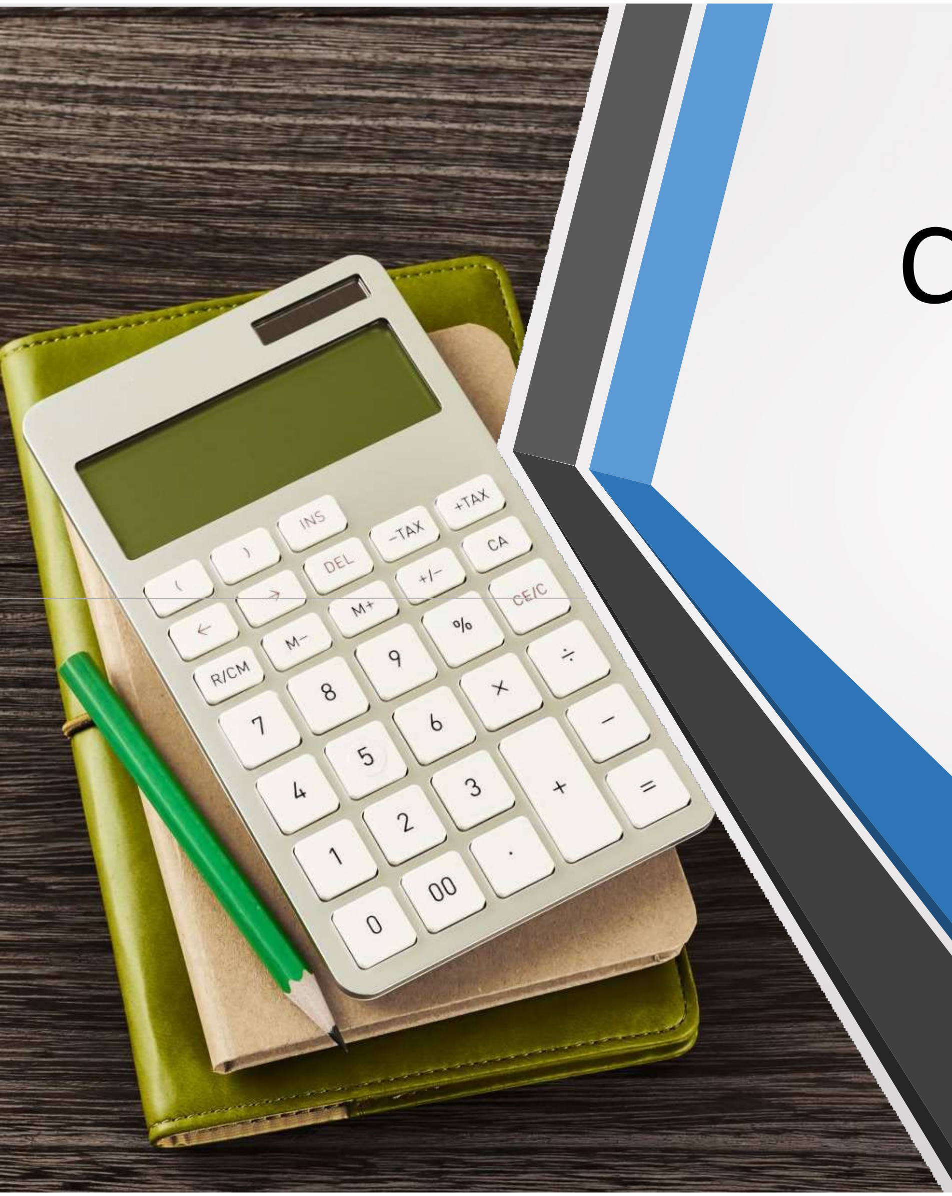


CREDIT EDA CASE STUDY

Hetal Khanapure – hetu.parmar@gmail.com

EXECUTIVE SUMMARY: Key Findings on Giving Loans

EDA for the banking dataset revealed that :

- The proportion of defaulters is 8.11%
- The bank lends more to females, proportion of female defaulter percentage is lower – could actively look for more male customers who satisfy other criteria.
- More Cash Loans go into default : bank should concentrate more on Revolving Loans
- Proportion of Working Clients in defaults is higher and State Servant is lower, should focus on State Servant for fresh loans.
- Clients living with their parents or in rented apartment have higher rate of default.
- Higher Education and Old Age default percentage less, bank should target them
- Married people are safer as Singles default percentage is very high
- The clients whose previous loans were approved are more likely to pay current loan in time, than the clients whose previous loans were rejected.

PROBLEM STATEMENT

This case study aims to identify patterns which indicate :

If an applicant has difficulty in paying his/her installments which may be used for taking actions such as :

- Denying the loan
- Reducing the amount of loan
- Lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected and the number of defaulters is also reduced.

Note :

- Financial Institution / Bank : wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default.
- The bank can utilize this knowledge for its portfolio and risk assessment.

Analysis of this dataset has been done in Python on a Jupyter Notebook.

TYPES OF ANALYSIS DONE

Steps :

- Check Missing Values
- Check Outliers
- Top 10 correlation for the Client with payment difficulties

Divided into following tasks :

Task 1 – Reading the dataset and finding information about the data

Task 2 – Inspecting the Data for Data Cleaning : Null Values, Which Columns to drop and which to impute

Task 3 – Imputing Values : Categorical – Mode ; Numeric – Depending on type of distribution

Task 4 – Checking Datatypes of Columns

Task 5 – Checking for Outliers and Handling

Task 6 – More Data Cleaning and Operations

Task 7 – Classifying Data into Bins

Task 8 – Finding Co-relations into data

Task 9 – Univariate, Bivariate and Multivariate Analysis

Task 10 – Checking History of clients with Previous Application

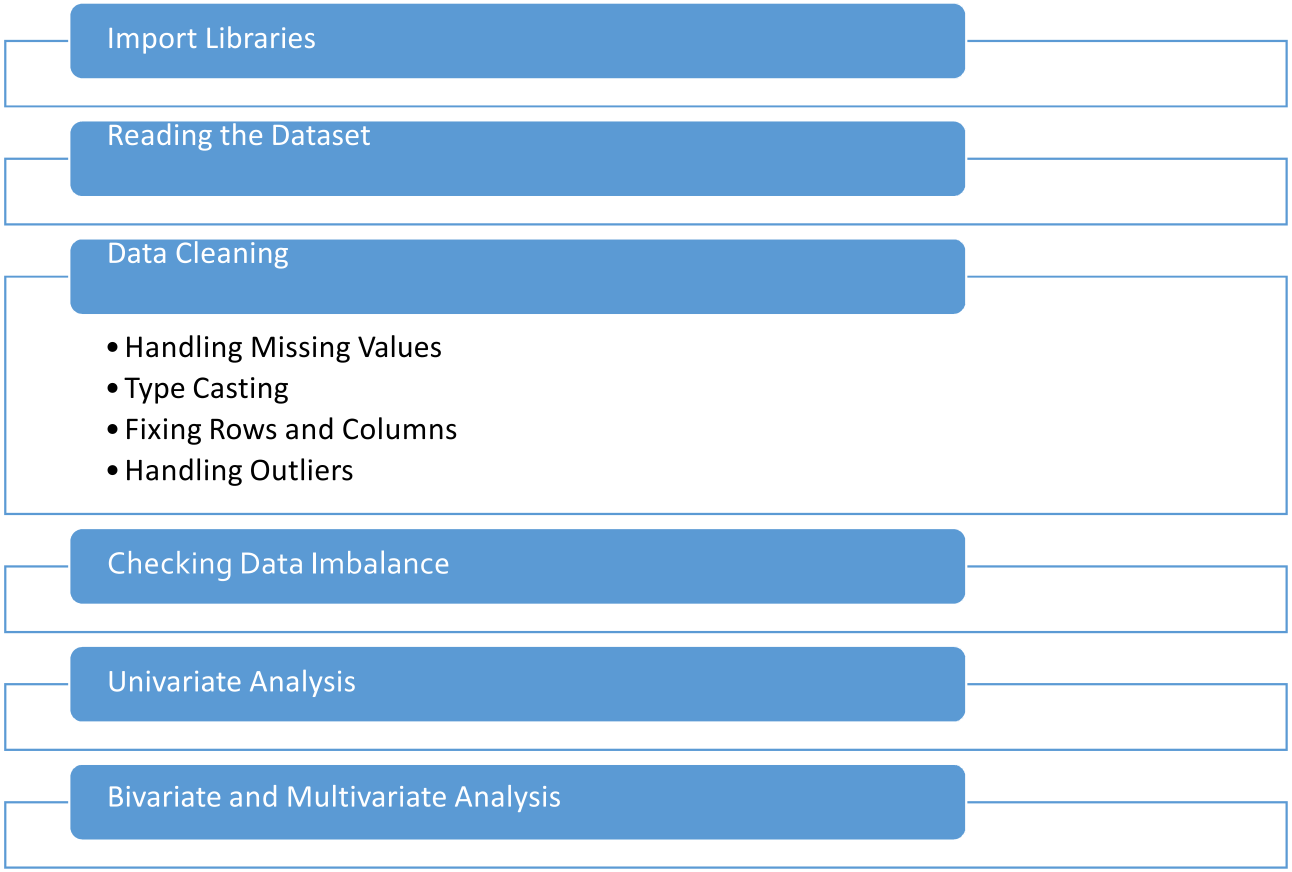
INFORMATION PROVIDED

Application Dataset: contains all the information of the client at the time of application. The data is about whether a client has payment difficulties.

Previous application Dataset : contains information about the applicant's previous loan data. It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer.

Columns description Dataset: It is a data dictionary which elaborates the meaning of the variables.

APPROACH FOR EDA



Import Libraries

Reading the Dataset

Data Cleaning

- Handling Missing Values
- Type Casting
- Fixing Rows and Columns
- Handling Outliers

Checking Data Imbalance

Univariate Analysis

Bivariate and Multivariate Analysis

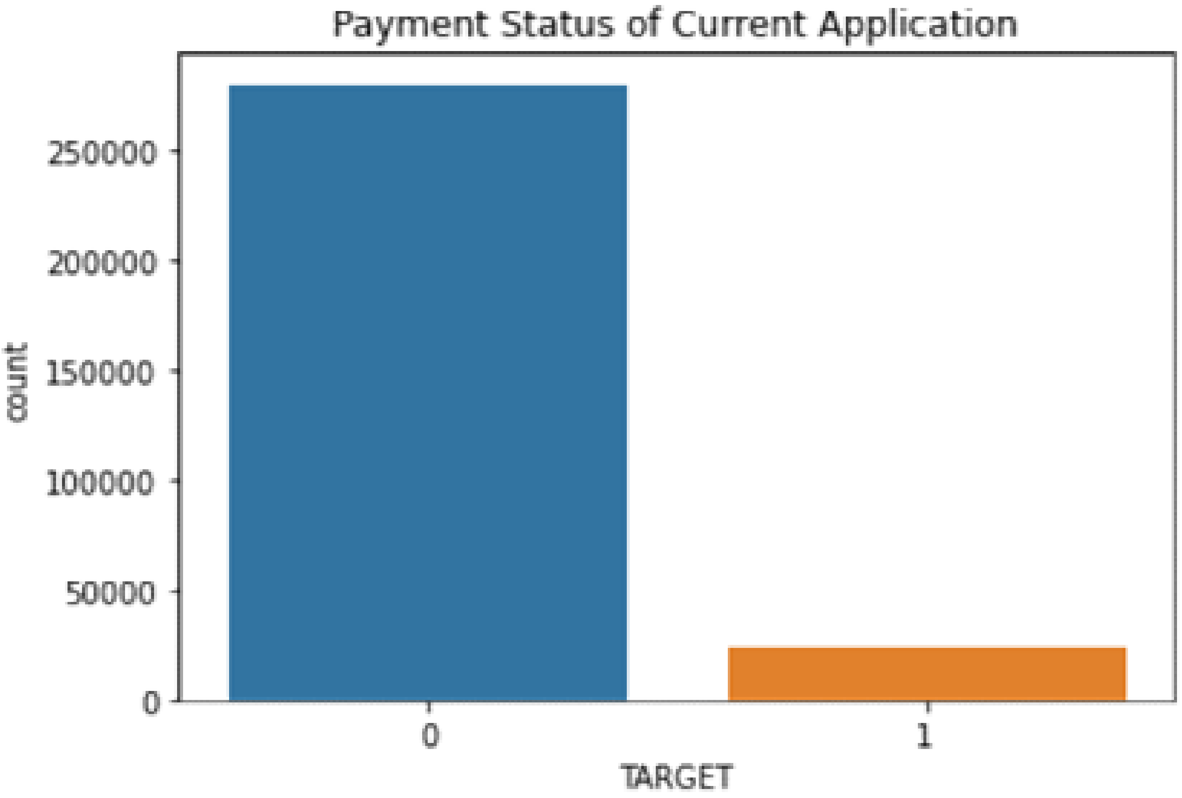
DATA IMBALANCE

Data is highly imbalanced as number of defaulter is very less in total population.

Data Imbalance Ratio is

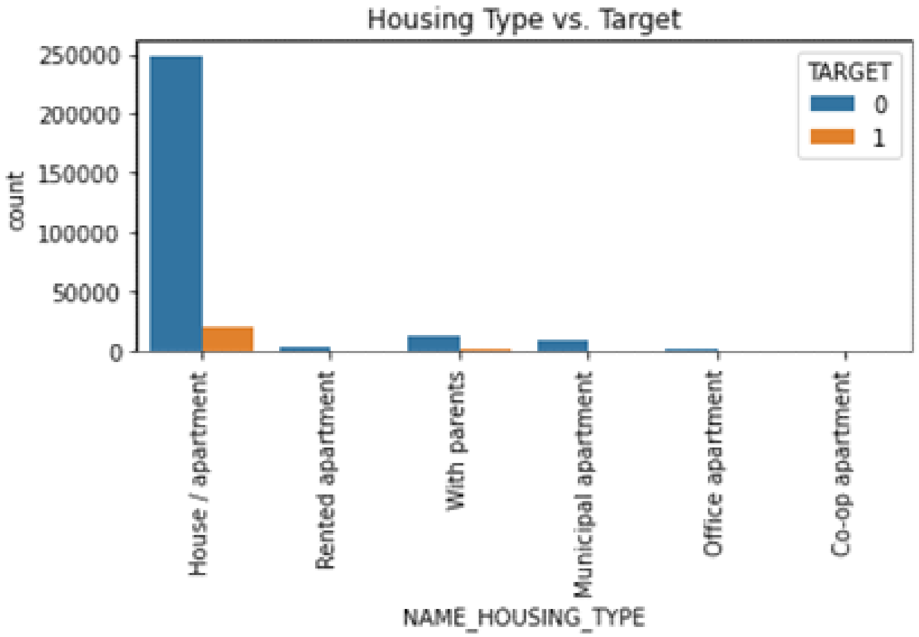
Defaulter : Non-Defaulter 8 : 92 = 2 : 23

Almost 92% loan amount returned to bank.



EDA - APPLICATION DATASET

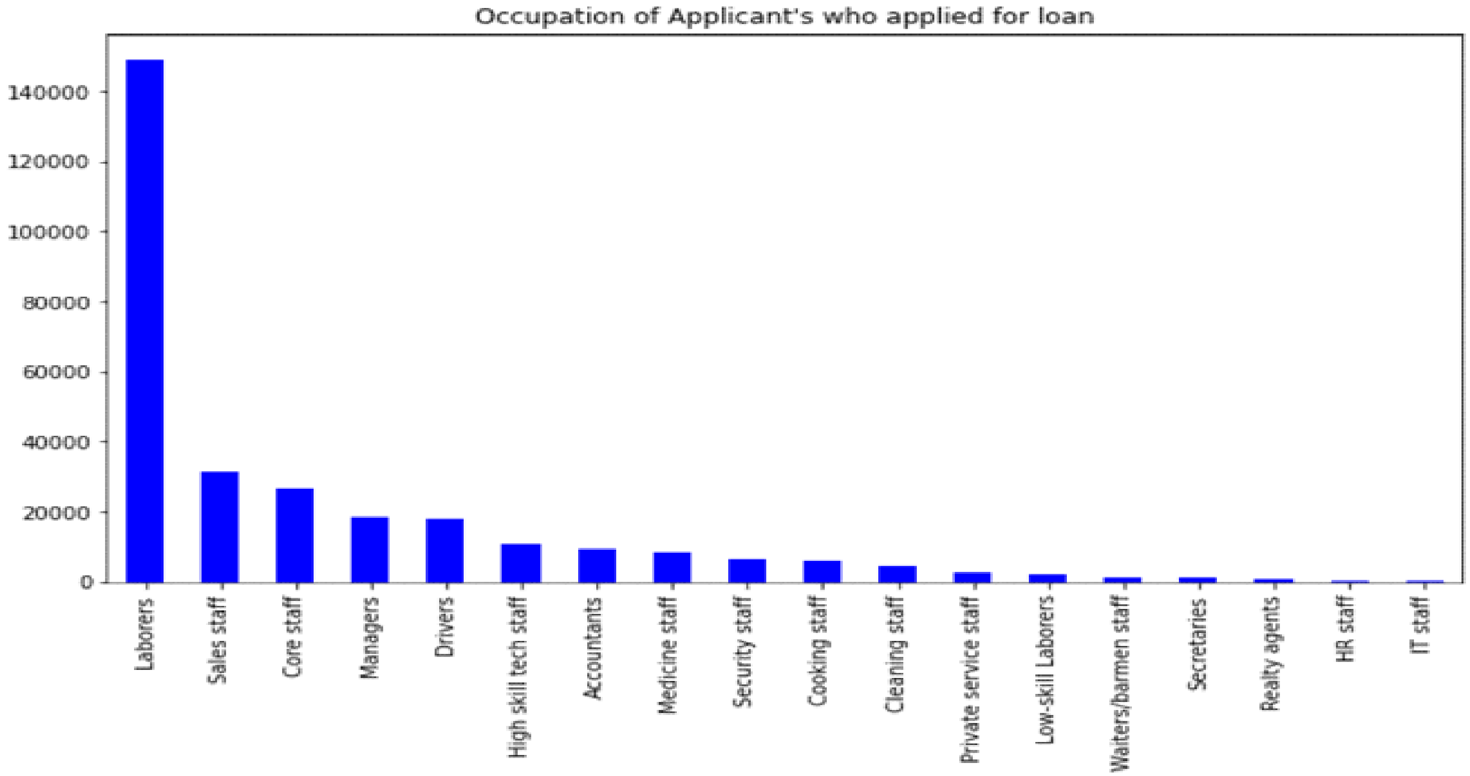
FINDING INSIGHTS IN DATA



Value	Percentage of Defaulter
Rented apartment	12.373477
With parents	11.710062
Municipal apartment	8.583536
Co-op apartment	7.996406
House / apartment	7.839354
Office apartment	6.602317

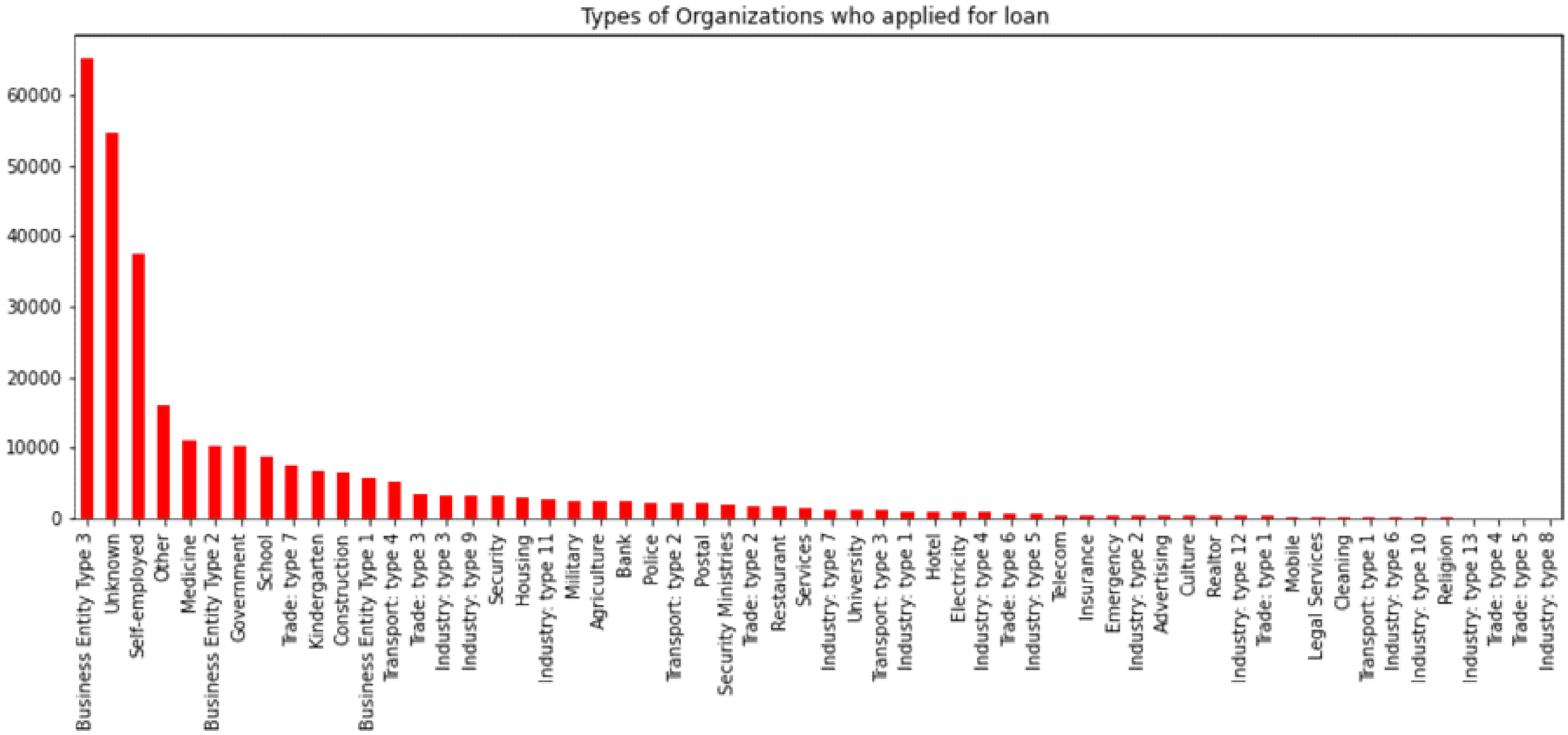
Most of the clients live in House/Apartment
Clients living with their parents or in rented apartment have higher rate of default.

FINDING INSIGHTS IN DATA



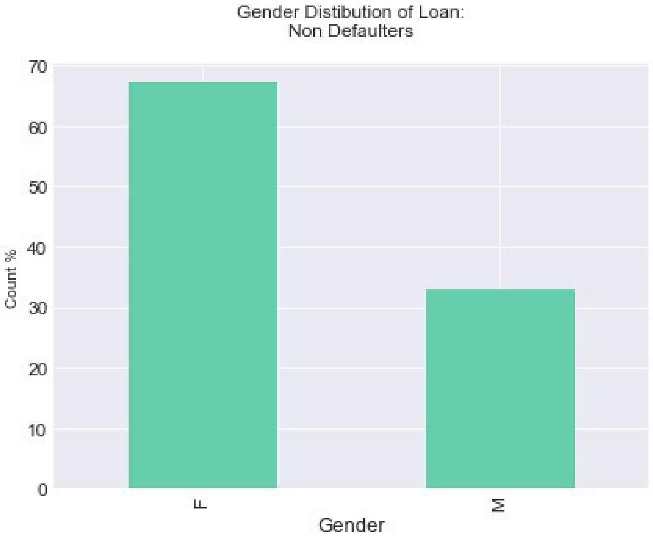
We can observer that Laborers, Sales Staff and Core Staff constitute the majority whereas IT Staff is on the lower side.

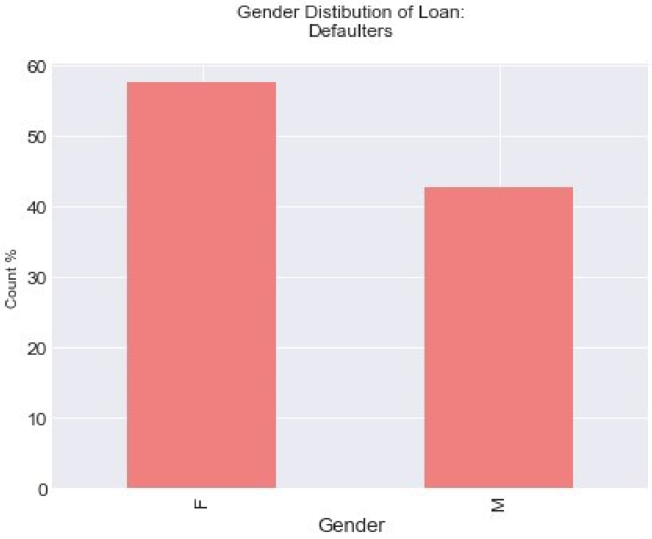
FINDING INSIGHTS IN DATA



We can observe that Business Entity Type 3 organizations have majority of the loan applications.

Univariate Analysis of Categorical Variables

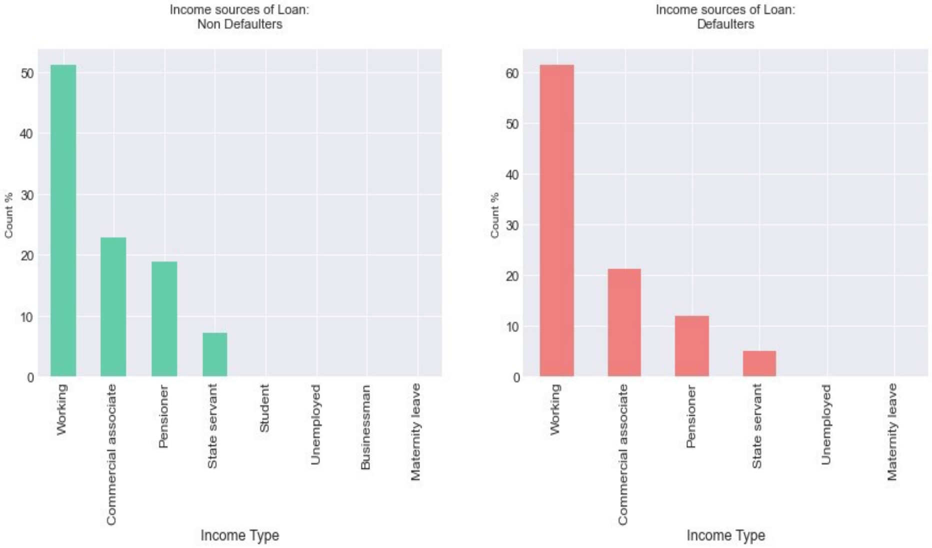




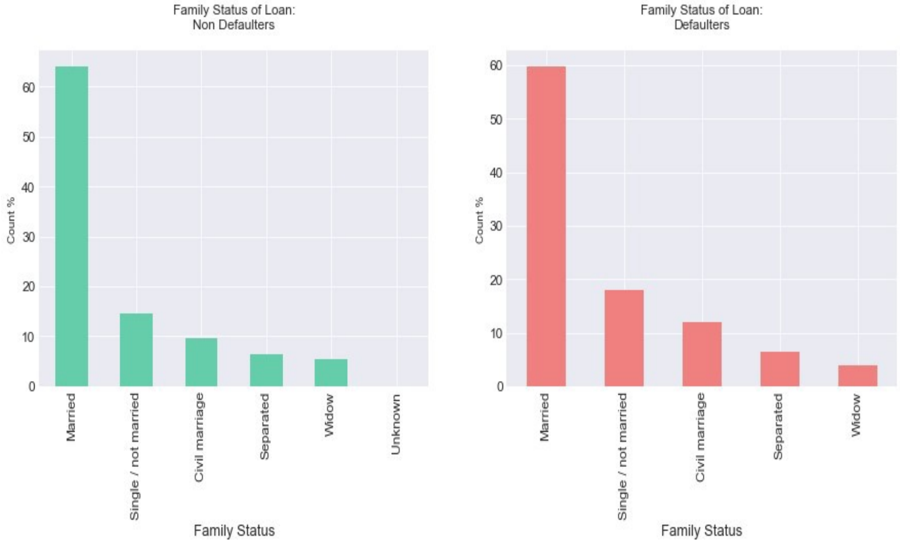
Value	Percentage of Defaulter
M	10.303457
F	7.058989

- Comparing the Payment Difficulties and Non Payment Difficulties on the basis of Gender, we observe that Females are the majority in both the cases although there is an increase in the percentage in Male Payment Difficulties from Non-Payment Difficulties.
- Overall Default % is higher for Male Clients

Univariate Analysis of Categorical Variables

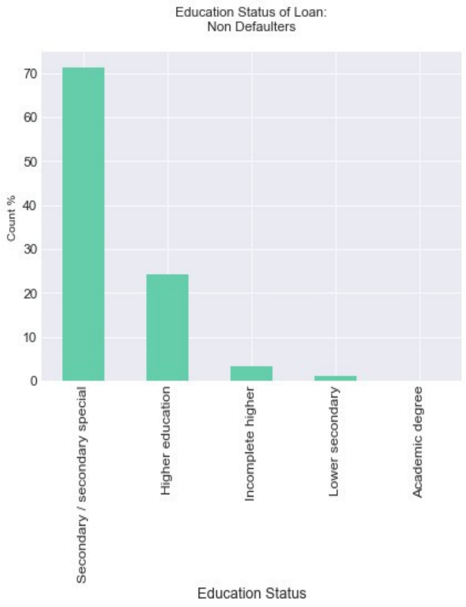


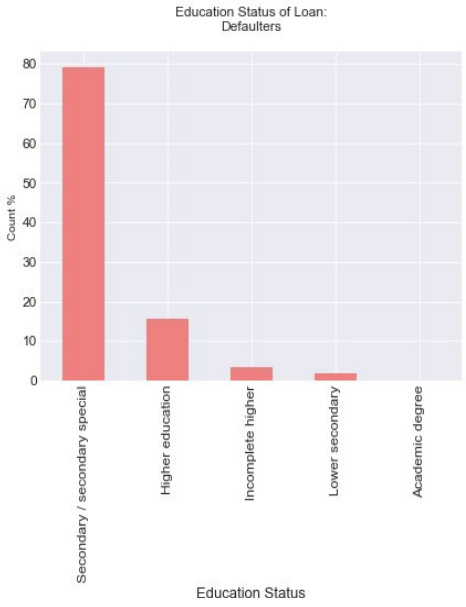
We observe a decrease in the percentage of Payment Difficulties who are pensioners and an increase in the percentage of Payment Difficulties who are working when compared

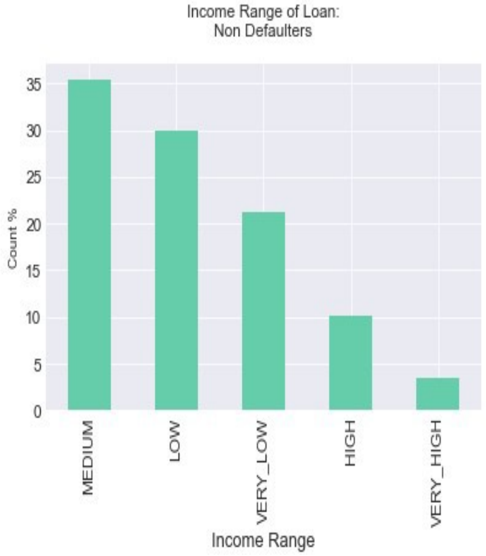


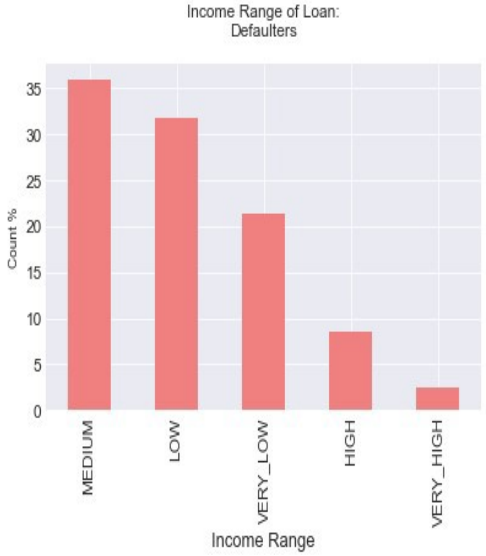
We observe a decrease in the percentage of married and widowed with Loan Payment Difficulties and an increase in the the percentage of single and civil married with Loan Payment Difficulties when compared with the percentages of both Loan Payment Difficulties and Loan Non-Payment Difficulties

Univariate Analysis of Categorical Variables





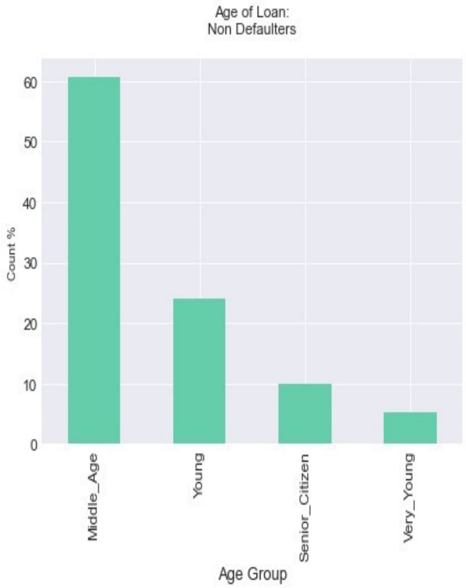


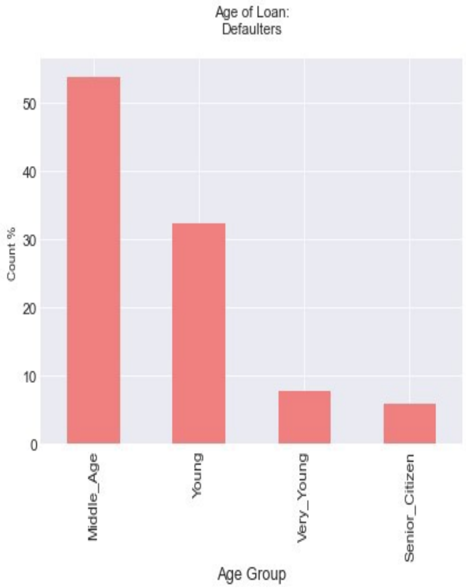


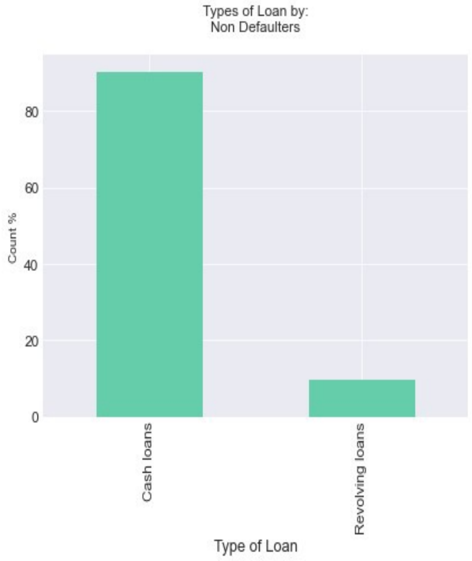
We observe an increase in percentage of Loan Payment Difficulties whose educational qualifications are secondary/secondary special and a decrease in the percentage of Loan Payment Difficulties who have completed higher education when compared with the percentages of Loan Payment Difficulties and Loan Non-Payment Difficulties

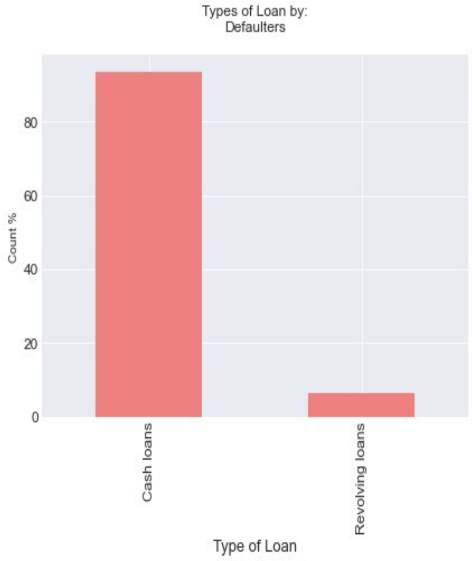
We observe an increase in the percentage of Loan Payment Difficulties whose income is low when compared with the percentages of Payment Difficulties and Loan-Non Payment Difficulties

Univariate Analysis of Categorical Variables





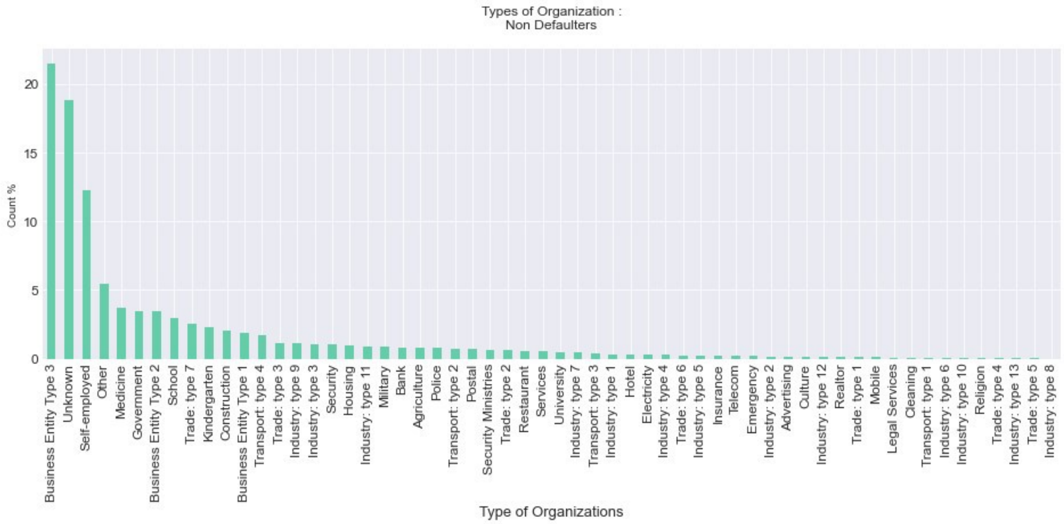




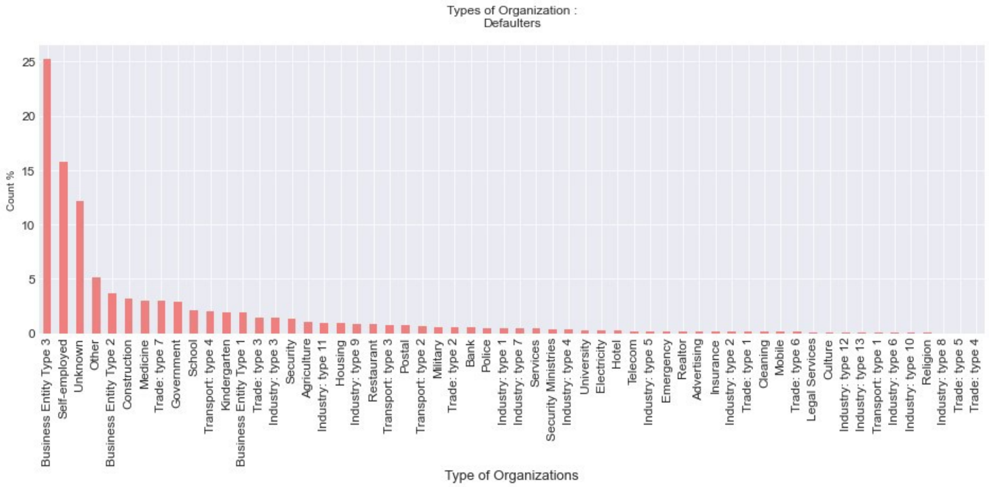
We observe that there is an increase in the percentage of Loan Payment Difficulties who are young in age when compared to the percentages of Payment Difficulties and Loan-Non Payment Difficulties.

We can observe that cash loans are preferred by both Loan Payment Difficulties and Loan-Non Payment Difficulties although there is a decrease in the percentage of Payment Difficulties who opt for revolving loans.

Univariate Analysis of Categorical Variables


Types of Organization : Non Defaulters

We observe that there is an increase in the percentage of Loan Payment Difficulties who are 'Self-Employed' in organization when compared to the percentages of Payment Difficulties and Loan-Non Payment Difficulties.

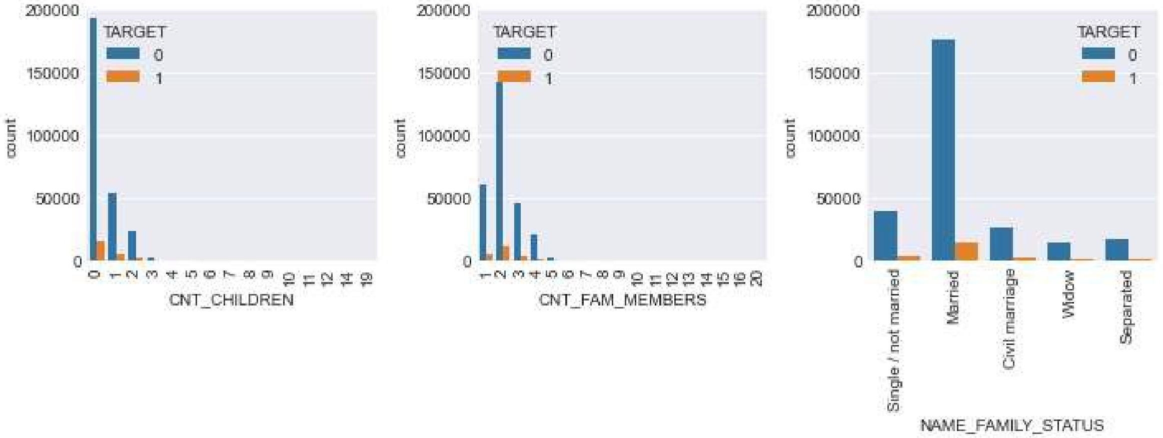

Types of Organization : Defaulters

Univariate Analysis of Categorical Variables

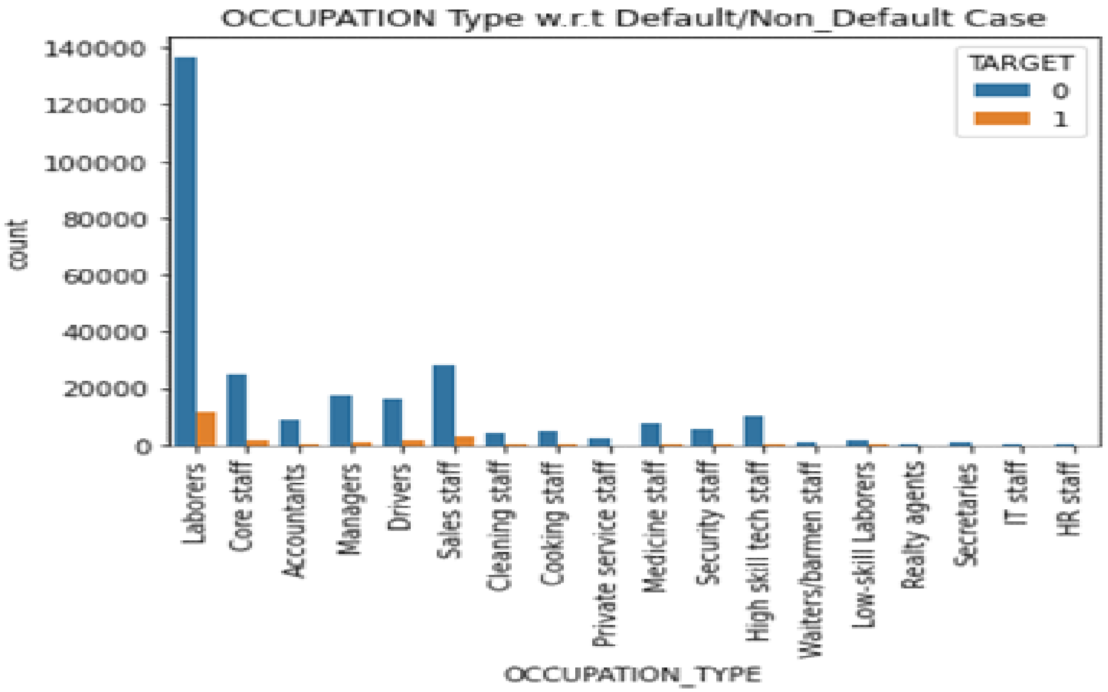
- Default rate is highest for Civil Marriage and Single clients.
- Maximum clients are married with family members as 2 and 0 children.
- As the number of children increases default percentage increases.
- As the number of family members increases default percentage increases.

Note :

- In some cases it has been observed that where the count of children/family member is high, default % is either very high or very low. This shows the imbalance in the dataset as the clients with very high count in family members there total count in dataset is low thereby cannot be taken as conclusion.



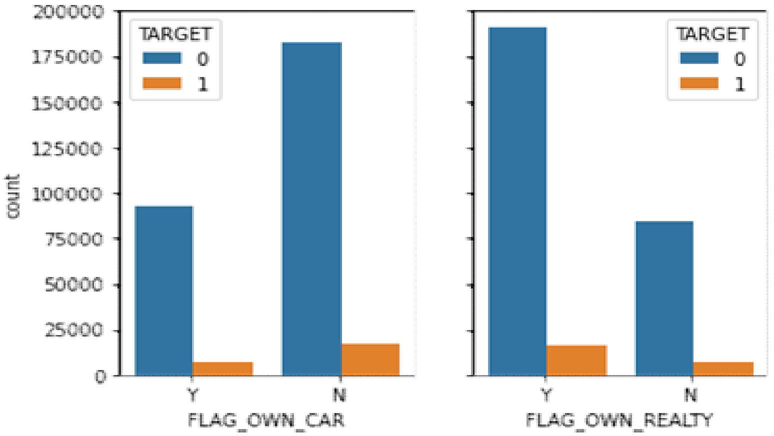
Univariate Analysis of Categorical Variables



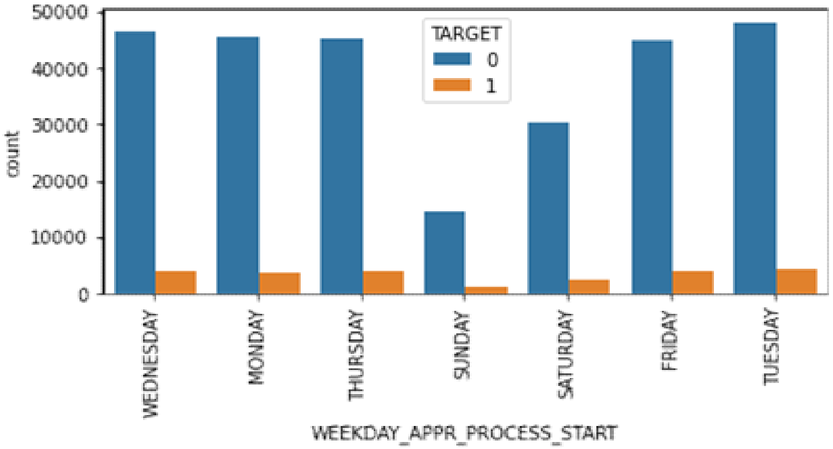
- Maximum clients occupation is Laborer as its count is high.
- Default % is high for Low-skill Laborers occupation of client as compared to other occupations.

Univariate Analysis of Categorical Variables

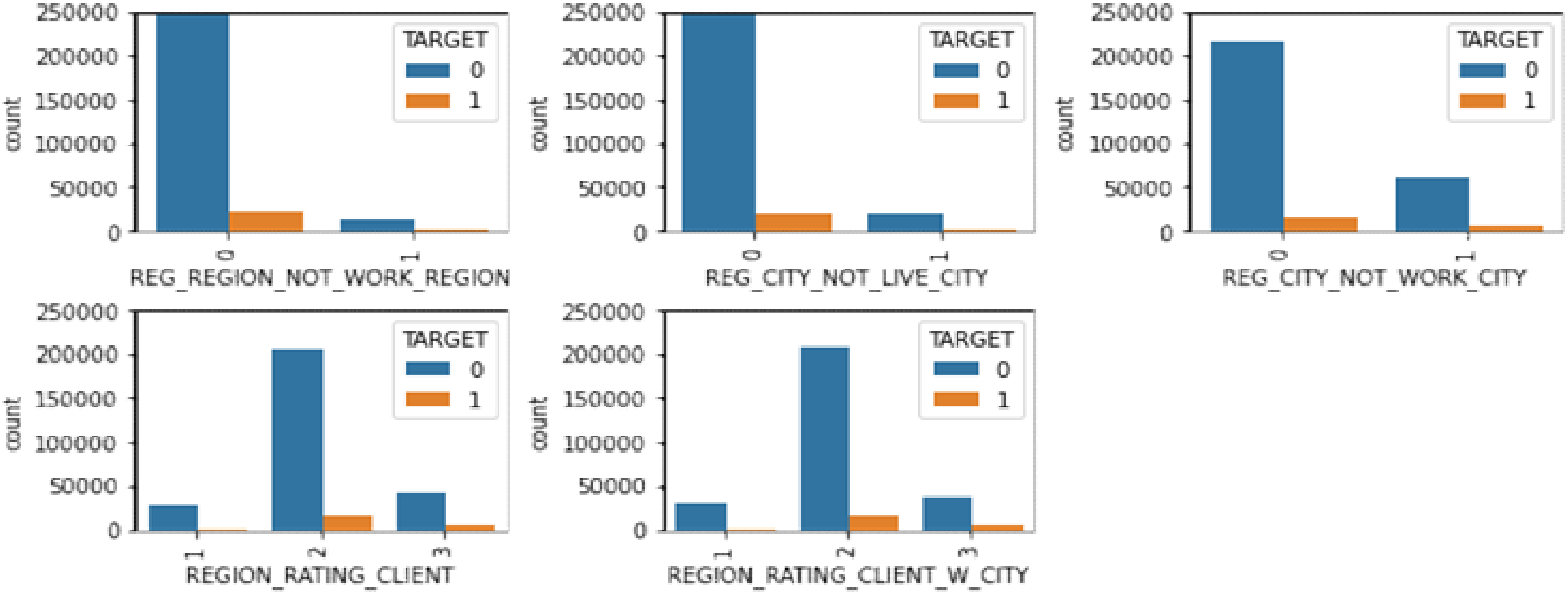
- Maximum Clients Own Realty.
- Maximum Clients do not Own Car
- Default case is approx. 8% in case of clients not owning realty or car



- All weekdays have similar number of applicants than weekend(Saturday and Sunday)

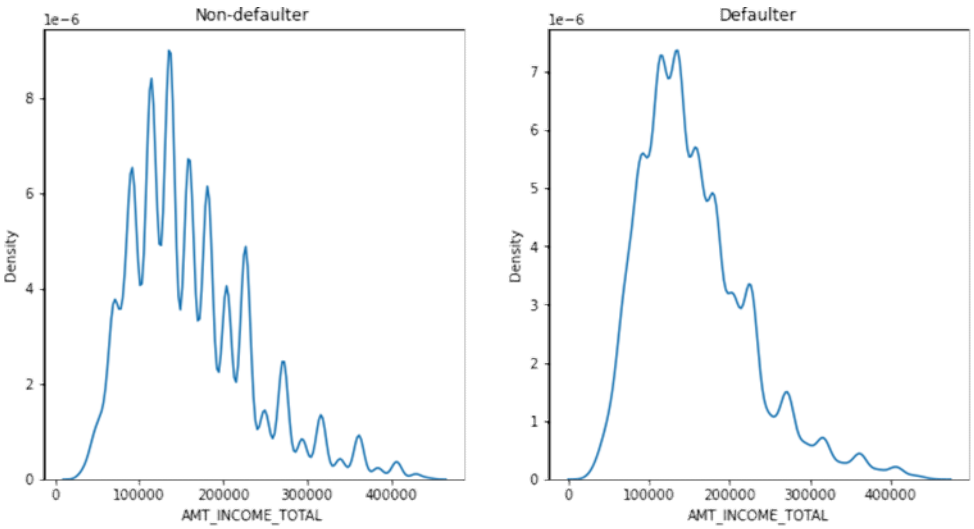


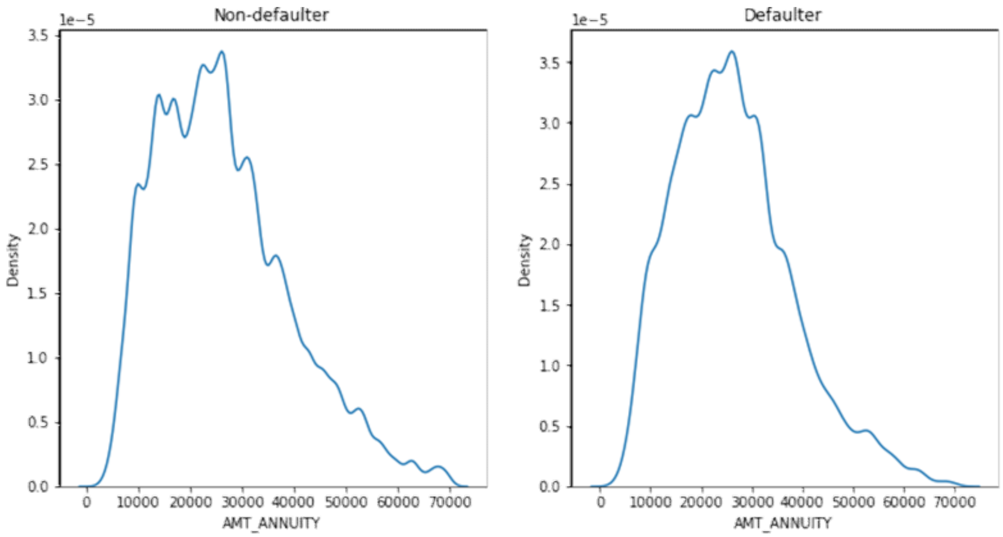
Univariate Analysis of Categorical Variables



- Defaulter rate is highest when REG_REGION_NOT_WORK_REGION=0 i.e. permanent address and working address is same.
- Highest Applicants have Region rating of 2

Univariate Analysis of Numerical Variables





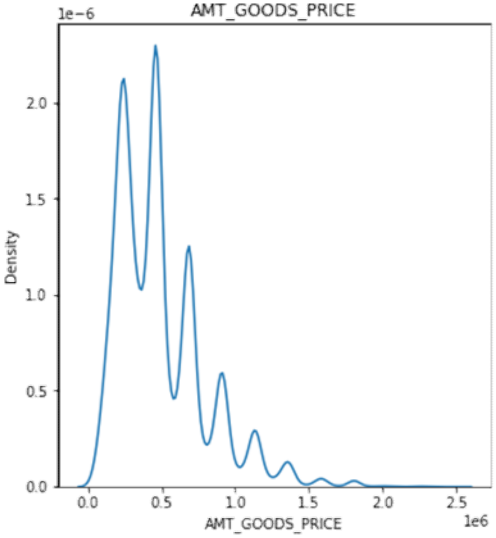
AMT_INCOME_TOTAL

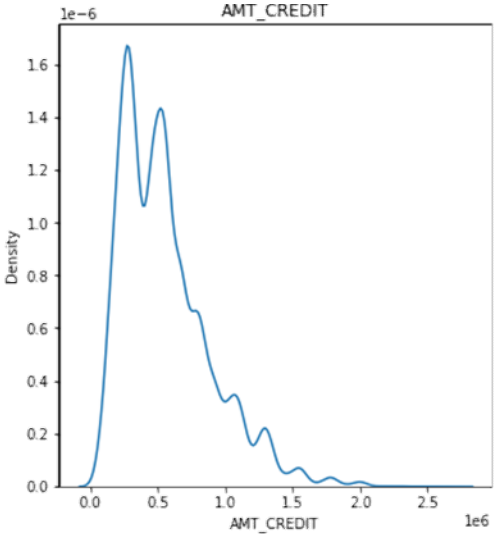
- As the client's income increases, default rate decreases

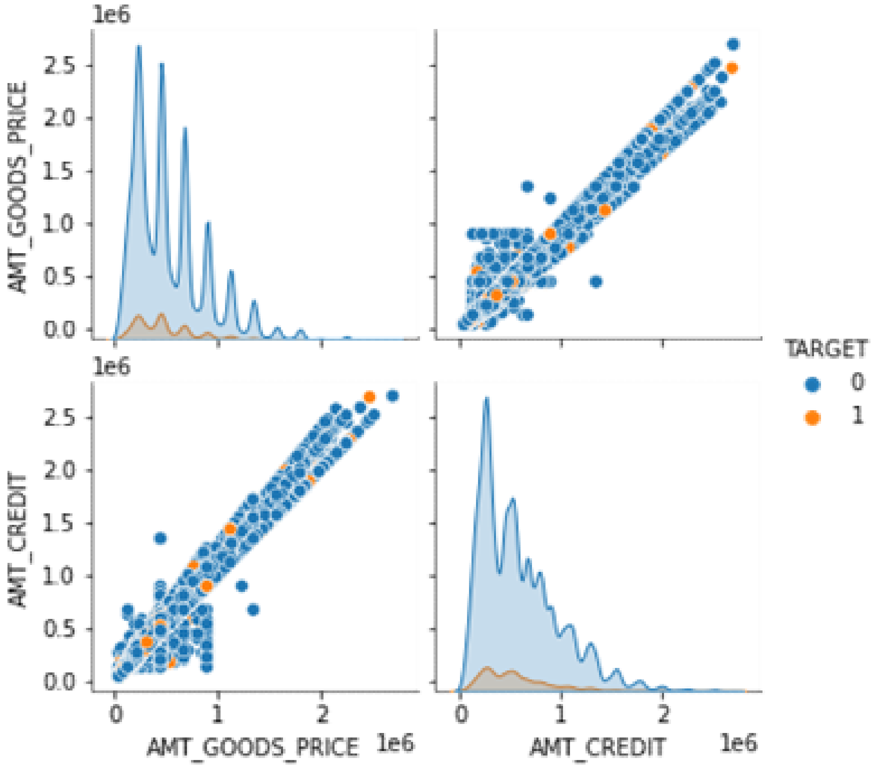
AMT_ANNUITY

- Quite similar distribution for both Defaulter and Non Defaulter

Univariate Analysis of Numerical Variables



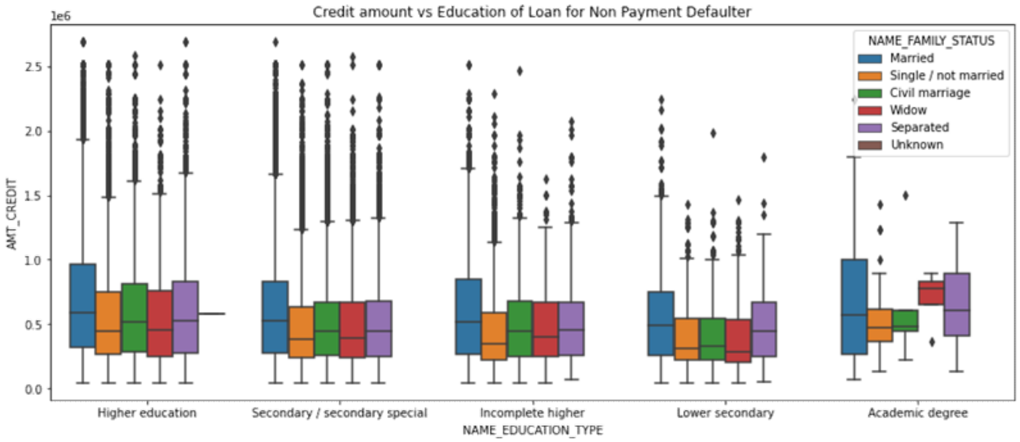


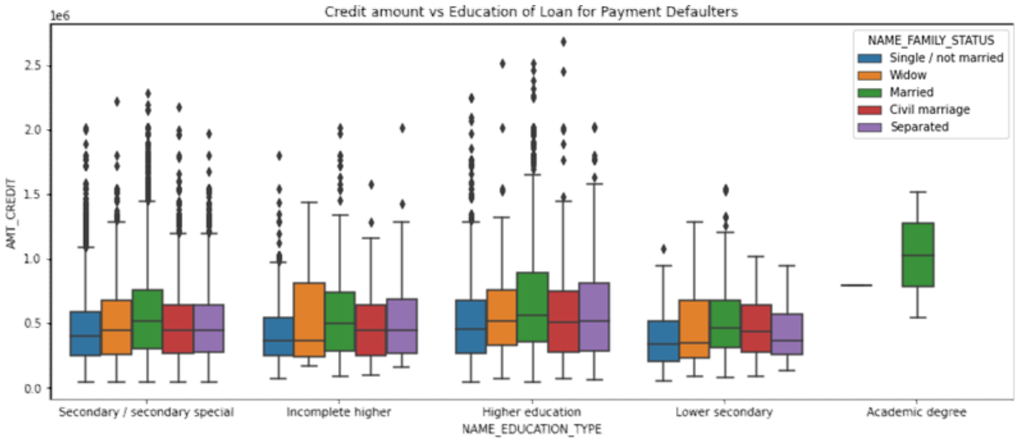


- AMT_CREDIT and AMT_GOODS_PRICE have linear relation.
- For lower range of AMT_CREDIT and AMT_GOODS_PRICE, default rate is high but amount of defaulters is less as distribution is narrow.

Bivariate Analysis of Categorical vs. Numerical Variables

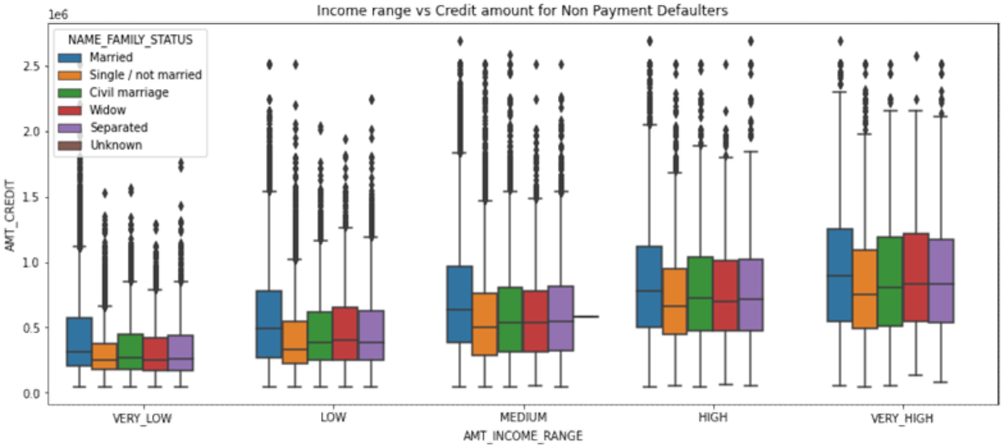
Most of the outliers are from Education type 'Higher education' and 'Secondary'. Civil marriage for Academic degree is having most of the credits in the third quartile.

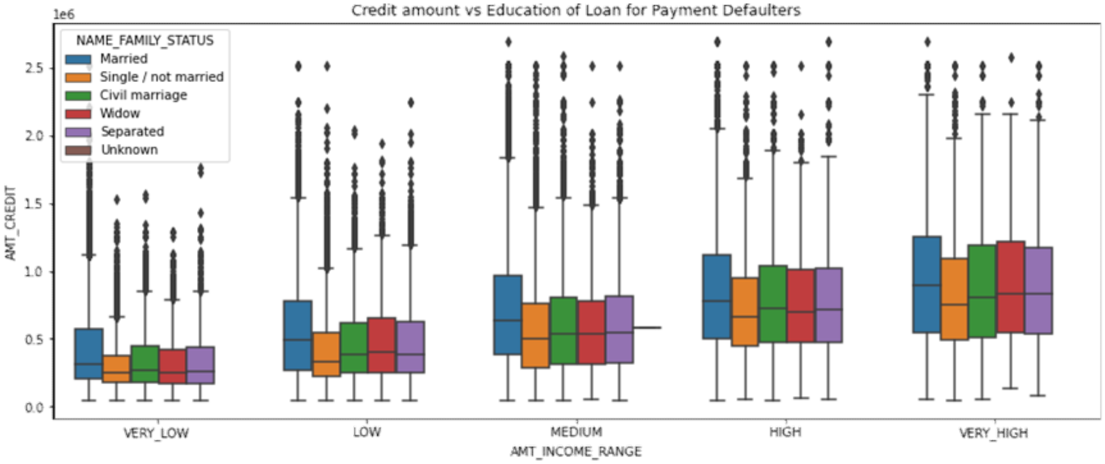




The graphs for Loan Payment Difficulties and Loan Non-Payment Difficulties appears to be similar. We observe that Family status of 'civil marriage', 'marriage' and 'separated' of Academic degree education are having higher number of credits than others.

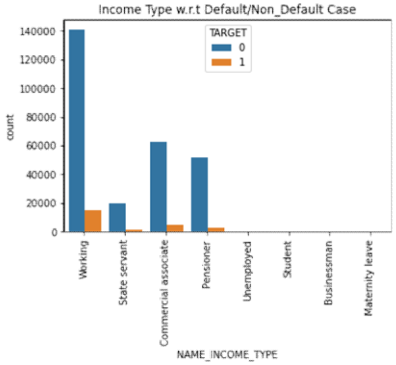
Bivariate Analysis of Categorical vs. Numerical Variables



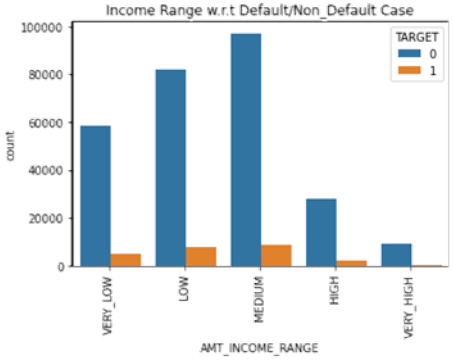


The graphs for Loan Payment Difficulties and Loan- Non Payment Difficulties appears to be similar. We observe that Family status of 'single', 'separated' and 'married' of income range very high are having higher number of credits than others.

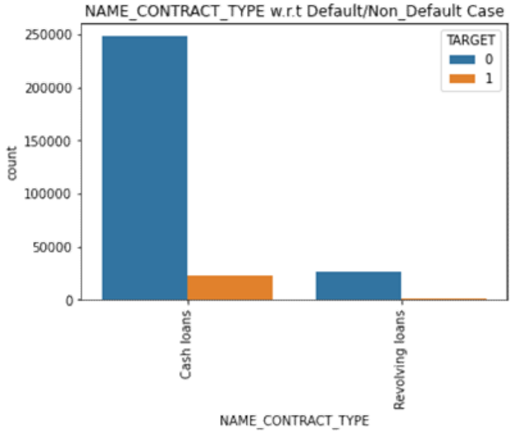
Bivariate Analysis of Categorical vs. Categorical Variables



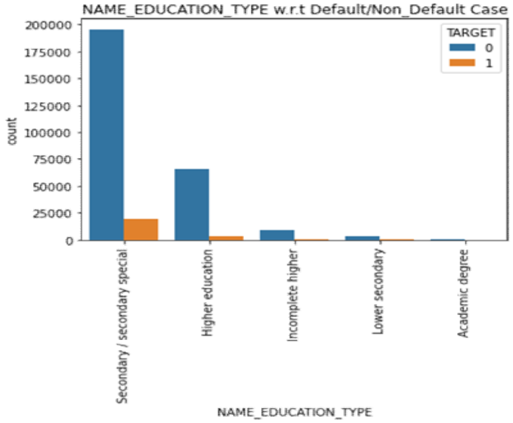
Maximum number of clients in dataset are Working. Default rate is high for clients who are Unemployed or are on Maternity leave



From the plot above we can say that clients with 'LOW' Income range have maximum % of Loan-Payment Difficulties.



Number Cash loans is quite higher than Revolving Loans. Clients with 'Cash loans' contract type have maximum % of Loan-Payment Difficulties.



Clients with 'Lower secondary' education type have maximum % of Loan-Payment Difficulties. Maximum clients are from Secondary/ secondary special and Higher Education background. Default rate is quite low for Higher Education clients.

Bivariate Analysis of Categorical vs. Categorical Variables

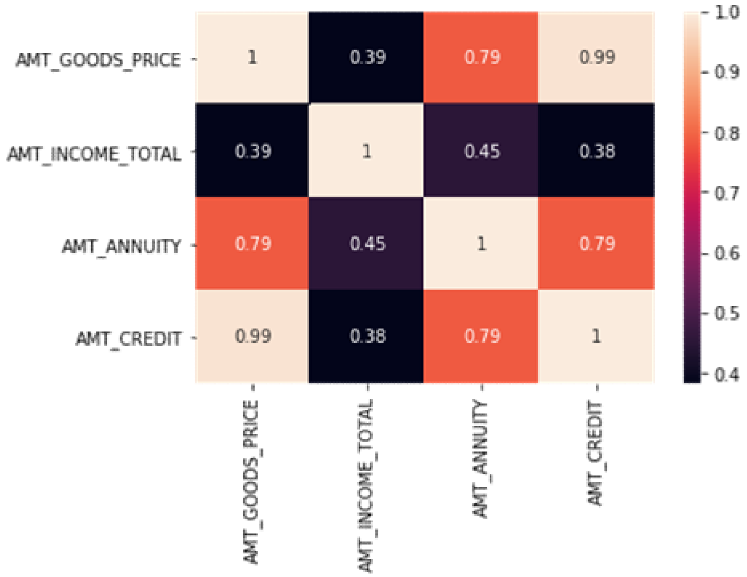
	NAME_EDUCATION_TYPE	Academic degree	Higher education	Incomplete higher	Lower secondary	Secondary / secondary special
CODE_GENDER	AMT_INCOME_RANGE					
F	VERY_LOW	0.000000	0.055954	0.086399	0.080271	0.076729
	LOW	0.000000	0.048882	0.080340	0.114206	0.079530
	MEDIUM	0.000000	0.050563	0.078257	0.097403	0.076118
	HIGH	0.071429	0.041831	0.077329	0.044118	0.071437
	VERY_HIGH	0.181818	0.041266	0.081395	0.000000	0.072107
M	VERY_LOW	0.000000	0.080411	0.123967	0.125000	0.118089
	LOW	0.000000	0.073346	0.096882	0.143451	0.123830
	MEDIUM	0.000000	0.070729	0.095836	0.150206	0.113625
	HIGH	0.000000	0.055804	0.075332	0.086022	0.094277
	VERY_HIGH	0.000000	0.048060	0.081897	0.105263	0.084774

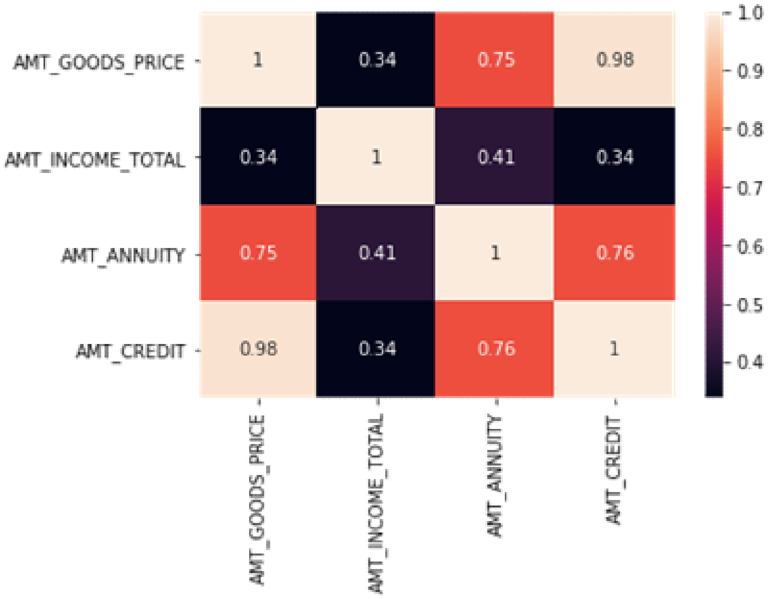
From Female and Male category Clients LOWER SECONDARY education have maximum % of Loan-Payment Difficulties.

In Females, it comes under Income Range : Low

In Males, it comes under Income Range : Medium

Bivariate Analysis of Numerical vs. Numerical Variables



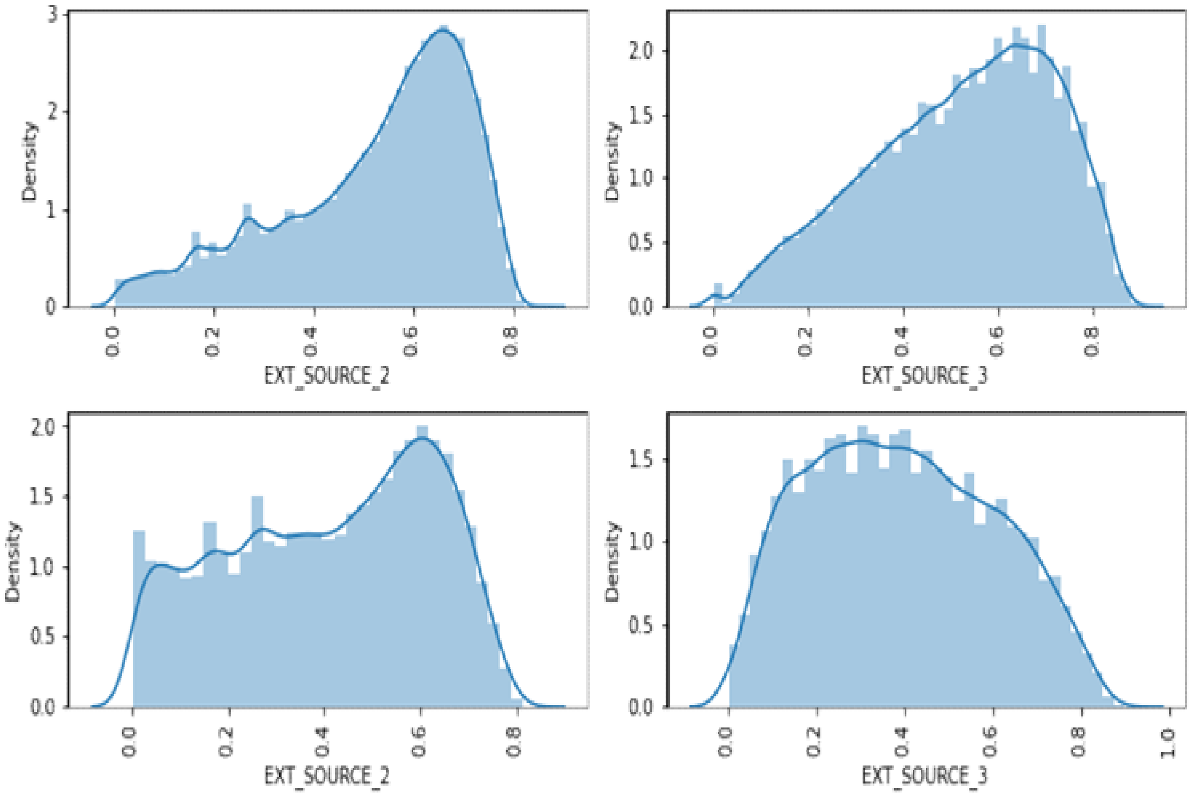


We observe that there is a high correlation between credit amount and goods price. There appears to be some deviancies in the correlation of Loan-Payment Difficulties and Loan- Non Payment Difficulties such as credit amount v/s income.

Bivariate Analysis of Numerical vs. Numerical Variables

EXT_SOURCE_2 has quite similar distribution for non defaulters and defaulter. Plus, no more description or info is present for this feature. Thereby dropping it will not affect the further analysis.

EXT_SOURCE_3 have very different distribution for defaulters and non-defaulters.



Top 10 Correlation for Defaulters

AMT_REQ_CREDIT_BUREAU_YEAR	AMT_REQ_CREDIT_BUREAU_YEAR	1.000000
OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.998262
AMT_CREDIT	AMT_GOODS_PRICE	0.982050
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.957471
CNT_FAM_MEMBERS	CNT_CHILDREN	0.884929
DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.868664
AMT_CREDIT	AMT_ANNUITY	0.757027
AMT_ANNUITY	AMT_GOODS_PRICE	0.753462
REG_CITY_NOT_LIVE_CITY	REG_CITY_NOT_WORK_CITY	0.471793
AMT_ANNUITY	AMT_INCOME_TOTAL	0.414203

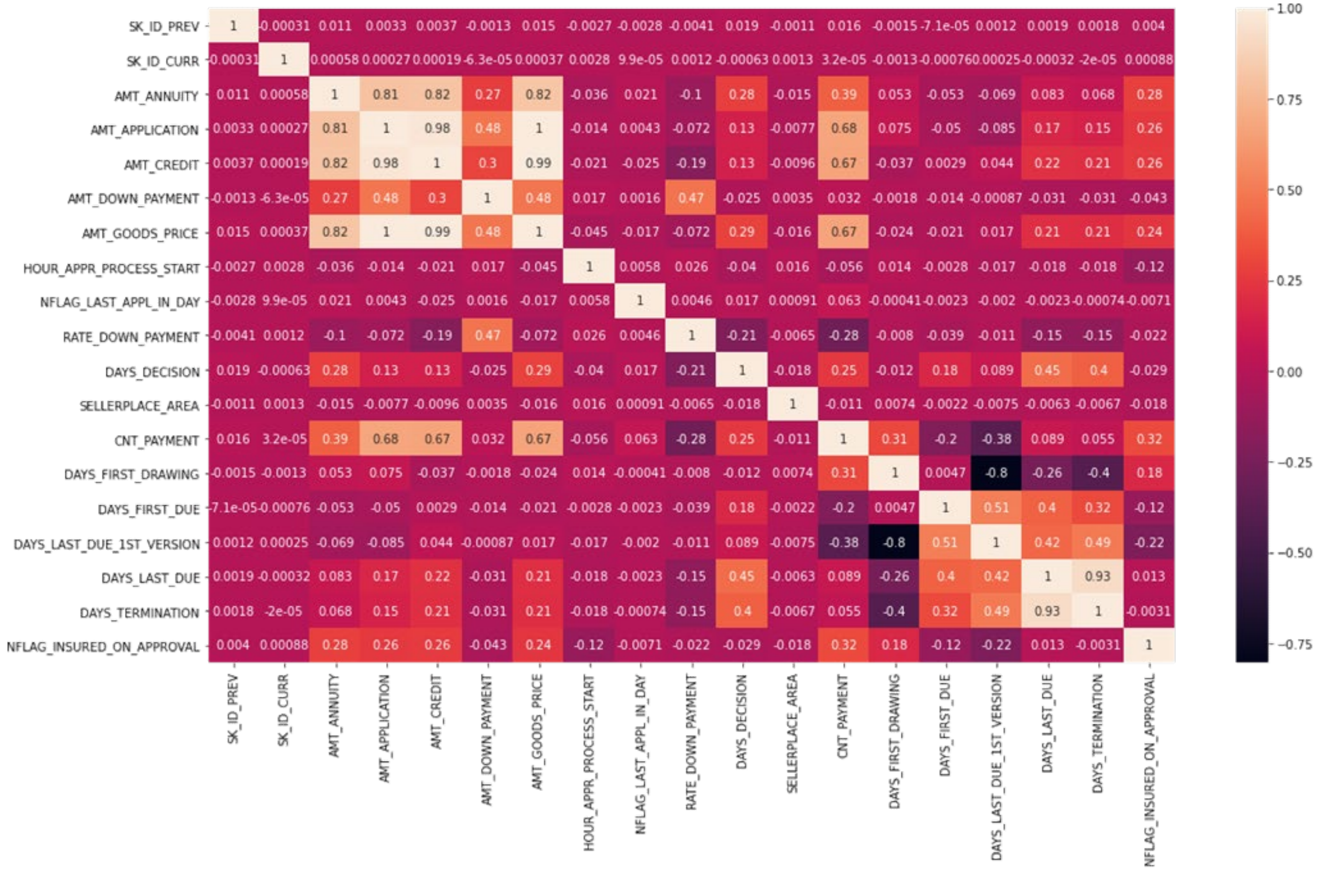
EDA – PREVIOUS APPLICATION DATASET

Correlation between numeric features of previous application data

- DAYS_LAST_DUE and DAYS_TERMINATION are highly correlated

- DAYS_FIRST_DRAWING and DAYS_LAST_DUE_1st_VERSION have high negative correlation

- AMT_ANNUITY , AMT_APPLICATION , AMT_CREDIT , AMT_GOODS_PRICE are highly correlated

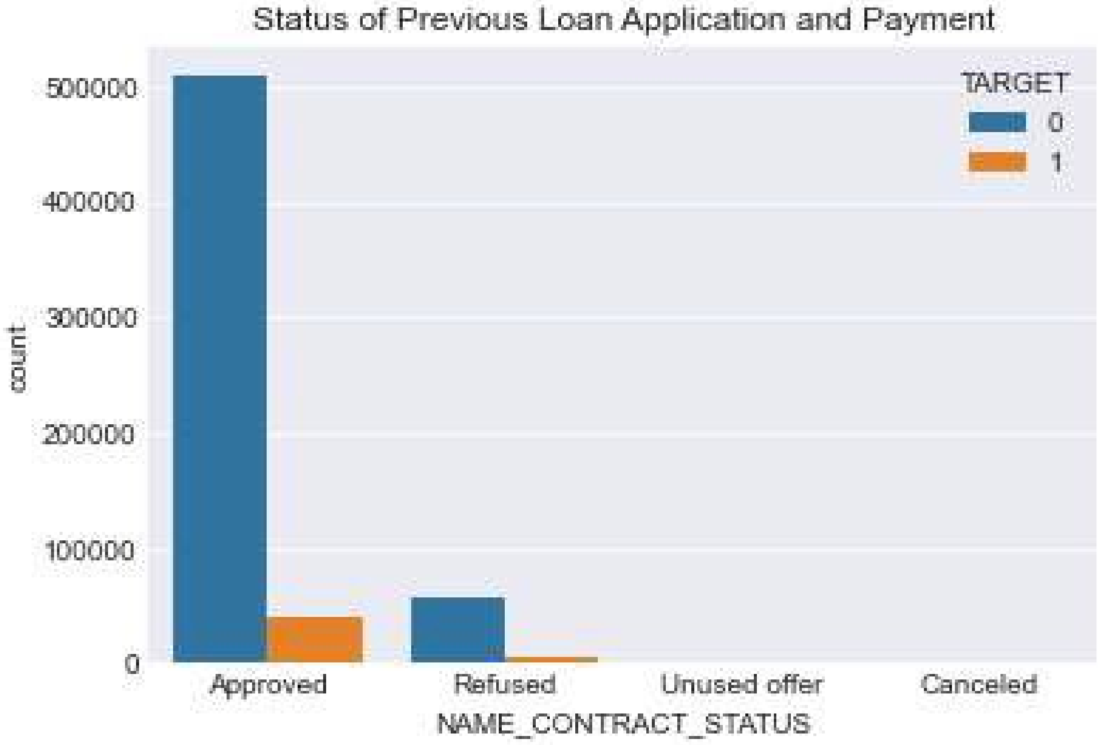


	SK_ID_PREV	SK_ID_CURR	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE	HOUR_APPR_PROCESS_START	NFLAG_LAST_APPL_IN_DAY	RATE_DOWN_PAYMENT	DAYS_DECISION	SELLERPLACE_AREA	CNT_PAYMENT	DAYS_FIRST_DRAWING	DAYS_FIRST_DUE	DAYS_LAST_DUE_1ST_VERSION	DAYS_LAST_DUE	DAYS_TERMINATION	NFLAG_INSURED_ON_APPROVAL
SK_ID_PREV	1	-0.00031	0.011	0.0033	0.0037	-0.0013	0.015	-0.0027	-0.0028	-0.0041	0.019	-0.0011	0.016	-0.0015	-7.1e-05	0.0012	0.0019	0.0018	0.004
SK_ID_CURR	-0.00031	1	0.00058	0.00027	0.00019	-6.3e-05	0.00037	0.0028	9.9e-05	0.0012	-0.00063	0.0013	3.2e-05	-0.0013	-0.00076	0.00025	-0.00032	-2e-05	0.00088
AMT_ANNUITY	0.011	0.00058	1	0.81	0.82	0.27	0.82	-0.036	0.021	-0.1	0.28	-0.015	0.39	0.053	-0.053	-0.069	0.083	0.068	0.28
AMT_APPLICATION	0.0033	0.00027	0.81	1	0.98	0.48	1	-0.014	0.0043	-0.072	0.13	-0.0077	0.68	0.075	-0.05	-0.085	0.17	0.15	0.26
AMT_CREDIT	0.0037	0.00019	0.82	0.98	1	0.3	0.99	-0.021	-0.025	-0.19	0.13	-0.0096	0.67	-0.037	0.0029	0.044	0.22	0.21	0.26
AMT_DOWN_PAYMENT	-0.0013	-6.3e-05	0.27	0.48	0.3	1	0.48	0.017	0.0016	0.47	-0.025	0.0035	0.032	-0.0018	-0.014	-0.00087	-0.031	-0.031	-0.043
AMT_GOODS_PRICE	0.015	0.00037	0.82	1	0.99	0.48	1	-0.045	-0.017	-0.072	0.29	-0.016	0.67	-0.024	-0.021	0.017	0.21	0.21	0.24
HOUR_APPR_PROCESS_START	-0.0027	0.0028	-0.036	-0.014	-0.021	0.017	-0.045	1	0.0058	0.026	-0.04	0.016	-0.056	0.014	-0.0028	-0.017	-0.018	-0.018	-0.12
NFLAG_LAST_APPL_IN_DAY	-0.0028	9.9e-05	0.021	0.0043	-0.025	0.0016	-0.017	0.0058	1	0.0046	0.017	0.00091	0.063	-0.00041	-0.0023	-0.002	-0.0023	-0.00074	-0.0071
RATE_DOWN_PAYMENT	-0.0041	0.0012	-0.1	-0.072	-0.19	0.47	-0.072	0.026	0.0046	1	-0.21	-0.0065	-0.28	-0.008	-0.039	-0.011	-0.15	-0.15	-0.022
DAYS_DECISION	0.019	-0.00063	0.28	0.13	0.13	-0.025	0.29	-0.04	0.017	-0.21	1	-0.018	0.25	-0.012	0.18	0.089	0.45	0.4	-0.029
SELLERPLACE_AREA	-0.0011	0.0013	-0.015	-0.0077	-0.0096	0.0035	-0.016	0.016	0.00091	-0.0065	-0.018	1	-0.011	0.0074	-0.0022	-0.0075	-0.0063	-0.0067	-0.018
CNT_PAYMENT	0.016	3.2e-05	0.39	0.68	0.67	0.032	0.67	-0.056	0.063	-0.28	0.25	-0.011	1	0.31	-0.2	-0.38	0.089	0.055	0.32
DAYS_FIRST_DRAWING	-0.0015	-0.0013	0.053	0.075	-0.037	-0.0018	-0.024	0.014	-0.00041	-0.008	-0.012	0.0074	0.31	1	0.0047	-0.8	-0.26	-0.4	0.18
DAYS_FIRST_DUE	-7.1e-05	-0.00076	-0.053	-0.05	0.0029	-0.014	-0.021	-0.0028	-0.0023	-0.039	0.18	-0.0022	-0.2	0.0047	1	0.51	0.4	0.32	-0.12
DAYS_LAST_DUE_1ST_VERSION	0.0012	0.00025	-0.069	-0.085	0.044	-0.00087	0.017	-0.017	-0.002	-0.011	0.089	-0.0075	-0.38	-0.8	0.51	1	0.42	0.49	-0.22
DAYS_LAST_DUE	0.0019	-0.00032	0.083	0.17	0.22	-0.031	0.21	-0.018	-0.0023	-0.15	0.45	-0.0063	0.089	-0.26	0.4	0.42	1	0.93	0.013
DAYS_TERMINATION	0.0018	-2e-05	0.068	0.15	0.21	-0.031	0.21	-0.018	-0.00074	-0.15	0.4	-0.0067	0.055	-0.4	0.32	0.49	0.93	1	-0.0031
NFLAG_INSURED_ON_APPROVAL	0.004	0.00088	0.28	0.26	0.26	-0.043	0.24	-0.12	-0.0071	-0.022	-0.029	-0.018	0.32	0.18	-0.12	-0.22	0.013	-0.0031	1

Data Imbalance in Previous Application Data

•The data is highly imbalance

•Majority of the loans were approved in past and very few are refused and quite null in terms of unused and canceled.



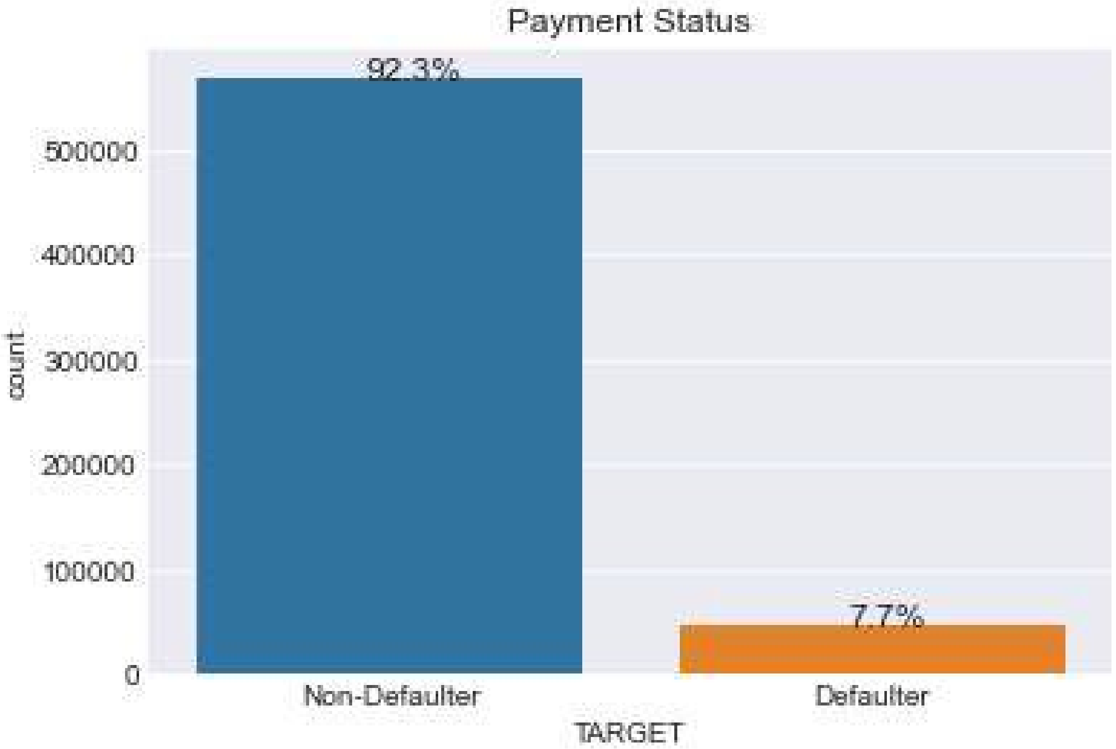
Data Imbalance in Previous Application Data

- The applicants whose previous loans were approved are more likely to pay current loan in time, than the applicants whose previous loans were rejected.

- 7% of the previously approved loan applicants that defaulted in current loan

- 90 % of the previously refused loan applicants that were able to pay current loan

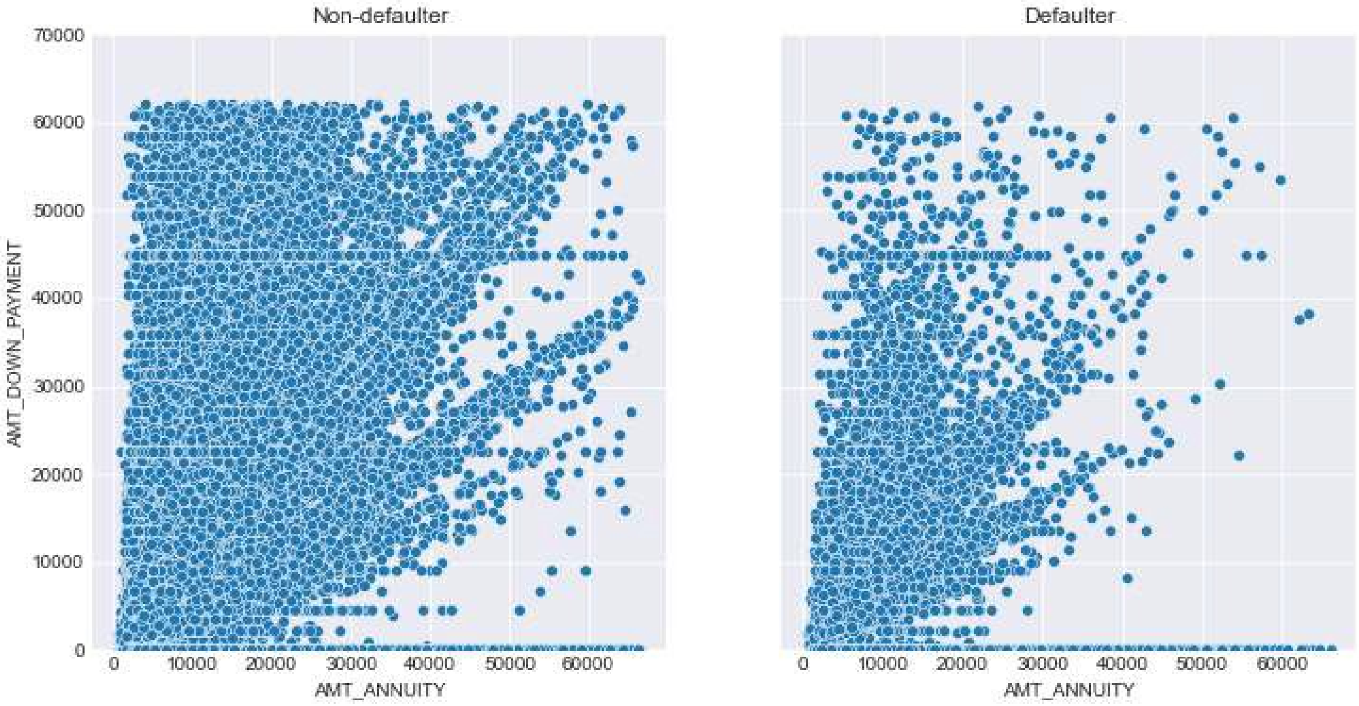
- This data is highly imbalanced as number of defaulter is very less in total population.



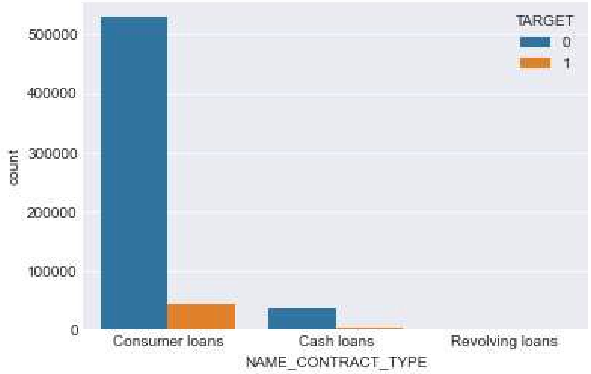
Analysis of Numeric Features of Previous Application Data

•The number of defaulters are getting less as the Amount of Annuity increases of previous application.

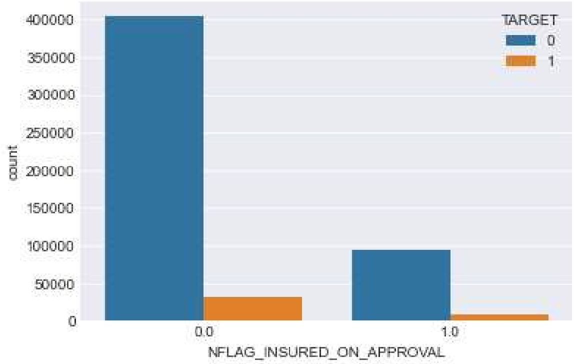
•As amount of down payment increases i.e. higher the down payment less chances of getting default.



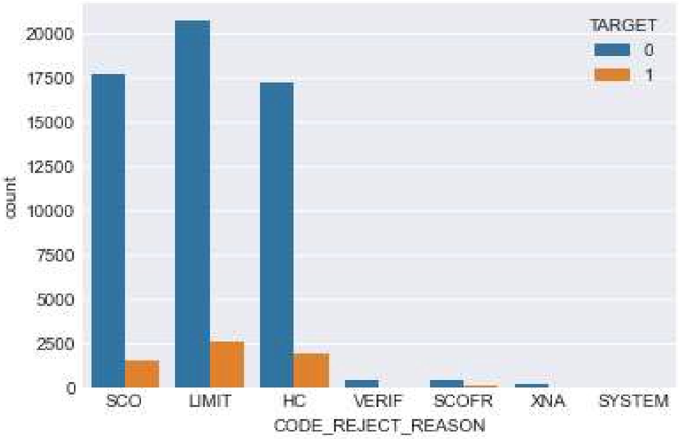
Analysis of Categorical Features of Previous Application Data



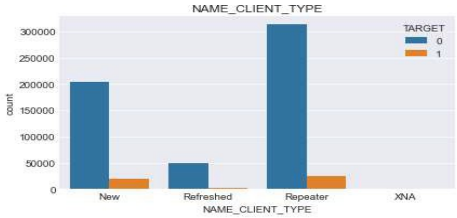
Highest number of loans are applied for Consumer Loans

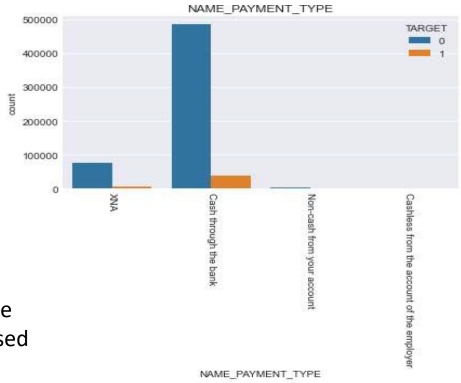


Most of the people did not request insurance during previous loan application.



SCO , LIMIT and HC are the most common reason of rejection.





Most of the applicants
are repeater. Cash through the
bank is the most frequently used
payment method

PREVIOUS APPLICATION CONCLUSION

- There are feature columns in the dataset that are highly correlated to each other. Which means both will have similar impact on the target value. Those features can be removed before feeding this data to a model to avoid collinearity.
- Feature columns with 50% or more missing data can be dropped.
- This dataset is highly imbalanced
- The applicants whose previous loans were approved are more likely to pay current loan in time, than the applicants whose previous loans were rejected. NAME_CONTRACT_STATUS is an important feature.
- 7% of the previously approved loan applicants that defaulted in current loan
- 90 % of the previously refused loan applicants that were able to pay current loan
- SCO, LIMIT and HC are the most common reason of rejection.
- Most of the people did not request insurance during previous loan application.
- For Cards defaulter percentage is highest (17%). NAME_PORTFOLIO is an important feature for analyzing 'TARGET' variable.
- 15% loan applicatiant defaulted for AP+ (Cash Loan). CHANNEL_TYPE is an important feature for analyzing 'TARGET' variable.
- Highest percentage (17%) of default cases is for Card Street. PRODUCT_COMBINATION is an important driving factor.

CURRENT APPLICATION CONCLUSION

- The count of ‘Low skilled Laborers’ in ‘OCCUPATION_TYPE’ is comparatively very less and it also has maximum % of payment difficulties. Hence, client with occupation type as ‘Low skilled Laborers’ are the driving factors for Loan Defaulters.

- The count of ‘Maternity Leave’ in ‘NAME_INCOME_TYPE’ is very less and it also has maximum % of payment difficulties. Hence, client with income type as ‘Maternity leave’ are the driving factors for Loan Defaulters.

- The count of ‘Lower Secondary’ in ‘NAME_EDUCATION_TYPE’ is comparatively very less and it also has maximum % of payment difficulties. Hence, client with education type as ‘Lower Secondary’ are the driving factors for Loan Defaulters.

TOP 5 IMPORTANT COLUMNS

Family Info: Important driving features : 'CNT_FAM_MEMBERS', 'CNT_CHILDREN'

- Most of the clients are married (and/or) no children (and/or) 2 family members.
- Clients with relatively more number of children (and/or) family members have higher default percentage. (For some of the cases where count children/family members is high, and the default rate is very high or very low. This cases cannot be considered for analysis as number of applicants having a large family is very low.)

Education and Occupation Info: Important driving features :'NAME_INCOME_TYPE', 'OCCUPATION_TYPE'

- Most of the clients are working.
- Clients on Maternity Leave and Unemployed has highest percentage of Defaulter
- Businessman have lowest (0) percentage of Defaulter However clients of income type('Unemployed', 'Student', 'Businessman', 'Maternity leave') are very few in the dataset to contribute in the analysis.

CODE_GENDER

- Female clients are more than male clients
- Defaulter percentage is higher for male clients
- XNA values can be replaced with "Female“

DAYS_BIRTH

- Changed the column stats in 'years' from this gave useful information.

•There is an increase in the percentage of Loan Payment Difficulties who are young in age when compared to the percentages of Payment Difficulties and Loan-Non Payment Difficulties.

- Default cases are less for Senior Citizens

EXT_SOURCE_3‘

- Have very different distribution for defaulters and non-defaulters. This can be important features.

SUMMARY

- This data is highly imbalanced as number of defaulter is very less in total population.
- 'CNT_FAM_MEMBERS', 'CNT_CHILDREN','NAME_INCOME_TYPE', 'OCCUPATION_TYPE',CODE_GENDER, and 'EXT_SOURCE_3' are some of the important driving factors.
- **Documents** : Considered features 'FLAG_DOCUMENT_2','FLAG_DOCUMENT_3',...,'FLAG_DOCUMENT_21' for this segment. Majority of the applicants did not submit any documents apart from DOCUMENT_3. FLAG_DOCUMENT_3 has similar impact on defaulters and non-defaulters. Hence these columns can be dropped.
- **Housing:** Plot of 'NAME_HOUSING_TYPE' vs 'TARGET' shows that
  - Most of the applicants live in House/Apartment
  - Applicants living with their parents or in rented apartment have higher rate of default.
- **Social Circle Info:** The features show similar trend for defaulters and non defaulters, can be dropped.
- **Regional Info:**
  - Defaulter rate is highest when REG_REGION_NOT_WORK_REGION=0 i.e. permanent address and working address is same
- **Contact Info :**
  - Considered 'FLAG_MOBIL','FLAG_EMP_PHONE' etc. for this segment. No impact on Target, features can be dropped.
- **Asset Info :**
  - Most of the clients own realty
  - Most of the clients do not own cars
  - People not owning reality and car and have a slightly higher default rate than the people who own reality and car
- **Organization Type :**
  - There is an increase in the percentage of Loan Payment Difficulties who are 'Self-Employed' in organization when compared to the percentages of Payment Difficulties and Loan-Non Payment Difficulties.
- **Occupation Type :**
  - Default % high for 'Low-skill Laborers' as compared to other Occupation Types

THANK YOU