EDA

Exploratory Data Analysis

Case Study

Consumer Finance Company

By Pinal Gajjar

Problem Statement

To analyse the patterns present in the data and identify a defaulter

- Identify if a person is likely to repay the loan and his loan is not rejected.
- Identify if a person is a defaulter so the loan is not approved.
- This statements are a part of risk analysis as either way if decision are wrong the company's business will face loss.

Dataset Understanding

Given 3 datasets: application_data.csv, previous_application.csv, *columns_description.csv*

Application_data

- Data size: 307511 rows , 122 cols
  - Contains data of loan application with the TARGET variable showing if a person is a defaulter or not.
  - Other variables like annual income, credit amount, loan annuity and personal details of client are present.

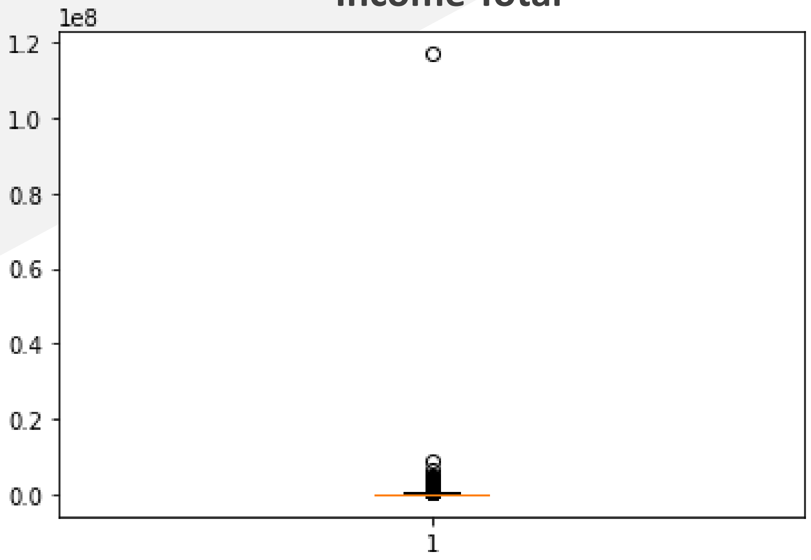
Previous_application

- Data size: 1670214 rows , 37 cols
  - Contains data of previous loan application with the status of loan Approved, Cancelled, Refused or Unused offer.
  - Other variables like annual income, credit amount, loan annuity and loan details of client are present.

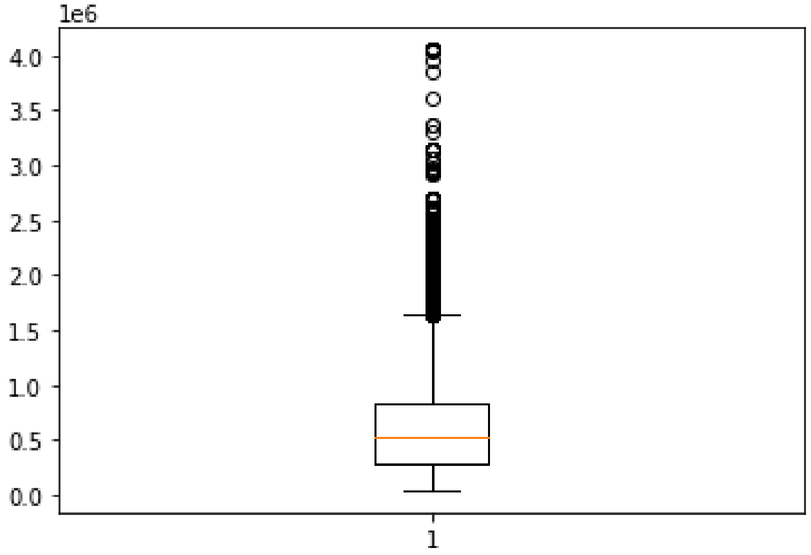
Analytics Approach

- Firstly, data columns with over 50% missing data are removed. We had total 122 cols now 75 are remaining.
- Data with less missing values and variables like annual amount we cannot simply replace with 0 as it will effect the analysis so we will drop such rows as they are in less compared to size of data.
- Finding the outliers of data. We cannot simply delete variables like income total and credit amount as there can be exception in incomes so we will create separate groups to distribute data imbalance.

Income Total



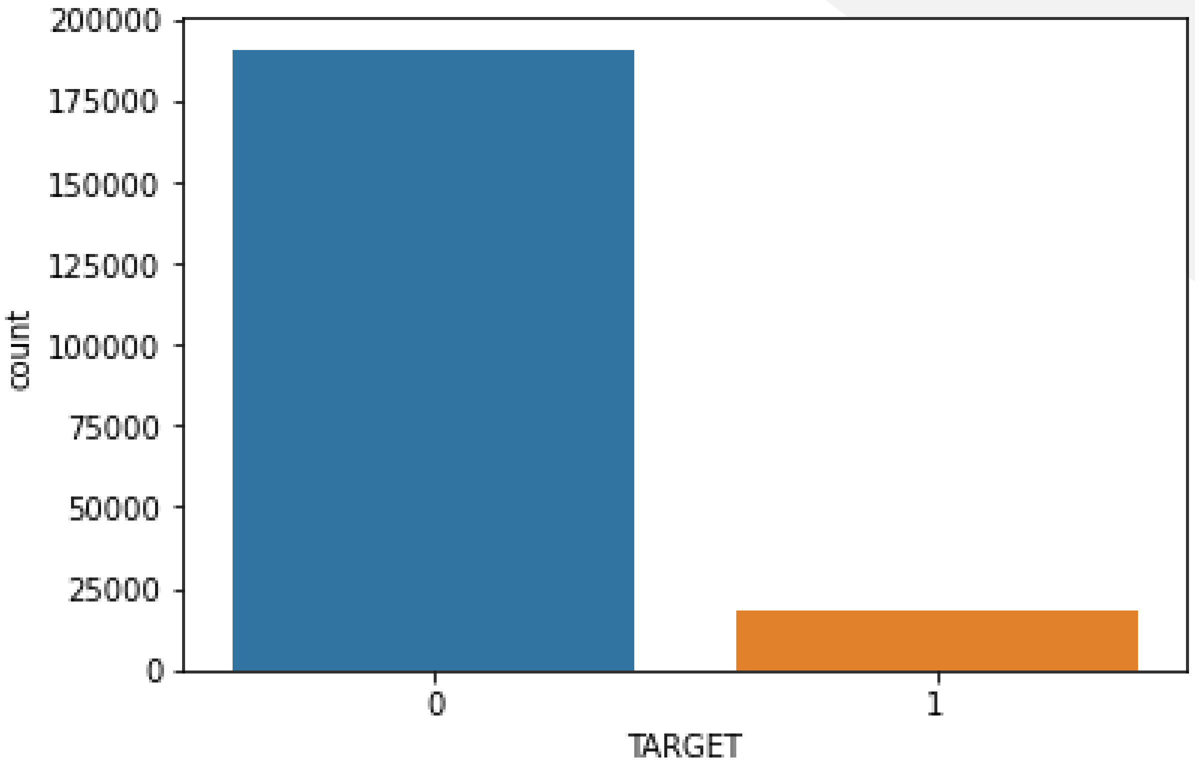
Credit Amount



Target Variable

1- defaulter: client with payment difficulty
0- other cases

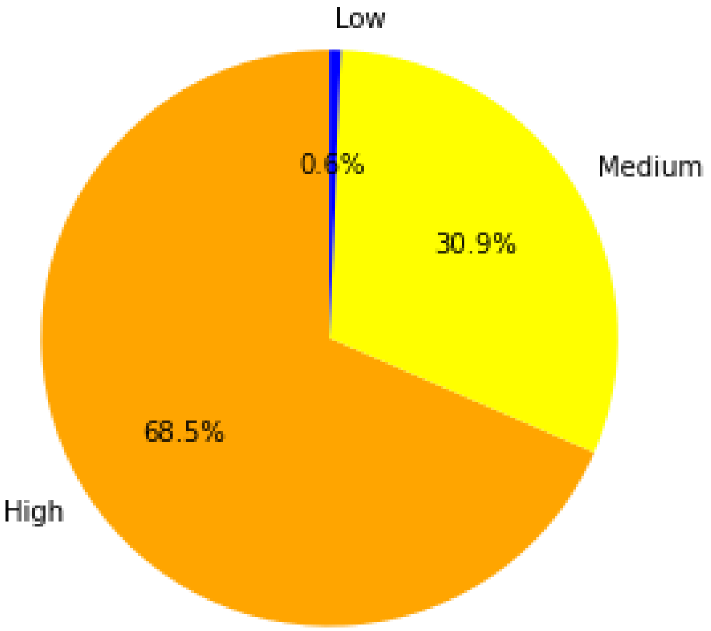
- As we can observe there is a huge difference in both values of TARGET variable.
- Furthermore, looking into data we can also find small differences in Income Category and Credit Category.
- Thus, for better understanding we will observe both values separately by dividing data into two.



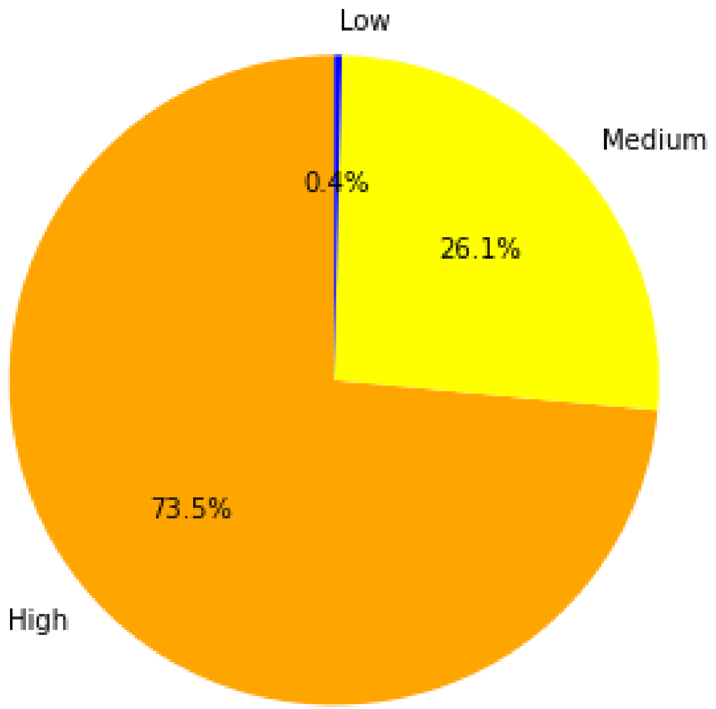
EDA

Income Category

Clients with payment issue(TARGET-1) are more in ‘high’ category in comparison to TARGET-0 concluding client with big businesses may have big risks.



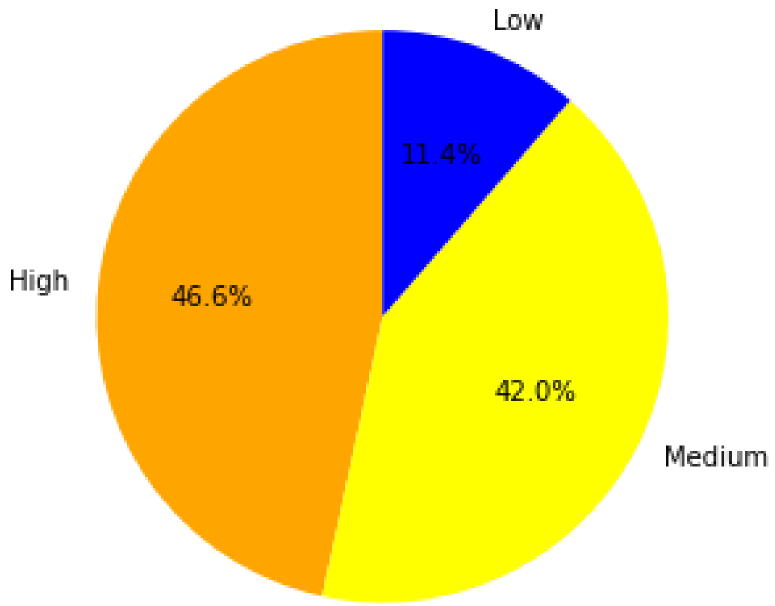
TARGET 0



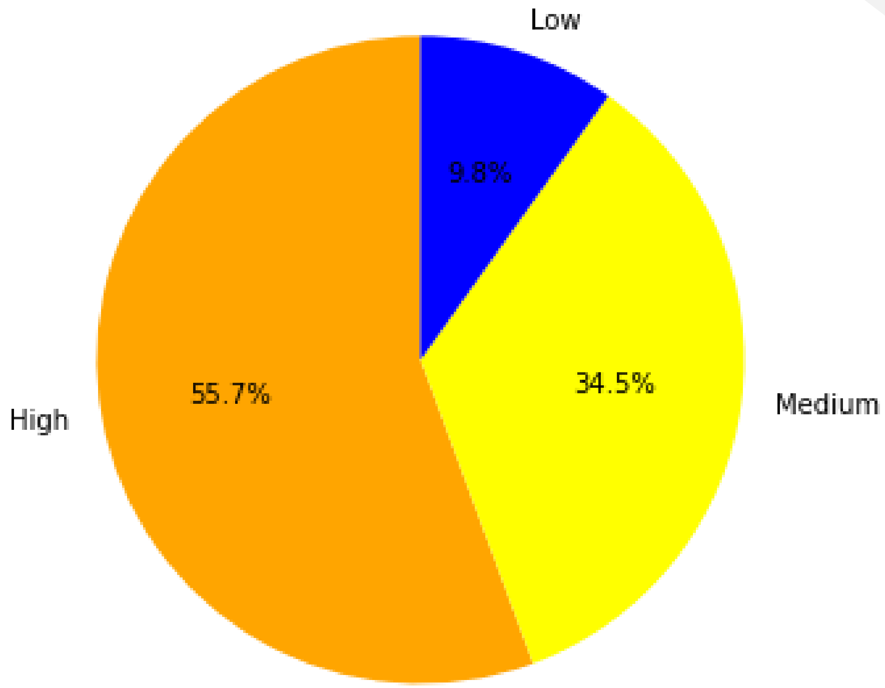
TARGET 1

Credit Category

Almost 9% difference is seen in 'high' loan amounts which can mean clients taking high amount loans can have difficulty in future.



TARGET 0

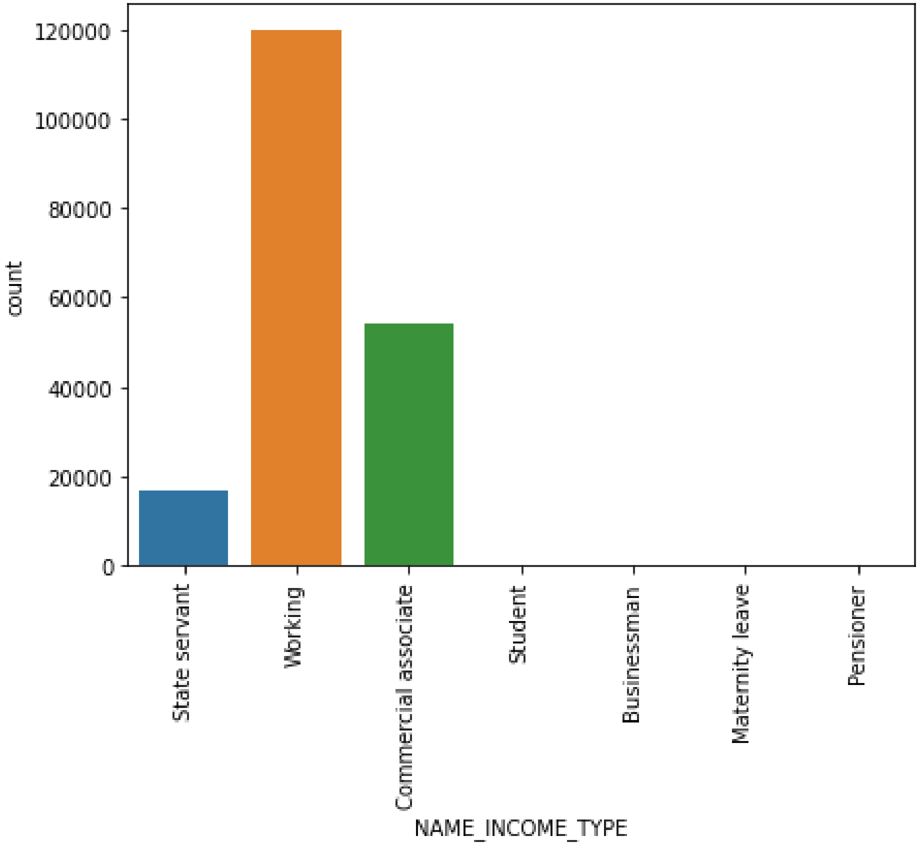


TARGET 1

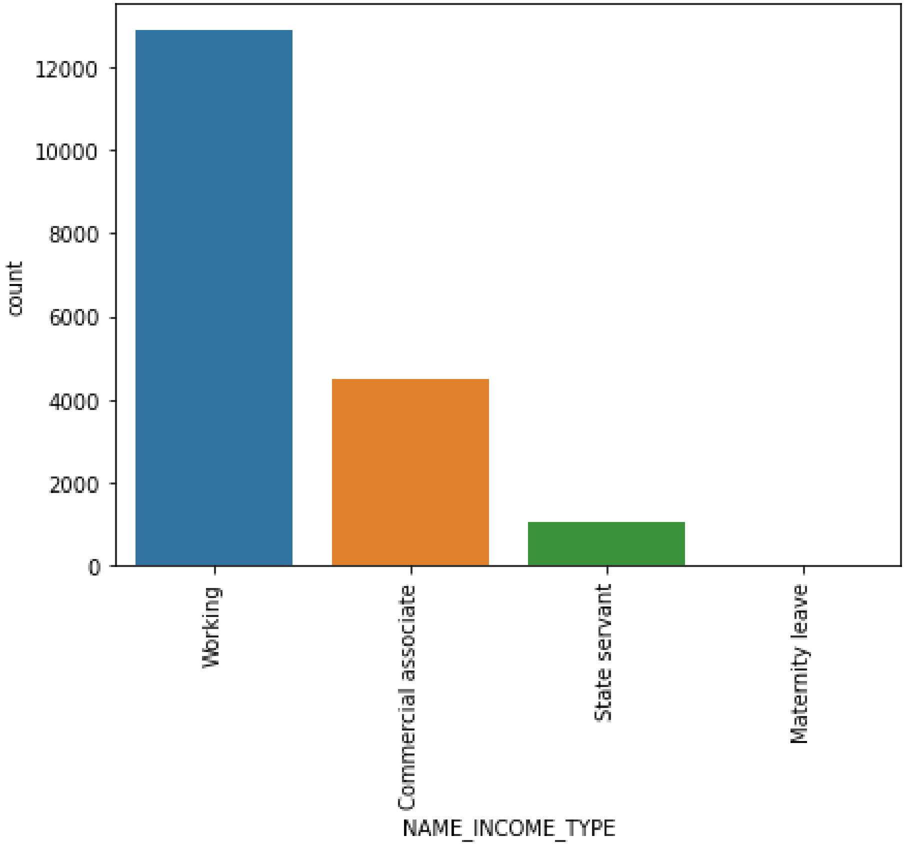
Occupation Category

We can observe students(supported by parents) and businessman usually don't have any problem in payments whereas other category have equal distribution.

TARGET 0



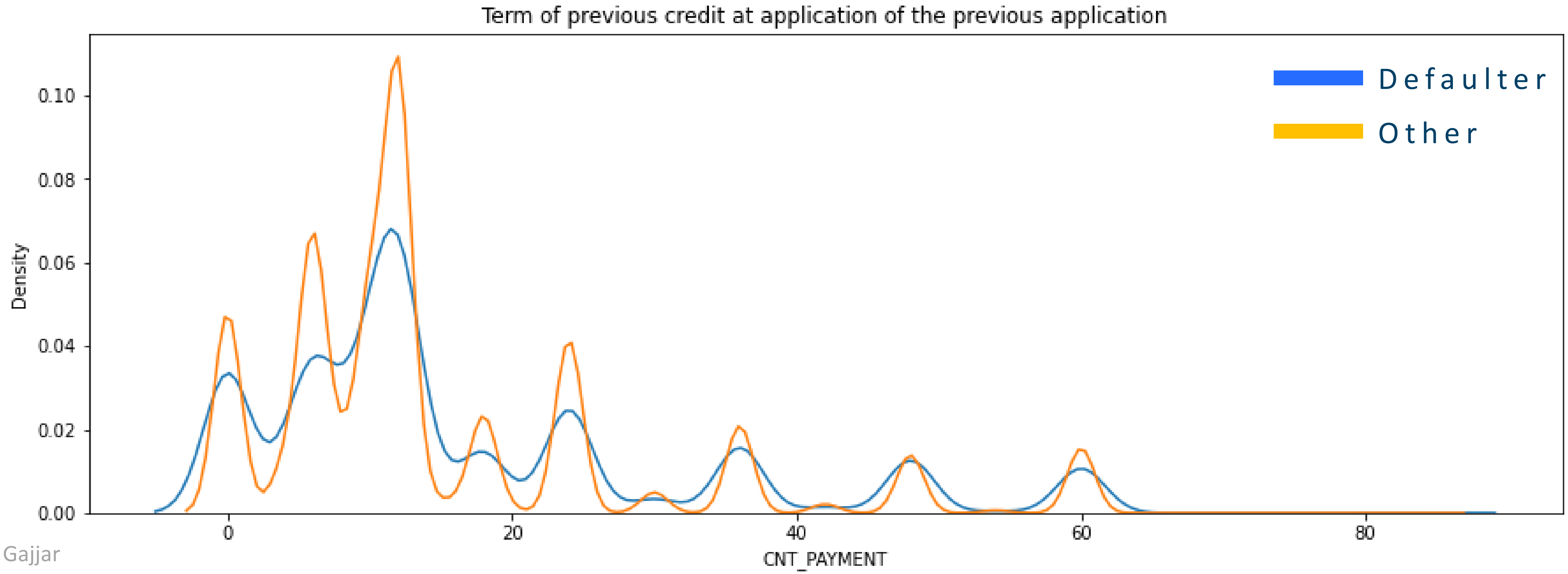
TARGET 1



Merging Datasets

Application_data and Previous_application

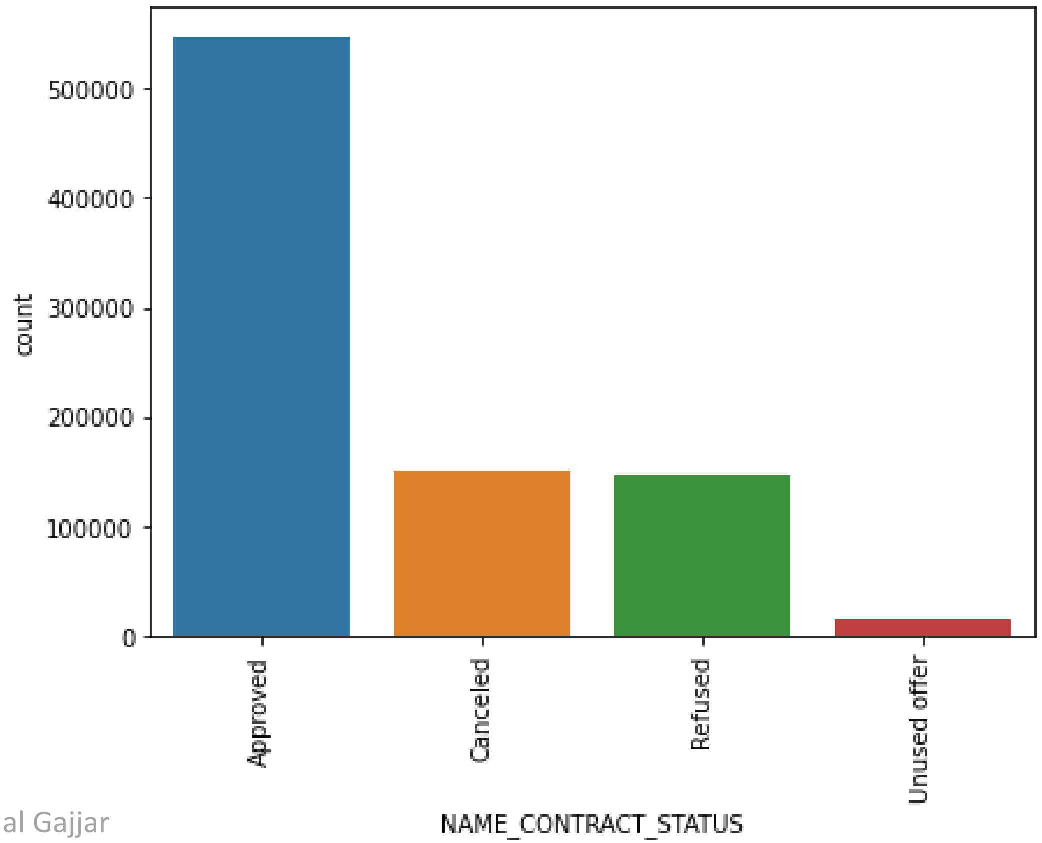
- Previous credit amount of application term is less for defaulters compared to other cases.



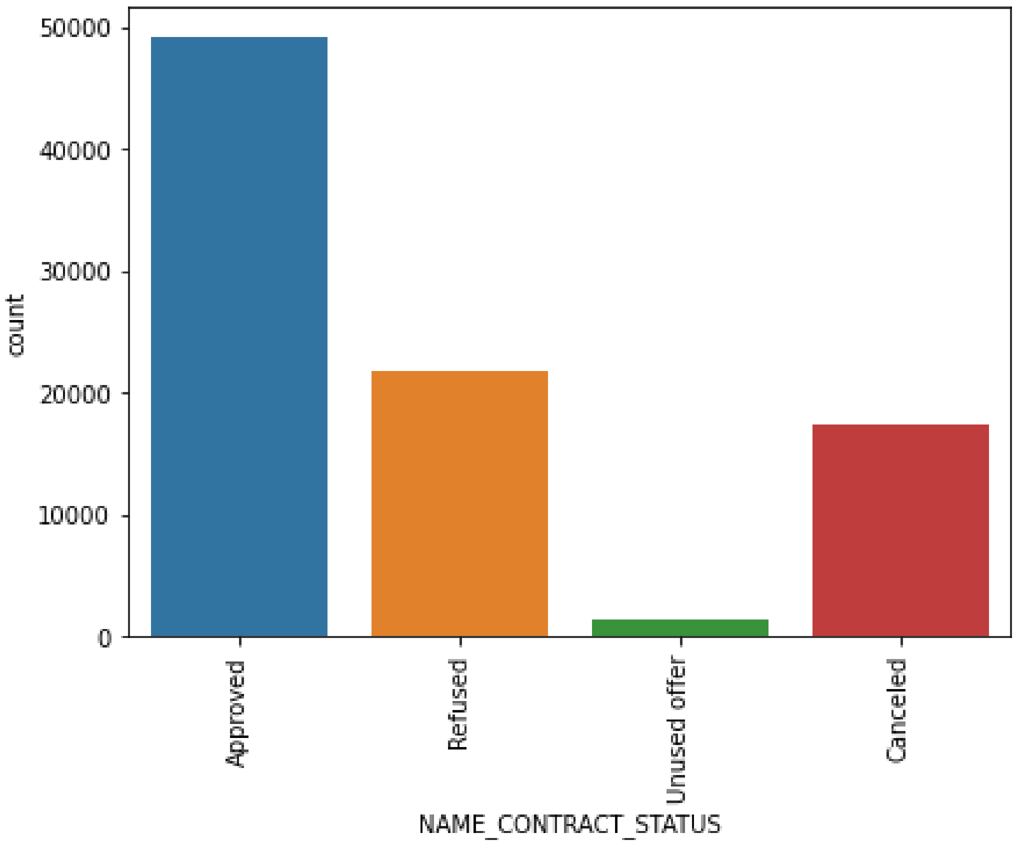
Loan Status

Mostly, previous loans of defaulters are Refused or Cancelled so we can use this variable to identify future applications.

TARGET 0



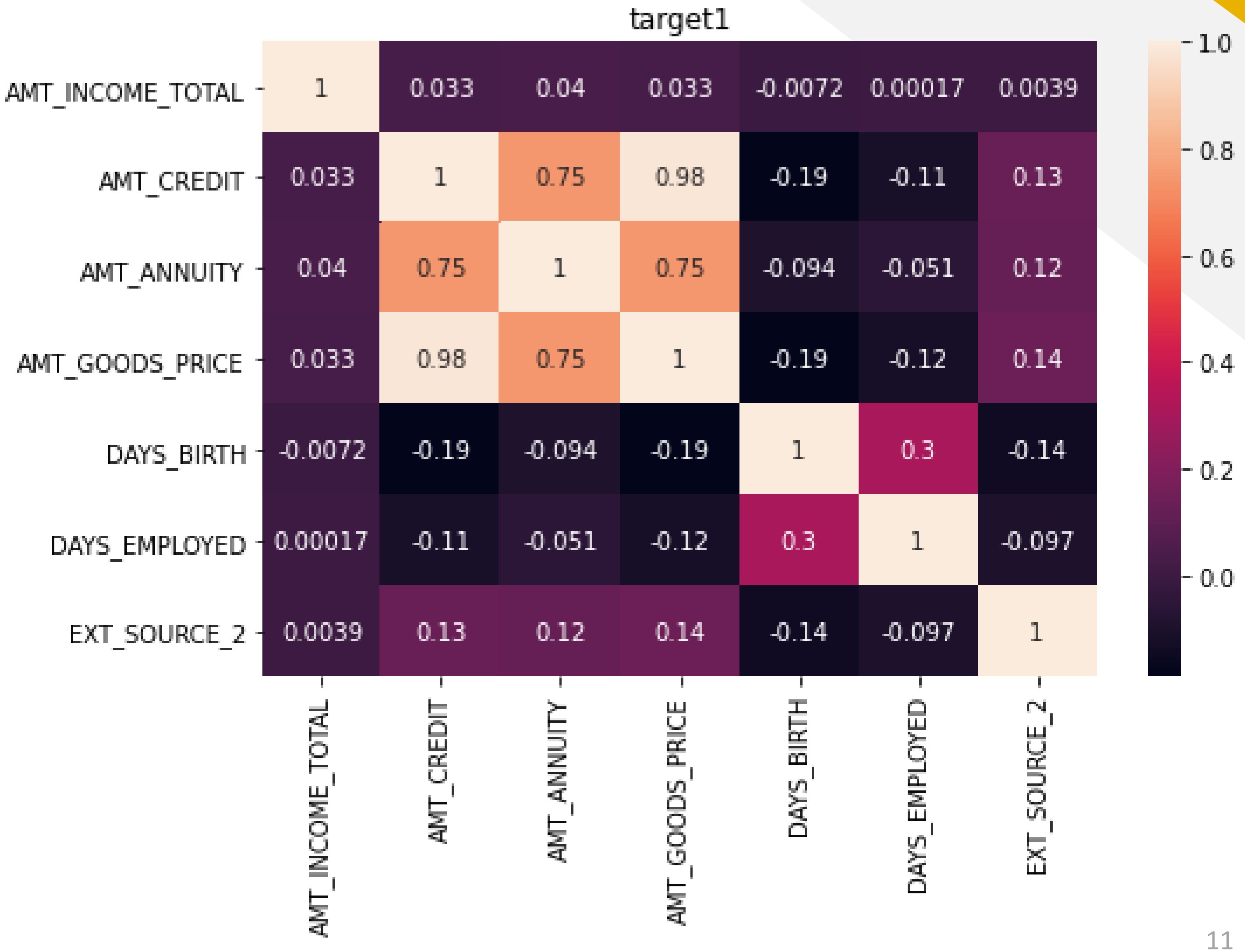
TARGET 1



Correlation

Correlation between variables

- We can observe that there is high correlation between Annuity, Loan credit and Goods price.
- This may lead to high loan amount which can turn out difficult for clients to pay as Income total is not correlated.



target1

	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	DAYS_BIRTH	DAYS_EMPLOYED	EXT_SOURCE_2
AMT_INCOME_TOTAL	1	0.033	0.04	0.033	-0.0072	0.00017	0.0039
AMT_CREDIT	0.033	1	0.75	0.98	-0.19	-0.11	0.13
AMT_ANNUITY	0.04	0.75	1	0.75	-0.094	-0.051	0.12
AMT_GOODS_PRICE	0.033	0.98	0.75	1	-0.19	-0.12	0.14
DAYS_BIRTH	-0.0072	-0.19	-0.094	-0.19	1	0.3	-0.14
DAYS_EMPLOYED	0.00017	-0.11	-0.051	-0.12	0.3	1	-0.097
EXT_SOURCE_2	0.0039	0.13	0.12	0.14	-0.14	-0.097	1

Conclusion

- We can conclude that in most scenarios high amount of loans can lead to defaulters where the income total is the main cause. Therefore, high incomes should be checked before giving large amount of loans.
- For some cases the previous data can also help if a clients loan is Rejected before then should not consider giving loans with high amount.
- It is recommended to check previous loan amount as client try to go for high amount of loan once they have completed previous loans.

EDA

Exploratory Data Analysis

Thank You.

Pinal Gajjar