Capstone Project – Final Report

Project Group Info:

BATCH DETAILS	PGPDSE-FT Online Nov22
TEAM MEMBERS	I. Shwetank Dhruva (L) II. Krishna Gupta III. Labani Debnath IV. Nimisha Dwivedi V. Pethu Raja
DOMAIN OF PROJECT	Finance & Risk Analytics
PROJECT TITLE	Predicting the Default Customers & Deciding the eligibility of the Customer for Loan on the basis of their Past Behaviour.
GROUP NUMBER	GROUP – 1
TEAM LEADER	Shwetank Dhruva
MENTOR NAME	Mrs. Vidhya Kannaiah

Abstract:

This capstone project presents a comprehensive analysis of loan data based on customer's banking history with the objective of identifying eligible candidates for loan approvals. The study enables financial institutions to make informed decisions when evaluating loan applications. The research aims to improve the loan approval process by leveraging the power of data-driven approaches and minimizing the risks associated with lending.

Objectives:

Analytics techniques allow lenders to develop customized loan products and pricing strategies based on individual borrower profiles. By segmenting customers according to their risk profiles and financial behaviors, lenders can tailor loan terms and interest rates to align with borrowers' needs and abilities to repay. This personalized approach not only benefits borrowers by offering more favorable loan terms, but also helps lenders mitigate risks and improves overall portfolio performance.

The objective of this Capstone Project is to revolutionize the lending industry, leading to fairer, more efficient, and inclusive financial services for individuals and businesses.

Industry Review:

The lending industry has witnessed a significant transformation in recent years with the advent of big data analytics and machine learning algorithms. Traditional methods of evaluating loan applications have relied on credit scores and limited financial information, often leading to inaccurate assessments and potential financial losses for lenders. However, with the availability of extensive banking data, including transactional records, income statements, and credit histories, financial institutions now have the opportunity to analyze a broader range of factors when assessing loan eligibility.

Data analytics techniques, such as exploratory data analysis, predictive modeling, and machine learning, are powerful tools for extracting meaningful patterns and insights of loan data. By applying these techniques, lenders can gain a deeper understanding of borrower’s financial behavior, repayment patterns, and risk profiles. This comprehensive approach helps lenders assess the stability of an applicant's income, the frequency of cash flow, and the consistency of financial obligations.

In conclusion, the analysis of loan data based on customer’s banking history provides financial institutions with valuable insights for identifying eligible candidates for loan approvals. By leveraging advanced data analytics techniques, lenders can make more informed decisions, reduce risks, and improve the accuracy of credit assessments.

Problem Statement:

The problem statement for this capstone project is to design and implement a loan data analysis system that utilizes customer's banking history to assess their creditworthiness and identify eligible candidates for loan approvals.

To incorporate advanced data analytics techniques, such as univariate and bivariate analysis, and other meaningful insights from banking history and transactional data.

Dataset & Domain:

This Dataset named Loan_dataset.csv has attributes about the Candidates who have applied for the Loan. It includes their Loan applications with banking history, credit history, and other details about the candidate.

The Loan_dataset.csv have 133889 rows and 145 columns.

Index (['loan_amnt', 'term', 'int_rate', 'grade', 'emp_length', 'home_ownership', 'annual_inc', ‘verification_status', 'issue_d', 'loan_status', 'pymnt_plan', 'purpose', 'addr_state', 'dti', ‘delinq_2yrs', 'inq_last_6mths', 'open_acc', 'pub_rec', 'revol_util', 'initial_list_status', 'out_prncp', 'total_rec_late_fee', 'recoveries', 'last_pymnt_d', 'collections_12_mths_ex_med', ‘application_type', 'acc_now_delinq', 'tot_coll_amt', 'tot_cur_bal', 'open_acc_6m', 'open_act_il', 'open_il_24m', 'mths_since_rcnt_il', 'total_bal_il', 'il_util', 'open_rv_24m', 'all_util', 'inq_fi', ‘total_cu_tl', 'inq_last_12m', 'bc_open_to_buy', 'bc_util', 'chargeoff_within_12_mths', 'delinq_amnt', 'mort_acc', 'mths_since_recent_inq', 'num_accts_ever_120_pd', 'num_actv_bc_tl', 'num_il_tl','num_tl_120dpd_2m','num_tl_90g_dpd_24m', ‘pct_tl_nvr_dlq',’pub_rec_bankruptcies', 'hardship_flag', 'debt_settlement_flag'], dtype='object')

We have narrowed down the dataset to 55 key attributes that are vital for our project.

These attributes have been carefully selected based on criteria such as null value handling, project requirements, and other relevant parameters.

Sr. No.	Feature	Description	Range	Data Type
1	***loan_amnt***	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.	Range (1000-40000)	Float
2	***term***	The number of payments on the loan. Values are in months and can be either 36 or 60.	60 Months', ' 36 Months', Nan]	Object/Categorical

3	***Int_rate***	Interest Rate on the loan Possible values	Range(5.32%-28.99%)	Object
4	***grade***	LC assigned loan unique 7, top B, freq=40267	Grade(A-G)	Object/Categorical
5	***emp_length-***	Employment length in years. In this 0 Means Less Than One Year And 10 Means Ten Or More Years.	Between 0 And 10	Object
6	***home_ownership-***	The homeownership status provided by the borrower during registration or obtained from the credit report.	'OWN', 'MORTGAGE', 'RENT'	Object /(Categorical)-
7	***annual_inc-***	The self-reported annual income provided by the borrower during registration.	0-6998721	Float
8	***verification_status -***	Indicates if income was verified by LC, not verified, or if the income source was verified	Source Verified, Verified, Not Verified	Object
9	***issue_d-***	The month which the loan was issued	Unique 3 Jan, Feb, March,	Object/ (Categorical)
10	***loan_status***	Current status of the loan	Multiple Categories	Object /Categorical
11	***pymnt_plan-***	Indicates if a payment plan has been put in place for the loan	'N', 'Y', Nan	Categorical/Object
12	***purpose-***	A category provided by the borrower for the loan request.		Object/ Categorical
13	***addr_state-***	The state provided by the borrower in the loan application		Object/ Categorical
14	***dti-***	Debt-to-income ratio is the ratio of total debt payment divided by gross income (before tax) expressed as a percentage, usually on either a monthly or annual basis.	-1000	Float
15	***delinq_2yrs***	how many times failure to make timely payments on debts or loans in last 2yrs	0-22	. (Categorical)/ Float
16	***inq_last_6mths -***	"inq_last_6mths" typically refers to the number of inquiries made by the borrower on their credit within the last six months.	(0-6)	(Categorical) Float

17	***open_acc***	- "open_acc" typically refers to the number of open credit accounts that a borrower has at the time of applying for the loan.	0-74	FLOAT
18	***pub_rec***	- The "pub_rec" variable is a numeric field that represents the count of public record	0-46	FLOAT
19	***Revol_util -***	The "revol_util" variable represents the revolving line utilization rate, which indicates the credit amount used relative to the total credit available to the borrower, ranging from 0% to 100%.	0-172%	OBJECT
20	***initial_list_status***	It indicates whether the loan was funded in its entirety by a single investor ('w') or if it was funded partially by multiple investors ('f').	w or f	Categorical
21	***out_prncp***	It represents the outstanding principal amount remaining on the loan, which is the remaining balance of the loan that the borrower still needs to repay.	0 to 3222.84	Float
22	***total_rec_late_fee***	Information about the additional charges incurred by the borrower due to late or missed payments.	0 to 609	Float
23	***recoveries***	Information about the amount of money that has been recovered from the borrower or through other means, such as collections efforts or legal action, to mitigate the loss incurred due to the loan default.	0 to 23271	Float
24	***last_pymnt_d***	Information about the timing of the borrower's most recent payment. It helps track the payment history and can be used to assess the borrower's repayment behavior.		Object
25	***last_credit_pull_d***	Information about the timing of the last credit check or credit report retrieval.		Object
26	***collections_12_mths_ex_med***	It represents the number of collections (excluding medical collections) reported by lenders in the last 12 months.	0 to 12	Float
27	***application_type***	It describes the type of application made by the borrower, indicating whether it was an individual application or a joint application with another borrower.	Individual and joint	Categorical

28	***acc_now_delinq***	The number of accounts on which the borrower is currently delinquent or has fallen behind on payments. It indicates the count of accounts that are currently in a state of delinquency.	0 to 3 Categorical	Float
29	***tot_coll_amt***	Information about the total debt that has been sent to collections for the borrower.	0 to 224107	Float
30	***Tot_cur_bal***	Total current balance of all accounts	-	Float
31	***Open_acc_6m***	Number of open trades in last 6 months	From 0 to 16.	Float
32	***Open_act_il***	Number of currently active installment trades	From 0 to 48.	Float
33	***Open_il_24m***	Number of installment accounts opened in past 24 months	From 0 to 30.	Float
34	***Mths_since_rcnt_il***	Months since most recent installment accounts opened	-	Float
35	***Total_bal_il***	Total current balance of all installment accounts	-	Float
36	***Il_util***	Ratio of total current balance to high credit/credit limit on all installation acct	-	Float
37	***Open_rv_24m***	Number of revolving trades opened in past 24 months	From 0 to 44.	Float
38	***All_util***	Balance to credit limit on all trades	-	Float
39	***Inq_fi***	Number of personal finance inquiries	From 0 to 28.	Float
40	***total_cu_tl***	Number of finance trades	From 0 to 79	Float
41	***inq_last_12m***	Number of credit inquiries in past 12 months	From 0 to 40	Float
42	***bc_open_to_buy***	Total open to buy on revolving bankcards	-	Float
43	***bc_util***	Ratio of total current balance to high credit/credit limit for all bankcard accounts.	-	Float
44	***chargeoff_within_12_mths***	Number of charge-offs within 12 months	From 0 to 7	Float
45	***delinq_amnt***	The past-due amount owed for the accounts on which the borrower is now delinquent.	From 0 to 112524	Float
46	***mort_acc***	Number of mortgage accounts.	From 0 to 51	Float
47	**mths_since_recent_inq**	Months elasped since customers most recent inquiry	Range (0 - 25)	Discrete/Float

48	**num_accts_ever_120_pd**	Number of accounts ever 120 or more days past due	Range (0 - 38)	Discrete/Float
49	**num_actv_bc_tl**	Number of currently active bankcard accounts with the customer	Range (0 - 36)	Discrete/Float
50	**num_il_tl**	Number of installment accounts created for that customer	Range (0 - 121)	Discrete/Float
51	**num_tl_90g_dpd_24m**	Number of accounts 90 or more days past due in last 24 months	Range (0 - 22)	Discrete/Float
52	**pct_tl_nvr_dlq**	Percent of trades never delinquent	Range (0 - 100)	Discrete/Float (Percentage)
53	**pub_rec_bankruptcies**	Number of public record bankruptcies that customer failed to pay the debt	Range (0 - 9)	Discrete/Float
54	***hardship_flag***	Flags whether the borrower is on a hardship plan	N,Y	Object / Categories (N, Y)
55	***debt_settlement_flag***	Flags whether the borrower, who has charged-off, is working with a debt-settlement company.	N,Y	Object / Two categories (“N”,” Y”)

Variable Categorization:

There are 40 Numerical columns and 14 Categorical columns.

Numerical Columns:

['loan_amnt','acc_now_delinq', 'out_prncp','int_rate','annual_inc', 'total_rec_late_fee', 'recoveries', 'tot_coll_amt', 'tot_cur_bal', 'open_acc_6m', 'open_act_il', 'open_il_24m', 'mths_since_rcnt_il', 'total_bal_il', 'il_util', 'open_rv_24m', 'all_util', 'inq_fi', 'total_cu_tl', 'inq_last_12m', 'bc_open_to_buy','bc_util','chargeoff_within_12_mths','delinq_amnt',‘mort_acc','mths_since_recent_inq','num_accts_ever_120_pd','num_actv_bc_tl','num_il_tl','num_tl_120dpd_2m','num_tl_90g_dpd_24m','pct_tl_nvr_dlq',‘pub_rec_bankruptcies','collections_12_mths_ex_med','dti',‘delinq_2yrs', 'inq_last_6mths', 'open_acc', 'pub_rec', 'revol_util']

Categorical Columns:

['addr_state', 'application_type', 'debt_settlement_flag', 'emp_length', 'grade', 'hardship_flag', 'home_ownership', 'initial_list_status', 'issue_d', 'last_pymnt_d', 'loan_status', 'purpose', 'pymnt_plan', 'term', 'verification_status']

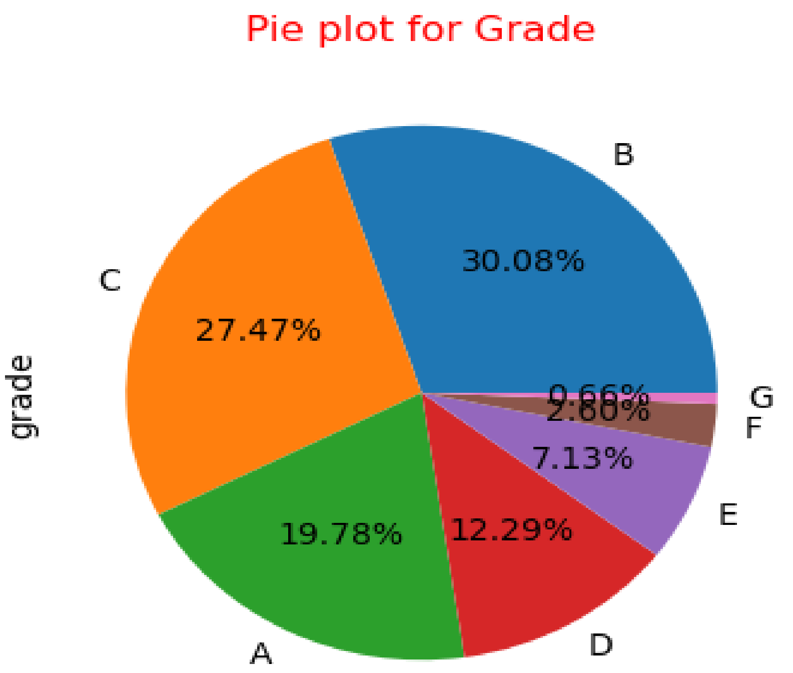
Exploratory Data Analysis:

During the exploratory data analysis phase of this project, we will thoroughly examine the 54 selected attributes to uncover patterns, relationships, and insights within the dataset. Through visualizations, statistical summaries, and data exploration techniques, we aim to gain a comprehensive understanding of the data and make informed decision-making on that basis.

1. Univariate Analysis:-

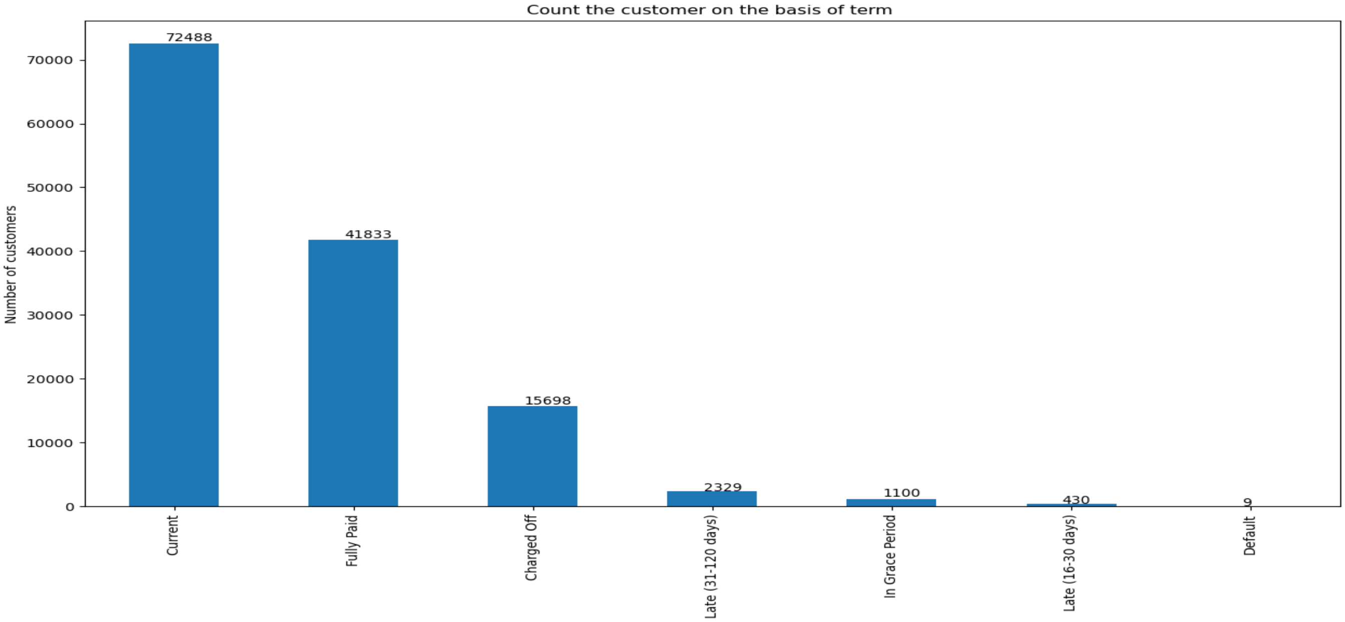
Univariate analysis in this project involves examining individual variables from the selected 54 attributes. By calculating summary statistics, visualizing distributions, and exploring variable characteristics, we gain insights into their standalone behavior and relevancy for the project's objectives.

Category Attribute :

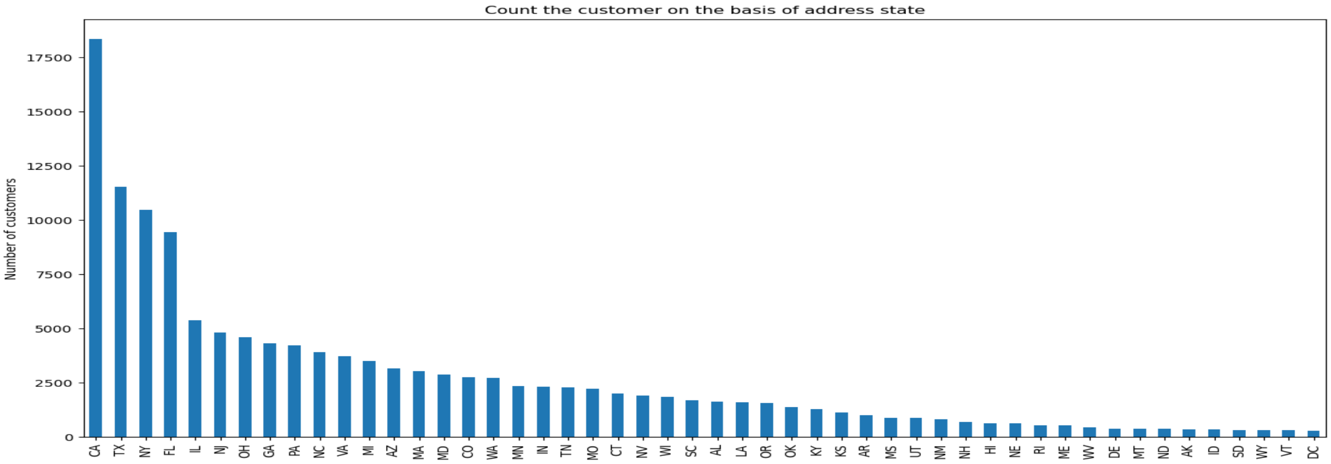


From the above plot, we can observe that more than 30% percentage of borrowers are prefer “Grade B” (Loan interest rate grade) and approx. 27% percentage of borrowers are prefer “Grade C” . Both grade B and C are cover more than 57% of data and both loan grade have low interest rate.

Type Attribute :

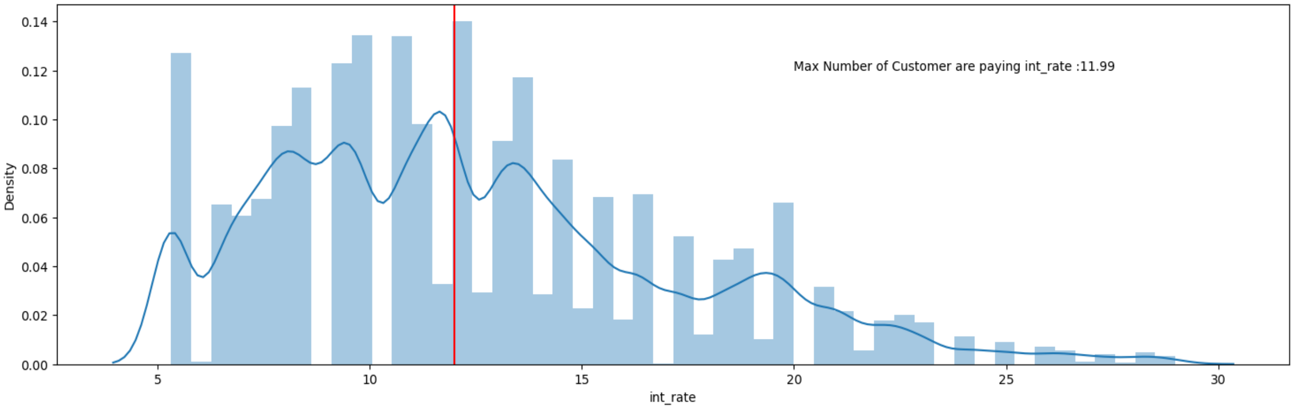


This is a count plot, we are counting the number of borrowers having different loan statuses and we found out that, 72488 borrow loan status is the current means, the customers are the recent customer or the current customer. And we also can see from the above graph that 41833 borrowers are already paid their loan amount. Other than these customers either the customers are not paying the loan or they are late by 31-120 days or 16-30 days, Which may form a risky customer in the future. Also from the above graph, we are able to find (defaulters are present in the given data).



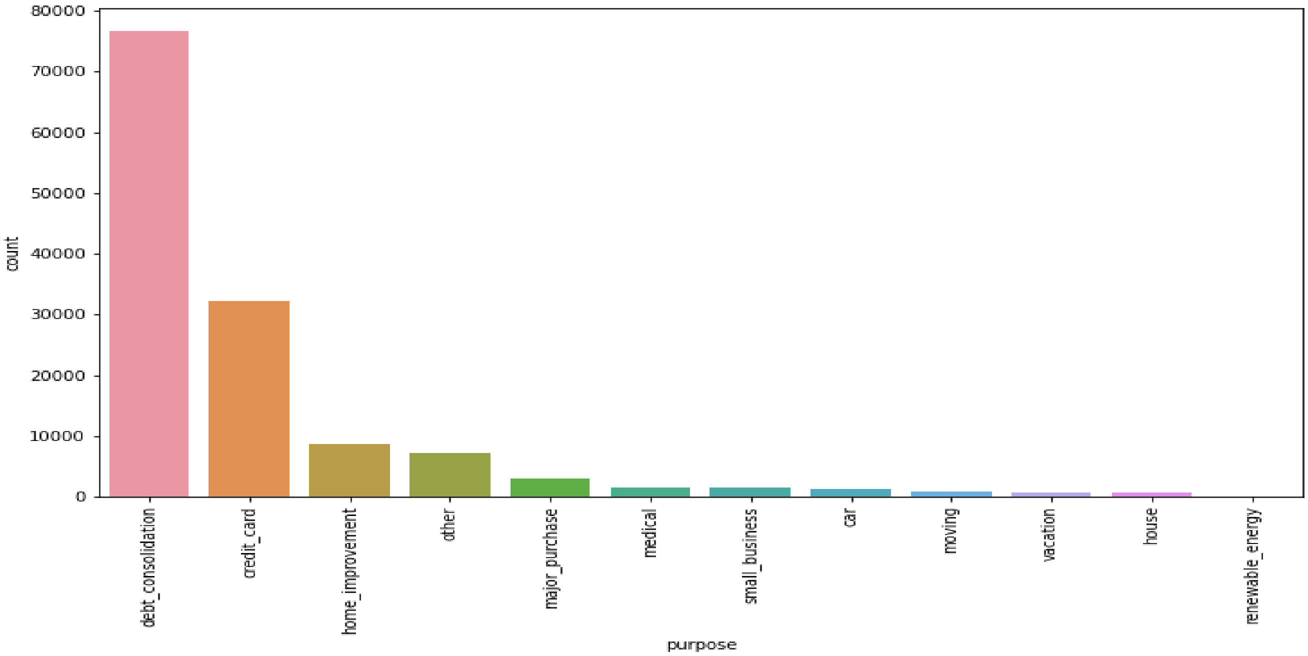
This graph is showing the number of borrowers present in different states and we found out that the maximum number of borrowers are from the "CA" state which is more than 17,500 borrowers and the minimum number of borrowers are from the "DE", "MT", "ND", "AK", “ID”, “SD”, “WY”, “VT” and “DC” which very less than 2,500 borrowers.

Data distribution of interest rate attribute:



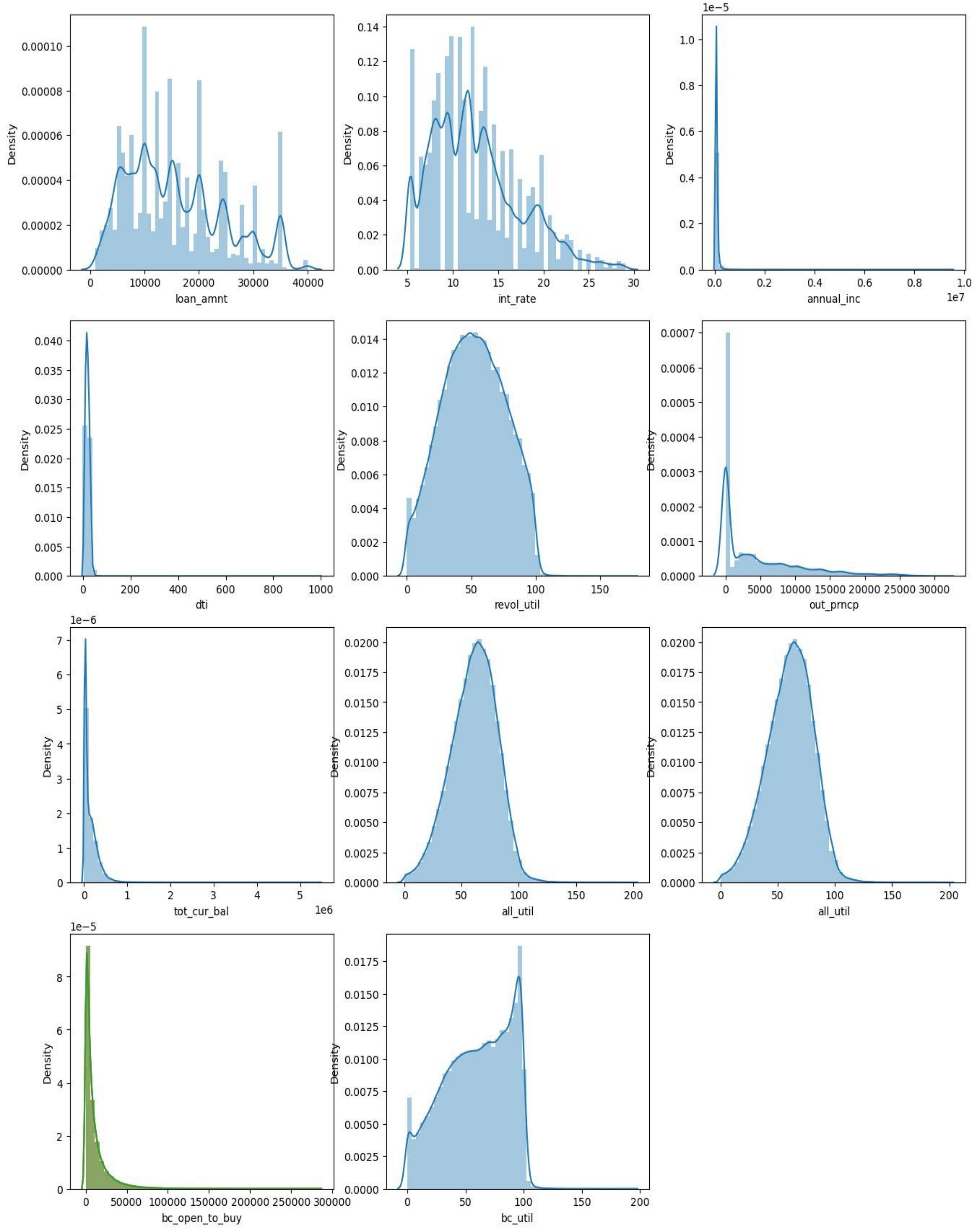
From the above graph, we can say that data is not normally distributed and also we can see data is right skewed. Also, the maximum number of borrowers are paying interest rates of more than 10% and less than 13%. The average is approx. 11%.

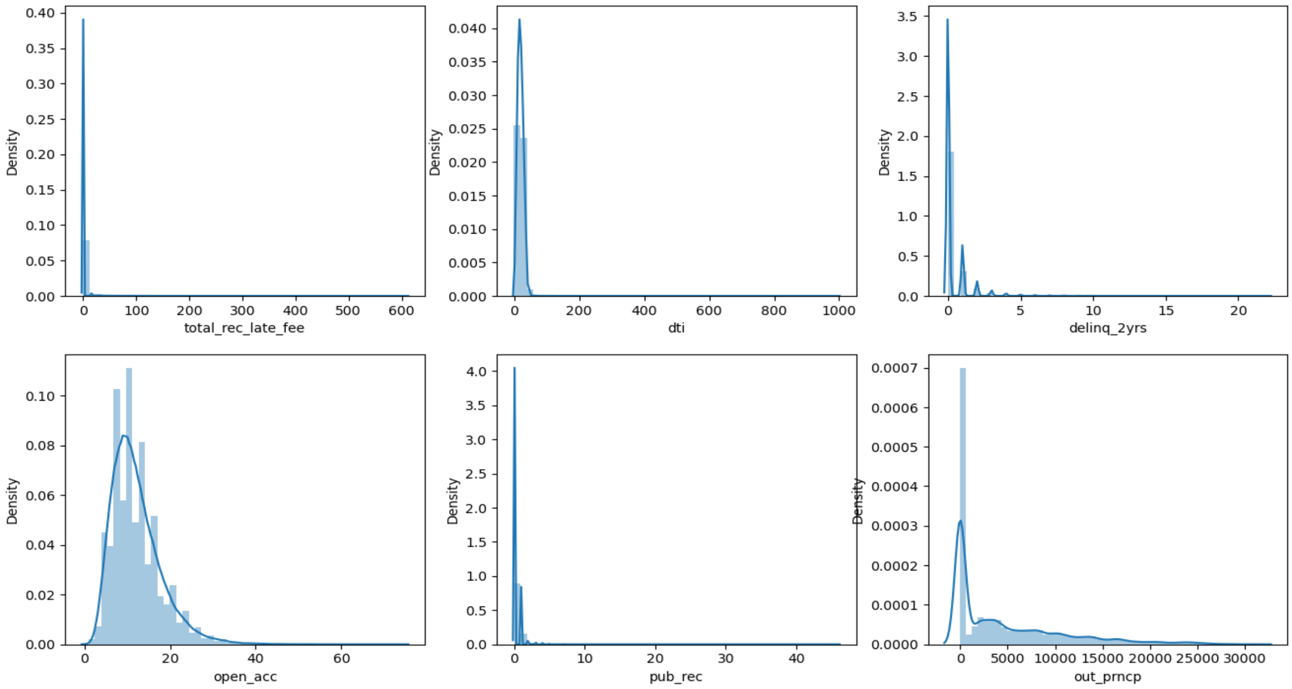
Finding out the purpose of most of the borrowers for taking the loan :



From the above graph, most borrowers are taking loans because of debt consolidation, which is more than 70,000 (more than 50%) of the data. And the minimum reason or purpose of taking the loan is renewable energy almost zero.

Data distribution of all the numeric data attribute :



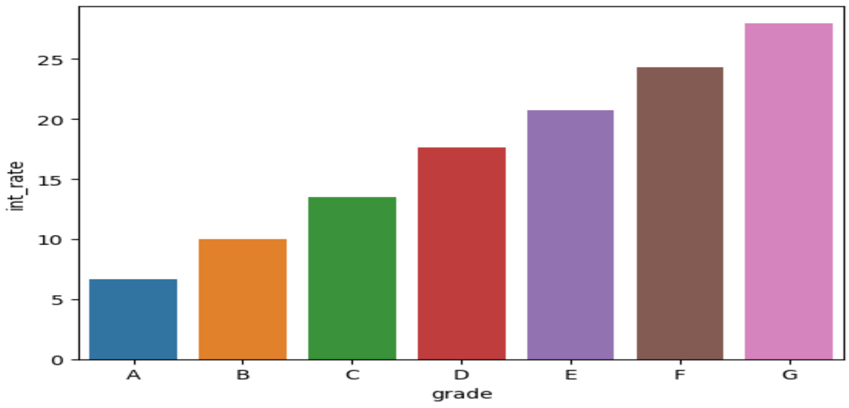


From the above distribution, we can say

1. Revol Util, all util are seems to be approximately normally distributed.
2. Loan amount, interest rate, annual income, out_prncp, open account, pub_ rec, dti and bc_open_to_buy are right skewed
3. Annual income, total_rec_late_fee, dti,delinq_2yrs, open account, out_prncp, bc_open_to_buy having high krutosis.
4. Also we can see the outliers are presemt in most of these attribites

2. Bivariate Analysis

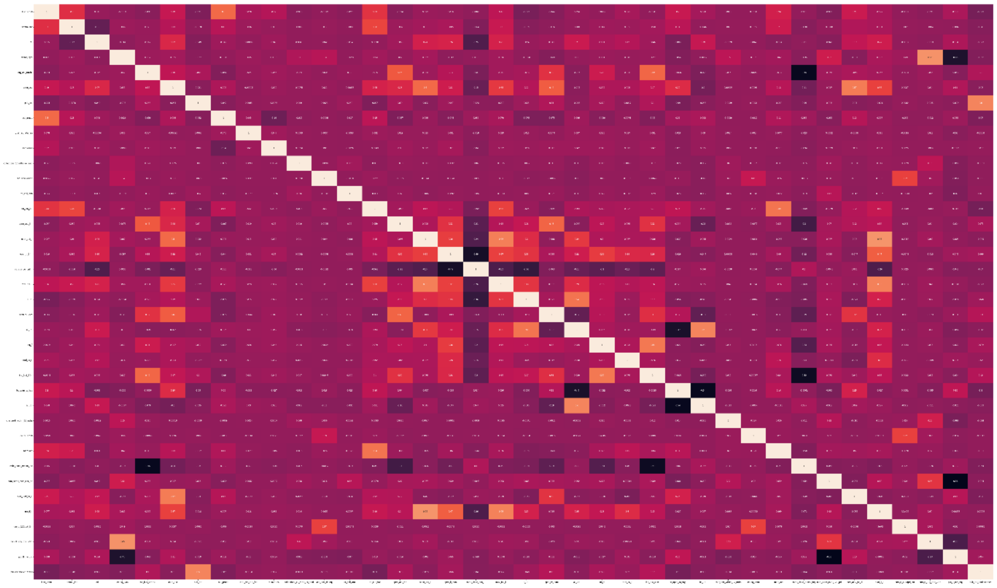
Bivariate analysis refers to the analysis of two variables to determine relationships between them. Bivariate analyses are often reported in quality-of-life research.



Grade Vs Interest Rate :

As per the above, we can say there is a direct relation between the int_rate and the grade attribute as we increase the rate of interest the grade is also changed from A to G, which means Grade “A” is a low-interest rate grade and grade “G” is the high interest rate grade.

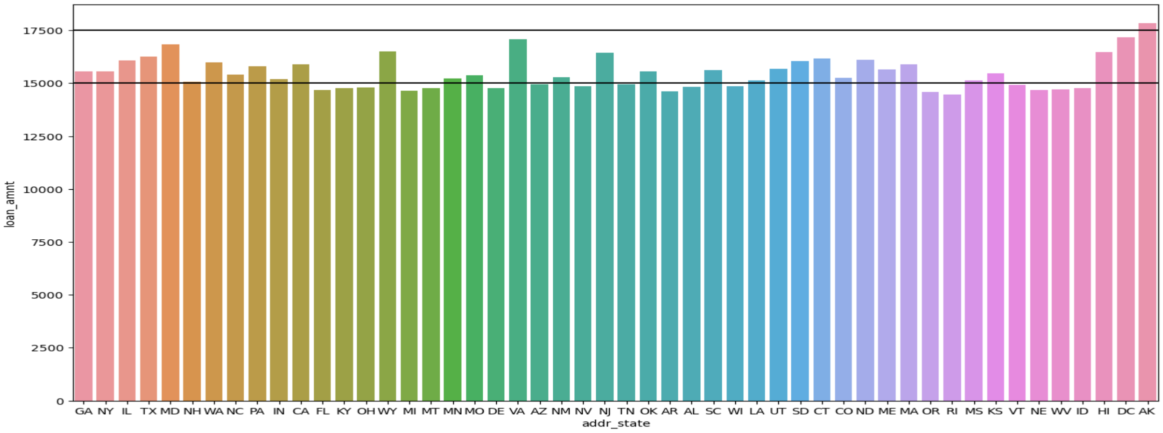
Heat Map For checking the correlation value between independent Attributes :



From this, we are able to find a high correlation between two independent columns, so we decided to drop those columns having a correlation value of more than 0.7. so we drop these columns.

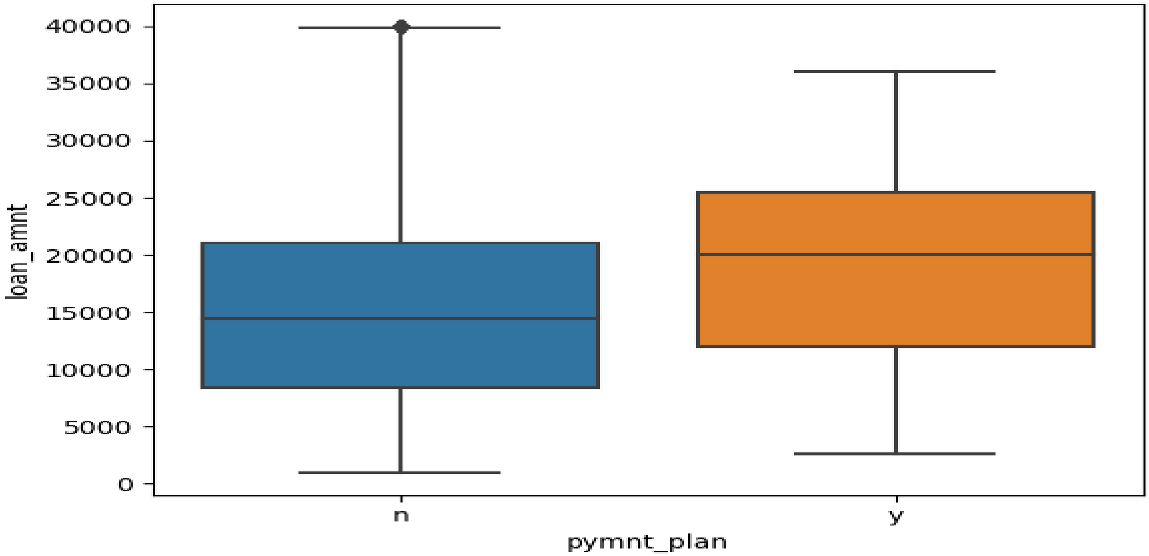
{'acc_open_past_24mths','avg_cur_bal','collection_recovery_fee', 'installment', 'last_pymnt_amnt', 'num_actv_rev_tl', 'num_bc_sats', 'm_bc_tl', 'num_op_rev_tl', 'num_rev_accts', 'num_rev_tl_bal_gt_0', 'num_sats',   'num_tl_30dpd','num_tl_op_past_12m', 'out_prncp_inv', 'percent_bc_gt_75',   'tax_liens', 'tot_hi_cred_lim', 'total_acc','total_bal_ex_mort', 'total_bc_limit', 'total_il_high_credit_limit', 'total_pymnt', 'total_rec_int', 'total_rec_prncp'}

State vs Loan Amount :



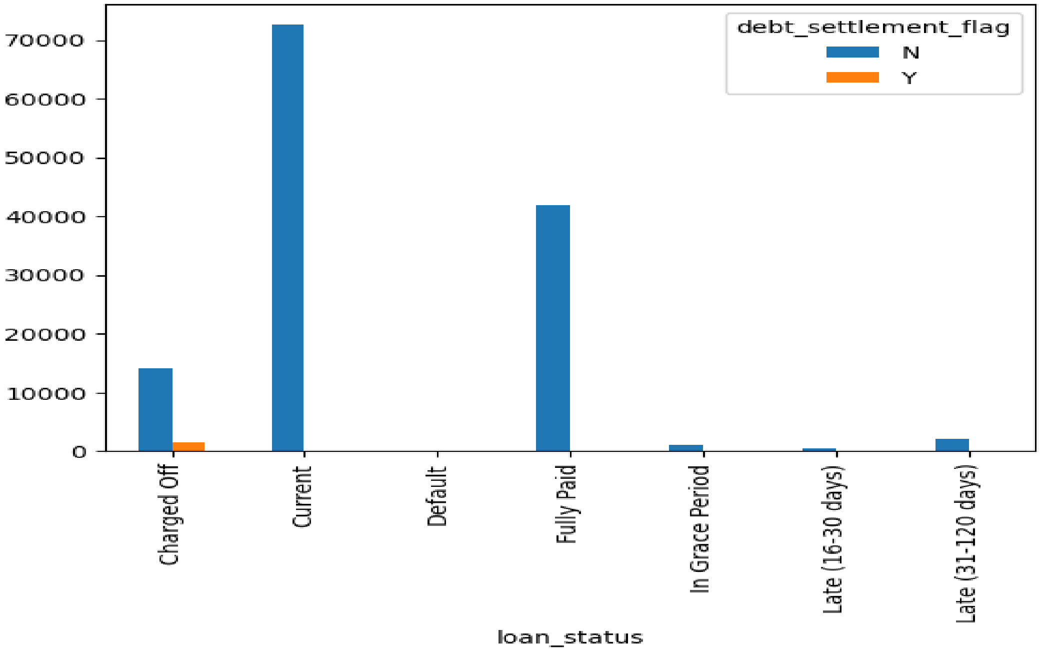
The borrower from AK state takes the maximum amount of loan which is more than 17,500 and more of the borrowers from different states are taking more than 15,000 loan amount from most of the states.

Pymnt_plan Vs Loan_amount :



We plot a boxplot for this and we found out there is an outlier present in pymnt_plan “n” above the loan amount of 40,000. Also, the “y” payment plan is slightly right skewed and the median of payment plan “n” is approx. 15,000 and the median loan amount for payment plan “y” is more which is equal to approx 20,000.

Finding the relationship between to categories (loan status and debt settlement flag) :

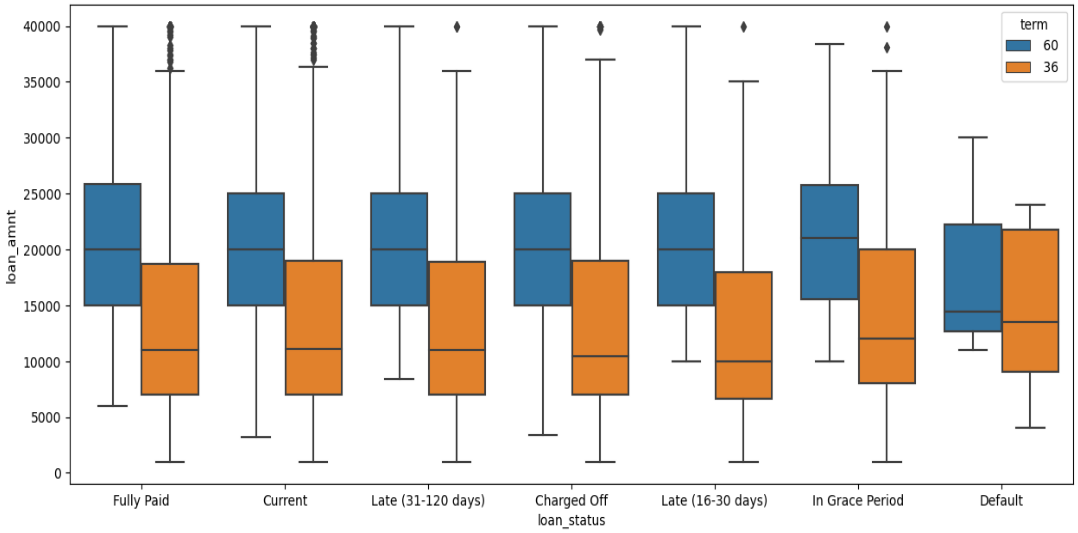


Most of the borrowers having debt settlement flags are in charged_off. And maximum the number of borrowers having no debt settlement flag is in the current loan status .

Multivariate Analysis :

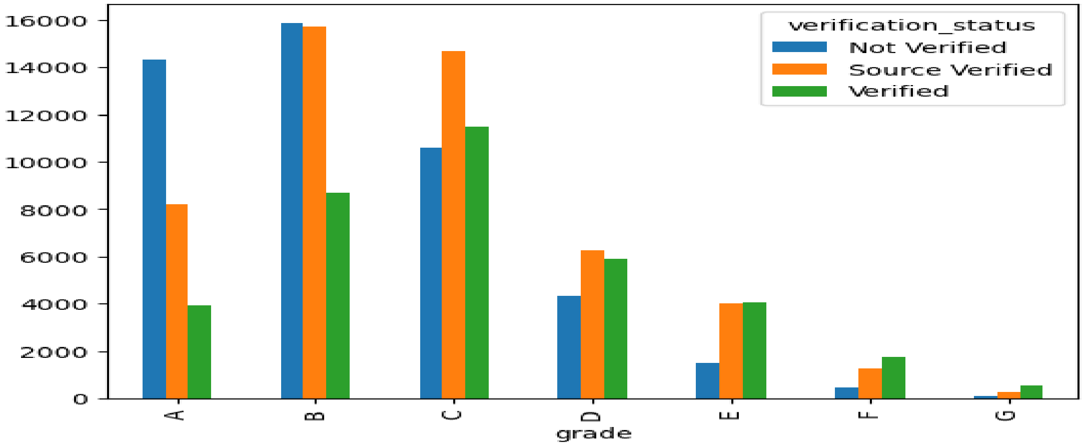
The statistical study of data where multiple measurements are made on each experimental unit and where the relationships among multivariate measurements and their structure are important.

Boxplot between Loan_status, loan_amount, and term :



We can see the outliers are present in each of the graphs, also most of the term plots are left skewed and others are right-skewed. The median of all the loan_status categories which comes in 60 months is same approx equal to 20,000 except for the default category which is approx. equal to 15,000, and the median for 36 months are less than the 60 months term which is approx. equal to 10,000.

Finding the relation between the verification status, grade ,and the number of borrowers :



From the above, we can directly say that most of the borrowers are in Grades “A”, “B” and Grade “C”. and grade B has the most number of not verified and source verified borrowers which is more than (14000+14000) approx.. 29000 or 30000 borrowers. Also, Grade A has more than 14000 borrowers who are not verified.

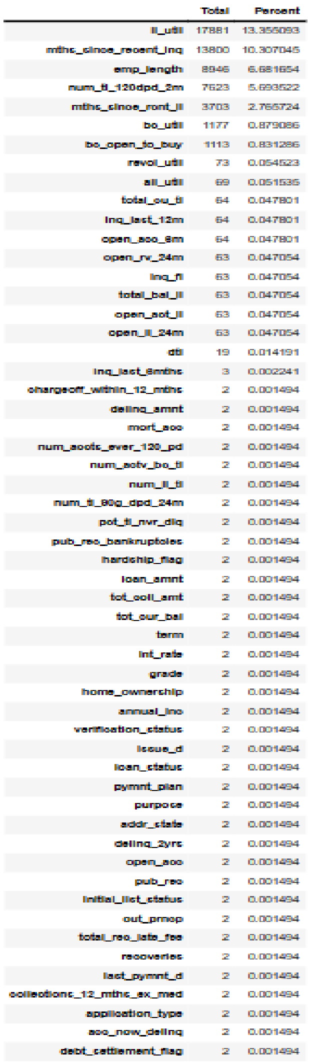
Pre-Processing Data Analysis (count of missing/ null values, redundant columns, etc.)

The dataset has missing values in most of the columns.

Dealing with Null values:

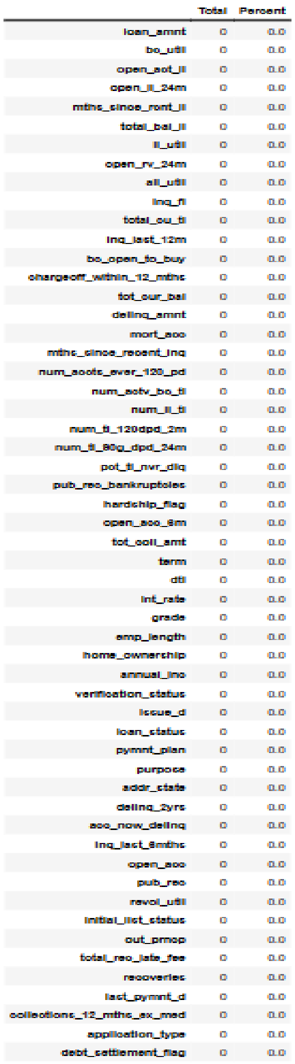
We have dropped the columns which has more than 80% null value. After that, the dataset has two columns more than 10% and the rest of the columns have less than 10% of null values.

Before



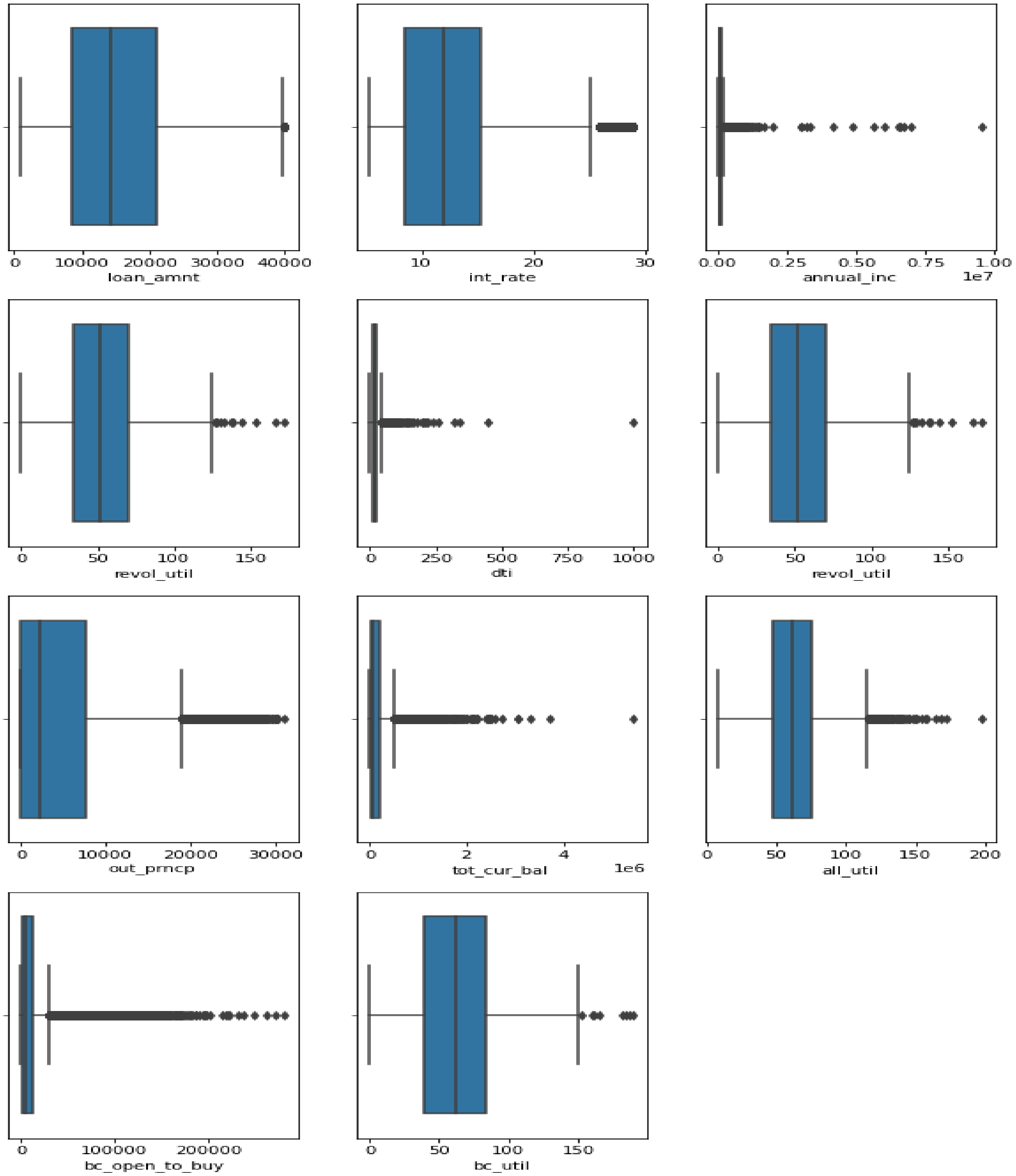
	Total	Percent
il_util	17881	13.356093
mths_since_recent_inq	13800	10.307045
emp_length	8946	6.681654
num_tl_120dpd_2m	7623	5.693522
mths_since_rcnt_il	3703	2.765724
bc_util	1177	0.879086
bc_open_to_buy	1113	0.831286
revol_util	73	0.054523
all_util	69	0.051535
total_cu_tl	64	0.047801
inq_last_12m	64	0.047801
open_acc_6m	64	0.047801
open_rv_24m	63	0.047054
inq_fi	63	0.047054
total_bal_il	63	0.047054
open_act_il	63	0.047054
open_il_24m	63	0.047054
dti	19	0.014191
inq_last_6mths	3	0.002241
chargeoff_within_12_mths	2	0.001494
delinq_amnt	2	0.001494
mort_acc	2	0.001494
num_accts_ever_120_pd	2	0.001494
num_actv_bc_tl	2	0.001494
num_il_tl	2	0.001494
num_tl_90g_dpd_24m	2	0.001494
pct_tl_nvr_dlq	2	0.001494
pub_rec_bankruptcies	2	0.001494
hardship_flag	2	0.001494
loan_amnt	2	0.001494
tot_coll_amt	2	0.001494
tot_cur_bal	2	0.001494
term	2	0.001494
int_rate	2	0.001494
grade	2	0.001494
home_ownership	2	0.001494
annual_inc	2	0.001494
verification_status	2	0.001494
issue_d	2	0.001494
loan_status	2	0.001494
pymnt_plan	2	0.001494
purpose	2	0.001494
addr_state	2	0.001494
delinq_2yrs	2	0.001494
open_acc	2	0.001494
pub_rec	2	0.001494
initial_list_status	2	0.001494
out_prncp	2	0.001494
total_rec_late_fee	2	0.001494
recoveries	2	0.001494
last_pymnt_d	2	0.001494
collections_12_mths_ex_med	2	0.001494
application_type	2	0.001494
acc_now_delinq	2	0.001494
debt_settlement_flag	2	0.001494

After



	Total	Percent
loan_amnt	0	0.0
bc_util	0	0.0
open_act_il	0	0.0
open_il_24m	0	0.0
mths_since_rcnt_il	0	0.0
total_bal_il	0	0.0
il_util	0	0.0
open_rv_24m	0	0.0
all_util	0	0.0
inq_fi	0	0.0
total_cu_tl	0	0.0
inq_last_12m	0	0.0
bc_open_to_buy	0	0.0
chargeoff_within_12_mths	0	0.0
tot_cur_bal	0	0.0
delinq_amnt	0	0.0
mort_acc	0	0.0
mths_since_recent_inq	0	0.0
num_accts_ever_120_pd	0	0.0
num_actv_bc_tl	0	0.0
num_il_tl	0	0.0
num_tl_120dpd_2m	0	0.0
num_tl_90g_dpd_24m	0	0.0
pct_tl_nvr_dlq	0	0.0
pub_rec_bankruptcies	0	0.0
hardship_flag	0	0.0
open_acc_6m	0	0.0
tot_coll_amt	0	0.0
term	0	0.0
dti	0	0.0
int_rate	0	0.0
grade	0	0.0
emp_length	0	0.0
home_ownership	0	0.0
annual_inc	0	0.0
verification_status	0	0.0
issue_d	0	0.0
loan_status	0	0.0
pymnt_plan	0	0.0
purpose	0	0.0
addr_state	0	0.0
delinq_2yrs	0	0.0
acc_now_delinq	0	0.0
inq_last_6mths	0	0.0
open_acc	0	0.0
pub_rec	0	0.0
revol_util	0	0.0
initial_list_status	0	0.0
out_prncp	0	0.0
total_rec_late_fee	0	0.0
recoveries	0	0.0
last_pymnt_d	0	0.0
collections_12_mths_ex_med	0	0.0
application_type	0	0.0
debt_settlement_flag	0	0.0

Skewness in numerical variables:



- High skewness columns are : Annual Income , Dti, out_prncp ,tot_cur_bal ,bc_open_to_buy
- Outliers are present in every numeric variable.

- We have to treat the outliers appropriately in order to get the reliable model.

To Reduce Skewness:

In our analysis, we observed that certain columns in the dataset exhibit a high degree of skewness, which can impact the accuracy and reliability of our predictive models. These columns include:

a. **Annual Income:** The distribution of annual income data is skewed, which might lead to biased predictions. Applying appropriate transformations could help mitigate this issue.

b. **Dti (Debt-to-Income Ratio):** High skewness in the debt-to-income ratio column could affect our understanding of its impact on loan eligibility. By addressing this skewness, we can ensure a more accurate representation of this crucial factor.

c. **Outstanding Principal (out_prncp)**: Skewed data in the outstanding principal column could influence our assessment of loan repayment patterns. Transforming this data could lead to more reliable insights.

d. **Total Current Balance (tot_cur_bal):** The skewed distribution of total current balance might hinder the accurate evaluation of an individual's financial stability. We should explore methods to reduce this skewness.

e. **Open to Buy Balance (bc_open_to_buy):** The skewness in the open-to-buy balance could distort our understanding of credit utilization. Addressing this skewness is vital for a well-informed analysis.

In our pursuit of refining the dataset for more accurate analysis, we undertook several crucial steps to mitigate the impact of high skewness within specific columns:

Skewness Reduction with Square Root Method: To rectify the skewed distribution within columns like Annual Income, Dti, out_prncp, tot_cur_bal, and bc_open_to_buy, we opted for the square root transformation. This transformation helped to normalize the data distribution, allowing for a more reliable and balanced representation of these variables.

Addressing Null Values: In addition to skewness, we recognized the presence of null values within these columns. To ensure completeness and data integrity, we meticulously addressed these null values through a combination of techniques such as imputation and removal, depending on the context and impact of each column.

Transformation of Skewed Columns: As part of our data preprocessing strategy, we applied the square root transformation not only to mitigate skewness but also to enhance the interpretability of the variables. By doing so, we aimed to align the distribution of these columns with the assumptions often made by statistical and machine learning models.

Our proactive approach toward skewness reduction and null value management underscores our commitment to producing accurate and reliable results.

Enhanced Analysis:

In our pursuit of refining the dataset for more effective analysis, we implemented a comprehensive data encoding strategy. This approach aimed to transform categorical variables into a suitable numerical format, facilitating the integration of these variables into our analytical processes.

1. Categorical Data Transformation:

As a preliminary step, we recognized the importance of categorizing specific columns for a clearer representation of the data. Columns such as "term," "grade," "emp_length," "home_ownership," "verification_status," "loan_status," "pymnt_plan," "purpose," "addr_state," "initial_list_status," "application_type," "acc_now_delinq," "hardship_flag," and "debt_settlement_flag" were transformed into the categorical data type. This transformation streamlined the data, ensuring that categorical attributes were treated appropriately.

2. One-Hot Encoding:

For categorical variables requiring a more extensive representation, we adopted one-hot encoding. By employing the `pd.get_dummies` function, we expanded the categorical columns, including "home_ownership," "verification_status," "issue_d," "pymnt_plan," "purpose," "addr_state," "initial_list_status," "last_pymnt_d," "application_type," "hardship_flag," and "debt_settlement_flag." This process generated binary-encoded columns, enabling a comprehensive analysis while avoiding multicollinearity through the `drop_first=True` parameter.

3. Label Encoding:

Complementing our data transformation efforts, we employed label encoding for specific variables. The "grade" column underwent label encoding using the `LabelEncoder` class from the Scikit-Learn library. This method translated categorical grades into numerical values, enhancing the compatibility of this attribute with our analytical models.

Through these meticulous data encoding techniques, we've not only standardized the data but prepared it for seamless integration into our machine learning and statistical models. These encoding strategies enhance the depth and accuracy of our analysis, enabling us to glean more meaningful insights from the dataset.

4. Feature Selection:

In pursuit of an optimized model architecture, we embarked on a meticulous feature selection journey. Our approach commenced with the division of our dataset into feature variables (X) and the target variable (y) using the train_test_split function. By isolating the 'loan_status' column as our target variable, we established a clear focal point for our analysis.

Subsequently, we curated a select subset of features that we deemed most influential in the prediction of loan status. These features encompassed critical financial indicators, including ["loan_amnt," "int_rate," "annual_inc," "dti," "delinq_2yrs," "open_acc," "pub_rec," "bc_open_to_buy," "bc_util," "revol_util," "out_prncp," "total_rec_late_fee," "recoveries," "tot_cur_bal," "tot_coll_amt," "total_bal_il," and "delinq_amnt."]. This curated feature set forms the cornerstone of our model's predictive capacity.

To ensure uniformity and accurate model convergence, we standardized the selected features using the **StandardScaler** from the **Scikit-learn library**. This scaling process eliminates the potential impact of differing magnitudes among features, resulting in a balanced and equitable contribution from each attribute during model training.

Our strategic approach to feature selection, curation, and scaling amplifies the predictive precision of our models. This careful refinement ensures that our models are trained on the most pertinent variables, thus empowering them to yield sharper insights into loan status prediction based on essential financial factors.

4. Balancing Dataset:

Achieving Balanced Data through Oversampling:

In our quest for robust model performance, addressing data imbalance emerged as a pivotal concern. To counter this challenge, we employed the **RandomOverSampler** from the **imbalanced-learn library**. With a sampling strategy set to balance the dataset, we **oversampled** the minority class, ensuring parity between classes. The resultant dataset, depicted through a pie chart, exhibited a harmonized distribution, laying the foundation for more accurate and unbiased model training.

Strategic Train-Test Split for Model Evaluation:

After achieving balanced data, we partitioned the dataset into training and testing subsets using the **train_test_split** function. A **70-30 split ratio** was chosen to ensure an optimal balance between model training and assessment. This process allowed us to evaluate model performance on unseen data, facilitating a robust evaluation of our predictive models.

Informed Feature Selection through Mutual Information:

Feature selection played a decisive role in enhancing model interpretability and performance. To this end, we utilized mutual information via the **mutual_info_classif** function to quantify the relationships between features and the target variable. By ranking features based on their mutual information scores, we gained insights into their predictive significance. This informed our selection of the top features, which were further refined using **SelectKBest**, yielding a curated set of features that demonstrated the strongest predictive potential.

Model Building:

Supervised Learning: Problem Statement I

Exploring Model Efficacy Through Diverse Approaches

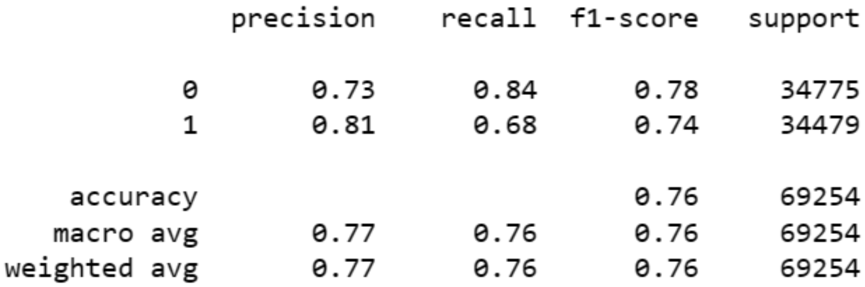
For comprehensive analysis, we now try model building in a range of supervised and unsupervised algorithms. By employing an ensemble of methodologies, we are poised to unravel the intricate patterns within our dataset. Through rigorous evaluation and comparison of these models, we aim to find the optimal predictive framework that attains the highest accuracy in forecasting loan status.

Model 1: Decision Tree Classifier using top 5 mutual info Features

Leveraging the Decision Tree Classifier, our investigation into loan status prediction took an insightful turn. Concentrating on features meticulously extracted through mutual information, this model showcased promising predictive potential.

Notably, our model yielded compelling results:
- AUC Score: 0.7636
- Cohen's Kappa Score: 0.527582
- Accuracy: 76.40%

These metrics signify a strong model fit, indicating its aptitude in classifying loan statuses. Additionally, with a balanced precision and recall trade-off, this model is adept at identifying positive cases while curbing false positives. This underscores the model's robustness in navigating the complex landscape of loan eligibility prediction.

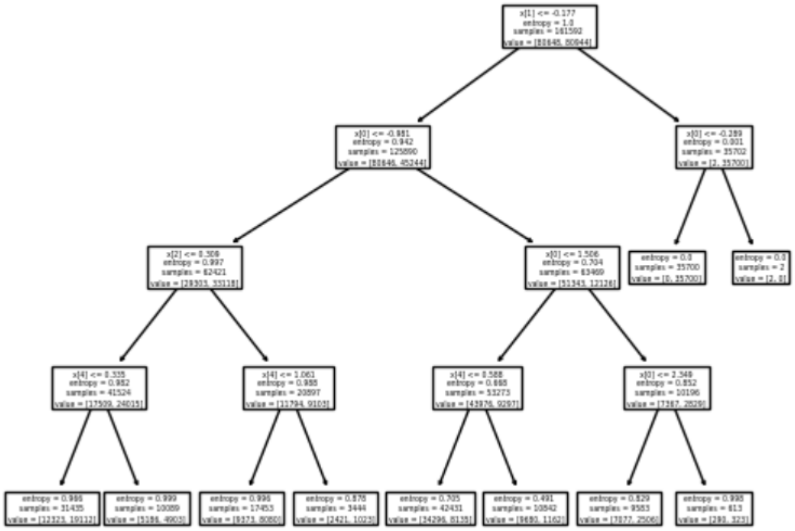


```
              precision    recall  f1-score   support

           0       0.73      0.84      0.78     34775
           1       0.81      0.68      0.74     34479

    accuracy                           0.76     69254
   macro avg       0.77      0.76      0.76     69254
weighted avg       0.77      0.76      0.76     69254
```

Classification Report



Decision Tree

Model 2: Decision Tree Classifier using all Features

Exploring a different approach, we utilized all available features in the Decision Tree Classifier, resulting in an impressive 95.07% accuracy on the test dataset, showcasing the model's exceptional precision in loan status prediction.

Furthermore, the metrics underscore this model's remarkable performance:
- AUC Score: 0.9509
- Cohen's Kappa Score: 0.901501
- Accuracy: 95.07%

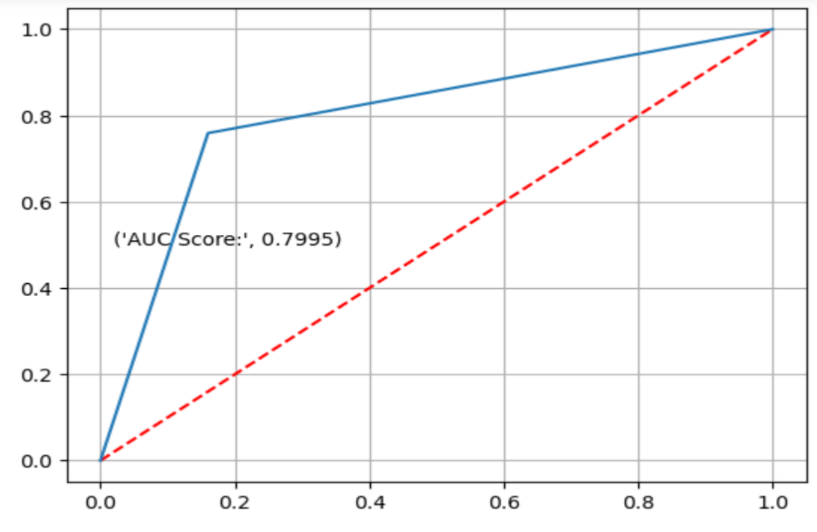
The model excels in reducing false positives and capturing positive cases, although signs of overfitting are observed due to a perfect training score of 1.0, underscoring the dataset's richness and guiding us towards the optimal predictive model.

Model 3: Decision Tree Classifier with Decision Tree best Features

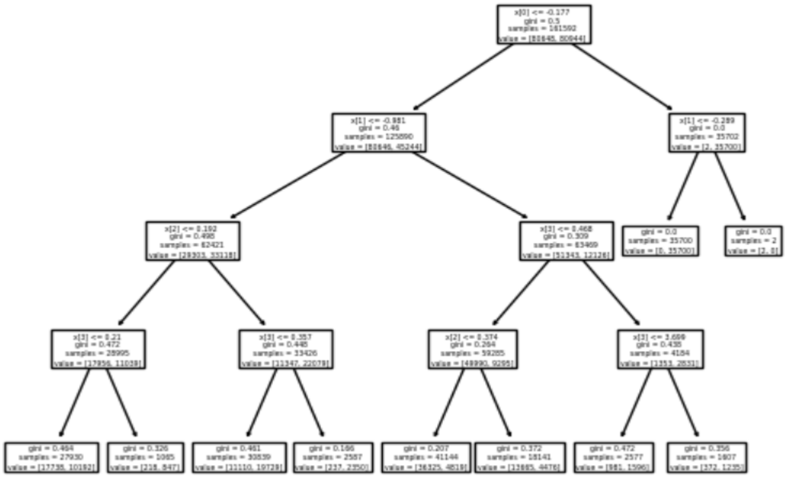
The Decision Tree (Feature selected by decision tree) model demonstrates an accuracy of 80.02%, effectively navigating data relationships, and maintaining a balance between precision and recall, thus forging an innovative predictive framework.

The updated model statistics:
- Training Score: 0.7995
- Test Score: 0.7997
- Cohen's Kappa Score: 0.599274

This reinforces the model's status as a good-fit for our predictive purposes.



AUC is good but there is chance of improvment .



Decision Tree

Model 4: Logistic Regression

Driven by probabilities, the Logistic Regression model provides crucial insights into loan outcomes. It maintains an impressive accuracy of 80.24%, balancing precision and recall effectively, thus demonstrating a dependable predictive ability.

The model's exploration of feature-outcome relationships, depicted through coefficients, and its ability to decipher intricate data patterns, including the 'intercept' term, contribute to its comprehensive understanding. Its precision is encapsulated by the confusion matrix, where 29,121 true negatives and 26,653 true positives reinforce its accuracy in case classification.

The updated model statistics:
- AUC Score: 0.8024
- Cohen's Kappa Score: 0.604937
- Accuracy: 80.25%

This firmly establishes the model as a good-fit for our predictive aspirations.

Model 5: Random Forest Classifier

The Random Forest Classifier shines with a remarkable accuracy of 79.71%, showcasing its consistent predictive strength. Its precision and recall scores of 79.72% each confirm its ability for accurate predictions and capturing true positives.
The model's harmonious balance between precision and recall is further supported by macro and weighted average F1-scores of 79.72%. This synergy of decision trees in the Random Forest manifests as a powerful instrument, deciphering loan status intricacies and solidifying its role as a reliable and insightful predictive tool.

The updated model statistics:
- Training Score: 0.7971
- Test Score: 0.7972
- Cohen's Kappa Score: 0.594334

This also indicates that the model exhibits a low degree of overfitting, further enhancing its credibility as a valuable predictive tool.

Model 6: Random Forest Classifier with important Feature

The Random Forest Classifier, featuring important attributes, showcases an exceptional precision with a training score of 100% and an impressive 98.5% test score. Its equilibrium between precision and recall, a Cohen's Kappa Score of 0.970488, and high F1-scores underline its robustness. This model's ability to generalize, low over-fit, and intricate understanding of loan statuses reinforce its pivotal role in data-driven decision-making.

The updated model statistics:
- Training Score: 1.0
- Test Score: 0.9853
- Cohen's Kappa Score: 0.970488

This reinforces its status as a dependable and powerful predictive model, equipped to navigate the complex landscape of loan eligibility prediction.

Model 7: Random Forest Classifier using Grid Search CV

In the pursuit of refining our predictive capabilities, we harnessed the power of the Random Forest Classifier through an advanced technique known as Grid Search CV. This approach allowed us to fine-tune the model's parameters for optimal performance.

The results were striking:
- Best Score: 98.17%
- Training Score: 100%
- Test Score: 100%
- Cohen's Kappa Score: 0.971787

This achievement attests to the model's extraordinary prowess, achieving impeccable accuracy and precision. The macro and weighted average F1-scores of 98.59% further validate its robustness.

By leveraging the synergy of the Random Forest with the finesse of Grid Search CV, this model exemplifies a strong fit for our predictive needs. It stands as a testament to our commitment to harnessing cutting-edge techniques for accurate and insightful loan status predictions.

Boosting :

Boosting 1: ADA Boost All Features

Employing the Ada Boost algorithm, our model construction journey reached new heights of accuracy and precision. By leveraging the power of boosting, we enhanced our predictive framework with a focus on all available features.

The model's performance speaks volumes:
- Accuracy: 81.43%
- Cohen's Kappa Score: 0.628924

This accomplishment highlights the model's adeptness in precise loan status prediction and its capacity to balance accuracy and precision, making it a valuable tool. Utilizing Ada Boost signifies a significant step toward a balanced and advanced predictive approach for loan eligibility assessment.

Boosting 2: ADA Boost using Grid Search CV

Our endeavor for predictive accuracy led us to Ada Boost algorithm coupled with Grid Search CV, resulting in impressive outcomes. This model's "Good Fit" showcases a harmonious balance between accuracy and precision, advancing loan status predictions and eligibility assessments.

Key Results:
- Accuracy: 81.41%
- Cohen's Kappa Score: 0.628374

Boosting 3: ADA Boost using RFE Rank Feature and Randomized CV

By incorporating RFE technique into Ada Boost, our model strategically selects features, achieving a balanced "Good Fit" performance that navigates intricate relationships between features and outcomes, enriching our predictive capabilities despite slightly lower accuracy.

Key Achievements:
- Accuracy: 57.76%
- Cohen's Kappa Score: 0.155250

Boosting 4 : XGBOOST All Numeric Features

Employing the XGBOOST algorithm, our model construction journey reached new heights of accuracy and precision. By leveraging the power of boosting, we enhanced our predictive framework with a focus on all available features.

Key Achievements:
- Accuracy: 86.96%
- Cohen's Kappa Score: 0.739358

Boosting 5 : XGBOOST With All Numeric Feature Randomized Search CV

Our endeavor for predictive accuracy led us to XGBOOST algorithm coupled with Randomized Search CV, resulting in impressive outcomes. This model's "Good Fit" showcases a harmonious balance between accuracy and precision, advancing loan status predictions and eligibility assessments.

Key Results:
- Accuracy: 87.74%
- Cohen's Kappa Score: 0.75493

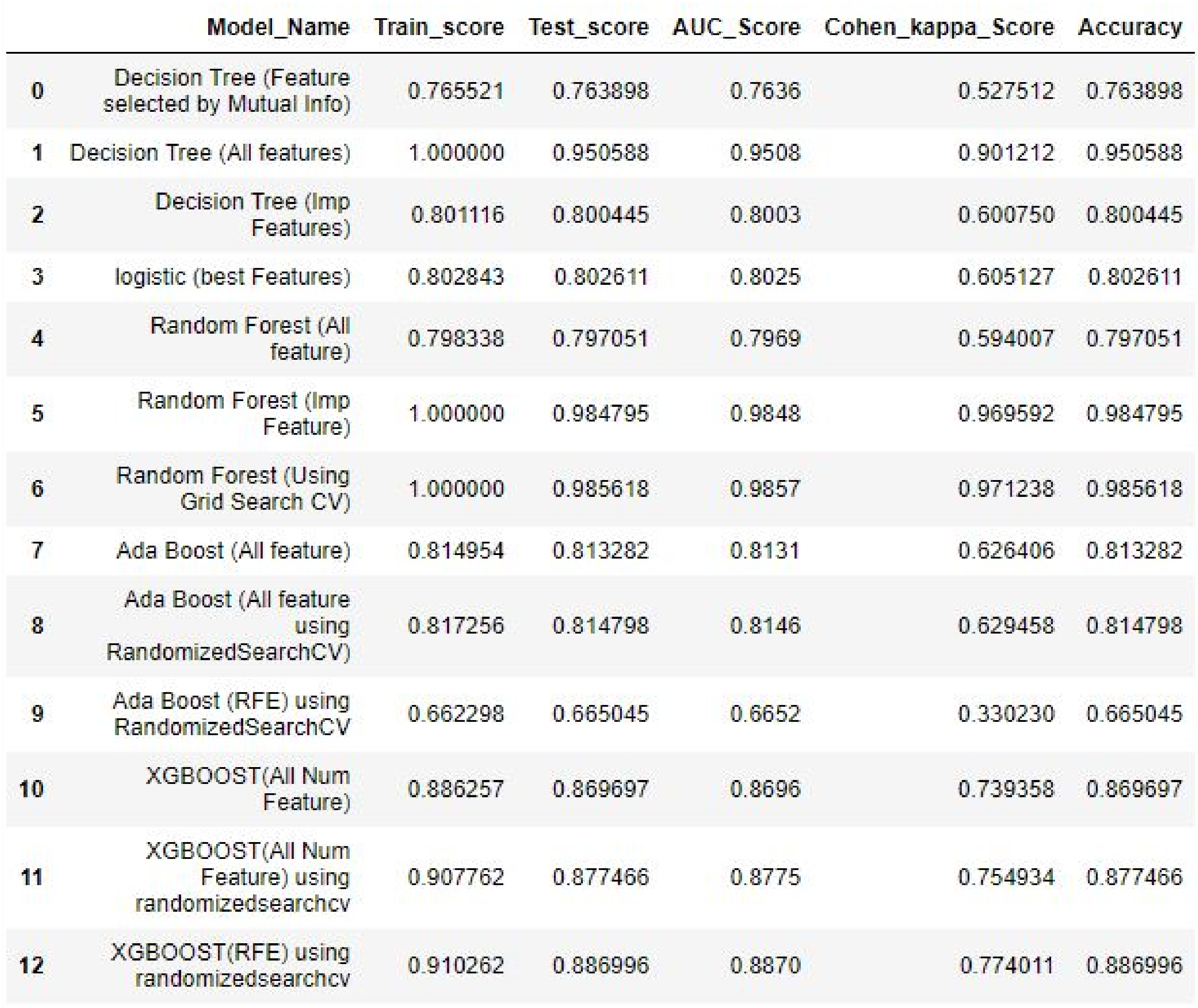
Boosting 6 : XGBOOST Using RFE Rank Feature with Randomized Search CV

By incorporating RFE technique into XGBOOST, our model strategically selects features, achieving a balanced "Good Fit" performance that navigates intricate relationships between features and outcomes, enriching our predictive capabilities despite slightly lower accuracy.

Key Achievements:
- Accuracy: 88.699%
- Cohen's Kappa Score: 0.774011

Model Performances :

Our model evaluation highlights varied predictive strengths. Decision Trees showed accuracy with precision-recall trade-offs. Logistic Regression exhibited balance, and ensemble methods like Random Forest excelled in capturing positives. Ada Boost, with and without Grid Search, showcased good predictive accuracy. Ada Boost with RFE indicated feature importance. These insights steer us towards a robust loan eligibility predictive framework, considering accuracy, precision, and balance.



	Model_Name	Train_score	Test_score	AUC_Score	Cohen_kappa_Score	Accuracy
0	Decision Tree (Feature selected by Mutual Info)	0.765521	0.763898	0.7636	0.527512	0.763898
1	Decision Tree (All features)	1.000000	0.950588	0.9508	0.901212	0.950588
2	Decision Tree (Imp Features)	0.801116	0.800445	0.8003	0.600750	0.800445
3	logistic (best Features)	0.802843	0.802611	0.8025	0.605127	0.802611
4	Random Forest (All feature)	0.798338	0.797051	0.7969	0.594007	0.797051
5	Random Forest (Imp Feature)	1.000000	0.984795	0.9848	0.969592	0.984795
6	Random Forest (Using Grid Search CV)	1.000000	0.985618	0.9857	0.971238	0.985618
7	Ada Boost (All feature)	0.814954	0.813282	0.8131	0.626406	0.813282
8	Ada Boost (All feature using RandomizedSearchCV)	0.817256	0.814798	0.8146	0.629458	0.814798
9	Ada Boost (RFE) using RandomizedSearchCV	0.662298	0.665045	0.6652	0.330230	0.665045
10	XGBOOST(All Num Feature)	0.886257	0.869697	0.8696	0.739358	0.869697
11	XGBOOST(All Num Feature) using randomizedsearchcv	0.907762	0.877466	0.8775	0.754934	0.877466
12	XGBOOST(RFE) using randomizedsearchcv	0.910262	0.886996	0.8870	0.774011	0.886996

From the about score card, I can clearly conclude that the random forest with best feature using Grid search CV is the best model for our project.

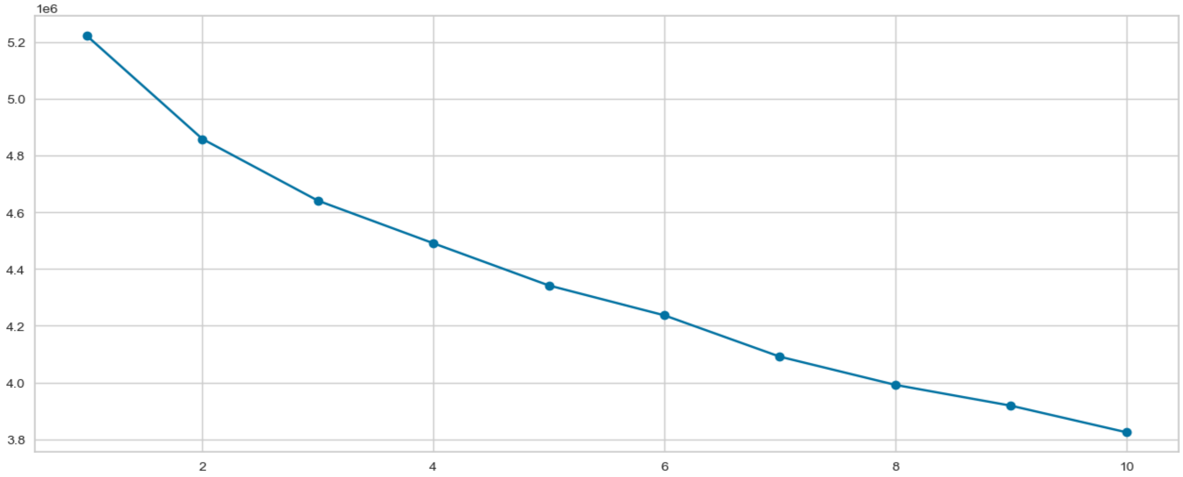
Unsupervised Learning Model: Problem Statement II

Clusters / Labels Formation:

1. **KMeans Clustering:** k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster.

   ### Finding out optimal number of cluster:

   1. **Elbow Plot:** WCSS is the sum of the squared distance between each point and the centroid in a cluster. When we plot the WCSS with the K value, the plot looks like an Elbow. As the number of clusters increases, the WCSS value will start to decrease.

      From the graph, It is hard to predict the optimal clusters. But we get an idea that the cluster may in the range or 3-8.

   2. **Silhoutte Score :** The value of the silhouette coefficient is between [-1, 1]. A score of 1 denotes the best, meaning that the data point i is very compact within the cluster to which it belongs and far away from the other clusters. The worst value is -1. Values near 0 denote overlapping clusters.
      From the Silhoutte Score we conclude that the optimal number of clusters is 4.

2. **DBSCAN:** DBSCAN is a density-based clustering algorithm that works on the assumption that clusters are dense regions in space separated by regions of lower density. It groups 'densely grouped' data points into a single cluster.
With the help of DBSCAN we create 7 labels {-1,0,1,2,3,4,5,} where the -1 is the noise (Outliers in data) and than we used label encoding to label the clusters.

PCA (Principal Component Analysis): Principal component analysis, or PCA, is a dimensionality reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

With the help of PCA , We reduced the dimensions from 39 features to 18 features.

Model Building: Using Kmeans Labels

1.Model 1: Random Forest Classifier

The Random Forest Classifier shines with a remarkable accuracy of 97.89%, showcasing its consistent predictive strength. Its precision and recall scores of approx 97% each confirm its ability for accurate predictions and capturing true positives.
The model's harmonious balance between precision and recall is further supported by macro and weighted average F1-scores of 97%. This synergy of decision trees in the Random Forest manifests as a powerful instrument, deciphering loan status intricacies and solidifying its role as a reliable and insightful predictive tool.

The updated model statistics:
- Training Score: 1.0
- Test Score: 0.978987
- Accuracy: 97.89%

2. Model 2: Random Forest with Important features using randomized search cv

With the help of Important features and randomized search cv, we are try to increase the accuracy of the model ans with the help of randomized search cv trying to make the model better fit . From this we get these score.

The updated model statistics:
- Training Score: .93872
- Test Score: 0.93616
- Accuracy: 93.616%

3. Model 3 : XGBOOST with all features

Appling the XGBOOST algorithm, our model construction journey reached new heights of accuracy and precision. By leveraging the power of boosting, we enhanced our predictive framework with a focus on all available features.

The updated model statistics:
- Training Score: .1.0
- Test Score: 0.990464
- Accuracy: 99.046%

4. Model 4 : XGBOOST using randomized search cv

Our endeavor for predictive accuracy led us to XGBOOST algorithm coupled with Randomized Search CV, resulting in impressive outcomes. This model's "Good Fit" showcases a harmonious balance between accuracy and precision, advancing loan status predictions and eligibility assessments.

The updated model statistics:
- Training Score: 0.9873
- Test Score: 0.98192
- Accuracy: 98.192%

Model Building: Using CBSCAN Labels

1. Model 1 : Random Forest with all feature

The Random Forest Classifier shines with a remarkable accuracy of 97.89%, showcasing its consistent predictive strength. Its precision and recall scores of approx 97% each confirm its ability for accurate predictions and capturing true positives.

The model's harmonious balance between precision and recall is further supported by macro and weighted average F1-scores of 97%. This synergy of decision trees in the Random Forest manifests as a powerful instrument, deciphering loan status intricacies and solidifying its role as a reliable and insightful predictive tool.

The updated model statistics:
- Training Score: 1.0
- Test Score: 0.9935
- Accuracy: 99.357%

2. Model 2 : Random Forest with best features using randomized search cv

With the help of Important features and randomized search cv, we are try to increase the accuracy of the model ans with the help of randomized search cv trying to make the model better fit . From this we get these score.

The updated model statistics:
- Training Score: 0.9871
- Test Score: 0.9871
- Accuracy: 98.712%

3. Model 3 : XGBOOST with all features

Appling the XGBOOST algorithm, our model construction journey reached new heights of accuracy and precision. By leveraging the power of boosting, we enhanced our predictive framework with a focus on all available features.

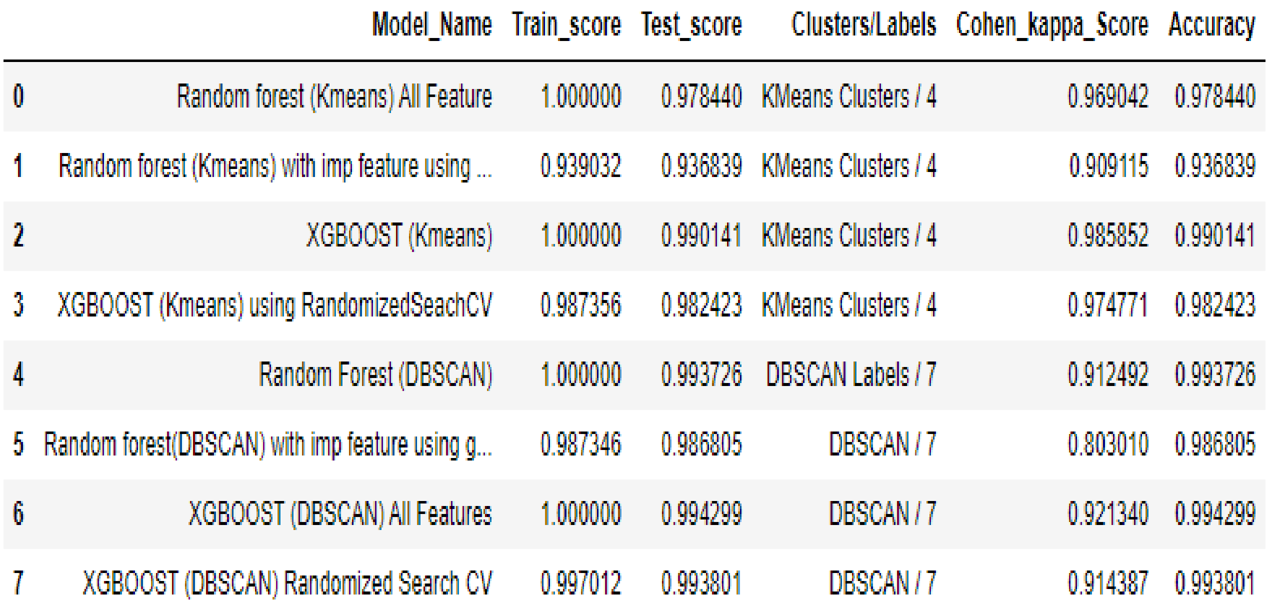
The updated model statistics:
- Training Score: .1.0
- Test Score: 0.99427
- Accuracy: 99.46%

4. Model 4 : XGBOOST using randomized search cv

Our endeavor for predictive accuracy led us to XGBOOST algorithm coupled with Randomized Search CV, resulting in impressive outcomes. This model's "Good Fit" showcases a harmonious balance between accuracy and precision, advancing loan status predictions and eligibility assessments.

The updated model statistics:
- Training Score: 0.9981
- Test Score: 0.9939
- Accuracy: 99.39%

Model Performances :



	Model_Name	Train_score	Test_score	Clusters/Labels	Cohen_kappa_Score	Accuracy
0	Random forest (Kmeans) All Feature	1.000000	0.978440	KMeans Clusters / 4	0.969042	0.978440
1	Random forest (Kmeans) with imp feature using ...	0.939032	0.936839	KMeans Clusters / 4	0.909115	0.936839
2	XGBOOST (Kmeans)	1.000000	0.990141	KMeans Clusters / 4	0.985852	0.990141
3	XGBOOST (Kmeans) using RandomizedSeachCV	0.987356	0.982423	KMeans Clusters / 4	0.974771	0.982423
4	Random Forest (DBSCAN)	1.000000	0.993726	DBSCAN Labels / 7	0.912492	0.993726
5	Random forest(DBSCAN) with imp feature using g...	0.987346	0.986805	DBSCAN / 7	0.803010	0.986805
6	XGBOOST (DBSCAN) All Features	1.000000	0.994299	DBSCAN / 7	0.921340	0.994299
7	XGBOOST (DBSCAN) Randomized Search CV	0.997012	0.993801	DBSCAN / 7	0.914387	0.993801

From the above score card, we conclude that the labels created by DBSCAN are the best label with XGBOOST technique using randomized search cv we get the best performance with more tha 99% accuracy.

We are able to divide the data on the basic for dbscan, here is the insights:

1. **Label "-1" :** Are the outliers in the data.
2. **Label "0" :** Average loan given to this group is approx =15624.34 , Minimum loan given =1000 , Maximum loan given=40000.
3. **Label "1" :** Average loan given to this group is approx= 13521.47, Minimum loan given =1000, Maximum loan given = 40000.
4. **Label "2" :** Average loan given to this group is approx=16147.36, Minimum loan given =1000, Maximum loan given = 40000.
5. **Label "3" :** Average loan given to this group is approx= 15373.92, Minimum loan given =1000, Maximum loan given = 40000.
6. **Label "4" :** Average loan given to this group is approx= 15866.89, Minimum loan given =2400, Maximum loan given = 35000.
7. **Label "5" :** Average loan given to this group is approx= 13587.11, Minimum loan given =2400, Maximum loan given = 35000

Summary:

Our capstone project journey:

Exploration of Loan Eligibility Prediction: We delved into the nuanced landscape of loan eligibility prediction, utilizing the intricate tapestry of banking history as a pivotal compass.

Multifaceted Predictive Modeling: Employing a rich spectrum of predictive models, our analysis spanned Decision Trees, Logistic Regression, and the formidable Random Forest Classifier.

Feature Identification and Refinement: Our quest led to the identification of key features, expertly harnessing advanced machine learning techniques to weave predictive frameworks.

Distinct Model Strengths: The diversity of models we employed unveiled a tapestry of strengths, each tailored to specific nuances within the data and lending domain.

Deeper Insights into Financial Variables: Our journey deepened our understanding of the intricate dance between financial variables and loan outcomes, shedding light on their pivotal role.

Recommendations:

Drawing from our findings, we present a set of strategic recommendations to elevate loan eligibility evaluation:

Leveraging Random Forest's Accuracy: Exploit the remarkable accuracy of the Random Forest Classifier. Employ it as a robust tool to expedite the approval process for applicants with low risk profiles, optimizing resource allocation and facilitating swifter decision-making.

Synergizing Logistic Regression and Decision Trees: Combine the probabilistic insights of Logistic Regression with the interpretability of Decision Tree models. This synthesis fosters a well-rounded strategy, achieving a delicate equilibrium between precision and recall. This balanced approach ensures accurate classification while minimizing false positives.

Enhancing Decision-Making Efficiency: Implementing these strategies not only accelerates the loan approval process but also empowers financial institutions to make informed decisions. Such informed decision-making resonates positively with applicants, fostering transparency and trust.

Empowering Loan Applicants: These measures offer loan applicants clearer visibility into the loan approval process. By refining the predictability of their application's outcome, applicants can make informed financial decisions and chart more definite pathways toward securing loans.

Limitations :

While our analysis has provided valuable insights, it's imperative to recognize inherent limitations:

Dataset Quality: The efficacy of our predictive models is intrinsically tied to the dataset's quality and comprehensiveness. Data integrity, accuracy, and completeness play a pivotal role in the models' performance and predictive power.

External Factors Omitted: Our analysis focused on the immediate banking history attributes. External influences, like macroeconomic trends, regulatory changes, or unforeseen events, weren't encompassed within our scope. These factors often exert a significant impact on loan decisions, necessitating future considerations.

Path to Refinement: To enhance the accuracy and robustness of our predictions, exploration into the omitted external factors is recommended. Incorporating these variables could elevate the models' ability to adapt and predict in dynamic real-world scenarios.

Continuous Iteration: Our limitations illuminate a path of ongoing refinement. Continued model iteration and augmentation with additional variables could pave the way for a more comprehensive and resilient predictive framework.

Implications & Applications:

The implications of our predictive models transcend statistics, yielding practical benefits:

Loan Processing Optimization: Financial institutions can harness our models to streamline loan processing. By accurately assessing loan eligibility, institutions can expedite approvals and allocate resources more efficiently, resulting in enhanced operational agility.

Risk Mitigation: These models offer a potent tool to mitigate risks linked to defaults. Informed decisions driven by predictive insights empower institutions to minimize potential losses and uphold a healthy loan portfolio.

Equitable Credit Access: A significant societal impact lies in the models potential to promote equitable credit access. By objectively evaluating loan applicants, regardless of subjective biases, financial institutions can extend timely support to deserving candidates, fostering economic inclusivity.

Prudent Lending Approach: The models deployment fosters a balanced lending approach. Timely financial support is extended to individuals with genuine creditworthiness, aligning with institutions' fiscal responsibility and ethical lending practices.

Our models offer a transformative potential that transcends numbers, shaping a financial landscape characterized by efficiency, fairness, and prudent risk management.

Future Directions:

Looking ahead, possibilities for exploration abound:

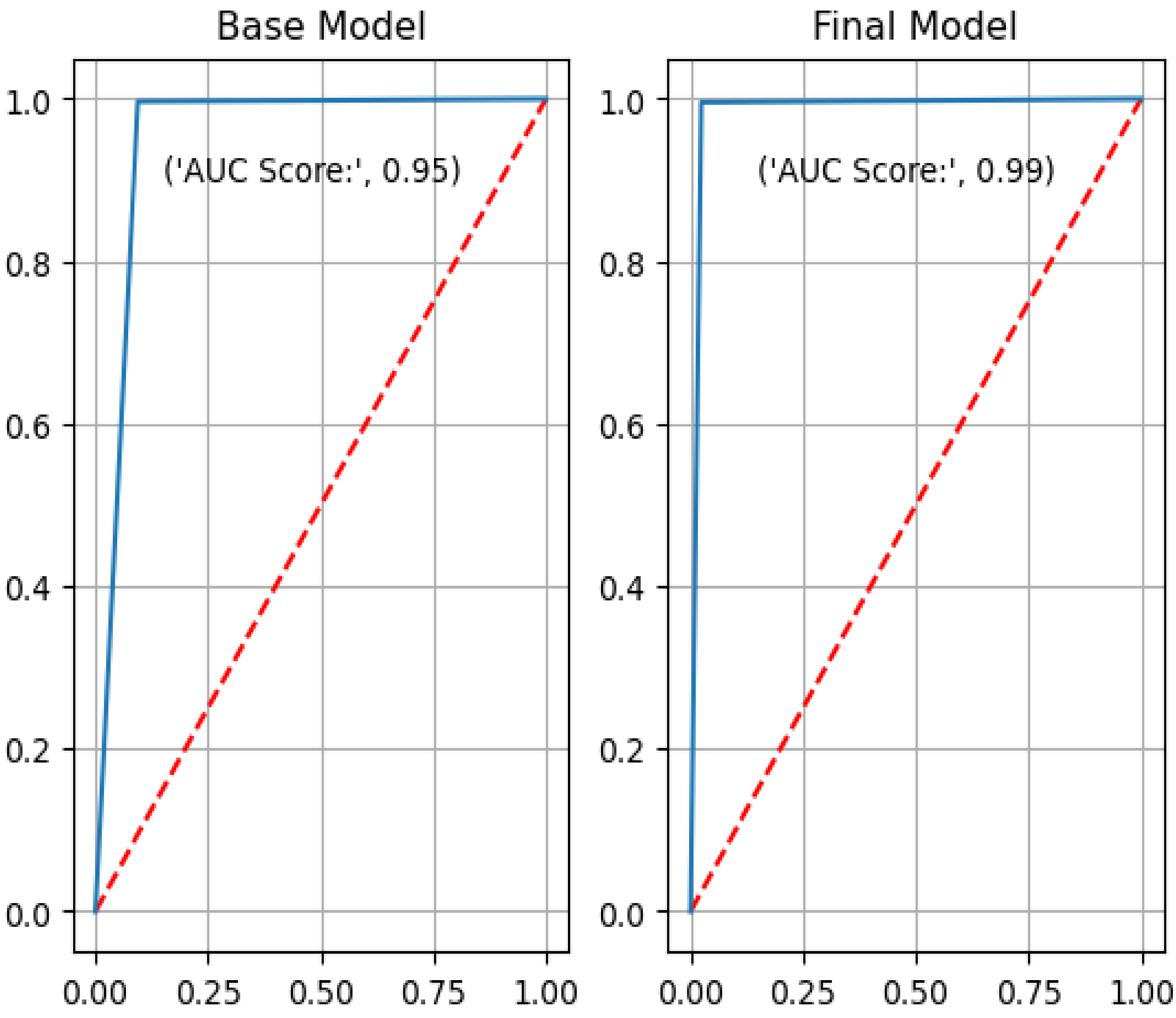
Macroeconomic Indicators: Integration of macroeconomic data could fortify model resilience, capturing broader economic influences.

Real-Time Data: Incorporating real-time data feeds would infuse dynamic accuracy into predictions, aligning with evolving scenarios.

Advanced Ensemble Techniques: Delving into advanced methodologies like Gradient Boosting could unlock heightened predictive potential, fine-tuning outcomes.

Deep Learning Algorithms: Harnessing deep learning algorithms offers a pathway to unravel intricate patterns within loan eligibility, enriching insights.

Comperison: Classification Model



- From here also We can say that our final model performing better than the base model. Cause Auc score final model is 0.99 which is greater than 0.95.
- We check the model score of base model and the final model to check the model fitness. So our base model score 1 in train data but gives .95 in the test data which conclude that the model is overfitted model. Where as our final model gives model score 1 in train and .99 in test which conclude that the model is a good fit model.
- From the cohen kappa score our base model gives a score = .9015 and our final model give a score =0.9717. Which show our final model performance is much better than the base model.
- Base model is made by 117 features where as out final model is made by only 15 features.

Conclusion:

In conclusion, our capstone project is like a big adventure where we used numbers and patterns to understand who should get loans. It's kind of like being a detective, but with math. We tried out different tools, like Decision Trees, Logistic Regression, and Random Forests, to help us see the future and decide who should get loans.

Imagine you have a tree and each branch helps us make a choice, like whether to give a loan or not. Then there's Logistic Regression, which uses a special math trick to tell us the probability of something happening, like if someone will pay back the loan. And then, there's the Random Forest, which is like a team of trees working together to make really smart choices.

All of this mathematical work with numbers showed us how important someone's banking history data is when deciding about loans. It's like looking at a person's money story to see if they can be trusted with a loan. Our work helps make money decisions smarter and fairer for everyone, using facts and patterns to guide the way.

Reference:

- Kaggle (Use to download the dataset for our project)
- Stakflow (We to use this website to correct our code and study the reason behind the code)
- GeeksforGeeks (For concept knowledge)
- Wikipidia to know the means of feature (For data dictionary)

Acknowledgments:

We extend our heartfelt gratitude to **Mrs.Vidhya Kannaiah** & **Great Learning** for their unwavering support, guidance, and invaluable contributions throughout this capstone endeavor. Their mentorship has been instrumental in shaping the course of this project.