PREDICTION OF LOAN DEFAULTER

Post Graduate Program in Data Science Engineering

Location: **Bangalore**

Batch: **PGPDSE-FT Sep21**

Submitted By:

Shashir Jattyeppa Gornal

Niranjan Gowda

Balaji Hari

Sangamesh gouda

Surajit Sasmal

Manas Sinha

Mentor:

Ms. Vidhya K

ACKNOWLEDGEMENT

Any endeavor in a specific field requires the guidance and support of many people for successful completion. The sense of achievement on completing anything remains incomplete if the people who were instrumental in its execution are not properly acknowledged. We would like to take this opportunity to verbalize our deepest sense of indebtedness to our project mentor, Ms. Vidhya K, who was a constant pillar of support and continually provided us with valuable insights to improve upon our project and make it a success. Further, we would like to thank our parents for encouraging us and providing us a platform wherein we got an opportunity to design our own project.

DECLARATION

We hereby declare, that this submission is our own work and that, to the best of our knowledge and belief, it contains no material previously published or written by another person nor material which has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

TABLE OF CONTENTS

INTRODUCTION:

- Numerous companies from the financial industry often invest considerable resources to improve their predictive models with the aim of having better insights into their customers. Such an interest in model improvement has intensified in recent years mostly because of the fast development of machine learning and artificial intelligence. For standard lending institutions default predictive model with high performance helps to considerably minimize Credit Loss, resulting in higher revenue and profits. Usually, the better the predictive model the more efficient is the underwriting policy and collection process. A well-functioning model should distinguish creditworthy customers from those that are credit risks. Often, the more-predictive credit-decisioning model can identify a greater number of customers within an institution's specified risk tolerance, which should expand revenues as well.
- In this project, the goal is to increase detection of defaulted loans before the loan is issued/offered by a P2P lending company - Lending Club. Peer-to-peer lending differs from traditional financial institutions like banks or commercial lending companies.

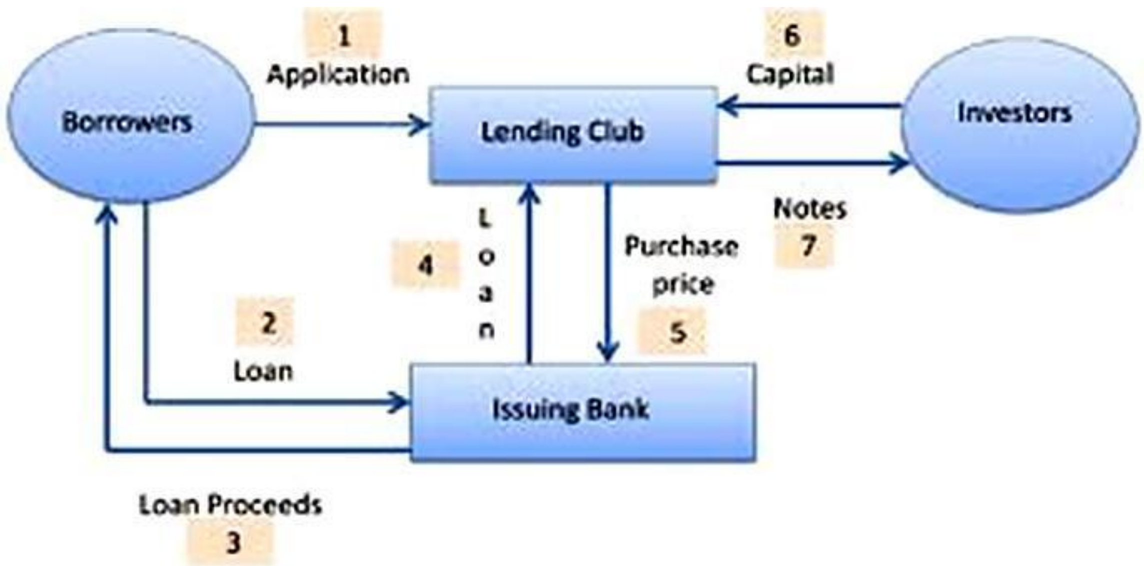
- So, Lending Club is a mediator between investors and borrowers, earning money by charging both. The main Lend Club interest is to attract more clients and maintain portfolio size. The motivation of borrowers is clear, they want to find as cheap capital as possible, so they're seeking for test offer at the market, which is available for them. In the case of investors, the motivation is obvious as well. Investors look for high ROI (return of investments), but remembering that returns are proportional to risks, we may formalize saying, that investors look for appropriate returns/risk ratio. If investors experience losses it may cause churn rate growth.

Problem Statement Analysis:

- Loan default occurs when a borrower fails to pay back a debt according to the initial arrangement. In the case of most consumer loans, this means that successive payments have been missed over the course of weeks or months. Fortunately, lenders and loan servicers usually allow a grace period before penalizing the borrower after missing one payment. The period between missing a loan payment and having the loan default is known as delinquency. The delinquency period gives the debtor time to avoid default by contacting their loan servicer or making up missed payments.

- Defaulting on a loan will cause a substantial and lasting drop in the debtor's credit score, as well as extremely high interest rates on any future loan. For loans secured with collateral, defaulting will likely result in the pledged asset being seized by the bank. The most popular types of consumer loans that are backed by collateral are mortgages, auto loans and secured personal loans.

- The loan is one of the most important products of the banking. All the banks are trying to figure out effective business strategies to persuade customers to apply their loans. However, **there are some customers behave negatively after their application are approved**.



BUSINESS OBJECTIVE:

The loan-providing companies find it hard to give loans to people due to their insufficient or non-existent credit history. Because of that, some consumers use it to their advantage by becoming a defaulter.

- This case study aims to identify patterns that indicate if a client has difficulty paying their 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.
- In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.
- This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicant's using EDA is the aim of this case study

ABOUT THE DATASET:

These files contain complete loan data for all loans issued from 2007-to 2015, including the current loan status (Current, Late, Fully Paid, etc.) and the latest payment information.

The file containing loan data through the "present" contains complete loan data for all loans issued through the previous completed calendar quarter. Additional features include credit scores, number of finance inquiries, addresses including zip codes, state, and collections among others.

Data Understanding:

The shape of the dataset –

Columns - 307511
Rows - 122

No of numerical variables columns - 106
No of categorical variables columns - 15
No of the variables with missing values -67
No of variables with no missing values – 55
Type of data – supervised classification data

COLUMNS DESCRIPTION:

Independent Variables: There are 122 independent variables listed below.

S.NO	Features	Description
1.	NAME_CONTRACT_TYPE	Identification if loan is cash or revolving
2.	CODE_GENDER	Gender of the client
3.	FLAG_OWN_CAR	Flag if the client owns a car
4.	FLAG_OWN_REALTY	Flag if client owns a house or flat
5.	CNT_CHILDREN	Number of children the client has
6.	AMT_INCOME_TOTAL	Income of the client
7.	AMT_CREDIT	Credit amount of the loan
8.	AMT_ANNUITY	Loan annuity
9.	AMT_GOODS_PRICE	Goods price of good that client asked for
10.	NAME_TYPE_SUITE	Who was accompanying client when he was applying for the loan
11.	NAME_INCOME_TYPE	Family status of the client
12.	NAME_EDUCATION_TYPE	Level of highest education the client achieved
13.	NAME_FAMILY_STATUS	Family status of the client
14.	NAME_HOUSING_TYPE	What is the housing situation of the client (renting, living with parents etc
15.	REGION_POPULATION_RE	Normalized population of region where client lives

	LATIVE	
16.	DAYS_BIRTH	Client's age in days at the time of application
17.	DAYS_EMPLOYED	How many days before the application the person started current employment
18.	DAYS_REGISTRATION	How many days before the application did client registration
19.	DAYS_ID_PUBLISH	How many days before the application did client change his registration
20.	FLAG_MOBIL	Did client provide mobile phone (1=YES, 0=NO)
21.	FLAG_EMP_PHONE	Did client provide home phone (1=YES, 0=NO)
22.	FLAG_WORK_PHONE	Did client provide home phone (1=YES, 0=NO)
23.	FLAG_CONT_MOBILE	Was mobile phone reachable (1=YES, 0=NO)
24.	FLAG_PHONE	Did client provide work phone (1=YES, 0=NO)
25.	FLAG_EMAIL	Did client provide Email (1=YES, 0=NO)
26.	OCCUPATION_TYPE	What kind of occupation does the client have
27.	CNT_FAM_MEMBERS	How many family members does client have
28.	REGION_RATING_CLIENT	Our rating of the region where client lives (1,2,3)
29.	REGION_RATING_CLIENT_W_CITY	Our rating of the region where client lives with taking city into account (1, 2,3)
30.	WEEKDAY_APPR_PROCESS_START	On which day of the week did the client apply
31.	HOUR_APPR_PROCESS_START	Approximately at what day hour did the client applied
32.	REG_REGION_NOT_LIVE_REGION	Flag if client's permanent address does not match contact address (1=different, 0=same, at region level)
33.	REG_REGION_NOT_WORK_REGION	Flag if client's permanent address does not match work address (1=different, 0=same, at region level)
34.	LIVE_REGION_NOT_WORK_REGION	Flag if client's contact address does not match work address (1=different, 0=same, at region level)
35.	REG_CITY_NOT_LIVE_CITY	Flag if client's permanent address does not match contact address (1=different, 0=same, at city level)
36.	REG_CITY_NOT_WORK_CITY	Flag if client's permanent address does not match work address (1=different, 0=same, at city level)
37.	LIVE_CITY_NOT_WORK_CITY	Flag if client's contact address does not match work address (1=different, 0=same, at city level)
38.	EXT_SOURCE_2	Normalized score from external data source
39.		
40.	EXT_SOURCE_3	Normalized score from external data source
41.	OBS_30_CNT_SOCIAL_CIRCLE	How many observation of client's social surroundings with observable 30 DPD (days past due) default
42.	DEF_30_CNT_SOCIAL_CIRCLE	How many observation of client's social surroundings defaulted on 30 DPD (days past due)
43.	OBS_60_CNT_SOCIAL_CIRCLE	How many observation of client's social surroundings with observable 60 DPD (days past due) default
44.	DEF_60_CNT_SOCIAL_CIRCLE	How many observation of client's social surroundings defaulted on 60 (days past due) DPD
45.	DAYS_LAST_PHONE_CHANGE	How many days before application did client change phone
46.	AMT_REQ_CREDIT_BUREAU_HOUR	Number of enquiries to Credit Bureau about the client one hour before application
47.	AMT_REQ_CREDIT_BUREAU_DAY	Number of enquiries to Credit Bureau about the client one day before application
48.	AMT_REQ_CREDIT_BUREAU_WEEK	Number of enquiries to Credit Bureau about the client week before application
49.	AMT_REQ_CREDIT_BUREAU_MON	Number of enquiries to Credit Bureau about the client one month before application

50.	AMT_REQ_CREDIT_BUREAU_QRT	Number of enquiries to Credit Bureau about the client one quater before application
51.	AMT_REQ_CREDIT_BUREAU_YEAR	Number of enquiries to Credit Bureau about the client one year before application
52.	AMT_INCOME_GROUP	Grouped Variable Amount Income
53	AMT_CREDIT_GROUP	Grouped Variable for Amount Credit

Target Variables:

S.N0	Row	Description
1.	TARGET	Target variable (1 - client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample, 0 - all other cases)

VARIABLE AND ITS DATA TYPE:

VARIBLES AND ITS TYPES

0	SK_ID_CURR	int64	30	REGION_RATING_CLIENT	int64
1	TARGET	int64	31	REGION_RATING_CLIENT_W_CITY	int64
2	NAME_CONTRACT_TYPE	object	32	WEEKDAY_APPR_PROCESS_START	object
3	CODE_GENDER	object	33	HOUR_APPR_PROCESS_START	int64
4	FLAG_OWN_CAR	object	34	REG_REGION_NOT_LIVE_REGION	int64
5	FLAG_OWN_REALTY	object	35	REG_REGION_NOT_WORK_REGION	int64
6	CNT_CHILDREN	int64	36	LIVE_REGION_NOT_WORK_REGION	int64
7	AMT_INCOME_TOTAL	float64	37	REG_CITY_NOT_LIVE_CITY	int64
8	AMT_CREDIT	float64	38	REG_CITY_NOT_WORK_CITY	int64
9	AMT_ANNUITY	float64	39	LIVE_CITY_NOT_WORK_CITY	int64
10	AMT_GOODS_PRICE	float64	40	ORGANIZATION_TYPE	object
11	NAME_TYPE_SUITE	object	41	EXT_SOURCE_1	float64
12	NAME_INCOME_TYPE	object	42	EXT_SOURCE_2	float64
13	NAME_EDUCATION_TYPE	object	43	EXT_SOURCE_3	float64
14	NAME_FAMILY_STATUS	object	44	APARTMENTS_AVG	float64
15	NAME_HOUSING_TYPE	object	45	BASEMENTAREA_AVG	float64
16	REGION_POPULATION_RELATIVE	float64	46	YEARS_BEGINEXPLUATATION_AVG	float64
17	DAYS_BIRTH	int64	47	YEARS_BUILD_AVG	float64
18	DAYS_EMPLOYED	int64	48	COMMONAREA_AVG	float64
19	DAYS_REGISTRATION	float64	49	ELEVATORS_AVG	float64
20	DAYS_ID_PUBLISH	int64	50	ENTRANCES_AVG	float64
21	OWN_CAR_AGE	float64	51	FLOORSMAX_AVG	float64
22	FLAG_MOBIL	int64	52	FLOORSMIN_AVG	float64
23	FLAG_EMP_PHONE	int64	53	LANDAREA_AVG	float64
24	FLAG_WORK_PHONE	int64	54	LIVINGAPARTMENTS_AVG	float64
25	FLAG_CONT_MOBILE	int64	55	LIVINGAREA_AVG	float64
26	FLAG_PHONE	int64	56	NONLIVINGAPARTMENTS_AVG	float64
27	FLAG_EMAIL	int64	57	NONLIVINGAREA_AVG	float64
28	OCCUPATION_TYPE	object	58	APARTMENTS_MODE	float64
29	CNT_FAM_MEMBERS	float64	59	BASEMENTAREA_MODE	float64

90	EMERGENCYSTATE_MODE	object	90	EMERGENCYSTATE_MODE	object
91	OBS_30_CNT_SOCIAL_CIRCLE	float64	91	OBS_30_CNT_SOCIAL_CIRCLE	float64
92	DEF_30_CNT_SOCIAL_CIRCLE	float64	92	DEF_30_CNT_SOCIAL_CIRCLE	float64
93	OBS_60_CNT_SOCIAL_CIRCLE	float64	93	OBS_60_CNT_SOCIAL_CIRCLE	float64
94	DEF_60_CNT_SOCIAL_CIRCLE	float64	94	DEF_60_CNT_SOCIAL_CIRCLE	float64
95	DAYS_LAST_PHONE_CHANGE	float64	95	DAYS_LAST_PHONE_CHANGE	float64
96	FLAG_DOCUMENT_2	int64	96	FLAG_DOCUMENT_2	int64
97	FLAG_DOCUMENT_3	int64	97	FLAG_DOCUMENT_3	int64
98	FLAG_DOCUMENT_4	int64	98	FLAG_DOCUMENT_4	int64
99	FLAG_DOCUMENT_5	int64	99	FLAG_DOCUMENT_5	int64
100	FLAG_DOCUMENT_6	int64	100	FLAG_DOCUMENT_6	int64
101	FLAG_DOCUMENT_7	int64	101	FLAG_DOCUMENT_7	int64
102	FLAG_DOCUMENT_8	int64	102	FLAG_DOCUMENT_8	int64
103	FLAG_DOCUMENT_9	int64	103	FLAG_DOCUMENT_9	int64
104	FLAG_DOCUMENT_10	int64	104	FLAG_DOCUMENT_10	int64
105	FLAG_DOCUMENT_11	int64	105	FLAG_DOCUMENT_11	int64
106	FLAG_DOCUMENT_12	int64	106	FLAG_DOCUMENT_12	int64
107	FLAG_DOCUMENT_13	int64	107	FLAG_DOCUMENT_13	int64
108	FLAG_DOCUMENT_14	int64	108	FLAG_DOCUMENT_14	int64
109	FLAG_DOCUMENT_15	int64	109	FLAG_DOCUMENT_15	int64
110	FLAG_DOCUMENT_16	int64	110	FLAG_DOCUMENT_16	int64
111	FLAG_DOCUMENT_17	int64	111	FLAG_DOCUMENT_17	int64
112	FLAG_DOCUMENT_18	int64	112	FLAG_DOCUMENT_18	int64
113	FLAG_DOCUMENT_19	int64	113	FLAG_DOCUMENT_19	int64
114	FLAG_DOCUMENT_20	int64	114	FLAG_DOCUMENT_20	int64
115	FLAG_DOCUMENT_21	int64	115	FLAG_DOCUMENT_21	int64
116	AMT_REQ_CREDIT_BUREAU_HOUR	float64	116	AMT_REQ_CREDIT_BUREAU_HOUR	float64
117	AMT_REQ_CREDIT_BUREAU_DAY	float64	117	AMT_REQ_CREDIT_BUREAU_DAY	float64
118	AMT_REQ_CREDIT_BUREAU_WEEK	float64	118	AMT_REQ_CREDIT_BUREAU_WEEK	float64
119	AMT_REQ_CREDIT_BUREAU_MON	float64	119	AMT_REQ_CREDIT_BUREAU_MON	float64
120	AMT_REQ_CREDIT_BUREAU_QRT	float64	120	AMT_REQ_CREDIT_BUREAU_QRT	float64
121	AMT_REQ_CREDIT_BUREAU_YEAR	float64	121	AMT_REQ_CREDIT_BUREAU_YEAR	float64

NUMERICAL VARIABLES: 105

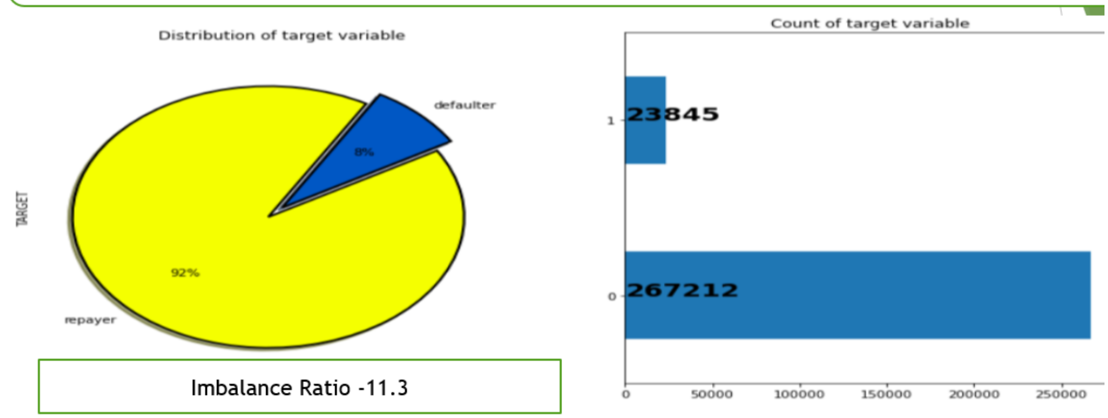
0	SK_ID_CURR	int64
1	TARGET	int64
2	CNT_CHILDREN	int64
3	AMT_INCOME_TOTAL	float64
4	AMT_CREDIT	float64
5	AMT_ANNUITY	float64
6	AMT_GOODS_PRICE	float64
7	REGION_POPULATION_RELATIVE	float64
8	DAYS_BIRTH	int64
9	DAYS_EMPLOYED	int64
10	DAYS_REGISTRATION	float64
11	DAYS_ID_PUBLISH	int64
12	OWN_CAR_AGE	float64
13	FLAG_MOBIL	int64
14	FLAG_EMP_PHONE	int64
15	FLAG_WORK_PHONE	int64
16	FLAG_CONT_MOBILE	int64
17	FLAG_PHONE	int64
18	FLAG_EMAIL	int64
19	CNT_FAM_MEMBERS	float64
20	REGION_RATING_CLIENT	int64
21	REGION_RATING_CLIENT_W_CITY	int64
22	HOUR_APPR_PROCESS_START	int64
23	REG_REGION_NOT_LIVE_REGION	int64
24	REG_REGION_NOT_WORK_REGION	int64
25	LIVE_REGION_NOT_WORK_REGION	int64
26	REG_CITY_NOT_LIVE_CITY	int64
27	REG_CITY_NOT_WORK_CITY	int64
28	LIVE_CITY_NOT_WORK_CITY	int64
29	EXT_SOURCE_1	float64
30	EXT_SOURCE_2	float64
31	EXT_SOURCE_3	float64
33	BASEMENTAREA_AVG	float64
34	YEARS_BEGINEXPLUATATION_AVG	float64
35	YEARS_BUILD_AVG	float64
36	COMMONAREA_AVG	float64
37	ELEVATORS_AVG	float64
38	ENTRANCES_AVG	float64
39	FLOORSMAX_AVG	float64
40	FLOORSMIN_AVG	float64
41	LANDAREA_AVG	float64
42	LIVINGAPARTMENTS_AVG	float64
43	LIVINGAREA_AVG	float64
44	NONLIVINGAPARTMENTS_AVG	float64
45	NONLIVINGAREA_AVG	float64
46	APARTMENTS_MODE	float64
47	BASEMENTAREA_MODE	float64
48	YEARS_BEGINEXPLUATATION_MODE	float64
49	YEARS_BUILD_MODE	float64
50	COMMONAREA_MODE	float64
51	ELEVATORS_MODE	float64
52	ENTRANCES_MODE	float64
53	FLOORSMAX_MODE	float64
54	FLOORSMIN_MODE	float64
55	LANDAREA_MODE	float64
56	LIVINGAPARTMENTS_MODE	float64
57	LIVINGAREA_MODE	float64
58	NONLIVINGAPARTMENTS_MODE	float64
59	NONLIVINGAREA_MODE	float64
60	APARTMENTS_MEDI	float64
61	BASEMENTAREA_MEDI	float64
62	YEARS_BEGINEXPLUATATION_MEDI	float64
63	YEARS_BUILD_MEDI	float64
64	COMMONAREA_MEDI	float64
65	ELEVATORS_MEDI	float64
66	ENTRANCES_MEDI	float64
72	NONLIVINGAPARTMENTS_MEDI	float64
73	NONLIVINGAREA_MEDI	float64
74	TOTALAREA_MODE	float64
75	OBS_30_CNT_SOCIAL_CIRCLE	float64
76	DEF_30_CNT_SOCIAL_CIRCLE	float64
77	OBS_60_CNT_SOCIAL_CIRCLE	float64
78	DEF_60_CNT_SOCIAL_CIRCLE	float64
79	DAYS_LAST_PHONE_CHANGE	float64
80	FLAG_DOCUMENT_2	int64
81	FLAG_DOCUMENT_3	int64
82	FLAG_DOCUMENT_4	int64
83	FLAG_DOCUMENT_5	int64
84	FLAG_DOCUMENT_6	int64
85	FLAG_DOCUMENT_7	int64
86	FLAG_DOCUMENT_8	int64
87	FLAG_DOCUMENT_9	int64
88	FLAG_DOCUMENT_10	int64
89	FLAG_DOCUMENT_11	int64
90	FLAG_DOCUMENT_12	int64
91	FLAG_DOCUMENT_13	int64
92	FLAG_DOCUMENT_14	int64
93	FLAG_DOCUMENT_15	int64
94	FLAG_DOCUMENT_16	int64
95	FLAG_DOCUMENT_17	int64
96	FLAG_DOCUMENT_18	int64
97	FLAG_DOCUMENT_19	int64
98	FLAG_DOCUMENT_20	int64
99	FLAG_DOCUMENT_21	int64
100	AMT_REQ_CREDIT_BUREAU_HOUR	float64
101	AMT_REQ_CREDIT_BUREAU_DAY	float64
102	AMT_REQ_CREDIT_BUREAU_WEEK	float64
103	AMT_REQ_CREDIT_BUREAU_MON	float64
104	AMT_REQ_CREDIT_BUREAU_QRT	float64
105	AMT_REQ_CREDIT_BUREAU_YEAR	float64

CATEGORICAL VARIABLES: 15

```
NAME_CONTRACT_TYPE            object
CODE_GENDER                   object
FLAG_OWN_CAR                  object
FLAG_OWN_REALTY               object
NAME_TYPE_SUITE               object
NAME_INCOME_TYPE              object
NAME_EDUCATION_TYPE           object
NAME_FAMILY_STATUS            object
NAME_HOUSING_TYPE             object
OCCUPATION_TYPE               object
WEEKDAY_APPR_PROCESS_START    object
ORGANIZATION_TYPE             object
FONDKAPREMONT_MODE            object
HOUSETYPE_MODE                object
WALLSMATERIAL_MODE            object
EMERGENCYSTATE_MODE           object
```

TARGET VARIABLES:

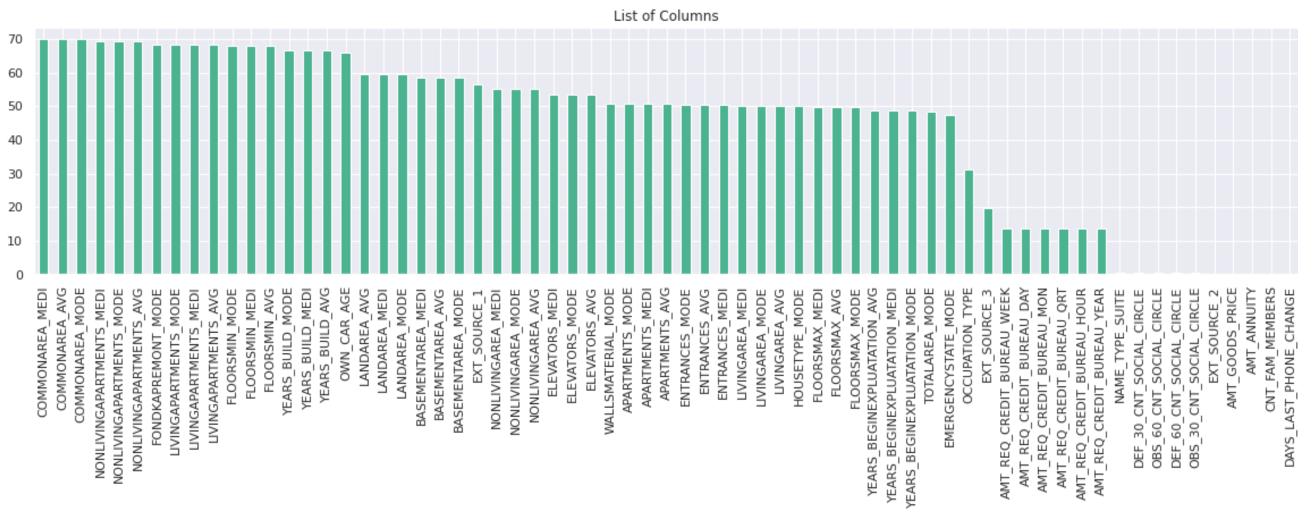
TARGET VARIABLES - IMBALANCE



'0' means the client has no problem in paying the installments or the loan amount
'1' means has difficulties in paying
92% of the customers have no problem in paying the amount (non-defaulter)
8% of the clients have a problem in paying (defaulter)
Clearly, it shows there is an imbalance in the data

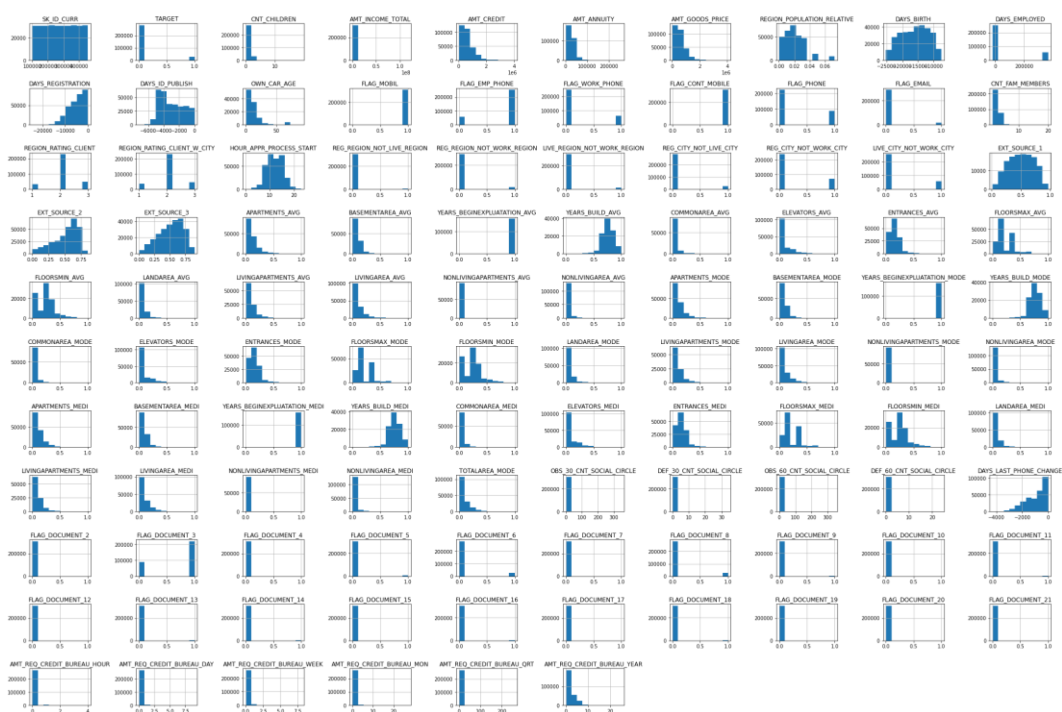
The Target Variable Is Imbalance So We Try To Use smote Technique To Counter The Imbalance Problem

VARIABLE WITH MISSING VALUES:



COUNT OF VARIABLES WITH MISSING VALUES - 67
COUNT OF VARIABLES WITH NO MISSING VALUES - 55

DISTRIBUTION OF NUMERICAL DATA:



- Most of the variables does follow the near-normal distribution

METHODOLOGY TO BE FOLLOWED

CRISP-DM which stands for Cross Industry Standard Process for Data Mining is a methodology created to help shape data mining projects. It describes the different phases/tasks involved in the project and provides an overview of data mining life cycle.

Business Understanding –

It focuses on determining the business requirements/objective and understanding what outcome to achieve. Also determine the business units being affected. Convert this business problem into a data mining problem and carve out an initial plan.

• Determine the business objectives: Understand what is needed to be accomplished for the customer.

• Assess the situation: Determine resources availability, project requirements, assess risks and contingencies, and conduct a cost-benefit analysis.

• Determine data mining goals: Convert business problem to a data mining problem and recognize the data mining problem type such as classification, regression or clustering, etc.

• Produce a project plan: Devise a step-to-step plan for executing the project.

Data understanding –

This phase starts with collecting the data and then examining the data for its surface properties like data format, number of records, etc. The next step is to better understand the data by understanding each attribute and perform basic statistics on them. Understand the relationship between different attributes. Determine the quality of data by checking the missing values, outliers, duplicates, etc.

• Collect initial data: Acquire the data and load it into the analysis tool to be used.

• Describe data: Examine the data and document its surface properties like data format, number of records, or field identities. Understand the meaning of each attribute and attribute value in business terms. For each attribute, compute basic statistics so as to get a higher-level understanding.

• Explore data: Find insights from the data. Query it, visualize it, and identify relationships among the data.

Data preparation –

This stage, which is often referred to as data wrangling, has the objective to develop the final data set for EDA and modelling. Covers all activities to construct the final dataset from the initial raw data. Some of the tasks include table, record and attribute selection as well as transformation and cleaning of data for modelling tools.

• Select data: Determine which attributes/features will be used and document reasons for inclusion/exclusion.

• Clean data: Correct, impute and remove the improper data.

• Extract data: Derive new attributes from the existing ones

• Integrate data: Create features by combining data from multiple sources.

Format data: Re-format data as necessary. For example, convert string values to numeric values so as to perform mathematical operations.

Modelling –

In this stage we build and assess different models built using various techniques from the training dataset.

• Select modelling technique: Determine the algorithms to be used to model the data based on the business requirement.

• Generate test design: In order to build and test the model, we need to divide the dataset into training and testing data set. In this step we divide the data into train and test data set.

• Build model: Based on the modelling technique selected, build the model on the input data set.

• Assess model: Compare the results of different models based on confusion matrix. The outcome of this step frequently leads to model tuning iterations until the best model is found.

Evaluation –

Evaluate the models and review the steps executed to construct the model to be certain it properly achieves the business objectives.

• Evaluate results: Understand the data mining results and check how impactful they are in achieving the data mining goal. Select appropriate model based on confusion matrix.

• Review process: Review the work accomplished and make sure that nothing was overlooked and all steps were properly executed. Summarize the findings and correct anything if needed.

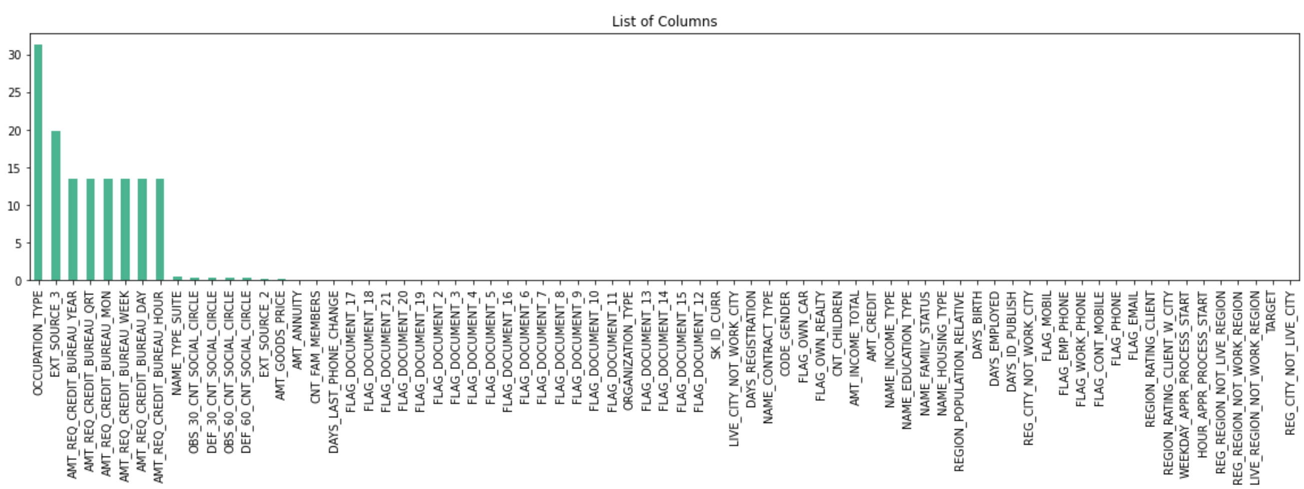
• Determine next steps: Based on the previous three tasks, determine whether to proceed to deployment, iterate further, or initiate new projects.

DATA PRE-PROCESSING

Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. When creating a machine learning project, it is not always a case that we come across clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put it in a formatted way. So for this, we use a data pre-processing task.
Real-world data generally contains noises, missing values, and maybe in an unusable format that cannot be directly used for machine learning models. Data pre-processing is a required task for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

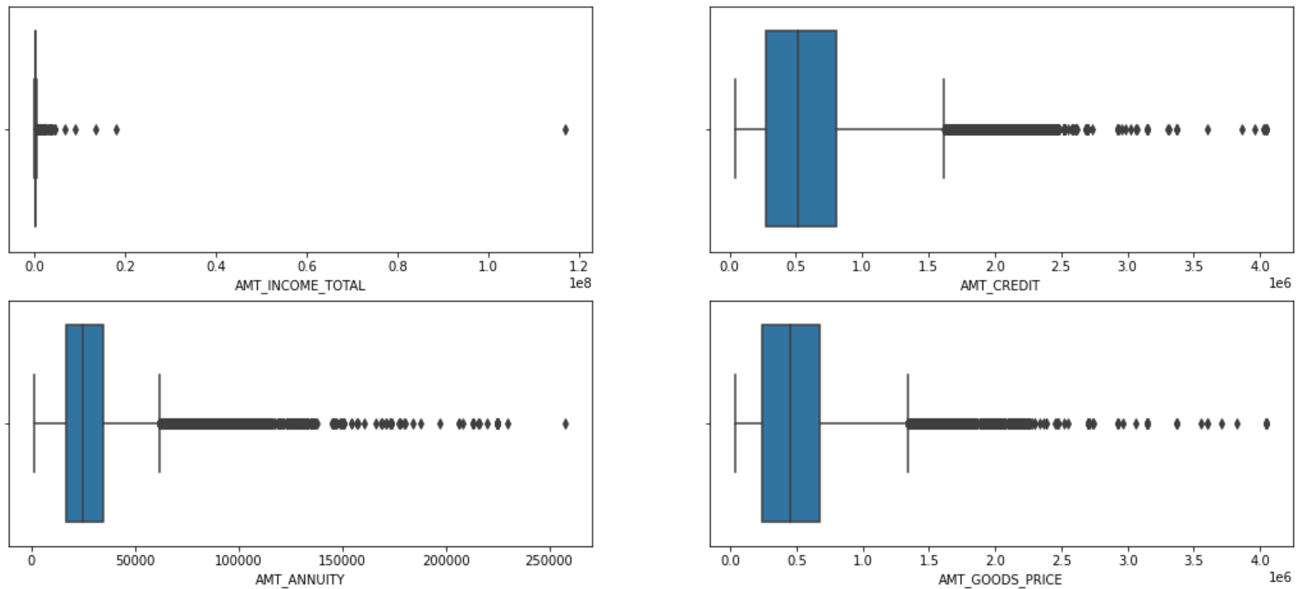
The data consists of 307511 rows and 122 columns. Out of these, we have 15 categorical columns and the rest as numerical.

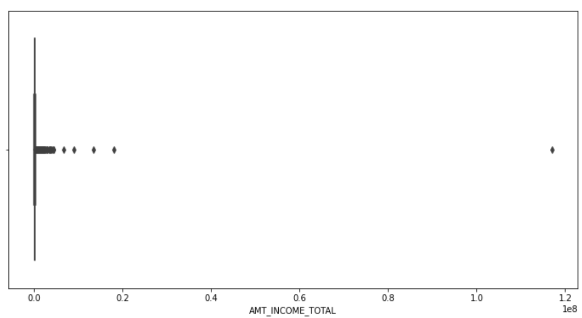
MISSING VALUE TREATMENT:



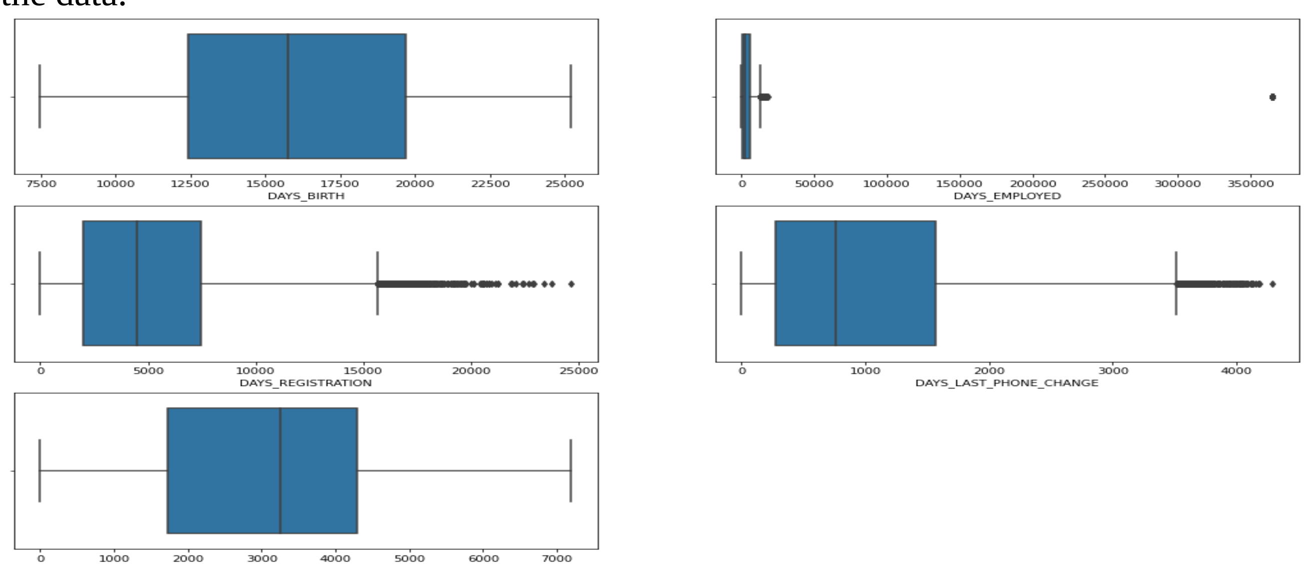
- Variable after removing the variable with a large number of missing values
Dropping the variables because after the many trail and errors these variables are affecting the modeling not being able to impute because of no business expertise.

OUTLIER TREATMENT:





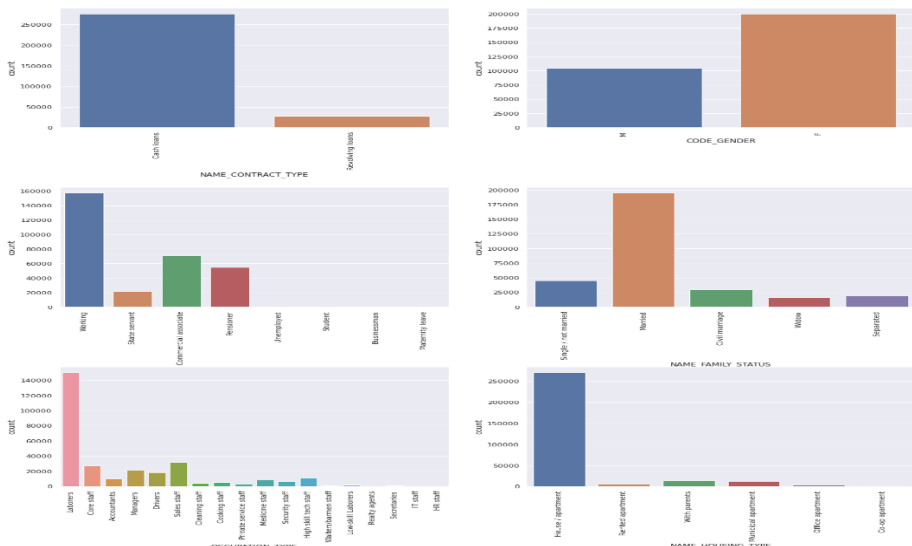
The above variable has outliers but in real-life, these data point is important in analyzing the data.



EXPLORATORY DATA ANALYSIS:

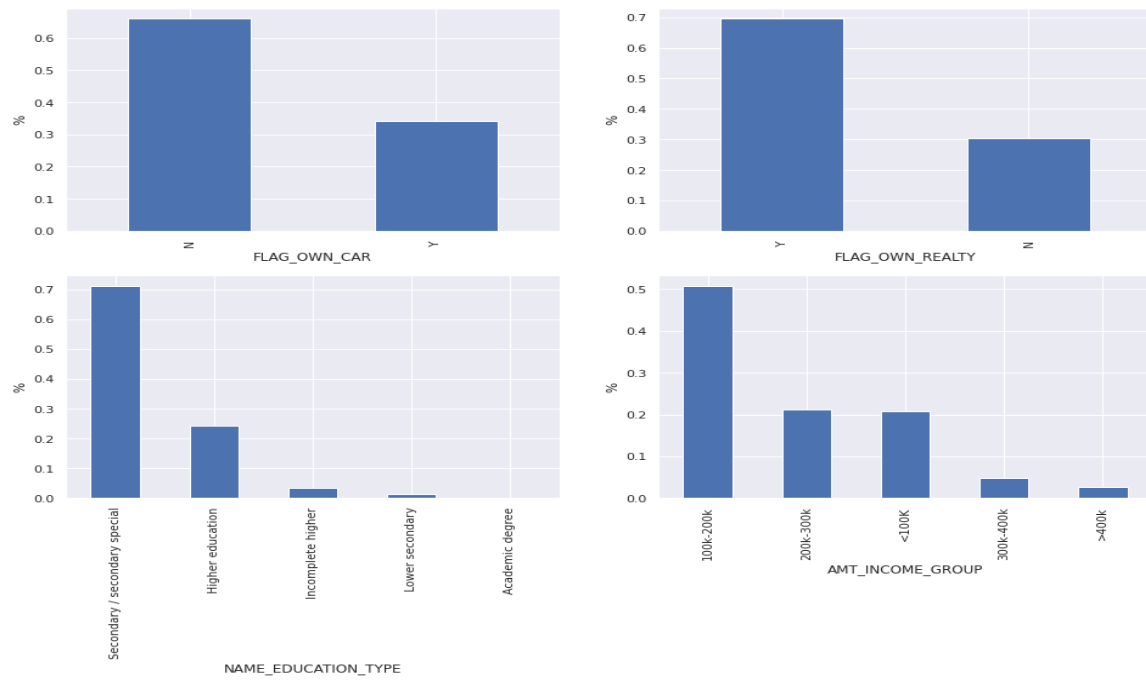
- UNIVARIATE ANALYSIS:

Univariate analysis is an analysis of a single variable. The univariate analysis explores each variable in a dataset.it looks at the range of the values as well as the central tendency of the values



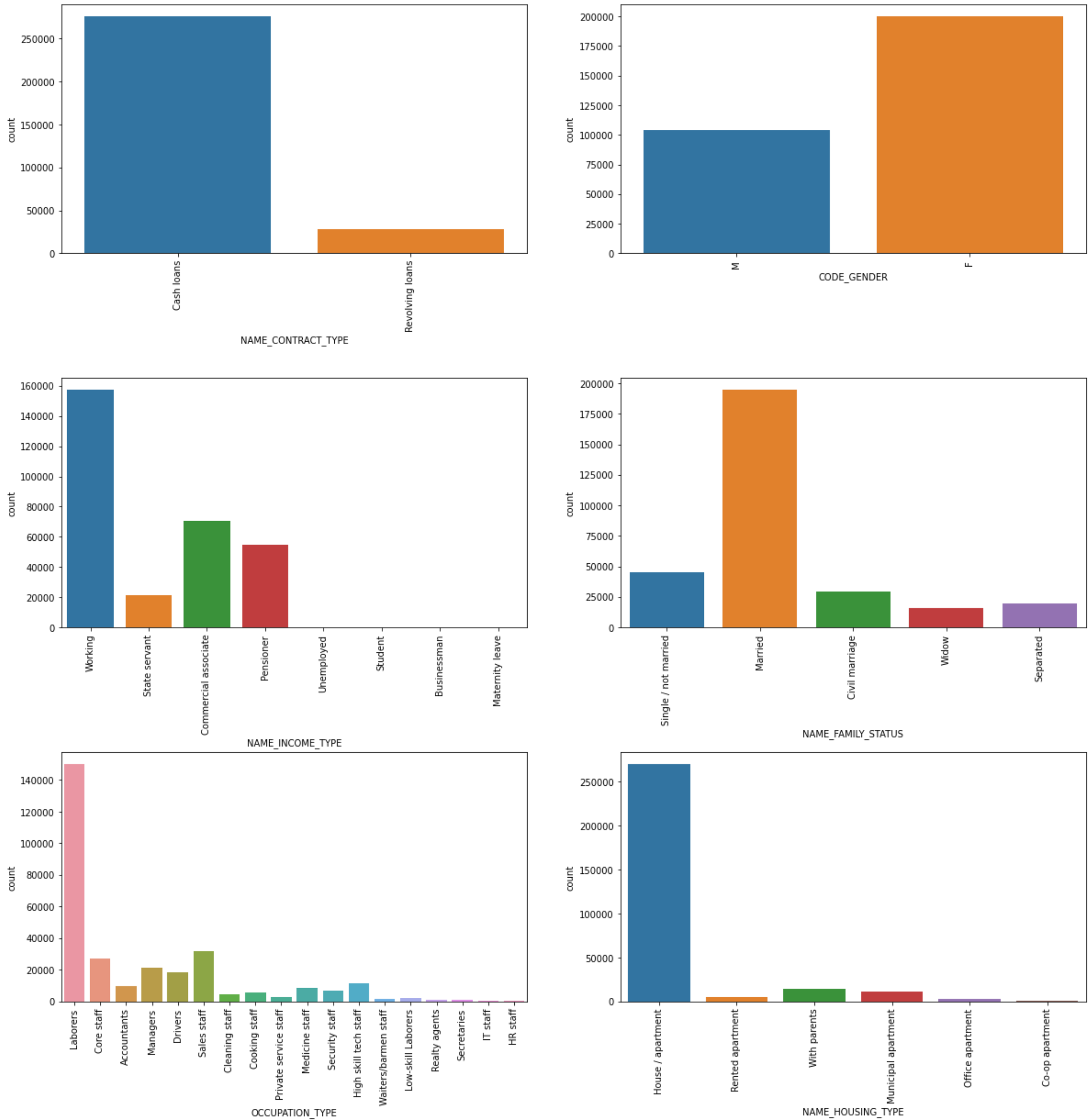
<u>Inference from the graphs:</u>

- Revolving loan* type is taken by very few clients
- The number of female clients is almost twice the male
- Number of loans taken by working professionals is more followed by commercial associate
- Higher number of loans are taken by married clients
- Laborers have more chances of taking loan
- Most of the clients live in their own house/apartment



Inference on the above graphs:

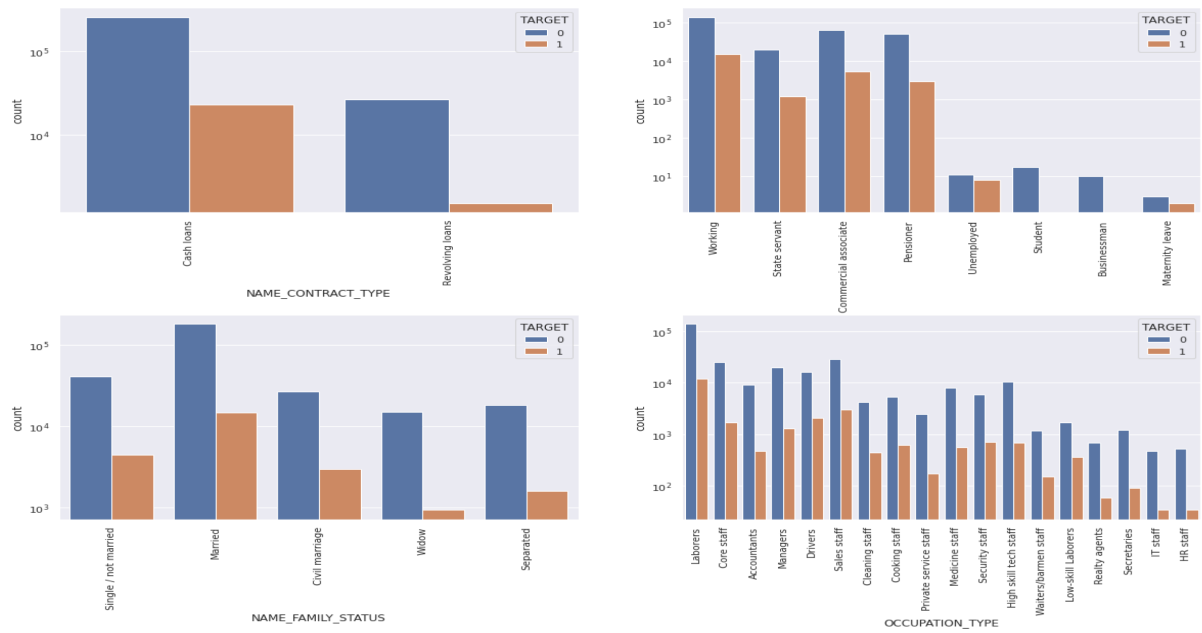
- More than 60% of the clients don't have a car
- Around 70% of them own a property
- Most of the clients have done only secondary education
- Income group of 100000 - 200000 are highest among the category who have taken the loans



<u>Inference from the graphs:</u>

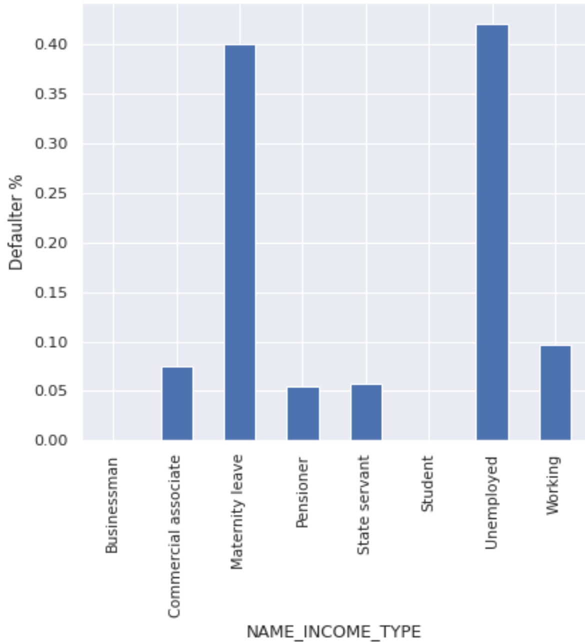
- Revolving loan* type are taken by very few clients
- The number of female clients are almost twice the male
- Number of loans taken by working professionals is more followed by commercial associate
- Higher number of loans are taken by married clients
- Laborers have more chances of taking loan

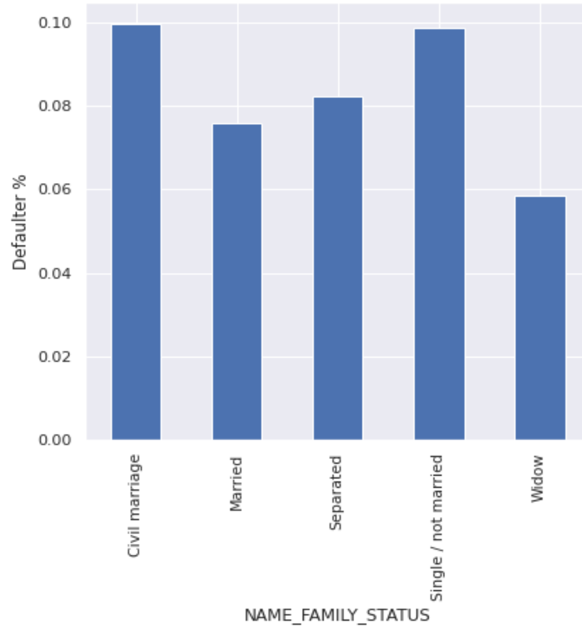
BIVARIATE ANALYSIS:



<u>Inference from the above graph:</u>

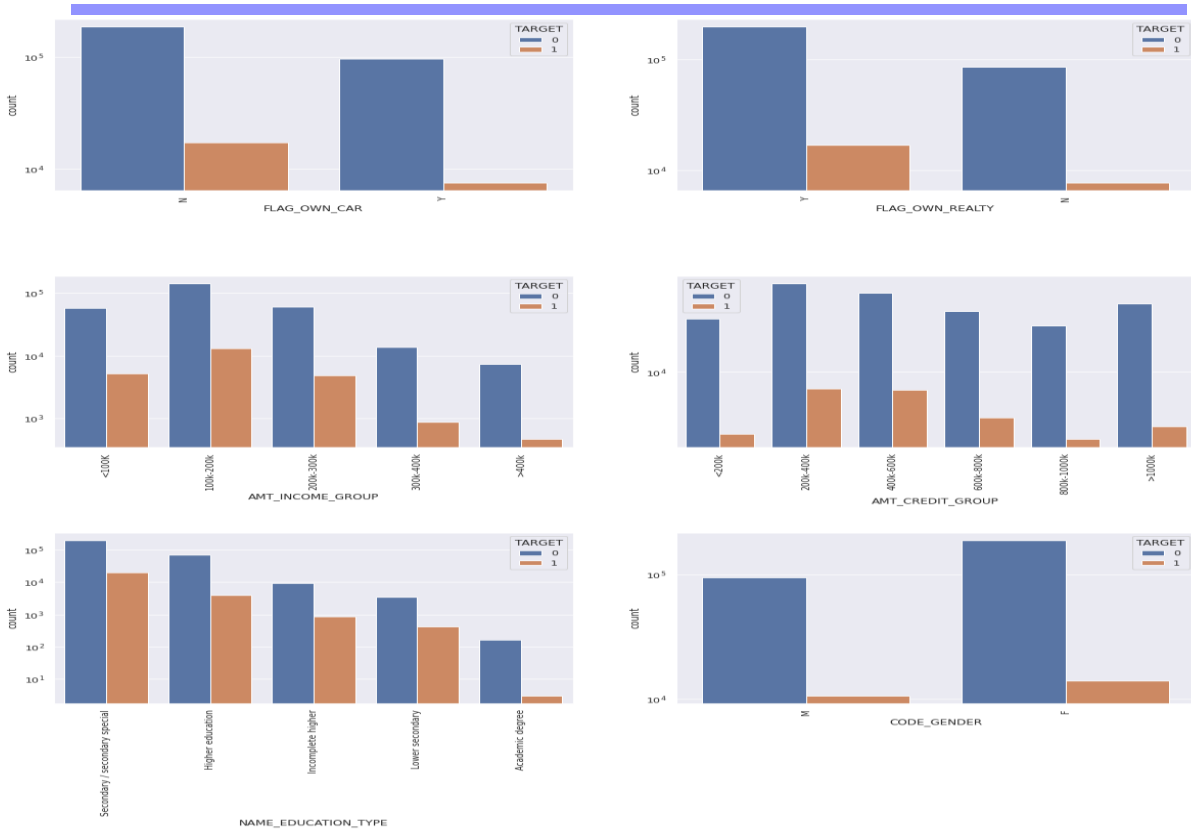
- NAME_CONTRACT_TYPE: Revolving loan type has less number of defaulters.
- NAME_INCOME_TYPE: Businessman and students groups have higher chances of repaying the loans. Chances that an unemployed become a defaulter is more
- NAME_FAMILY_STATUS: Widow category has fewer chances of becoming a defaulter whereas married has high chances. The civil marriage category have fewer defaulters compared to the single
- OCCUPATION_TYPE: IT staff, HR staff, and Realty agents have fewer defaulters whereas sales staff, laborers, and drivers have more defaulters





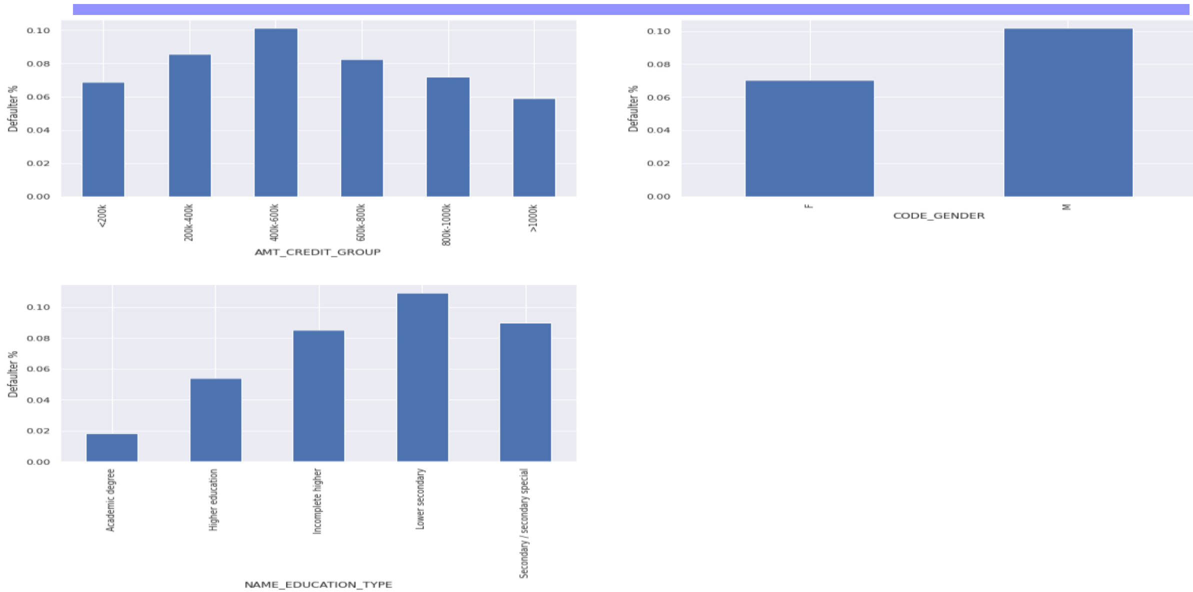
Inference of the above graphs:

- NAME_INCOME_TYPE: Around 40% of the clients who are unemployed or on maternity leave are defaulters
- NAME_FAMILY_STATUS: 10% of clients with civil marriage or single are defaulters
- Example; Out of 100 unemployed clients, around 40 clients fail to pay the loan



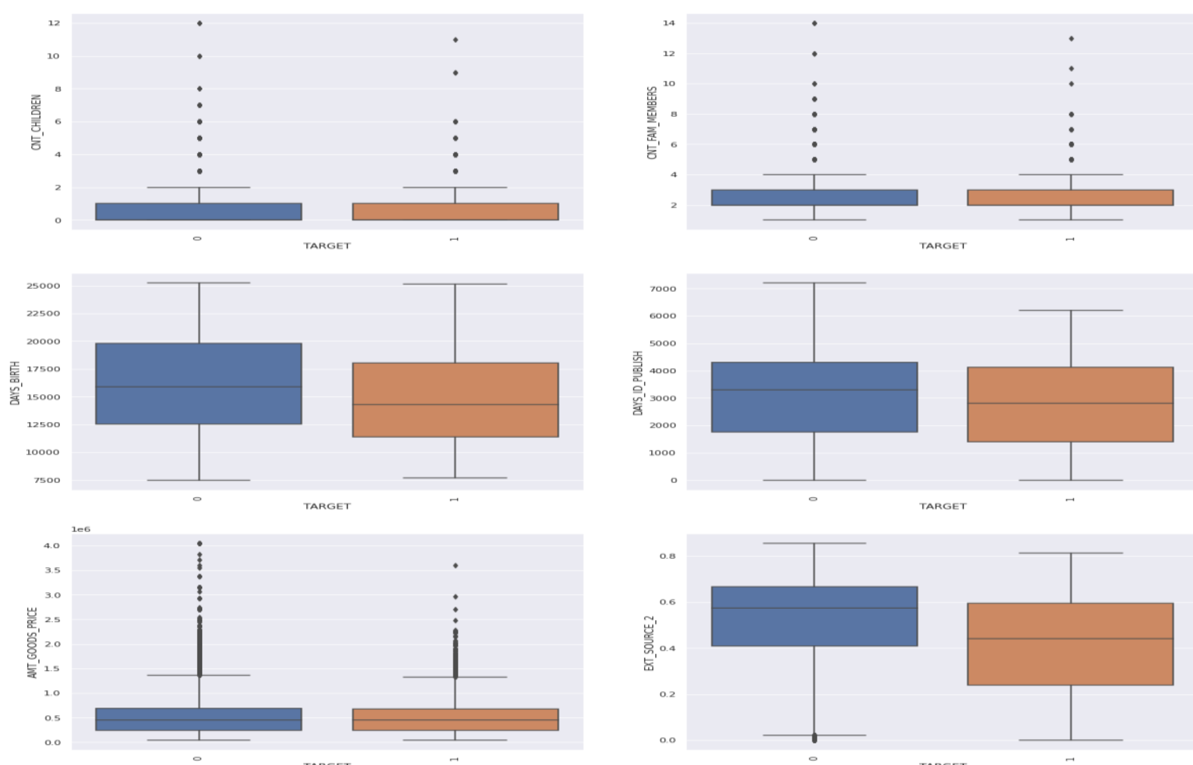
<u>Inference from the above graphs:</u>

- FLAG_OWN_CAR: Clients having a car have comparatively less number of defaulters
- FLAG_OWN_REALTY: Clients with no property have lesser chance to become defaulters
- AMT_INCOME_GROUP: Income group of 1,00,000-2,00,000 has higher number of defaulters whereas >4,00,000 has least
- AMT_CREDIT_GROUP: Credit Groups 2,00,000-4,00,000 and 4,00,000-6,00,000 have higher number of defaulters whereas 8,00,000-10,00,000 have less number of defaulters
- NAME_EDUCATION_TYPE: Academic degree group has lowest number of defaulters



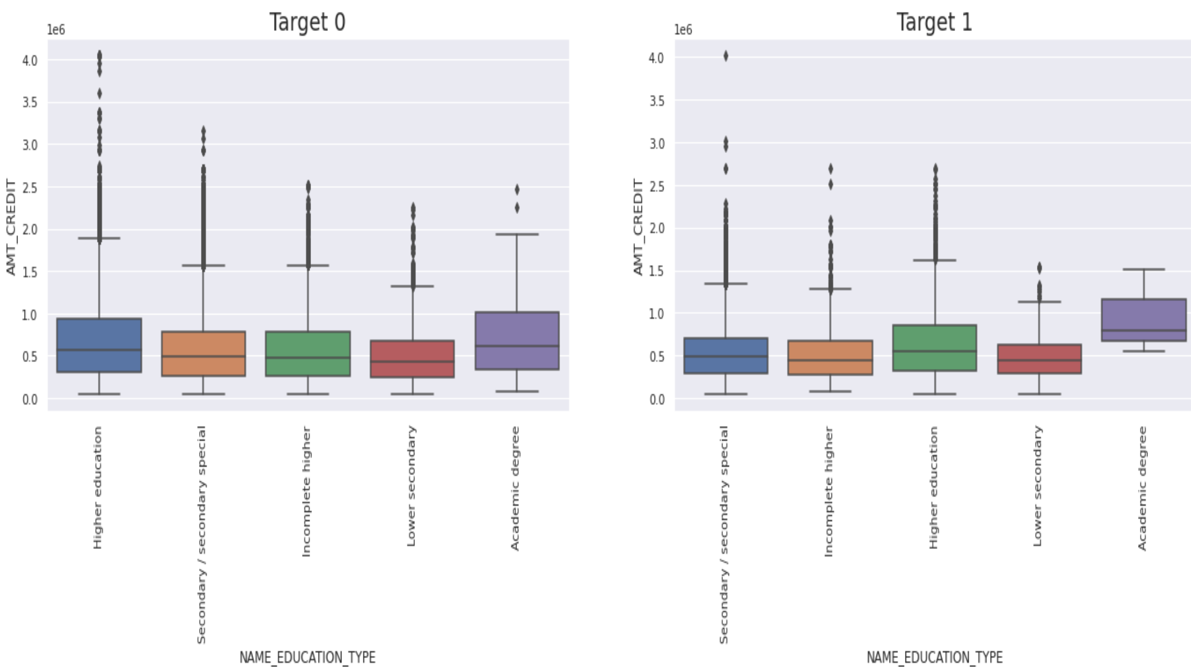
<u>Inference from the above graphs:</u>

- AMT_CREDIT_GROUP: 10% of the clients who have taken loan amounts in the range of 4,00,000-5,00,000 became defaulters
- CODE_GENDER: 10% of Male clients become defaulters
- NAME_EDUCATION_TYPE: More than 10% of the clients with lower secondary education become defaulters
- Explanation: Example- Out of 100 male clients, 10 males fail to repay the loan. The same follows for the AMT_CREDIT_GROUP and NAME_EDUCATION_TYPE

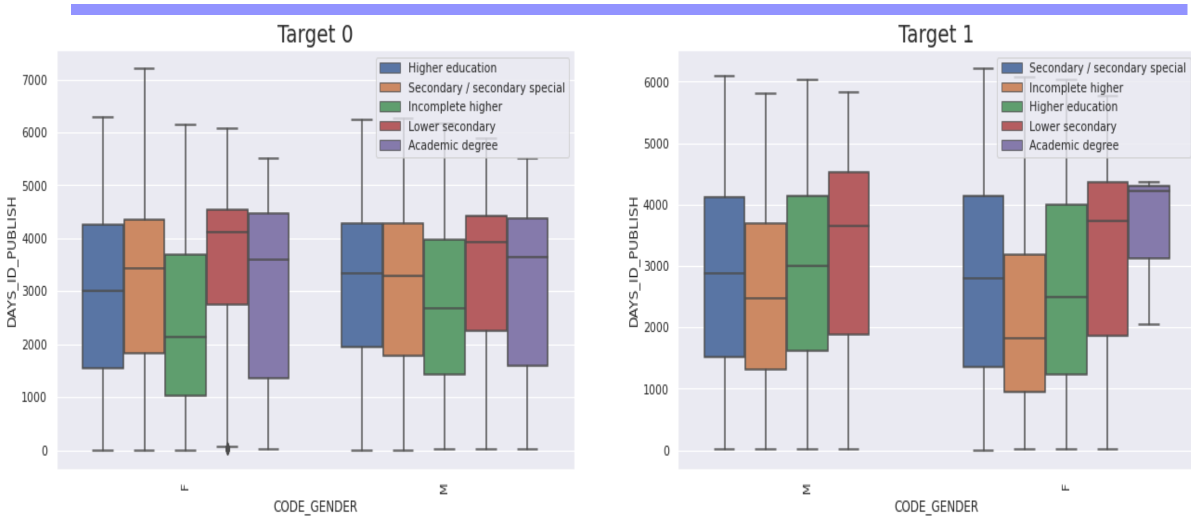


<u>Inference on the above graphs:</u>

- DAYS_BIRTH: If the days_birth (age of client) is more than 15000 (around 40 years of age) then the chances of becoming a defaulter are less. Or 40+ age clients are most likely to repay the loan
- DAYS_ID_PUBLISH: If the change of the ID is done a few days before the application chances are that he becomes a defaulter
- AMT_GOODS_PRICE: Clients with loans taken above 35,00,000 goods price are fewer defaulters. (But this is not a strong trend because of the outliers)
- EXT_SOURCE_2: Low score indicates signs of the defaulter. A score above 0.55 have a high chance of repayment

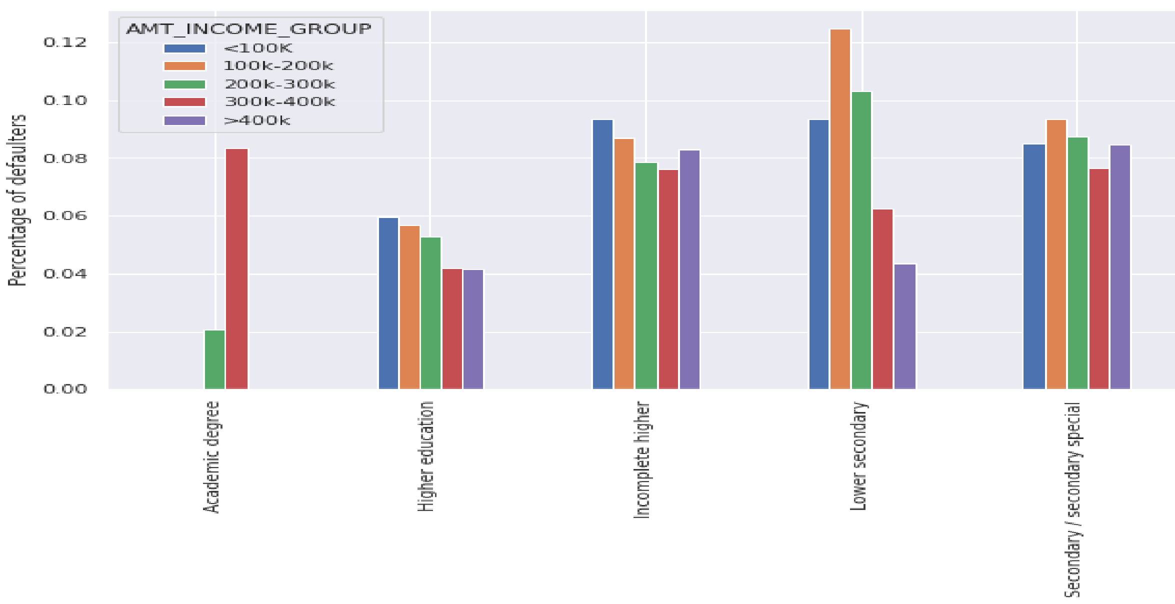


- **<u>Name education type with respect to target 0 and target 1</u>**
- NAME_EDUCATION_TYPE vs AMT_CREDIT: Clients with higher education and academic degree find it hard to pay the loan if the loan amount is above 5,00,000
- NAME_FAMILY_STATUS vs DAYS_BIRTH: If the client is single and has days birth (or age) below 12500 days (35 years age) has more chances of becoming a defaulter.



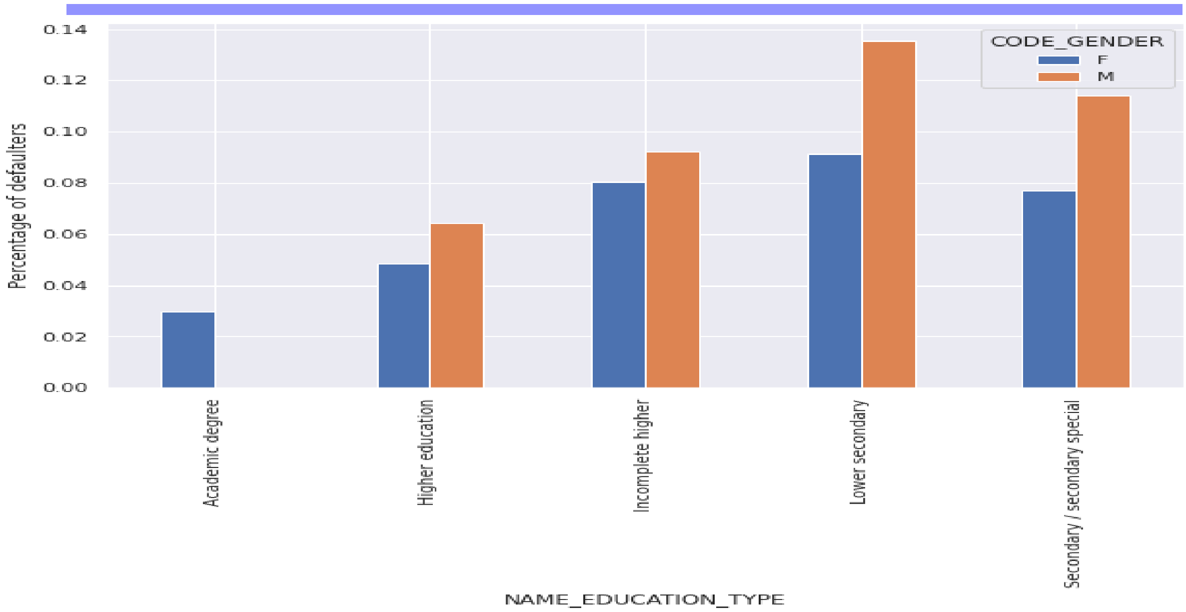
- CODE_GENDER vs DAYS_ID_PUBLISH: Female clients with incomplete higer education and have changed their identity document below 2000 days before the application have high number of defaulters

MULTIVARIATE ANALYSIS:



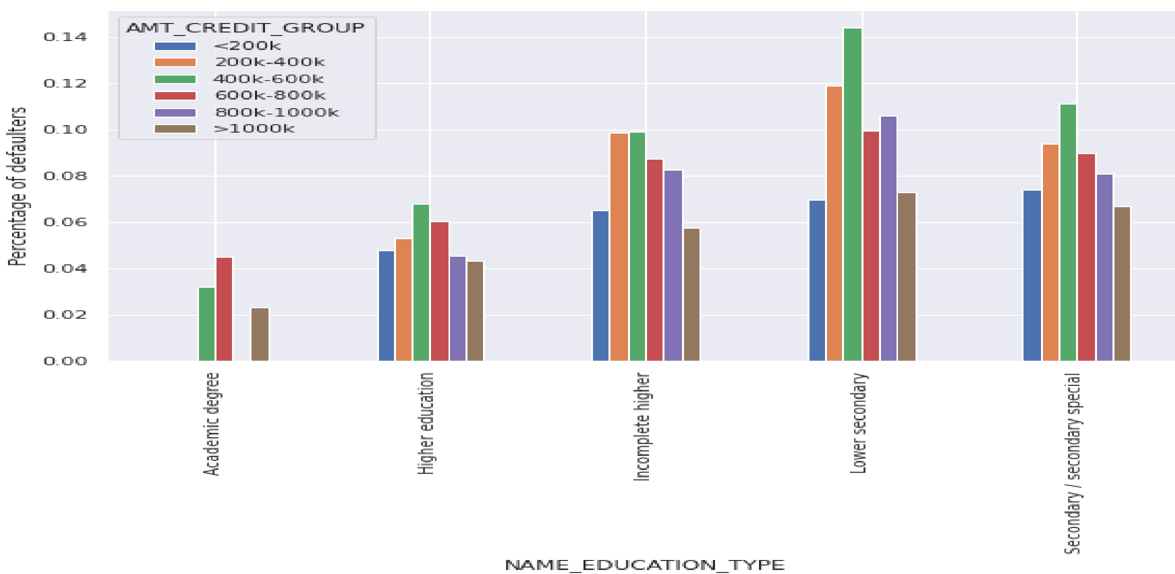
<u>Inference on the above graphs:</u>

- AMT_INCOME_TOTAL vs NAME_EDUCATION_TYPE vs Target:
- Clients with academic degree in the income range of 3,00,000-4,00,000 have higher defaulters. Clients with lower scondary education in the income range of 1,00,000-2,00,000 have higher defaulters.



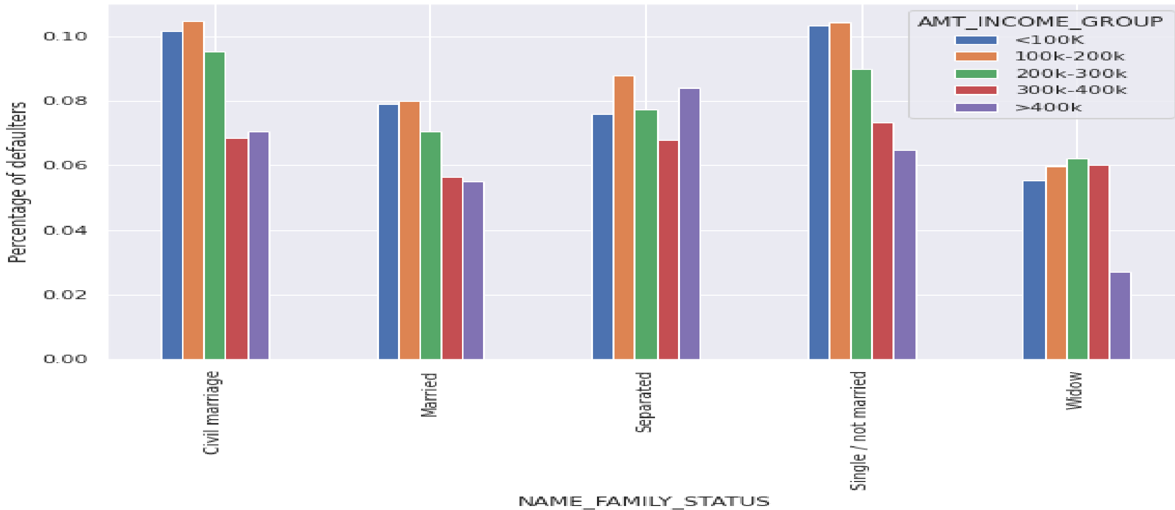
Inference on the above graphs

- ➢ NAME_EDUCATION_TYPE vs CODE_GENDER vs TARGET:
- ➢ Male and Female with lower secondary have high percentage of defaulters whereas Male and Female with academic degree the lowest



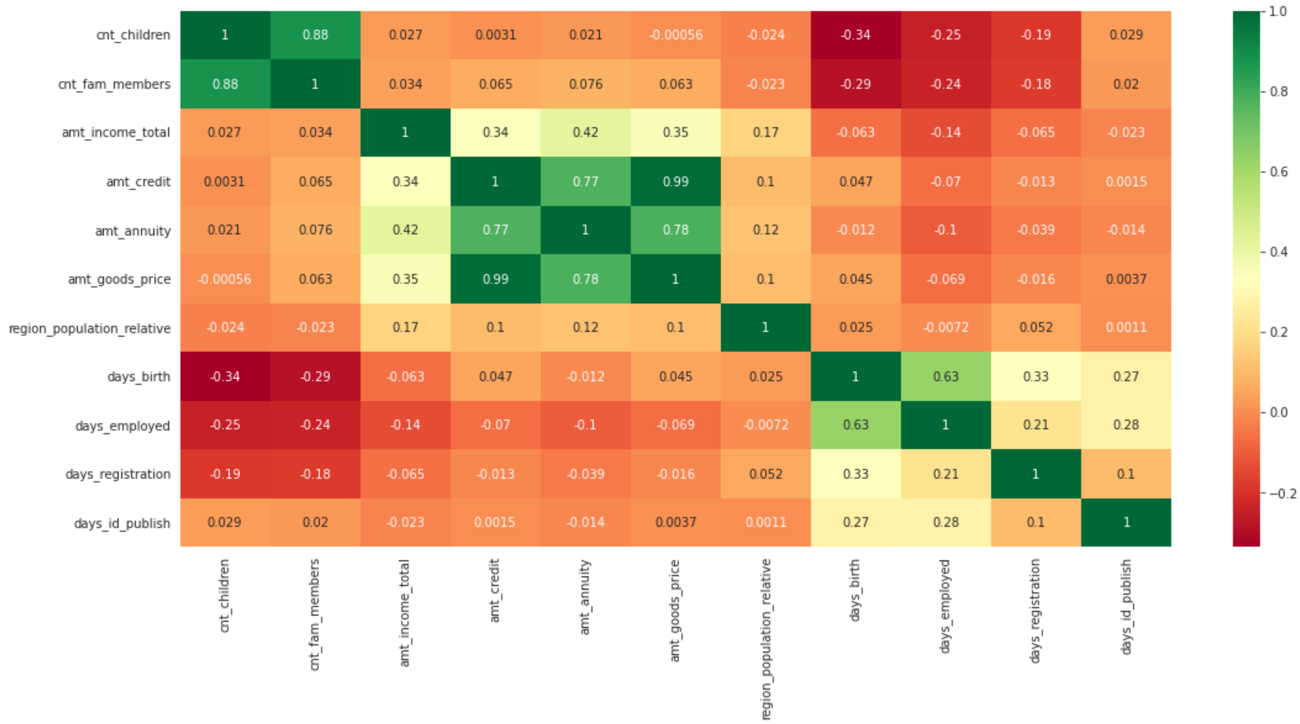
Inference on the above graphs:

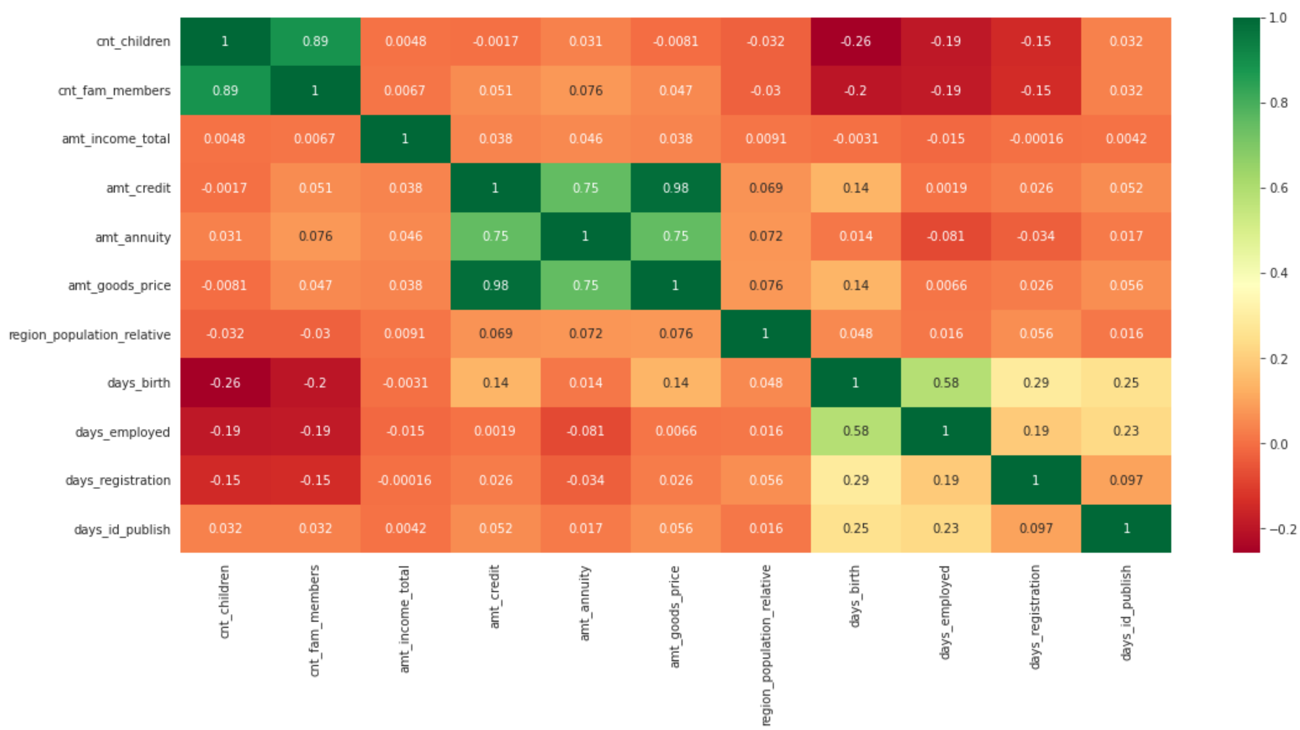
- ➢ NAME_EDUCATION_TYPE vs AMT_CREDIT_GROUP vs Target:
- ➢ More than 14% of the clients with lower secondary education and have taken loan amount between 4,00,000-6,00,000 are defaulters

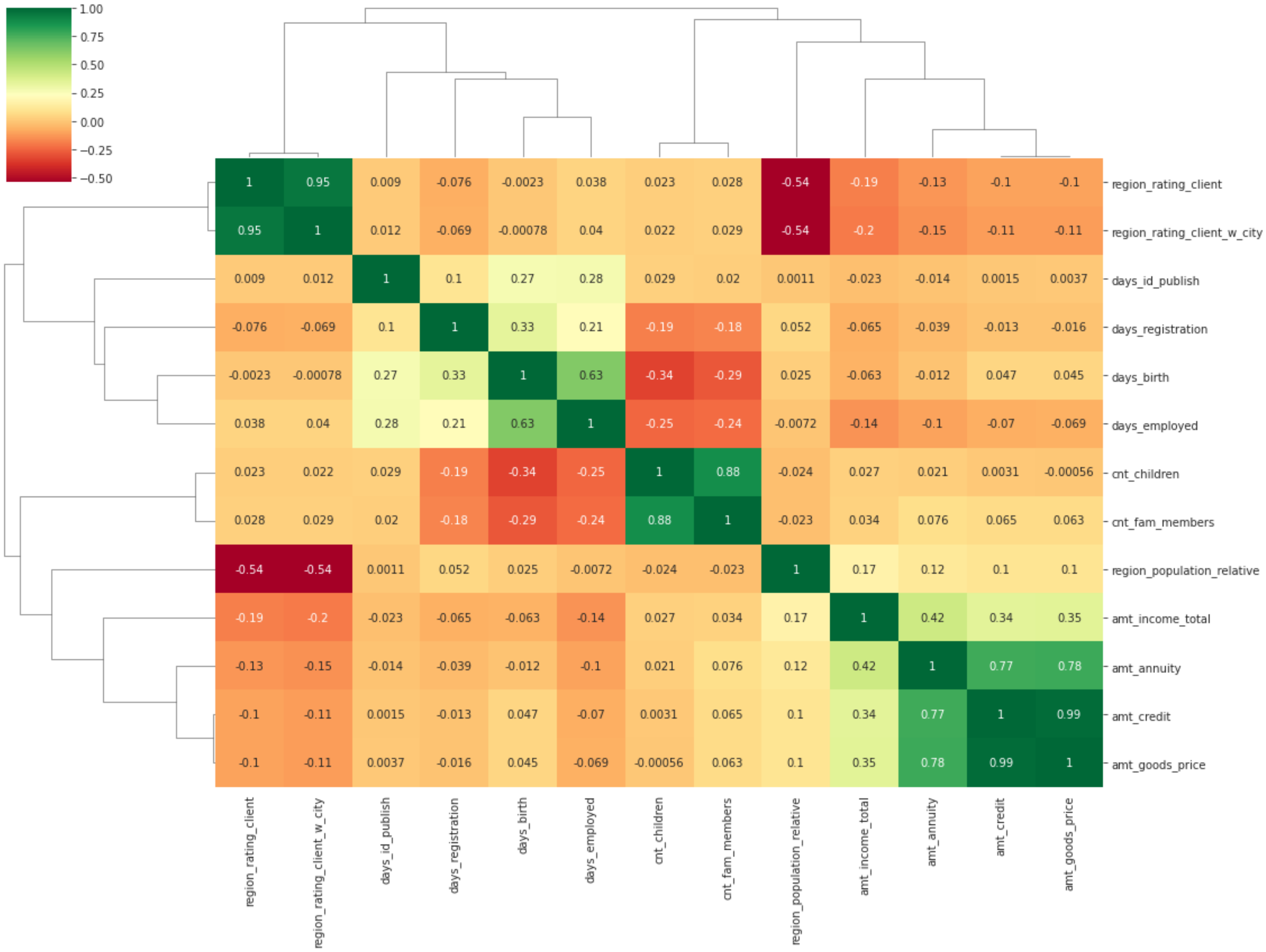


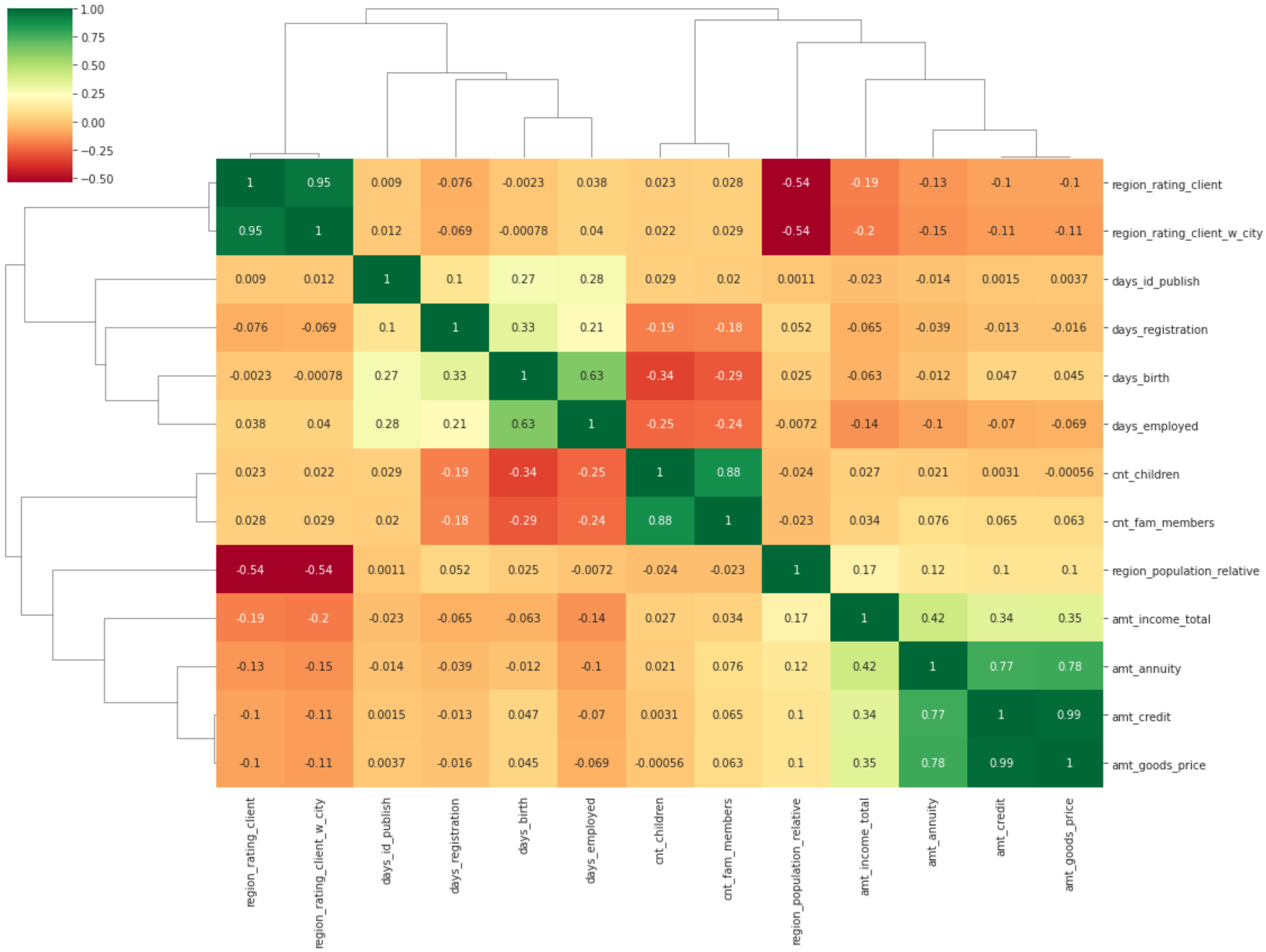
- NAME_FAMILY_STATUS vs AMT_INCOME_GROUP vs target: Clients with civil marriage or who is single find it hard to repay the loan

<u>Correlation Matrix for different variables</u> :






1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
region_rating_client: 1 0.95 0.009 -0.076 -0.0023 0.038 0.023 0.028 -0.54 -0.19 -0.13 -0.1 -0.1
region_rating_client_w_city: 0.95 1 0.012 -0.069 -0.00078 0.04 0.022 0.029 -0.54 -0.2 -0.15 -0.11 -0.11
days_id_publish: 0.009 0.012 1 0.1 0.27 0.28 0.029 0.02 0.0011 -0.023 -0.014 0.0015 0.0037
days_registration: -0.076 -0.069 0.1 1 0.33 0.21 -0.19 -0.18 0.052 -0.065 -0.039 -0.013 -0.016
days_birth: -0.0023 -0.00078 0.27 0.33 1 0.63 -0.34 -0.29 0.025 -0.063 -0.012 0.047 0.045
days_employed: 0.038 0.04 0.28 0.21 0.63 1 -0.25 -0.24 -0.0072 -0.14 -0.1 -0.07 -0.069
cnt_children: 0.023 0.022 0.029 -0.19 -0.34 -0.25 1 0.88 -0.024 0.027 0.021 0.0031 -0.00056
cnt_fam_members: 0.028 0.029 0.02 -0.18 -0.29 -0.24 0.88 1 -0.023 0.034 0.076 0.065 0.063
region_population_relative: -0.54 -0.54 0.0011 0.052 0.025 -0.0072 -0.024 -0.023 1 0.17 0.12 0.1 0.1
amt_income_total: -0.19 -0.2 -0.023 -0.065 -0.063 -0.14 0.027 0.034 0.17 1 0.42 0.34 0.35
amt_annuity: -0.13 -0.15 -0.014 -0.039 -0.012 -0.1 0.021 0.076 0.12 0.42 1 0.77 0.78
amt_credit: -0.1 -0.11 0.0015 -0.013 0.047 -0.07 0.0031 0.065 0.1 0.34 0.77 1 0.99
amt_goods_price: -0.1 -0.11 0.0037 -0.016 0.045 -0.069 -0.00056 0.063 0.1 0.35 0.78 0.99 1
region_rating_client
region_rating_client_w_city
days_id_publish
days_registration
days_birth
days_employed
cnt_children
cnt_fam_members
region_population_relative
amt_income_total
amt_annuity
amt_credit
amt_goods_price



- From the heatmap we can see that some variables are having a high and partial correlation with other variables
- Dropping the variables to reduce the multicollinearity effect in the modeling

FINAL COLUMNS

0	TARGET	304526 non-null	int64
1	NAME_CONTRACT_TYPE	304526 non-null	uint8
2	CODE_GENDER	304526 non-null	uint8
3	FLAG_OWN_CAR	304526 non-null	uint8
4	FLAG_OWN_REALTY	304526 non-null	uint8
5	CNT_CHILDREN	304526 non-null	int64
6	AMT_INCOME_TOTAL	304526 non-null	float64
7	AMT_CREDIT	304526 non-null	float64
8	AMT_ANNUITY	304526 non-null	float64
9	AMT_GOODS_PRICE	304526 non-null	float64
10	NAME_TYPE_SUITE	304526 non-null	float64
11	NAME_INCOME_TYPE	304526 non-null	float64
12	NAME_EDUCATION_TYPE	304526 non-null	float64
13	NAME_FAMILY_STATUS	304526 non-null	float64
14	NAME_HOUSING_TYPE	304526 non-null	float64
15	REGION_POPULATION_RELATIVE	304526 non-null	float64
16	DAYS_BIRTH	304526 non-null	int64
17	DAYS_EMPLOYED	304526 non-null	int64
18	DAYS_REGISTRATION	304526 non-null	int64
19	DAYS_ID_PUBLISH	304526 non-null	int64
20	FLAG_MOBIL	304526 non-null	int64
21	FLAG_EMP_PHONE	304526 non-null	int64
22	FLAG_WORK_PHONE	304526 non-null	int64
23	FLAG_CONT_MOBILE	304526 non-null	int64
24	FLAG_PHONE	304526 non-null	int64
25	FLAG_EMAIL	304526 non-null	int64
26	OCCUPATION_TYPE	304526 non-null	float64
27	CNT_FAM_MEMBERS	304526 non-null	float64
28	REGION_RATING_CLIENT	304526 non-null	int64
29	REGION_RATING_CLIENT_W_CITY	304526 non-null	int64
30	WEEKDAY_APPR_PROCESS_START	304526 non-null	float64
31	HOUR_APPR_PROCESS_START	304526 non-null	int64
32	REG_REGION_NOT_LIVE_REGION	304526 non-null	int64
33	REG_REGION_NOT_WORK_REGION	304526 non-null	int64
34	LIVE_REGION_NOT_WORK_REGION	304526 non-null	int64
35	REG_CITY_NOT_LIVE_CITY	304526 non-null	int64
36	REG_CITY_NOT_WORK_CITY	304526 non-null	int64
37	LIVE_CITY_NOT_WORK_CITY	304526 non-null	int64
38	EXT_SOURCE_2	304526 non-null	float64
39	EXT_SOURCE_3	304526 non-null	float64
40	OBS_30_CNT_SOCIAL_CIRCLE	304526 non-null	float64
41	DEF_30_CNT_SOCIAL_CIRCLE	304526 non-null	float64
42	OBS_60_CNT_SOCIAL_CIRCLE	304526 non-null	float64
43	DEF_60_CNT_SOCIAL_CIRCLE	304526 non-null	float64
44	DAYS_LAST_PHONE_CHANGE	304526 non-null	int64
45	AMT_REQ_CREDIT_BUREAU_HOUR	304526 non-null	float64
46	AMT_REQ_CREDIT_BUREAU_DAY	304526 non-null	float64
47	AMT_REQ_CREDIT_BUREAU_WEEK	304526 non-null	float64
48	AMT_REQ_CREDIT_BUREAU_MON	304526 non-null	float64
49	AMT_REQ_CREDIT_BUREAU_QRT	304526 non-null	float64
50	AMT_REQ_CREDIT_BUREAU_YEAR	304526 non-null	float64
51	AMT_INCOME_GROUP	304526 non-null	float64
52	AMT_CREDIT_GROUP	304526 non-null	float64

- the list of columns is obtained after doing all EDA like outlier treatment, missing value treatment, removing the redundant columns and multicollinearity treatment in the dataset
- the final dataset contains:

The shape of the dataset –

Columns - 52
Rows - 304526

Type of data – supervised classification data

BASE MODEL:

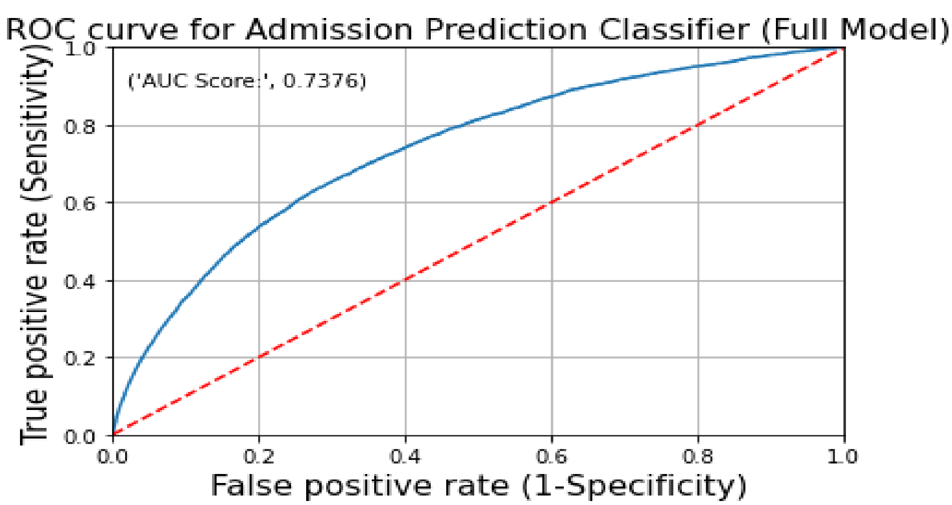
Logistic Regression Model:

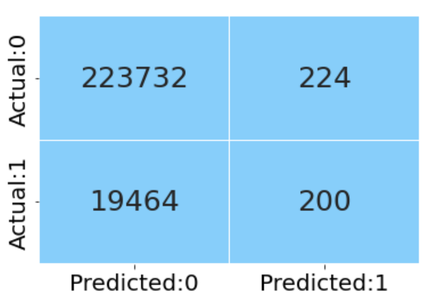
We have selected Logistic Regression as our base model. For this, we have encoded all the categorical variables using Encoder and have kept the numerical columns as it is.

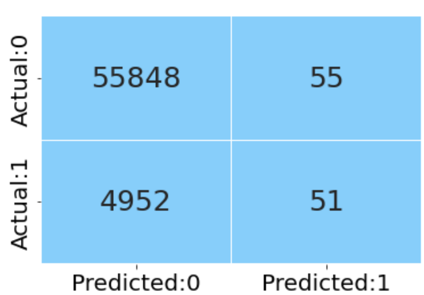
Classification Report:

	precision	recall	f1-score	support
0	0.92	1.00	0.96	223956
1	0.47	0.01	0.02	19664
accuracy			0.92	243620
macro avg	0.70	0.50	0.49	243620
weighted avg	0.88	0.92	0.88	243620

	precision	recall	f1-score	support
0	0.92	1.00	0.96	55903
1	0.48	0.01	0.02	5003
accuracy			0.92	60906
macro avg	0.70	0.50	0.49	60906
weighted avg	0.88	0.92	0.88	60906







The false negative is higher for the base model. The objective of our model prediction is to increase the recall score.

The model is not able to find the actual defaulters, it predicted 19464 as non-defaulters.

MODEL BUILDING AND METHODS:

From EDA, we observed the presence of high cardinality is certain categorical variables. In order to build models, we need to use appropriate encoding techniques to address this issue. Also, for building better models, we need to transform the numerical variables

FEATURE ENGINEERING:

- **Dummy Encoding** – variable has only two unique values so the dummy encoding is the best way to encode the variables

- **Ordinal encoding -** variable has more than two unique values and has the order in the variables so we use the ordinal encoding

-
- **Label encoding -** variable has more than two unique values and no has the order in the variables so we use the label encoding.

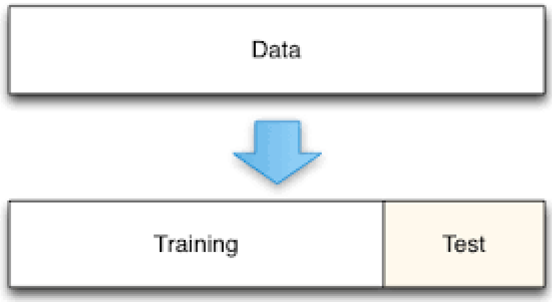
FEATURE SCALING:

Min-Max scalar (Numerical variables) – the min-max normalization is used for the numerical variables which will convert the values between 0 and 1

MODEL BUILDING:

Step by step approach for model building: -

1. After performing encoding for the categorical features and transforming the numerical variables, we split the data into train data and test data. Model data uses train data to learn whereas test data is used to evaluate or validate the trained model.



2. The baseline model which we built was a logistic regression with performing any transformation on numerical variables and using an encoder for categorical variables.
3. Next, we build non-linear models such as Decision Tree, Random Forest, Gradient Boost, Ada Boost, and XG Boost Classifier. For these models, we perform hyper-parameter tuning. Also, since there is a presence of a moderate amount of class imbalance, we perform oversampling.

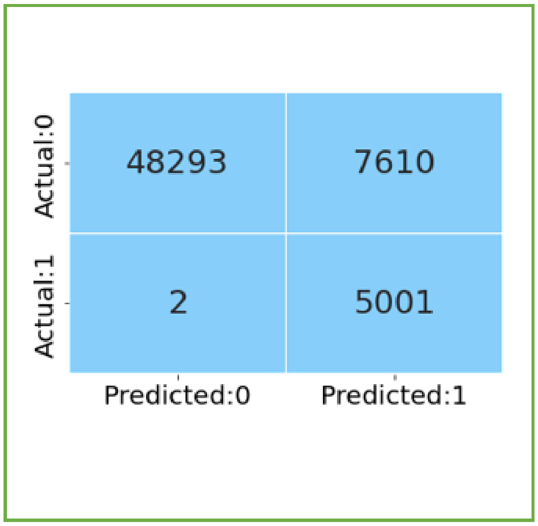
4. From these models, we do not achieve the desired amount of accuracy, precision, and recall even though we achieve a moderate level of accuracy for the model, we get low precision and recall values. The KNN model gives the good accuracy

5. In order to further improve the model, we perform over-sampling to address the presence of the moderate amount of class imbalance and again build the model. From here we observe that we obtain a better model with over-sampling. Even though over sampling leads to an increase in the size of the dataset, it contributes towards realistic data

FINAL MODEL – KNN

	precision	recall	f1-score	support
0.0	0.96	0.86	0.93	223956
1.0	0.17	$\boxed{1.00}$	0.57	19664
accuracy			$\boxed{0.88}$	243620
macro avg	0.57	0.93	0.75	243620
weighted avg	0.90	0.88	0.90	243620

	precision	recall	f1-score	support
0.0	1.00	0.86	0.93	55903
1.0	0.40	$\boxed{1.00}$	0.57	5003
accuracy			$\boxed{0.88}$	60906
macro avg	0.70	0.93	0.75	60906
weighted avg	0.95	0.88	0.90	60906



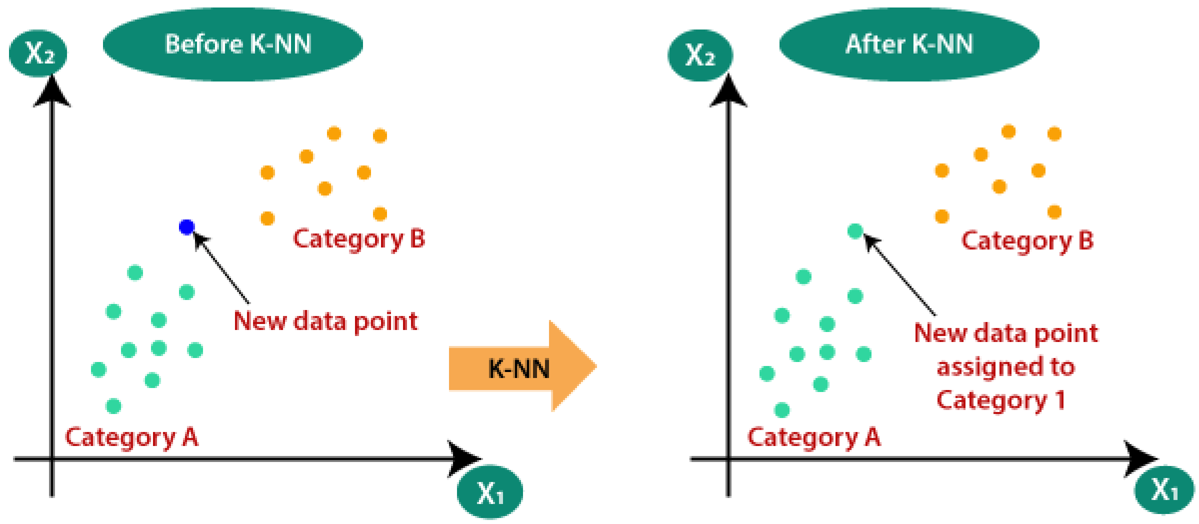
MODEL UNDERSTANDING:

K-Nearest Neighbour (KNN) Algorithm for Machine Learning.

- K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on the Supervised Learning technique.
- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
- K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for Classification problems.
- K-NN is a **non-parametric algorithm**, which means it does not make any assumptions on underlying data.

<u>Why do we need a K-NN Algorithm?</u>

Suppose there are two categories, i.e., Category A and Category B, and we have a new data point x1, so this data point will lie in which of these categories. To solve this type of problem, we need a K-NN algorithm. With the help of K-NN, we can easily identify the category or class of a particular dataset. Consider the below diagram:

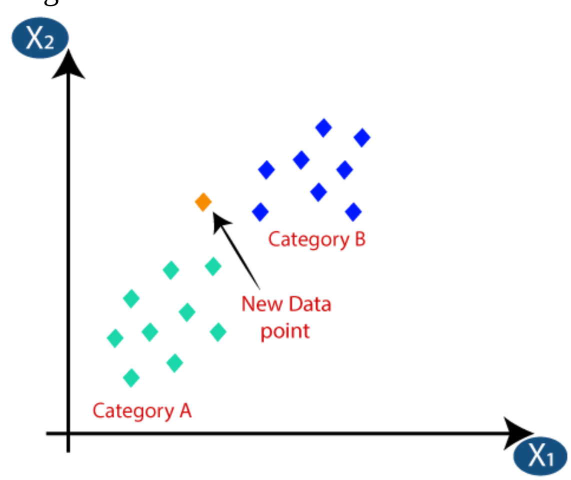


How does K-NN work?

The K-NN working can be explained on the basis of the below algorithm:

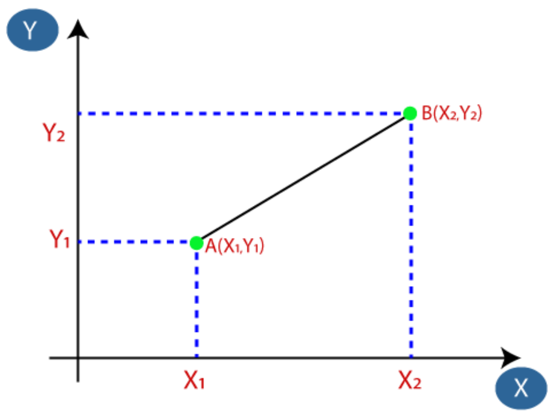
- **Step-1:** Select the number K of the neighbours
- **Step-2:** Calculate the Euclidean distance of **K number of neighbours**
- **Step-3:** Take the K nearest neighbours as per the calculated Euclidean distance.
- **Step-4:** Among these k neighbours, count the number of the data points in each category.
- **Step-5:** Assign the new data points to that category for which the number of the neighbour is maximum.
- **Step-6:** Our model is ready.

Suppose we have a new data point and we need to put it in the required category. Consider the below image:



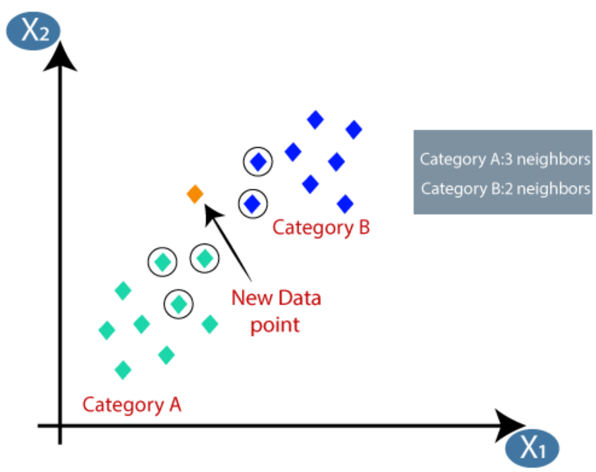
- Firstly, we will choose the number of neighbors, so we will choose the k=5.

- Next, we will calculate the **Euclidean distance** between the data points. The Euclidean distance is the distance between two points, which we have already studied in geometry. It can be calculated as:



Euclidean Distance between A_1 and B_2 = $\sqrt{(X_2-X_1)^2+(Y_2-Y_1)^2}$

- By calculating the Euclidean distance we got the nearest neighbors, as three nearest neighbors in category A and two nearest neighbors in category B. Consider the below image:



- As we can see the 3 nearest neighbours are from category A, hence this new data point must belong to category A.

<u>How to select the value of K in the K-NN Algorithm?</u>
Below are some points to remember while selecting the value of K in the K-NN algorithm:

- There is no particular way to determine the best value for "K", so we need to try some values to find the best out of them. The most preferred value for K is 5.
- A very low value for K such as K=1 or K=2, can be noisy and lead to the effects of outliers in the model.
- Large values for K are good, but it may find some difficulties.

<u>Advantages of KNN Algorithm:</u>

- It is simple to implement.
- It is robust to the noisy training data

- It can be more effective if the training data is large.

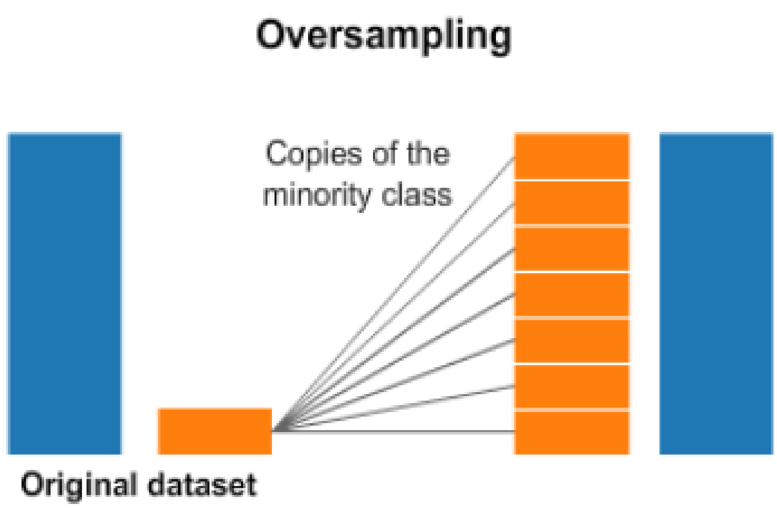
Disadvantages of KNN Algorithm:

- Always needs to determine the value of K which may be complex some time.
- The computation cost is high because of calculating the distance between the data points for all the training samples.

OVER-SAMPLING :

SMOTE(**Synthetic Minority Oversampling Technique**)

Random over-sampling involves randomly duplicating examples from the minority class and adding them to the training dataset. Examples from the training dataset are selected randomly with replacement. This means that examples from the minority class can be chosen and added to the new "more balanced" training 50 dataset multiple times; they are selected from the original training dataset, added to the new training dataset, and then returned or "replaced" in the original dataset, allowing them to be selected again. In some cases, seeking a balanced distribution for a severely imbalanced dataset can cause affected algorithms to overfit the minority class, leading to increased generalization error. The effect can be better performance on the training dataset, but worse performance on the holdout or test dataset.



Prominent Parameters:

Accuracy: Accuracy is the ratio of the total number of correct predictions and the total number of predictions.

$$Accuracy = \frac{True\ Positive}{True\ Postive + False\ Negative + False\ Positive + False\ Negative}.$$

Using accuracy as a defining metric for our model does make sense intuitively, but more often than not, it is always advisable to use Precision and Recall too. There might be other situations where our accuracy is very high, but our precision or recall is low.

Precision: It is the accuracy of positive predictions.

$$Precision = \frac{True\ Positive}{True\ Postive + False\ Positive}$$

That means, when the model predicts that a Customer will be Defaulter, it is correct around %precision times.

Recall:

It is the ratio of positive instance that are correctly detected. It is also called sensitivity.

$$Recall = \frac{True\ Positive}{True\ Postive + False\ Negative}$$

Hence, for all the Defaulted that were actually Defaulter, recall tells us how many of the models correctly identified as to be Defaulter

F1-score: F1-score is the Harmonic mean of the Precision and Recall.

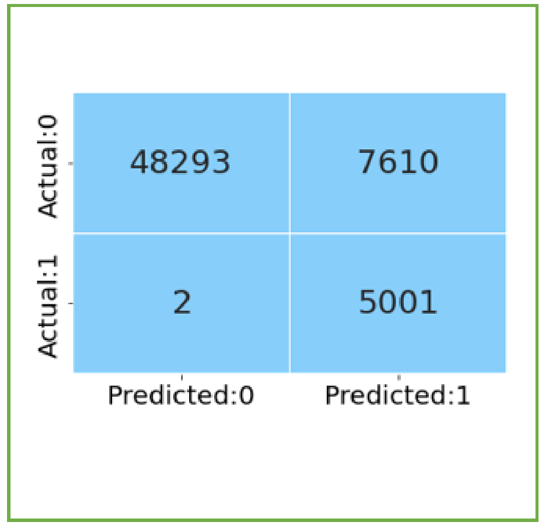
$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

Unfortunately, we can’t have both precision and recall high. If we increase precision, it will reduce recall and vice versa. This is called the precision/recall trade-off.

FINAL MODEL – KNN

	precision	recall	f1-score	support
0.0	0.96	0.86	0.93	223956
1.0	0.17	1.00	0.57	19664
accuracy			0.88	243620
macro avg	0.57	0.93	0.75	243620
weighted avg	0.90	0.88	0.90	243620

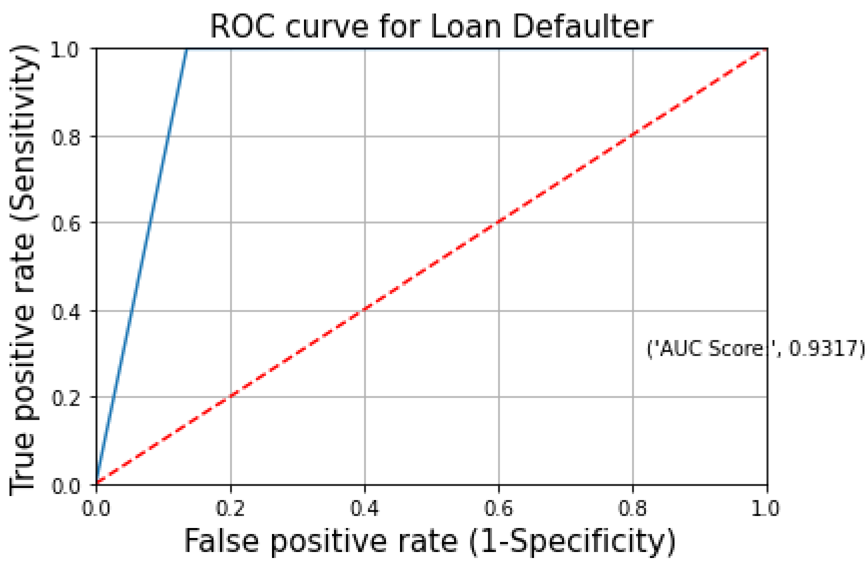
	precision	recall	f1-score	support
0.0	1.00	0.86	0.93	55903
1.0	0.40	1.00	0.57	5003
accuracy			0.88	60906
macro avg	0.70	0.93	0.75	60906
weighted avg	0.95	0.88	0.90	60906



- ✓ HIGH RECALL
  - The lending club is able to classify who is having the difficulty in paying the loan. By predicting who is having the difficulty in paying the loan, the lending club is able to make the decision on distributing the loan to the appropriate customers.
  - The higher recall measure tells that out of total defaulters how many are actual defaulters.
  - But if the recall of the model is low, this might lead to distributing the loan, who has the difficulty in paying the loan. Hence for a good business of the loan distribution, we would like to preserve the recall of the model with a decent amount of accuracy.

- ✓ HIGHER ACCURACY
  - The accuracy is 0.88. this means the model is 88 percent accurate.
  - This measure gives the loan lending club able understanding that how good the model is.
  - With that in mind, you might think that for any sample (regardless of its class) the model is likely to make a correct prediction 88 % of the time.

✓ Higher AOC_ROC Score



- The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.
- When AUC = 0.93, then the classifier is able to perfectly distinguish between all the Positive and the Negative class points correctly.

Inference:

- The True positive and True negative is higher in the KNN model
- The False negative is low compared to other models
- The AUC score is higher than other models
- The Recall is higher for both the class

From the above KNN model, we can able clearly classify we can able to distinguish between the defaulter and non-defaulters. This helps the lending club able to make better business decisions and increase the performance of the lending business.

COMPARISON AND IMPLICATIONS:

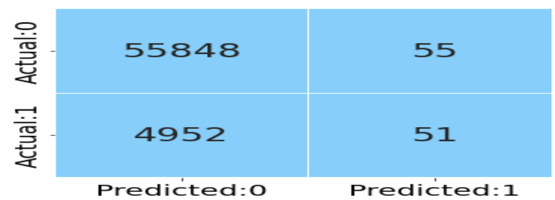
Applying the training and test data into the various model and our objective to sees there increase in the accuracy and recall for our model

LOGISTIC REGRESSION:

	precision	recall	f1-score	support
0	0.92	1.00	0.96	223956
1	0.47	0.01	0.02	19664
accuracy			0.92	243620
macro avg	0.70	0.50	0.49	243620
weighted avg	0.88	0.92	0.88	243620

	precision	recall	f1-score	support
0	0.92	1.00	0.96	55903
1	0.48	0.01	0.02	5003
accuracy			0.92	60906
macro avg	0.70	0.50	0.49	60906
weighted avg	0.88	0.92	0.88	60906

Test set



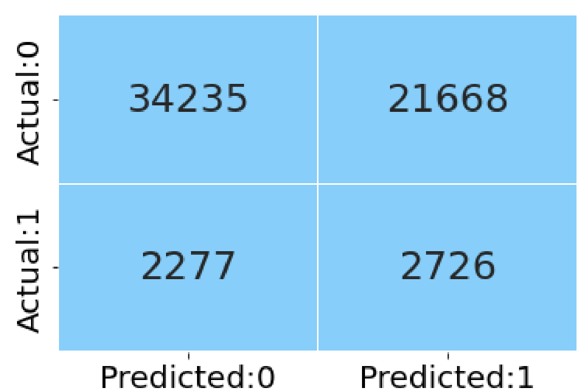
Training set

- The base model without feature scaling and trar
- Recall is very high only for class 1 and accuracy is higher
- The model is not able to find the actual defaulters, it predicted 4952 as non-defaulters.
-

1. LOGISTIC REGRESSION ON BALANCED DATASET:

	precision	recall	f1-score	support
0.0	0.96	0.69	0.80	223956
1.0	0.16	0.67	0.26	19664
accuracy			0.69	243620
macro avg	0.56	0.68	0.53	243620
weighted avg	0.90	0.69	0.76	243620

	precision	recall	f1-score	support
0.0	0.96	0.68	0.80	55903
1.0	0.16	0.67	0.26	5003
accuracy			0.68	60906
macro avg	0.56	0.67	0.53	60906
weighted avg	0.89	0.68	0.75	60906

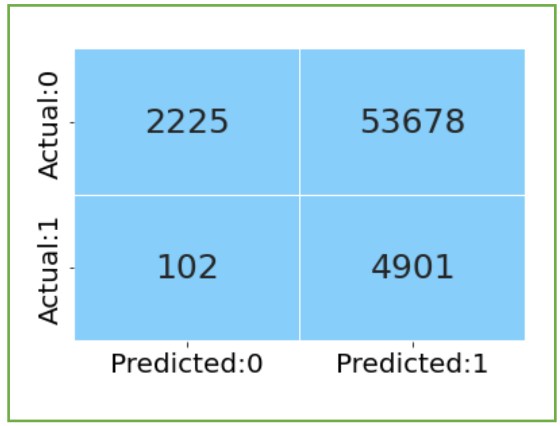


Logistic Regression Model with hyper-parameter tuning using Randomized Search – We perform hyperparameter tuning for the final baseline model obtained above using Randomised Search.
The model is not able to find the actual defaulters, it predicted 2277 as non-defaulters.

2. NAÏVE BAYES:

	precision	recall	f1-score	support
0.0	0.95	0.04	0.08	223956
1.0	0.08	0.98	0.15	19664
accuracy			0.12	243620
macro avg	0.52	0.51	0.11	243620
weighted avg	0.88	0.12	0.08	243620

	precision	recall	f1-score	support
0.0	0.96	0.04	0.08	55903
1.0	0.08	0.98	0.15	5003
accuracy			0.12	60906
macro avg	0.52	0.51	0.12	60906
weighted avg	0.88	0.12	0.08	60906

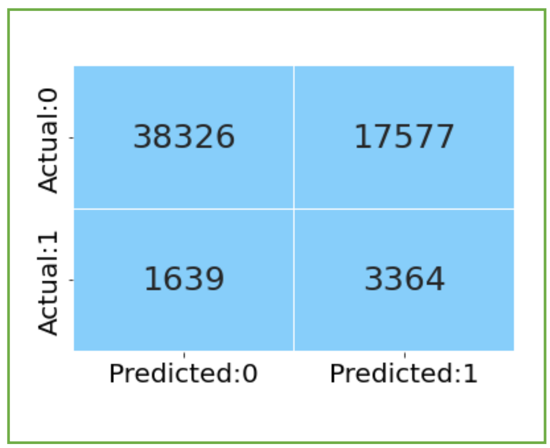


The f1- score is very low and recall is very high compare to other models
The model is able to find the actual defaulters, it predicted 102 as non-defaulters.
But the accuracy is very low.

3. ADA BOOST:

	precision	recall	f1-score	support
0.0	0.96	0.69	0.80	223956
1.0	0.16	0.67	0.26	19664
accuracy			0.69	243620
macro avg	0.56	0.68	0.53	243620
weighted avg	0.90	0.69	0.76	243620

	precision	recall	f1-score	support
0.0	0.96	0.69	0.80	55903
1.0	0.16	0.67	0.26	5003
accuracy			0.68	60906
macro avg	0.56	0.68	0.53	60906
weighted avg	0.89	0.68	0.76	60906

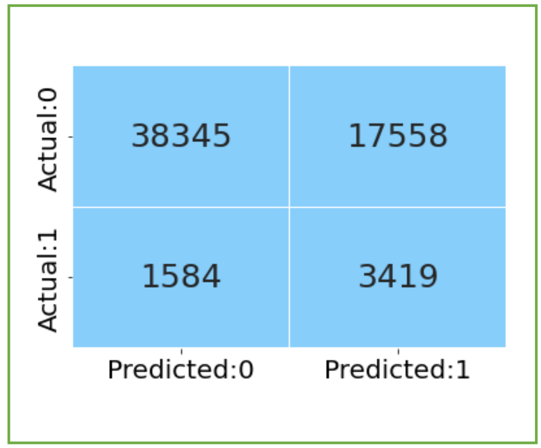


The model is not able to find the actual defaulters, it predicted 1639 as non-defaulters.

4. XG – BOOST:

	precision	recall	f1-score	support
0.0	0.96	0.69	0.80	223956
1.0	0.16	0.69	0.26	19664
accuracy			0.69	243620
macro avg	0.56	0.69	0.53	243620
weighted avg	0.90	0.69	0.76	243620

	precision	recall	f1-score	support
0.0	0.96	0.69	0.80	55903
1.0	0.16	0.68	0.26	5003
accuracy			0.69	60906
macro avg	0.56	0.68	0.53	60906
weighted avg	0.89	0.69	0.76	60906

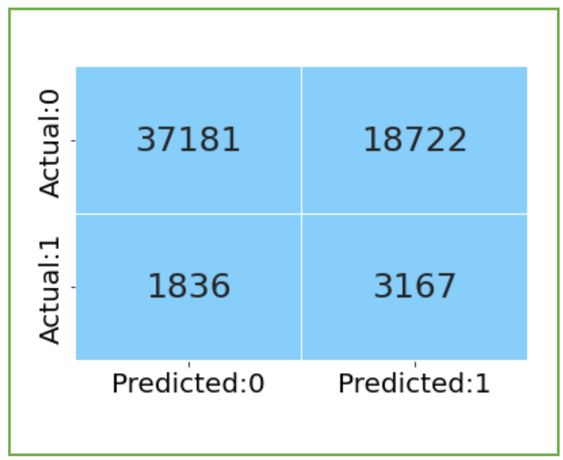


The model is not able to find the actual defaulters, it predicted 1584 as non-defaulters.

5. GRADIENT BOOSTING:

	precision	recall	f1-score	support
0.0	0.95	0.67	0.79	223956
1.0	0.14	0.64	0.24	19664
accuracy			0.67	243620
macro avg	0.55	0.65	0.51	243620
weighted avg	0.89	0.67	0.74	243620

	precision	recall	f1-score	support
0.0	0.95	0.67	0.78	55903
1.0	0.14	0.63	0.24	5003
accuracy			0.66	60906
macro avg	0.55	0.65	0.51	60906
weighted avg	0.89	0.66	0.74	60906



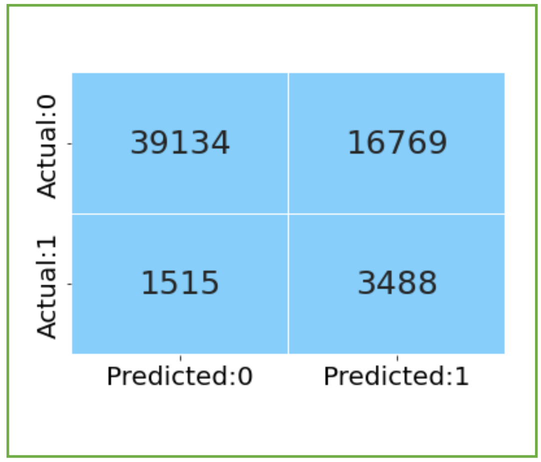
The model is not able to find the actual defaulters, it predicted 1836 as non-defaulters.

REGULARIZATION:

XGBOOST:

	precision	recall	f1-score	support
0.0	0.96	0.70	0.81	223956
1.0	0.17	0.70	0.28	19664
accuracy			0.70	243620
macro avg	0.57	0.70	0.55	243620
weighted avg	0.90	0.70	0.77	243620

	precision	recall	f1-score	support
0.0	0.96	0.70	0.81	55903
1.0	0.17	0.70	0.28	5003
accuracy			0.70	60906
macro avg	0.57	0.70	0.54	60906
weighted avg	0.90	0.70	0.77	60906

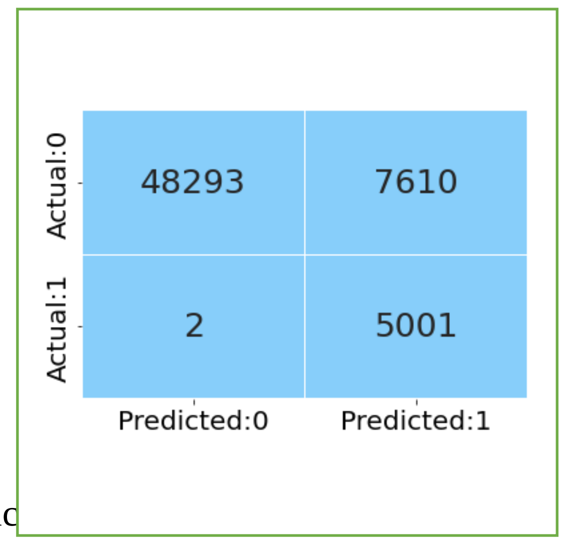


The model is not able to find the actual defaulters, it predicted 1515 as non-defaulters.

6. FINAL MODEL – KNN:

	precision	recall	f1-score	support
0.0	0.96	0.86	0.93	223956
1.0	0.17	1.00	0.57	19664
accuracy			0.88	243620
macro avg	0.57	0.93	0.75	243620
weighted avg	0.90	0.88	0.90	243620

	precision	recall	f1-score	support
0.0	1.00	0.86	0.93	55903
1.0	0.40	1.00	0.57	5003
accuracy			0.88	60906
macro avg	0.70	0.93	0.75	60906
weighted avg	0.95	0.88	0.90	60906

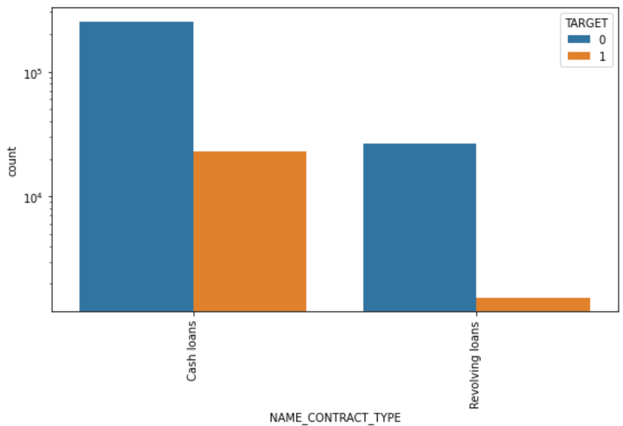


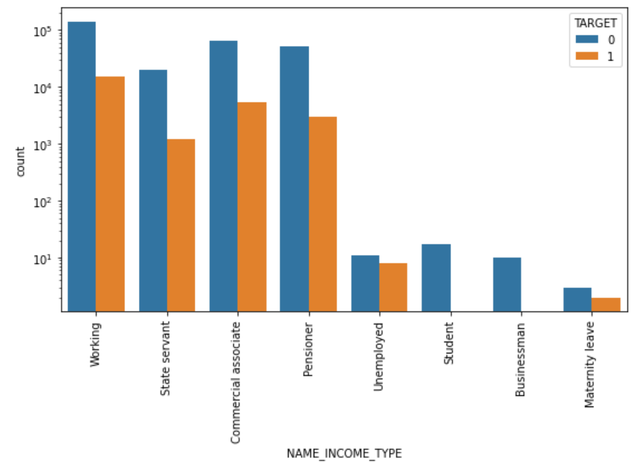
The model is able to find the actual defaulters, it predic

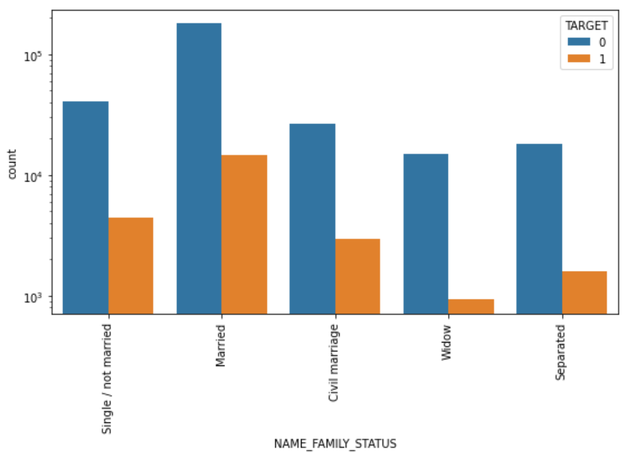
Important points:

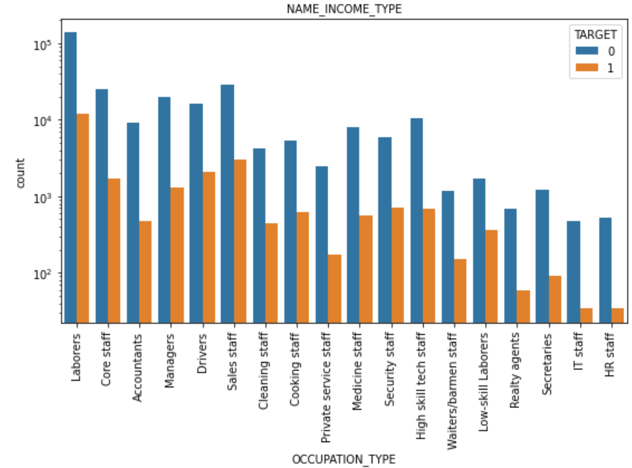
- The True positive and True negative is higher in the KNN model
- The False negative is low compared to other models
- F1-score is higher than other models
- The recall for both the class is higher
- The false-positive is lower compared to another model.

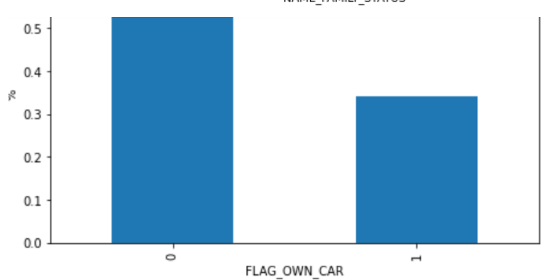
Inferences and Recommendations:

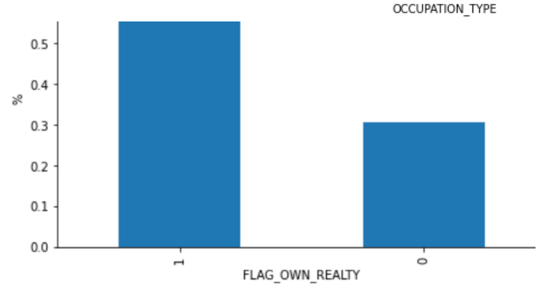


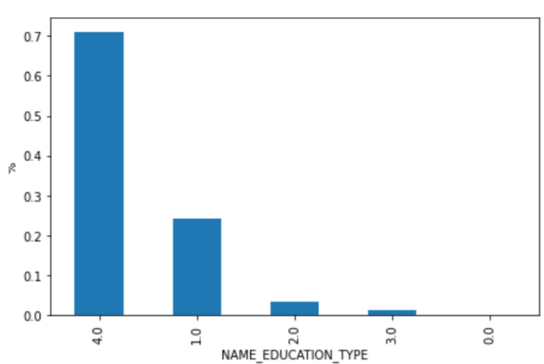


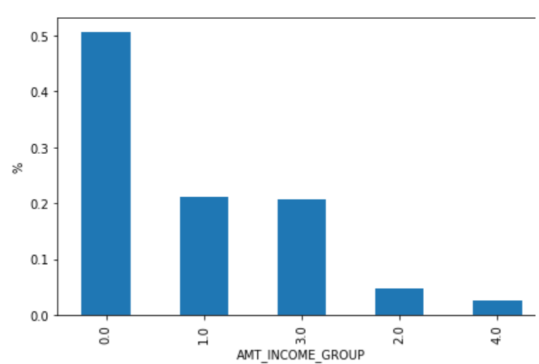


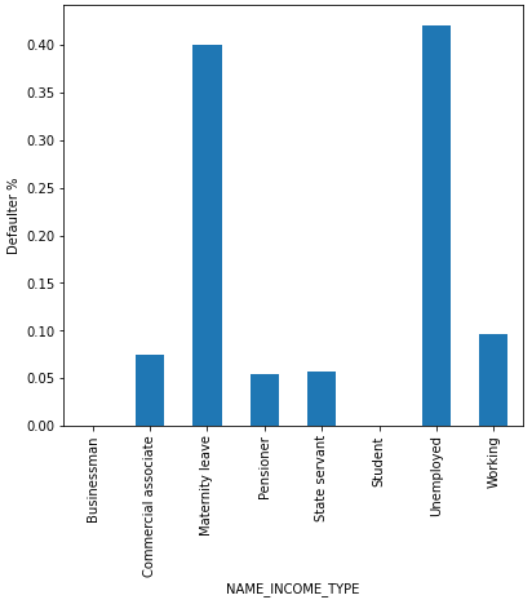


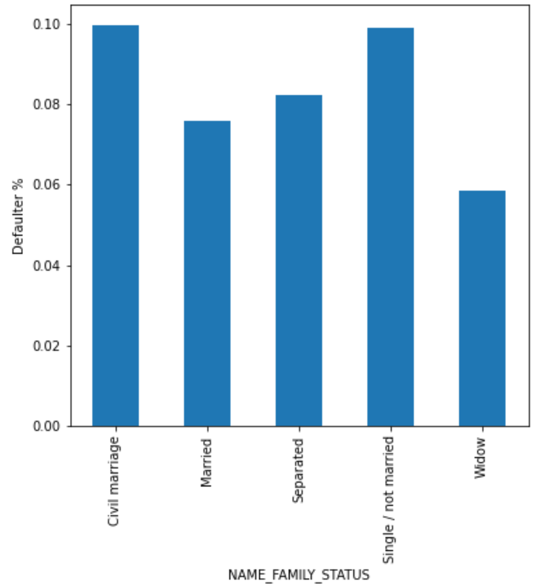


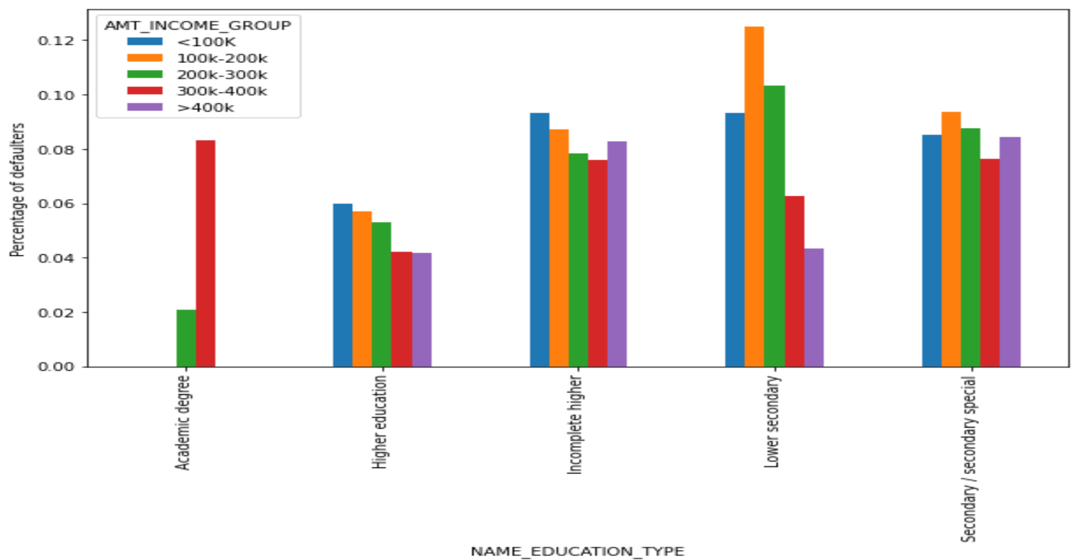












INFERENCE:

- Business man and IT STAFF have higher chances of repaying the loans. Chances that an unemployed become a defaulter is more
- Widow category has fewer chances of becoming defaulter whereas married has high chances. The civil marriage category has fewer defaulters compared to the single
- More than 60% of the clients don't have a car is no defaulters
- The defaulter percentage is higher in a customer in maternity leave and unemployed
- Most of the clients have done only secondary education
- The customer is defaulters who owns the property
- The customer who education of lower secondary is more defaulter than others

RECOMMENDATION:

- Education plays a major role in repaying the loan in time, so we recommend verifying the educational background before distribution loan
- The high earning customers have a chance of paying the loan in time so we can give some attractive offers to them.
- Gender doesn't play important role in the difficulty in repaying the loan.
- Some customers have the highest degree but earn very low, So have to be careful with the customer with higher education.
- Check the credit score before distribution of loan

LIMITATIONS:

A few of the limitations are: -

- The dataset is about the loan lending club which is collected between 2007 to 2015. Due to this, the information regarding the current data is missing

- The dataset belongs to the USA and consists of data from only two hotels. The model will be more robust if the data would have belonged from different regions of the world.

CHALLENGES:

- High cardinality results in huge training effort in model tuning due to an increase in model complexity (i.e. more number of features)
- The model dimension is very large, there are some challenges in processing the data for modeling.
- The percentage of the variable with missing value Is higher. Processing it is challenging
- We also faced challenges on robust model tuning on all the models. Due to computational limitations, we are limited to using Randomized Search as a hyperparameter tuning technique instead of using Grid Search, etc.

SCOPE:

Scope for some future work is: -

- Perform hyperparameter tuning for the ensemble model since due to the lower processing power of our laptops, we couldn't do that.
- Exploring Google collab as an option for model training and tuning with a faster lead time.
- Exploring some robust data sampling techniques as part of choosing a smaller sample (a true representation of population data) from the population data.
- The missing value is higher in the dataset which can be imputed for further analysis using industrial knowledge and try the prediction.