LaonTap - Logistic Regression

LoanTap is an online platform committed to delivering customized loan products to millennials. They innovate in an otherwise dull loan segment, to deliver instant, flexible loans on consumer-friendly terms to salaried professionals and businessmen.

The data science team at LoanTap is building an underwriting layer to determine the creditworthiness of MSMEs as well as individuals.

Problem Statement: Given a set of attributes for an Individual, determine if a credit line should be extended to them. The main challenge is to minimise the risk of NPAs by flagging defaulters while maximising opportunity to earn interest by disbursing loans to as many customers as possible.

Data dictionary:

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.
2. term : The number of payments on the loan. Values are in months and can be either 36 or 60.
3. int_rate : Interest Rate on the loan
4. installment : The monthly payment owed by the borrower if the loan originates.

5. grade : LoanTap assigned loan grade
6. sub_grade : LoanTap assigned loan subgrade
7. emp_title :The job title supplied by the Borrower when applying for the loan.
8. 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.
9. home_ownership : The home ownership status provided by the borrower during registration or obtained from the credit report.
10. annual_inc : The self-reported annual income provided by the borrower during registration.
11. verification_status : Indicates if income was verified by LoanTap, not verified, or if the income source was verified
12. issue_d : The month which the loan was funded
13. loan_status : Current status of the loan - Target Variable
14. purpose : A category provided by the borrower for the loan request.
15. title : The loan title provided by the borrower
16. dti : A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LoanTap loan, divided by the borrower's self-reported monthly income.
17. earliest_cr_line :The month the borrower's earliest reported credit line was opened
18. open_acc : The number of open credit lines in the borrower's credit file.
19. pub_rec : Number of derogatory public records
20. revol_bal : Total credit revolving balance
21. revol_util : Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
22. total_acc : The total number of credit lines currently in the borrower's credit file
23. initial_list_status : The initial listing status of the loan. Possible values are – W, F
24. application_type : Indicates whether the loan is an individual application or a joint application with two co-borrowers
25. mort_acc : Number of mortgage accounts.
26. pub_rec_bankruptcies : Number of public record bankruptcies
27. Address: Address of the individual

Importing Libraries

```
#Data processing
import pandas as pd
import numpy as np

#Data Visualisation
import matplotlib
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
%matplotlib inline
```

```
#Seting option for full column view of Data
pd.set_option('display.max_columns', None)

#Stats & model building
from scipy import stats
from sklearn.preprocessing import OneHotEncoder
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import MinMaxScaler
from sklearn.metrics import (accuracy_score, confusion_matrix,
                             roc_curve, auc, ConfusionMatrixDisplay,
                             f1_score, recall_score,
                             precision_score, precision_recall_curve,
                             average_precision_score,
classification_report)
from statsmodels.stats.outliers_influence import
variance_inflation_factor
from imblearn.over_sampling import SMOTE

#Hide warnings
import warnings
warnings.filterwarnings("ignore")
```

```
df = pd.read_csv("logistic_regression.csv")
df.head()
```

```
   loan_amnt        term  int_rate  installment grade sub_grade  \
0    10000.0   36 months     11.44       329.48     B        B4
1     8000.0   36 months     11.99       265.68     B        B5
2    15600.0   36 months     10.49       506.97     B        B3
3     7200.0   36 months      6.49       220.65     A        A2
4    24375.0   60 months     17.27       609.33     C        C5

                 emp_title emp_length home_ownership  annual_inc  \
0                Marketing  10+ years           RENT    117000.0
1          Credit analyst     4 years       MORTGAGE     65000.0
2             Statistician   < 1 year           RENT     43057.0
3           Client Advocate    6 years           RENT     54000.0
4  Destiny Management Inc.    9 years       MORTGAGE     55000.0

  verification_status   issue_d  loan_status             purpose  \
0        Not Verified  Jan-2015   Fully Paid            vacation
1        Not Verified  Jan-2015   Fully Paid  debt_consolidation
2     Source Verified  Jan-2015   Fully Paid         credit_card
```

```
3        Not Verified  Nov-2014   Fully Paid         credit_card
4            Verified Apr-2013  Charged Off          credit_card

                       title    dti earliest_cr_line  open_acc  pub_rec
\
0                   Vacation  26.24         Jun-1990      16.0      0.0

1         Debt consolidation  22.05         Jul-2004      17.0      0.0

2    Credit card refinancing  12.79         Aug-2007      13.0      0.0

3    Credit card refinancing   2.60         Sep-2006       6.0      0.0

4      Credit Card Refinance  33.95         Mar-1999      13.0      0.0


   revol_bal revol_util total_acc initial_list_status
application_type \
0    36369.0        41.8       25.0                   w
INDIVIDUAL
1    20131.0        53.3       27.0                   f
INDIVIDUAL
2    11987.0        92.2       26.0                   f
INDIVIDUAL
3     5472.0        21.5       13.0                   f
INDIVIDUAL
4    24584.0        69.8       43.0                   f
INDIVIDUAL

   mort_acc  pub_rec_bankruptcies  \
0       0.0                   0.0
1       3.0                   0.0
2       0.0                   0.0
3       0.0                   0.0
4       1.0                   0.0

                                             address
0     0174 Michelle Gateway\r\nMendozaberg, OK 22690
1  1076 Carney Fort Apt. 347\r\nLoganmouth, SD 05113
2  87025 Mark Dale Apt. 269\r\nNew Sabrina, WV 05113
3            823 Reid Ford\r\nDelacruzside, MA 00813
4             679 Luna Roads\r\nGreggshire, VA 11650
```

```
#shape of data

df.shape
```

```
(396030, 27)
```

```python
# Statistical summary

df.describe()
```

```
             loan_amnt       int_rate    installment    annual_inc  \
count  396030.000000  396030.000000  396030.000000  3.960300e+05
mean    14113.888089      13.639400     431.849698  7.420318e+04
std      8357.441341       4.472157     250.727790  6.163762e+04
min       500.000000       5.320000      16.080000  0.000000e+00
25%      8000.000000      10.490000     250.330000  4.500000e+04
50%     12000.000000      13.330000     375.430000  6.400000e+04
75%     20000.000000      16.490000     567.300000  9.000000e+04
max     40000.000000      30.990000    1533.810000  8.706582e+06

                   dti       open_acc        pub_rec     revol_bal  \
count  396030.000000  396030.000000  396030.000000  3.960300e+05
mean       17.379514      11.311153       0.178191  1.584454e+04
std        18.019092       5.137649       0.530671  2.059184e+04
min         0.000000       0.000000       0.000000  0.000000e+00
25%        11.280000       8.000000       0.000000  6.025000e+03
50%        16.910000      10.000000       0.000000  1.118100e+04
75%        22.980000      14.000000       0.000000  1.962000e+04
max      9999.000000      90.000000      86.000000  1.743266e+06

            revol_util      total_acc       mort_acc
pub_rec_bankruptcies
count  395754.000000  396030.000000  358235.000000
395495.000000
mean       53.791749      25.414744       1.813991
0.121648
std        24.452193      11.886991       2.147930
0.356174
min         0.000000       2.000000       0.000000
0.000000
25%        35.800000      17.000000       0.000000
0.000000
50%        54.800000      24.000000       1.000000
0.000000
75%        72.900000      32.000000       3.000000
0.000000
max       892.300000     151.000000      34.000000
8.000000
```

Data Cleaning

```python
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 396030 entries, 0 to 396029
Data columns (total 27 columns):
 #   Column                Non-Null Count   Dtype

 0   loan_amnt             396030 non-null  float64
 1   term                  396030 non-null  object
 2   int_rate              396030 non-null  float64
 3   installment           396030 non-null  float64
 4   grade                 396030 non-null  object
 5   sub_grade             396030 non-null  object
 6   emp_title             373103 non-null  object
 7   emp_length            377729 non-null  object
 8   home_ownership        396030 non-null  object
 9   annual_inc            396030 non-null  float64
 10  verification_status   396030 non-null  object
 11  issue_d               396030 non-null  object
 12  loan_status           396030 non-null  object
 13  purpose               396030 non-null  object
 14  title                 394275 non-null  object
 15  dti                   396030 non-null  float64
 16  earliest_cr_line      396030 non-null  object
 17  open_acc              396030 non-null  float64
 18  pub_rec               396030 non-null  float64
 19  revol_bal             396030 non-null  float64
 20  revol_util            395754 non-null  float64
 21  total_acc             396030 non-null  float64
 22  initial_list_status   396030 non-null  object
 23  application_type      396030 non-null  object
 24  mort_acc              358235 non-null  float64
 25  pub_rec_bankruptcies  395495 non-null  float64
 26  address               396030 non-null object
dtypes: float64(12), object(15)
memory usage: 81.6+ MB
```

Checking Column Datatypes

```
# Non-numeric columns
cat_cols = df.select_dtypes(include='object').columns
cat_cols

Index(['term', 'grade', 'sub_grade', 'emp_title', 'emp_length',
       'home_ownership', 'verification_status', 'issue_d',
'loan_status',
       'purpose', 'title', 'earliest_cr_line', 'initial_list_status',
       'application_type', 'address'],
      dtype='object')
```

```
# Number of unique values in all non-numeric columns
for col in cat_cols:
  print(f"No. of unique values in {col}: {df[col].nunique()}")
```

```
No. of unique values in term: 2
No. of unique values in grade: 7
No. of unique values in sub_grade: 35
No. of unique values in emp_title: 173105
No. of unique values in emp_length: 11
No. of unique values in home_ownership: 6
No. of unique values in verification_status: 3
No. of unique values in issue_d: 115
No. of unique values in loan_status: 2
No. of unique values in purpose: 14
No. of unique values in title: 48817
No. of unique values in earliest_cr_line: 684
No. of unique values in initial_list_status: 2
No. of unique values in application_type: 3
No. of unique values in address: 393700
```

```
# Convert earliest credit line & issue date to datetime
df['earliest_cr_line'] = pd.to_datetime(df['earliest_cr_line'])
df['issue_d'] = pd.to_datetime(df['issue_d'])

#Convert employment length to numeric
d = {'10+ years':10, '4 years':4, '< 1 year':0,
     '6 years':6, '9 years':9,'2 years':2, '3 years':3,
     '8 years':8, '7 years':7, '5 years':5, '1 year':1}
df['emp_length']=df['emp_length'].replace(d)

#Convert columns with less number of unique values to categorical
columns
cat_cols = ['term', 'grade','sub_grade','home_ownership',
            'verification_status','loan_status','purpose',
            'initial_list_status','application_type']

df[cat_cols] = df[cat_cols].astype('category')

df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 396030 entries, 0 to 396029
Data columns (total 27 columns):
 #   Column                Non-Null Count   Dtype

 0   loan_amnt             396030 non-null  float64
 1   term                  396030 non-null  category
 2   int_rate              396030 non-null  float64
 3   installment           396030 non-null  float64
 4   grade                 396030 non-null  category
 5   sub_grade             396030 non-null  category
```

```
 6   emp_title             373103 non-null  object
 7   emp_length            377729 non-null  float64
 8   home_ownership        396030 non-null  category
 9   annual_inc            396030 non-null  float64
 10  verification_status   396030 non-null  category
 11  issue_d               396030 non-null  datetime64[ns]
 12  loan_status           396030 non-null  category
 13  purpose               396030 non-null  category
 14  title                 394275 non-null  object
 15  dti                   396030 non-null  float64
 16  earliest_cr_line      396030 non-null  datetime64[ns]
 17  open_acc              396030 non-null  float64
 18  pub_rec               396030 non-null  float64
 19  revol_bal             396030 non-null  float64
 20  revol_util            395754 non-null  float64
 21  total_acc             396030 non-null  float64
 22  initial_list_status   396030 non-null  category
 23  application_type      396030 non-null  category
 24  mort_acc              358235 non-null  float64
 25  pub_rec_bankruptcies  395495 non-null  float64
 26  address               396030 non-null  object
dtypes: category(9), datetime64[ns](2), float64(13), object(3)
memory usage: 57.8+ MB
```

Check for Duplicate Values

```
df.duplicated().sum()
```

```
0
```

There are no duplicate instances in the data

Handling Missing Values

```
df.isna().sum()
```

```
loan_amnt                  0
term                       0
int_rate                   0
installment                0
grade                      0
sub_grade                  0
emp_title              22927
emp_length             18301
home_ownership             0
annual_inc                 0
verification_status        0
issue_d                    0
loan_status                0
```

```
purpose                      0
title                     1755
dti                          0
earliest_cr_line             0
open_acc                     0
pub_rec                      0
revol_bal                    0
revol_util                 276
total_acc                    0
initial_list_status          0
application_type             0
mort_acc                 37795
pub_rec_bankruptcies       535
address                      0
dtype: int64
```

```
#Filling missing values with 'Unknown' for object dtype
fill_values = {'title': 'Unknown', 'emp_title': 'Unknown'}
df.fillna(value=fill_values, inplace=True)

#Mean aggregation of mort_acc by total_acc to fill missing values

avg_mort = df.groupby('total_acc')['mort_acc'].mean()

def fill_mort(total_acc, mort_acc):
  if np.isnan(mort_acc):
    return avg_mort[total_acc].round()
  else:
    return mort_acc

df['mort_acc'] = df.apply(lambda x:
fill_mort(x['total_acc'],x['mort_acc']), axis=1)

df.dropna(inplace=True)

df.isna().sum()
```

```
loan_amnt               0
term                    0
int_rate                0
installment             0
grade                   0
sub_grade               0
emp_title               0
emp_length              0
home_ownership          0
annual_inc              0
verification_status     0
issue_d                 0
loan_status             0
purpose                 0
```

```
title                      0
dti                        0
earliest_cr_line           0
open_acc                   0
pub_rec                    0
revol_bal                  0
revol_util                 0
total_acc                  0
initial_list_status        0
application_type           0
mort_acc                   0
pub_rec_bankruptcies       0
address                    0
dtype: int64
```

```
df.shape
```

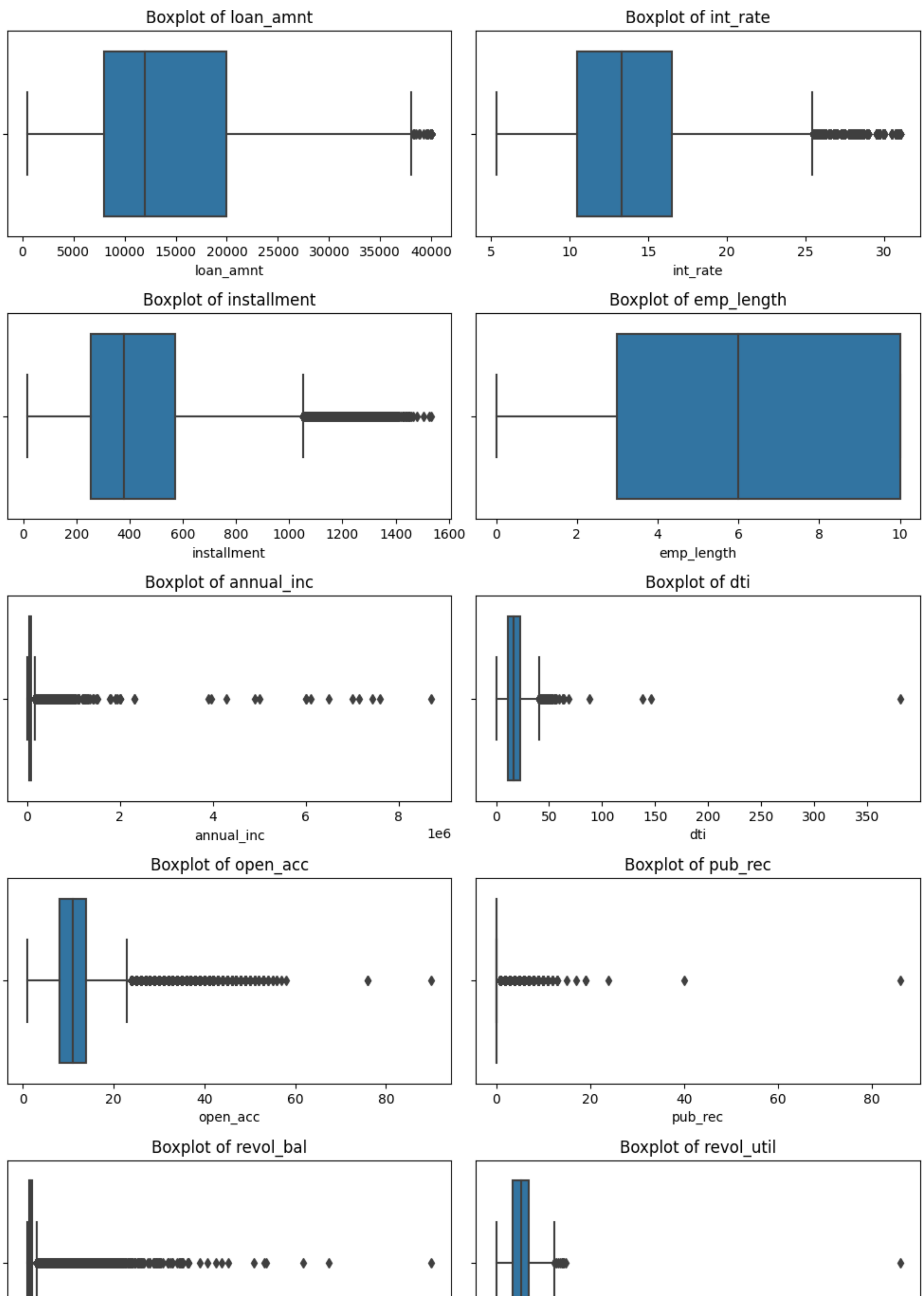
```
(376929, 27)
```

Outlier Treatment

```
num_cols = df.select_dtypes(include='number').columns
num_cols
```

```
Index(['loan_amnt', 'int_rate', 'installment', 'emp_length',
'annual_inc',
       'dti', 'open_acc', 'pub_rec', 'revol_bal', 'revol_util',
'total_acc',
       'mort_acc', 'pub_rec_bankruptcies'],
      dtype='object')
```

```
fig = plt.figure(figsize=(10,21))
i=1
for col in num_cols:
  ax = plt.subplot(7,2,i)
  sns.boxplot(x=df[col])
  plt.title(f'Boxplot of {col}')
  i += 1

plt.tight_layout()
plt.show()
```



Boxplot of loan_amnt
0 5000 10000 15000 20000 25000 30000 35000 40000
loan_amnt
Boxplot of int_rate
5 10 15 20 25 30
int_rate
Boxplot of installment
0 200 400 600 800 1000 1200 1400 1600
installment
Boxplot of emp_length
0 2 4 6 8 10
emp_length
Boxplot of annual_inc
0 2 4 6 8
1e6
annual_inc
Boxplot of dti
0 50 100 150 200 250 300 350
dti
Boxplot of open_acc
0 20 40 60 80
open_acc
Boxplot of pub_rec
0 20 40 60 80
pub_rec
Boxplot of revol_bal
Boxplot of revol_util

Here we can see that many columns have outliers. Lets remove the rows with outliers using standard deviation (99% data is within 3 standard deviations in case of normally distributed data).
For pub_Rec and pub_rec_bankruptcies, we can apply the 0 or 1 approach

```
# Convert pub_rec and pub_rec_bankruptcies to categorical variables

df['pub_rec_bankruptcies'] =
np.where(df['pub_rec_bankruptcies']>0,'yes','no')
df['pub_rec'] = np.where(df['pub_rec']>0,'yes','no')
df[['pub_rec_bankruptcies','pub_rec']] =
df[['pub_rec_bankruptcies','pub_rec']].astype('category')

# Numeric columns after converting public records to category
num_cols = df.select_dtypes(include='number').columns
num_cols
```

```
Index(['loan_amnt', 'int_rate', 'installment', 'emp_length',
'annual_inc',
       'dti', 'open_acc', 'revol_bal', 'revol_util', 'total_acc',
'mort_acc'],
      dtype='object')
```

```
#Removing outliers using standard deviation
for col in num_cols:
  mean=df[col].mean()
  std=df[col].std()
  upper = mean + (3*std)
  df = df[~(df[col]>upper)]

df.shape
```

```
(350845, 27)
```

Feature Engineering

```
df['address'].sample(10)
```

```
285569                    Unit 9894 Box 9319\r\nDPO AA 05113
101576      2867 Lindsey Shoal\r\nWilliamschester, LA 00813
139688          008 Alicia Gateway\r\nLake Stacey, VT 30723
160700          2315 Pamela Park\r\nNew Aaronbury, HI 05113
219251    3828 Jack Squares Suite 231\r\nRodriguezhaven,...
373303            0136 Tina Inlet\r\nNew Frankton, MO 30723
299038           949 Adam Track\r\nNorth Ryanberg, HI 00813
109538          12924 White Island\r\nLisaborough, WI 30723
360921            0203 Keith Neck\r\nEast Brandon, SC 30723
319103       607 Jennifer Path\r\nNew Williamtown, HI 48052
Name: address, dtype: object
```

```
# Deriving zip code and state from address
df[['state', 'zip_code']] = df['address'].apply(lambda x:
pd.Series([x[-8:-6], x[-5:]]))

#Drop address
df.drop(["address"], axis = 1, inplace=True)

df.zip_code.nunique()
```

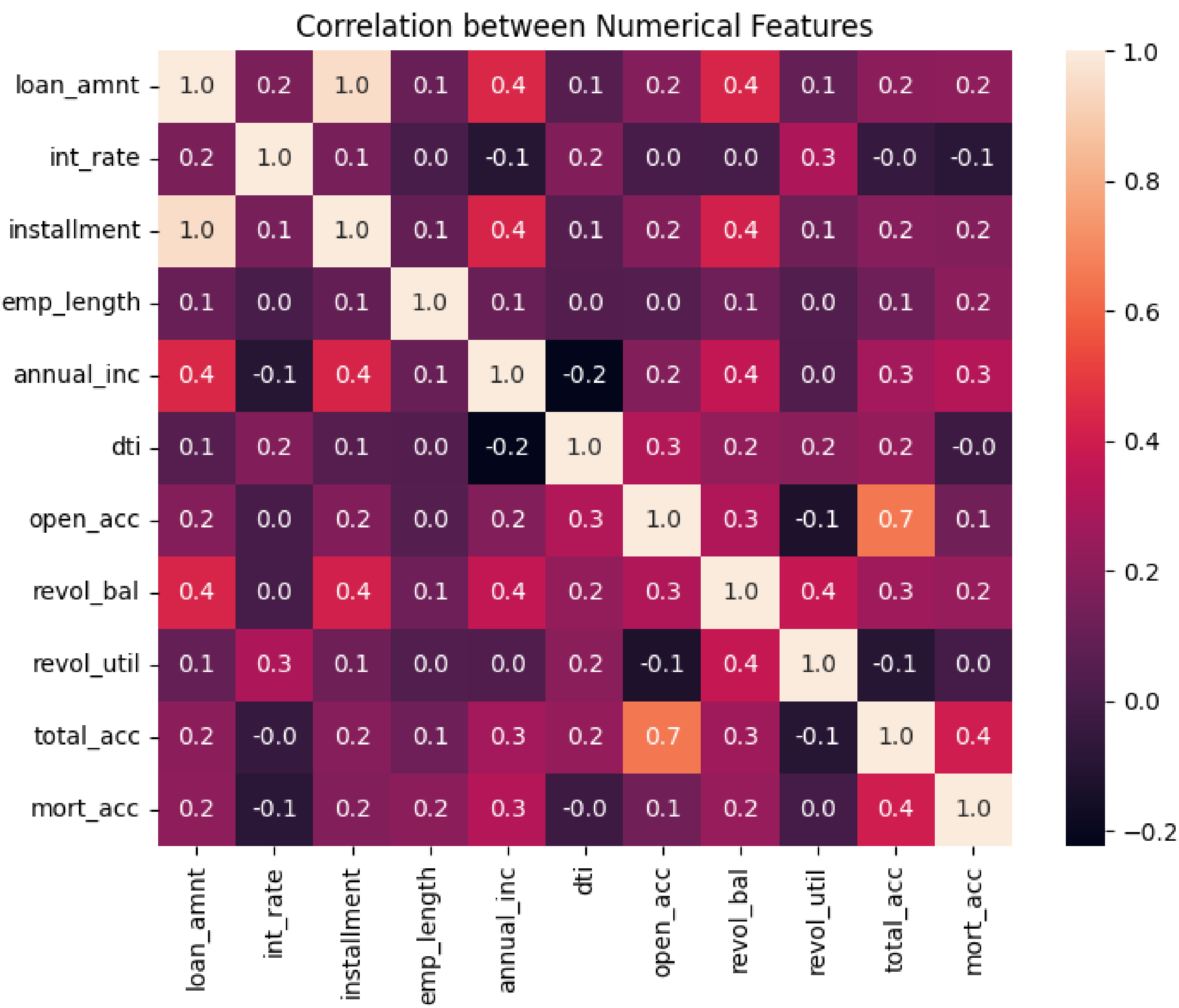
```
10
```

Since there are only 10 zipcodes, we can change the datatype of zipcodes to categorical

```
df['zip_code'] = df['zip_code'].astype('category')
```

Exploratory Data Analysis

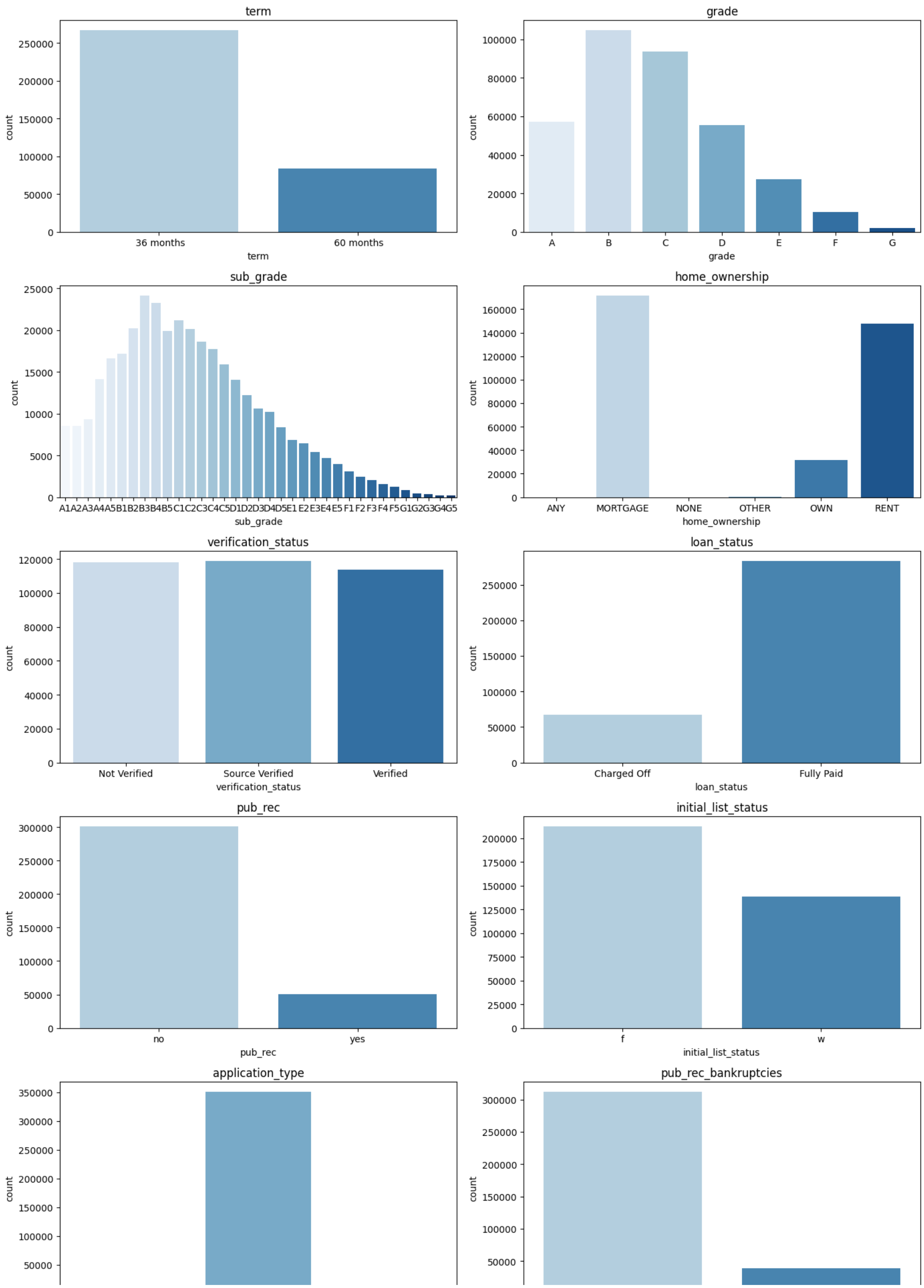
```
#Correlation between numerical features
plt.figure(figsize=(8,6))
sns.heatmap(df.corr(), annot=True, fmt=".1f")
plt.title('Correlation between Numerical Features')
plt.show()
```



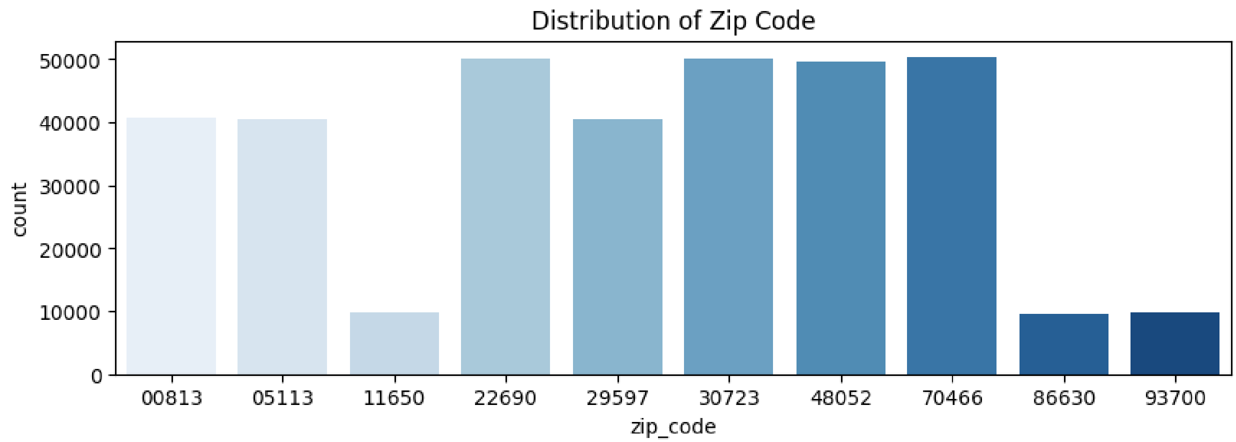
1. loan_amnt and installment are perfectly correlated
2. total_acc is highly correlated with open_acc
3. total_acc is moderately correlated with mort_acc We can remove some of these correlated features to avoid multicolinearity

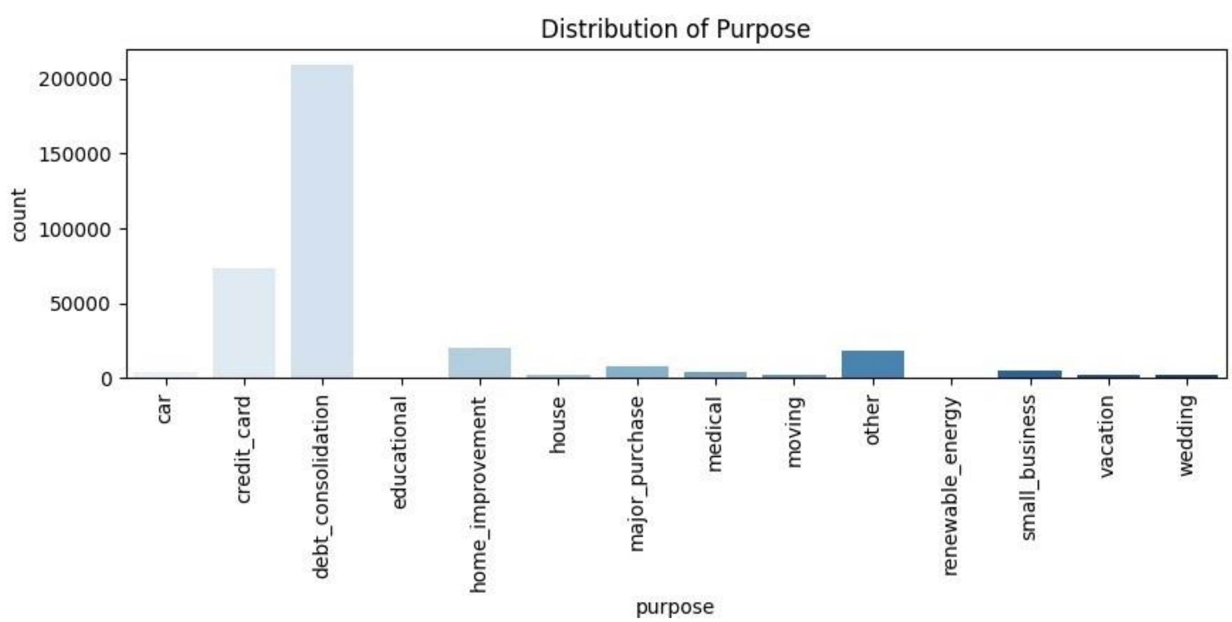
```
#Drop installment
df.drop(columns=['installment'], inplace=True)

#Distribution of categorical variables
plot = ['term', 'grade', 'sub_grade', 'home_ownership',
'verification_status',
       'loan_status', 'pub_rec', 'initial_list_status',
       'application_type', 'pub_rec_bankruptcies']

plt.figure(figsize=(14,20))
i=1
for col in plot:
  ax=plt.subplot(5,2,i)
```

```
  sns.countplot(x=df[col], palette='Blues')
  plt.title(f'{col}')
  i += 1

plt.tight_layout()
plt.show()
```


term
count
250000
200000
150000
100000
50000
0
36 months
60 months
term
grade
count
100000
80000
60000
40000
20000
0
A
B
C
D
E
F
G
grade
sub_grade
count
25000
20000
15000
10000
5000
0
A1A2A3A4A5B1B2B3B4B5C1C2C3C4C5D1D2D3D4D5E1E2E3E4E5 F1 F2 F3 F4 F5G1G2G3G4G5
sub_grade
home_ownership
count
160000
140000
120000
100000
80000
60000
40000
20000
0
ANY
MORTGAGE
NONE
OTHER
OWN
RENT
home_ownership
verification_status
count
120000
100000
80000
60000
40000
20000
0
Not Verified
Source Verified
Verified
verification_status
loan_status
count
250000
200000
150000
100000
50000
0
Charged Off
Fully Paid
loan_status
pub_rec
count
300000
250000
200000
150000
100000
50000
0
no
yes
pub_rec
initial_list_status
count
200000
175000
150000
125000
100000
75000
50000
25000
0
f
w
initial_list_status
application_type
count
350000
300000
250000
200000
150000
100000
50000
pub_rec_bankruptcies
count
300000
250000
200000
150000
100000
50000

```
plt.figure(figsize=(10,3))
sns.countplot(x=df['zip_code'], palette='Blues')
plt.title('Distribution of Zip Code')

plt.figure(figsize=(10,3))
sns.countplot(x=df['purpose'], palette='Blues')
plt.xticks(rotation=90)
plt.title('Distribution of Purpose')

plt.show()
```



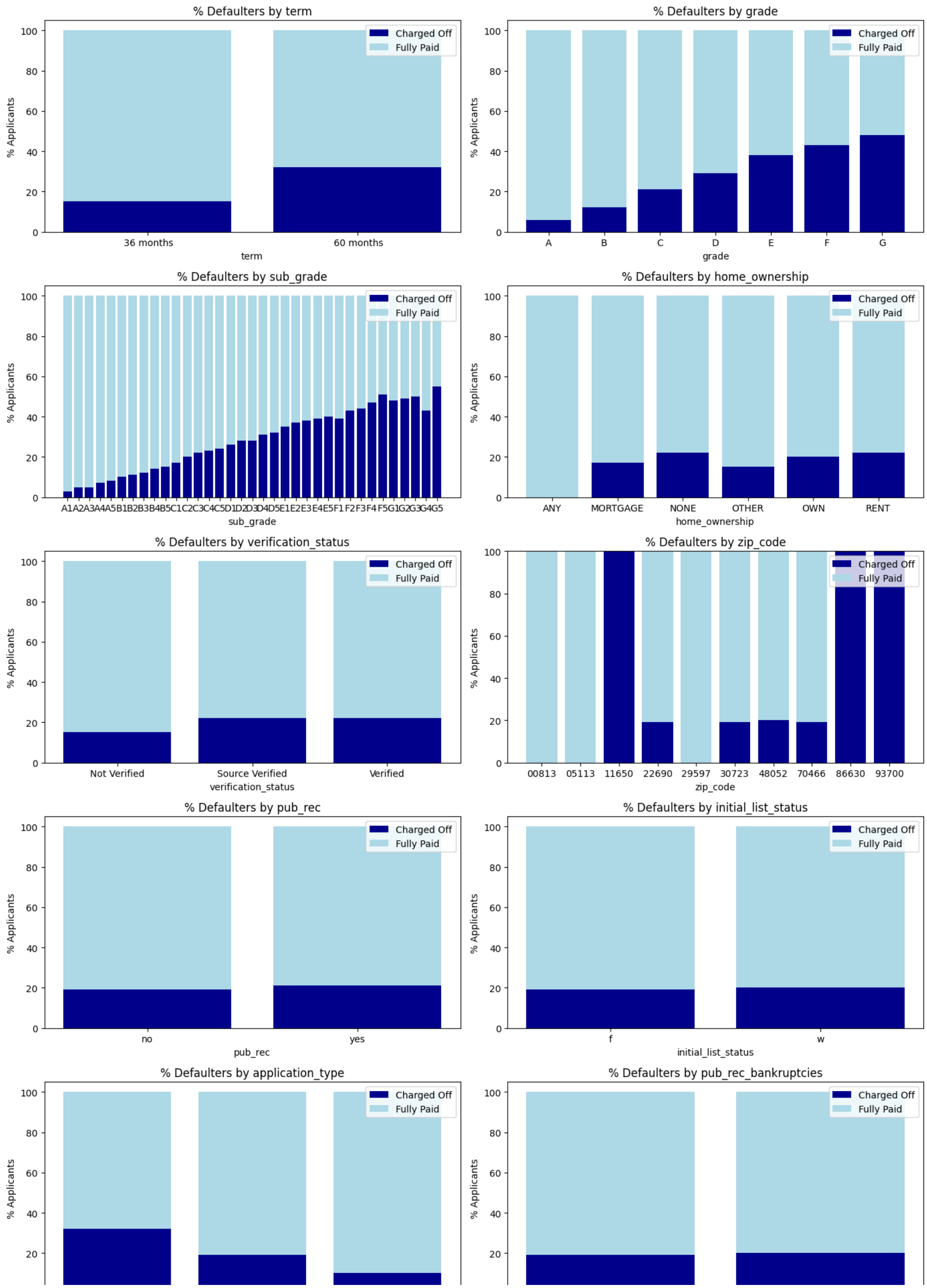


Observations:

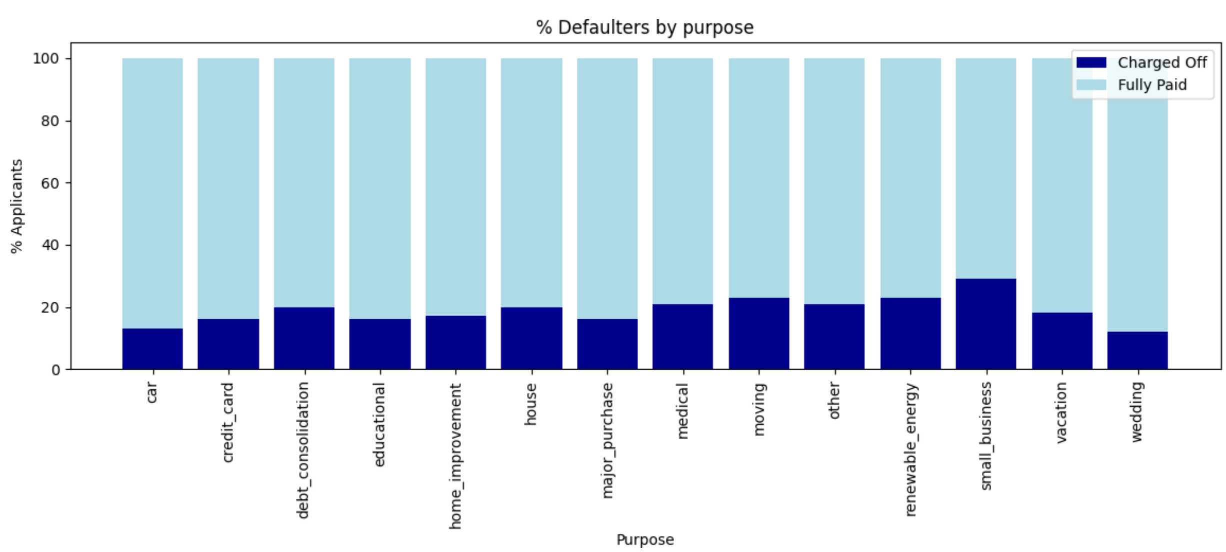
- Almost 80% loans are of 36 months term
- Maximum loans (30%) fall in B grade, followed by C,A & D respectively
- The type of home ownership for 50% cases is mortgage

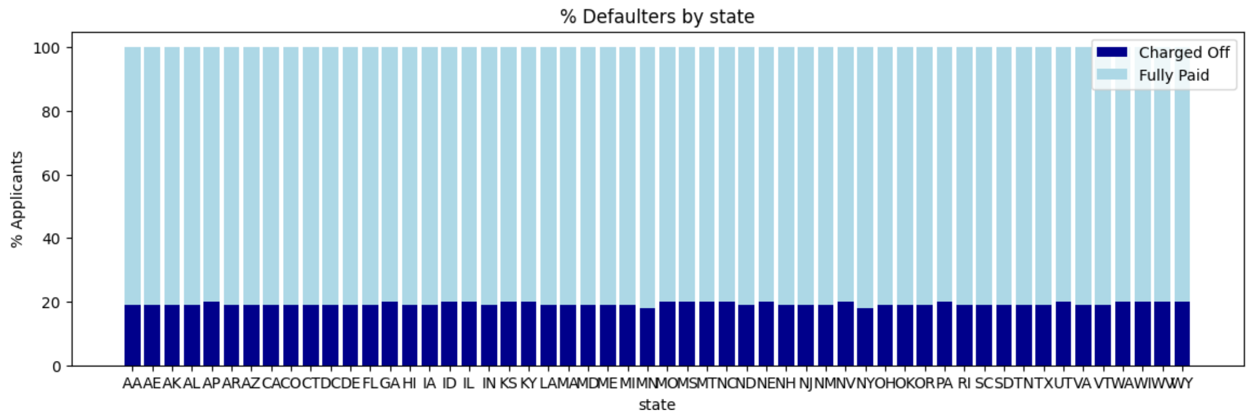
- The target variable (loan status) is imbalanced in the favour of fully-paid loans. Defaulters are approx 25% of fully paid instances.
- 85% of applicants don't have a public record/haven't filled for bankruptcy
- 99% applicants have applied under 'individual' application type
- 55% of loans are taken for the purpose of debt consolidation followed by 20% on credit card

```
# Impact of categorical factors on loan status

plot = ['term', 'grade', 'sub_grade', 'home_ownership',
'verification_status',
       'zip_code', 'pub_rec', 'initial_list_status',
       'application_type', 'pub_rec_bankruptcies']

plt.figure(figsize=(14,20))
i=1
for col in plot:
  ax=plt.subplot(5,2,i)

  data = df.pivot_table(index=col, columns='loan_status',
aggfunc='count', values='purpose')
  data = data.div(data.sum(axis=1), axis=0).multiply(100).round()
  data.reset_index(inplace=True)

  plt.bar(data[col],data['Charged Off'], color='#00008b')
  plt.bar(data[col],data['Fully Paid'], color='#add8e6',
bottom=data['Charged Off'])
  plt.xlabel(f'{col}')
  plt.ylabel('% Applicants')
  plt.title(f'% Defaulters by {col}')
  plt.legend(['Charged Off','Fully Paid'])
  i += 1

plt.tight_layout()
plt.show()
```


% Defaulters by term
Charged Off
Fully Paid
% Applicants
36 months
60 months
term
% Defaulters by grade
A B C D E F G
grade
% Defaulters by sub_grade
A1A2A3A4A5B1B2B3B4B5C1C2C3C4C5D1D2D3D4D5E1E2E3E4E5F1F2F3F4F5G1G2G3G4G5
sub_grade
% Defaulters by home_ownership
ANY MORTGAGE NONE OTHER OWN RENT
home_ownership
% Defaulters by verification_status
Not Verified
Source Verified
Verified
verification_status
% Defaulters by zip_code
00813 05113 11650 22690 29597 30723 48052 70466 86630 93700
zip_code
% Defaulters by pub_rec
no
yes
pub_rec
% Defaulters by initial_list_status
f
w
initial_list_status
% Defaulters by application_type
% Defaulters by pub_rec_bankruptcies

```python
# Impact of Purpose/state on loan status

purpose = df.pivot_table(index='purpose', columns='loan_status',
aggfunc='count', values='sub_grade')
purpose = purpose.div(purpose.sum(axis=1),
axis=0).multiply(100).round()
purpose.reset_index(inplace=True)

plt.figure(figsize=(14,4))
plt.bar(purpose['purpose'],purpose['Charged Off'], color='#00008b')
plt.bar(purpose['purpose'],purpose['Fully Paid'], color='#add8e6',
bottom=purpose['Charged Off'])
plt.xlabel('Purpose')
plt.ylabel('% Applicants')
plt.title('% Defaulters by purpose')
plt.legend(['Charged Off','Fully Paid'])
plt.xticks(rotation=90)
plt.show()

state = df.pivot_table(index='state', columns='loan_status',
aggfunc='count', values='sub_grade')
state = state.div(state.sum(axis=1), axis=0).multiply(100).round()
state.reset_index(inplace=True)

plt.figure(figsize=(14,4))
plt.bar(state['state'],state['Charged Off'], color='#00008b')
plt.bar(state['state'],state['Fully Paid'], color='#add8e6',
bottom=state['Charged Off'])
plt.xlabel('state')
plt.ylabel('% Applicants')
plt.title('% Defaulters by state')
plt.legend(['Charged Off','Fully Paid'])
plt.show()
```

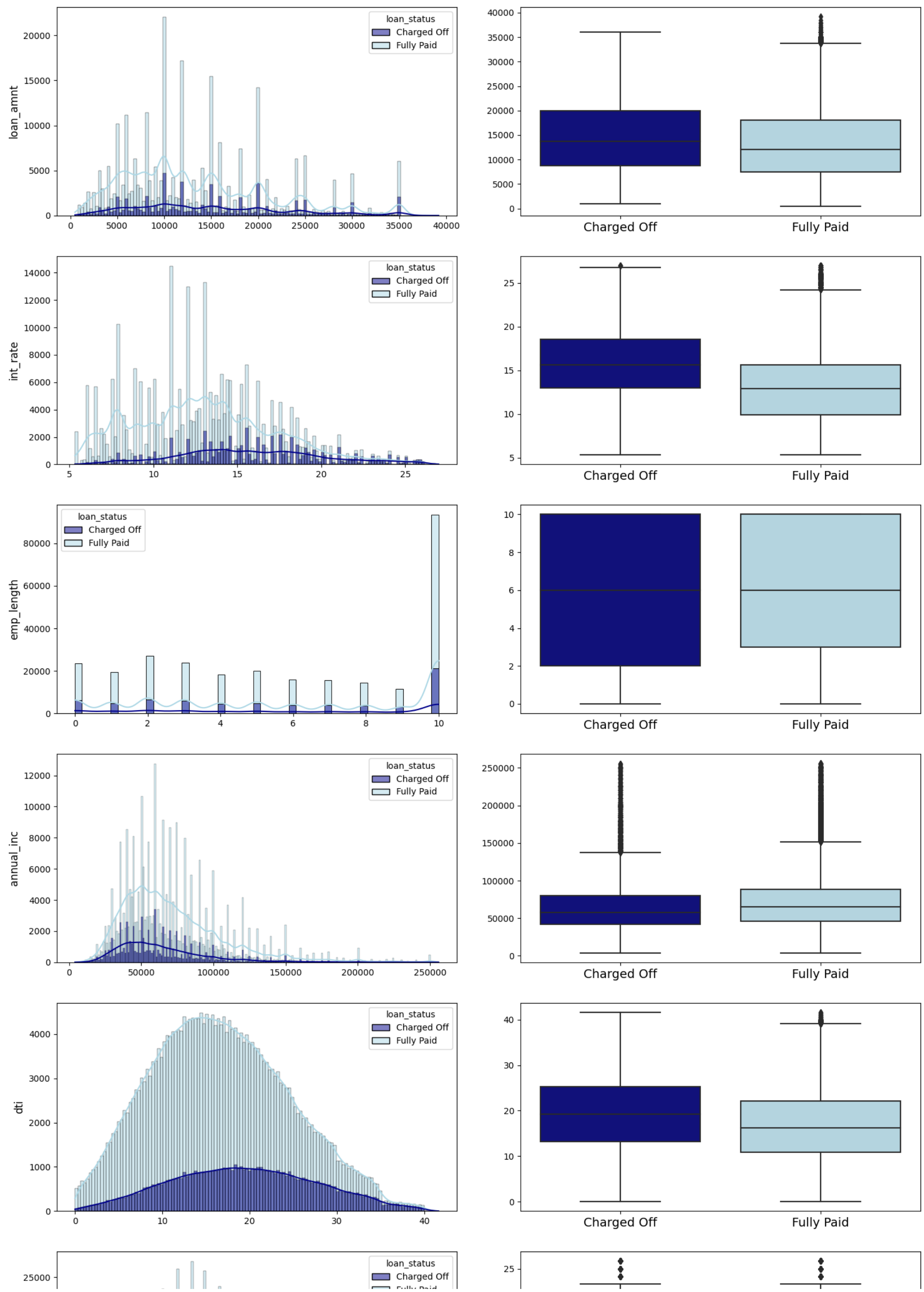




Observations:

- The % of defaulters is much higher for longer (60-month) term
- As expected, grade/sub-grade has the maximum impact on loan_status with highest grade having maximum defaulters
- Zip codes such as 11650, 86630 and 93700 have 100% defaulters
- We can remove initial_list_status and state as they have no impact on loan_status
- public records also don't seem to have any impact on loan_status surprisingly
- Direct pay application type has higher default rate compared to individual/joint
- Loan taken for the purpose of small business has the highest rate of default

```
# Impact of numerical features on loan_status

num_cols = df.select_dtypes(include='number').columns

fig, ax = plt.subplots(10,2,figsize=(15,40))
i=0
color_dict = {'Fully Paid': matplotlib.colors.to_rgba('#add8e6', 0.5),
              'Charged Off': matplotlib.colors.to_rgba('#00008b', 1)}
for col in num_cols:
    sns.histplot(data=df, x=col, hue='loan_status', ax=ax[i, 0],
legend=True,
                palette=color_dict, kde=True, fill=True)
    sns.boxplot(data=df, y=col, x='loan_status', ax=ax[i,1],
               palette=('#00008b', '#add8e6'))
    ax[i,0].set_ylabel(col, fontsize=12)
    ax[i,0].set_xlabel(' ')
    ax[i,1].set_xlabel(' ')
    ax[i,1].set_ylabel(' ')
    ax[i,1].xaxis.set_tick_params(labelsize=14)
    i += 1

plt.tight_layout()
plt.show()
```



loan_status
Charged Off
Fully Paid
loan_amnt
int_rate
emp_length
annual_inc
dti
Charged Off
Fully Paid

Observations:

- From the boxplots, it can be observed that the mean loan_amnt, int_rate, dti, open_acc and revol_util are slightly higher for defaulters while annual income is lower

```
# Remove columns which do not have an impact on loan_status
df.drop(columns=['initial_list_status','state',
                 'emp_title', 'title','earliest_cr_line',
                 'issue_d','sub_grade'], inplace=True)

# Subgrade is removed because grade and subgrade are similar features
```

Data Pre-Processing

```
# Encoding Target Variable

df['loan_status']=df['loan_status'].map({'Fully Paid': 0, 'Charged 
Off':1}).astype(int)

x = df.drop(columns=['loan_status'])
x.reset_index(inplace=True, drop=True)
y = df['loan_status']
y.reset_index(drop=True, inplace=True)

# Encoding Binary features into numerical dtype

x['term']=x['term'].map({' 36 months': 36, ' 60 
months':60}).astype(int)
x['pub_rec']=x['pub_rec'].map({'no': 0, 'yes':1}).astype(int)
x['pub_rec_bankruptcies']=x['pub_rec_bankruptcies'].map({'no': 0, 
'yes':1}).astype(int)
```

One Hot Encoding of Categorical Features

```
cat_cols = x.select_dtypes('category').columns

encoder = OneHotEncoder(sparse=False)
encoded_data = encoder.fit_transform(x[cat_cols])
encoded_df = pd.DataFrame(encoded_data, 
columns=encoder.get_feature_names_out(cat_cols))
x = pd.concat([x,encoded_df], axis=1)
x.drop(columns=cat_cols, inplace=True)
x.head()
```

```
   loan_amnt  term  int_rate  emp_length  annual_inc    dti  open_acc  \
0    10000.0    36     11.44        10.0    117000.0  26.24      16.0   

1     8000.0    36     11.99         4.0     65000.0  22.05      17.0   
```

```
2    15600.0    36     10.49          0.0     43057.0  12.79       13.0

3     7200.0    36      6.49          6.0     54000.0   2.60        6.0

4    24375.0    60     17.27          9.0     55000.0  33.95       13.0


   pub_rec revol_bal  revol_util  total_acc  mort_acc
pub_rec_bankruptcies  \
0        0    36369.0        41.8       25.0       0.0
0
1        0    20131.0        53.3       27.0       3.0
0
2        0    11987.0        92.2       26.0       0.0
0
3        0     5472.0        21.5       13.0       0.0
0
4        0    24584.0        69.8       43.0       1.0
0

   grade_A  grade_B  grade_C  grade_D  grade_E  grade_F  grade_G  \
0      0.0      1.0      0.0      0.0      0.0      0.0      0.0
1      0.0      1.0      0.0      0.0      0.0      0.0      0.0
2      0.0      1.0      0.0      0.0      0.0      0.0      0.0
3      1.0      0.0      0.0      0.0      0.0      0.0      0.0
4      0.0      0.0      1.0      0.0      0.0      0.0      0.0

   home_ownership_ANY  home_ownership_MORTGAGE  home_ownership_NONE  \
0                 0.0                      0.0                  0.0
1                 0.0                      1.0                  0.0
2                 0.0                      0.0                  0.0
3                 0.0                      0.0                  0.0
4                 0.0                      1.0                  0.0

   home_ownership_OTHER  home_ownership_OWN  home_ownership_RENT  \
0                   0.0                 0.0                  1.0
1                   0.0                 0.0                  0.0
2                   0.0                 0.0                  1.0
3                   0.0                 0.0                  1.0
4                   0.0                 0.0                  0.0

   verification_status_Not Verified verification_status_Source
Verified \
0                                1.0
0.0
1                                1.0
0.0
2                                0.0
1.0
```

```
3                                    1.0
0.0
4                                    0.0
0.0

   verification_status_Verified  purpose_car  purpose_credit_card  \
0                           0.0          0.0                  0.0
1                           0.0          0.0                  0.0
2                           0.0          0.0                  1.0
3                           0.0          0.0                  1.0
4                           1.0          0.0                  1.0

   purpose_debt_consolidation purpose_ed ucational
purpose_home_improvement \
0                         0.0                  0.0
0.0
1                         1.0                  0.0
0.0
2                         0.0                  0.0
0.0
3                         0.0                  0.0
0.0
4                         0.0                  0.0
0.0

   purpose_house  purpose_major_purchase  purpose_medical
purpose_moving \
0            0.0                     0.0              0.0
0.0
1            0.0                     0.0              0.0
0.0
2            0.0                     0.0              0.0
0.0
3            0.0                     0.0              0.0
0.0
4            0.0                     0.0              0.0
0.0

   purpose_other  purpose_renewable_energy  purpose_small_business
0            0.0                       0.0                     0.0
1            0.0                       0.0                     0.0
2            0.0                       0.0                     0.0
3            0.0                       0.0                     0.0
4            0.0                       0.0                     0.0

   purpose_vacation  purpose_wedding  application_type_DIRECT_PAY  \
0               1.0              0.0                          0.0
1               0.0              0.0                          0.0
2               0.0              0.0                          0.0
3               0.0              0.0                          0.0
```

```
4                  0.0               0.0                           0.0

   application_type_INDIVIDUAL  application_type_JOINT  zip_code_00813
\
0                          1.0                     0.0             0.0

1                          1.0                     0.0             0.0

2                          1.0                     0.0             0.0

3                          1.0                     0.0             1.0

4                          1.0                     0.0             0.0


   zip_code_05113  zip_code_11650  zip_code_22690  zip_code_29597  \
0             0.0             0.0             1.0             0.0
1             1.0             0.0             0.0             0.0
2             1.0             0.0             0.0             0.0
3             0.0             0.0             0.0             0.0
4             0.0             1.0             0.0             0.0

   zip_code_30723  zip_code_48052  zip_code_70466  zip_code_86630  \
0             0.0             0.0             0.0             0.0
1             0.0             0.0             0.0             0.0
2             0.0             0.0             0.0             0.0
3             0.0             0.0             0.0             0.0
4             0.0             0.0             0.0             0.0

   zip_code_93700
0             0.0
1             0.0
2             0.0
3             0.0
4             0.0
```

Train-Test Split

```
x_train, x_test, y_train, y_test =
train_test_split(x,y,test_size=0.20,stratify=y,random_state=42)

x_train.shape, y_train.shape, x_test.shape, y_test.shape

((280676, 56), (280676,), (70169, 56), (70169,))
```

Scaling Numeric Features

```
scaler = MinMaxScaler()
x_train = pd.DataFrame(scaler.fit_transform(x_train),
columns=x_train.columns)
```

```
x_test = pd.DataFrame(scaler.transform(x_test),
columns=x_test.columns)

x_train.tail()

        loan_amnt  term  int_rate  emp_length  annual_inc       dti
open_acc  \
280671   0.167959   0.0  0.141671         0.7    0.194444  0.255954
0.60
280672   0.497416   0.0  0.445778         0.4    0.182540  0.414482
0.24
280673   0.064599   0.0  0.686664         0.7    0.238095  0.220111
0.32
280674   0.245478   1.0  0.177665         0.9    0.313492  0.134953
0.92
280675   0.646641   1.0  0.885095         0.6    0.349206  0.747173
0.88

        pub_rec  revol_bal  revol_util  total_acc  mort_acc  \
280671      0.0   0.104275    0.271695   0.578947  0.428571
280672      0.0   0.224536    0.670722   0.263158  0.285714
280673      0.0   0.249454    0.622871   0.385965  0.428571
280674      0.0   0.080701    0.039740   0.842105  0.428571
280675      1.0   0.213775    0.543390   0.596491  0.714286

        pub_rec_bankruptcies  grade_A  grade_B  grade_C  grade_D
grade_E  \
280671                   0.0      1.0      0.0      0.0      0.0
0.0
280672                   0.0      0.0      0.0      1.0      0.0
0.0
280673                   0.0      0.0      0.0      0.0      1.0
0.0
280674                   0.0      0.0      1.0      0.0      0.0
0.0
280675                   1.0      0.0      0.0      0.0      0.0
0.0

        grade_F  grade_G  home_ownership_ANY  home_ownership_MORTGAGE
\
280671      0.0      0.0                 0.0                      0.0

280672      0.0      0.0                 0.0                      0.0

280673      0.0      0.0                 0.0                      1.0

280674      0.0      0.0                 0.0                      1.0

280675      1.0      0.0                 0.0                      1.0
```

```
        home_ownership_NONE  home_ownership_OTHER  home_ownership_OWN
\
280671                  0.0                   0.0                 0.0

280672                  0.0                   0.0                 0.0

280673                  0.0                   0.0                 0.0

280674                  0.0                   0.0                 0.0

280675                  0.0                   0.0                 0.0


        home_ownership_RENT  verification_status_Not Verified  \
280671                  1.0                               1.0
280672                  1.0                               0.0
280673                  0.0                               0.0
280674                  0.0                               0.0
280675                  0.0                               0.0

        verification_status_Source Verified
verification_status_Verified \
280671                                  0.0
0.0
280672                                  0.0
1.0
280673                                  1.0
0.0
280674                                  0.0
1.0
280675                                  0.0
1.0

        purpose_car purpose_credit_card
purpose_debt_consolidation \
280671          0.0                 1.0                        0.0

280672          0.0                 0.0                        1.0

280673          0.0                 0.0                        0.0

280674          0.0                 0.0                        0.0

280675          0.0                 0.0                        0.0


        purpose_educational purpose_home_improvement
purpose_house \
280671                  0.0                      0.0           0.0

280672                  0.0                      0.0           0.0
```

```
280673                  0.0                        0.0             0.0

280674                  0.0                        0.0             0.0

280675                  0.0                        1.0             0.0


        purpose_major_purchase  purpose_medical  purpose_moving  \
280671                     0.0              0.0             0.0
280672                     0.0              0.0             0.0
280673                     0.0              0.0             1.0
280674                     0.0              0.0             0.0
280675                     0.0              0.0             0.0

        purpose_other purpose_renewable_energy
purpose_small_business \
280671            0.0                        0.0
0.0
280672            0.0                        0.0
0.0
280673            0.0                        0.0
0.0
280674            0.0                        1.0
0.0
280675            0.0                        0.0
0.0

        purpose_vacation  purpose_wedding  application_type_DIRECT_PAY
\
280671               0.0              0.0                          0.0

280672               0.0              0.0                          0.0

280673               0.0              0.0                          0.0

280674               0.0              0.0                          0.0

280675               0.0              0.0                          0.0


        application_type_INDIVIDUAL application_type_JOINT
zip_code_00813 \
280671                          1.0                    0.0
0.0
280672                          1.0                    0.0
1.0
280673                          1.0                    0.0
0.0
280674                          1.0                    0.0
```

```
1.0
280675                          1.0                          0.0
0.0

        zip_code_05113  zip_code_11650  zip_code_22690  zip_code_29597
\
280671             0.0             0.0             0.0             0.0

280672             0.0             0.0             0.0             0.0

280673             0.0             0.0             0.0             0.0

280674             0.0             0.0             0.0             0.0

280675             0.0             1.0             0.0             0.0


        zip_code_30723  zip_code_48052  zip_code_70466  zip_code_86630
\
280671             0.0             1.0             0.0             0.0

280672             0.0             0.0             0.0             0.0

280673             0.0             1.0             0.0             0.0

280674             0.0             0.0             0.0             0.0

280675             0.0             0.0             0.0             0.0


        zip_code_93700
280671             0.0
280672             0.0
280673             0.0
280674             0.0
280675             0.0
```

Oversampling with SMOTE

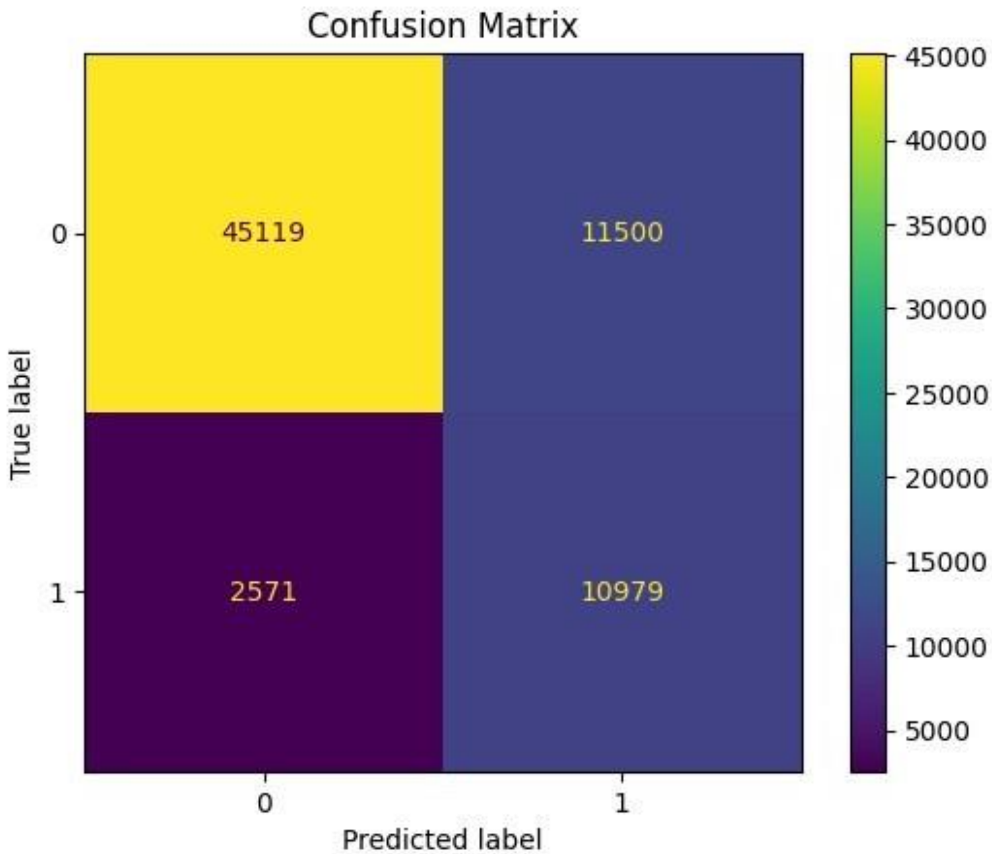
```
# Oversampling to balance the target variable

sm=SMOTE(random_state=42)
x_train_res, y_train_res = sm.fit_resample(x_train,y_train.ravel())

print(f"Before OverSampling, count of label 1: {sum(y_train == 1)}")
print(f"Before OverSampling, count of label 0: {sum(y_train == 0)}")
print(f"After OverSampling, count of label 1: {sum(y_train_res ==
1)}")
print(f"After OverSampling, count of label 0: {sum(y_train_res ==
0)}")
```

```
Before OverSampling, count of label 1: 54200
Before OverSampling, count of label 0: 226476
After OverSampling, count of label 1: 226476
After OverSampling, count of label 0: 226476
```

Logistic Regression

```
model = LogisticRegression()
model.fit(x_train_res, y_train_res)
train_preds = model.predict(x_train)
test_preds = model.predict(x_test)

#Model Evaluation
print('Train Accuracy :', model.score(x_train, y_train).round(2))
print('Train F1 Score:',f1_score(y_train,train_preds).round(2))
print('Train Recall
Score:',recall_score(y_train,train_preds).round(2))
print('Train Precision
Score:',precision_score(y_train,train_preds).round(2))

print('\nTest Accuracy :',model.score(x_test,y_test).round(2))
print('Test F1 Score:',f1_score(y_test,test_preds).round(2))
print('Test Recall Score:',recall_score(y_test,test_preds).round(2))
print('Test Precision
Score:',precision_score(y_test,test_preds).round(2))

# Confusion Matrix
cm = confusion_matrix(y_test, test_preds)
disp = ConfusionMatrixDisplay(cm)
disp.plot()
plt.title('Confusion Matrix')
plt.show()
```

```
Train Accuracy : 0.8
Train F1 Score: 0.61
Train Recall Score: 0.81
Train Precision Score: 0.49

Test Accuracy : 0.8
Test F1 Score: 0.61
Test Recall Score: 0.81
Test Precision Score: 0.49
```



Model Coefficient with column name

```
coef = model.coef_[0]
imp_features = pd.DataFrame({"Features" : cols , "Weights": coef})
imp_features = imp_features.sort_values(["Weights"] , ascending = False)
print(model.coef_)
print(imp_features)
```

```
[[ 0.67086675  0.46294261  0.52257871 -0.01778102 -1.63378648  1.06860903
   0.84908845  0.05776276 -0.58046684  0.60909815 -0.7658847  -0.3039249
  -0.05502705 -0.97401973 -0.50760769 -0.11422166  0.09561079  0.23385738
   0.27905452  0.32370805 -0.02843792 -0.20084565  0.06421717 -0.31565695
  -0.15167948 -0.03121551 -0.29875822 -0.09959339 -0.26526672 -0.43969934
   0.02089984  0.10007305  0.21091204  0.16489038 -0.19889648 -0.02792455
  -0.0150925  -0.0195308   0.11368533 -0.0129795   0.50502374 -0.2612935
  -0.80368605 -0.35193795  0.66377059 -0.97545098 -8.53933083 -8.52879241
   8.71237456 -0.28498983 -8.52156752 -0.29044261 -0.2508412  -0.28609553
   8.58983296  8.73623407]]
```

```
                              Features   Weights
55                      zip_code_93700  8.736234
48                      zip_code_11650  8.712375
54                      zip_code_86630  8.589833
5                                  dti  1.068609
6                             open_acc  0.849088
0                            loan_amnt  0.670867
44        application_type_INDIVIDUAL  0.663771
9                           revol_util  0.609098
2                             int_rate  0.522579
40              purpose_small_business  0.505024
1                                 term  0.462943
19                             grade_G  0.323708
18                             grade_F  0.279055
17                             grade_E  0.233857
32                 purpose_educational  0.210912
33            purpose_home_improvement  0.164890
38                       purpose_other  0.113685
31          purpose_debt_consolidation  0.100073
16                             grade_D  0.095611
22                 home_ownership_NONE  0.064217
7                              pub_rec  0.057763
30                 purpose_credit_card  0.020900
39            purpose_renewable_energy -0.012979
36                     purpose_medical -0.015093
3                           emp_length -0.017781
37                      purpose_moving -0.019531
35              purpose_major_purchase -0.027925
20                  home_ownership_ANY -0.028438
25                 home_ownership_RENT -0.031216
12                pub_rec_bankruptcies -0.055027
27  verification_status_Source Verified -0.099593
15                             grade_C -0.114222
24                  home_ownership_OWN -0.151679
```

Classification Report

```
print(classification_report(y_test, test_preds))
```

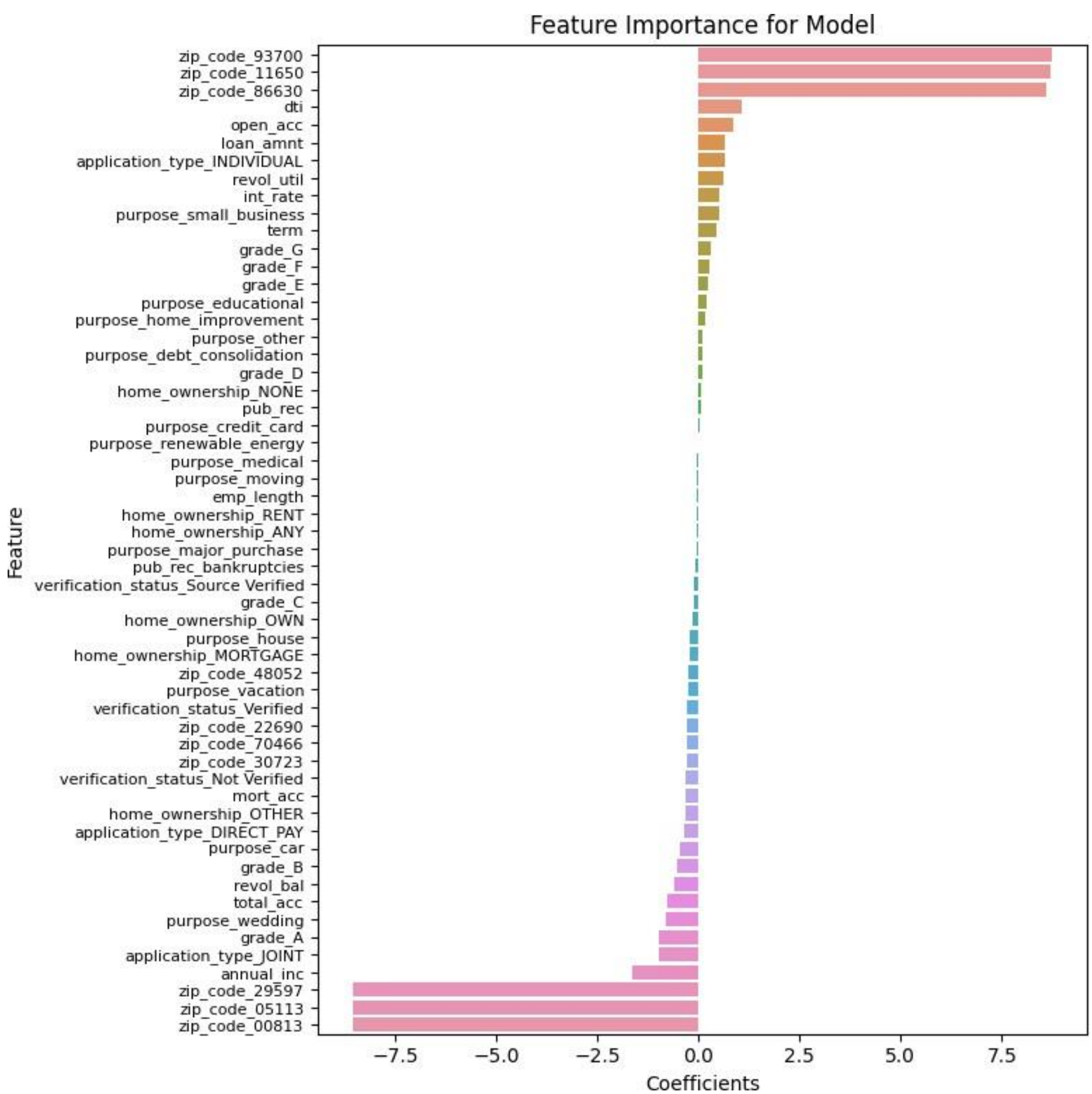
```
              precision    recall  f1-score   support

           0       0.95      0.80      0.87     56619
```

1	0.49	0.81	0.61	13550
accuracy			0.80	70169
macro avg	0.72	0.80	0.74	70169
weighted avg	0.86	0.80	0.82	70169

- It can be observed that the recall score is very high (our model is able to identify 80% of actual defaulters) but the precision is low for positive class (of all the predicted defaulters, only 50% are actually defaulters).
- Although this model is effective in reducing NPAs by flagging most of the defaulters, it may cause loantap to deny loans to many deserving customers due to low precision (false positives)
- Low precision has also caused F1 score to drop to 60% even though accuracy is 80%

Feature Importance

```
feature_imp = pd.DataFrame({'Columns':x_train.columns,
'Coefficients':model.coef_[0]}).round(2).sort_values('Coefficients',
ascending=False)

plt.figure(figsize=(8,8))
sns.barplot(y = feature_imp['Columns'],
            x = feature_imp['Coefficients'])
plt.title("Feature Importance for Model")
plt.yticks(fontsize=8)
plt.ylabel("Feature")
plt.tight_layout()
plt.show()
```



- The model has assigned large weightage to zip_code features followed by dti, open_acc, loan_amnt
- Similarly, large negative coefficients are assigned to a few zip codes, followed by annual income and joint application type

ROC Curve & AUC

The Receiver Operating Characteristic (ROC) curve is a graphical representation of the performance of a binary classification model. It helps evaluate and compare different models by illustrating the trade-off between the true positive rate (TPR) and false positive rate (FPR) at various classification thresholds.

The ROC curve is created by plotting the TPR on the y-axis against the FPR on the x-axis for different threshold values.

- TPR: Also known as sensitivity or recall, is the proportion of true positive predictions out of all actual positive instances.
- FPR: Proportion of false positive predictions out of all actual negative instances.

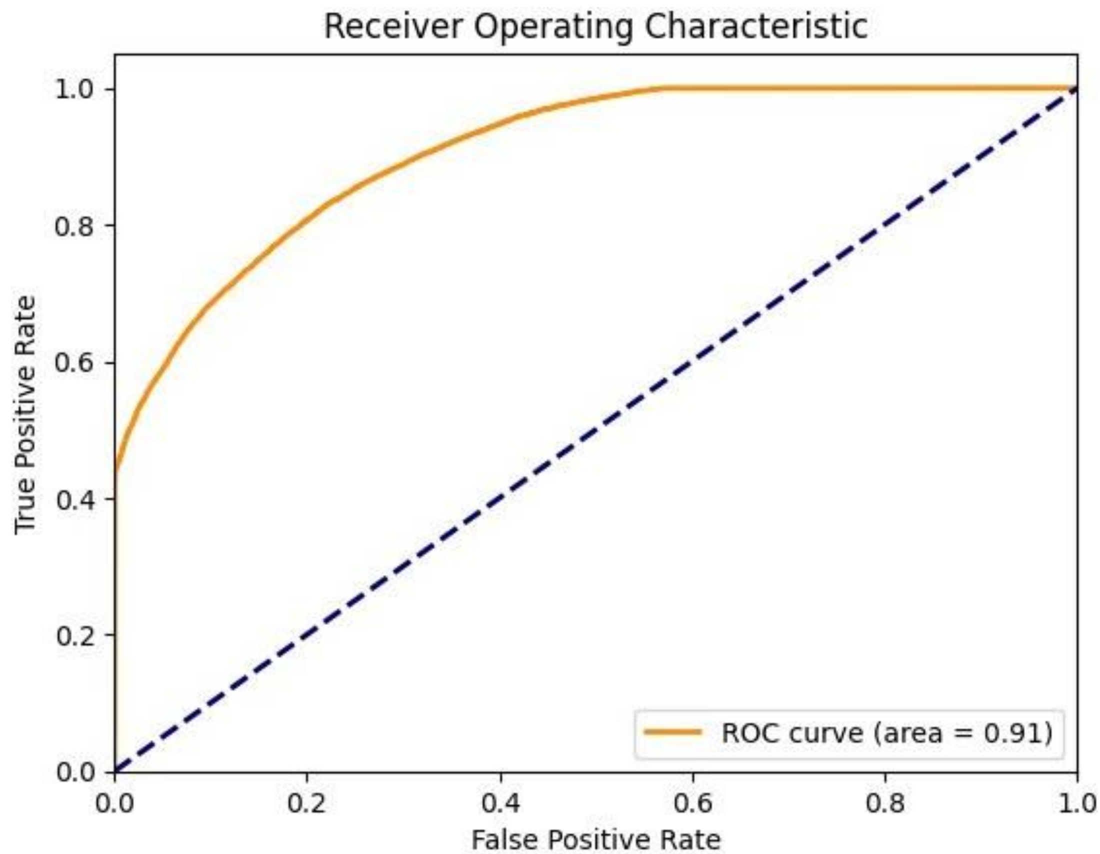
A perfect classifier would have a TPR of 1 and an FPR of 0, resulting in a point at the top-left corner of the ROC curve. On the other hand, a random classifier would have an ROC curve following the diagonal line, as it has an equal chance of producing true positive and false positive predictions.

The area under the ROC curve (AUC) is a commonly used metric to quantify the overall performance of a classifier.

A perfect classifier would have an AUC of 1, while a random classifier would have an AUC of 0.5. The higher the AUC value, the better the classifier's performance in distinguishing between positive and negative instances.

```
# Predict probabilities for the test set
probs = model.predict_proba(x_test)[:,1]

# Compute the false positive rate, true positive rate, and thresholds
fpr, tpr, thresholds = roc_curve(y_test, probs)

# Compute the area under the ROC curve
roc_auc = auc(fpr, tpr)

# Plot the ROC curve
plt.figure()
plt.plot(fpr, tpr, color='darkorange', lw=2, label='ROC curve (area =
%0.2f)' % roc_auc)
plt.plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic')
```

```
plt.legend(loc="lower right")
plt.show()
```



- AUC of 0.91 signifies that the model is able to discriminate well between the positive and the negative class.
- But it is not a good measure for an imbalanced target variable because it may be high even when the classifier has a poor score on the minority class.
- This can happen when the classifier performs well on the majority class instances, which dominate the dataset. As a result, the AUC may appear high, but the model may not effectively identify the minority class instances.

Lets plot the Precision-Recall curve which is more suited for evaluation of imbalanced data

Precision Recall Curve

The Precision-Recall (PR) curve is another graphical representation commonly used to evaluate the performance of a binary classification model. It provides insights into the trade-off between precision and recall at various classification thresholds.

- Precision represents the proportion of correctly predicted positive instances out of all instances predicted as positive. It focuses on the accuracy of positive predictions.

- Recall, also known as sensitivity or true positive rate, represents the proportion of correctly predicted positive instances out of all actual positive instances. It focuses on capturing all positive instances.

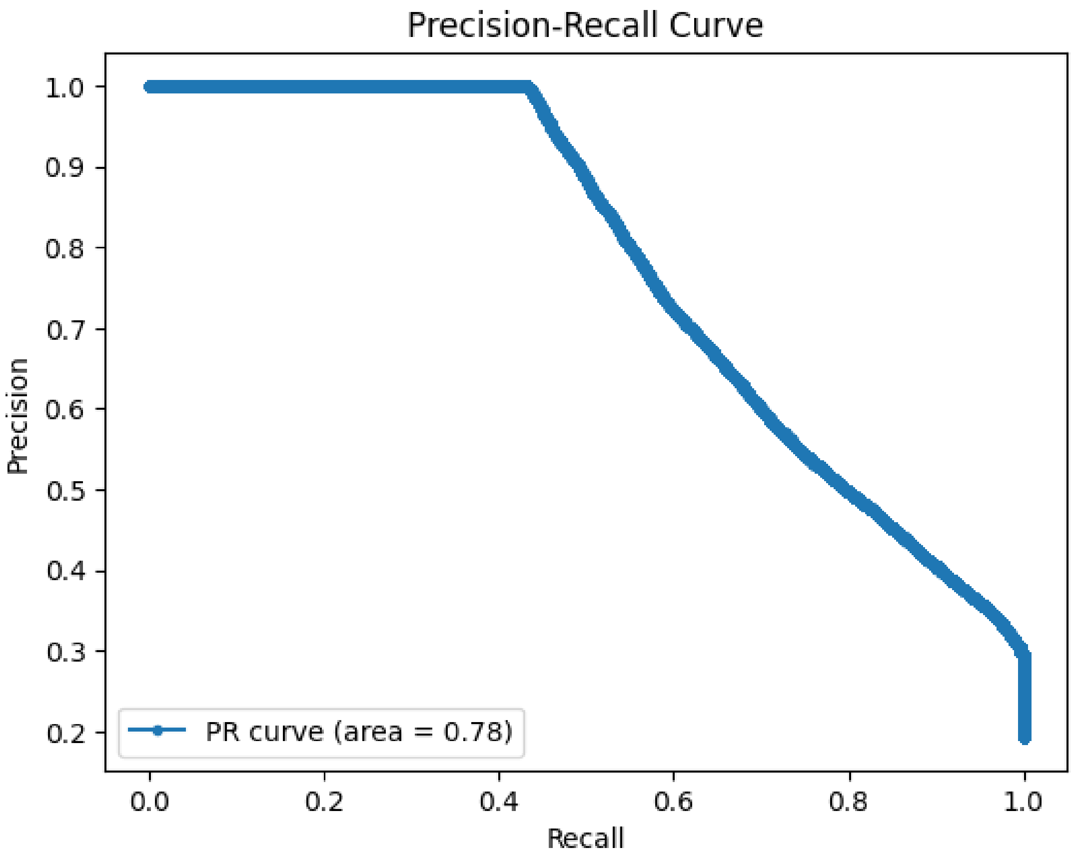
Similar to the ROC curve, the PR curve is created by plotting recall on the x-axis and precision on the y-axis for different threshold values. The curve illustrates the relationship between precision and recall as the classification threshold changes.

A perfect classifier would have a precision of 1 and a recall of 1, resulting in a point at the top-right corner of the PR curve. Conversely, a random classifier would have a PR curve following the horizontal line defined by the ratio of positive instances in the dataset.

The PR curve is useful when dealing with imbalanced datasets, where the number of negative instances far outweighs the positives. In such cases, the PR curve provides a more comprehensive evaluation of the model's performance compared to the ROC curve. This is because the ROC curve can be misleading when the majority of instances are negative, as it primarily focuses on the true negative rate.

The area under the PR curve (AUPRC) is a commonly used metric to quantify the overall performance of a classifier. A perfect classifier would have an AUPRC of 1, while a random classifier would have an AUPRC equal to the ratio of positive instances. Generally, a higher AUPRC indicates better performance.

```
# Compute the false precision and recall at all thresholds
precision, recall, thresholds = precision_recall_curve(y_test, probs)

# Area under Precision Recall Curve
auprc = average_precision_score(y_test, probs)

# Plot the precision-recall curve
plt.plot(recall, precision, marker='.', label='PR curve (area =
%0.2f)' % auprc)
plt.xlabel('Recall')
plt.ylabel('Precision')
plt.title('Precision-Recall Curve')
plt.legend(loc="lower left")
plt.show()
```



As expected, the area under precision recall curve is not as high. It is a decent model as the area is more than 0.5 (random model benchmark) but there is still scope for improvement

Tradeoff Questions

1. How can we make sure that our model can detect real defaulters and there are less false positives? This is important as we can lose out on an opportunity to finance more individuals and earn interest on it.
    - `Answer` - Since data is imbalances by making the data balance, we can try to avoid false positives. For evaluation metrics, we should be focusing on the macro average f1-score because we don't want to make false positive prediction and at the same, we want to detect the defaulters.
2. Since NPA (non-performing asset) is a real problem in this industry, it’s important we play safe and shouldn’t disburse loans to anyone
    - `Answer` - Below are the most features and their importance while making the prediction. So, these variables can help the managers to identify which are customers who are more likely to pay the loan amount fully,

Insights:

- 396030 data points, 26 features, 1 label.

- 80% belongs to the class 0: which is loan fully paid.
- 20% belongs to the class 1: which were charged off.
- Loan Amount distribution / media is slightly higher for Charged_off loanStatus.
- Probability of CHarged_off status is higher in case of 60 months term.
- Interest Rate mean and media is higher for Charged_off LoanStatus.
- Probability of Charged_off LoanStatus is higher for Loan Grades are E, F, G.
- G grade has the highest probability of having defaulter.
- Similar pattern is visible in sub_grades probability plot.
- Employement Length has overall same probability of Loan_status as fully paid and defaulter.
- That means Defaulters has no relation with their Emoployement length.
- For those borrowers who have rental home, has higher probability of defaulters.
- borrowers having their home mortgage and owns have lower probability of defaulter.
- Annual income median is lightly higher for those whos loan status is as fully paid.
- Somehow, verified income borrower's probability of defaulter is higher than those who are not verified by loan tap.
- Most of the borrowers take loans for dept-consolidation and credit card payoffs.
- the probability of defaulters is higher in the small_business owner borrowers.
- debt-to-income ratio is higher for defaulters.
- number of open credit lines in the borrower's credit file is same as for loan status as fully paid and defaulters.
- Number of derogatory public records increases, the probability of borrowers declared as defaulters also increases
- specially for those who have higher than 12 public_records.
- Total credit revolving balance is almost same for both borrowers who had fully paid loan and declared defaulter
- but Revolving line utilization rate is higher for defaulter borrowers.
- Application type Direct-Pay has higher probability of defaulter borrowers than individual and joint.
- Number of public record bankruptcies increases, higher the probability of defaulters.

- Most important features/ data for prediction, as per Logistic Regression, Decision tree classifier and Random Forest model are: Employee Title, Loan Grade and Sub-Grade, Interest rate and dept-to-income ratio.