Batch details	PGPDSE-FT Offline BLR April22
Team members	Swarnashree Taniya C Mathews Pratiksha Patil Nawaz S R Praveen
Domain of Project	Finance And Risk Analytics
Proposed project title	Detection Of Loan Defaulters
Group Number	6
Team Leader	Swarnashree
Mentor Name	Ms. Vidhya K

Date: 13-August-2022

Ms Vidhya K

Signature of the Mentor

Signature of the Team Leader

TABLE OF CONTENTS

INDUSTRY REVIEW

The banking industry includes systems of financial institutions called banks that help people store and use their money. Banks offer clients the opportunity to open accounts for different purposes, like saving or investing their money.

A possible effect of loan defaults is on shareholders' earnings. Dividend payments are based on the bank's performance in terms of net profit. Thus, since loan defaults have an adverse effect on the profitability of banks; it can affect the number of dividends to be paid to shareholders.

Why do People Take Loans? Why does Lending exist?

- Many individuals utilize debt to pay for things they wouldn't be able to buy otherwise, such as a home or a vehicle.
- Lending is a vital tool that propels all enterprises and individuals worldwide to greater financial success.
- In recent years, there has been an increase in loan defaults, which has already begun to affect the bottom lines of several financial institutions.

Major reasons for loan default:

- A secured debt default can occur, such as with a mortgage loan secured by a property or a business loan secured by the company's assets. If you do not make your mortgage payments on time, the loan may default.
- If a corporation issues bonds (essentially borrows money from investors) and cannot fulfil coupon payments to bondholders, the company is in default.

Predicting loan defaults has become important as banks try to follow laws and regulations, grant credits to qualified customers, mitigate credits to unqualified customers and to make their application processes efficient. This research studies credit risk in banking, discusses banking regulations which affect loan granting and presents how machine learning is utilized in lending. In addition, the literature review explains machine learning and the steps in building machine learning models. The empirical study is conducted with a loan data set retrieved from hackerearth.com.

OBJECTIVES

- Study The Bank Indessa data and gather insights about the business.
- Predicting every defaulter with high accuracy and precision.
- Estimating the grounds for granting a loan, such that the loan defaulters can be decreased.
- To strengthen the banking ecosystem of The Bank Indessa &
- Strengthen the loan sanctioning thereby enhancing the performance of the bank which is not doing well since the past three quarters.

DATASET AND DOMAIN

- Data is collected from hackerearth.com. There are two datasets, train_indessa.csv and valid_indessa.csv.

- The number of records in the training set is 372,699 records and the testing set is 159,729 records.

- Total number of Numerical columns - 27
- Total number of Categorical columns - 18

FEATURES UNDERSTANDING

Feature Name	Description
member_id	unique ID assigned to each member
loan_amnt	loan amount ($) applied by the member
funded_amnt	loan amount ($) sanctioned by the bank
funded_amnt_inv	loan amount ($) sanctioned by the investors
term	term of loan (in months)
batch_enrolled	batch numbers allotted to members
int_rate	interest rate (%) on loan
grade	grade assigned by the bank
sub_grade	grade assigned by the bank
emp_title	job / Employer title of member
emp_length	employment length, where 0 means less than one year and 10 means ten or more years
home_ownership	status of home ownership
annual_inc	annual income ($) reported by the member
verification_status	status of income verified by the bank
pymnt_plan	indicates if any payment plan has started against loan
desc	loan description provided by member
purpose	purpose of loan
title	loan title provided by member
zip_code	first three digits of area zip code of member
addr_state	living state of member
dti	ratio of member's total monthly debt repayment excluding mortgage divided by self-reported monthly income
delinq_2yrs	number of 30+ days delinquency in past 2 years
inq_last_6mths	number of inquiries in last 6 months
mths_since_last_delinq	number of months since last delinq
mths_since_last_record	number of months since last public record
open_acc	number of open credit line in member's credit line
pub_rec	number of derogatory public records
revol_bal	total credit revolving balance
revol_util	amount of credit a member is using relative to revol_bal
total_acc	total number of credit lines available in members credit line
initial_list_status	unique listing status of the loan - W(Waiting), F(Forwarded)
total_rec_int	Interest received till date
total_rec_late_fee	Late fee received till date
recoveries	post charge off gross recovery
collection_recovery_fee	post charge off collection fee
collections_12_mths_ex_med	number of collections in last 12 months excluding medical collections
mths_since_last_major_derog	months since most recent 90 day or worse rating
application_type	indicates when the member is an individual or joint
verification_status_joint	indicates if the joint members income was verified by the bank
last_week_pay	indicates if the joint members income was verified by the bank
acc_now_delinq	number of accounts on which the member is delinquent
tot_coll_amt	total collection amount ever owed
tot_cur_bal	total current balance of all accounts
total_rev_hi_lim	total revolving credit limit
loan_status	status of loan amount, 1 = Defaulter, 0 = non-Defaulters

DATA EXPLORATION (EDA)

DATATYPES INFO

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 532428 entries, 0 to 532427
Data columns (total 45 columns):
 #   Column                       Non-Null Count   Dtype  
---  ------                       --------------   -----  
 0   member_id                    532428 non-null  int64  
 1   loan_amnt                    532428 non-null  int64  
 2   funded_amnt                  532428 non-null  int64  
 3   funded_amnt_inv              532428 non-null  float64
 4   term                         532428 non-null  object 
 5   batch_enrolled               447279 non-null  object 
 6   int_rate                     532428 non-null  float64
 7   grade                        532428 non-null  object 
 8   sub_grade                    532428 non-null  object 
 9   emp_title                    501595 non-null  object 
 10  emp_length                   505537 non-null  object 
 11  home_ownership               532428 non-null  object 
 12  annual_inc                   532425 non-null  float64
 13  verification_status          532428 non-null  object 
 14  pymnt_plan                   532428 non-null  object 
 15  desc                         75599 non-null   object 
 16  purpose                      532428 non-null  object 
 17  title                        532338 non-null  object 
 18  zip_code                     532428 non-null  object 
 19  addr_state                   532428 non-null  object 
 20  dti                          532428 non-null  float64
 21  delinq_2yrs                  532412 non-null  float64
 22  inq_last_6mths               532412 non-null  float64
 23  mths_since_last_delinq       259874 non-null  float64
 24  mths_since_last_record       82123 non-null   float64
 25  open_acc                     532412 non-null  float64
 26  pub_rec                      532412 non-null  float64
 27  revol_bal                    532428 non-null  float64
 28  revol_util                   532141 non-null  float64
 29  total_acc                    532412 non-null  float64
 30  initial_list_status          532428 non-null  object 
 31  total_rec_int                532428 non-null  float64
 32  total_rec_late_fee           532428 non-null  float64
 33  recoveries                   532428 non-null  float64
 34  collection_recovery_fee      532428 non-null  float64
 35  collections_12_mths_ex_med   532333 non-null  float64
 36  mths_since_last_major_derog  132980 non-null  float64
 37  application_type             532428 non-null  object 
 38  verification_status_joint    305 non-null     object 
 39  last_week_pay                532428 non-null  object 
 40  acc_now_delinq               532412 non-null  float64
 41  tot_coll_amt                 490424 non-null  float64
 42  tot_cur_bal                  490424 non-null  float64
 43  total_rev_hi_lim             490424 non-null  float64
 44  loan_status                  532428 non-null  int64  
dtypes: float64(23), int64(4), object(18)
memory usage: 182.8+ MB
```

INFERENCE

The Dataset consists of int, float and object data types and also many features consist of null values.

DESCRIPTION OF NUMERICAL FEATURES

```
df.describe().T
```

	count	mean	std	min	25%	50%	75%	max
member_id	532428.000000	35005472.347129	24121476.515915	70473.000000	10866882.500000	37095895.000000	58489200.750000	73544841.00
loan_amnt	532428.000000	14757.595722	8434.420080	500.000000	8000.000000	13000.000000	20000.000000	35000.00
funded_amnt	532428.000000	14744.271291	8429.139277	500.000000	8000.000000	13000.000000	20000.000000	35000.00
funded_amnt_inv	532428.000000	14704.926696	8441.290381	0.000000	8000.000000	13000.000000	20000.000000	35000.00
int_rate	532428.000000	13.242969	4.379611	5.320000	9.990000	12.990000	16.200000	28.99
annual_inc	532425.000000	75029.843289	65199.845014	1200.000000	45000.000000	65000.000000	90000.000000	9500000.00
dti	532428.000000	18.138767	8.369074	0.000000	11.930000	17.650000	23.950000	672.52
delinq_2yrs	532412.000000	0.314448	0.860045	0.000000	0.000000	0.000000	0.000000	30.00
inq_last_6mths	532412.000000	0.694603	0.997025	0.000000	0.000000	0.000000	1.000000	31.00
mths_since_last_delinq	259874.000000	34.055735	21.884797	0.000000	15.000000	31.000000	50.000000	180.00
mths_since_last_record	82123.000000	70.093068	28.139219	0.000000	51.000000	70.000000	92.000000	121.00
open_acc	532412.000000	11.545594	5.311442	0.000000	8.000000	11.000000	14.000000	90.00
pub_rec	532412.000000	0.194858	0.583822	0.000000	0.000000	0.000000	0.000000	86.00
revol_bal	532428.000000	16921.280323	22423.215835	0.000000	6444.000000	11876.000000	20843.000000	2568995.00
revol_util	532141.000000	55.057189	23.853436	0.000000	37.700000	56.000000	73.600000	892.30
total_acc	532412.000000	25.267357	11.843211	1.000000	17.000000	24.000000	32.000000	162.00
total_rec_int	532428.000000	1753.428788	2093.199837	0.000000	441.600000	1072.690000	2234.735000	24205.62
total_rec_late_fee	532428.000000	0.394954	4.091546	0.000000	0.000000	0.000000	0.000000	358.68
recoveries	532428.000000	45.717832	409.647467	0.000000	0.000000	0.000000	0.000000	33520.27
collection_recovery_fee	532428.000000	4.859221	63.123361	0.000000	0.000000	0.000000	0.000000	7002.19
collections_12_mths_ex_med	532333.000000	0.014299	0.133005	0.000000	0.000000	0.000000	0.000000	16.00
mths_since_last_major_derog	132980.000000	44.121462	22.198410	0.000000	27.000000	44.000000	61.000000	180.00
acc_now_delinq	532412.000000	0.005015	0.079117	0.000000	0.000000	0.000000	0.000000	14.00
tot_coll_amt	490424.000000	213.562222	1958.571538	0.000000	0.000000	0.000000	0.000000	496651.00
tot_cur_bal	490424.000000	139554.110792	153914.877437	0.000000	29839.750000	80669.500000	208479.250000	8000078.00
total_rev_hi_lim	490424.000000	32080.572919	38053.035312	0.000000	14000.000000	23700.000000	39800.000000	9999999.00
loan_status	532428.000000	0.236327	0.424826	0.000000	0.000000	0.000000	0.000000	1.00

INFERENCE

- The features 'delinq_2yrs', 'inq_last_6mths', 'pub_rec', 'collections_12_mths_ex_med', 'acc_now_delinq' have very low standard deviation.
- The average loan amount borrowed by the people is 14744.27 dollars, average interest rate at which the bank lent is around 13.242969 and the average annual income of the customers is 75029.84 dollars.
- 75% of the data in the features 'delinq_2yrs', 'pub_rec','total_rec_late_fee', 'collection_recovery_fee', 'collections_12_mths_ex_med', 'acc_now_delinq' and 'tot_coll_amt'.

The maximum amount of loan ever taken is 73,544,841 dollars.

DESCRIPTION OF CATEGORICAL FEATURES

```
df.select_dtypes(include='object').describe().T
```

	count	unique	top	freq
term	532428	2	36 months	372793
batch_enrolled	447279	104		106079
grade	532428	7	B	152713
sub_grade	532428	35	B3	33844
emp_title	501595	190124	Teacher	8280
emp_length	505537	11	10+ years	175105
home_ownership	532428	6	MORTGAGE	265940
verification_status	532428	3	Source Verified	197750
pymnt_plan	532428	2	n	532420
desc	75599	70638	> Debt consolidation 	576
purpose	532428	14	debt_consolidation	314989
title	532338	39693	Debt consolidation	248967
zip_code	532428	917	945xx	5845
addr_state	532428	51	CA	77911
initial_list_status	532428	2	f	274018
application_type	532428	2	INDIVIDUAL	532123
verification_status_joint	305	3	Not Verified	170
last_week_pay	532428	98	13th week	30333

INFERENCE

- Most of the applicants have chosen a span of 36 months for the repayment of their loan
- Most of the applicant's profession is teaching and they have a 10+ years of experience.
- Most applicants are based out of state of California.
- Most loans were taken for debt consolidation.

Most applicants live in a mortgaged house.

NULL VALUES

	Count	Percentage
verification_status_joint	532123	99.942715
desc	456829	85.801085
mths_since_last_record	450305	84.575755
mths_since_last_major_derog	399448	75.023853
mths_since_last_delinq	272554	51.190771
batch_enrolled	85149	15.992585
total_rev_hi_lim	42004	7.889142
tot_cur_bal	42004	7.889142
tot_coll_amt	42004	7.889142
emp_title	30833	5.791018
emp_length	26891	5.050636
revol_util	287	0.053904
collections_12_mths_ex_med	95	0.017843
title	90	0.016904
open_acc	16	0.003005
pub_rec	16	0.003005
delinq_2yrs	16	0.003005
inq_last_6mths	16	0.003005
acc_now_delinq	16	0.003005
total_acc	16	0.003005
annual_inc	3	0.000563
recoveries	0	0.000000
total_rec_late_fee	0	0.000000
total_rec_int	0	0.000000
collection_recovery_fee	0	0.000000
initial_list_status	0	0.000000
application_type	0	0.000000
last_week_pay	0	0.000000
member_id	0	0.000000
revol_bal	0	0.000000
loan_amnt	0	0.000000
dti	0	0.000000
addr_state	0	0.000000
zip_code	0	0.000000
purpose	0	0.000000
pymnt_plan	0	0.000000
verification_status	0	0.000000
home_ownership	0	0.000000
sub_grade	0	0.000000
grade	0	0.000000
int_rate	0	0.000000
term	0	0.000000
funded_amnt_inv	0	0.000000
funded_amnt	0	0.000000
loan_status	0	0.000000

NULL VALUE IMPUTATION

- 'verification_status_joint', 'mths_since_last_record' and 'mths_since_last_major_derog' have high percentage of null values.
- Percentage of Null Values in the feature total_rev_hi_lim, tot_cur_bal is close to 8% and it is highly right skewed thus we impute the null values with the median.
- Percentage of Null Values in the feature tot_coll_amt is close to 8%, it is also highly right skewed and 85% data of this feature consists of 0 thus we impute it with 0.
- Since the feature 'mths_since_last_delinq' consists 51% of null values, imputing all of them with mean or median or mode will add wrong information and this may lead to erroneous classification thus we group the data based on the grade and assign the mean of each grade to its respective null value.

REDUNDANT FEATURE ELIMINATION

- 'verification_status_joint','mths_since_last_record','mths_since_last_major_derog' have high percentage of null values, thus we drop them.
- 'pymnt_plan' and 'application_type' feature contains all the observations as only one category which will not be of any significance in our analysis. So, proceeding to drop them.
- 'member_id','batch_enrolled', 'emp_title','title','zip_code','addr_state','desc' do not have any significance in prediction thus, we drop them.
- Since 'loan_amnt','funded_amnt','funded_amnt_inv' are highly correlated and have 99% similar values we keep 'funded_amnt' out of the three features.
- After dropping redundant features, we are left with 30 features.

FEATURE ENGINEERING

- We used features 'collection_recovery_fee' and 'recoveries' to create a new feature 'rec_and_col_fee' and dropped those two.
- Similarly, we used features 'last_week_pay' and 'term' to create a new feature 'emi_paid_progress_perc'.

PROJECT JUSTIFICATION

PROJECT STATEMENT

The Bank Indessa has not done well in last 3 quarters. Their NPAs (Non-Performing Assets) have reached all-time high. It is starting to lose the confidence of its investors. As a result, its stock has fallen by 20% in the previous quarter alone.

After careful analysis, it was found that the majority of NPA was contributed by loan defaulters. With the messy data collected over the years, this bank has decided to use machine learning to figure out a way to find these defaulters and devise a plan to reduce them.

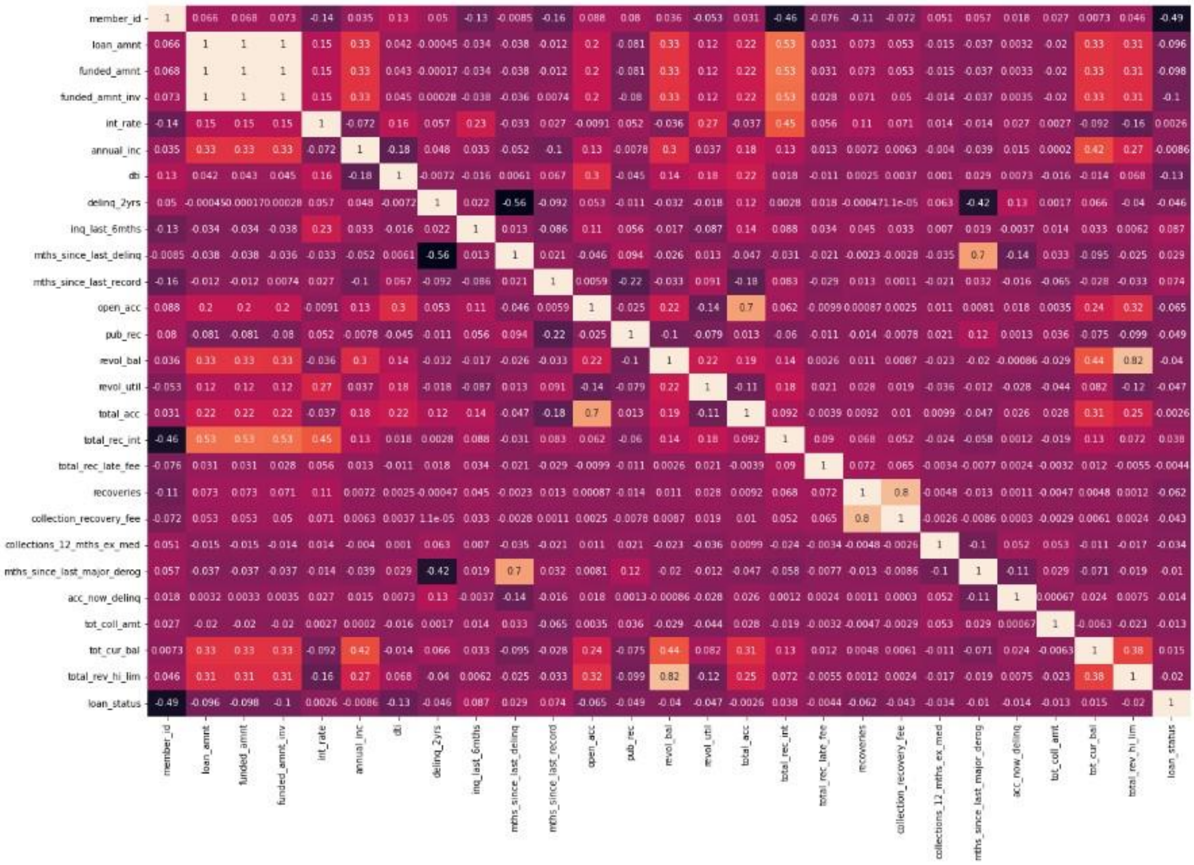
We have built a Logistic Regression Model which will help the Bank to predict if an applicant will default the loan or not and thus, help the bank to take an informed decision.

COMPLEXITY INVOLVED

- The target variable is imbalanced.
- Many redundant features present in the dataset.
- Collinearity amongst several features.
- Huge outliers present in the dataset.
- Since it is a high dimensional dataset, it will lead to high computational complexities.

EXPLORATORY DATA ANALYSIS (EDA)

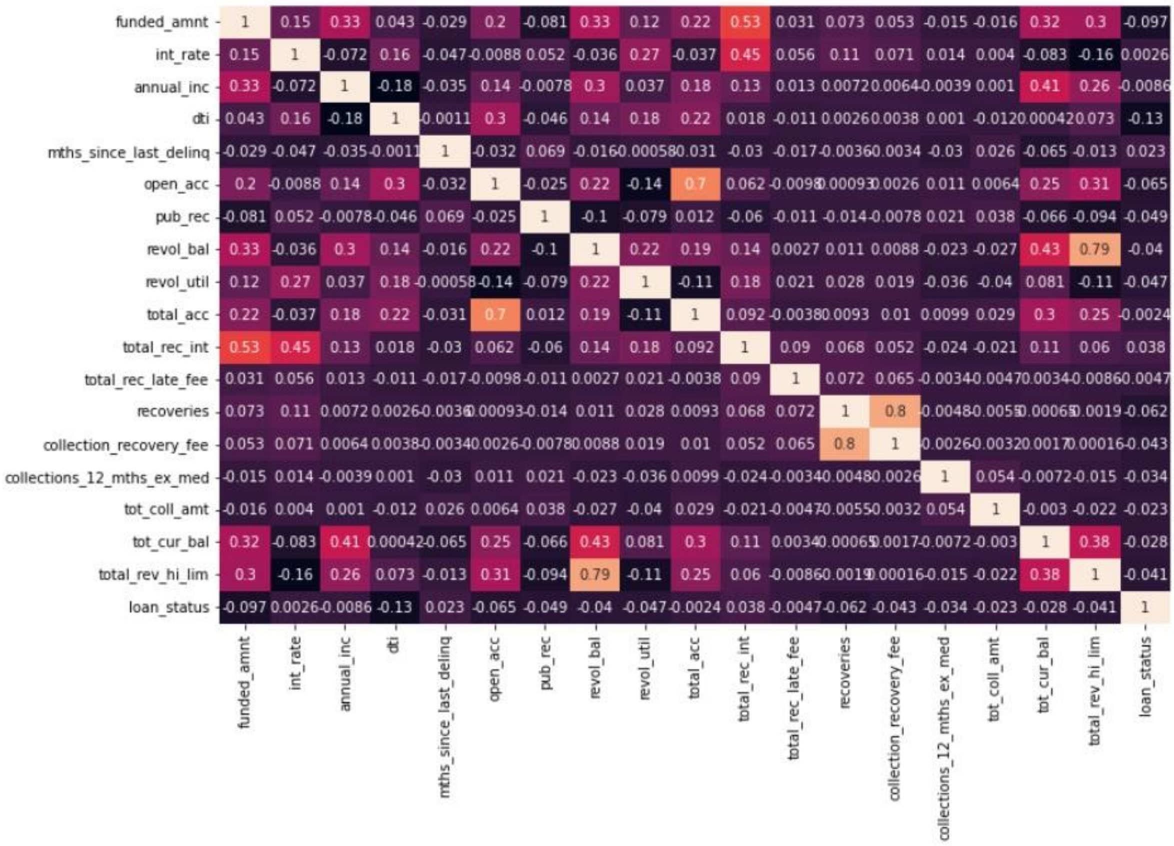
HEATMAP BEFORE NULL IMPUTATION AND DROPPING REDUNDANT FEATURES



INFERENCE

- ‘member_id’, ‘delinq_2_yrs’, ‘delinq_last_6_mths’ feature has very low correlation with all other features.
- ‘loan_amnt’ , ‘funded_amnt’, ‘funded_amnt_inv’ all three have correlation close to 1.
- ‘collection_recovery_fee’, ‘recoveries’ has a correlation of almost 1.
- Total_rev_hi_lim , revolving_bal have a correlation of 0.8.
- ‘mths_since_last_major_derog’, ‘mths_since_last_delinq’ have a correlation of 0.7

HEATMAP AFTER NULL IMPUTATION AND DROPPING REDUNDANT FEATURES



INFERENCE

After removing the null values, dropping redundant features and adding new features, the correlation between the features is reduced.

CHECK FOR MULTICOLLINEARITY (VIF)

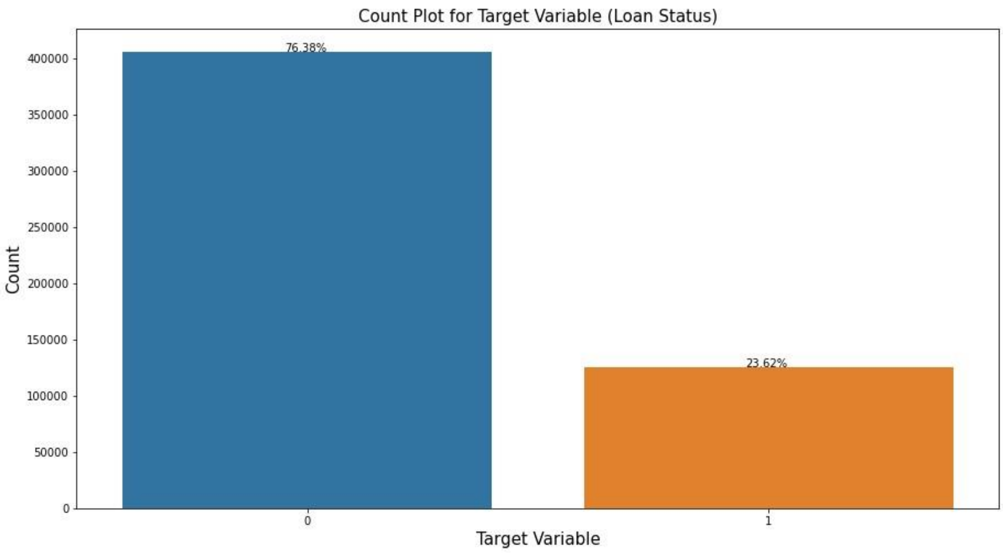
	Feature	VIF
5	open_acc	12.272925
9	total_acc	11.295260
1	int_rate	11.237203
8	revol_util	8.644506
0	funded_amnt	7.600825
3	dti	7.212553
15	total_rev_hi_lim	5.989397
7	revol_bal	5.577459
4	mths_since_last_delinq	5.224795
10	total_rec_int	3.589924
18	emi_paid_progress_perc	3.403595
2	annual_inc	3.102006
14	tot_cur_bal	2.710881
16	loan_status	1.499903
6	pub_rec	1.159956
17	rec_and_col_fee	1.035207
11	total_rec_late_fee	1.025416
13	tot_coll_amt	1.020274
12	collections_12_mths_ex_med	1.018679

INFERENCE

VIF is not very high for any features, so we decided to use all the features to build our basic model and we shall build the upcoming models based on the p_value of each feature we'll decide it's significance for our prediction. If we find it insignificant, we’ll drop it.

DISTRIBUTION OF VARIABLES

DISTRIBUTION OF TRAGET COLUMN

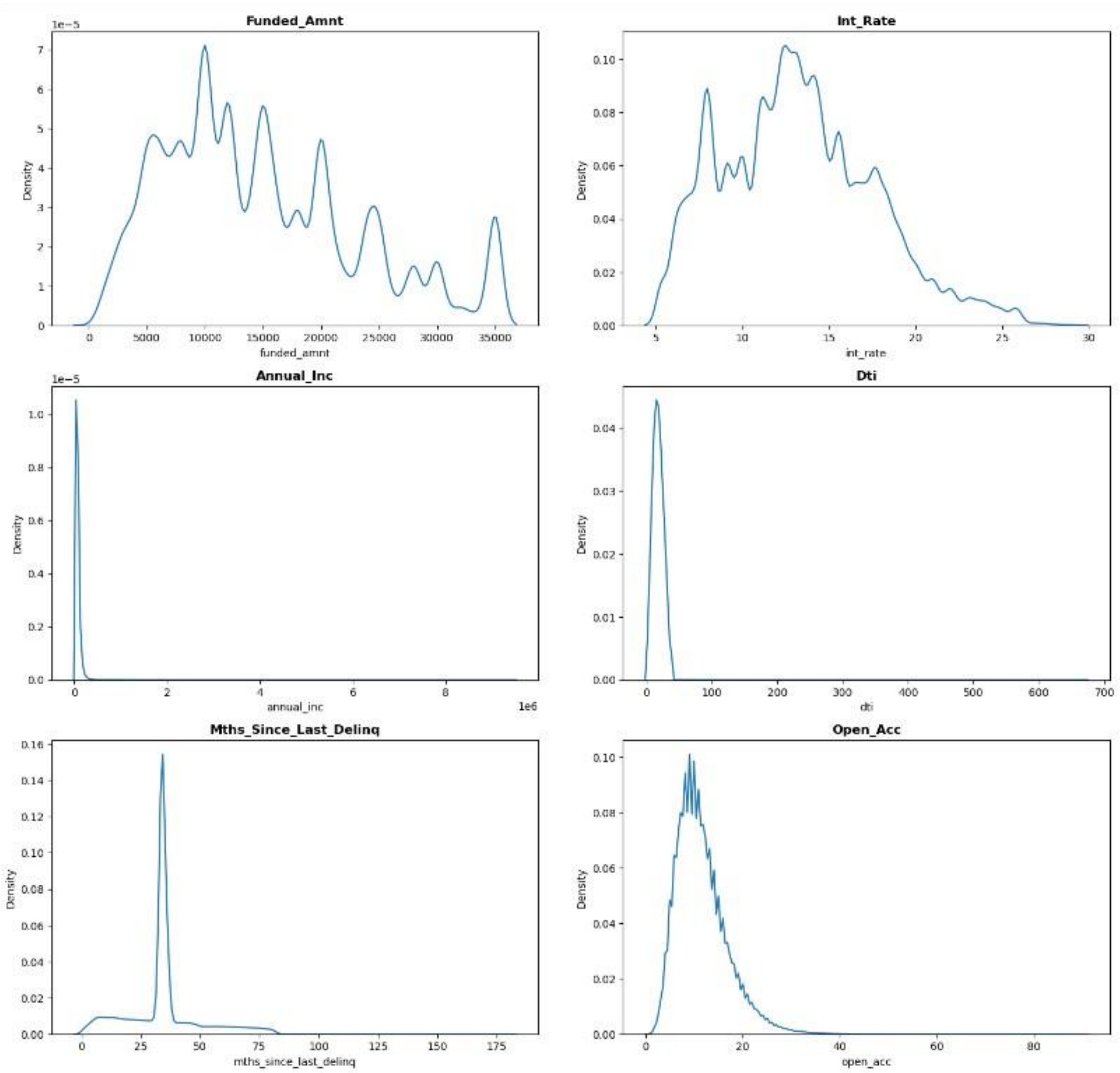


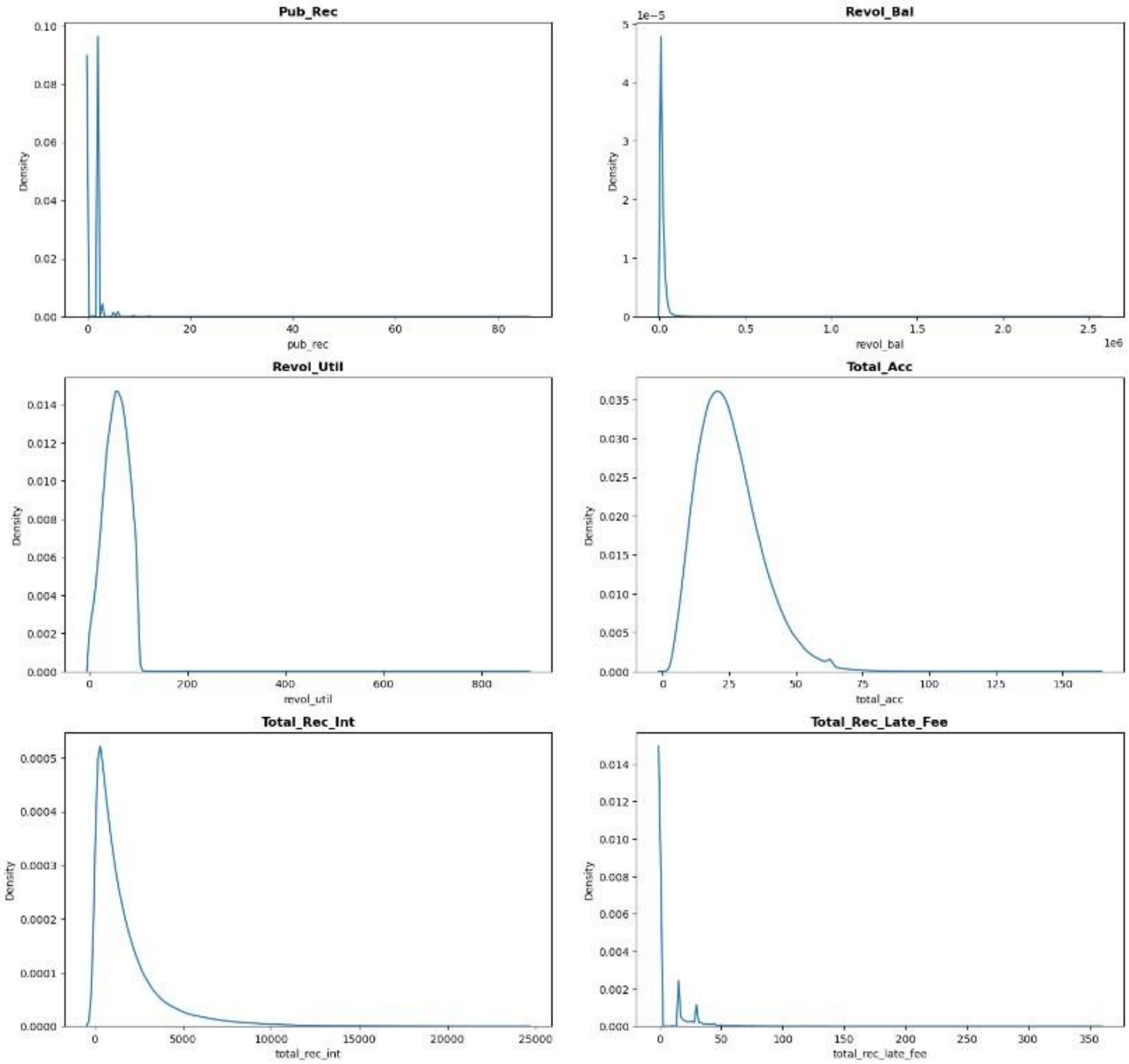
76.38% people are non-defaulters, 23.62% are defaulters in the target variable 'loan_status'

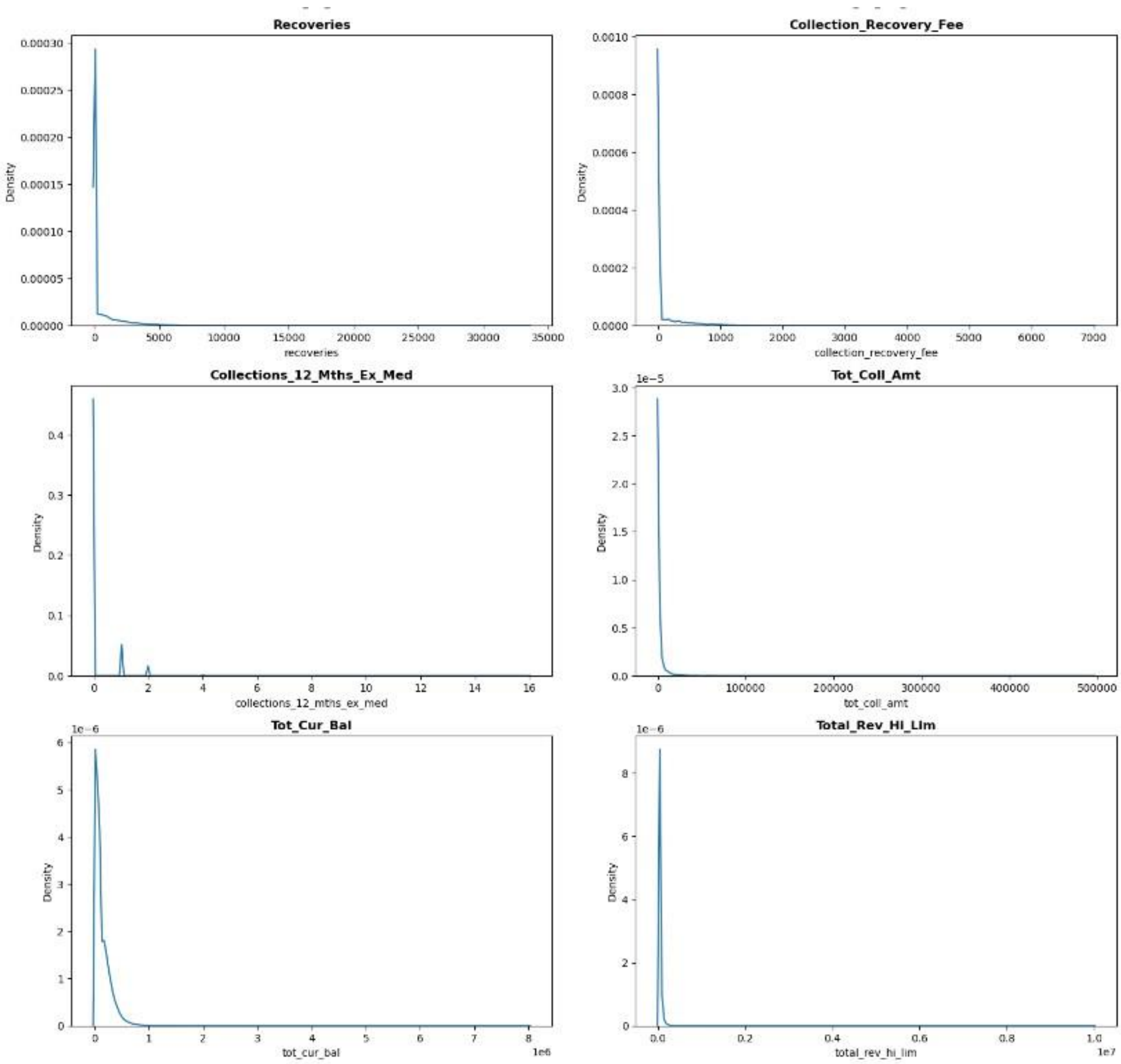
CLASS IMBALANCE AND ITS TREATMENT

- The target column is imbalanced and this can lead to bias during prediction if we go ahead building a model with this data set.
- Initially, we have gone ahead with building our model without treatment.
- In the further model's built, we shall be using SMOTE Techniques to overcome this Class Imbalance in the target column.

DISTRIBUTION OF NUMERICAL FEATURES



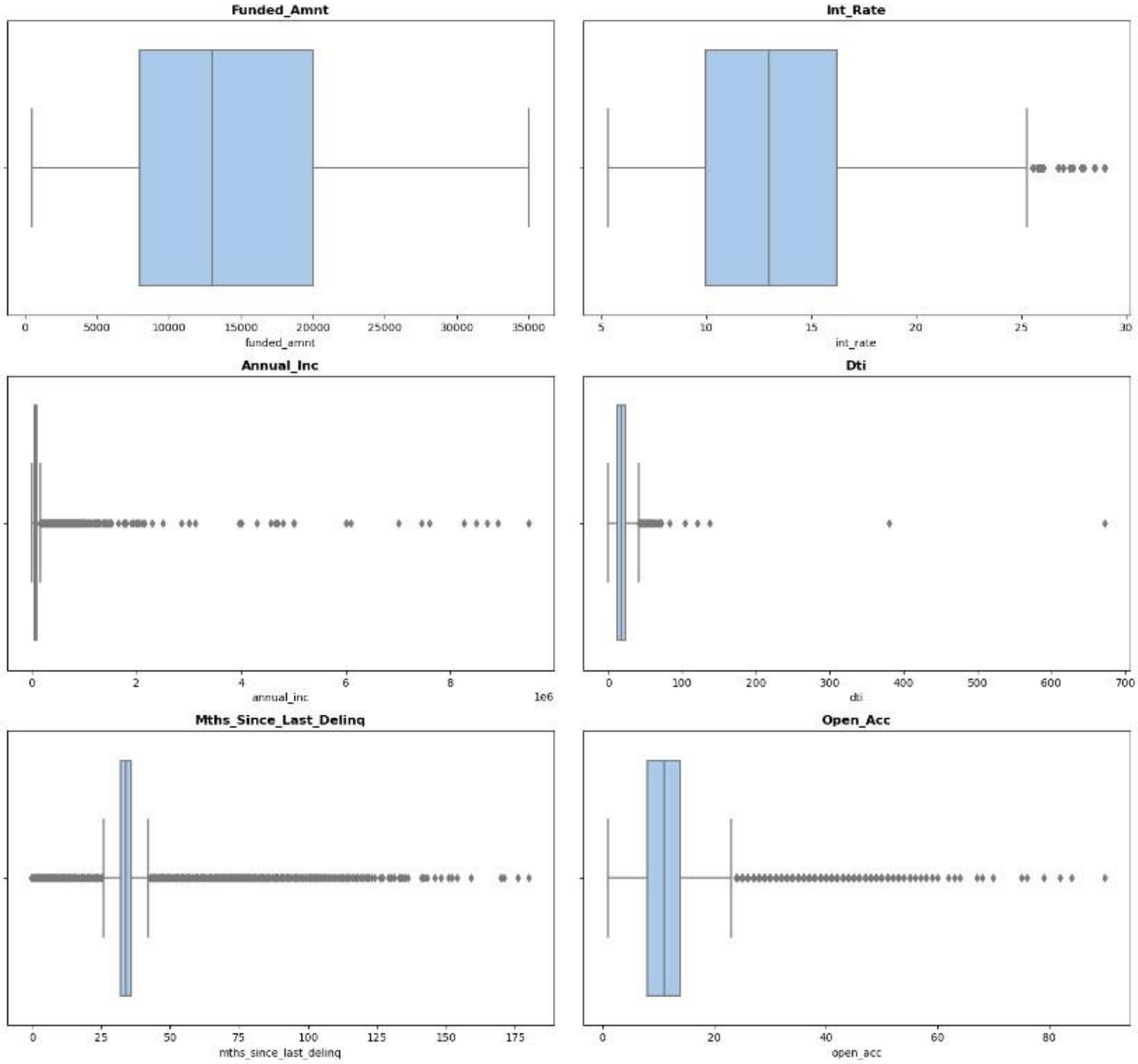

Pub_Rec
Density
pub_rec
Revol_Bal
1e−5
Density
revol_bal
1e6
Revol_Util
Density
revol_util
Total_Acc
Density
total_acc
Total_Rec_Int
Density
total_rec_int
Total_Rec_Late_Fee
Density
total_rec_late_fee

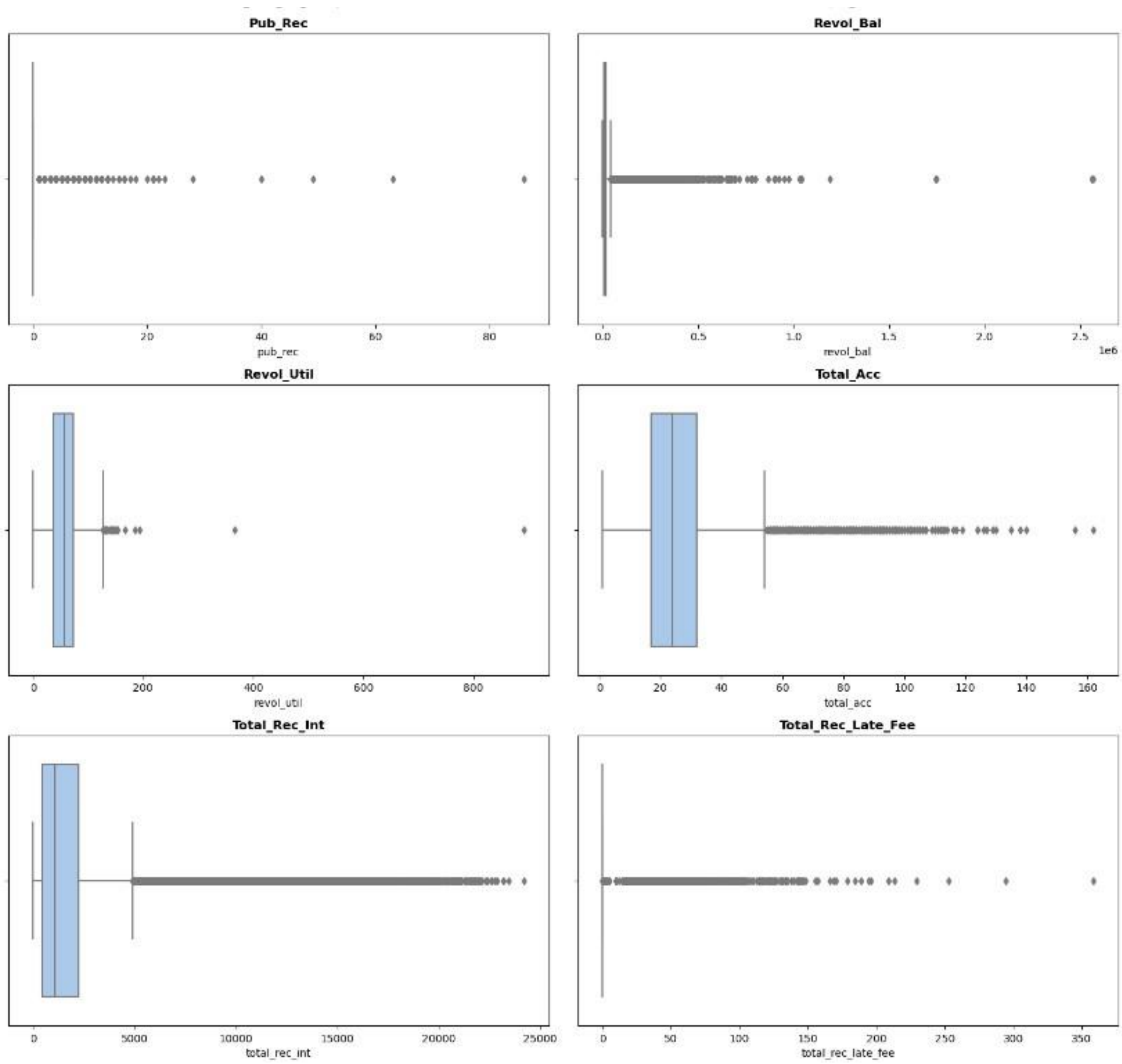


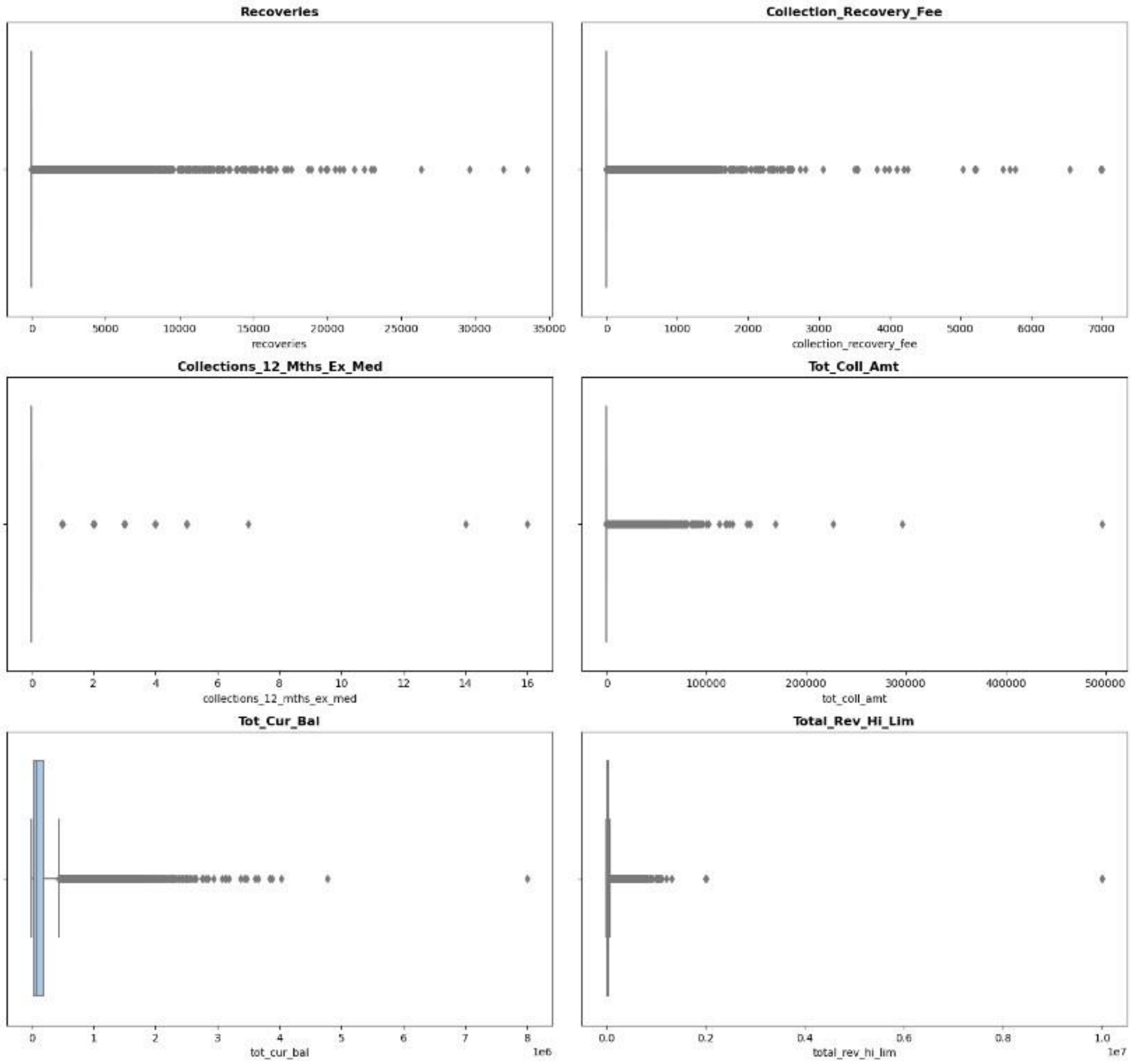
INFERENCE

- All of the numerical features are highly right-skewed.
- None of them has negative values.
- Most of the data points are 0s in a few features

PRESENCE OF OUTLIERS AND ITS TREATMENT




Pub_Rec
pub_rec
Revol_Bal
revol_bal
1e6
Revol_Util
revol_util
Total_Acc
total_acc
Total_Rec_Int
total_rec_int
Total_Rec_Late_Fee
total_rec_late_fee



INFERENCE

Since people with very low salary and high loan amount are outliers and they are more likely to default, there are many such data points which are outliers, removing them will lead to loss of important information, therefore we keep the outliers.

STATISTICAL SIGNIFICANCE OF CATEGORICAL FEATURES

CHI-SQUARE TEST OF INDEPENDENCE

```
from scipy.stats import chi2_contingency
from statsmodels.stats.anova import anova_lm
```

```
Dependent_features=pd.DataFrame(columns=['Feature','Target','P_Value','Dependency'])
for i in df_cat.columns:
    table=pd.crosstab(df_cat[i],df_target)
    observed_table=table.values
    test_stat,p_value,dof,expected_value=chi2_contingency(observed=table,correction=False)
    if p_value<0.05:
        Dependent_features=Dependent_features.append({'Feature':i,'Target':'Loan_Status','P_Value':round(p_value,3),
                                                      'Dependency':'Dependent'},ignore_index=True)
    else:
        Dependent_features=Dependent_features.append({'Feature':i,'Target':'Loan_Status','P_Value':round(p_value,3),
                                                      'Dependency':'Independent'},ignore_index=True)
print('Target variable is dependent on the following Features: ')
Dependent_features
```

```
Target variable is dependent on the following Features:
```

	Feature	Target	P_Value	Dependency
0	term_60 months	Loan_Status	0.000000	Dependent
1	grade_B	Loan_Status	0.000000	Dependent
2	grade_C	Loan_Status	0.000000	Dependent
3	grade_D	Loan_Status	0.000000	Dependent
4	grade_others	Loan_Status	0.000000	Dependent
5	emp_length_Low	Loan_Status	0.000000	Dependent
6	emp_length_Medium	Loan_Status	0.000000	Dependent
7	home_ownership_OTHERS	Loan_Status	0.000000	Dependent
8	home_ownership_OWN	Loan_Status	0.000000	Dependent
9	home_ownership_RENT	Loan_Status	0.000000	Dependent
10	verification_status_Verified	Loan_Status	0.000000	Dependent
11	purpose_debt_consolidation	Loan_Status	0.000000	Dependent
12	purpose_home_improvement	Loan_Status	0.000000	Dependent
13	purpose_other	Loan_Status	0.000000	Dependent
14	initial_list_status_w	Loan_Status	0.000000	Dependent

INFERENCE

We tested if every categorical feature is dependent on the Target Variable, the p-value of all the features came out to be less than 0.05 (Level of Significance) and thus we reject Null hypothesis. Therefore, all the features are dependent on the Target Variable.

FEATURE ENGINEERING

TRANSFORMATION

Since our data has 0s and non-negative values we decided to go ahead with SQRT, POWER and p1log Transformation. Our Base model is built on untransformed data. We would be building our upcoming models with different transformations depending on how effective they are.

NUMERICAL FEATURES SCALING

We scale the variables to get all the variables in the same range. With this, we can avoid a problem in which some features come to dominate solely because they tend to have larger values than others.

```
# initialize the standard scalar
X_scaler = StandardScaler()

# scale all the numerical columns
# standardize all the columns of the dataframe 'num_f'
num_scaled = X_scaler.fit_transform(num_f)

# create a dataframe of scaled numerical variables
# pass the required column names to the parameter 'columns'
df_num_scaled = pd.DataFrame(num_scaled, columns = num_f.columns)
```

BASE MODEL (LOGISTIC REGRESSION)

BUILD A FULL LOGISTIC MODEL ON A TRAINING DATASET

```
# build the model on train data (x_train and y_train)
# use fit() to fit the logistic regression model
logreg = sm.Logit(y_train,x_train).fit()

# print the summary of the model
print(logreg.summary())
```

```
Warning: Maximum number of iterations has been exceeded.
         Current function value: 0.458857
         Iterations: 35
```

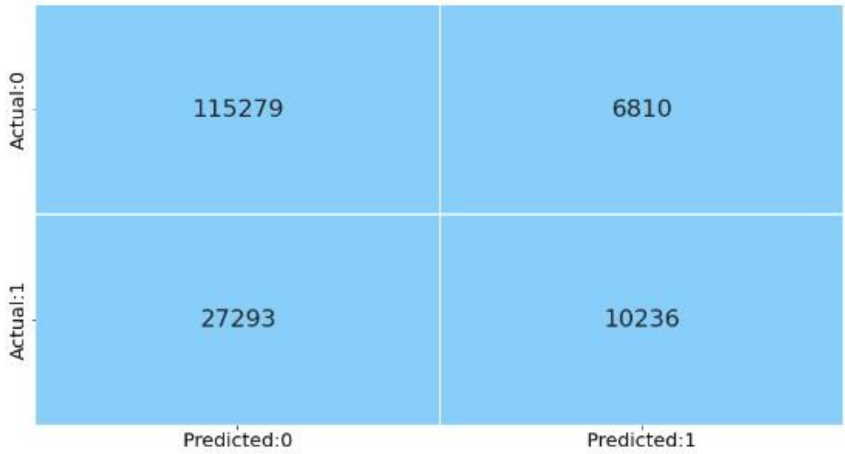
```
C:\Users\User\anaconda3\lib\site-packages\statsmodels\base\model.py:604: ConvergenceWarning: Maximum Likelihood optimizatio
n failed to converge. Check mle_retvals
  warnings.warn("Maximum Likelihood optimization failed to "
```

```
                           Logit Regression Results
==============================================================================
Dep. Variable:            loan_status   No. Observations:               372439
Model:                          Logit   Df Residuals:                   372405
Method:                           MLE   Df Model:                           33
Date:                Thu, 11 Aug 2022   Pseudo R-squ.:                  0.1614
Time:                        02:11:32   Log-Likelihood:            -1.7090e+05
converged:                      False   LL-Null:                   -2.0380e+05
Covariance Type:            nonrobust   LLR p-value:                     0.000
===============================================================================================
                                  coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
const                         -48.2482     69.097     -0.698      0.485    -183.676      87.179
term_60 months                 -0.2542      0.013    -19.589      0.000      -0.280      -0.229
grade_B                        -0.7825      0.017    -45.210      0.000      -0.816      -0.749
grade_C                        -1.5258      0.025    -60.699      0.000      -1.575      -1.477
grade_D                        -2.2127      0.035    -64.122      0.000      -2.280      -2.145
grade_others                   -3.0440      0.048    -63.856      0.000      -3.137      -2.951
emp_length_Low                  0.0960      0.010      9.264      0.000       0.076       0.116
emp_length_Medium               0.0722      0.011      6.803      0.000       0.051       0.093
home_ownership_OTHERS           2.4443      0.364      6.721      0.000       1.732       3.157
home_ownership_OWN             -0.2077      0.016    -13.184      0.000      -0.239      -0.177
home_ownership_RENT            -0.1300      0.011    -12.213      0.000      -0.151      -0.109
verification_status_Verified   -0.1192      0.010    -12.240      0.000      -0.138      -0.100
purpose_debt_consolidation      0.1358      0.011     12.430      0.000       0.114       0.157
purpose_home_improvement        0.0223      0.020      1.094      0.274      -0.018       0.062
purpose_other                   0.1796      0.016     11.378      0.000       0.149       0.211
initial_list_status_w          -0.8174      0.009    -87.868      0.000      -0.836      -0.799
funded_amnt                     0.0271      0.007      4.024      0.000       0.014       0.040
int_rate                        1.1523      0.014     81.905      0.000       1.125       1.180
annual_inc                     -0.0050      0.005     -1.021      0.307      -0.015       0.005
dti                            -0.2918      0.005    -57.314      0.000      -0.302      -0.282
mths_since_last_delinq          0.0702      0.004     16.453      0.000       0.062       0.079
open_acc                       -0.2135      0.007    -32.806      0.000      -0.226      -0.201
pub_rec                        -0.1944      0.006    -33.708      0.000      -0.206      -0.183
revol_bal                       0.1129      0.009     11.936      0.000       0.094       0.131
revol_util                     -0.1688      0.006    -30.471      0.000      -0.180      -0.158
total_acc                       0.3005      0.006     50.096      0.000       0.289       0.312
total_rec_int                  -0.2645      0.007    -35.592      0.000      -0.279      -0.250
total_rec_late_fee             -0.0331      0.005     -6.922      0.000      -0.042      -0.024
collections_12_mths_ex_med     -0.0940      0.006    -15.341      0.000      -0.106      -0.082
tot_coll_amt                   -0.1214      0.008    -14.494      0.000      -0.138      -0.105
tot_cur_bal                    -0.0125      0.006     -2.169      0.030      -0.024      -0.001
total_rev_hi_lim               -0.1148      0.012     -9.823      0.000      -0.138      -0.092
rec_and_col_fee              -444.7749    631.033     -0.705      0.481   -1681.576     792.027
emi_paid_progress_perc          0.5157      0.005     93.852      0.000       0.505       0.526
===============================================================================================
```

INFERENCE

- LLR p-value of the model is 0.000, which is less than 0.05 therefore Null hypothesis is rejected and thus, there is at least one feature which is significant.
- Pseudo R-square of the model is: 0.1614, it's far from 1 therefore we conclude that there are many improvements to be done on the model.
- Log-Likelihood of the model is: -1.7090e+05 which is greater than the Log-Likelihood of the Null Model i.e., -2.0380e+05. Indicating our model has performed quite better.
- 'purpose_home_improvement', annual_inc','rec_and_col_fee' are insignificant features.

CONFUSION MATRIX



INFERENCE

- True Negative: 115279
- True Positive: 10236
- False Positive: 6810
- False Negative: 27293

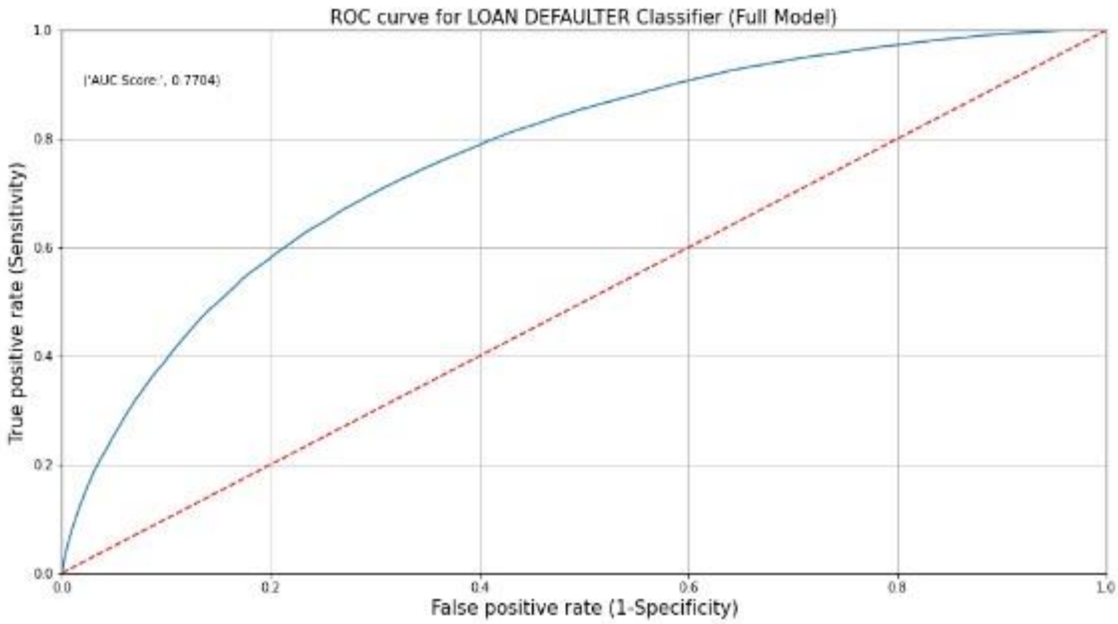
CLASSIFICATION REPORT

	precision	recall	f1-score	support
0	0.81	0.94	0.87	122089
1	0.60	0.27	0.38	37529
accuracy			0.79	159618
macro avg	0.70	0.61	0.62	159618
weighted avg	0.76	0.79	0.75	159618

INFERENCE

- Since, we wouldn’t want to wrongly classify the actual defaulters as non-defaulters i.e., reduce Type-II Error as much as possible.
- This can be done by focussing on the recall i.e., 0.27 for the model, since it is a component of Type-II Error which effects the prediction of the model.

ROC CURVE



INTERPRETATION

- The red dotted line represents the ROC curve of a purely random classifier; a good classifier stays as far away from that line as possible (toward the top-left corner).
- From the above plot, we can see that our classifier (logistic regression) is away from the dotted line; with the AUC score 0.7704.

CONCLUSIONS

SCORE CARD FOR LOGISTIC REGRESSION

	Probability Cutoff	AUC Score	Precision Score	Recall Score	Accuracy Score	Kappa Score	f1-score
0	0.100000	0.770387	0.296274	0.941778	0.460362	0.144859	0.450747
1	0.200000	0.770387	0.388207	0.764715	0.661329	0.295137	0.514983
2	0.300000	0.770387	0.471529	0.582083	0.748355	0.352896	0.521006
3	0.400000	0.770387	0.540862	0.413360	0.779567	0.332653	0.468593
4	0.500000	0.770387	0.600493	0.272749	0.786346	0.267537	0.375117
5	0.600000	0.770387	0.667381	0.157558	0.783464	0.181412	0.254931
6	0.700000	0.770387	0.733008	0.064084	0.774462	0.083203	0.117863
7	0.800000	0.770387	0.765957	0.011511	0.766762	0.015829	0.022681
8	0.900000	0.770387	0.945946	0.000933	0.765089	0.001401	0.001863

- The Logistic Regression model we've built will help the bank to predict the defaulters on the basis of their details with an accuracy of 78%.
- Threshold of 0.3 has the best AUC Score, Kappa Score and f-1 score thus, we'll use 0.3 as a threshold for building the future Logistic Regression models.
- Since, the initial model we built is based on imbalanced target feature, we shouldn't be focussing on the accuracy score to decide the best threshold value, because the model would be biased towards a particular sub class.