

**Guidelines for PGPDSE FT Capstone Project – Interim Report**

**Project Group Info:**

|  |  |
| --- | --- |
| BATCH DETAILS | PGPDSE-FT Online Sep21 |
| TEAM MEMBERS | * RAVITEJA CHODAVARAPU * LEON MENDES * SUPRATICK MOHORAR * PRADEEPTA SAHOO |
| DOMAIN OF PROJECT | Finance & Risk Analytics |
| PROJECT TITLE | Loan Default Prediction using Machine Learning |
| GROUP NUMBER | 8 |
| TEAM LEADER | RAVITEJA CHODAVARAPU |
| MENTOR NAME | MR. JAYVEER NANDA |

Date: 12/06/2022

JAYVEER NANDA RAVITEJA CHODAVARAPU

Signature of the Mentor Signature of the Team Leader

**Abstract:**

One of the primary sources of income of a financial institution is through interests of loaned money. It’s quite tricky for a financial institution to sanction loans because they need to ensure that they get back the loaned amount along with the interests. The banks need a method where they can predict and ensure whether one can pay that amount back. To tackle such a reallife problem, we are building a machine learning model that can predict this with ease.

**Objectives:**

The objective of our project is to predict whether a loan will default or not based on objective financial data only and whether investors should lend to a customer or not. Data from 2007-2011 will be used because most of the loans from that period have already been repaid or defaulted on. With the above objective our aim is to develop a prediction model which assists the Bank to Understand the optimal credit limit and reduce NPA by classifying customer who are most likely to default and who will pay back the loan amount on time. The main objective is to build a machine learning model: o The prediction whether banks or investors should lend to a customer or not o The prediction of future traits o To Predict Optimal Credit Limit for customer’s o To reduce NPA (Non- performing Assets)

**Industry Review:**

Loans always been a crucial part of the banking industry. In the earlier days it was the social reputation and image which were determining factors of one being eligible for a loan. But as the industry grew and times move forward this technique of ‘trust’ soon turned into something we know as guarantee of collaterals. Needless to say with times, the eligibility criteria’s started getting more and more complex.

One thing that didn’t change in the industry from the very beginning is the practice of ensuring that the person/company pays back the loaned amount, whether it was the reputation, assets or cibil score.

Nowadays, organizations heavily rely on data and IT services for efficient internal and external operations making the promptness an important factor. The quicker a bank predicts one’s eligibility the quicker they can sanction the loans and quicker they can reel in another customer.

**Current practices in Industry:**

With changing times, the complexity of analysing what a person is worth or their assets are worth are getting quite tricky. People are slowly away from conventional investments and moving towards digitized one. These portfolios are getting tricky and difficult to analyse for the banks.

To build a machine learning model now for this problem we have to ensure that all possible criteria’s (that are available in the market right now) are included and also design the model in such a way that when new criteria’s come into play the model can be easily tweaked in the future.

**Problem Statement:**

If a model can identify credit-worthy customers that were not recognized by traditional credit scores, while minimizing their risk of default on the loans, this can be a lucrative niche market or micro-market, pushing higher the profit margin of the financial institution or investor. Although the prospect of more customers seems positive, it is important to be careful as to not lend to people that will default on the loan. Thus, a conservative approach and strict evaluation metrics were kept in mind throughout the project. The loan default prediction is a problem of binary classification (should the investor lend or not).

**Project Outcome:**

* By implementing the resultant models in the working dataset, we can get the rejection/approval list of all everyone who applied for loans in an accurate and efficient way.

**Dataset and Domain:**

**Data Dictionary:**

Real-time Loan dataset obtained from an organization.

The Data\_Dictionary.xlsx consists of all the loans i.e., both approved and rejected ones.

The dataset has 39,717 records and 111 attributes

# Column Dtype

--- ------ -----

0 id int64

1 member\_id int64

2 loan\_amnt int64

3 funded\_amnt int64

4 funded\_amnt\_inv float64

5 term object

6 int\_rate object

7 installment float64

8 grade object

9 sub\_grade object

10 emp\_title object

11 emp\_length object

12 home\_ownership object

13 annual\_inc float64

14 verification\_status object

15 issue\_d object

16 loan\_status object

17 pymnt\_plan object

18 url object

19 desc object

20 purpose object

21 title object

22 zip\_code object

23 addr\_state object

24 dti float64

25 delinq\_2yrs int64

26 earliest\_cr\_line object

27 inq\_last\_6mths int64

28 mths\_since\_last\_delinq float64

29 mths\_since\_last\_record float64

30 open\_acc int64

31 pub\_rec int64

32 revol\_bal int64

33 revol\_util object

34 total\_acc int64

35 initial\_list\_status object

36 out\_prncp float64

37 out\_prncp\_inv float64

38 total\_pymnt float64

39 total\_pymnt\_inv float64

40 total\_rec\_prncp float64

41 total\_rec\_int float64

42 total\_rec\_late\_fee float64

43 recoveries float64

44 collection\_recovery\_fee float64

45 last\_pymnt\_d object

46 last\_pymnt\_amnt float64

47 next\_pymnt\_d object

48 last\_credit\_pull\_d object

49 collections\_12\_mths\_ex\_med float64

50 mths\_since\_last\_major\_derog float64

51 policy\_code int64

52 application\_type object

53 annual\_inc\_joint float64

54 dti\_joint float64

55 verification\_status\_joint float64

56 acc\_now\_delinq int64

57 tot\_coll\_amt float64

58 tot\_cur\_bal float64

59 open\_acc\_6m float64

60 open\_il\_6m float64

61 open\_il\_12m float64

62 open\_il\_24m float64

63 mths\_since\_rcnt\_il float64

64 total\_bal\_il float64

65 il\_util float64

66 open\_rv\_12m float64

67 open\_rv\_24m float64

68 max\_bal\_bc float64

69 all\_util float64

70 total\_rev\_hi\_lim float64

71 inq\_fi float64

72 total\_cu\_tl float64

73 inq\_last\_12m float64

74 acc\_open\_past\_24mths float64

75 avg\_cur\_bal float64

76 bc\_open\_to\_buy float64

77 bc\_util float64

78 chargeoff\_within\_12\_mths float64

79 delinq\_amnt int64

80 mo\_sin\_old\_il\_acct float64

81 mo\_sin\_old\_rev\_tl\_op float64

82 mo\_sin\_rcnt\_rev\_tl\_op float64

83 mo\_sin\_rcnt\_tl float64

84 mort\_acc float64

85 mths\_since\_recent\_bc float64

86 mths\_since\_recent\_bc\_dlq float64

87 mths\_since\_recent\_inq float64

88 mths\_since\_recent\_revol\_delinq float64

89 num\_accts\_ever\_120\_pd float64

90 num\_actv\_bc\_tl float64

91 num\_actv\_rev\_tl float64

92 num\_bc\_sats float64

93 num\_bc\_tl float64

94 num\_il\_tl float64

95 num\_op\_rev\_tl float64

96 num\_rev\_accts float64

97 num\_rev\_tl\_bal\_gt\_0 float64

98 num\_sats float64

99 num\_tl\_120dpd\_2m float64

100 num\_tl\_30dpd float64

101 num\_tl\_90g\_dpd\_24m float64

102 num\_tl\_op\_past\_12m float64

103 pct\_tl\_nvr\_dlq float64

104 percent\_bc\_gt\_75 float64

105 pub\_rec\_bankruptcies float64

106 tax\_liens float64

107 tot\_hi\_cred\_lim float64

108 total\_bal\_ex\_mort float64

109 total\_bc\_limit float64

110 total\_il\_high\_credit\_limit float64

**Data Description:**

|  |  |
| --- | --- |
| **LoanStatNew** | **Description** |
| acc\_now\_delinq | The number of accounts on which the borrower is now delinquent. |
| acc\_open\_past\_24mths | Number of trades opened in past 24 months. |
| addr\_state | The state provided by the borrower in the loan application |
| all\_util | Balance to credit limit on all trades |
| annual\_inc | The self-reported annual income provided by the borrower during registration. |
| annual\_inc\_joint | The combined self-reported annual income provided by the co-borrowers during registration |
| application\_type | Indicates whether the loan is an individual application or a joint application with two co-borrowers |
| avg\_cur\_bal | Average current balance of all accounts |
| bc\_open\_to\_buy | Total open to buy on revolving bankcards. |
| bc\_util | Ratio of total current balance to high credit/credit limit for all bankcard accounts. |
| chargeoff\_within\_12\_mths | Number of charge-offs within 12 months |
| collection\_recovery\_fee | post charge off collection fee |
| collections\_12\_mths\_ex\_med | Number of collections in 12 months excluding medical collections |
| delinq\_2yrs | The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years |
| delinq\_amnt | The past-due amount owed for the accounts on which the borrower is now delinquent. |
| desc | Loan description provided by the borrower |
| dti | A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower’s self-reported monthly income. |
| dti\_joint | A ratio calculated using the co-borrowers' total monthly payments on the total debt obligations, excluding mortgages and the requested LC loan, divided by the co-borrowers' combined self-reported monthly income |
| earliest\_cr\_line | The month the borrower's earliest reported credit line was opened |
| emp\_length | Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years. |
| emp\_title | The job title supplied by the Borrower when applying for the loan.\* |
| fico\_range\_high | The upper boundary range the borrower’s FICO at loan origination belongs to. |
| fico\_range\_low | The lower boundary range the borrower’s FICO at loan origination belongs to. |
| funded\_amnt | The total amount committed to that loan at that point in time. |
| funded\_amnt\_inv | The total amount committed by investors for that loan at that point in time. |
| grade | LC assigned loan grade |
| home\_ownership | The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER. |
| id | A unique LC assigned ID for the loan listing. |
| il\_util | Ratio of total current balance to high credit/credit limit on all install acct |
| initial\_list\_status | The initial listing status of the loan. Possible values are – W, F |
| inq\_fi | Number of personal finance inquiries |
| inq\_last\_12m | Number of credit inquiries in past 12 months |
| inq\_last\_6mths | The number of inquiries in past 6 months (excluding auto and mortgage inquiries) |
| installment | The monthly payment owed by the borrower if the loan originates. |
| int\_rate | Interest Rate on the loan |
| issue\_d | The month which the loan was funded |
| last\_credit\_pull\_d | The most recent month LC pulled credit for this loan |
| last\_fico\_range\_high | The upper boundary range the borrower’s last FICO pulled belongs to. |
| last\_fico\_range\_low | The lower boundary range the borrower’s last FICO pulled belongs to. |
| last\_pymnt\_amnt | Last total payment amount received |
| last\_pymnt\_d | Last month payment was received |
| 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. |
| loan\_status | Current status of the loan |
| max\_bal\_bc | Maximum current balance owed on all revolving accounts |
| member\_id | A unique LC assigned Id for the borrower member. |
| mo\_sin\_old\_il\_acct | Months since oldest bank installment account opened |
| mo\_sin\_old\_rev\_tl\_op | Months since oldest revolving account opened |
| mo\_sin\_rcnt\_rev\_tl\_op | Months since most recent revolving account opened |
| mo\_sin\_rcnt\_tl | Months since most recent account opened |
| mort\_acc | Number of mortgage accounts. |
| mths\_since\_last\_delinq | The number of months since the borrower's last delinquency. |
| mths\_since\_last\_major\_derog | Months since most recent 90-day or worse rating |
| mths\_since\_last\_record | The number of months since the last public record. |
| mths\_since\_rcnt\_il | Months since most recent installment accounts opened |
| mths\_since\_recent\_bc | Months since most recent bankcard account opened. |
| mths\_since\_recent\_bc\_dlq | Months since most recent bankcard delinquency |
| mths\_since\_recent\_inq | Months since most recent inquiry. |
| mths\_since\_recent\_revol\_delinq | Months since most recent revolving delinquency. |
| next\_pymnt\_d | Next scheduled payment date |
| num\_accts\_ever\_120\_pd | Number of accounts ever 120 or more days past due |
| num\_actv\_bc\_tl | Number of currently active bankcard accounts |
| num\_actv\_rev\_tl | Number of currently active revolving trades |
| num\_bc\_sats | Number of satisfactory bankcard accounts |
| num\_bc\_tl | Number of bankcard accounts |
| num\_il\_tl | Number of installment accounts |
| num\_op\_rev\_tl | Number of open revolving accounts |
| num\_rev\_accts | Number of revolving accounts |
| num\_rev\_tl\_bal\_gt\_0 | Number of revolving trades with balance >0 |
| num\_sats | Number of satisfactory accounts |
| num\_tl\_120dpd\_2m | Number of accounts currently 120 days past due (updated in past 2 months) |
| num\_tl\_30dpd | Number of accounts currently 30 days past due (updated in past 2 months) |
| num\_tl\_90g\_dpd\_24m | Number of accounts 90 or more days past due in last 24 months |
| num\_tl\_op\_past\_12m | Number of accounts opened in past 12 months |
| open\_acc | The number of open credit lines in the borrower's credit file. |
| open\_acc\_6m | Number of open trades in last 6 months |
| open\_il\_12m | Number of installment accounts opened in past 12 months |
| open\_il\_24m | Number of installment accounts opened in past 24 months |
| open\_il\_6m | Number of currently active installment trades |
| open\_rv\_12m | Number of revolving trades opened in past 12 months |
| open\_rv\_24m | Number of revolving trades opened in past 24 months |
| out\_prncp | Remaining outstanding principal for total amount funded |
| out\_prncp\_inv | Remaining outstanding principal for portion of total amount funded by investors |
| pct\_tl\_nvr\_dlq | Percent of trades never delinquent |
| percent\_bc\_gt\_75 | Percentage of all bankcard accounts > 75% of limit. |
| policy\_code | publicly available policy\_code=1 new products not publicly available policy\_code=2 |
| pub\_rec | Number of derogatory public records |
| pub\_rec\_bankruptcies | Number of public record bankruptcies |
| purpose | A category provided by the borrower for the loan request. |
| pymnt\_plan | Indicates if a payment plan has been put in place for the loan |
| recoveries | post charge off gross recovery |
| revol\_bal | Total credit revolving balance |
| revol\_util | Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit. |
| sub\_grade | LC assigned loan subgrade |
| tax\_liens | Number of tax liens |
| term | The number of payments on the loan. Values are in months and can be either 36 or 60. |
| title | The loan title provided by the borrower |
| tot\_coll\_amt | Total collection amounts ever owed |
| tot\_cur\_bal | Total current balance of all accounts |
| tot\_hi\_cred\_lim | Total high credit/credit limit |
| total\_acc | The total number of credit lines currently in the borrower's credit file |
| total\_bal\_ex\_mort | Total credit balance excluding mortgage |
| total\_bal\_il | Total current balance of all installment accounts |
| total\_bc\_limit | Total bankcard high credit/credit limit |
| total\_cu\_tl | Number of finance trades |
| total\_il\_high\_credit\_limit | Total installment high credit/credit limit |
| total\_pymnt | Payments received to date for total amount funded |
| total\_pymnt\_inv | Payments received to date for portion of total amount funded by investors |
| total\_rec\_int | Interest received to date |
| total\_rec\_late\_fee | Late fees received to date |
| total\_rec\_prncp | Principal received to date |
| total\_rev\_hi\_lim | Total revolving high credit/credit limit |
| url | URL for the LC page with listing data. |
| verification\_status | Indicates if income was verified by LC, not verified, or if the income source was verified |
| verified\_status\_joint | Indicates if the co-borrowers' joint income was verified by LC, not verified, or if the income source was verified |
| zip\_code | The first 3 numbers of the zip code provided by the borrower in the loan application. |

**Variable categorization:**

There are 87 numerical columns and 24 categorical columns

**Numerical Columns:**

[Index ['id',

'member\_id',

'loan\_amnt',

'funded\_amnt',

'funded\_amnt\_inv',

'installment',

'annual\_inc',

'dti',

'delinq\_2yrs',

'inq\_last\_6mths',

'mths\_since\_last\_delinq',

'mths\_since\_last\_record',

'open\_acc',

'pub\_rec',

'revol\_bal',

'total\_acc',

'out\_prncp',

'out\_prncp\_inv',

'total\_pymnt',

'total\_pymnt\_inv',

'total\_rec\_prncp',

'total\_rec\_int',

'total\_rec\_late\_fee',

'recoveries',

'collection\_recovery\_fee',

'last\_pymnt\_amnt',

'collections\_12\_mths\_ex\_med',

'mths\_since\_last\_major\_derog',

'policy\_code',

'annual\_inc\_joint',

'dti\_joint',

'verification\_status\_joint',

'acc\_now\_delinq',

'tot\_coll\_amt',

'tot\_cur\_bal',

'open\_acc\_6m',

'open\_il\_6m',

'open\_il\_12m',

'open\_il\_24m',

'mths\_since\_rcnt\_il',

'total\_bal\_il',

'il\_util',

'open\_rv\_12m',

'open\_rv\_24m',

'max\_bal\_bc',

'all\_util',

'total\_rev\_hi\_lim',

'inq\_fi',

'total\_cu\_tl',

'inq\_last\_12m',

'acc\_open\_past\_24mths',

'avg\_cur\_bal',

'bc\_open\_to\_buy',

'bc\_util',

'chargeoff\_within\_12\_mths',

'delinq\_amnt',

'mo\_sin\_old\_il\_acct',

'mo\_sin\_old\_rev\_tl\_op',

'mo\_sin\_rcnt\_rev\_tl\_op',

'mo\_sin\_rcnt\_tl',

'mort\_acc',

'mths\_since\_recent\_bc',

'mths\_since\_recent\_bc\_dlq',

'mths\_since\_recent\_inq',

'mths\_since\_recent\_revol\_delinq',

'num\_accts\_ever\_120\_pd',

'num\_actv\_bc\_tl',

'num\_actv\_rev\_tl',

'num\_bc\_sats',

'num\_bc\_tl',

'num\_il\_tl',

'num\_op\_rev\_tl',

'num\_rev\_accts',

'num\_rev\_tl\_bal\_gt\_0',

'num\_sats',

'num\_tl\_120dpd\_2m',

'num\_tl\_30dpd',

'num\_tl\_90g\_dpd\_24m',

'num\_tl\_op\_past\_12m',

'pct\_tl\_nvr\_dlq',

'percent\_bc\_gt\_75',

'pub\_rec\_bankruptcies',

'tax\_liens',

'tot\_hi\_cred\_lim',

'total\_bal\_ex\_mort',

'total\_bc\_limit',

'total\_il\_high\_credit\_limit']

**Categorical Columns:**

[Index ['term',

'int\_rate',

'grade',

'sub\_grade',

'emp\_title',

'emp\_length',

'home\_ownership',

'verification\_status',

'issue\_d',

'loan\_status',

'pymnt\_plan',

'url',

'desc',

'purpose',

'title',

'zip\_code',

'addr\_state',

'earliest\_cr\_line',

'revol\_util',

'initial\_list\_status',

'last\_pymnt\_d',

'next\_pymnt\_d',

'last\_credit\_pull\_d',

'application\_type']

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

The dataset has missing values in most of the columns.

**Redundant Columns:**

**Before:**

# Column Dtype

--- ------ -----

0 id int64

1 member\_id int64

2 loan\_amnt int64

3 funded\_amnt int64

4 funded\_amnt\_inv float64

5 term object

6 int\_rate object

7 installment float64

8 grade object

9 sub\_grade object

10 emp\_title object

11 emp\_length object

12 home\_ownership object

13 annual\_inc float64

14 verification\_status object

15 issue\_d object

16 loan\_status object

17 pymnt\_plan object

18 url object

19 desc object

20 purpose object

21 title object

22 zip\_code object

23 addr\_state object

24 dti float64

25 delinq\_2yrs int64

26 earliest\_cr\_line object

27 inq\_last\_6mths int64

28 mths\_since\_last\_delinq float64

29 mths\_since\_last\_record float64

30 open\_acc int64

31 pub\_rec int64

32 revol\_bal int64

33 revol\_util object

34 total\_acc int64

35 initial\_list\_status object

36 out\_prncp float64

37 out\_prncp\_inv float64

38 total\_pymnt float64

39 total\_pymnt\_inv float64

40 total\_rec\_prncp float64

41 total\_rec\_int float64

42 total\_rec\_late\_fee float64

43 recoveries float64

44 collection\_recovery\_fee float64

45 last\_pymnt\_d object

46 last\_pymnt\_amnt float64

47 next\_pymnt\_d object

48 last\_credit\_pull\_d object

49 collections\_12\_mths\_ex\_med float64

50 mths\_since\_last\_major\_derog float64

51 policy\_code int64

52 application\_type object

53 annual\_inc\_joint float64

54 dti\_joint float64

55 verification\_status\_joint float64

56 acc\_now\_delinq int64

57 tot\_coll\_amt float64

58 tot\_cur\_bal float64

59 open\_acc\_6m float64

60 open\_il\_6m float64

61 open\_il\_12m float64

62 open\_il\_24m float64

63 mths\_since\_rcnt\_il float64

64 total\_bal\_il float64

65 il\_util float64

66 open\_rv\_12m float64

67 open\_rv\_24m float64

68 max\_bal\_bc float64

69 all\_util float64

70 total\_rev\_hi\_lim float64

71 inq\_fi float64

72 total\_cu\_tl float64

73 inq\_last\_12m float64

74 acc\_open\_past\_24mths float64

75 avg\_cur\_bal float64

76 bc\_open\_to\_buy float64

77 bc\_util float64

78 chargeoff\_within\_12\_mths float64

79 delinq\_amnt int64

80 mo\_sin\_old\_il\_acct float64

81 mo\_sin\_old\_rev\_tl\_op float64

82 mo\_sin\_rcnt\_rev\_tl\_op float64

83 mo\_sin\_rcnt\_tl float64

84 mort\_acc float64

85 mths\_since\_recent\_bc float64

86 mths\_since\_recent\_bc\_dlq float64

87 mths\_since\_recent\_inq float64

88 mths\_since\_recent\_revol\_delinq float64

89 num\_accts\_ever\_120\_pd float64

90 num\_actv\_bc\_tl float64

91 num\_actv\_rev\_tl float64

92 num\_bc\_sats float64

93 num\_bc\_tl float64

94 num\_il\_tl float64

95 num\_op\_rev\_tl float64

96 num\_rev\_accts float64

97 num\_rev\_tl\_bal\_gt\_0 float64

98 num\_sats float64

99 num\_tl\_120dpd\_2m float64

100 num\_tl\_30dpd float64

101 num\_tl\_90g\_dpd\_24m float64

102 num\_tl\_op\_past\_12m float64

103 pct\_tl\_nvr\_dlq float64

104 percent\_bc\_gt\_75 float64

105 pub\_rec\_bankruptcies float64

106 tax\_liens float64

107 tot\_hi\_cred\_lim float64

108 total\_bal\_ex\_mort float64

109 total\_bc\_limit float64

110 total\_il\_high\_credit\_limit float64

**After:**

loan\_amnt int64

funded\_amnt int64

funded\_amnt\_inv float64

term object

int\_rate object

installment float64

grade object

sub\_grade object

emp\_title object

emp\_length object

home\_ownership object

annual\_inc float64

verification\_status object

issue\_d object

loan\_status object

pymnt\_plan object

purpose object

title object

addr\_state object

dti float64

earliest\_cr\_line object

inq\_last\_6mths int64

mths\_since\_last\_delinq float64

mths\_since\_last\_record float64

open\_acc int64

pub\_rec int64

revol\_util object

total\_acc int64

initial\_list\_status object

out\_prncp\_inv float64

total\_pymnt\_inv float64

last\_credit\_pull\_d object

collections\_12\_mths\_ex\_med float64

mths\_since\_last\_major\_derog float64

policy\_code int64

application\_type object

annual\_inc\_joint float64

dti\_joint float64

verification\_status\_joint float64

acc\_now\_delinq int64

tot\_coll\_amt float64

tot\_cur\_bal float64

open\_acc\_6m float64

open\_il\_6m float64

open\_il\_12m float64

open\_il\_24m float64

mths\_since\_rcnt\_il float64

total\_bal\_il float64

il\_util float64

open\_rv\_12m float64

open\_rv\_24m float64

max\_bal\_bc float64

all\_util float64

total\_rev\_hi\_lim float64

inq\_fi float64

total\_cu\_tl float64

inq\_last\_12m float64

acc\_open\_past\_24mths float64

avg\_cur\_bal float64

bc\_open\_to\_buy float64

bc\_util float64

delinq\_amnt int64

mo\_sin\_old\_il\_acct float64

mo\_sin\_old\_rev\_tl\_op float64

mo\_sin\_rcnt\_rev\_tl\_op float64

mo\_sin\_rcnt\_tl float64

mort\_acc float64

mths\_since\_recent\_bc float64

mths\_since\_recent\_bc\_dlq float64

mths\_since\_recent\_inq float64

mths\_since\_recent\_revol\_delinq float64

num\_accts\_ever\_120\_pd float64

num\_actv\_bc\_tl float64

num\_actv\_rev\_tl float64

num\_bc\_sats float64

num\_bc\_tl float64

num\_il\_tl float64

num\_op\_rev\_tl float64

num\_rev\_accts float64

num\_rev\_tl\_bal\_gt\_0 float64

num\_sats float64

num\_tl\_120dpd\_2m float64

num\_tl\_30dpd float64

num\_tl\_90g\_dpd\_24m float64

num\_tl\_op\_past\_12m float64

pct\_tl\_nvr\_dlq float64

percent\_bc\_gt\_75 float64

pub\_rec\_bankruptcies float64

tax\_liens float64

tot\_hi\_cred\_lim float64

total\_bal\_ex\_mort float64

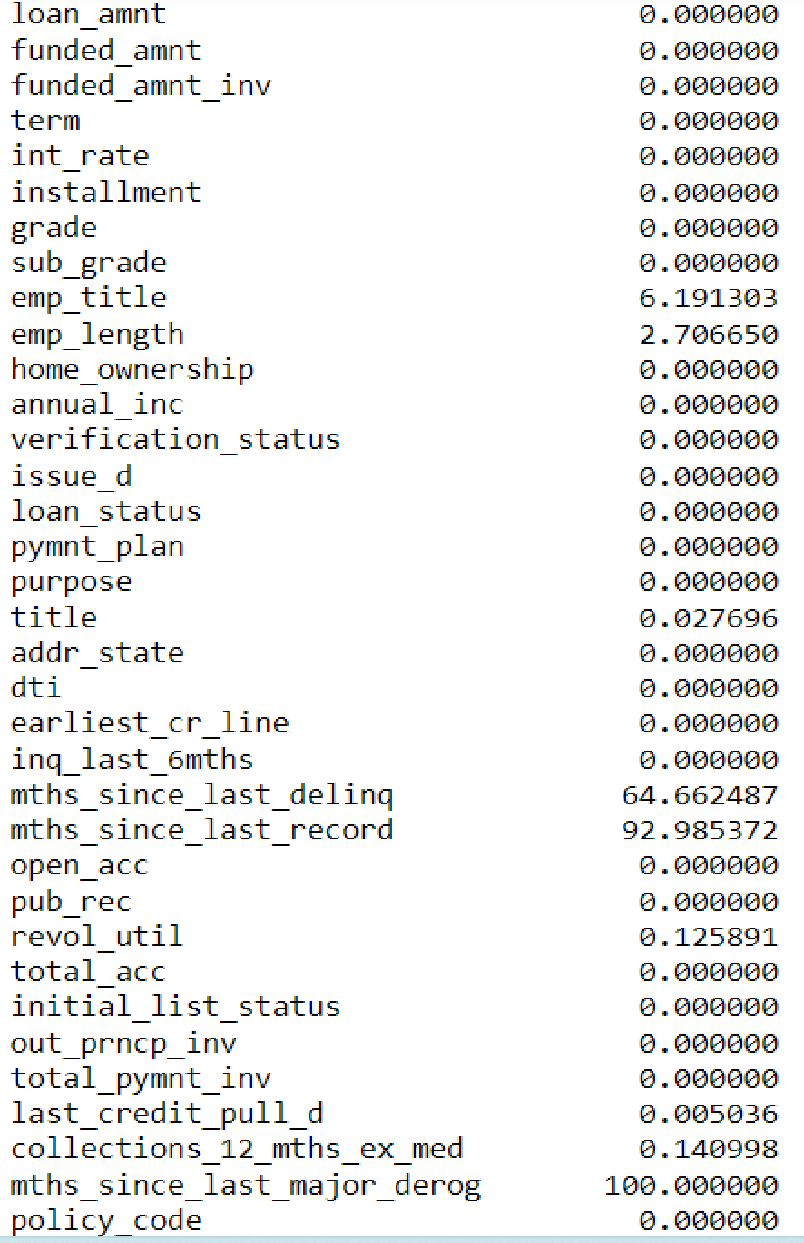
total\_bc\_limit float64

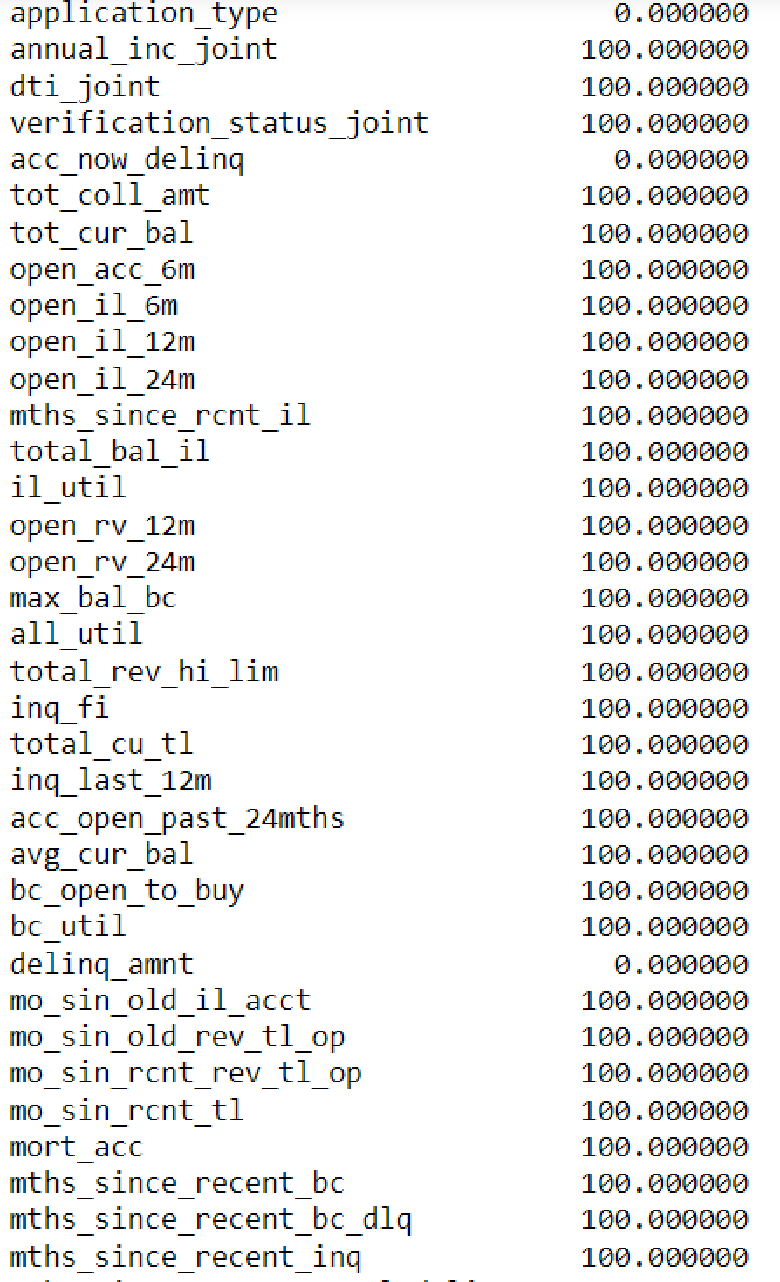
total\_il\_high\_credit\_limit float64

**Dealing with Null values:**

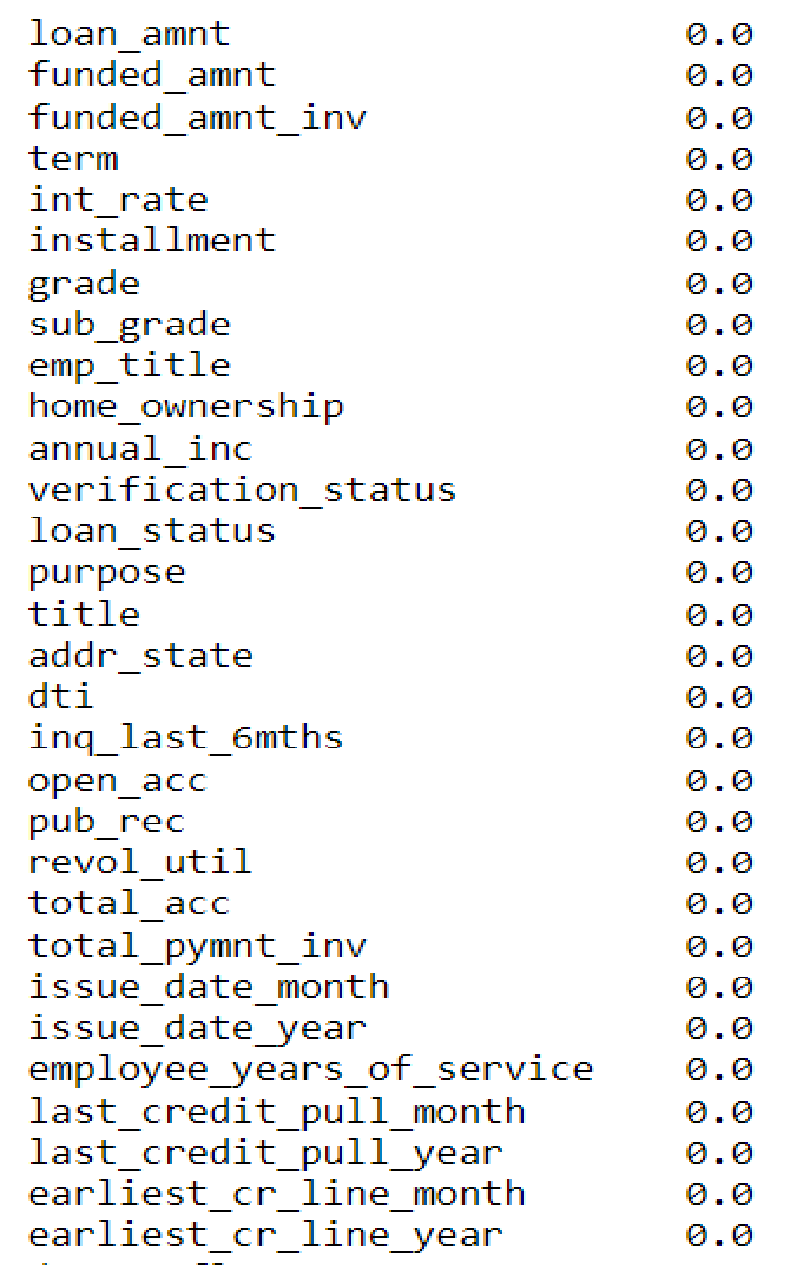
The dataset has multiple columns with 100% and the 2 columns have more than 50% of null values.

**Before**





**After**



**Changing Datatypes:**

Changing the datatypes as required.

**Before:**

loan\_amnt int64

funded\_amnt int64

funded\_amnt\_inv float64

term object

int\_rate object

installment float64

grade object

sub\_grade object

emp\_title object

emp\_length object

home\_ownership object

annual\_inc float64

verification\_status object

issue\_d object

loan\_status object

pymnt\_plan object

purpose object

title object

addr\_state object

dti float64

earliest\_cr\_line object

inq\_last\_6mths int64

open\_acc int64

pub\_rec int64

revol\_util object

total\_acc int64

initial\_list\_status object

out\_prncp\_inv float64

total\_pymnt\_inv float64

last\_credit\_pull\_d object

collections\_12\_mths\_ex\_med float64

policy\_code int64

application\_type object

acc\_now\_delinq int64

delinq\_amnt int64

tax\_liens float64

**After:**

oan\_amnt int64

funded\_amnt int64

funded\_amnt\_inv float64

term int32

int\_rate float64

installment float64

grade object

sub\_grade object

emp\_title object

home\_ownership object

annual\_inc float64

verification\_status object

loan\_status object

purpose object

title object

addr\_state object

dti float64

inq\_last\_6mths int64

open\_acc int64

revol\_util float64

total\_acc int64

total\_pymnt\_inv float64

issue\_date\_month int32

issue\_date\_year int32

employee\_years\_of\_service float64

last\_credit\_pull\_month Int64

last\_credit\_pull\_year Int64

earliest\_cr\_line\_month Int64

earliest\_cr\_line\_year Int64

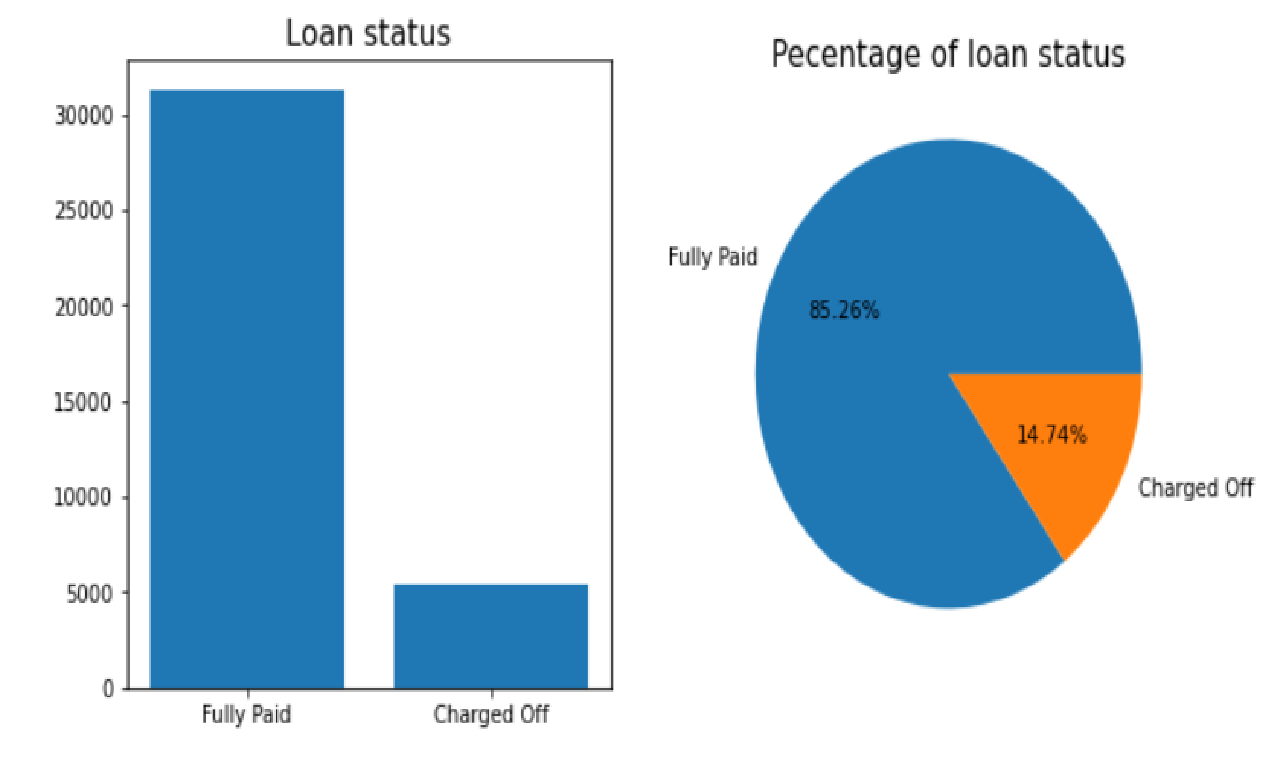
**UNIVARIATE & BIVARIATE ANALYSIS :**

**Analysis of Categorical Variables:**

In the below plots, each of the categorical variables are analysed individually**.**

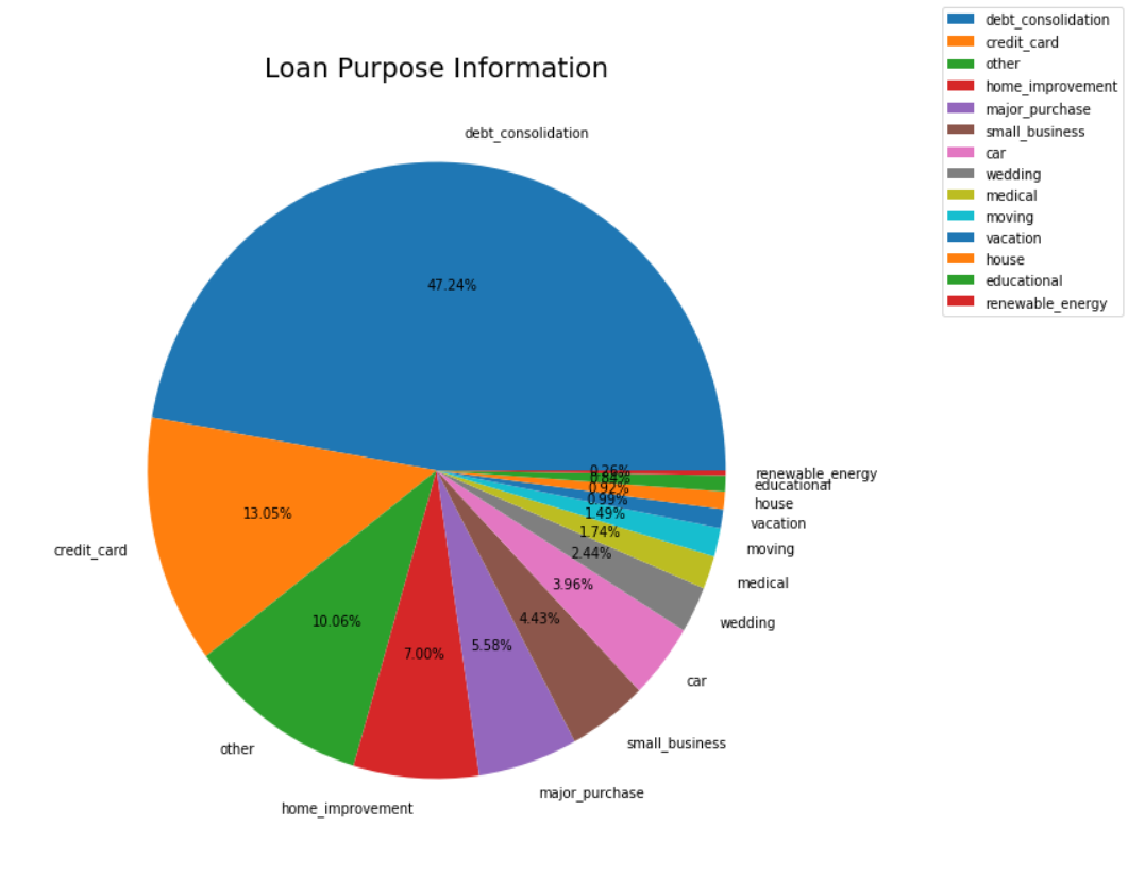
**Loan Status:**

In the below plot, each distribution of the target variable is shown**.**

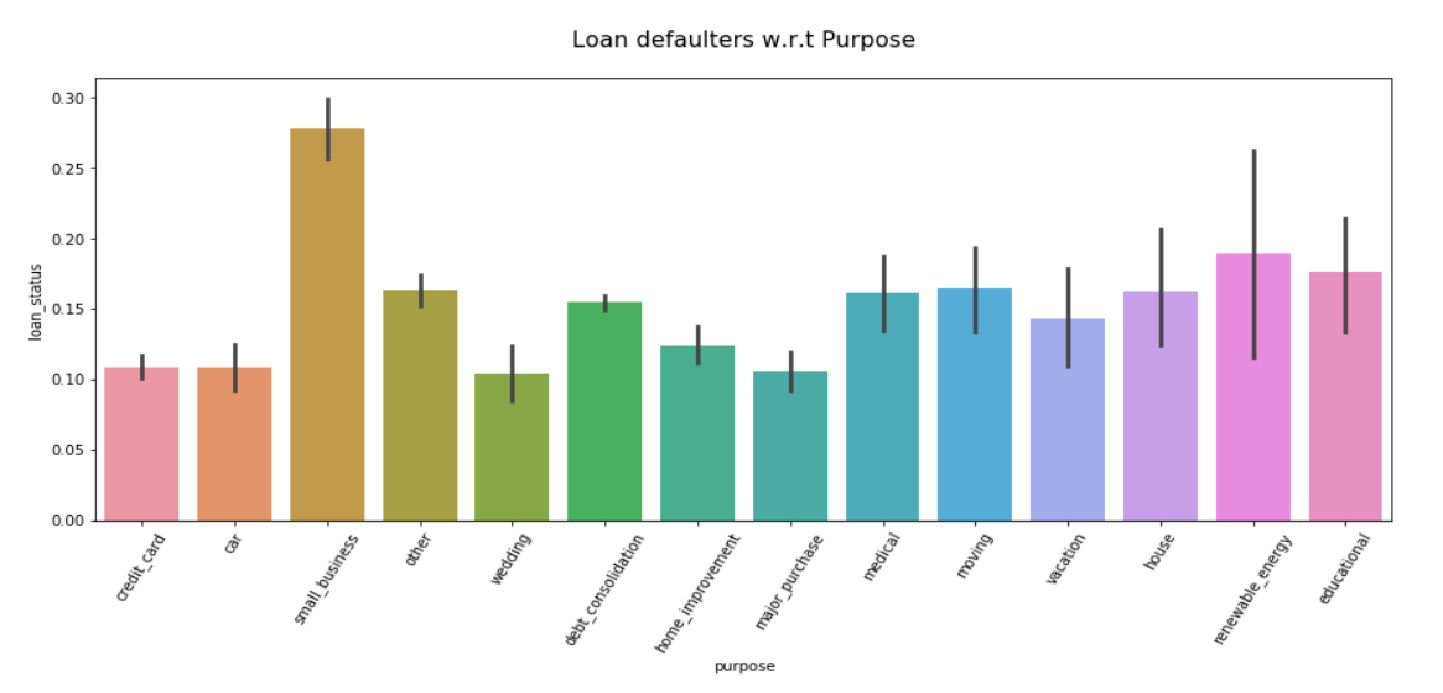


* From the above plot , we ca say that 14.46 % people are the loan defaulters.
* There are around 5232 people of total 36194 people who are loan defaulters.

**Loan Purpose:**

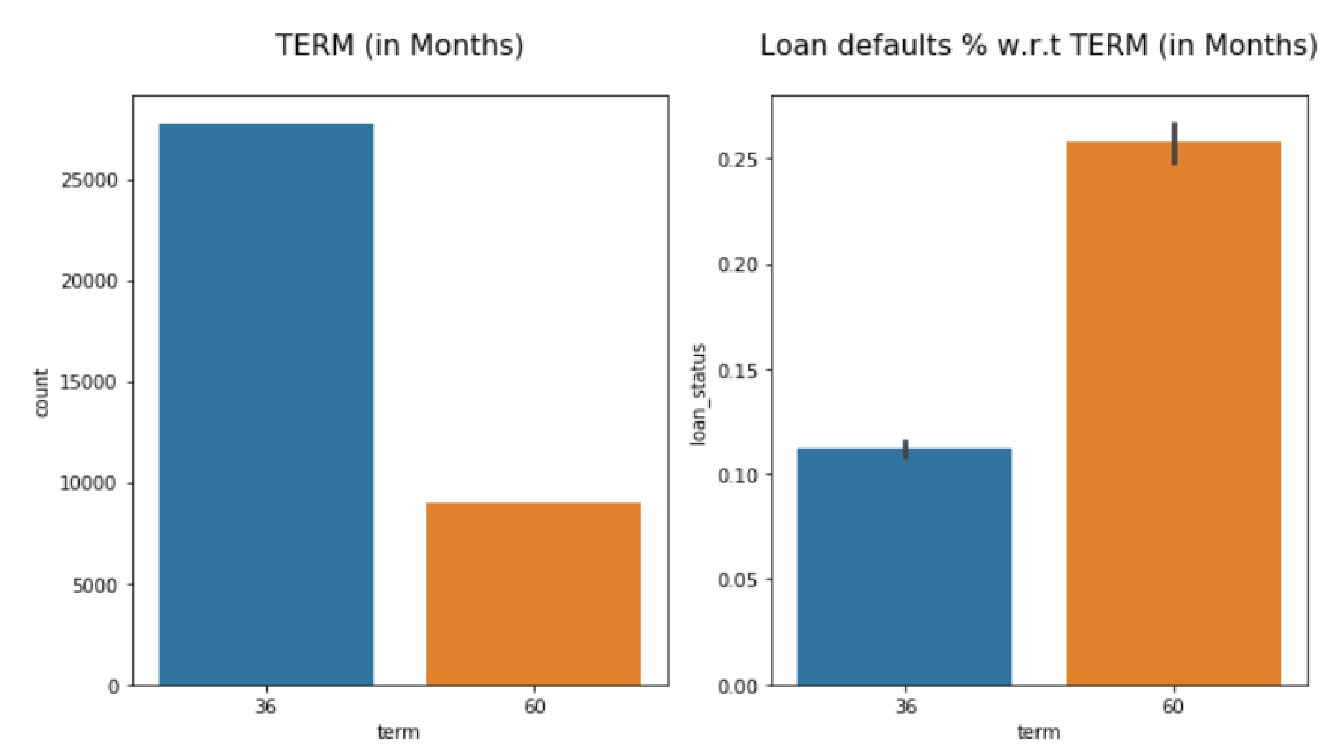


**Loan Defaulters:**



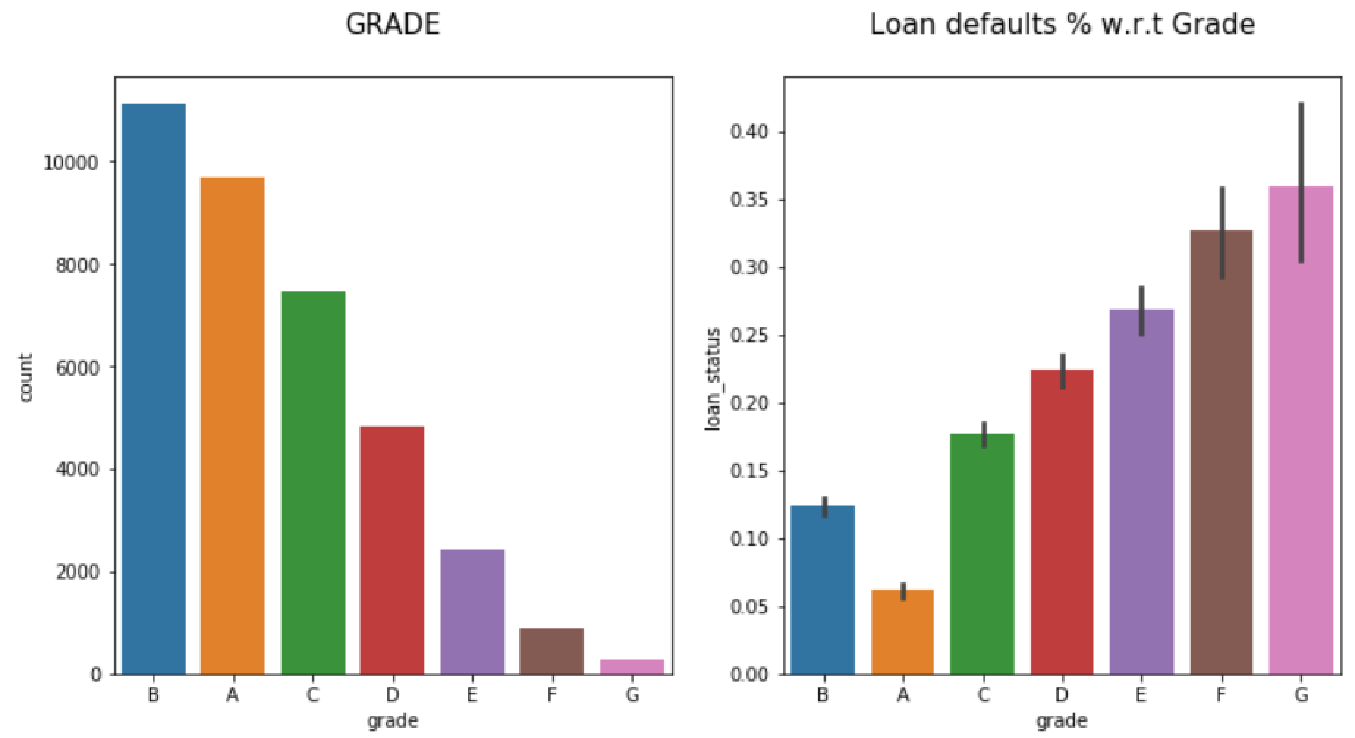
* Loans are taken for majority for debt consolidation, Other,Home improvement , major purchase & small buisiness purpose.
* From the bar plot, we can observe that the probability of persong being loan defaulter is more in small buisiness, Defaulter rate is more when person take loan for small business purpose.
* Defaulter Rate will depend on the Purpose of the loan.

**Term:**



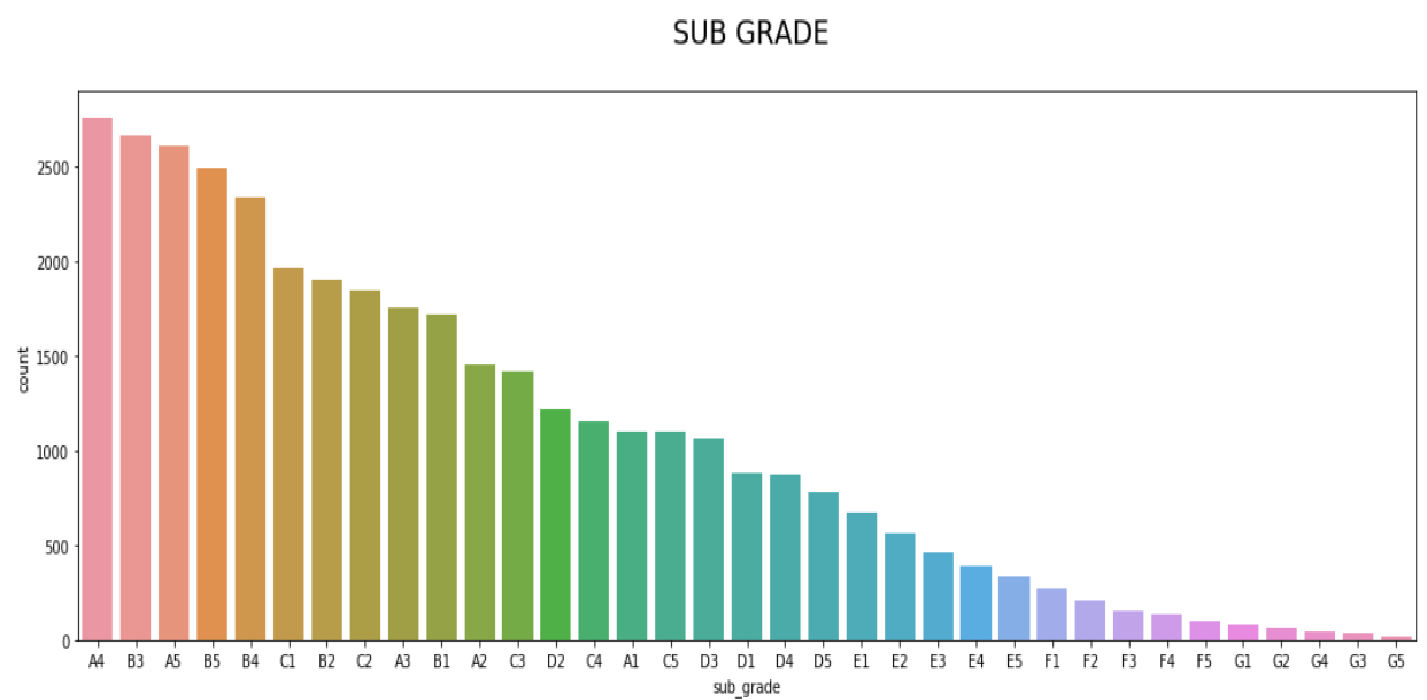
* In the givan dataset, we are having more number of people with term 36 months than 60 months.
* But The probability of loan getting defaulted is more for 60 months than 30 months.

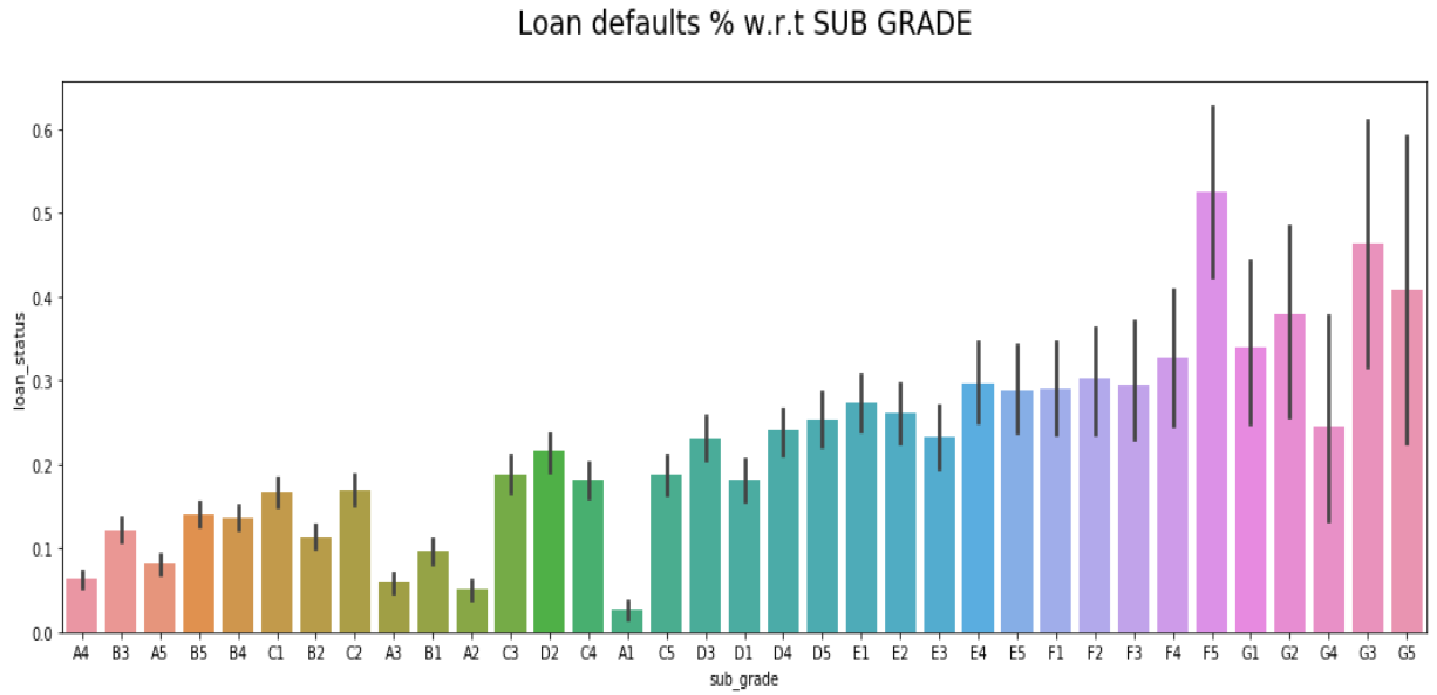
**Grade:**



* There are more number of people with grade B.
* Defaulter rate is high with the grade G and less for grade A.

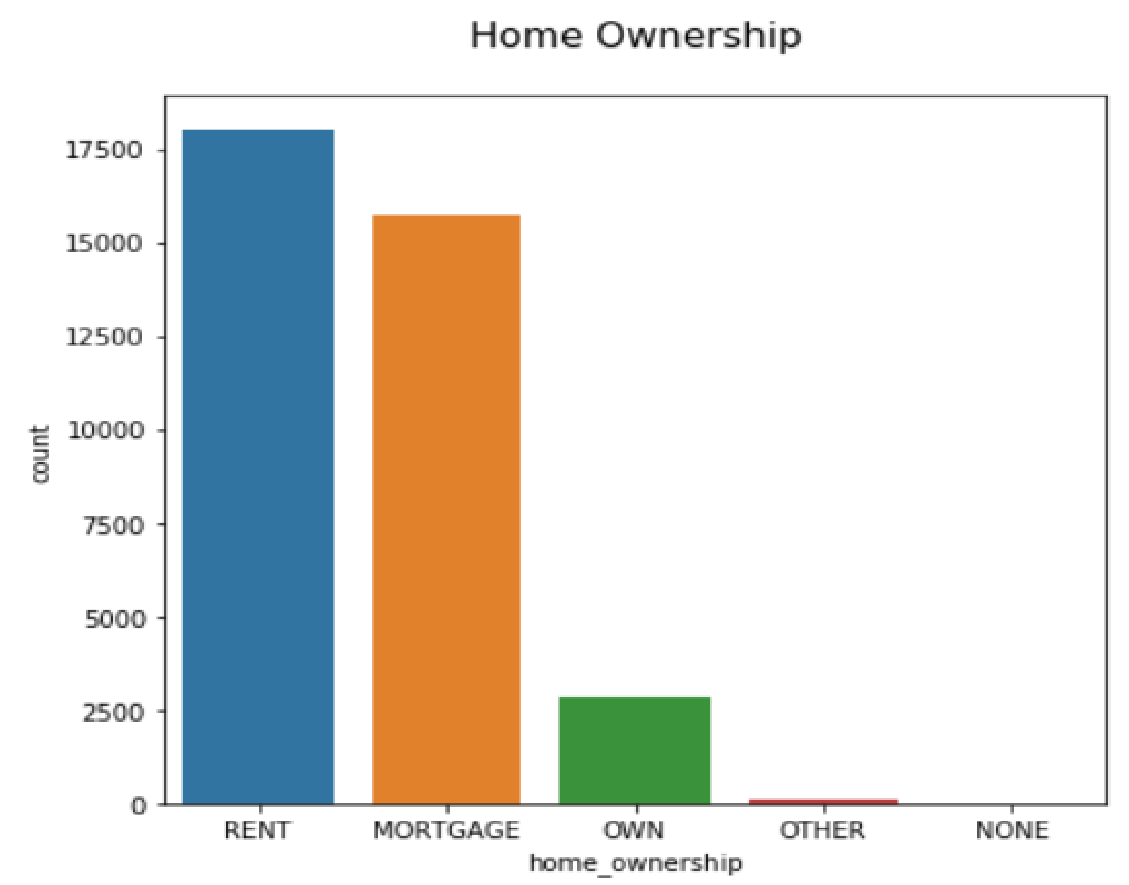
**Sub-Grade:**

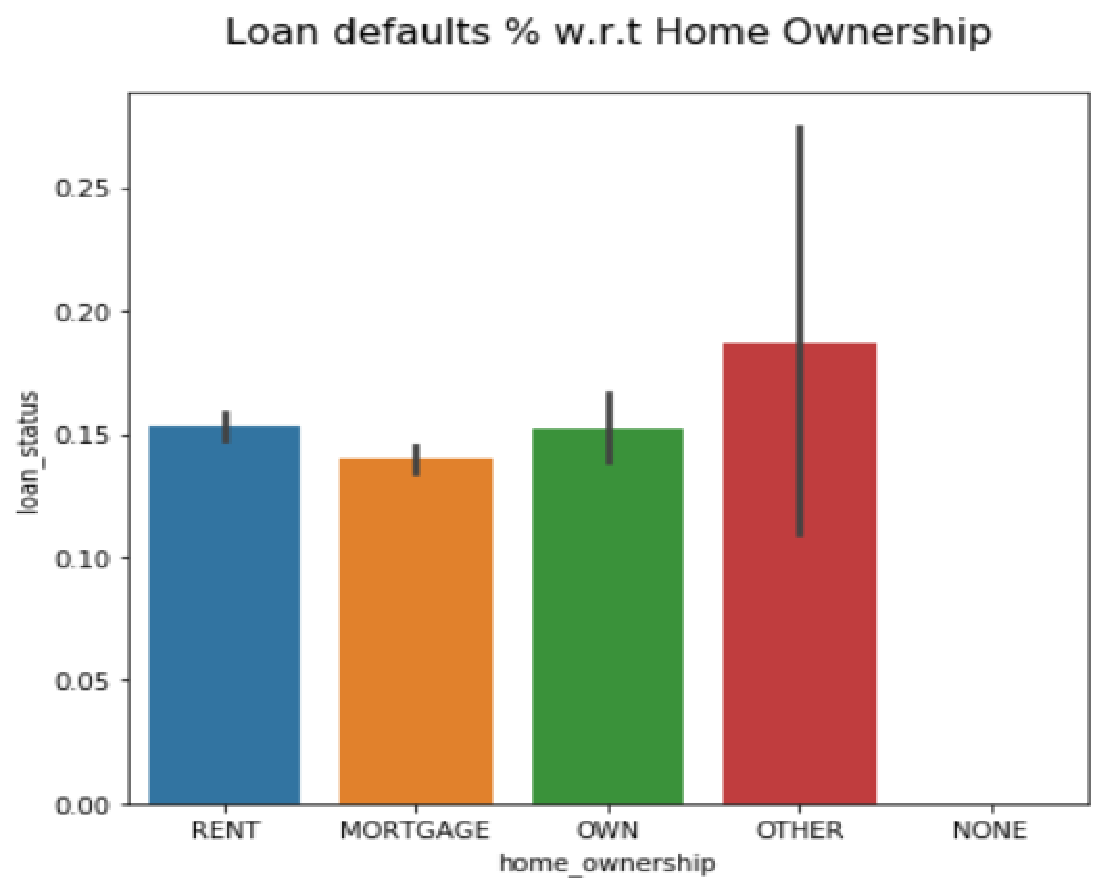




* There are more number of people with sub grade A4.
* Defaulter rate is increasing with the sub grade , loan defaulter rate is more for F5 grade and less for grade A1.
* Sub Grade is useful for further analysis.

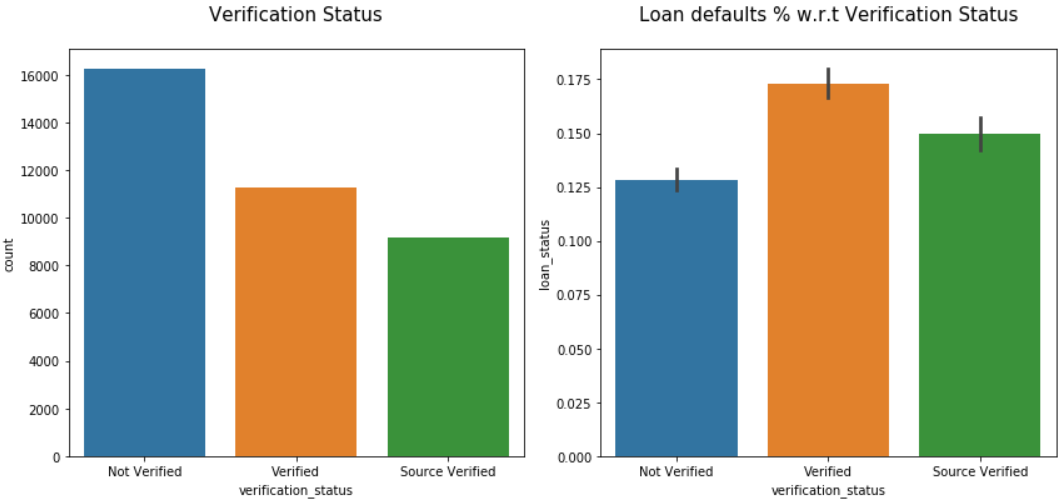
**Home Ownership:**

****



* Loan defaulter rate is almost constant for all the home ownerships, slightly more for OTHER home ownership.
* We can say that loan defaulters does not depends on home ownership.
* Home ownership is not useful for further analysis.

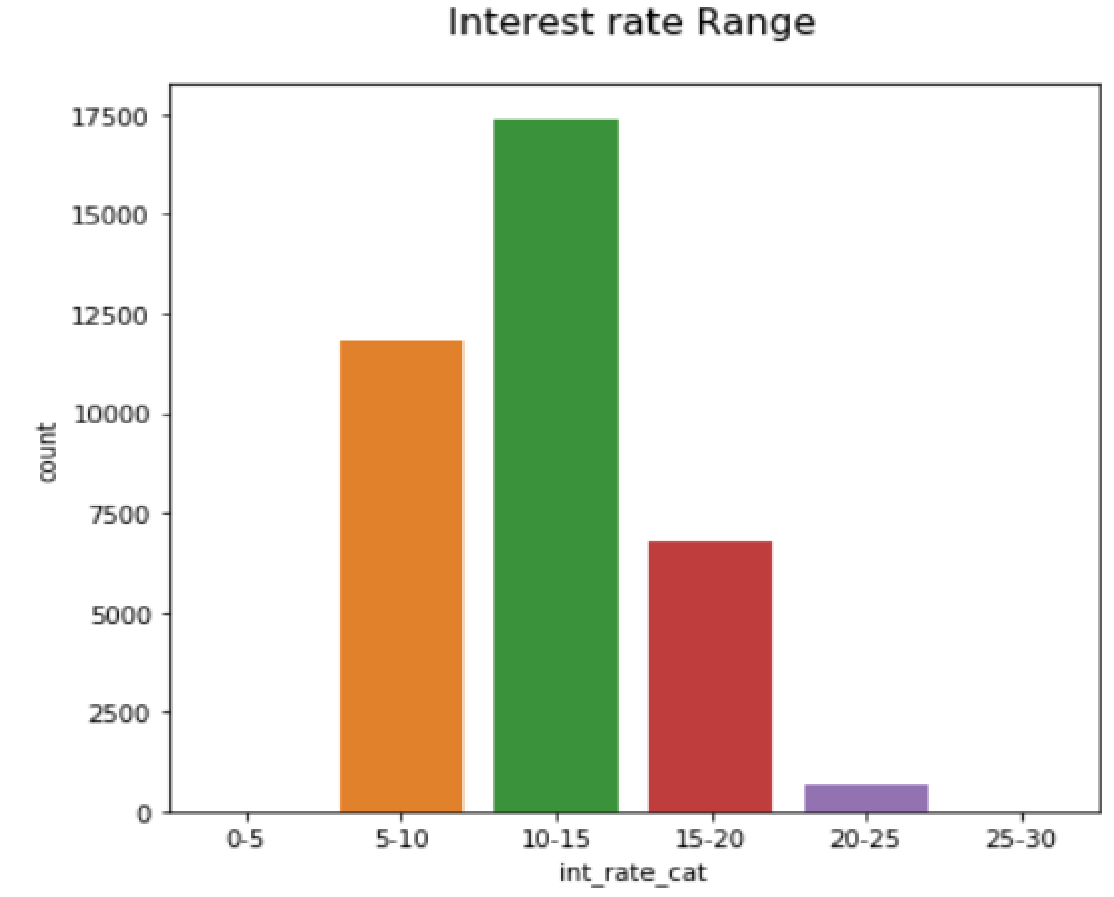
**Verification Status:**

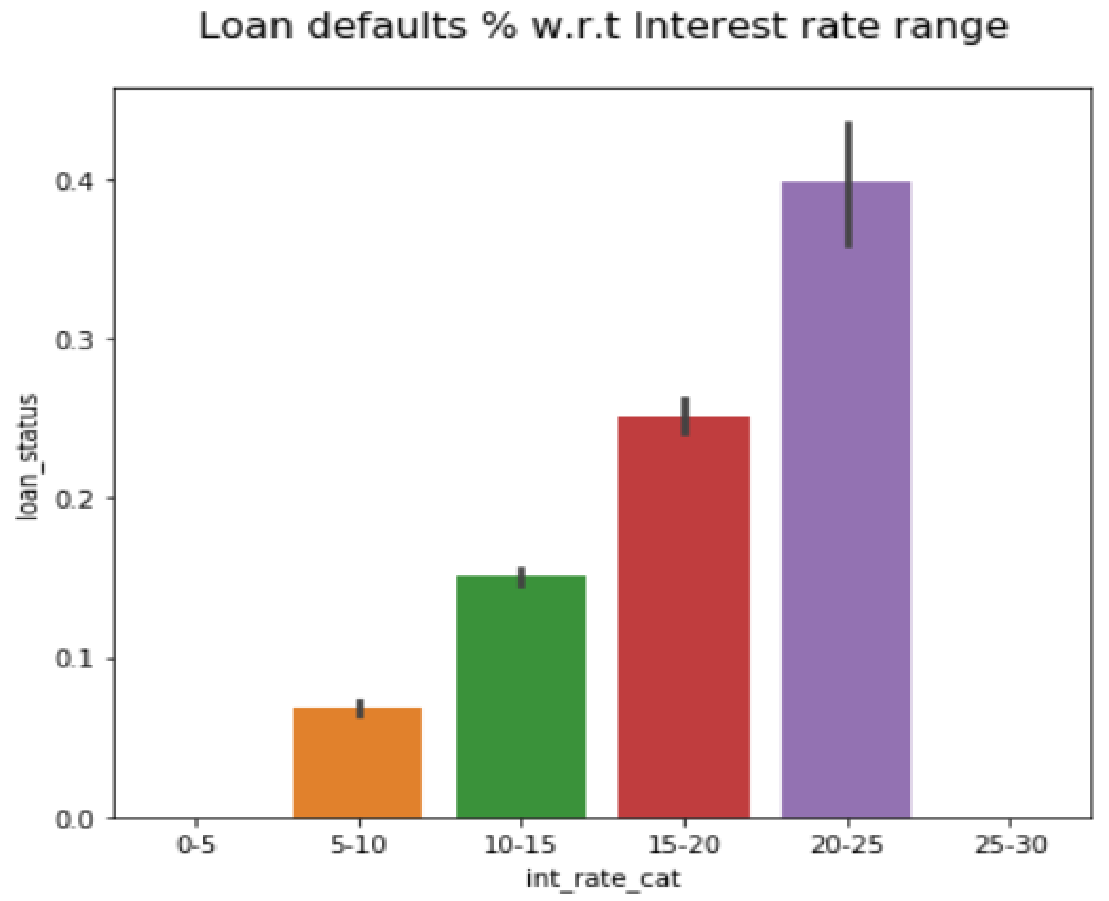


* In the given data ,There are more records for which the ststus is non verified.
* But the defaulter rate is more for verified status.

**Analysis of Numerical Variables:**

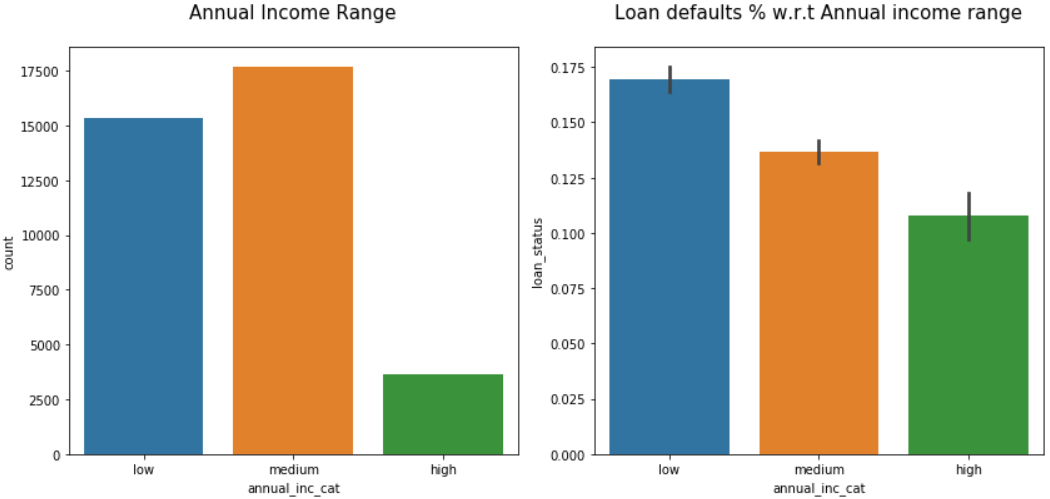
**Interest Rate:**





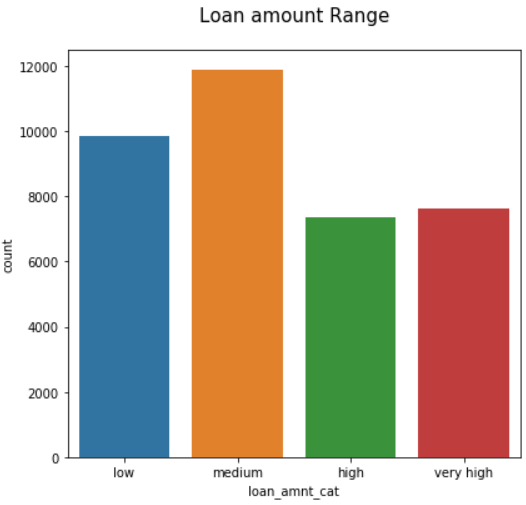
* There are more records with interest rate between 10-15%.
* The rate of loan defaulter is more for highest interest rate & less for lowest interest rate.

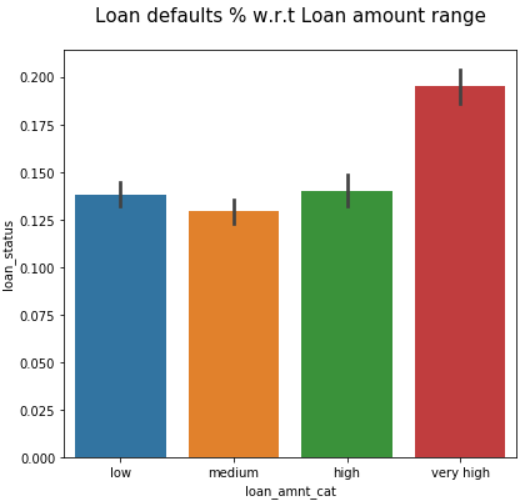
**Annual Income:**

****

* Defaulter rate is increasing with the annual income value, defaulter rate will depend on loan amount.
* The annual income variable is useful for further analysis.

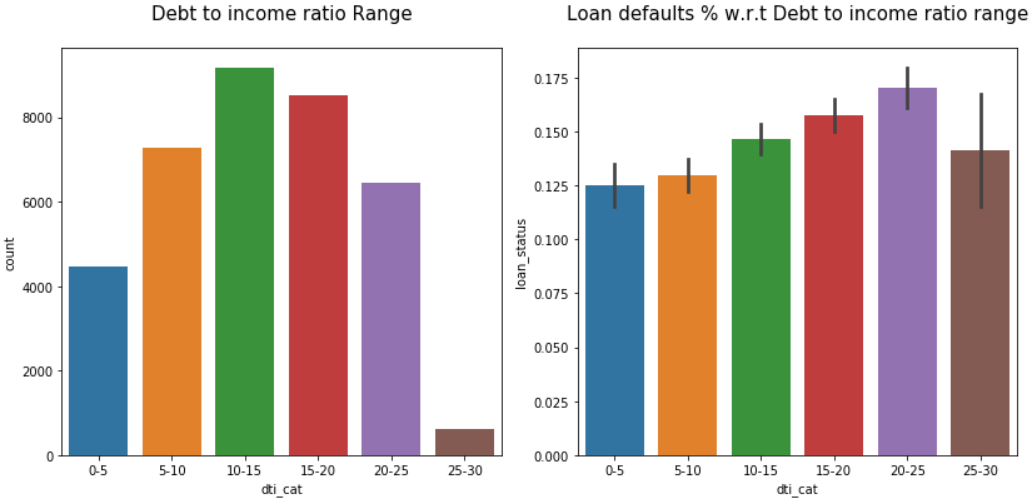
**Loan Amount:**

****

****

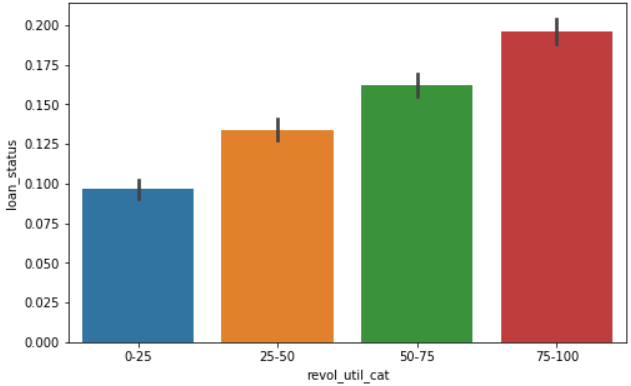
* Loan defaulter rate is increasing with loan amount range.
* This feature is useful for further analysis.

**Debt To Income Ratio:**

****

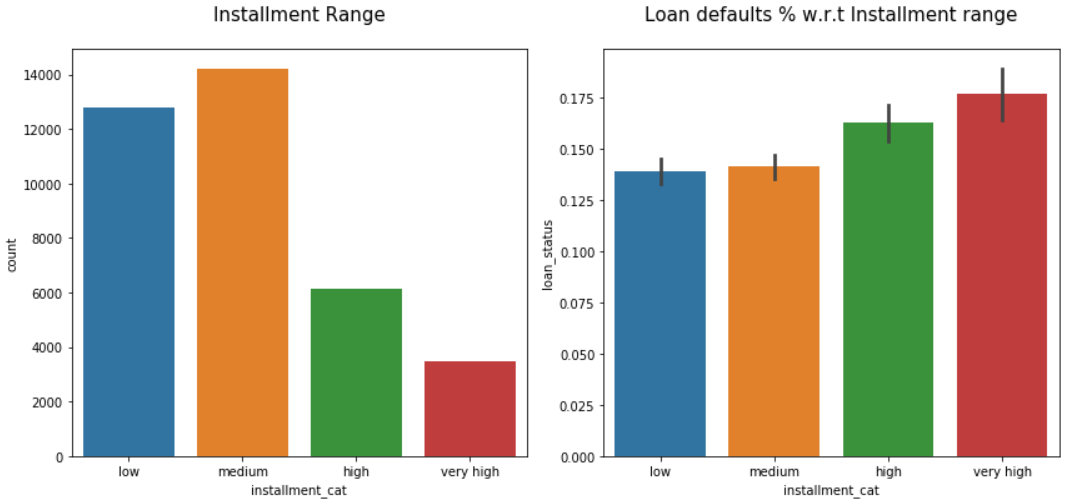
* Loan defaulter rate is increasing with the debt to income ratio.
* Debt to income ratio is useful for finding the loan defaulter.

**Revolving Line Utilisation Rate:**



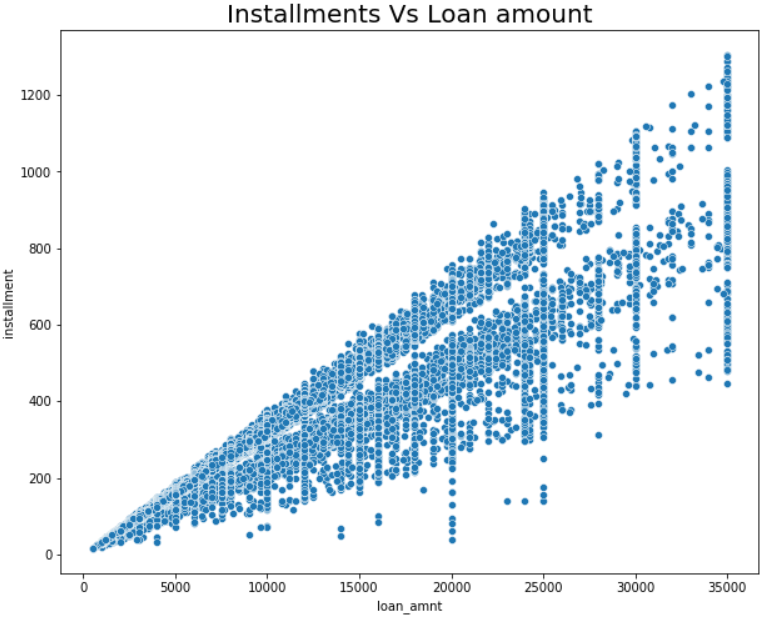
* Loan defaulter rate is increasing with the revolving line utilization rate.
* This is useful for finding the loan defaulter.

**Installments:**

****

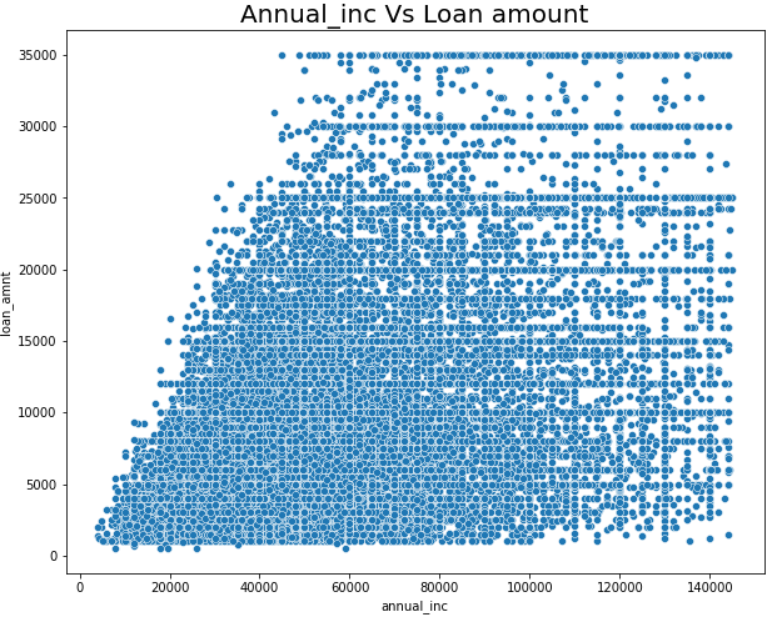
* Loan defaulter rate is increasing with the installment values.
* Dti is useful for finding the loan defaulter.

**Installments VS Loan Amount:**

****

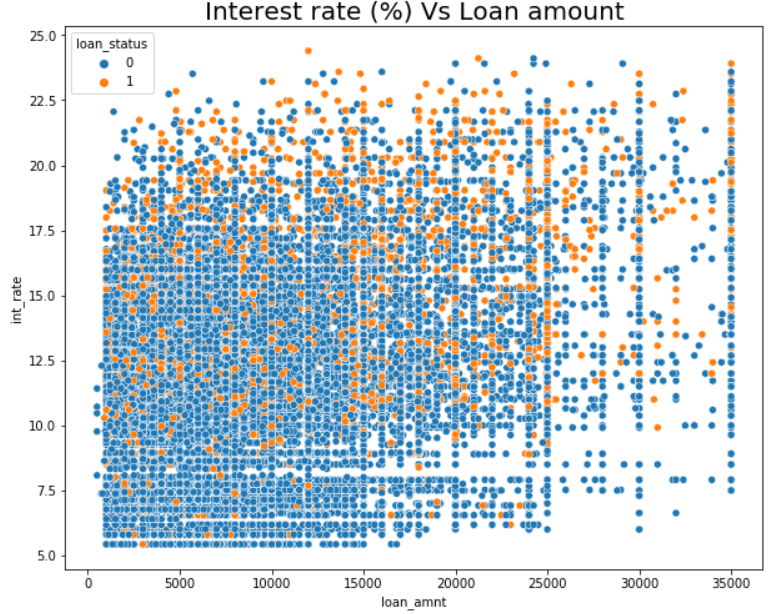
* It is obvious that if the loan amount is high, the installment amount will also be high.

**Annual Income VS Loan Amount:**

****

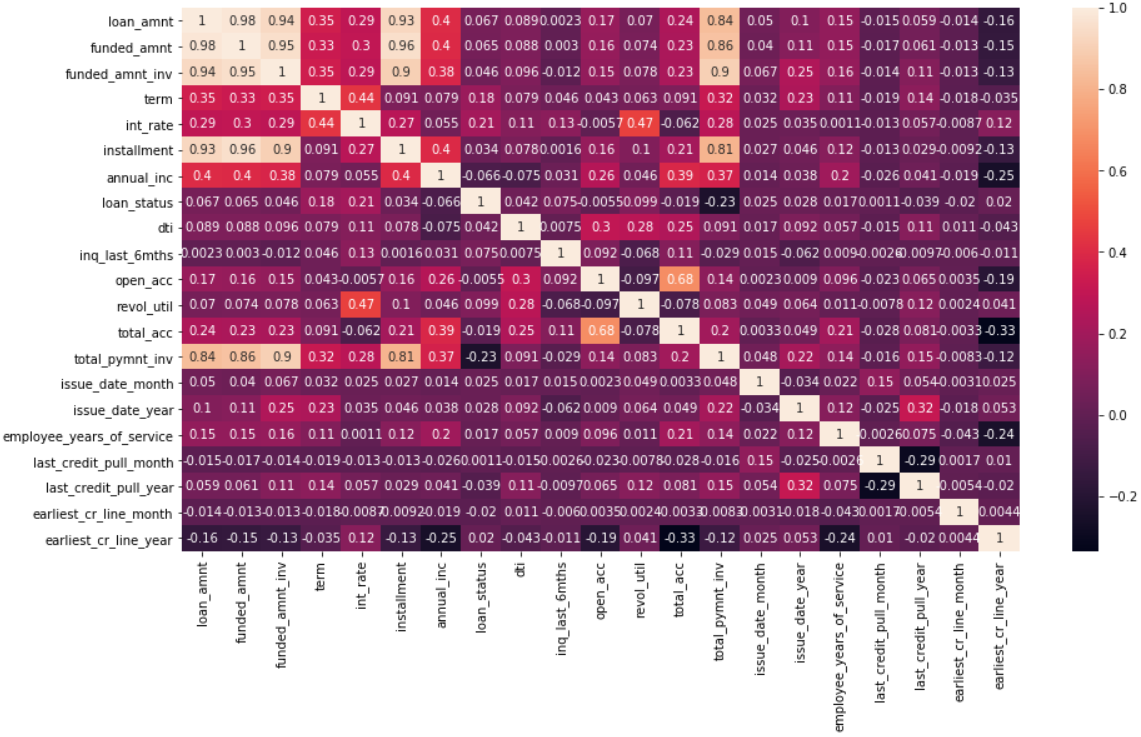
* People with annual income between 0 to 1,00,000 tend to apply more for loan.

**Loan Amount & Interest Rate:**

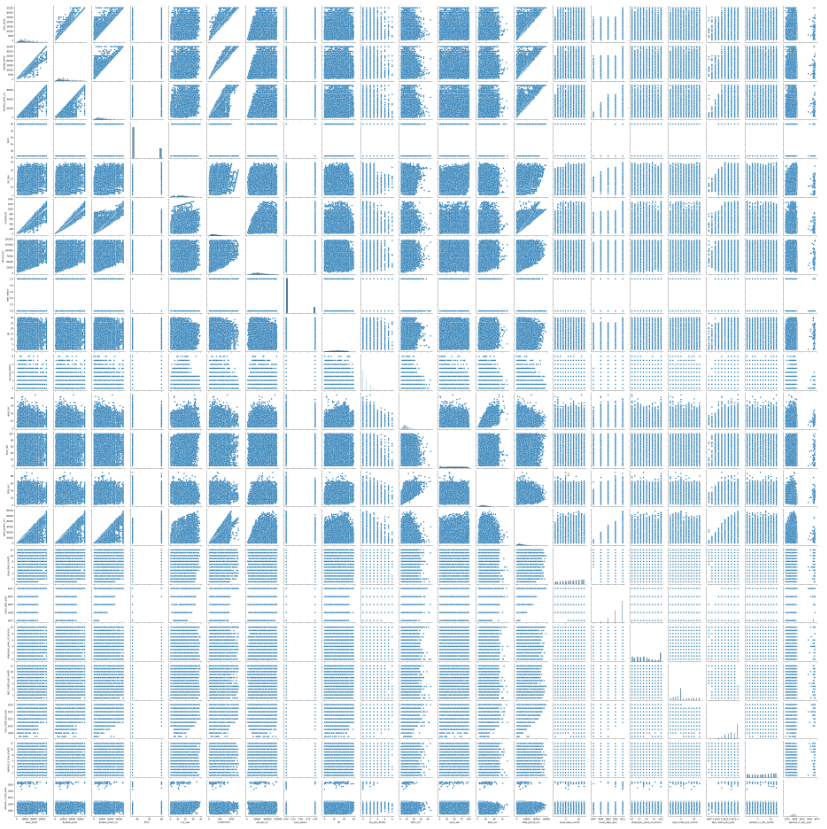


* Defaulters are present in all the amounts.

**MULTIVARIATE ANALYSIS :**

****

1. There is a strong correlation between loan amount and funded amount.
2. There is a strong correlation between loan amount and funded amount invested.
3. There is a strong correlation between loan amount and installment.
4. There is a strong correlation between installment and funded\_amount,funded\_amount\_invested.
5. There is a strong correlation between total\_pymnt\_inv and total\_amount,funded\_amount , funded\_amnt\_inv.

**Pairplot:**

* The above pair plot shows the pairwise relationship between all the numerical variables.

**Model Building:**

**Assumptions Before Linear Regression (OLS):**

* Assumption on Dependent Variable: Target variable is numerical.
* Normality of target variable:
* It is apparent that the p-value is less than 0.05. So we have enough evidence to reject the null hypothesis.
* It can be concluded that the data is not normally distributed. So, we can do log transformation on the target variable.
* But first, the base model is built without log transformation.

**Train and Test data:**

The shape of X\_train is: (133113, 72)

The shape of X\_test is: (44371, 72)

The shape of y\_train is: (133113,)

The shape of y\_test is: (44371,)

**Assumptions for Logistic Regression:**

● The Output is a categorical variable which is ‘0’ and ‘1’.

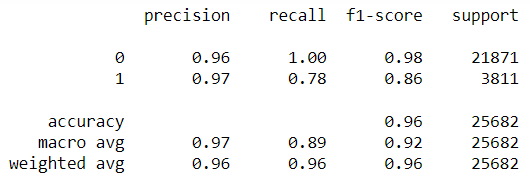
● There is no multicollinearity in the data as we have checked for the correlation in heatmap and removed the columns that have an impact on multicollinearity. (treated accordingly beforehand)

**Logistic Regression:**

● We have used a logistic regression function in python LogisticRegression().

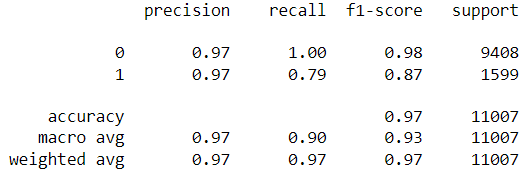
● We have fit and trained the model on the train dataset and checked the performance score of the train data.

**Training Data Performance:**

****

**F1 score : 0.8619**

**Test Data Performance:**

****

● As the next step we have predicted the values for the test dataset and checked the performance metrics to check for overfitting or underfitting.

**Auc: 0.8954163396127681**

**Recall: 0.7948717948717948**

**Precision: 0.970970206264324**

**f1\_score: 0.87414030261348**

**Accuracy: 0.9667484328154811**

● The model is performing well on the train as well as the test data and it is neither overfitting nor underfitting.

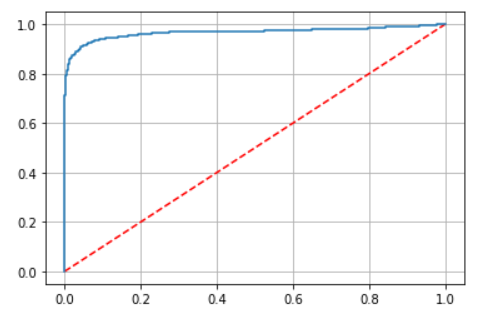
● Using the f1\_score function, we have got an f1\_score of 0.87.

● The scores can further be improved by selecting the features using different techniques.

● The scores can be checked and compared against different supervised learning models.

**ROC-AUC Curve:**

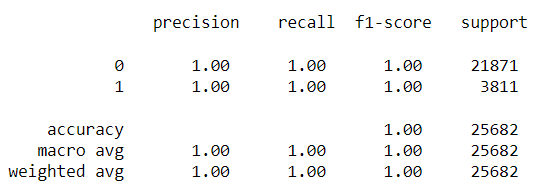
We plotted the receiver operating characteristic(roc) curve for our model and we can see that our model is able to distinguish between positive and negative class data. And out AUC is also high.



**Decision Tree:**

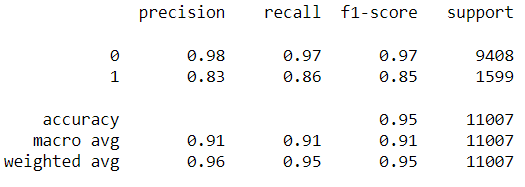
Applying decision tree classifier to the dataset.

**Train Data Performance:**



**F1 score : 1.0**

**Test Data Performance:**



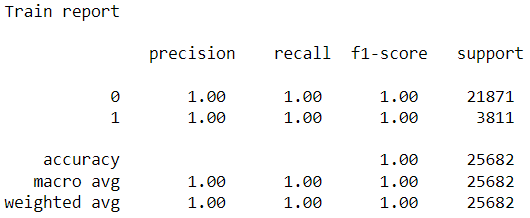
**F1 score of 0.8511.**

Overfitting can be seen in the model.

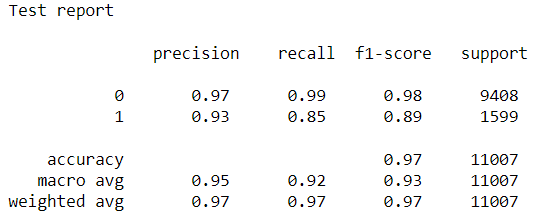
As compared to the previous model the F1 score has reduced a little from 0.87 to 0.85.

The mean of 10 fold cross validation for this model is 0.94.

To improve the decision tree based model, GridSearchCV was applied to obtain the best parameters. These parameters were then passed in the decision tree classifier to obtain the following results:



**F1 score: 1.0**



**F1 score: 0.88765**

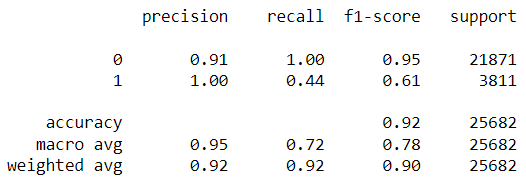
Clearly overfitting can still be seen.

The mean of 10 fold cross validation for this model is 0.968. So, the mean of the cross validation score has increased by 2%.

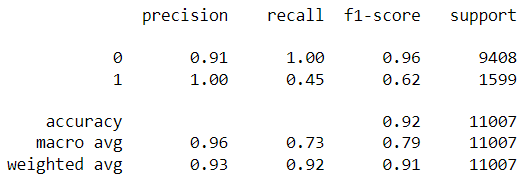
Recall has increased by quite a margin as compared to the base (Logistic Regression) model.

**Ridge Classifier:**

**Train Data Performance:**

****

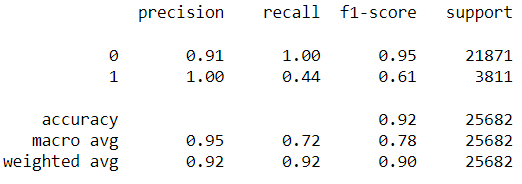
**Test Data Performance:**

****

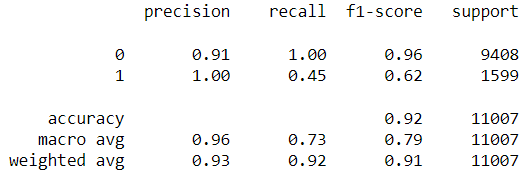
**Ridge Classifier CV:**

RidgeClassifierCV was used to give multiple alpha values, in the range 0 to 1, as an input to get the best value of alpha. The best alpha value was then used on the train and test data.

**Train Data Performance:**

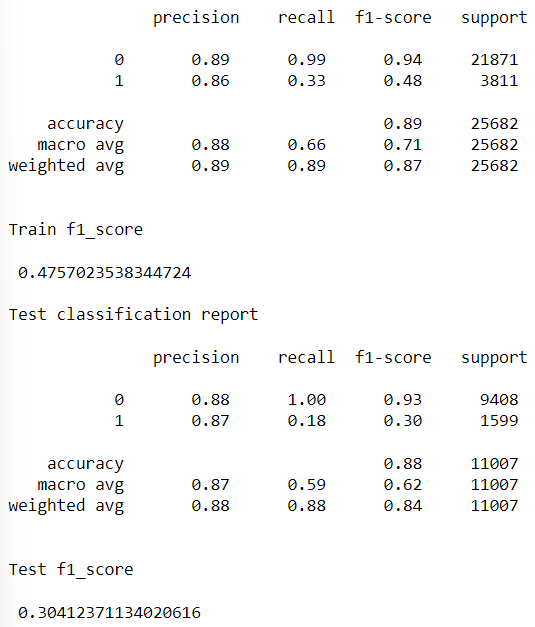


**Test Data Performance:**

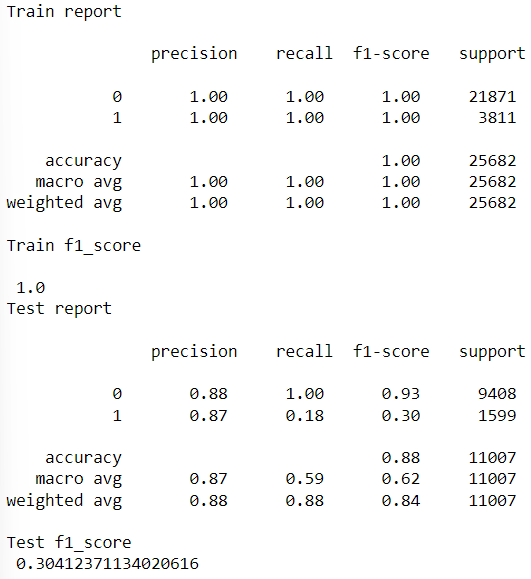
****

**K-Nearest Neighbors(KNN):**

KNN was applied to the dataset after scaling the attributes.

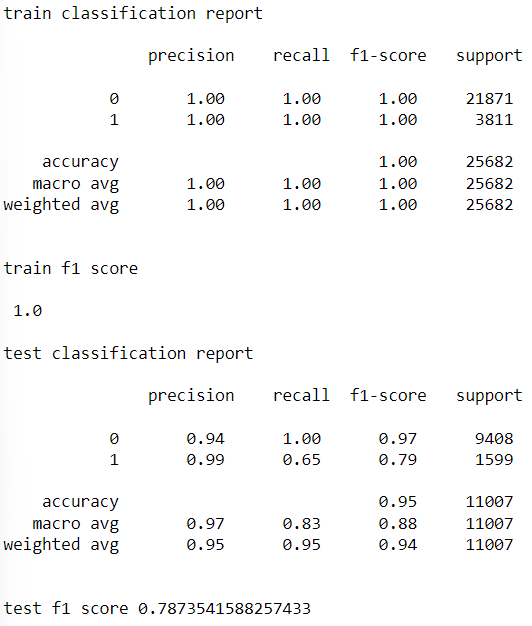


To improve the KNN base model, GridSearchCV was applied to obtain the best parameters. These parameters were then passed in KNN classifier to obtain the following results:



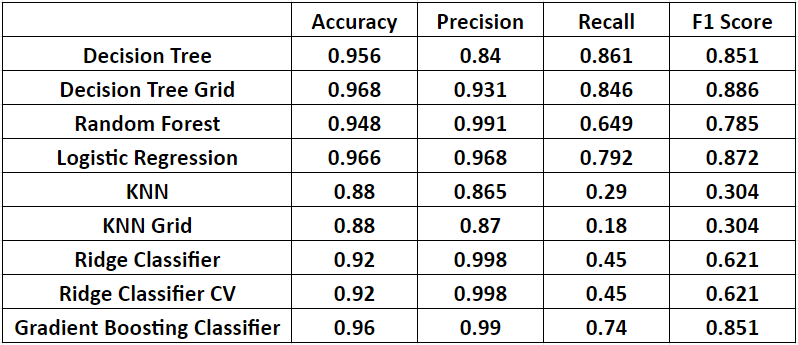
Grid search on KNN has led to the overfitting of data.

**Random Forest :**

****

Overfitting can be seen here.

**Model Comparisons:**

****

**Conclusion:**

To summarize our project, we can say that we have successfully created an efficient model which will determine whether a loan will be defaulted or not based on the bare minimum number of strong constraints.

We compared the results of our dataset by fitting on 8 models which included Logistic Regression, Decision Tree Classifier, Grid Search using Decision Tree, Ridge Classifier, KNN, Grid Search using KNN, Random Forest Classifier and Gradient Boosting.

Of these models except for Logistic Regression, Ridge Classifier, KNN and Gradient Boosting in all other models overfitting was observed. However, though KNN was not overfitting, it had a low recall and F1 score.

Among Logistic regression, Ridge classifier and Gradient boosting, Logistic regression returned the best results overall and hence we finalised it as our final model.

|  |  |
| --- | --- |
|  |  |