GOODS_PRICE



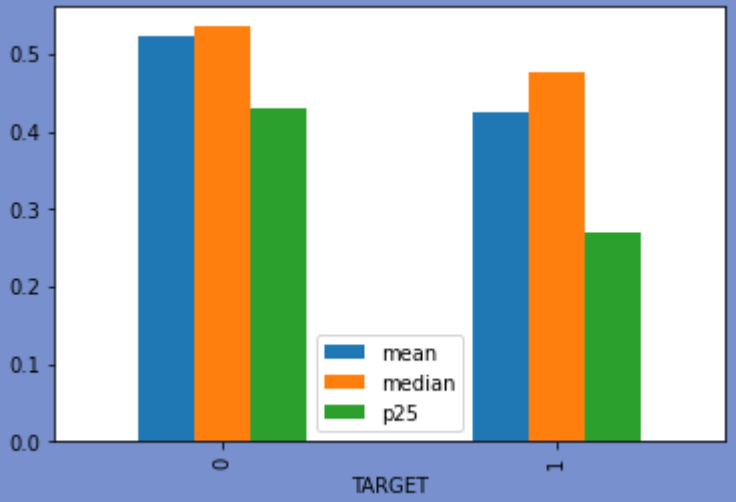
- The amount of loan taken increases with the goods price amount

EXT_SOURCE_2 vs TARGET



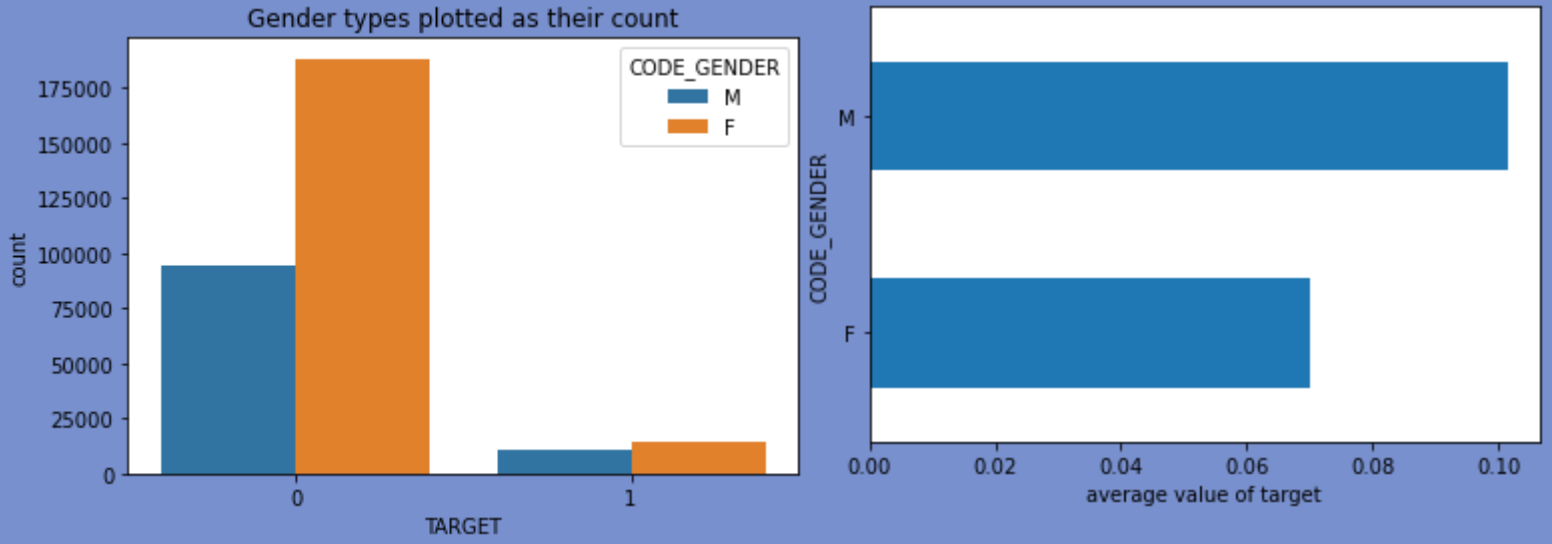
- People who had payment issues have mean, median and 25th percentile values of EXT_SOURCE_2 lower than that of people who paid on time

EXT_SOURCE_3 vs TARGET



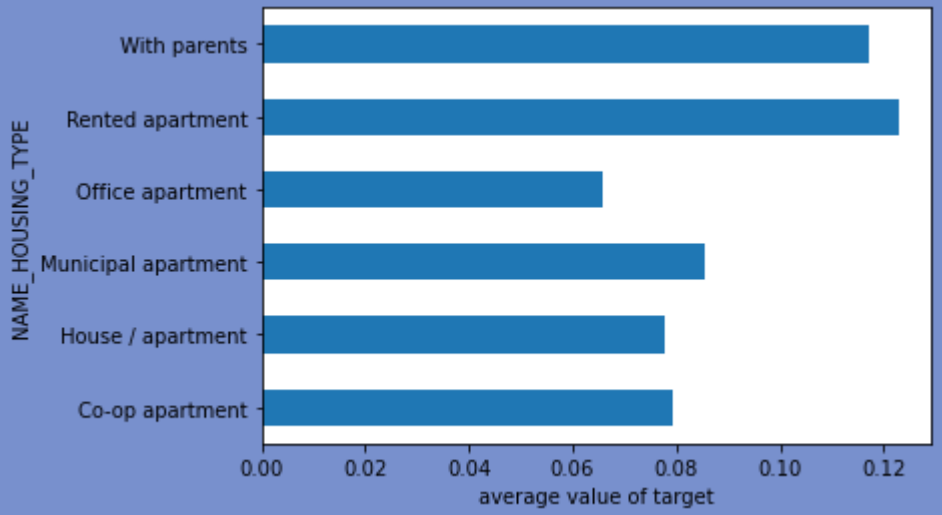
- People who had payment issues have mean, median and 25th percentile values of EXT_SOURCE_3 lower than that of people who paid on time

Gender type vs TARGET



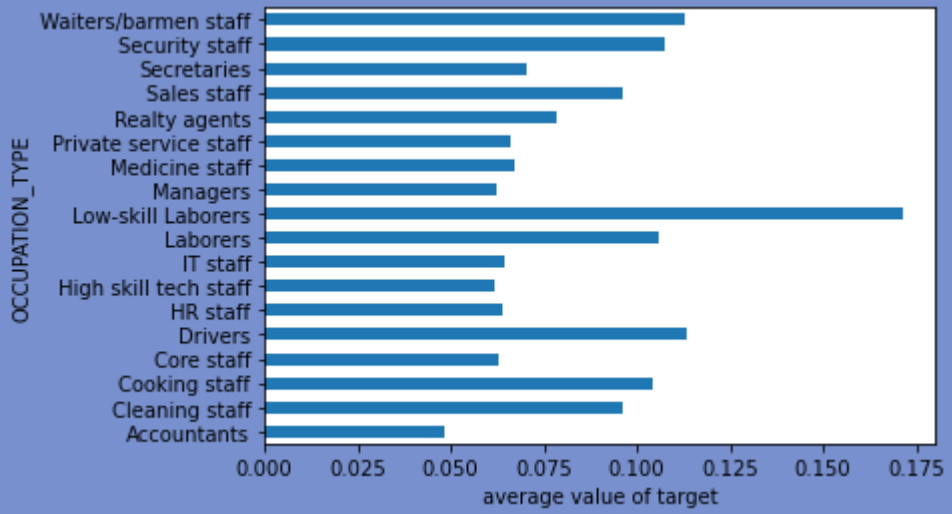
Even though the number of males apply for loan are less, the chances of males defaulting is higher than female applicants

NAME_HOUSING vs TARGET



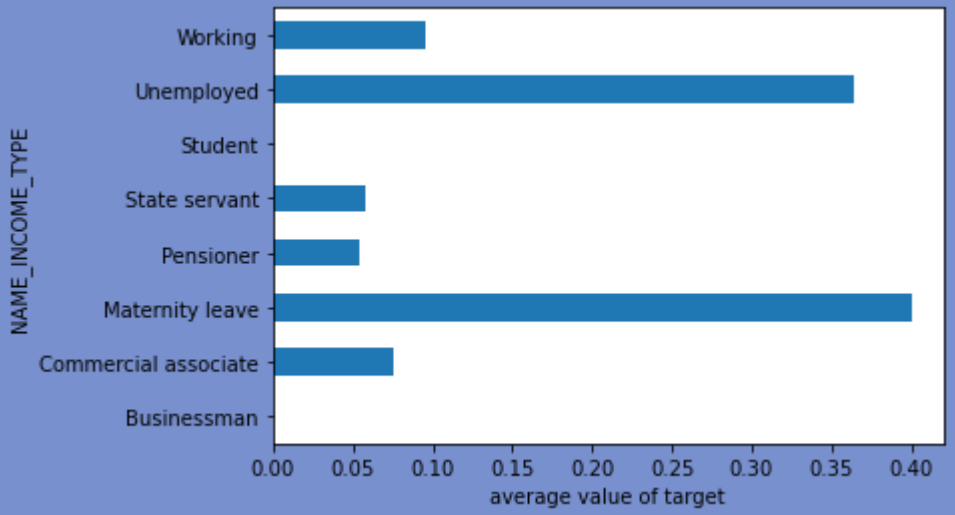
People who live in rented apartment or with their parents have more chances of defaulting

OCCUPATION_TYPE vs TARGET



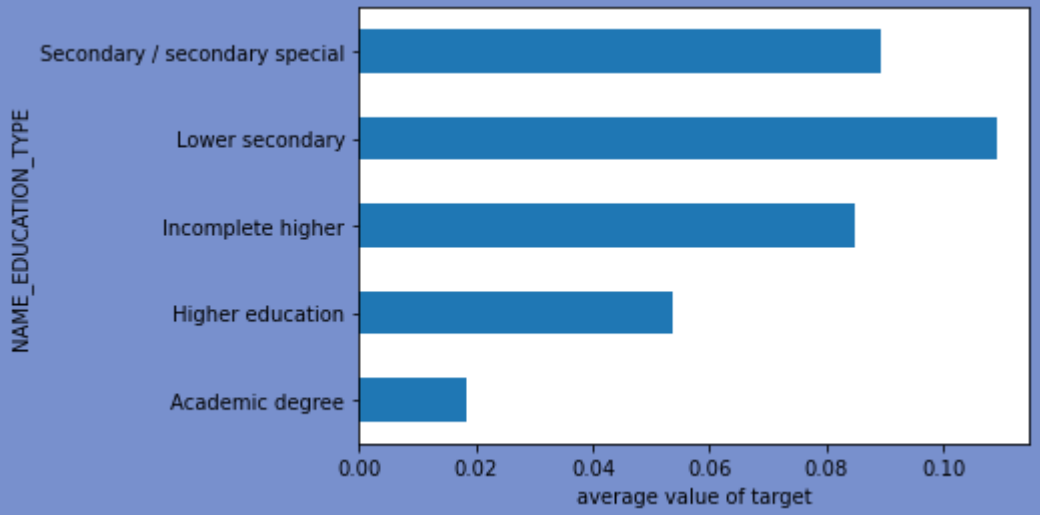
- Low skilled laborers are more likely to default

Income typeVS TARGET



- Unemployed applicants and those who are on maternity leave struggles to pay the amount back(more likely to default)

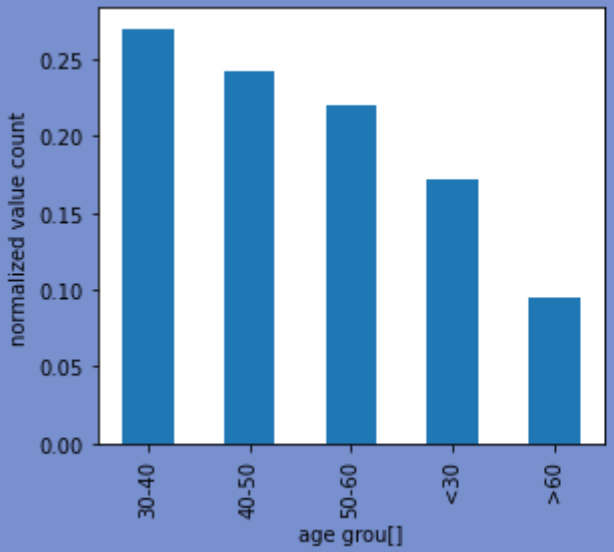
Education type vs Target

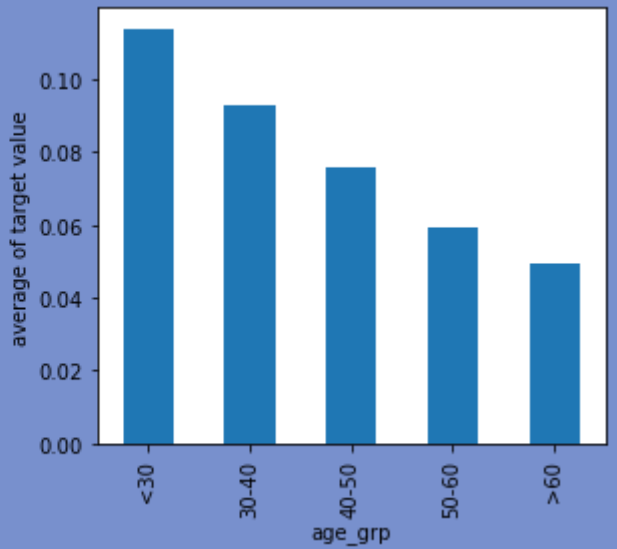


-Clients with academic degree are less likely to default

- Clients with lower secondary education are more likely to default

Age vs Target

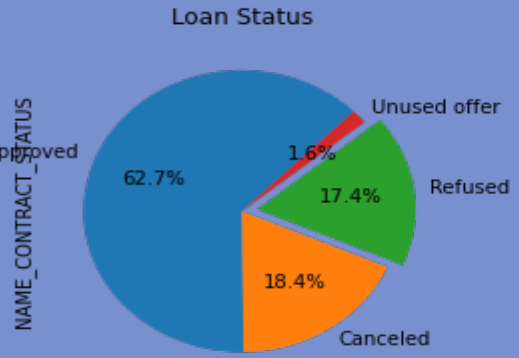




-Among the applicants people who are <30 yrs old are less, but defaulting by them is higher than any other age group

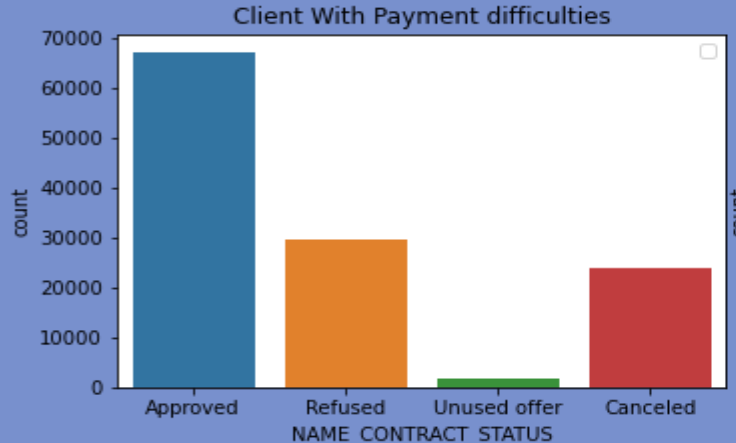
- As applicants age increases the chance of defaulting decreases.

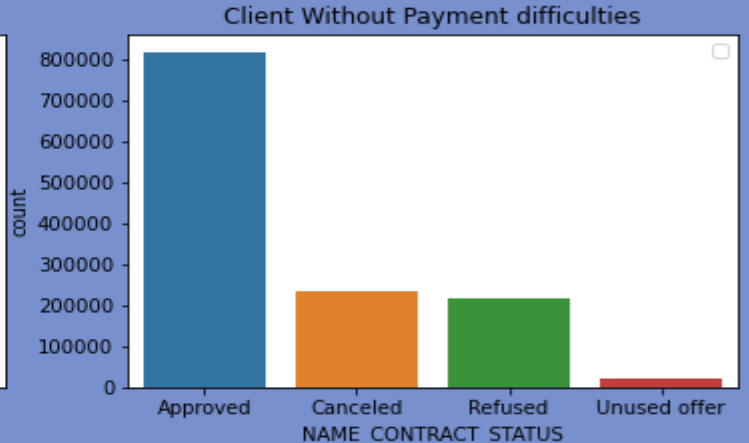
CONTRACT STATUS



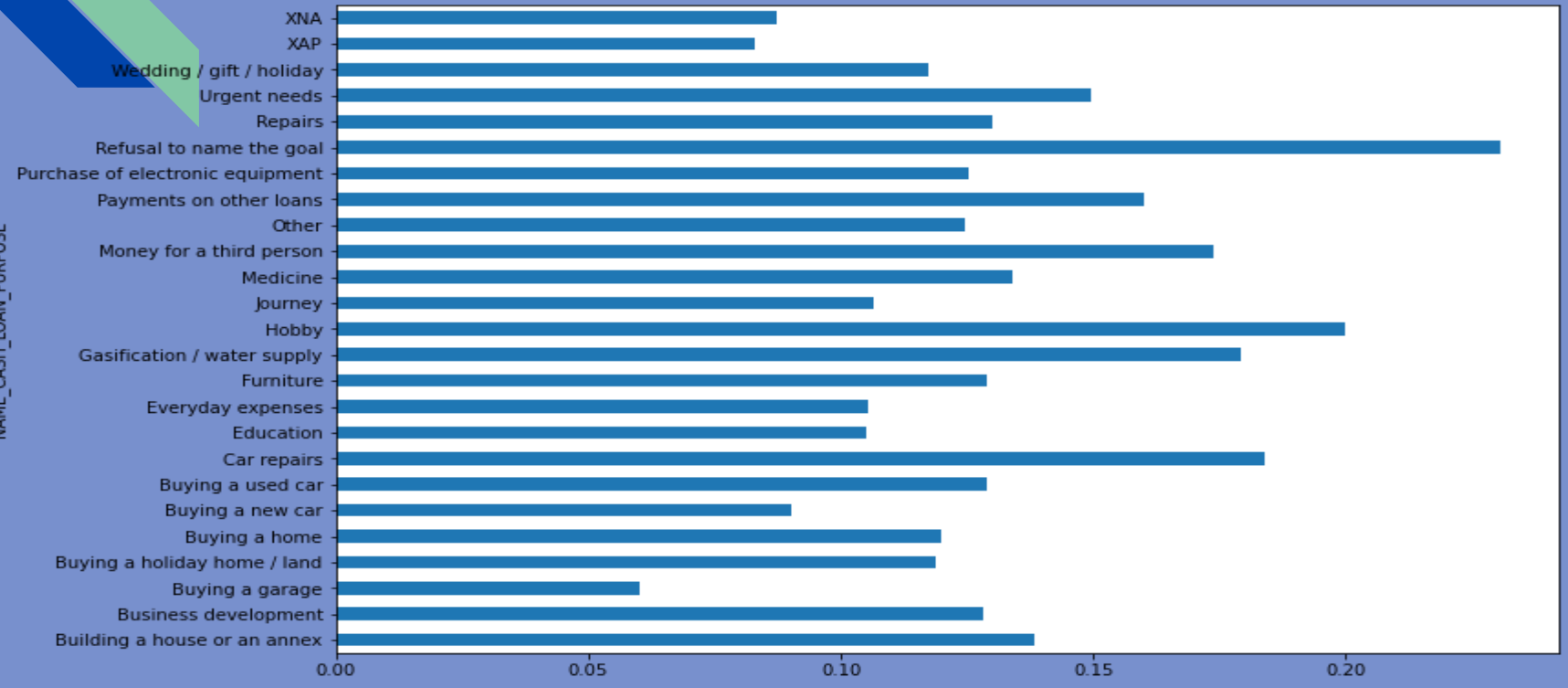
- Applications of clients with payment difficulties are often approved than refused.

- A good fraction of clients without payment difficulties were refused by the institution.





PURPOSE VS TARGET



- People who refuses to name the purpose are highly likely to deafult compared to people from other categories.
- Clients who apply for loan to buy a garage are less likely to default

CONCLUSIONS

The financial institution has to be careful while lending cash to the following categories as the likelihood of defaulting in these groups are higher

1. People who refuse to provide purpose of loan
2. People who are less than 30
3. People with lower secondary education
4. People who are unemployed
5. To Low skill laborers
6. Who lives in rented apartment/with their parents