greatlearning

CREDIT RISK ASSESSMENT FOR HOME CREDIT GROUP

GROUP-8

Final Report

Mentored by:

Sravan Malla

Submitted by:

Sandeep

Sasikiran

Shashank

Vivek

Vyshnavi

Contents

Part 1 Project Background:

Home Credit is an international Non-Banking Financial Institution (NBFC) founded in 1997 in the Czech Republic. The company operates in 14 countries and focuses on lending primarily to people with little or no credit history.

Home Credit strives to broaden financial inclusion for the unbanked population by providing a positive and safe borrowing experience. In order to make sure this underserved population has a positive loan experience, Home Credit makes use of a variety of alternative data--including Telco and transactional information--to predict their clients' repayment abilities.

Home Credit is currently using various statistical and machine learning methods to make these predictions successful. This will ensure that clients capable of repayment are not rejected and that loans are given with a principal, maturity, and repayment calendar that will empower their clients to be successful.

Part 2 About the Data:

There are 307511 observations in the dataset with 122 columns that stand for both qualitative and quantitative attributes of which 67 columns have missing values.

Out of 122, 16 columns are categorical and 106 are numerical columns.

There is a binary output variable that denotes “Delay in payments” (1) or “No Delay in payments (0)”.

Source:

This dataset is taken from Kaggle competition.

https://www.kaggle.com/c/home-credit-default-risk

Project Justification:

1) Credit risk management is the practice of mitigating losses by understanding the adequacy of a bank's capital and loan loss reserves at any given time – a process that has long been a challenge for financial institutions.

2) Conservative credit risk management policies, fast loan decisions and reasonable loan pricing achieve this balance of protecting loan portfolios while keeping bank customers satisfied with the institution.

3) The objective of this project is to predict the home loan credit risk for the financial institution. The project will enable the bank to reduce their risk of loan loss by gaining an apt understanding its customer base, thus minimizing the loss of capital for the financial institution while reaping optimal profit.

4) By analyzing the customer features such as transaction history, annual income, demographics etc., and the bank will be able to estimate the risk of the loan repayment.

Part 3 Data Description and Cleaning:

3.1 Data Description:

S.No	Row	Description
1.	SK_ID_CURR	ID of loan in our sample
2.	TARGET	Target variable (1 - client with payment difficulties, 0 - all other cases)
3.	NAME_CONTRACT_TYPE	Identification if loan is cash or revolving
4.	CODE_GENDER	Gender of the client
5.	FLAG_OWN_CAR	Flag if the client owns a car
6.	FLAG_OWN_REALTY	Flag if client owns a house or flat
7.	CNT_CHILDREN	Number of children the client has
8.	AMT_INCOME_TOTAL	Income of the client
9.	AMT_CREDIT	Credit amount of the loan
10.	AMT_ANNUITY	Loan annuity
11.	AMT_GOODS_PRICE	For consumer loans it is the price of the goods for which the loan is given
12.	NAME_TYPE_SUITE	Who was accompanying client when he was applying for the loan
13.	NAME_INCOME_TYPE	Client's income type (businessman, working, maternity leave etc.)
14.	NAME_EDUCATION_TYPE	Level of highest education the client achieved
15.	NAME_FAMILY_STATUS	Family status of the client
16.	NAME_HOUSING_TYPE	What is the housing situation of the client (renting, living with parents etc.)
17.	REGION_POPULATION_RELATIVE	Normalized population of region where client lives (higher number means the client lives in more populated region)
18.	DAYS_BIRTH	Client's age in days at the time of application
19.	DAYS_EMPLOYED	How many days before the application the person started current employment
20.	DAYS_REGISTRATION	How many days before the application did client change his registration
21.	DAYS_ID_PUBLISH	How many days before the application did client change the identity document with which he applied for the loan
22.	OWN_CAR_AGE	Age of client's car
23.	FLAG_MOBIL	Did client provide mobile phone (1=YES, 0=NO)
24.	FLAG_EMP_PHONE	Did client provide work phone (1=YES, 0=NO)
25.	FLAG_WORK_PHONE	Did client provide home phone (1=YES, 0=NO)
26.	FLAG_CONT_MOBILE	Was mobile phone reachable (1=YES, 0=NO)
27.	FLAG_PHONE	Did client provide home phone (1=YES, 0=NO)
28.	FLAG_EMAIL	Did client provide email (1=YES, 0=NO)
29.	OCCUPATION_TYPE	What kind of occupation does the client have
30.	CNT_FAM_MEMBERS	How many family members does client have
31.	REGION_RATING_CLIENT	Our rating of the region where client lives (1,2,3)
32.	REGION_RATING_CLIENT_W_CITY	Our rating of the region where client lives with taking city

		into account (1,2,3)
33.	WEEKDAY_APPR_PROCESS_START	On which day of the week did the client apply for the loan
34.	HOUR_APPR_PROCESS_START	Approximately at what hour did the client apply for the loan
35.	REG_REGION_NOT_LIVE_REGION	Flag if client's permanent address does not match contact address (1=different, 0=same, at region level)
36.	REG_REGION_NOT_WORK_REGION	Flag if client's permanent address does not match work address (1=different, 0=same, at region level)
37.	LIVE_REGION_NOT_WORK_REGION	Flag if client's contact address does not match work address (1=different, 0=same, at region level)
38.	REG_CITY_NOT_LIVE_CITY	Flag if client's permanent address does not match contact address (1=different, 0=same, at city level)
39.	REG_CITY_NOT_WORK_CITY	Flag if client's permanent address does not match work address (1=different, 0=same, at city level)
40.	LIVE_CITY_NOT_WORK_CITY	Flag if client's contact address does not match work address (1=different, 0=same, at city level)
41.	ORGANIZATION_TYPE	Type of organization where client works
42.	EXT_SOURCE_1	Normalized score from external data source
43.	EXT_SOURCE_2	Normalized score from external data source
44.	EXT_SOURCE_3	Normalized score from external data source
45.	APARTMENTS_AVG	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
46.	BASEMENTAREA_AVG	Normalized information about building where the client lives
47.	YEARS_BEGINEXPLUATATION_AVG	Normalized information about building where the client lives
48.	YEARS_BUILD_AVG	Normalized information about building where the client lives
49.	COMMONAREA_AVG	Normalized information about building where the client lives
50.	ELEVATORS_AVG	Normalized information about building where the client lives
51.	ENTRANCES_AVG	Normalized information about building where the client lives
52.	FLOORSMAX_AVG	Normalized information about building where the client lives
53.	FLOORSMIN_AVG	Normalized information about building where the client lives
54.	LANDAREA_AVG	Normalized information about building where the client lives
55.	LIVINGAPARTMENTS_AVG	Normalized information about building where the client lives
56.	LIVINGAREA_AVG	Normalized information about building where the client lives
57.	NONLIVINGAPARTMENTS_AVG	Normalized information about building where the client lives
58.	NONLIVINGAREA_AVG	Normalized information about building where the client lives

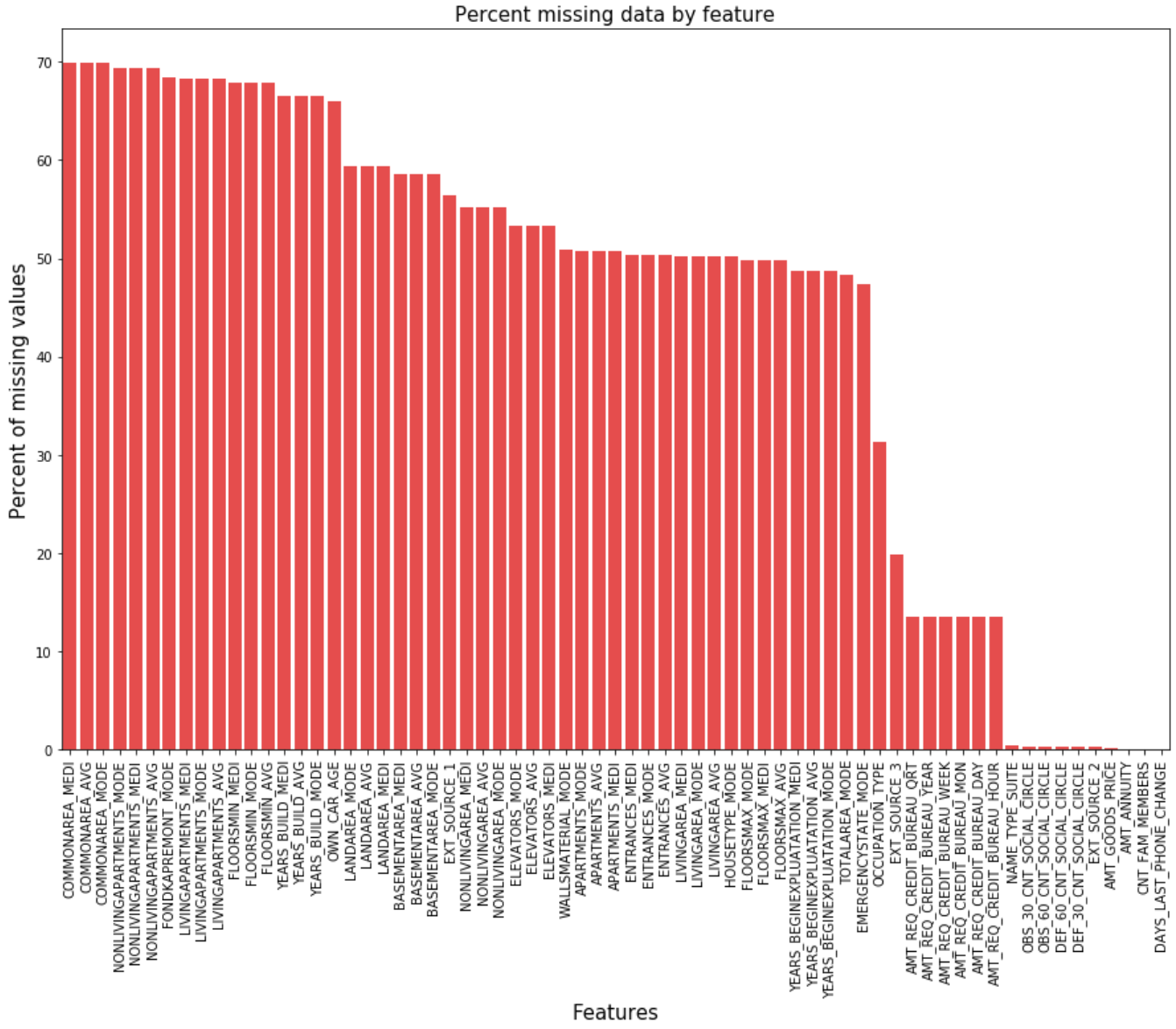
59.	APARTMENTS_MODE	Normalized information about building where the client lives
60.	BASEMENTAREA_MODE	Normalized information about building where the client lives
61.	YEARS_BEGINEXPLUATATION_MODE	Normalized information about building where the client lives
62.	YEARS_BUILD_MODE	Normalized information about building where the client lives
63.	COMMONAREA_MODE	Normalized information about building where the client lives
64.	ELEVATORS_MODE	Normalized information about building where the client lives
65.	ENTRANCES_MODE	Normalized information about building where the client lives
66.	FLOORSMAX_MODE	Normalized information about building where the client lives
67.	FLOORSMIN_MODE	Normalized information about building where the client lives
68.	LANDAREA_MODE	Normalized information about building where the client lives
69.	LIVINGAPARTMENTS_MODE	Normalized information about building where the client lives
70.	LIVINGAREA_MODE	Normalized information about building where the client lives
71.	NONLIVINGAPARTMENTS_MODE	Normalized information about building where the client lives
72.	NONLIVINGAREA_MODE	Normalized information about building where the client lives
73.	APARTMENTS_MEDI	Normalized information about building where the client lives
74.	BASEMENTAREA_MEDI	Normalized information about building where the client lives
75.	YEARS_BEGINEXPLUATATION_MEDI	Normalized information about building where the client lives
76.	YEARS_BUILD_MEDI	Normalized information about building where the client lives
77.	COMMONAREA_MEDI	Normalized information about building where the client lives
78.	ELEVATORS_MEDI	Normalized information about building where the client lives
79.	ENTRANCES_MEDI	Normalized information about building where the client lives
80.	FLOORSMAX_MEDI	Normalized information about building where the client lives
81.	FLOORSMIN_MEDI	Normalized information about building where the client lives
82.	LANDAREA_MEDI	Normalized information about building where the client lives
83.	LIVINGAPARTMENTS_MEDI	Normalized information about building where the client lives
84.	LIVINGAREA_MEDI	Normalized information about building where the client lives
85.	NONLIVINGAPARTMENTS_MEDI	Normalized information about building where the client lives

86.	NONLIVINGAREA_MEDI	Normalized information about building where the client lives
87.	FONDKAPREMONT_MODE	Normalized information about building where the client lives
88.	HOUSETYPE_MODE	Normalized information about building where the client lives
89.	TOTALAREA_MODE	Normalized information about building where the client lives
90.	WALLSMATERIAL_MODE	Normalized information about building where the client lives
91.	EMERGENCYSTATE_MODE	Normalized information about building where the client lives
92.	OBS_30_CNT_SOCIAL_CIRCLE	How many observation of client's social surroundings with observable 30 DPD (days past due) default
93.	DEF_30_CNT_SOCIAL_CIRCLE	How many observation of client's social surroundings defaulted on 30 DPD (days past due)
94.	OBS_60_CNT_SOCIAL_CIRCLE	How many observation of client's social surroundings with observable 60 DPD (days past due) default
95.	DEF_60_CNT_SOCIAL_CIRCLE	How many observation of client's social surroundings defaulted on 60 (days past due) DPD
96.	DAYS_LAST_PHONE_CHANGE	How many days before application did client change phone
97.	FLAG_DOCUMENT_2	Did client provide document 2
98.	FLAG_DOCUMENT_3	Did client provide document 3
99.	FLAG_DOCUMENT_4	Did client provide document 4
100.	FLAG_DOCUMENT_5	Did client provide document 5
101.	FLAG_DOCUMENT_6	Did client provide document 6
102.	FLAG_DOCUMENT_7	Did client provide document 7
103.	FLAG_DOCUMENT_8	Did client provide document 8
104.	FLAG_DOCUMENT_9	Did client provide document 9
105.	FLAG_DOCUMENT_10	Did client provide document 10
106.	FLAG_DOCUMENT_11	Did client provide document 11
107.	FLAG_DOCUMENT_12	Did client provide document 12
108.	FLAG_DOCUMENT_13	Did client provide document 13
109.	FLAG_DOCUMENT_14	Did client provide document 14
110.	FLAG_DOCUMENT_15	Did client provide document 15
111.	FLAG_DOCUMENT_16	Did client provide document 16
112.	FLAG_DOCUMENT_17	Did client provide document 17
113.	FLAG_DOCUMENT_18	Did client provide document 18
114.	FLAG_DOCUMENT_19	Did client provide document 19
115.	FLAG_DOCUMENT_20	Did client provide document 20
116.	FLAG_DOCUMENT_21	Did client provide document 21
117.	AMT_REQ_CREDIT_BUREAU_HOUR	Number of enquiries to Credit Bureau about the client one hour before application
118.	AMT_REQ_CREDIT_BUREAU_DAY	Number of enquiries to Credit Bureau about the client one day before application (excluding one hour before application)

119.	AMT_REQ_CREDIT_BUREAU_WEEK	Number of enquiries to Credit Bureau about the client one week before application (excluding one day before application)
120.	AMT_REQ_CREDIT_BUREAU_MON	Number of enquiries to Credit Bureau about the client one month before application (excluding one week before application)
121.	AMT_REQ_CREDIT_BUREAU_QRT	Number of enquiries to Credit Bureau about the client 3 month before application (excluding one month before application)
122.	AMT_REQ_CREDIT_BUREAU_YEAR	Number of enquiries to Credit Bureau about the client one day year (excluding last 3 months before application)

3.2 Data Cleaning:

Missing Values in dataset:



Imputation:

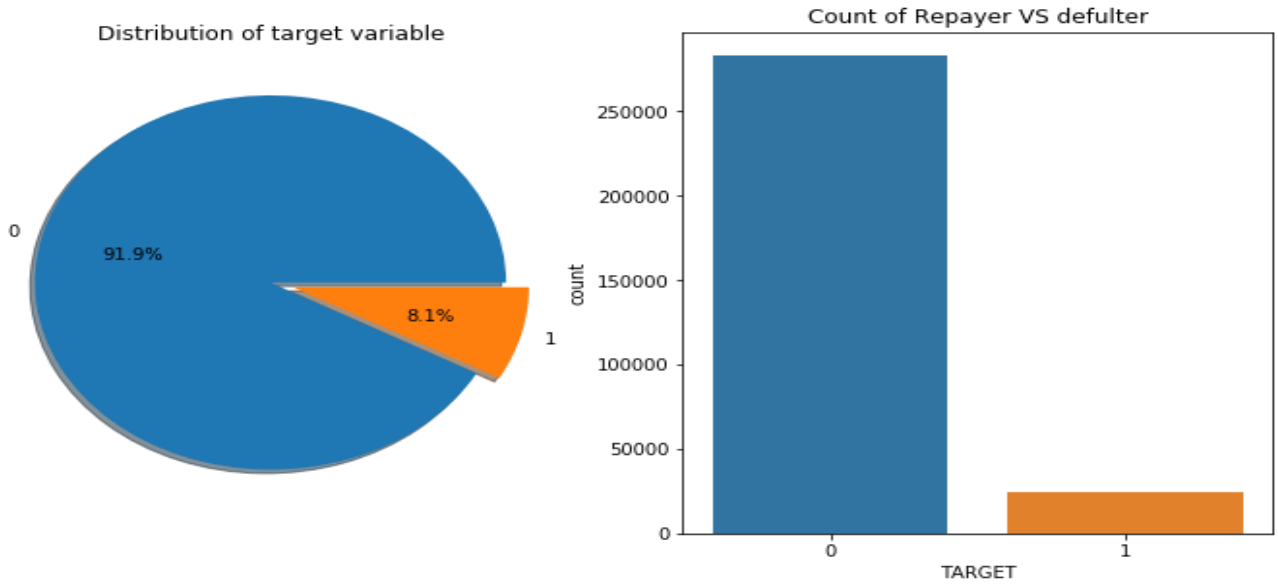
Treating missing values plays a vital role in building machine learning models. It is required to fill these missing values (known as imputation). Later, we can use models such as XGBoost that can handle missing values with no need for imputation. Another option would be to drop columns with a high percentage of missing values, although it is difficult to know whether these columns would be helpful to our model. Therefore, we will keep all of the columns for now. For continuous variables, we used median imputation and for categorical variables, we used mode imputation.

Part 4 Exploratory Data Analysis:

Relationship between the variables:

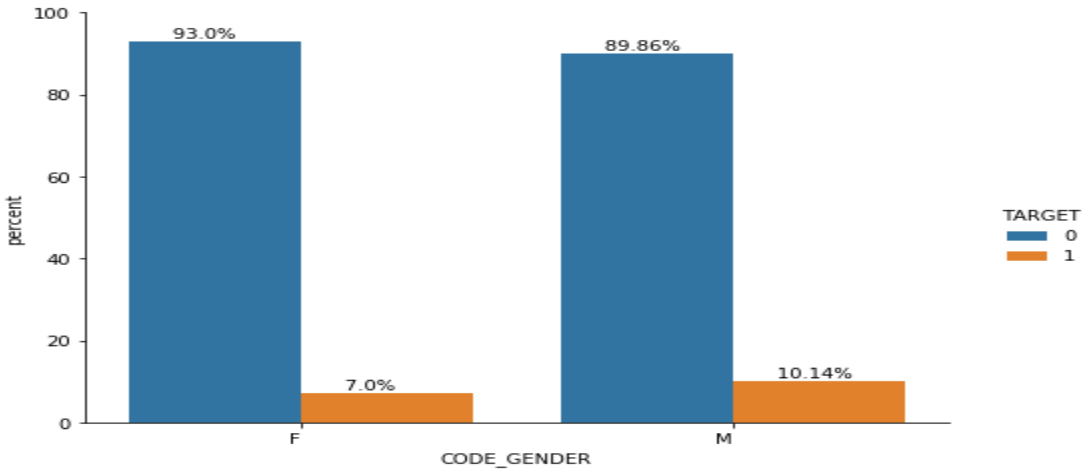
There are total of 122 features and since visualizations of all 122 is difficult to interpret. So we considered some variables with respect to Target variable. Let us see the various visual analytics as follows:

4.1 Visualize the Target variable:



From the count plot we can observe that the target variable is imbalanced with 91.9% of 0 and 8.1% of 1 are there. In the future steps we are treating the target variable with oversampling method called SMOTE.

4.2 Visualize the relationship of Gender vs. Target variable:

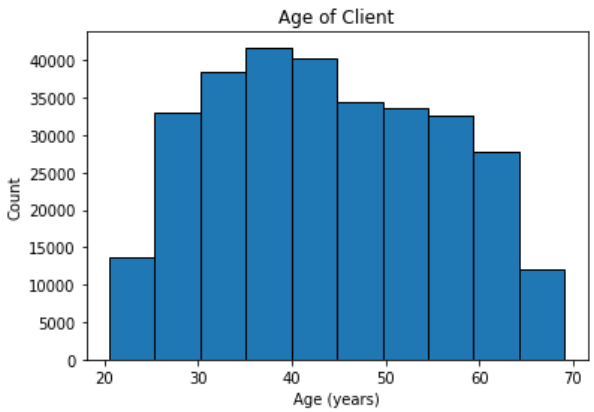


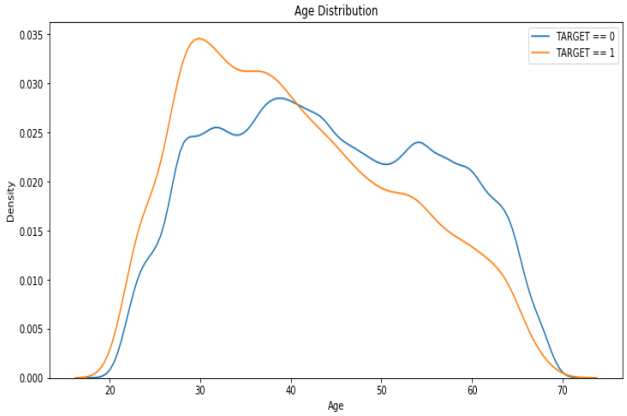
We see that male clients have high percentage of difficulties to repay the loan compared to female clients.

93% of female customers are repaying loan whereas 7% are facing difficulties in repaying the loan.

89.86% of male customers are repaying loan whereas 10.14% are facing difficulties in repaying the loan.

4.3 Visualize the relationship of Age vs. Target Variable:





The correlation between Age and the TARGET is -0.0782

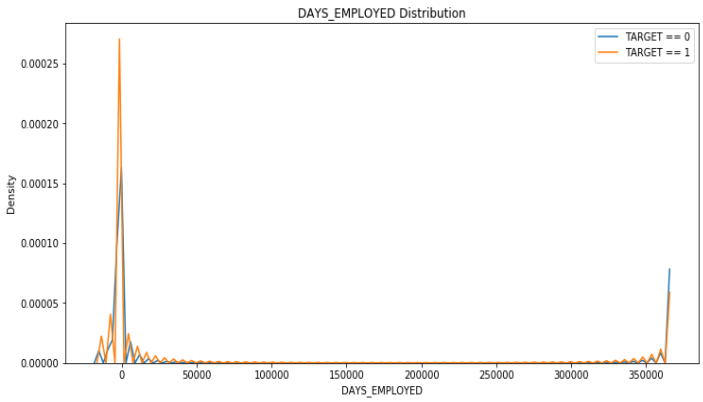
Median value for loan that was not repaid = 39.1288

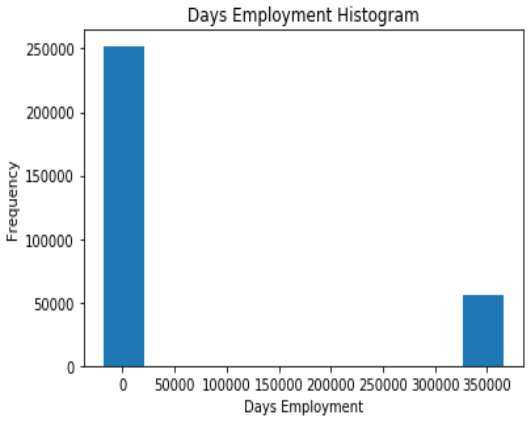
Median value for loan that was repaid = 43.4986

The numbers in the DAYS_BIRTH column are negative because they are recorded relative to the current loan application. To see these stats in years, we multiplied by -1 and divided by the number of days in a year.

As the client gets older than 40, he seems to repay the loan on time and the clients younger than 40 are mostly delaying the repayment of the loan, as the younger clients are less likely to repay the loan, instead of discriminating against them, it would be smart to provide more guidance and financial planning tips to them. By doing this we can take precautionary measures to facilitate younger clients to pay on time.

4.4 Visualize the distribution of Days employed vs. Target Variable:



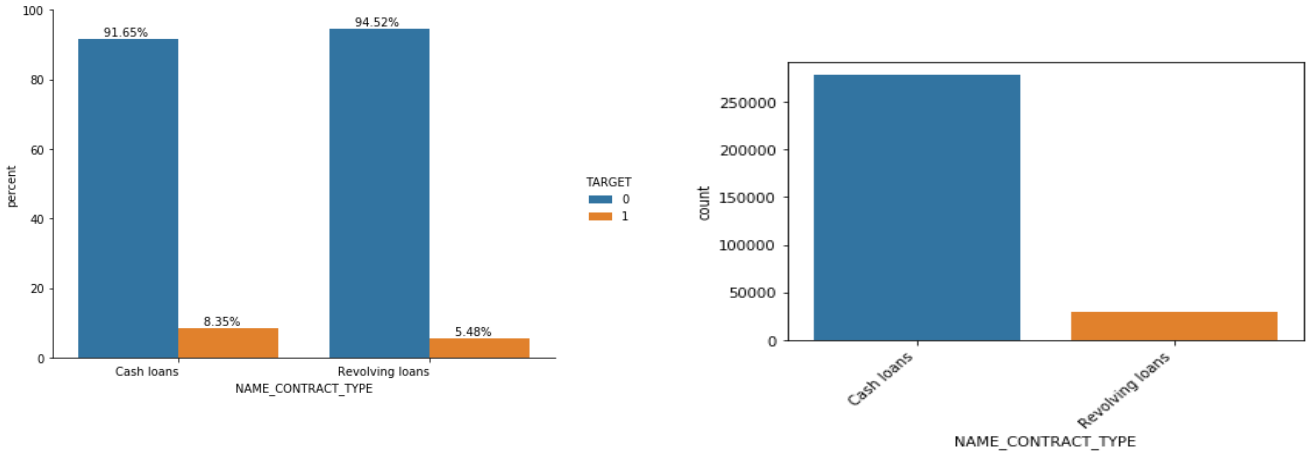


The correlation between DAYS_EMPLOYED and the TARGET is -0.0449

Median value for loan that was not repaid = -1034.0000, Median value for loan that was repaid = -1235.0000

The numbers in the DAYS_EMPLOYED column are negative because they are recorded relative to the current loan application. To see these stats in years, we multiplied by -1 and divided by the number of days in a year. Quite often, we will be dealing with anomalies in the dataset. These may be due to mistyped numbers, errors in measuring equipment, or they could be valid but extreme measurements. One such anomaly is seen in the above graph where we have 55374 records with 365243 as the DAYS_EMPLOYED. So we replaced these 55k records with Nan values and then replaced with the median value.

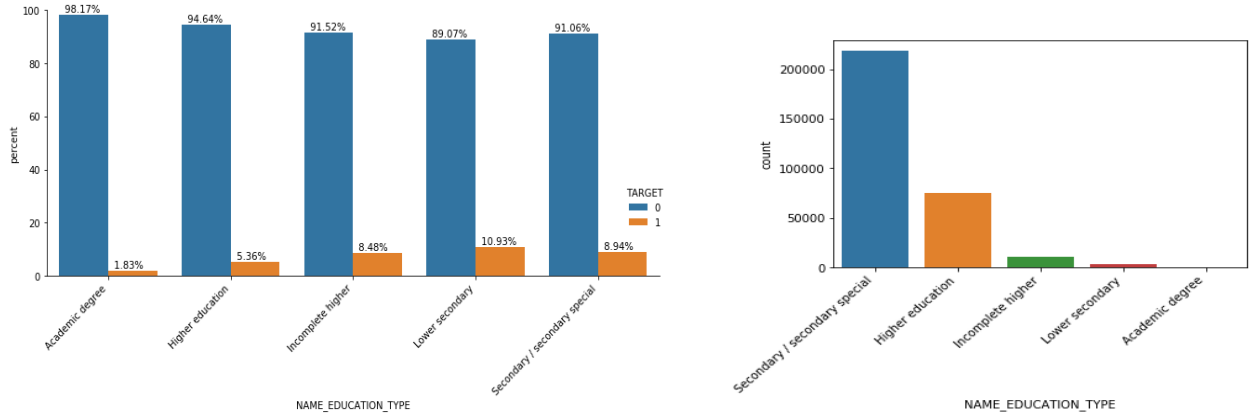
4.5 Visualize the relationship of Loan Type vs. Target variable:



91.65% of cash loans are repaid properly whereas 8.35% of the loan payments are delayed.

94.52% of revolving loans are repaid properly whereas 5.48% of the loan payments are delayed.

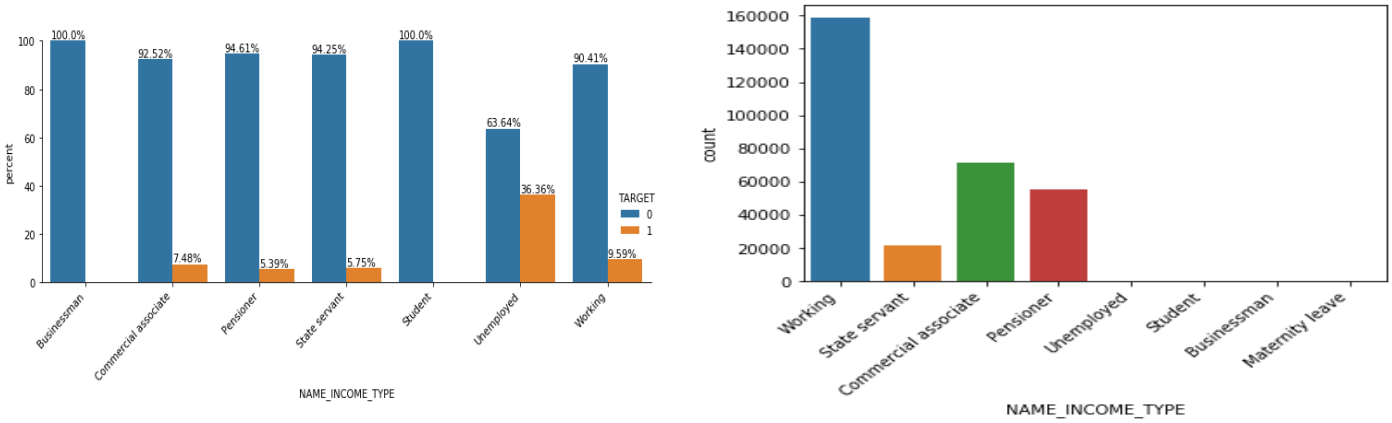
4.6 Visualize the relationship of Education Type vs. Target Variable:



We observe that the applicants with Lower Secondary education status have the highest percentage of payment related problems followed by applicants with Secondary/secondary special status.

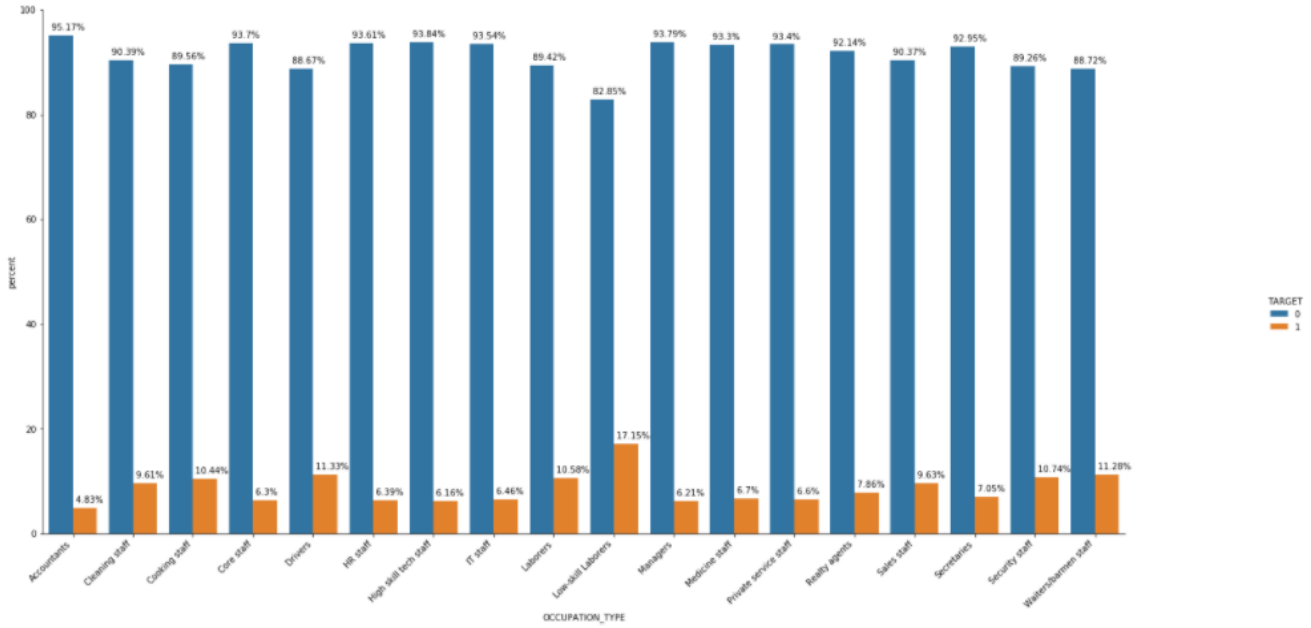
A large number of applications (above 2Lakh) are filed by people having secondary education followed by people with Higher Education (with nearly 75K applications).

4.7 Visualize the relationship of Income Type vs. Target variable:



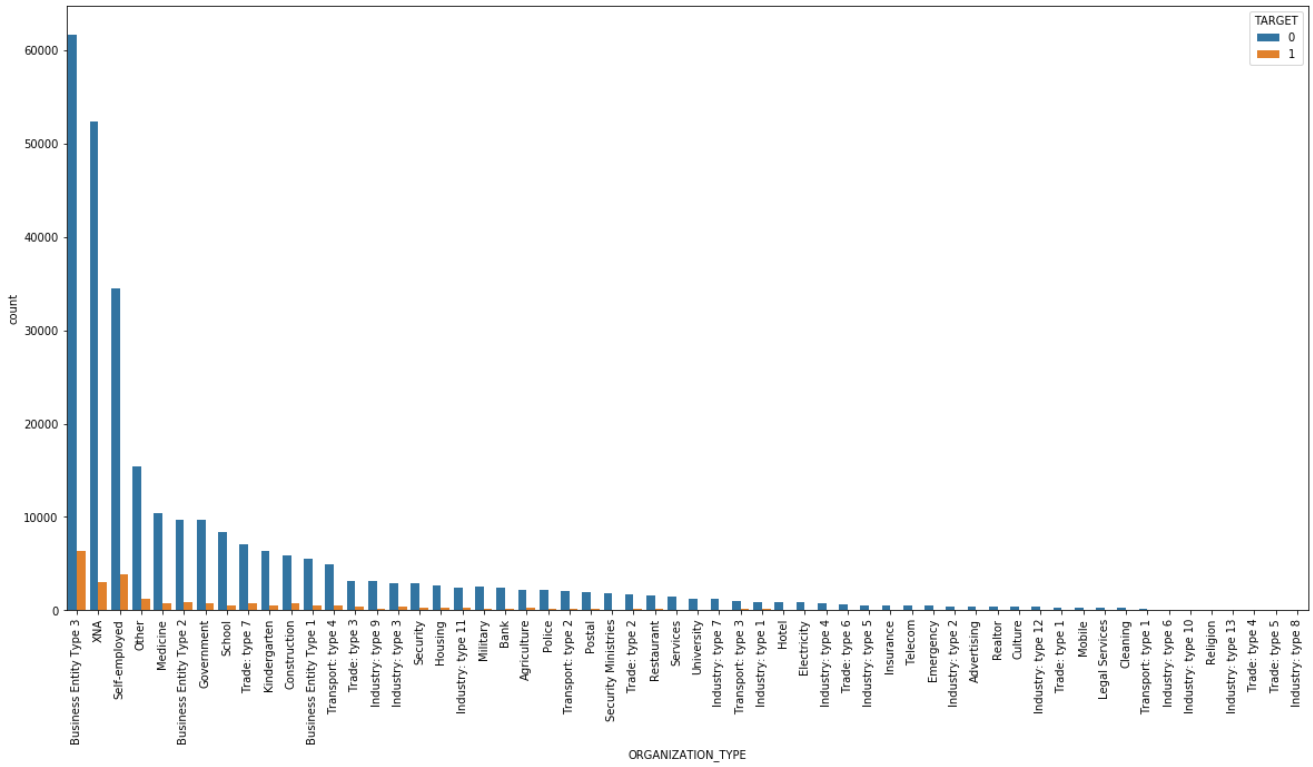
We observe that the applicants who are Unemployed have the highest percentage of payment related problems followed by applicants who are working.

4.8 Visualize the relationship of Occupation Type vs. Target Variable:



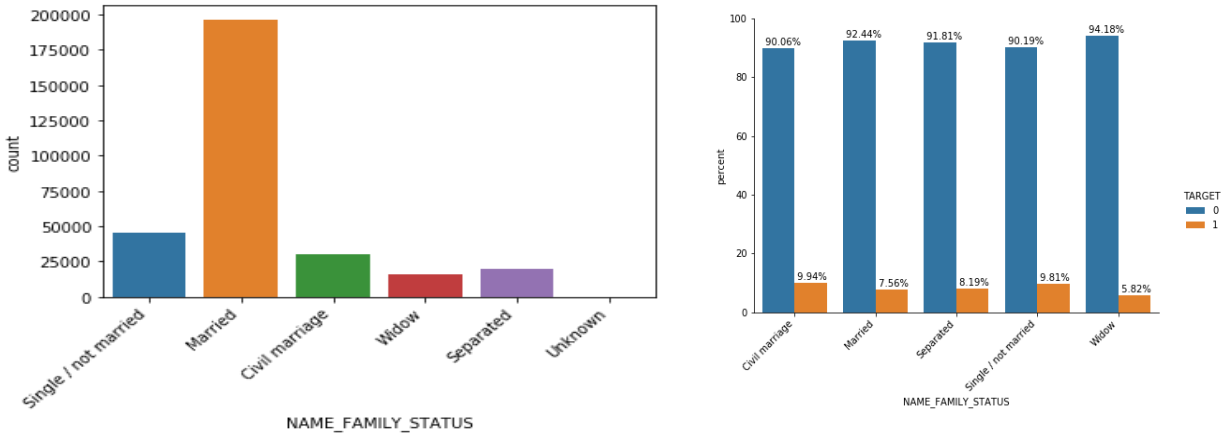
We observe that the applicants who are Low skill Laborers have the highest percentage of payment related problems followed by applicants who are drivers and waiters/barmen staff.

4.9 Visualize the relationship of Organization Type vs. Target Variable:



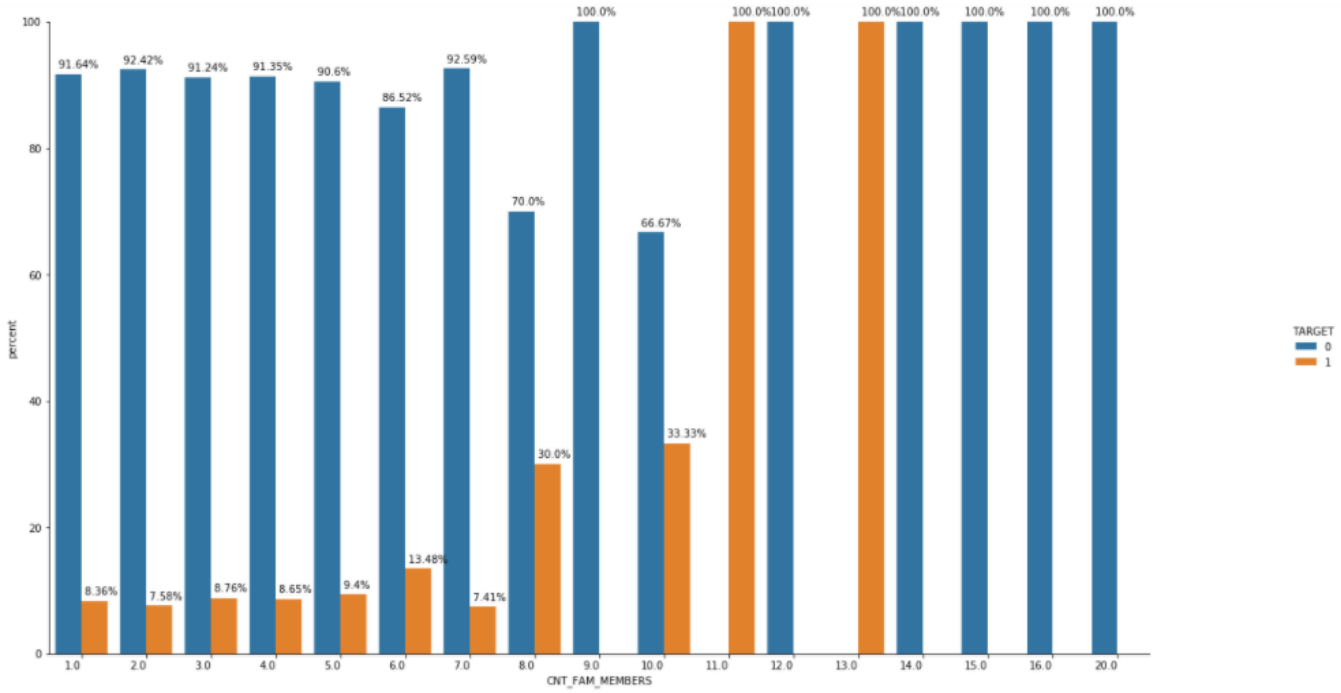
We observe that the applicants who are in organization type of Business Entity type 3 have the highest percentage of payment related problems followed by applicants who are self-employed.

4.10 Visualize the relationship of Family Status vs. Target variable



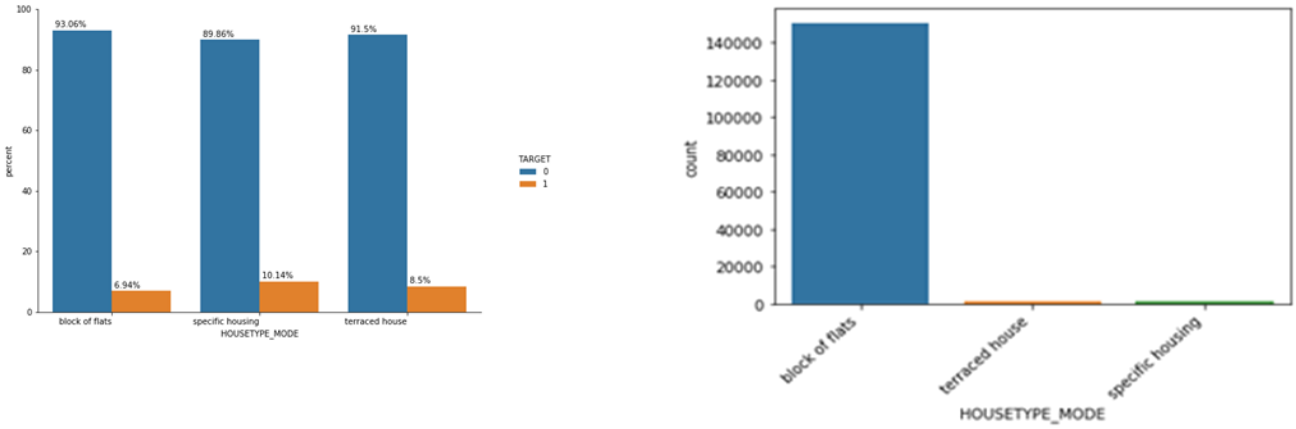
We observe that the applicants with Civil marriage have the highest percentage of payment related problems followed by applicants who are Single/ not married.

4.11 Visualize the relationship of Family members count vs. Target variable:



It is clear by looking at the above graph that Families with lesser size are repaying the loan on time and as the size of the family increased, there are delays in loan repayment. As the records of the families with sizes 14, 15, 16, 20 are only a few for each size; there is no proper interpretability for these sizes.

4.12 Visualize the relationship of House Type vs. Target Variable:



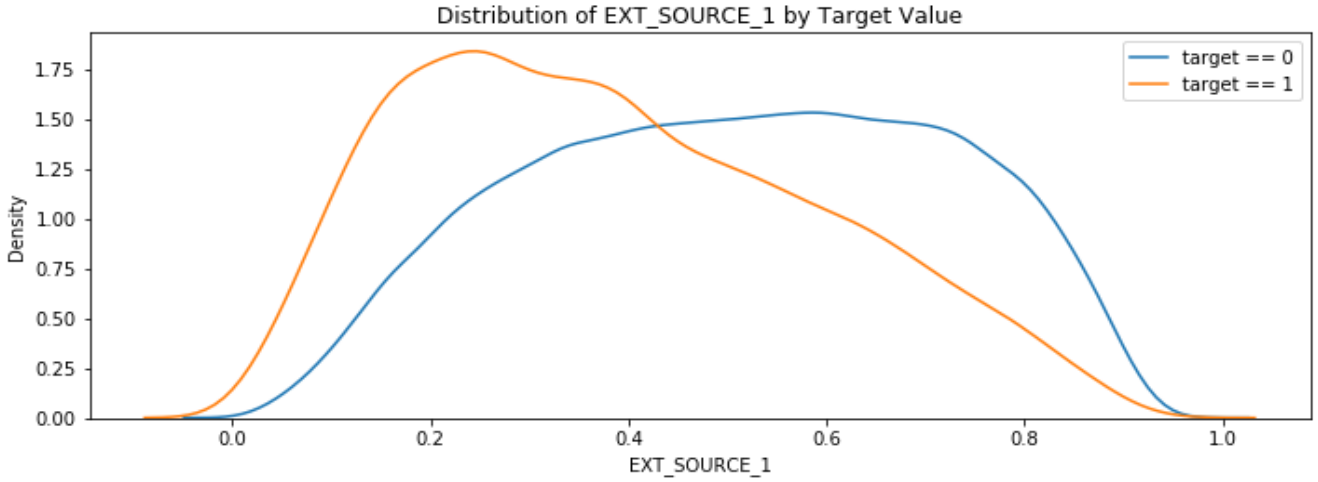
It is clear by looking at the above given plot that clients with house type mode as Specific housing were not repay the loan compared to other house types.

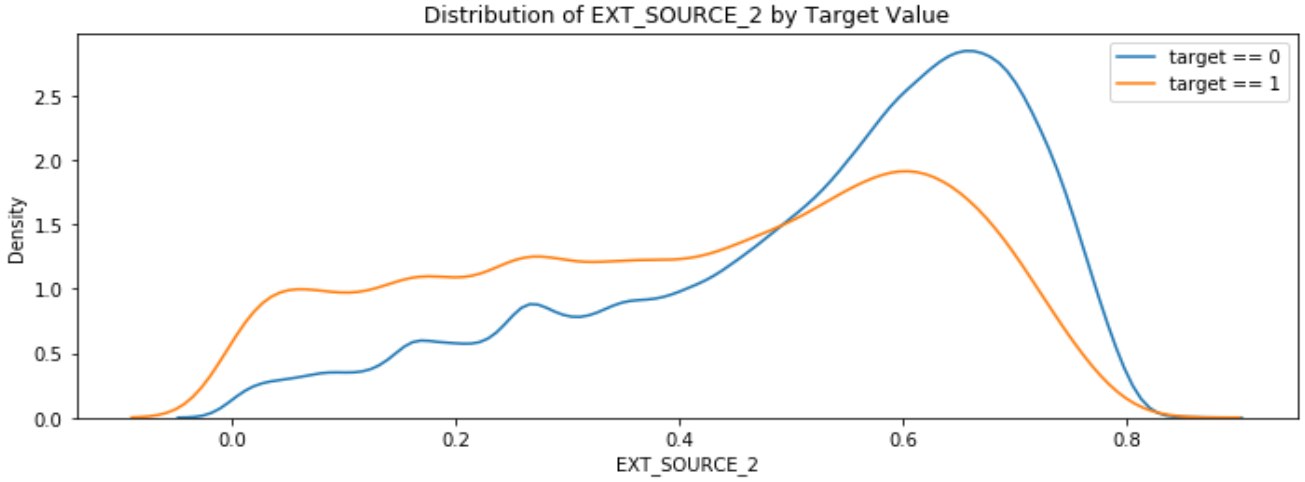
Also, a large number of applications were filed by clients who have house type mode as Block of flats.

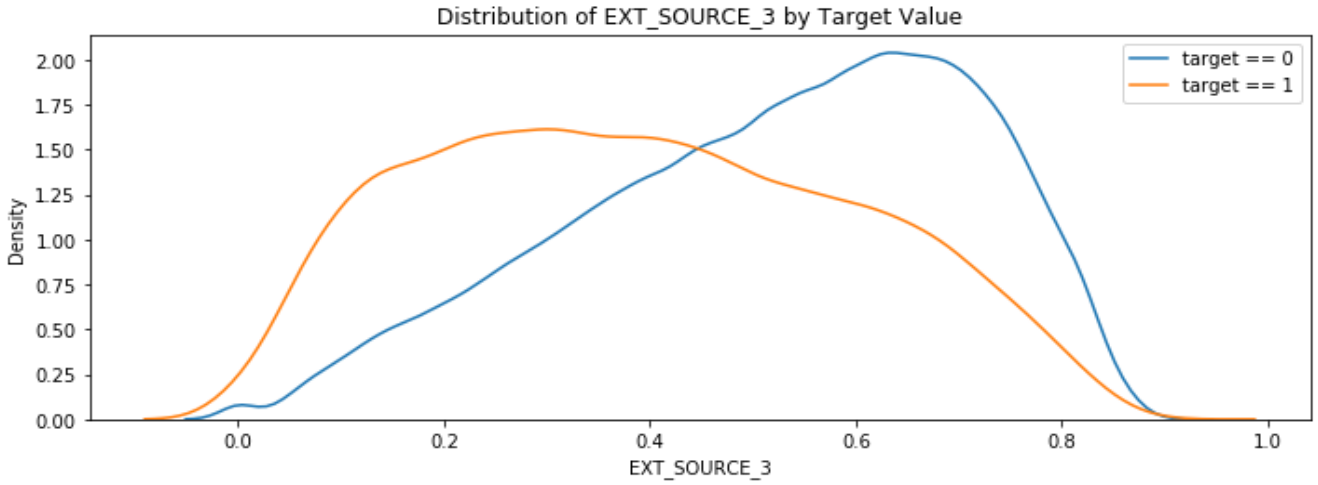
4.13 Visualize the distribution of External Sources vs. Target Variable:

All three EXT_SOURCE features have negative correlations with the target, which shows that as the value of the EXT_SOURCE increases, the client is more likely to repay the loan.

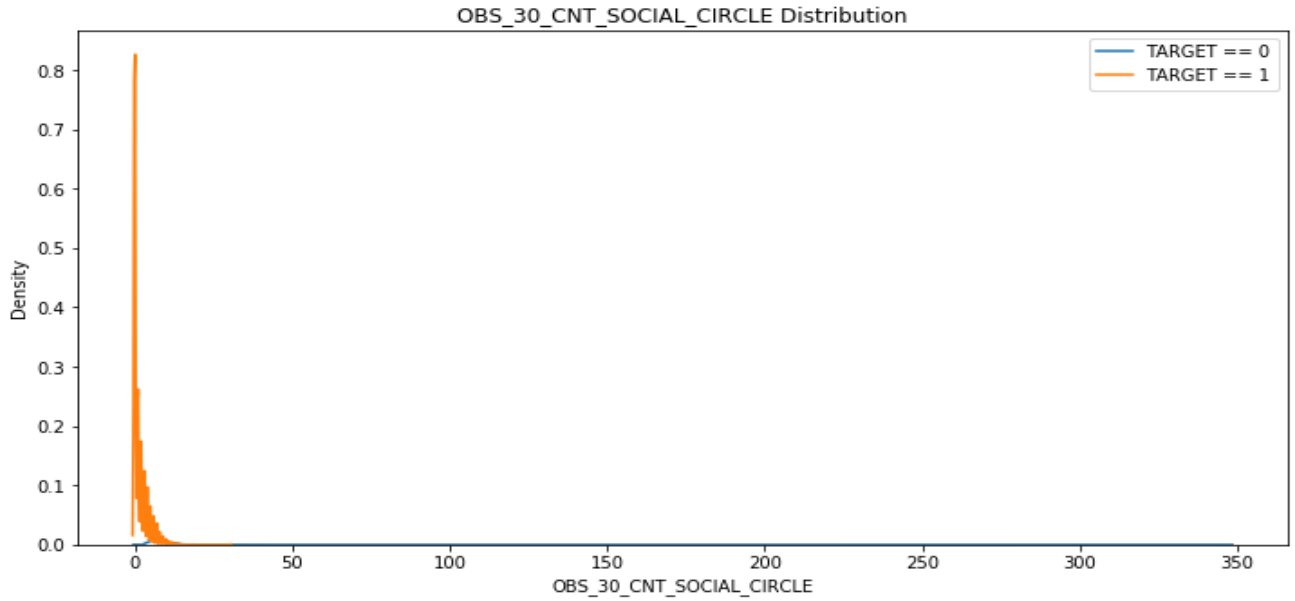
Next we can look at the distribution of these features colored by the value of the target. This will let us visualize the effect of this variable on the target.





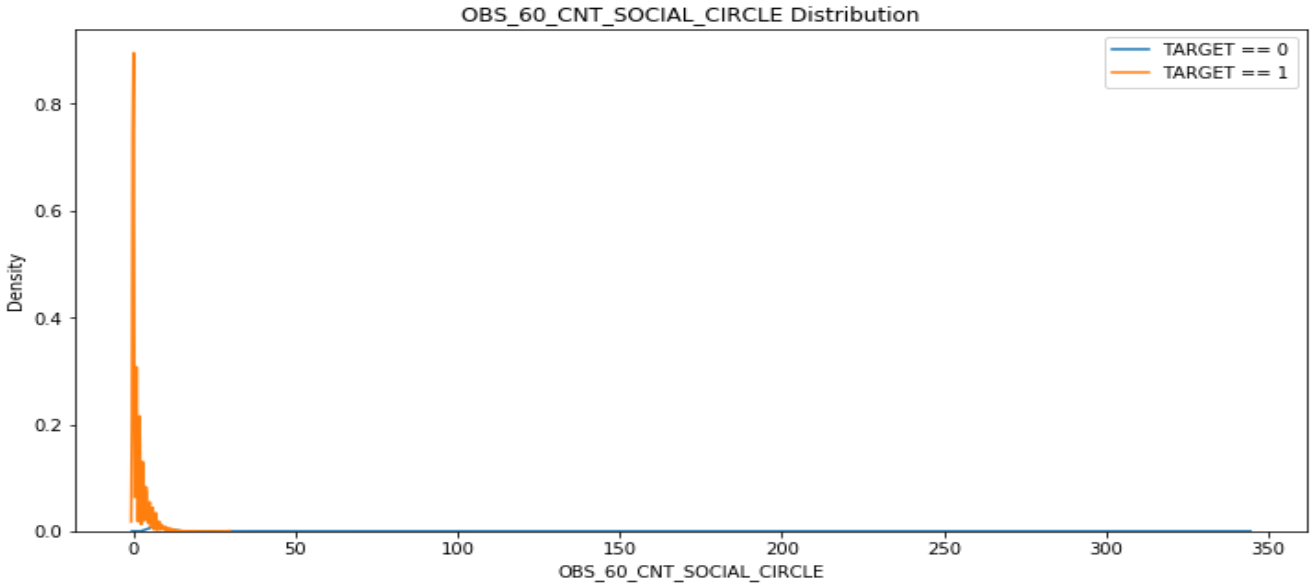


4.14 Visualize the distribution of 30 DPD in social circle vs. Target variable:



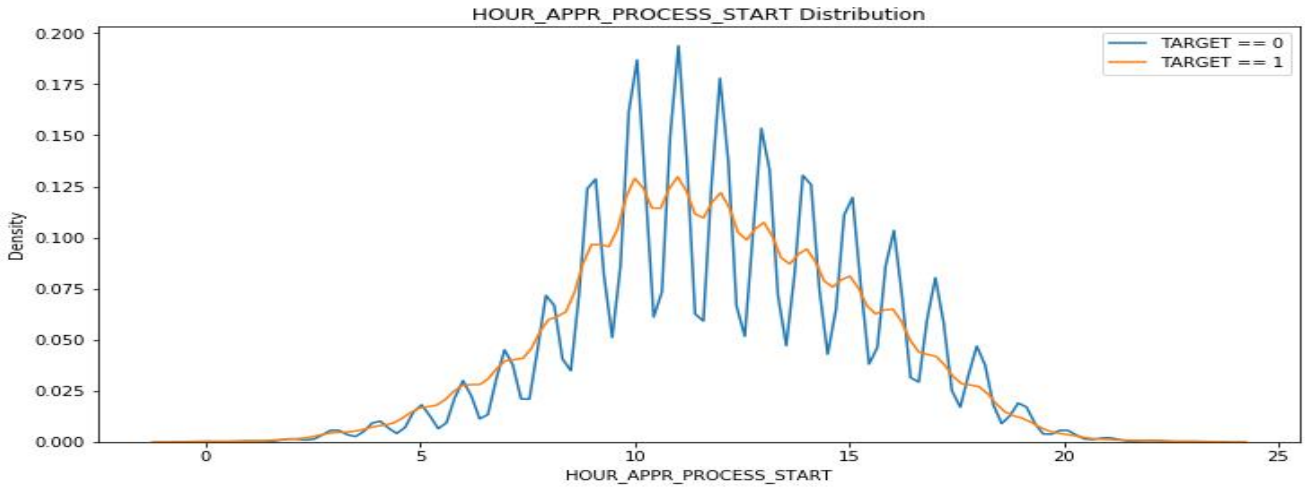
In social circle of customers, there are less than 15 people with 30 days past due, some are having very high number of people and this can be one of the attribute for financial institute to take a decision on customer.

4.15 Visualize the distribution of 60 DPD in social circle vs. Target variable:



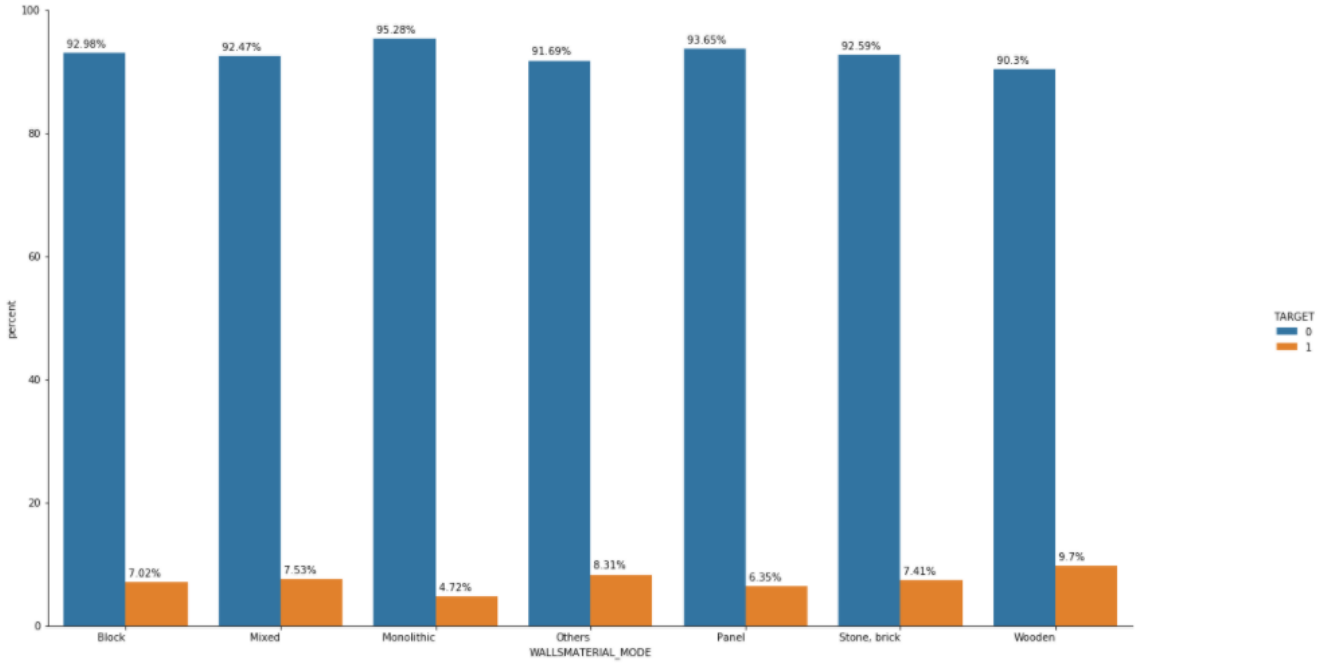
In social circle of customers, there are less than 15 people with 60 days past due, some are having very high number of people and this can be one of the attribute for financial institute to take a decision on customer.

4.16 Visualize the relationship of Application Process starting time vs. Target variable:



We observe that the most of the application process is starting between 10 A.M to 1 P.M and few are after 5 P.M.

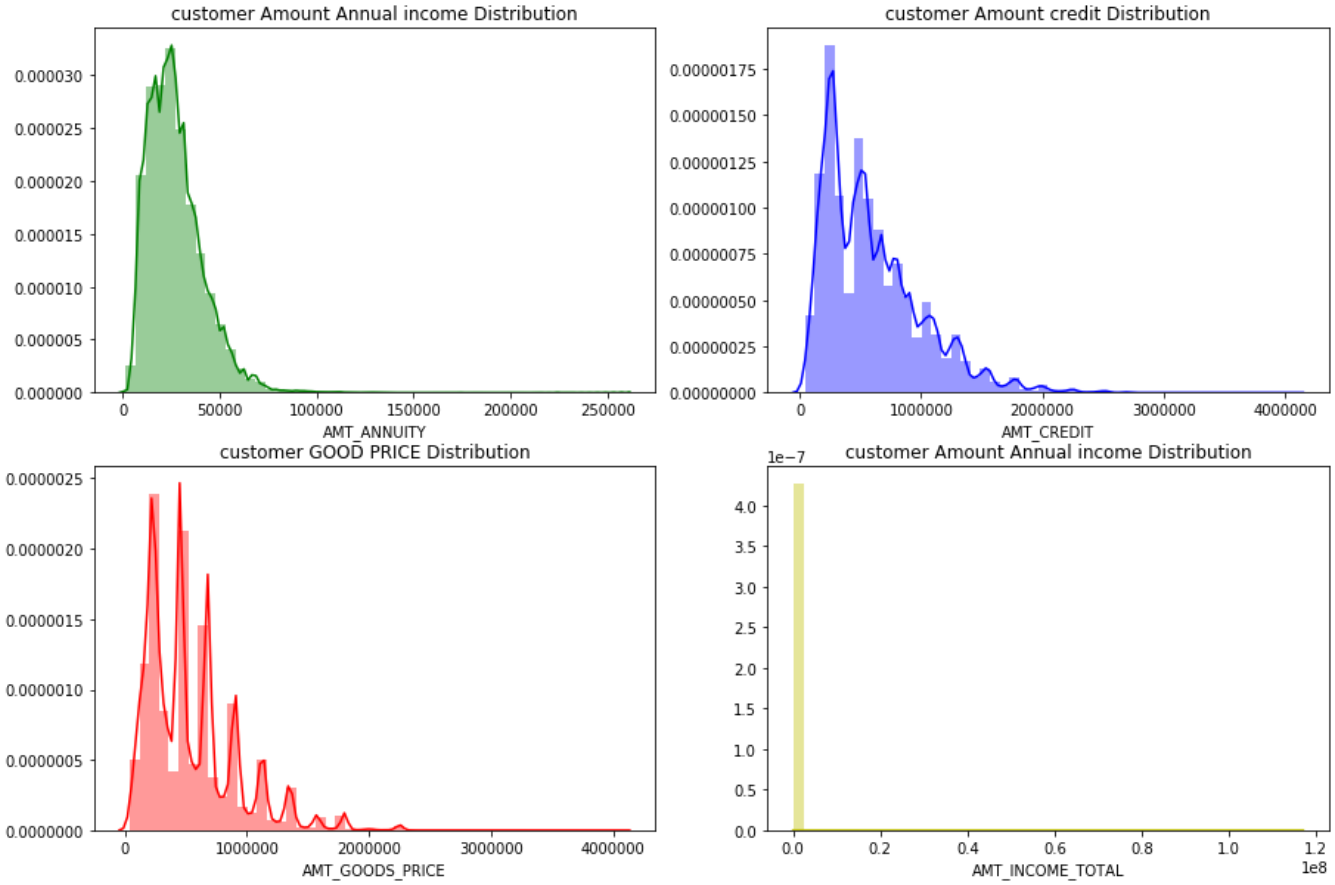
4.17 Visualize the relationship of Walls Material vs. Target Variable:



It is clear by looking at the above given plot that clients with walls material mode as Wooden housing were facing difficulties in repaying the loan.

Also, a large number of applications were filed by clients who have walls material mode as panel followed by those who have Stone,Brick as wallsmaterial mode.

4.18 Visualize the distribution of Annuity, Credit, Goods price, Income vs. Target Variable



Most of the customers are having an annuity of 50k with credit of around 10L and the goods price worth 10L.

Majority of the customers are having the income around 1-2 lakhs.

Part 5 Statistical Analysis:

Statistical tests were performed to see the whether the independent variables have a significant relationship with the dependent variable, TARGET.

5.1 Chi-square Test:

For the Categorical Columns, a Chi-square Test of independence was performed with the target variable, TARGET which is also a categorical column.

Null Hypothesis H0: There is NO association between the two variables.

Alternate Hypothesis Ha: There is an association between the two variables.

```
Columns Fails to Reject H0 : []

Columns Rejected H0 : ['NAME_CONTRACT_TYPE', 'CODE_GENDER', 'FLAG_OWN_CAR', 'FLAG_OWN_REALTY', 'NAME_TYPE_SUITE', 'NAME_INCOME_
TYPE', 'NAME_EDUCATION_TYPE', 'NAME_FAMILY_STATUS', 'NAME_HOUSING_TYPE', 'OCCUPATION_TYPE', 'WEEKDAY_APPR_PROCESS_START', 'ORGA
NIZATION_TYPE', 'FONDKAPREMONT_MODE', 'HOUSETYPE_MODE', 'WALLSMATERIAL_MODE', 'EMERGENCYSTATE_MODE']
```

Based on the above results it is observed that all the features p-values is less than 0.05 which indicates that all features are statistically significant with the target variable and helps in the prediction.

5.2 Two-sample t test:

For all the numeric variables, two-sample unpaired t tests were performed between values of the variable for two classes of target variables to compare their means.

Null Hypothesis H0: The means of the two samples are EQUAL.

Alternate Hypothesis Ha: The means of the two samples are NOT EQUAL.

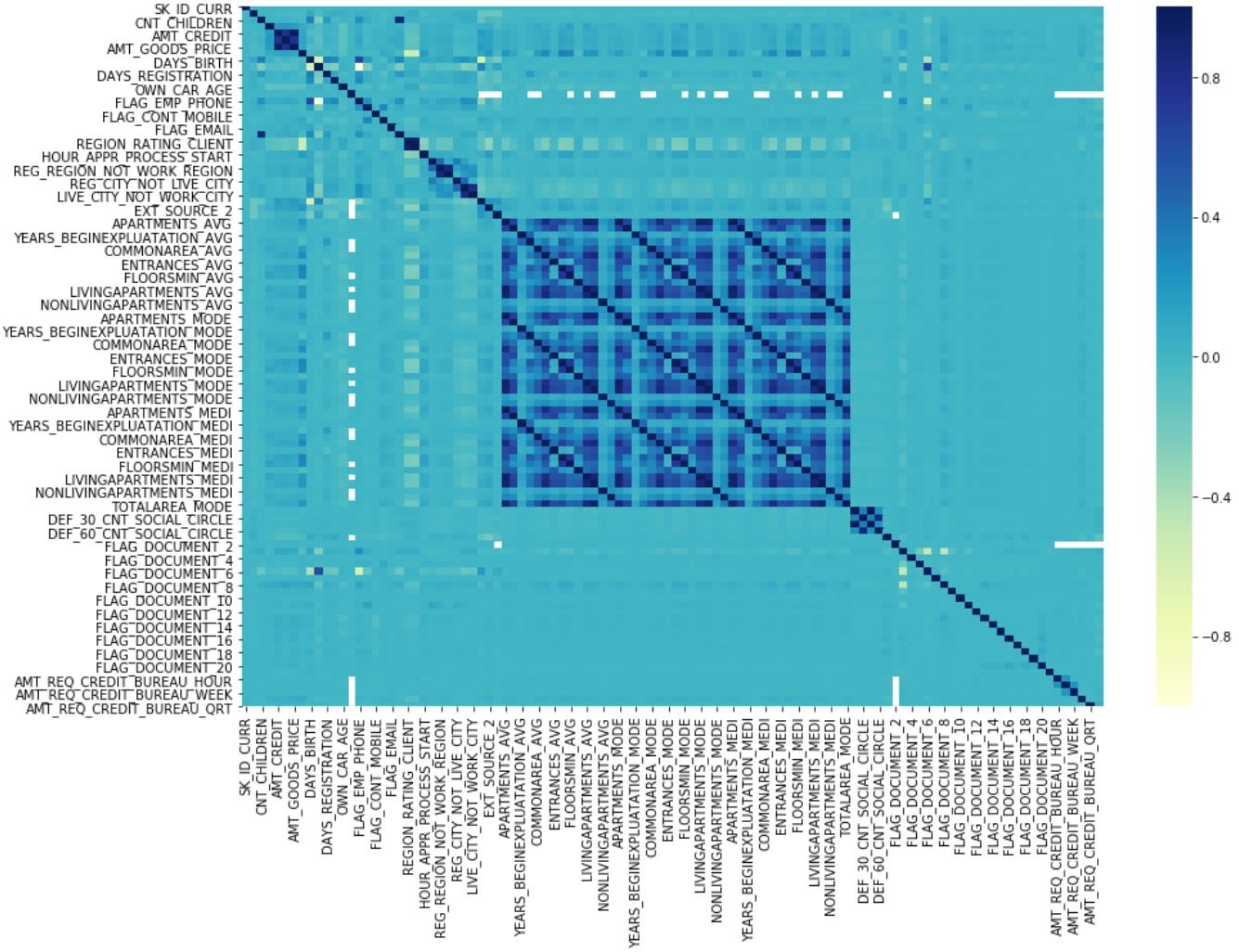
```
Columns Fails to reject H0 : []

Columns Reject H0 : ['CNT_CHILDREN', 'AMT_INCOME_TOTAL', 'AMT_CREDIT', 'AMT_ANNUITY', 'AMT_GOODS_PRICE', 'REGION_POPULATION_REL
ATIVE', 'DAYS_REGISTRATION', 'DAYS_ID_PUBLISH', 'OWN_CAR_AGE', 'FLAG_CONT_MOBILE', 'FLAG_PHONE', 'FLAG_EMAIL', 'CNT_FAM_MEMBER
S', 'REGION_RATING_CLIENT', 'REGION_RATING_CLIENT_W_CITY', 'HOUR_APPR_PROCESS_START', 'REG_REGION_NOT_LIVE_REGION', 'REG_REGION
_NOT_WORK_REGION', 'LIVE_REGION_NOT_WORK_REGION', 'REG_CITY_NOT_LIVE_CITY', 'REG_CITY_NOT_WORK_CITY', 'LIVE_CITY_NOT_WORK_CIT
Y', 'EXT_SOURCE_1', 'EXT_SOURCE_2', 'EXT_SOURCE_3', 'APARTMENTS_AVG', 'BASEMENTAREA_AVG', 'YEARS_BEGINEXPLUATATION_AVG', 'YEARS
_BUILD_AVG', 'COMMONAREA_AVG', 'ELEVATORS_AVG', 'ENTRANCES_AVG', 'FLOORSMAX_AVG', 'FLOORSMIN_AVG', 'LANDAREA_AVG', 'LIVINGAPART
MENTS_AVG', 'LIVINGAREA_AVG', 'NONLIVINGAPARTMENTS_AVG', 'NONLIVINGAREA_AVG', 'APARTMENTS_MODE', 'BASEMENTAREA_MODE', 'YEARS_BE
GINEXPLUATATION_MODE', 'YEARS_BUILD_MODE', 'COMMONAREA_MODE', 'ELEVATORS_MODE', 'ENTRANCES_MODE', 'FLOORSMAX_MODE', 'FLOORSMIN_
MODE', 'LANDAREA_MODE', 'LIVINGAPARTMENTS_MODE', 'LIVINGAREA_MODE', 'NONLIVINGAPARTMENTS_MODE', 'NONLIVINGAREA_MODE', 'APARTMEN
TS_MEDI', 'BASEMENTAREA_MEDI', 'YEARS_BEGINEXPLUATATION_MEDI', 'YEARS_BUILD_MEDI', 'COMMONAREA_MEDI', 'ELEVATORS_MEDI', 'ENTRAN
CES_MEDI', 'FLOORSMAX_MEDI', 'FLOORSMIN_MEDI', 'LANDAREA_MEDI', 'LIVINGAPARTMENTS_MEDI', 'LIVINGAREA_MEDI', 'NONLIVINGAPARTMENT
S_MEDI', 'NONLIVINGAREA_MEDI', 'TOTALAREA_MODE', 'OBS_30_CNT_SOCIAL_CIRCLE', 'DEF_30_CNT_SOCIAL_CIRCLE', 'OBS_60_CNT_SOCIAL_CIR
CLE', 'DEF_60_CNT_SOCIAL_CIRCLE', 'DAYS_LAST_PHONE_CHANGE', 'FLAG_DOCUMENT_15', 'YEARS_EMPLOYED', 'AGE', 'FLAG_DOCUMENT_2', 'FL
AG_DOCUMENT_3', 'FLAG_DOCUMENT_4', 'FLAG_DOCUMENT_5', 'FLAG_DOCUMENT_6', 'FLAG_DOCUMENT_7', 'FLAG_DOCUMENT_8', 'FLAG_DOCUMENT_
9', 'FLAG_DOCUMENT_10', 'FLAG_DOCUMENT_11', 'FLAG_DOCUMENT_12', 'FLAG_DOCUMENT_13', 'FLAG_DOCUMENT_14', 'FLAG_DOCUMENT_15', 'FL
AG_DOCUMENT_16', 'FLAG_DOCUMENT_17', 'FLAG_DOCUMENT_18', 'FLAG_DOCUMENT_19', 'FLAG_DOCUMENT_20', 'FLAG_DOCUMENT_21', 'AMT_REQ_C
REDIT_BUREAU_HOUR', 'AMT_REQ_CREDIT_BUREAU_DAY', 'AMT_REQ_CREDIT_BUREAU_WEEK', 'AMT_REQ_CREDIT_BUREAU_MON', 'AMT_REQ_CREDIT_BUR
EAU_QRT', 'AMT_REQ_CREDIT_BUREAU_YEAR', 'FLAG_MOBIL', 'FLAG_EMP_PHONE', 'FLAG_WORK_PHONE']
```

If the means of the two samples are significantly different form each other, and then we can conclude that the variable does have a significant relationship with the target variable.

Based on the above results it is observed that all the features p-values is less than 0.05 which indicates that all the features are statistically significant with the target variable and helps in the prediction of target variable .

5.3 Multicollinearity:



Though the correlation plots is not very clear to interpret, we can vaguely make a sense that there are few columns which are highly correlated as we can evidently see "Dark Blue" blocks and "light Green" blocks explaining positive and negative correlation in the Heat Map of Correlation plot.

Part 6 Machine Learning: Classification:

The main objective of this project is to identify the customers whose repayment records are good so that the bank will be able to efficiently filter them, saving time and resources. To achieve this objective, classification algorithms will be employed. By analyzing customer statistics, a classification model will be built to classify all clients into two groups: "0" for “No Delay in payments” and "1" for “Delay in payments”.

Prepare data for classification:

Dummy variables were used instead of continuous integers because these categorical variables are not ordinal. They simply represent different types rather than levels, so dummy variables are ideal to distinguish the effect of different categories. Feature selection: all customer statistics were selected as features while the Target feature was set as target. 70% of the data was used to build the classification model and 30% was reserved for testing the model.

Part 7 Evaluation Metrics:

The Evaluation Metrics that can be used for a Binary Classification problem are:

- Accuracy - Proportion of correctly identified instances
- Precision - proportion of positive predictions that are correct
- Recall - Proportion of Actual positives predicted correctly
- F1 Score - Harmonic mean of Precision and Recall
- ROC AUC - Area Under Receiver’s Operating Characteristics Curve (tradeoff between sensitivity and specificity for different thresholds)

Part 8 Base Model with Imbalanced target variable:

Base classification algorithms (Logistic Regression) were run on the dataset which was under sampled and identified the classification metrics such as Accuracy, Precision, Recall, and F1-score.

Metrics	Training Scores	Testing Scores
Accuracy	91.92	91.92
Precision	NA*	NA*
Recall	NA*	NA*
F1-Score	NA*	NA*
ROC AUC	49.99	50

*NA – It’s an edge case, model haven’t predicted any positive cases due to class imbalance.

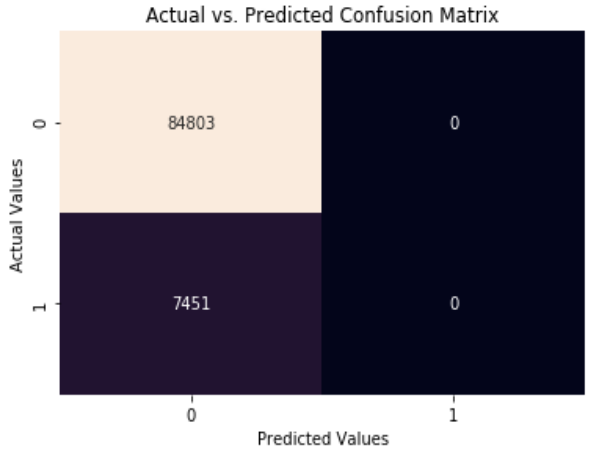
From the above results it is observed that Accuracy is good. This is due to imbalance in the target variable. So by doing oversampling there is a chance to significantly increase the classification metrics (precision, recall, and f1-score) along with accuracy.

Metrics:

```
Precision: nan
Recall: 0.0
roc_auc_Score: 0.5
f1_score: 0.0
accuracy_score: 0.9192338543586186
```

Nan is the output for precision since there is class imbalance in dataset.

Confusion matrix:

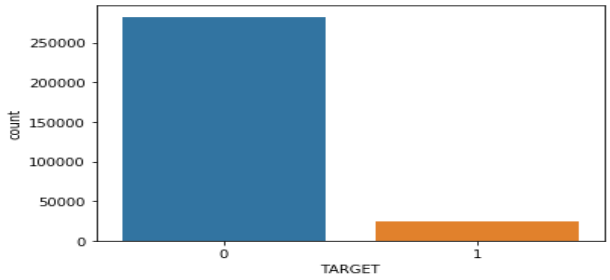


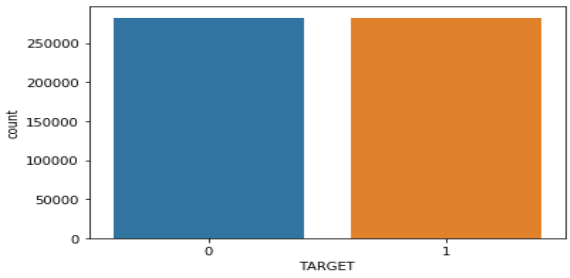
Above picture displays confusion matrix for test data, we can observe that the True Positives and False Positives are zero because the data set is imbalanced.

8.1 Oversampling the target variable by using SMOTE:

Classification using class-imbalanced data is biased in favor of the majority class. Over sampling is the process of converting the minority class into majority class. i.e., increase the minority class to majority class count.

The Synthetic Minority Over-sampling Technique (SMOTE) is an oversampling approach that creates synthetic minority class samples. It potentially performs better than simple oversampling and it is widely used.





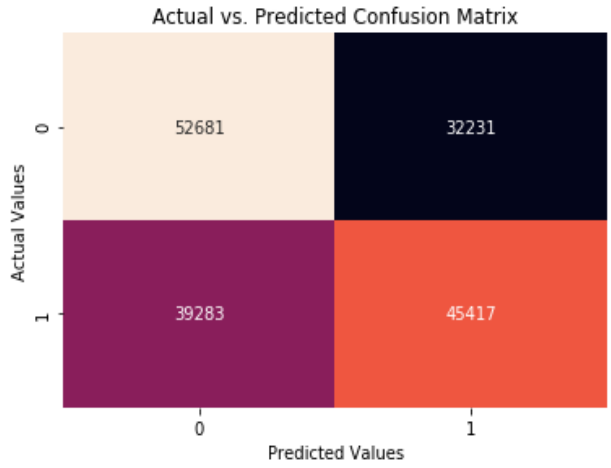
Part 9 Base Model with oversampled data (SMOTE):

Base classification algorithms (Logistic Regression) were run on the dataset which is balanced by doing over sampling and identified the classification metrics such as Accuracy, Precision, Recall, and F1-score.

Metrics	Training Scores	Testing Scores
Accuracy	57.74	57.83
Precision	58.47	58.48
Recall	53.63	53.63
F1-Score	55.94	55.95
ROC AUC	57.75	57.82

From the above results it is observed that Accuracy, precision, recall and f1-score are present. This is due to generating balance in the target variable by using smote technique. So by doing oversampling the data there is an increase in the classification metrics (precision, recall, and f1-score).

Confusion Matrix:



Here we can see that there are TP and FP in the confusion matrix after oversampling.

Metrics:

```
Precision: 0.584908819287039
Recall: 0.5362101534828807
roc_auc_Score: 0.578314469995633
f1_score: 0.5595018109246802
accuracy_score: 0.5783670966676886
```

After performing over-sampling, we were successfully able to overcome edge case of Zero TP's and Zero FP's and have some valid values in confusion matrix with balanced data. This can further be improved with model fine tuning.

Part 10 Different Approaches:

Different Approaches:

Approach 1:

Min-max scaler for non-normalized features, get dummies for all features and feature integration.

Metrics/Model	Logistic Regression		Decision Tree	
	Train Scores	Test Scores	Train Scores	Test Scores
Accuracy	70.55	70.88	100	89.67
Precision	70.22	70.53	100	88.76
Recall	71.43	71.59	100	90.81
f1-score	70.82	71.06	100	89.78
Roc-Auc Score	70.55	70.88	100	89.67

Approach 2:

Label encoder for multi-class variables and get dummies for binary variables, with feature engineering and Standard Scaler.

Metrics/Model	Logistic Regression		Decision Tree	
	Train Scores	Test Scores	Train Scores	Test Scores
Accuracy	69.71	69.65	100	88.59
Precision	69.48	69.41	100	87.43
Recall	70.35	70.13	100	90.10
f1-score	69.91	69.77	100	88.74
Roc-Auc Score	69.71	69.65	100	88.59

Part 11 Building Best Model:

Till the base model, we performed get dummies and done some feature engineering which gave us ROC_AUC score around 70% (considering Logistic Regression only).

Since ROC_AUC score is considered as the best metrics in industry, we took that as our important metrics to be considered amongst Accuracy and F1_score. We prepared our data in the following manner for model building.

Data Preparation:

11.1 Outlier Treatment:

We observed there is necessity to treat the outliers and proceed for the model building. Considering that, we used capping method for outlier treatment.

Q1 = First quartile

Q2 = Second quartile

IQR = inter quartile range = Q3 - Q1

Lower bound = Q1 - 1.5(IQR)

Upper bound = Q3 + 1.5(IQR)

We have substituted all the values above the upper bound with the upper bound and all the values below the lower bound with the lower bound.

11.2 Feature Integration:

There are 122 columns in dataset. After building base model, we integrated some of the columns in below fashion.

Flag-Documents:

There are total 20 Flag-document columns, which gives the information whether the customer submitted the required documents or not.

Merging has been done for these 20 columns by adding row - wise to get the information in single feature.

Education-Type:

Education type feature has been manually given labels because higher education type should have higher value.

Enquiries:

There are six columns in the dataset, which say how many enquiries have been done with respect to the client from the past year till the past hour of the application. Merging has been done for these six columns by adding row - wise to get the information in single feature.

Phone:

There are three columns which say whether the customer has provided work phone, house phone and personal phone numbers. Merged these columns by adding row - wise to see how many phone numbers he submitted out of the above three.

11.3 Treating null values:

As there are 25% of null values in the whole dataset, we treated them as follows:

Numerical columns:

There are many methods for imputing missing values like KNN imputer, mean imputation, median imputation.

KNN imputer is distance based method for filling null values. Since we are having 3L rows before oversampling and around 6L rows after oversampling, it was computationally difficult for us to go with this method.

Mean imputation can be effected by outliers while median imputation can't be effected by outliers. So we chose median imputation for numerical columns.

Categorical Columns:

We applied mode imputation for missing categorical values.

Get-Dummies:

Other than Education type of customer, all other categorical columns are converted into dummy variables by using get_dummies.

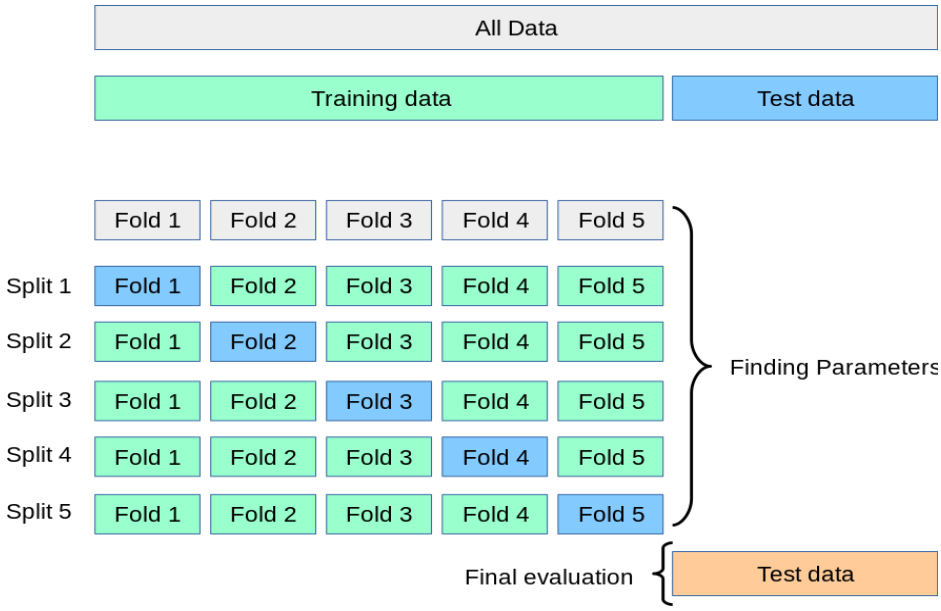
By performing all the above cleaning and feature integration, size of the dataset is (307509, 198)

After performing oversampling method (SMOTE) the size of dataset is (565370, 196)

Cross-Validation:

Since train test split doesn't ensure that all the records of the data are used for training of the model, we used cross validation which will ensure each record is used for training the model.

K fold cross validation splits the dataset into K folds and each time it uses separate set for the testing purpose and uses remaining sets for training purpose.



11.4 PCA model:

Since we observe multi collinearity between the dependent variables in the dataset, we have tried PCA which can help in reducing the redundancy in the data set**.** When we choose to analyze our data using PCA, part of the process involves checking to make sure that the data we want to analyze can actually be analyzed using PCA. We need to do this because it is only appropriate to use PCA if our data "passes" assumptions that are required for PCA to give us valid results.

1) Linear relationship between variables within data i.e. presence of Multi-collinearity in data
2) We should have sampling adequacy, which simply means that for PCA to produce a reliable result, large enough sample sizes are required
3) Our data should be suitable for data reduction

Model Building:

Since the data is huge, we tried PCA to see whether this can improve our model performance.

Logistic Regression	Train Scores	Test Scores
Accuracy	94.28	94.30
Precision	97.29	97.30
Recall	91.09	91.16
F1_score	94.09	94.13
Roc_Auc score	94.28	94.31

From the above results it is observed that Accuracy, precision, recall and f1-score are increased when compared with base model.

After performing Cross Validation for PCA model with CV=5 and scoring = 'roc_auc', below are the results.

```
[0.97176648 0.97093312 0.97008974 0.97173741 0.96994854]
Bias_error: 0.029104944712872283
VE: 0.0008681624871695027
```

Although we got good results, we want to try with "Select KBEST" method for feature selection and proceed with model building because after implementing PCA on the dataset, our original features will turn into Principal Components. Principal Components are the linear combination of our original features. Principal Components are not as readable and interpretable as original features.

11.5 Univariate Selection:

Univariate Feature Selection is a statistical method used to select the features which have the strongest relationship with our correspondent labels. Using the SelectKBest method we can decide which metrics to use to evaluate our features and the number of K best features we want to keep.

Different types of scoring functions are available depending on our needs:

- Classification = chi2, f_classif, mutual_info_classif
- Regression = f_regression, mutual_info_regression

When performed statistical tests such as Chi square and t-test for the dataset, the result of test shown that all the features are significant with respect to target variable. In SelectKBest method we specify the number of features and the method returns the most significant amongst them.

We had number of trials with k=80, 90,100 and so on. We got good score for k=100 value. Although the scores will be more if we go beyond k=120, but there would be much variance error in the scores. So choosing optimal k value can be done only through trial and error method.

SelectKBest:

For k=100:

Logistic Regression	Train Scores	Test Scores
Accuracy	93.92	93.99
Precision	96.90	97.08
Recall	90.75	90.71
F1_score	93.72	93.78
Roc_Auc score	93.92	93.99

After performing Cross Validation for the model with CV=5 and scoring = 'roc_auc', below are the results.

```
[0.97050352 0.96949795 0.96883082 0.97040282 0.9684298 ]
Bias_error: 0.030467017435086063
VE: 0.0009232839447959737
```

As our scores are almost same as PCA, here we are able to interpret the features by backing them with statistical analysis. So going further we tried different ensemble methods on KBest model to see if we can improve the score.

11.6 Ensemble Methods:

Random Forest:

Random Forest	Train Scores	Test Scores
Accuracy	99.99	95.02
Precision	100	99.48
Recall	99.99	90.52
F1_score	99.99	94.79
Roc_Auc score	99.99	95.02

After performing Cross Validation for the model with CV=5 and scoring = 'roc_auc', below are the results.

```
[0.97647807 0.97591306 0.97501687 0.97632086 0.97471726]
Bias_error: 0.024310777138270567
VE: 0.0007855131847472758
```

AdaBoost Classifier:

AdaBoost Classifier	Train Scores	Test Scores
Accuracy	91.74	91.76
Precision	100	92.60
Recall	99.99	90.77
F1_score	91.67	91.68
Roc_Auc score	91.74	91.76

After performing Cross Validation for the model with CV=5 and scoring = 'roc_auc', below are the results.

```
[0.96245328 0.96257487 0.96085019 0.96350985 0.96087502]
Bias_error: 0.03794735744251487
VE: 0.0011607509466992076
```

Bagging Classifier:

Bagging Classifier	Train Scores	Test Scores
Accuracy	99.31	93.45
Precision	99.97	97.69
Recall	98.64	89.01
F1_score	99.30	93.15
Roc_Auc score	99.31	93.45

After performing Cross Validation for the model with CV=5 and scoring = 'roc_auc', below are the results.

```
[0.96665783 0.96567886 0.96454217 0.96577547 0.96401803]
Bias_error: 0.03466552724531158
VE: 0.0010517983445753076
```

Gradient Boosting:

Gradient Boosting	Train Scores	Test Scores
Accuracy	93.48	93.48
Precision	96.65	96.67
Recall	90.07	90.05
F1_score	93.25	93.24
Roc_Auc score	93.48	93.48

After performing Cross Validation for the model with CV=5 and scoring = 'roc_auc', below are the results.

```
[0.9682385  0.96759655 0.96617713 0.96847877 0.96602999]
Bias_error: 0.03269581092801255
VE: 0.0011436670662469627
```

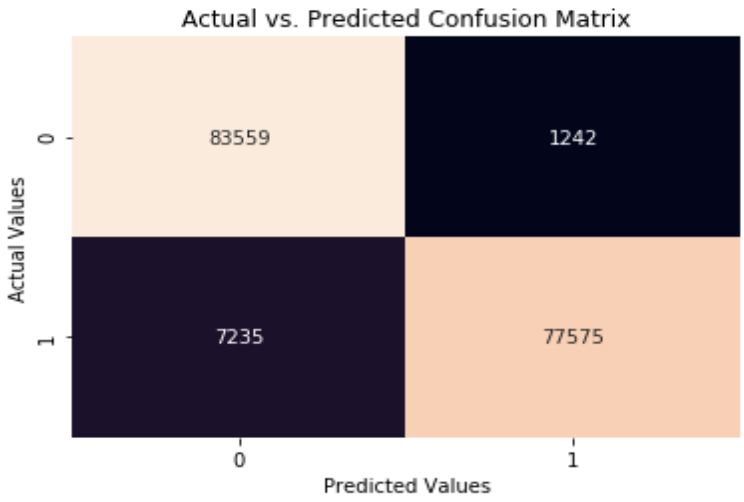
XGBC:

XGBC	Train Scores	Test Scores
Accuracy	95.42	95.00
Precision	98.85	98.42
Recall	91.90	91.46
F1_score	95.25	94.81
Roc_Auc score	95.42	95.00

After performing Cross Validation for the model with CV=5 and scoring = 'roc_auc', below are the results.

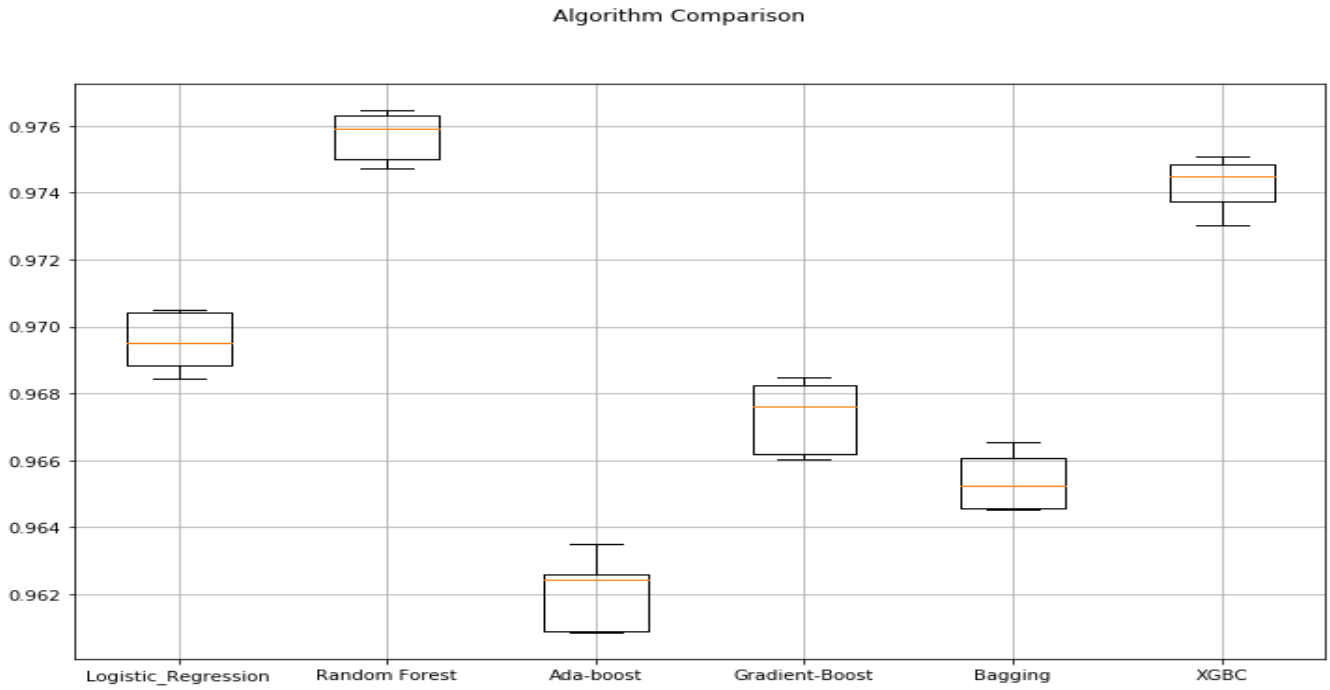
```
[0.97509985 0.97449464 0.97372718 0.97484438 0.97302458]
Bias_error: 0.025761873585473438
VE: 0.00085288686631961
```

Confusion Matrix:



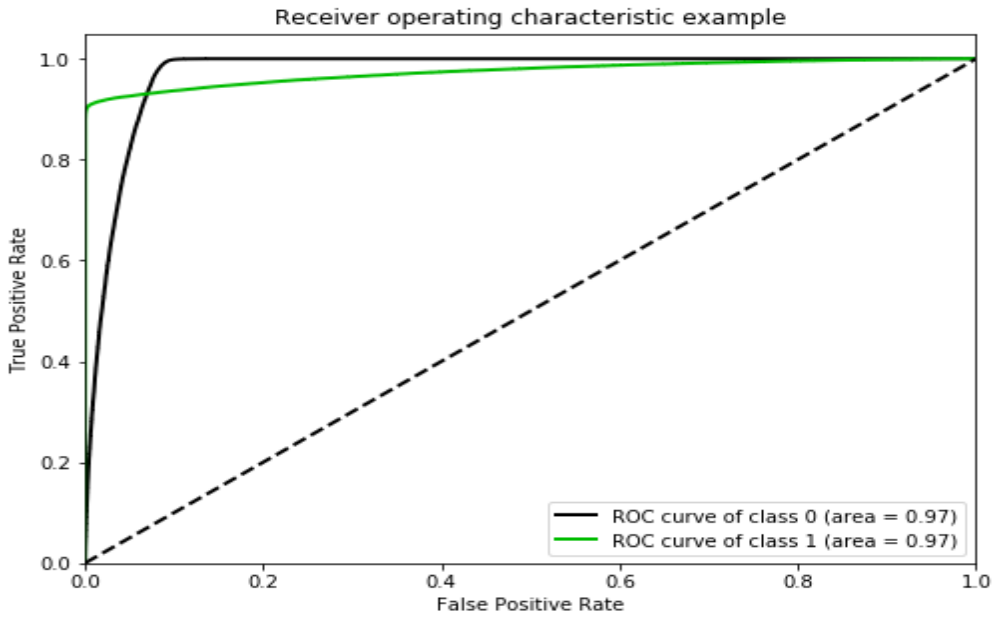
11.7 Box Plot Comparison:

As we built these models, scores are almost equal. In order to find out best model we can go with boxplot comparison for all the models.



From the box plot comparison diagram, we can see that Random Forest and XGBC are better models amongst all of the models. But when we observe train and test scores of Random Forest and XGBC, we can see that XGBC model is performing better on both train and test data. So we are finalizing best model as XGBC model.

11.8 Roc_Auc curve for XGBC:



Part 12 Assumptions:

Since the records of the dataset comprises of the data from different countries, we assumed that the currency of all the monetary features is in same units.

Part 13 Business Insights and Summary:

Business Insights:

- There are some columns stating the rating of the regions with labels as 1, 2, 3 and the highest rating is for 2. So, the focus should be on the regions with rating 1 and this can be achieved by analyzing why the rating of that particular region is low and in these regions, the product can be marketed by offering additional benefits.
- Company is focusing more on cash loans. It also provides revolving loans, but customers who applied for revolving loan accounts are significantly less. This can be considered as better option for company to market revolving loans to customers by conducting more marketing activities. Also company can have brainstorming sessions with employees regarding this. As they are the front end people who will be the barriers between company and customer, they understand market well and they can give inputs regarding improvement in business.
- As there are more defaulters in civil marriage cases, when a new application gets registered with family type as civil marriage, the bank should be vigilant.
- Most of the people who live in apartments and cooperative apartments generally are able to repay the loan on time. Better schemes and awareness should be created among the people who live in single and independent houses that would facilitate them to repay the loan on time.
- Clients with education type as academic degree have very less default rate compared to the clients with other types of education. Also, the average income is higher for the clients who have education type as Academic degree. More awareness should be created for the clients who have lower education types and may be more knowledge should be provided about the perks of repaying the loan on time.

Summary:

- As the target variable of the dataset is highly imbalanced, base model has been fit with both with and without sampling of the data to see the effect of sampling on the performance of the model.
- As the count of features are 122 which is considered as high, visualisations have been done for the features which are significantly effecting the class separation of the target.
- Statistical tests (chi square for qualitative features and two-sample t test for quantitative features) have been performed to select the features which are statistically significant.
- Feature integration has been done for few features which are feasible to be combined into one feature.
- There are few normalised columns in the dataset, performed normalisation for the remaining numerical features, performed one hot encoding for all the categorical features as our 1st approach to see if there is improvement in the performance of the model.
- Further, applied standard scaler for all the features, label encoding for multi class qualitative features, one hot encoding for binary qualitative features to see if this approach is helping in increasing the performance of the model.
- As another approach manually encoded education type feature (replaced higher education type with higher value) and applied get dummies on all the remaining features. Applied the same feature integration and Standard Scaler.
- On the above approach applied PCA and Best feature selection methods separately. Got good scores after applying feature selection methods than PCA.
- Applied ensemble methods on the model built with KBest feature selection method and performed cross validation for each model.

- Compared the performance of all the models built through K Best feature selection method and observed XB boost has good scores with low variance error.

Part 14 Limitations:

- As our data set volume is large, we were confined to few imputation methods but not advanced methods like KNN imputer because of computational difficulties.
- We were able to perform only with 5 folds as dataset is huge and this process also took hours to execute.

Part 15 Implications:

The solution in the project determine whether the customer may/may not repay the loan amount based on several factors like 'Age', 'Gender', 'Education type', 'Occupation type', 'Annual Income', 'Amount annuity', 'region', 'Organization type', 'Type of loan', 'Type of House', 'Marital Status', 'Work experience', 'Family status', 'Family count', 'Children count', 'Social circle count', 'wall material' . With 95 % confidence, our models can predict the credit risk of Home credit group with bias error of 2.57% and accuracy of 95%.

Part 16 Conclusion:

Credit risk assessment is only possible by means of measurement. Machine learning models can be used as tools to measure the credit risk exposure of various financial institutions. With the correct prediction of credit risk, its management will become effective and efficient. This project work concentrates on evaluation of different machine learning classifier models to predict the credit risks associated with various borrowers of an institution. For this the major assessment parameters of the institution are taken as the predictor variables. There are many classifier models we have approached which are discussed in the report. We can say conclusively that XGBoost is the model that performed well in our project. On the other hand different statistical techniques like chi square test, 2 sample t-tests etc. are performed to determine the important features. However we have also tried KBest technique to determine the feature importance. Feature integration has also been implemented because of its high dimensionality. Normalization, one hot encoding and standard scalar methods have been executed to check the improvement of the model performance. KBest feature selection was enacted to check the important features and PCA was applied on the data because of its high dimensionality. Bagging and Boosting methods has been tried to check whether it improves the score or not. After comparing the performance of all the models it is concluded that XGBoost is the best model which is performing well. This project opens the doors for further research in the credit risk using the deep learning models and there is a scope of smoothing the process of decision making in credit risk.