A PROJECT REPORT ON
PREDICTION OF LOAN DEFAULT
BY

Abhay Mhatre
Abhishek Bhaste
Giridhar K S
Hemant Sharma
Sgar Sarkar

PREPARED IN PARTIAL FULLFILMENT
OF
CAPSTONE PROJECT UNDER THE SUPERVISION
OF
MR. SRIKAR MUPPIDI

DSE Capstone Project Group -1

Project Summary

Batch details	PGPDSE-FT Mumbai Jul20
Team members	Abhay Mhatre Abhishek Bhaste Giridhar K S Hemanth Sharma Sagar Sarkar
Domain of Project	Finance
Proposed Project title	Loan default prediction
Group Number	01
Team Leader	Hemant Sharma
Mentor Name	Mr. Srikar Muppidi

Date:

Signature of the Mentor

Signature of the Team Leader

Project Details

Introduction

Typically, commercial and non-commercial banks are the main lenders for loans. In contrast of those, Lending Club (LC) is a peer-to-peer online lending platform. It is world's largest marketplace connecting borrowers and investors, where consumers and small business owners lower the cost of their credit and enjoy a better experience than traditional bank lending, and investors earn attractive risk-adjusted returns.

Business Problem

- Many loans aren't completely paid off on time, however, and some borrowers default on the loan. We'll be building a model to predict whether borrowers are likely to pay or default on their loans
- Credit risk is something all peer-to-peer lending investors (and bond investors in general) must carefully consider when making informed investment decisions; it is the risk of default as a result of borrowers failing to make required payments, leading to loss of principal and interest. In this project, we build machine-learned models trained on Lending Club (a leading P2P lending platform) historical loan data.
- We are using Lending Club's data for this analysis. The data set is for the period from 2007 to 2011. There are more than 1 lakh observations and more than 100 variables. Loan status is the dependent variable for our analysis which is a categorical variable that takes only two values: Fully paid or charged off. We recoded this variable as 1 for fully paid and 0 for charged off.

Methodology

- Data collection:
  The data has been collected from https://www.kaggle.com/wordsforthewise/lending-club.

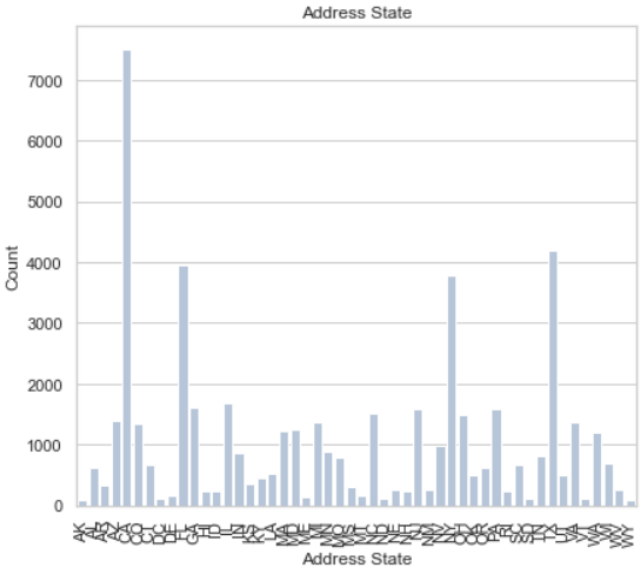
- Data Understanding:
  The dataset contains information regarding the loan defaulters. The features would be available during the application filed by the potential customer. These features are Annual Income, loan amount, employment title, grade, etc. Based on these features we would be predictingwhether that particular customer is a potential defaulter.

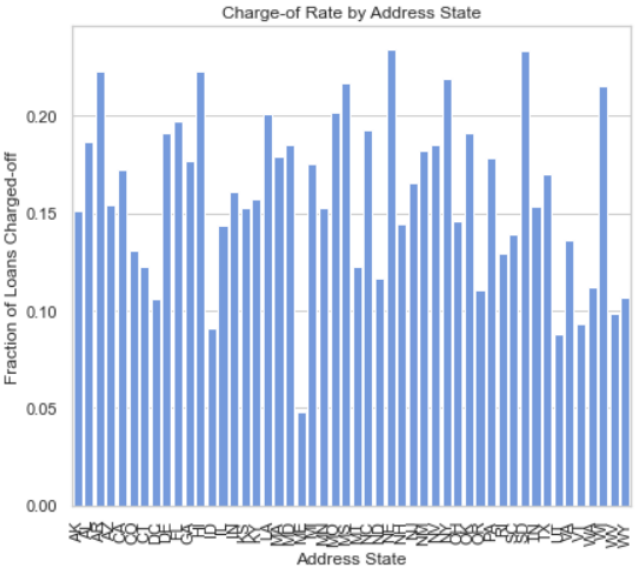
- Data Cleaning:
  The dataset consisted of 151 columns in which some features had more than 50% of missing values hence we removed them. We had 7 different types of categories in the target variable in which we have taken fully paid and charged off categories. The approach consisted of further removing the features as the EDA progressed.
  Also for the sake of the machine's capability, we decided to train data on the latest data. The fiscal year in the USA is from 1st October to 31st September. The latest date we had was December 2018. So we sampled our data for the last 3 quarters out of which we will train the data on the first 2 quarters and test our model on the last quarter.

Exploratory Data Analysis and Preprocessing

➢ Addr_state

The state provided by the borrower in the loan application

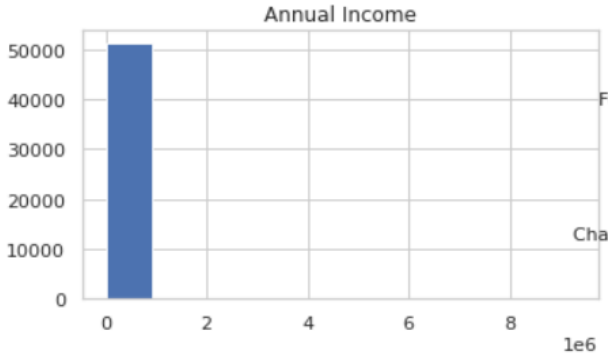


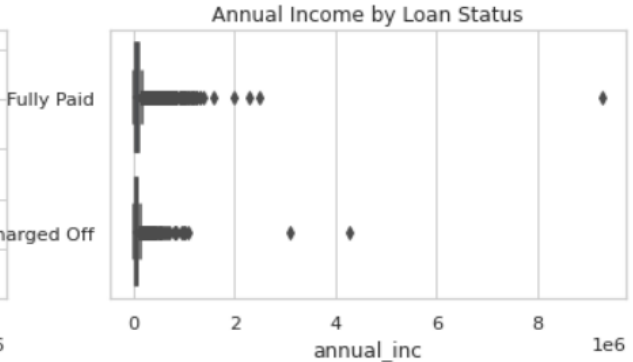


- We are having 50 unique values in address state

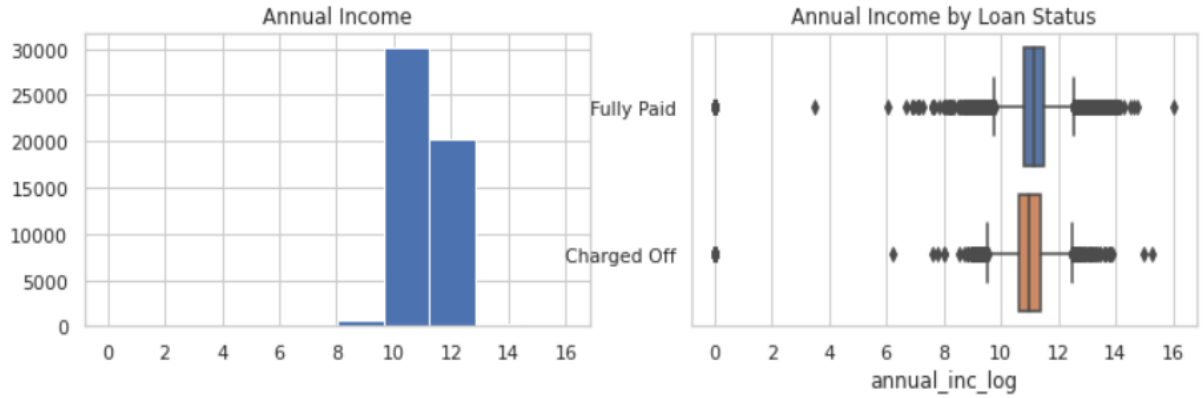
➢ Annual_inc

The self-reported annual income provided by the borrower during registration.



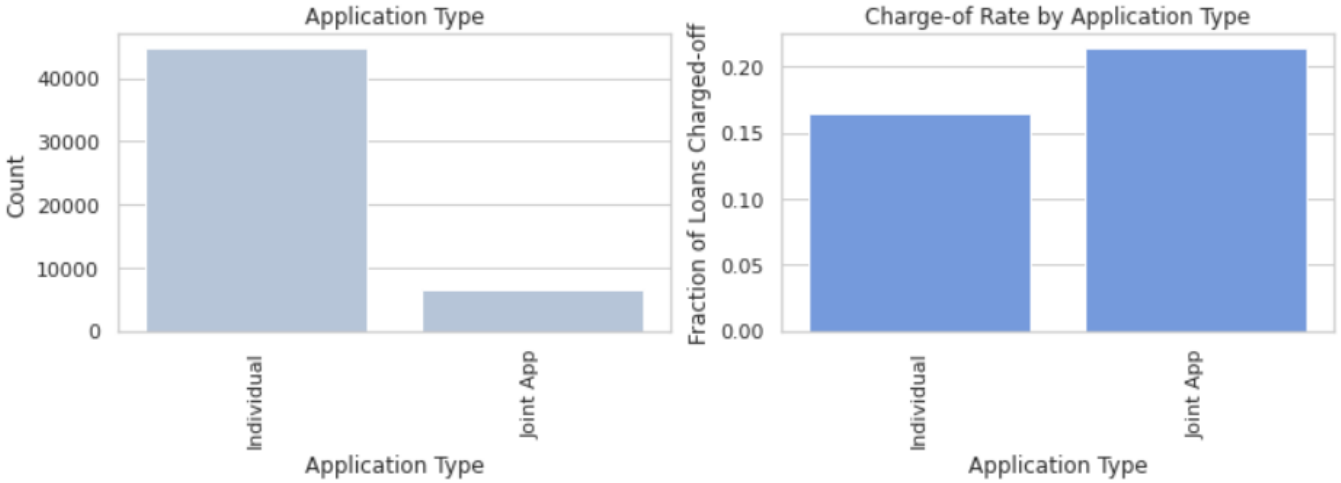


- The income range is from 0 to 9300000 with median of 66000.
- Since it is showing skewness we can take log_transformation.



➢ Application_type

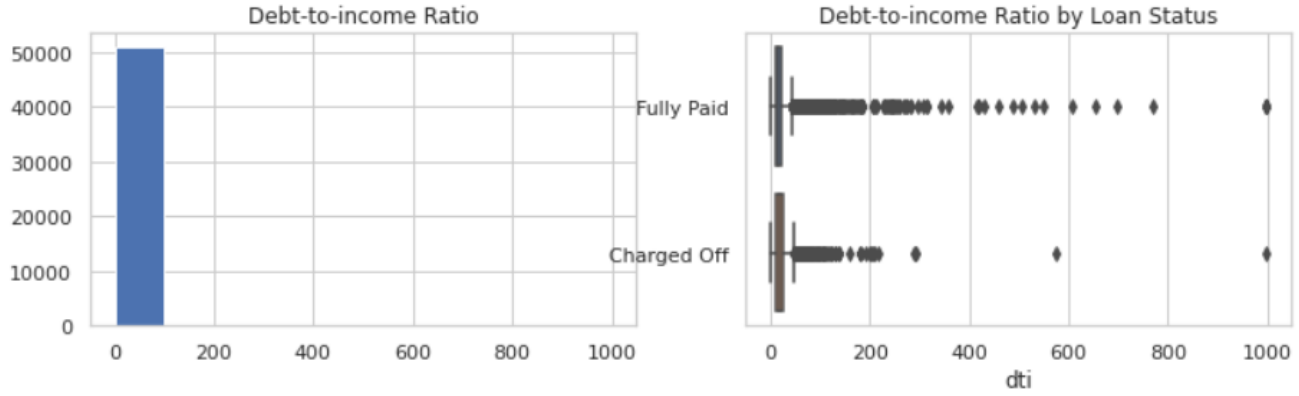
Indicates whether the loan is an individual application or a joint application with two co-borrowers



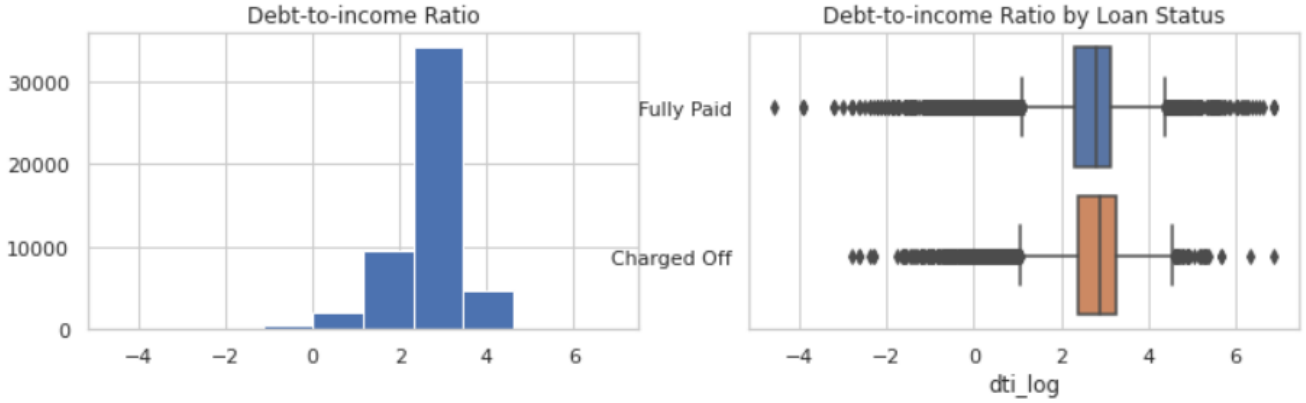
- We are having 2 types of Application type Individual and Joint. In which Joint Applicant wrt Target column is more in number.

➢ Dti

- Debt-to-income ratio.
- Dti is 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.
- This features tells us about applicant's repayment capability.
- Lower dti is better, as it means that applicant has a good monthly income.

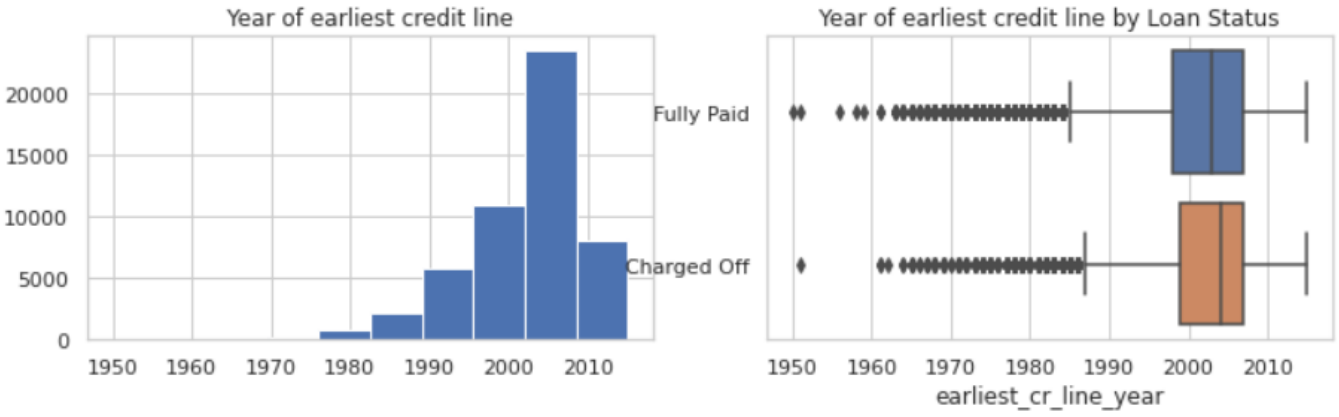


- There are some values which goes as high as 999.
- The data is skewed, we need to do log transformation



➢ Earliest_cr_line

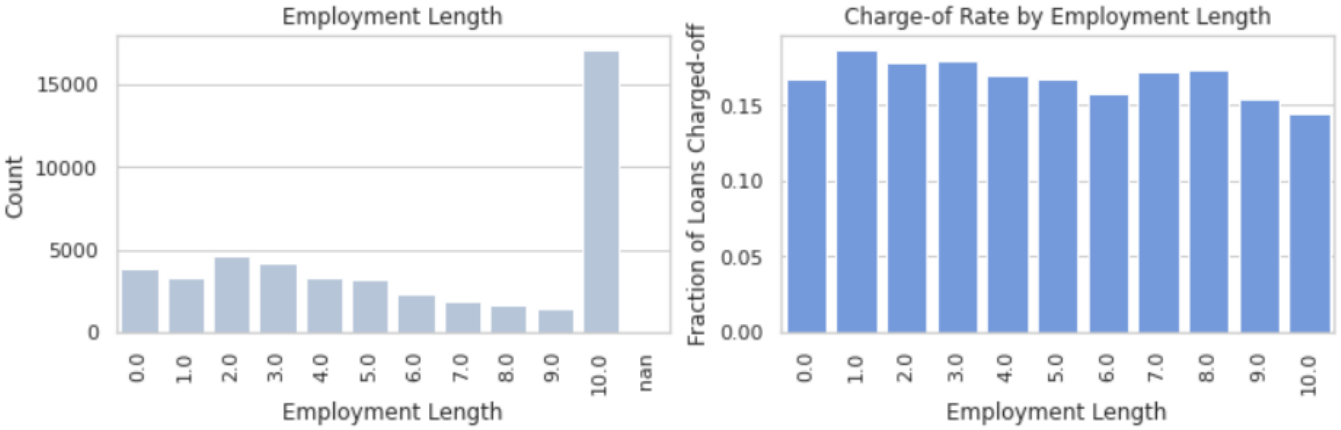
The month the borrower's earliest reported credit line was opened



- Most of applicants have earliest credit line between 2000 to 2010

➢ Emp_length

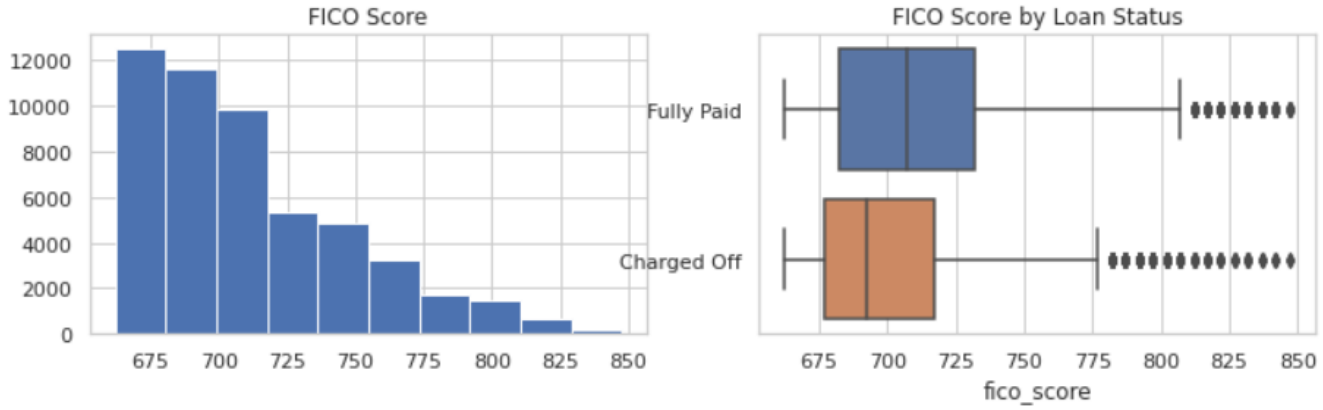
Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.



- We can see that loan status does not vary much w.r.t. employment lenght.
- Only a slight dip for employment length 10.
- So we'll not keep this column.

➢ fico_range_high & fico_range_low

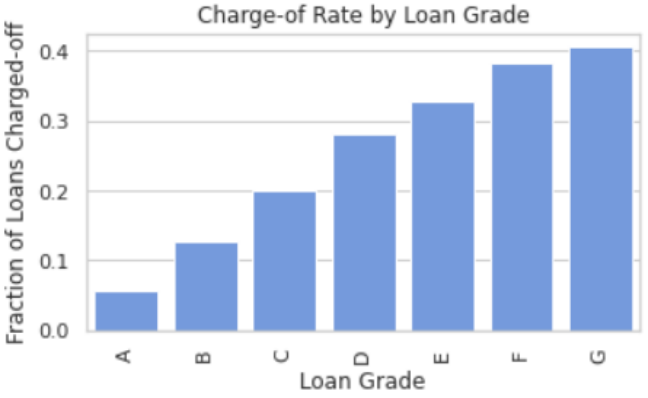
The upper boundary range the borrower's FICO at loan origination belongs to. The lower boundary range the borrower's FICO at loan origination belongs to.



- We can see that for fully paid, fico score tends to be more.

➢ Grade

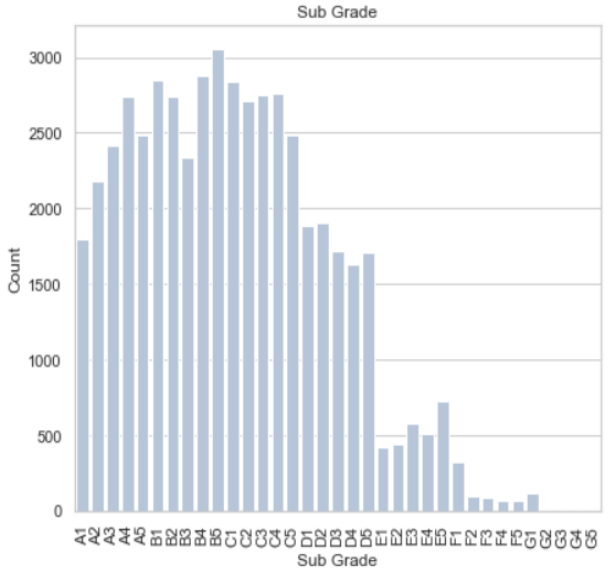
LC assigned loan grade

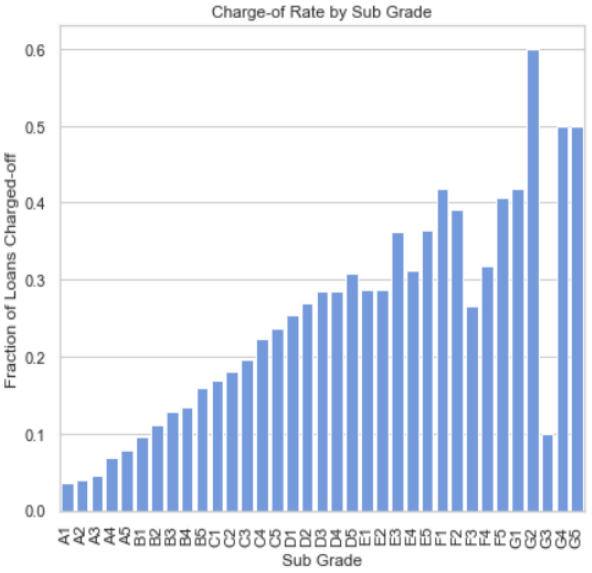


- We can see that there is huge impact of Loan grade on charged off.

➢ Sub Grade

Grade is sub divided into 5 parts

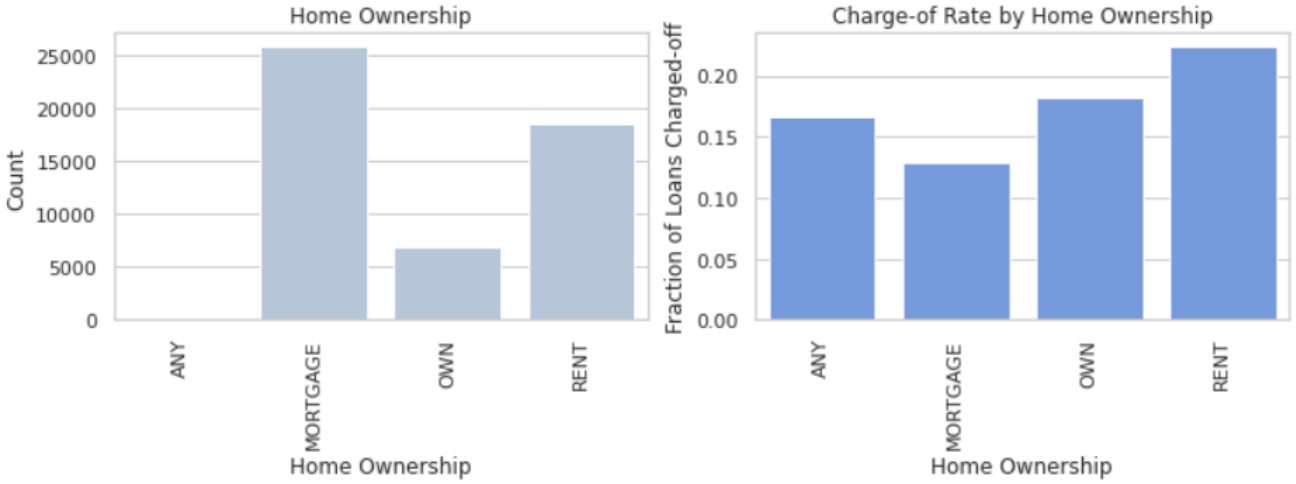




- We can see that sub grade G2 have more defaults.
- Instead of using grade, we'll use sub-grade.

➢ Home_ownership

The home ownership status provided by the borrower during registration. Our values are: RENT, OWN,MORTGAGE, OTHER.



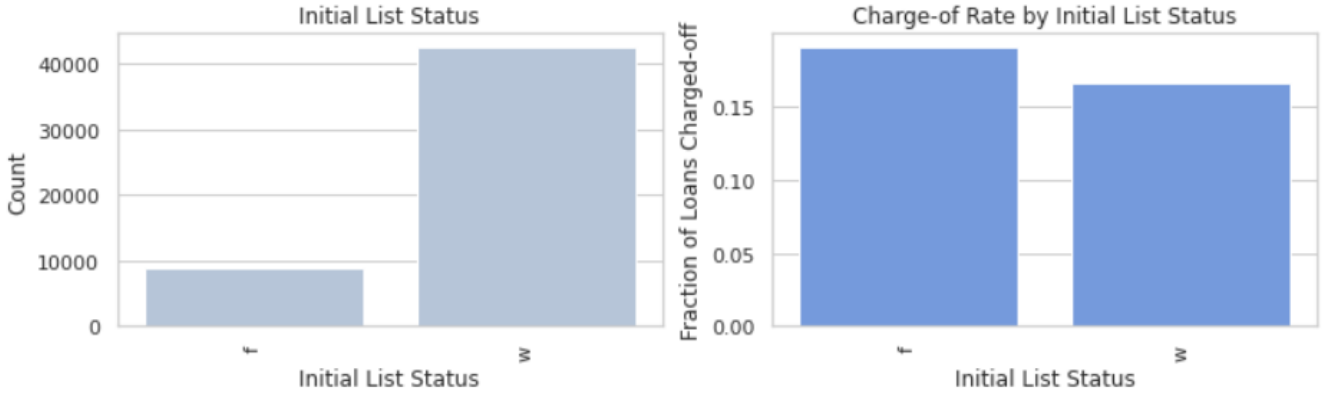
- RENT have a slightly high percentage of defaulting the loan

➢ Initial_list_status

The initial listing status of the loan. Possible values are – w, f.

w: whole - means whole loan is sponsored by a particular investor

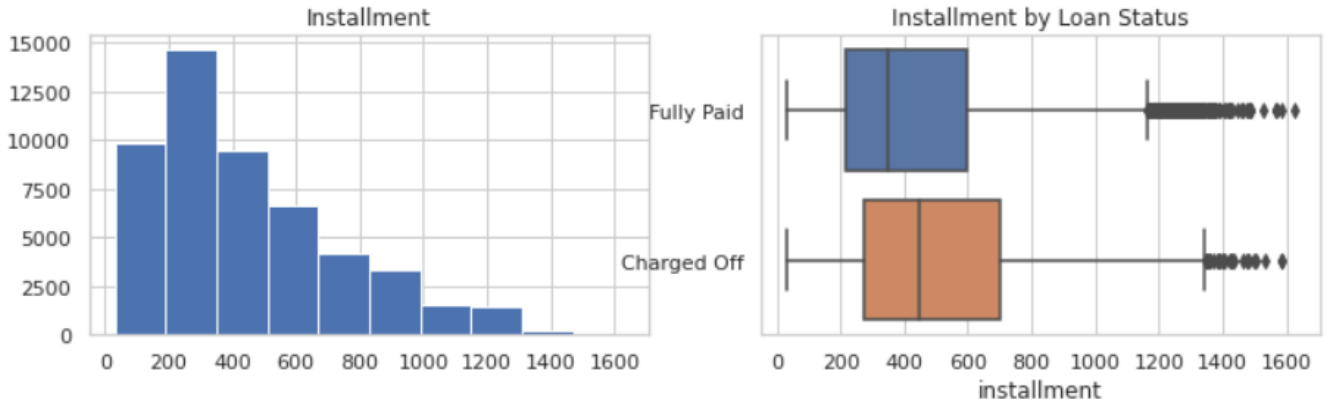
f: fraction - the loan amount consists of the amount taken from pool of investors, i.e. a small fraction from each investor



- We can see that the loan given in fraction has a slightly more chances of being default, but it's not much

➢ Installment

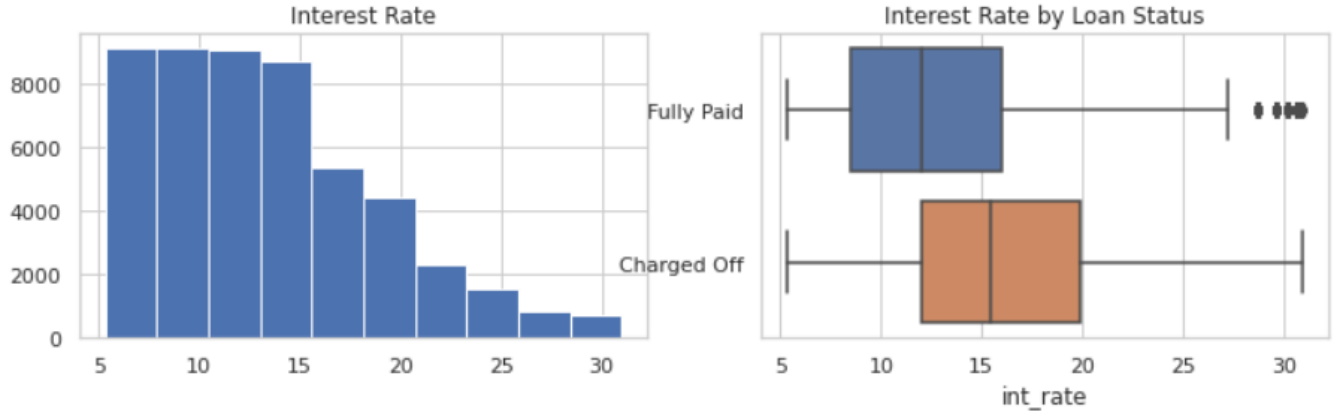
Fixed monthly installment to payback the loan.



- The minimum installment amount is 30 and it goes as high as 1628.
- Low installment amount have less chance of default.

➢ Int_rate

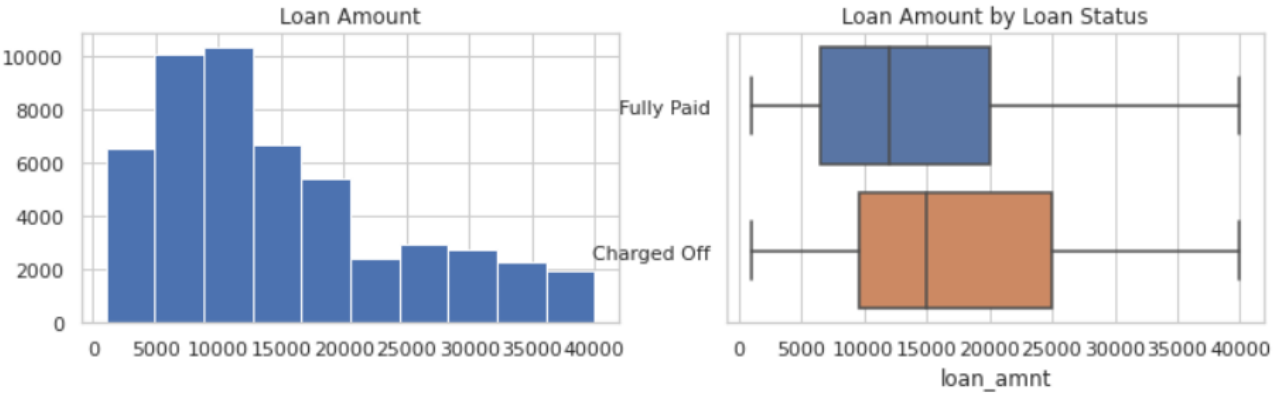
Interest rate charged on the loan amount.



- For most of the people the interest rate is from 5 to 15%, but it goest as high as 30 for some people.

➢ Loan_amnt

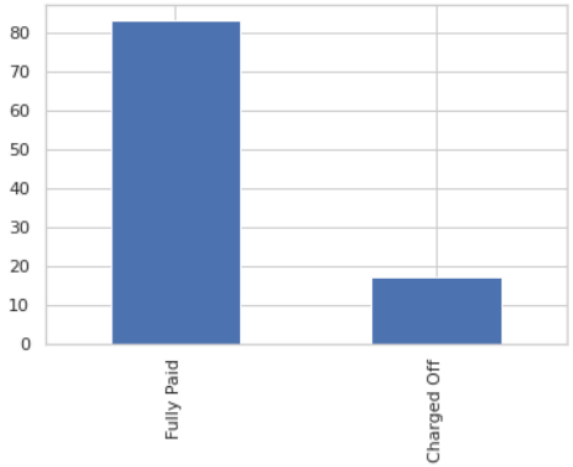
The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.



- Most of the peolpe have applies for loan between 5000 and 10000, with 12000 as median.
- It goes as high as 40000.

➢ Loan_status(Target Column)

Current status of the loan. 2 sub categories Fully Paid and Charged off



- We can see there is huge imbalance in the data.
- We may have to use some oversampling technique in order to level this imbalance.

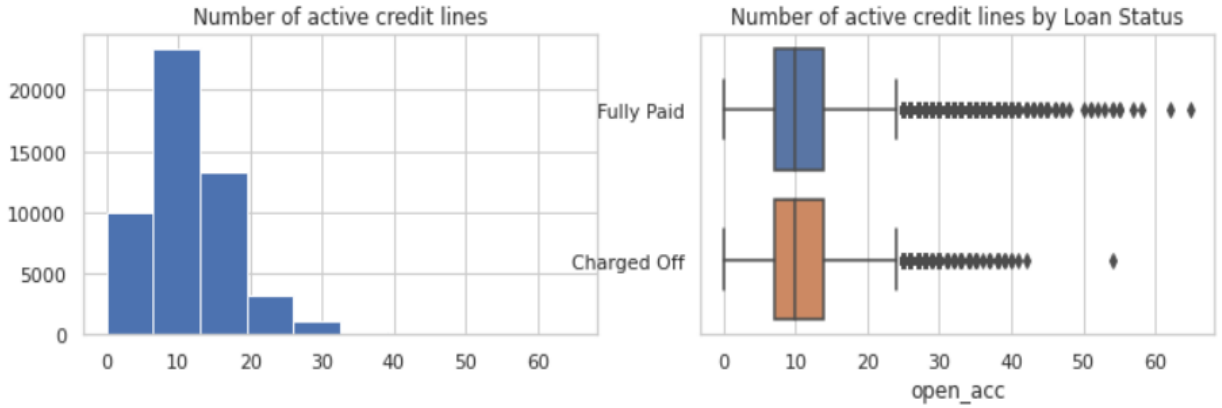
➢ Mort_acc

Number of mortgage account applicant have.This tells us about applicant's credit history, as you usually get mortgage loan only when you were good paying back previous loans.

- More than 25% of applicant never had a mortgage account.
- Again we can see that people who had a mortgage account tends not to default on the loan.

➢ Open_acc

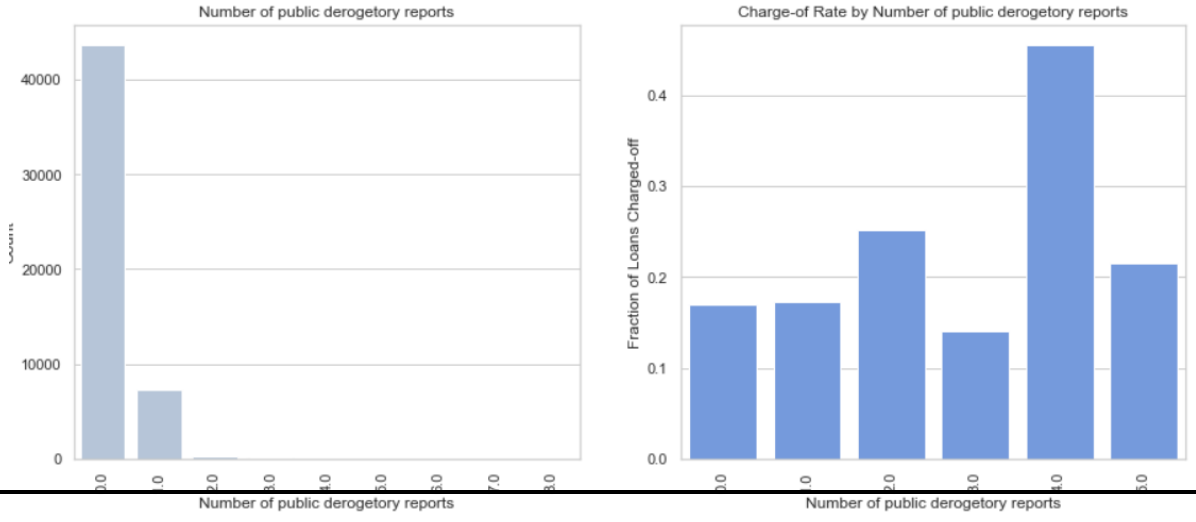
Number of active credit lines on the applicants account.



- As we can see most of the applicants have about 10 active credit lines.
- Number of credit lines can tell us how good the applicant is paying back the credit amount.
- We can see the defaulters have less number of active credit lines.

➢ Pub_rec

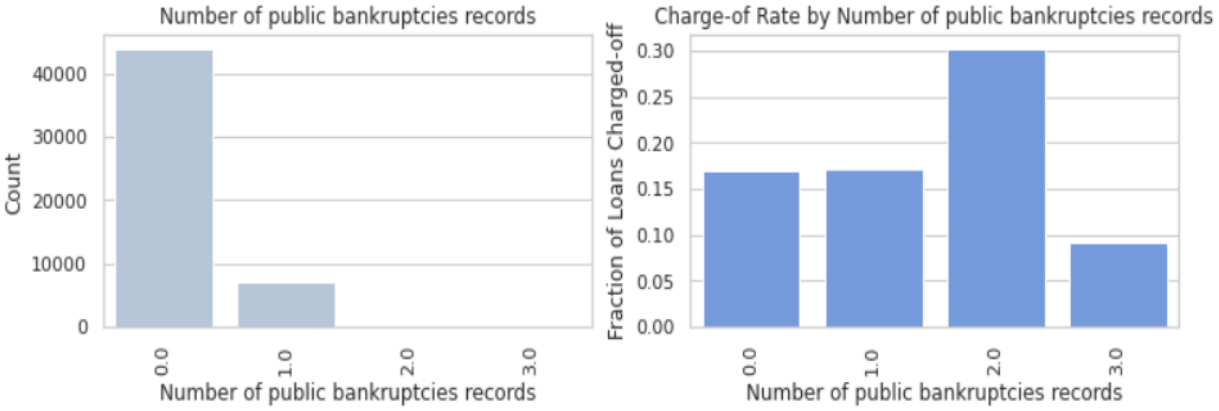
Number of public derogetories.



- Most of the people don't have any public derogetory report.
- After applying group by with loan status we saw applicants having 4 public derogetory reports are having about 45% defaulters.

➢ Pub_rec_bankruptcies

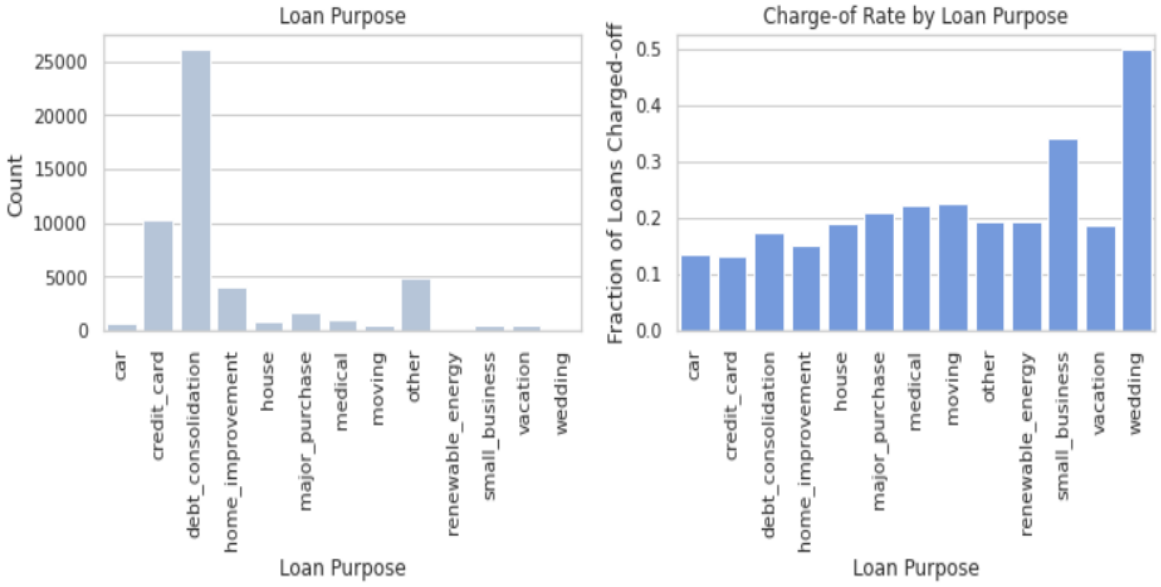
- Number of public bankruptcies records.



- We can see that most of the peopl have 0 public bankruptices records.
- Applicants with 2 bankruptcies are having more percentage of defaults.

➢ Purpose

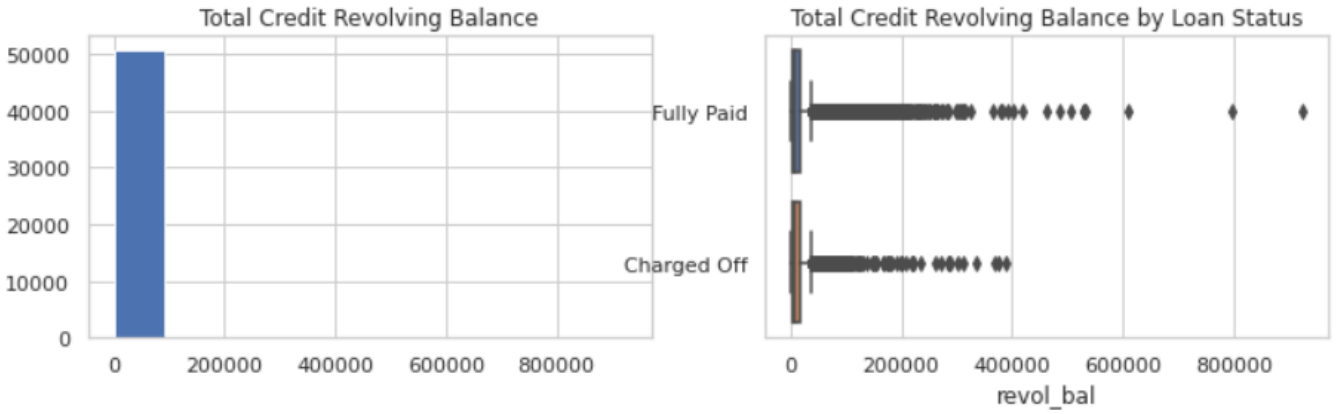
Purpose of the loan



- We can see that small business and weddings seems to have more loan default.
- Default in small business can be explained by the fact that some small business are getting hit by huge dominance of e-commerce services, as these services have entered into practically every segment.

➢ Revol_bal

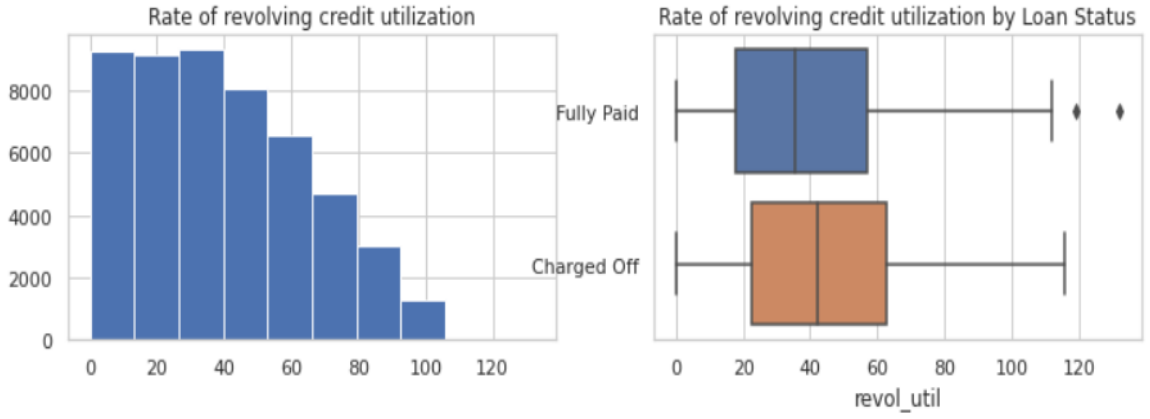
This is total credit revolving balance.After utilizing the available credit balance how much balance is left.



- Those who have defaluted have less credit revolving balance.
- This feature have very high variance, and also most of the people are having 0 credit revolving balance.

➢ Revol_util

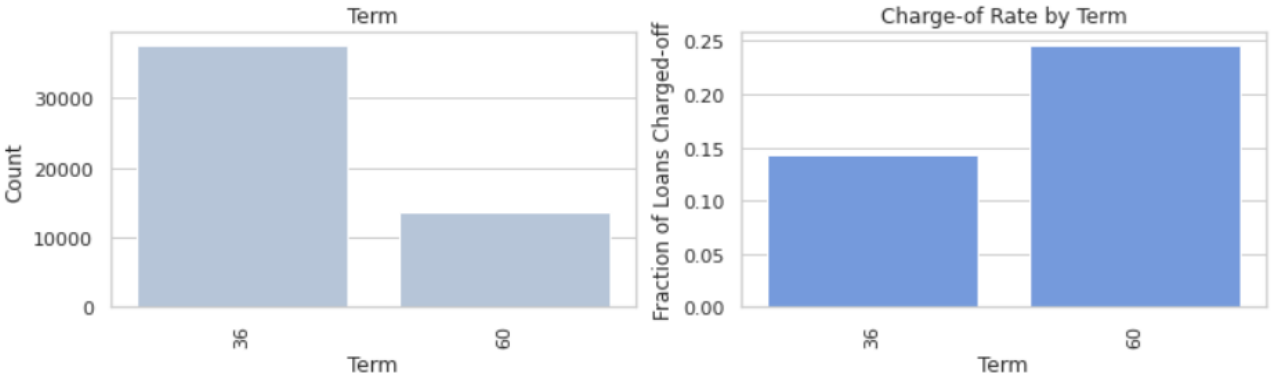
This tells us the rate of credit utilization by the applicant



- Mostly people are having 40% Revolving Credit utilization and there are some outliers in case of Fully paid customers.

➢ Term

This is period in months, loan has to repay in these period only.



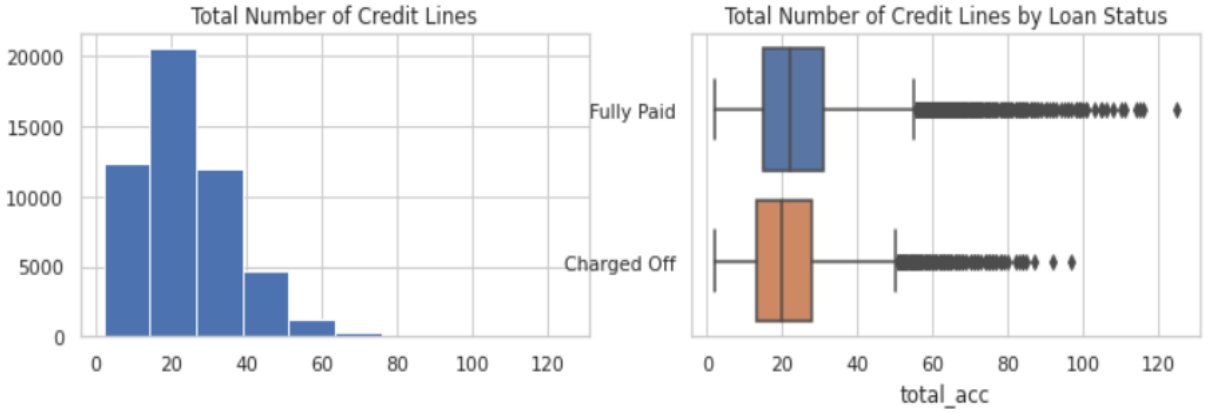
- We can see that applicants having 60 months as loan term have more percentage of default.

➢ Title

- This feature is same as purpose.
- We're dropping this feature.

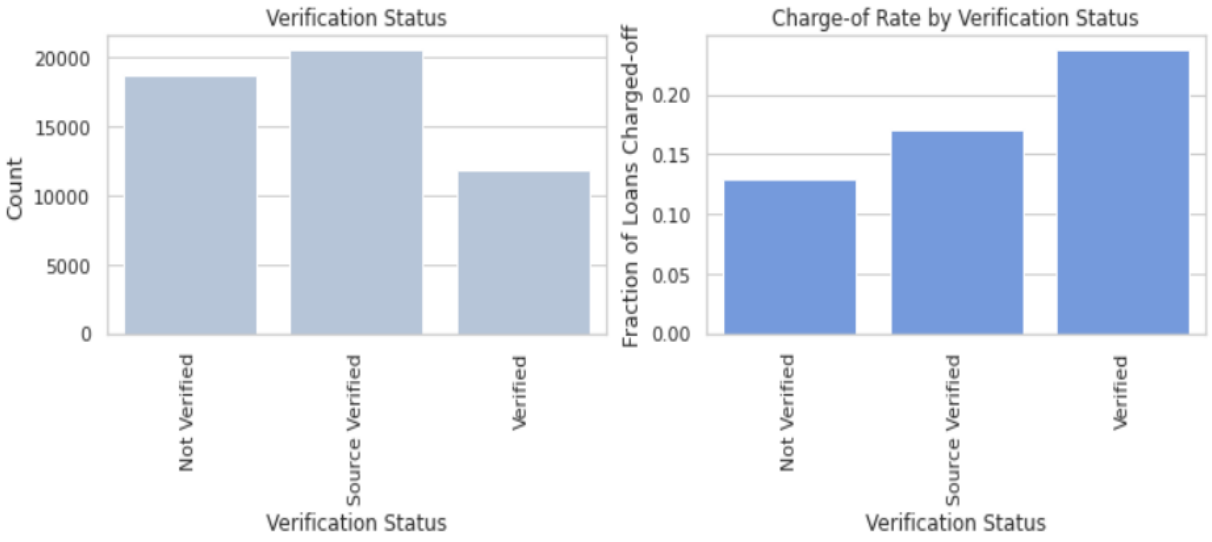
➢ total_acc:

This is the total number of credit lines till date applicant ever had.



- Defaulters have less number of total number of credit lines.
- This can mean that they didn't get more credit line because they were not consistent in repaying the loan.

➢ verification_status



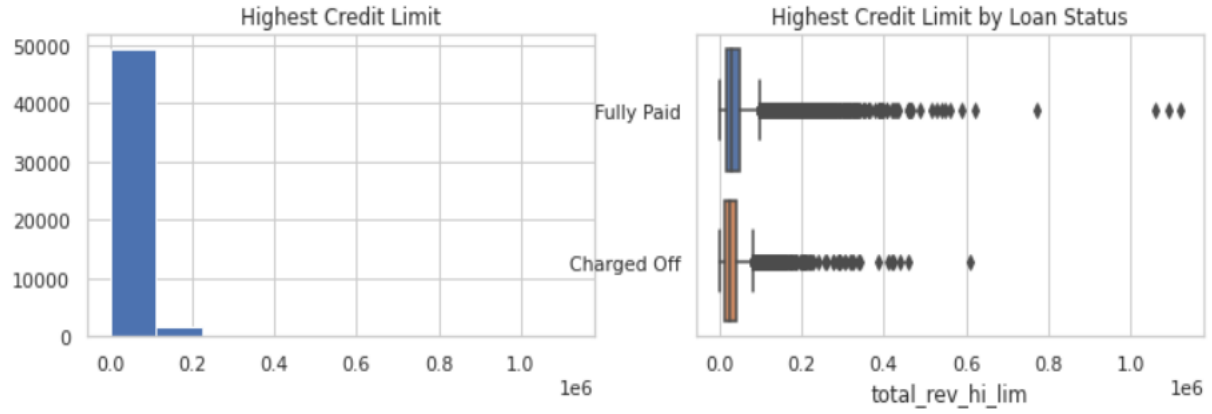
- It is interesting, that verified income have more defaulters.

➢ zip_code

There are 864 unique values, dropping this feature

➢ Total_rev_hi_lim

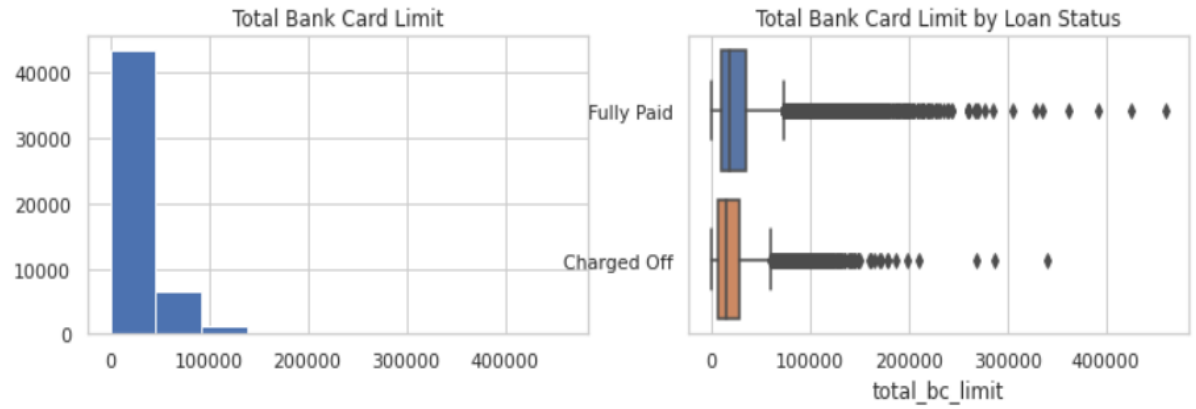
Revolving account is essentially a loan account having varying credit limit.Similar to credit card.This feature tells about the highest credit limit applicant has/had.



- Most of the people have 0 credit limit, which is true as most of applicant don't have any revolving account either.
- Again defaulters tends to have lower credit limit as compared to who didn't defaulted the loan.

➢ Total_bc_limit

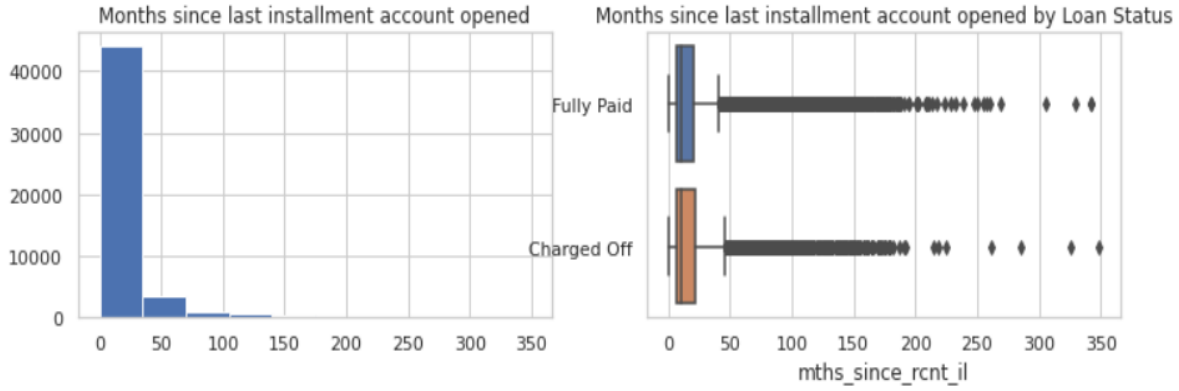
This is total limit of credit card.



- Charged Off have lower bank card limit.

➢ Mths_since_rcnt_il

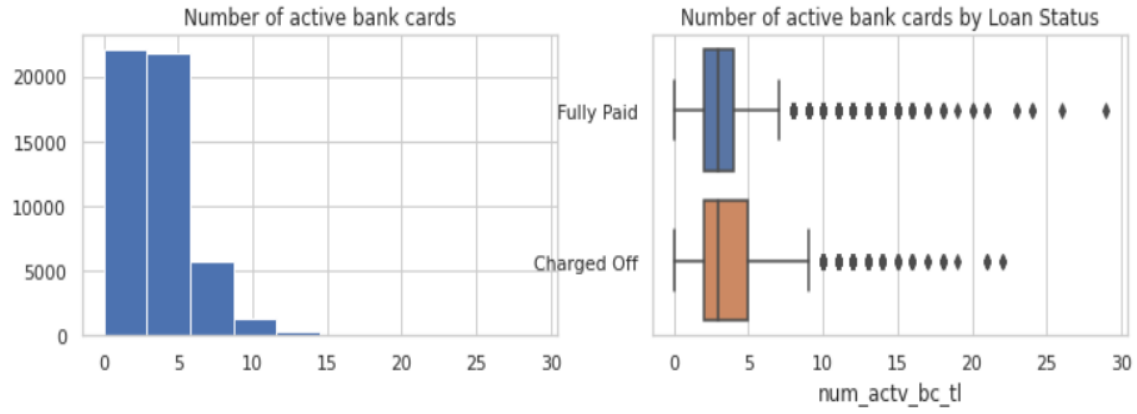
This is months since last installment account opened



- Most of the applicant never had any installment account.
- Fully paid tend to have more number of installment account.

➢ num_actv_bc_tl

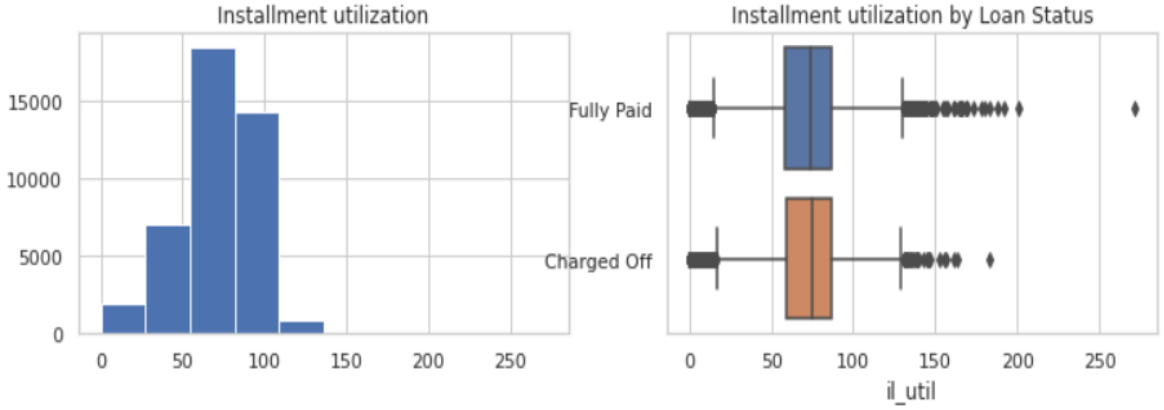
Number of total active bank cards on applicant's account as of now.



- Most applicants have active bank cards between 0 to 4.
- Those who have charged off tend to have less number of active bank cards

➢ il_util

It is the ratio of current balance to highest credit limit on all installment accounts.This ratio, if bigger means applicant have more balance and less credit is utilized.



- Installment utilization by Loan Status is having same mean but outliers are more in case of Fully Paid.

BASE MODEL:

Final Dataframe operations:

- Column **mths_since_rcnt_il**, **revol_util**, **il_util**, **dti_log** has some missing values in it.
- We imputed this missing values with **KNN IMPUTER**

TRAIN TEST SPLIT :

Before applying various classification techniques to predict the results, let us split the dataset in train and test set.

- We have divided splits into two parts as , first two quarter as training set and next quarter as Test set

Train shape : 40973 rows and 95 columns
Test shape : 10308 rows and 95 columns

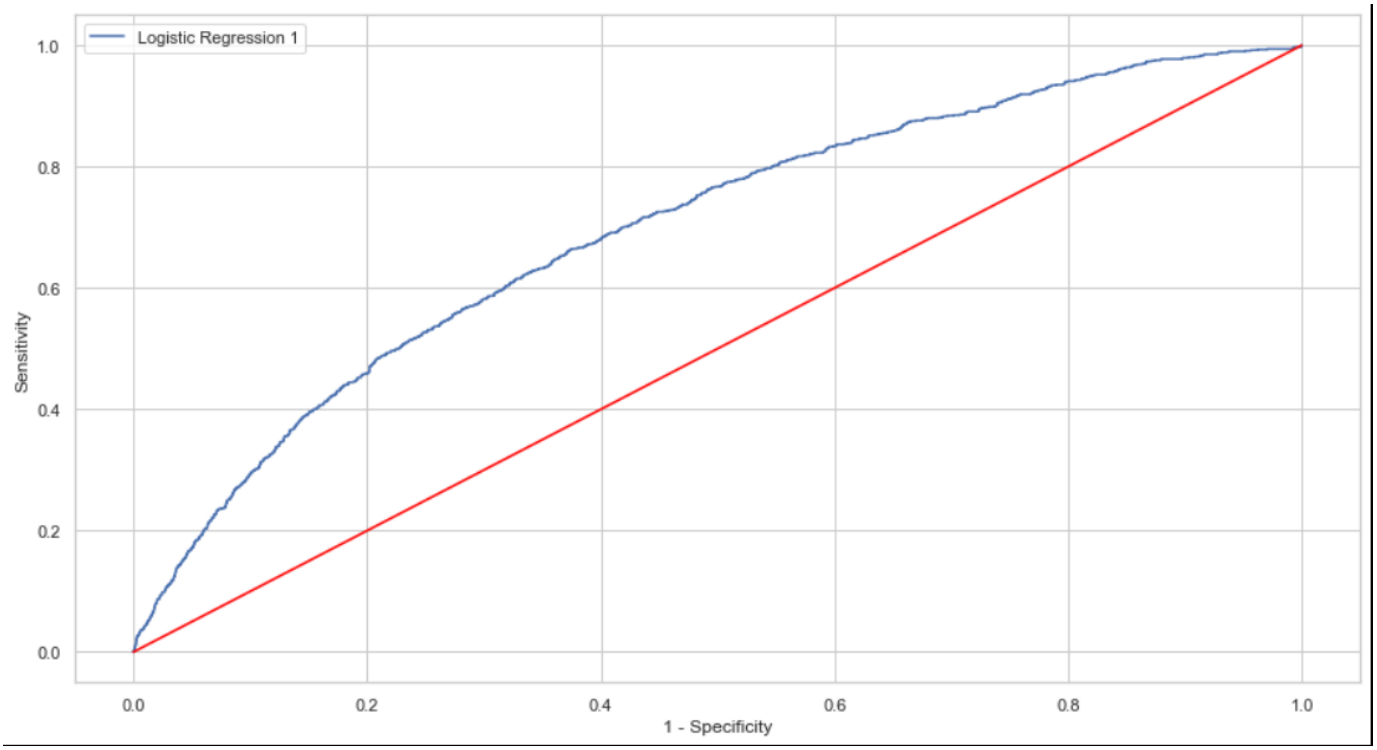
Base model building :

- We have used logistic regression for our base model where we got 83% train accuracy and 89% test accuracy.

```
Accuracy =  0.8961001164144354
Recall =  0.06096361848574238
Precision =  0.34831460674157305
F1_score =  0.10376569037656905
```

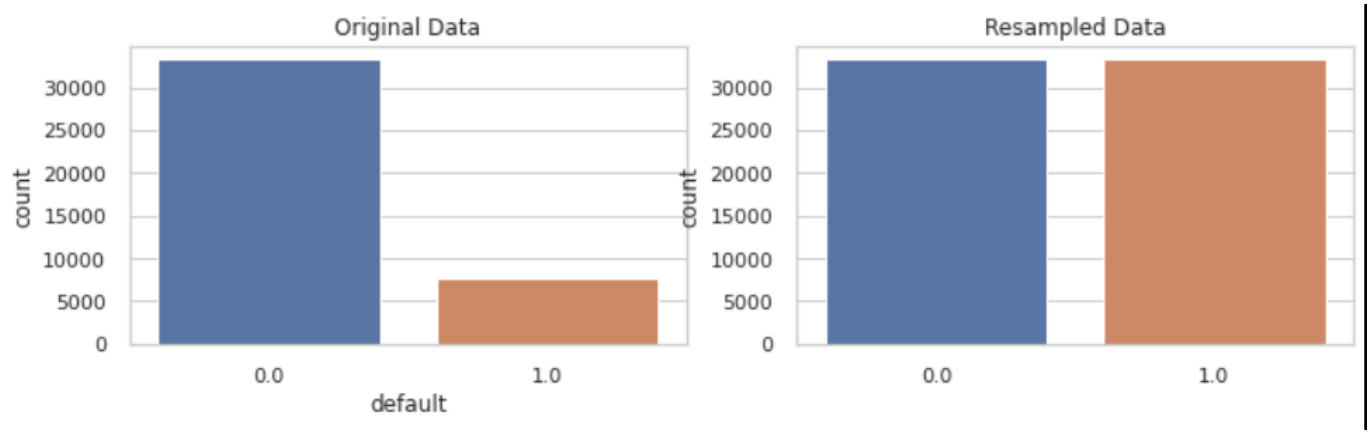
We can see that, Even though getting good accuracy our F1 score is not that good.

ROC CURVE :



The red line represents the ROC curve of a purely random classifier; a good classifier stays as far away from that line as possible
From above plot ,we can see that our classifier (Logistic Regression) is not to far from red line. This means we have to improve our model for better performance. We will use SMOTE Technique for improvement.

IMPROVE MODEL PERFORMANCE:



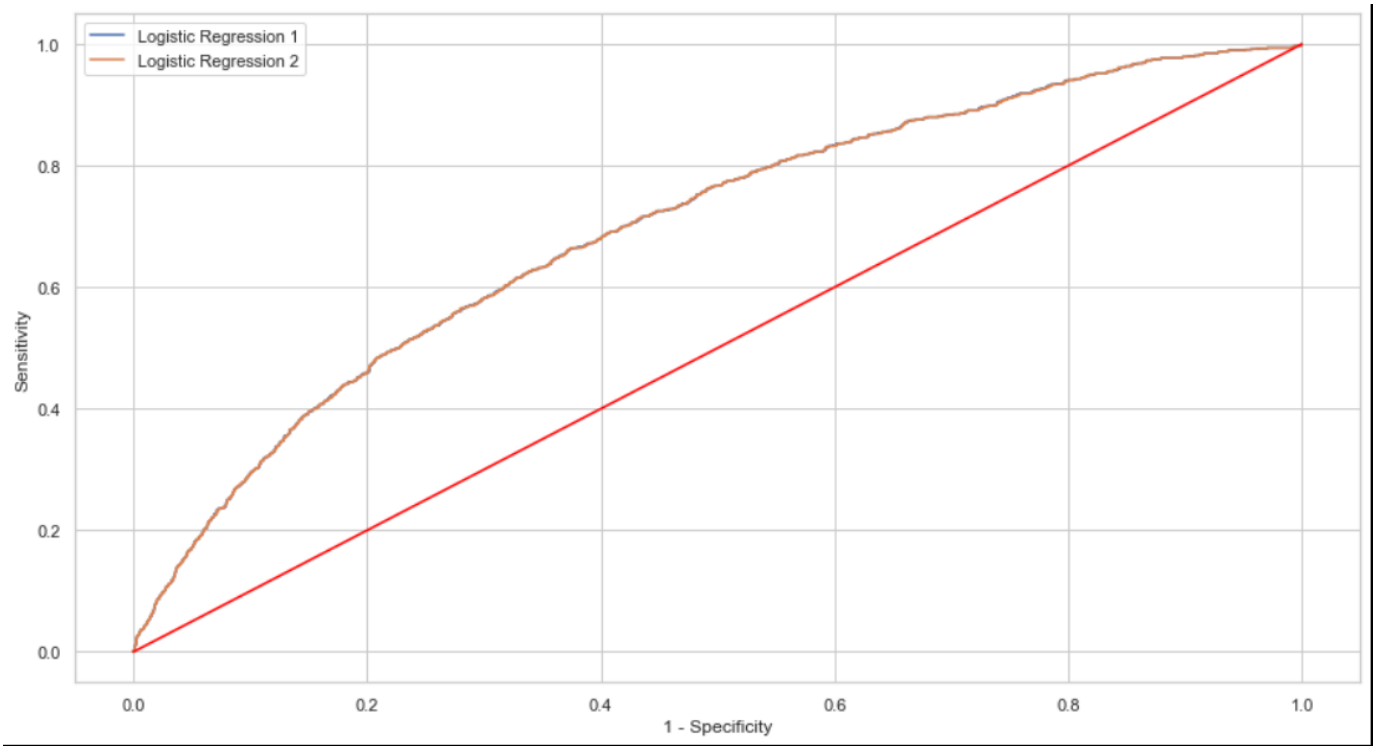
We can see from above plots, we have imbalance data in left plot and balanced data in right plot. We have used SMOTE technique to improve this defect.

SMOTE (synthetic minority oversampling **technique**) is one of the most commonly used oversampling methods to solve the imbalance problem. It aims to balance class distribution by randomly increasing minority class examples by replicating them. **SMOTE** synthesises new minority instances between existing minority instances

RESULTS AFTER APPLYING SMOTE TECHNIQUE:

```
# After SMOTE
# Accuracy =  0.6753977493209158
# Recall =  0.6312684365781711
# Precision =  0.1776916689731525
# F1_score =  0.27732181425485963
```

After applying SMOTE technique our accuracy has reduced but we have really good improvement in Recall and F1 score as compare to previous scores.

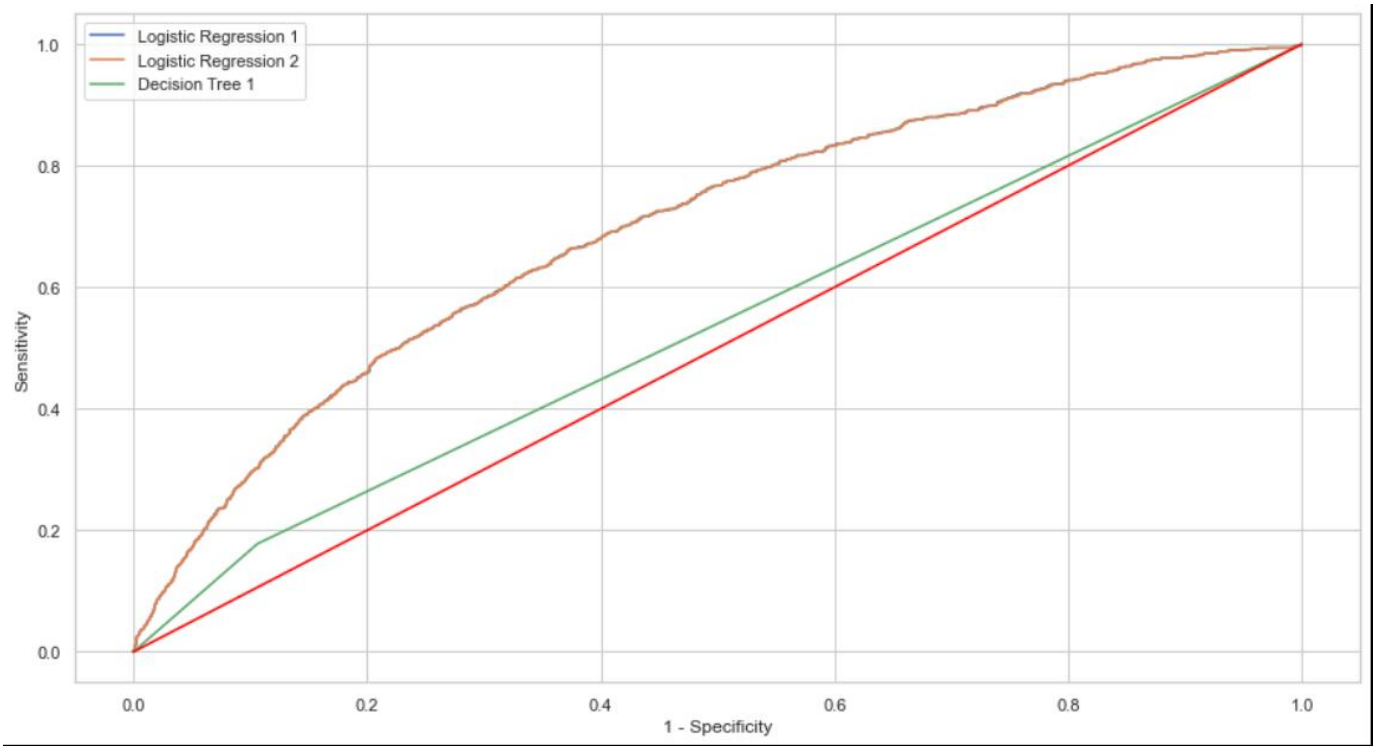


We can observe from above graph that our classifier2 (Logistic Regression 2) is Overlaaping on classifier1 and both the curves are far away from red line in some amount.

Decision Tree for performance improvment:

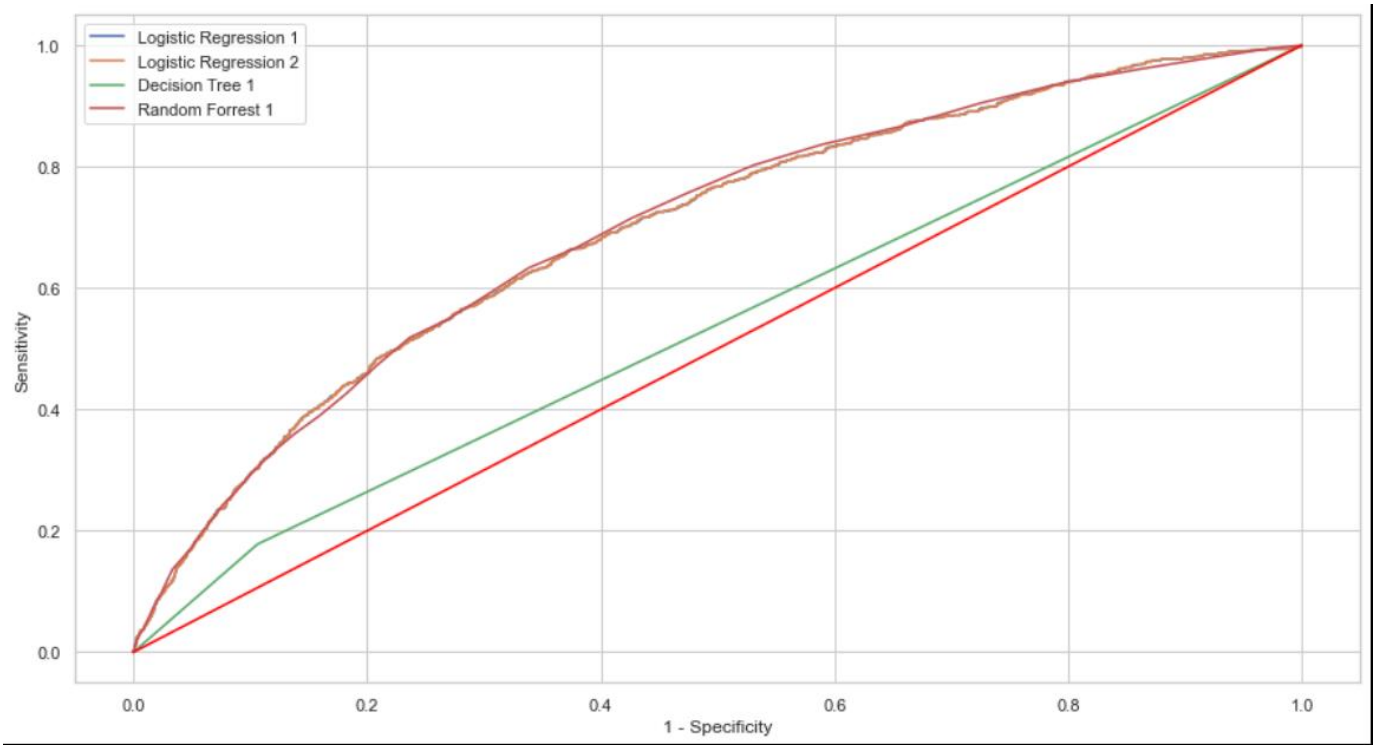
```
Accuracy =  1.0
Recall =  1.0
Precision =  1.0
F1_score =  1.0
```

The results we have got after appling Decision Tree seems that, the mode is overfitting.



The ROC curve which we got after applying Decision Tree, is closer to red line, which is not a ideal condition for perfect model.
To counter this ,will go for another ML algorithem implication , which is RANDOM FOREST.

Using Random Forrest:



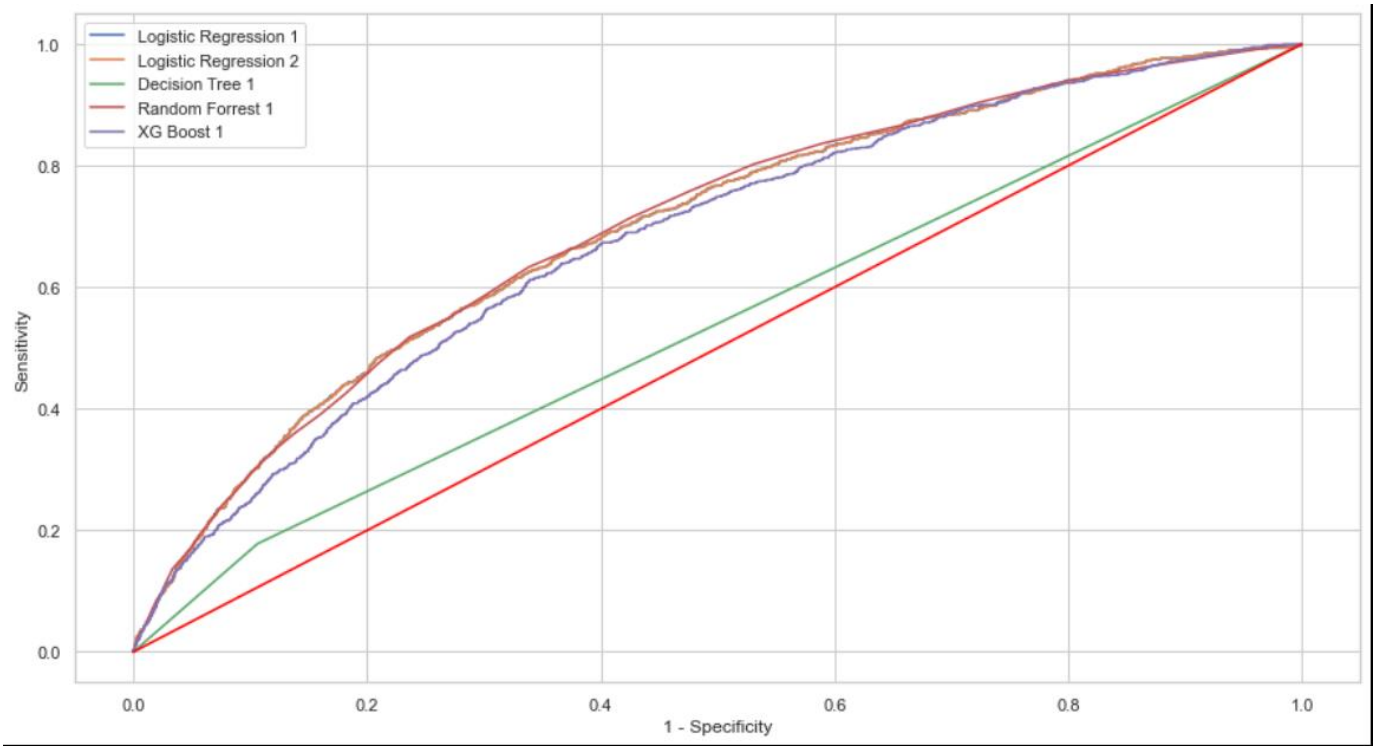
After using random forest ,the results were little better as compared to Decision Tree but still they are on the line of overfitting which is not a good thing. So we tried XGBOOST next.

XGBOOST:

```
Accuracy =  0.8775708187815289
Recall =  0.16027531956735497
Precision =  0.28546409807355516
F1_score =  0.2052896725440806
```

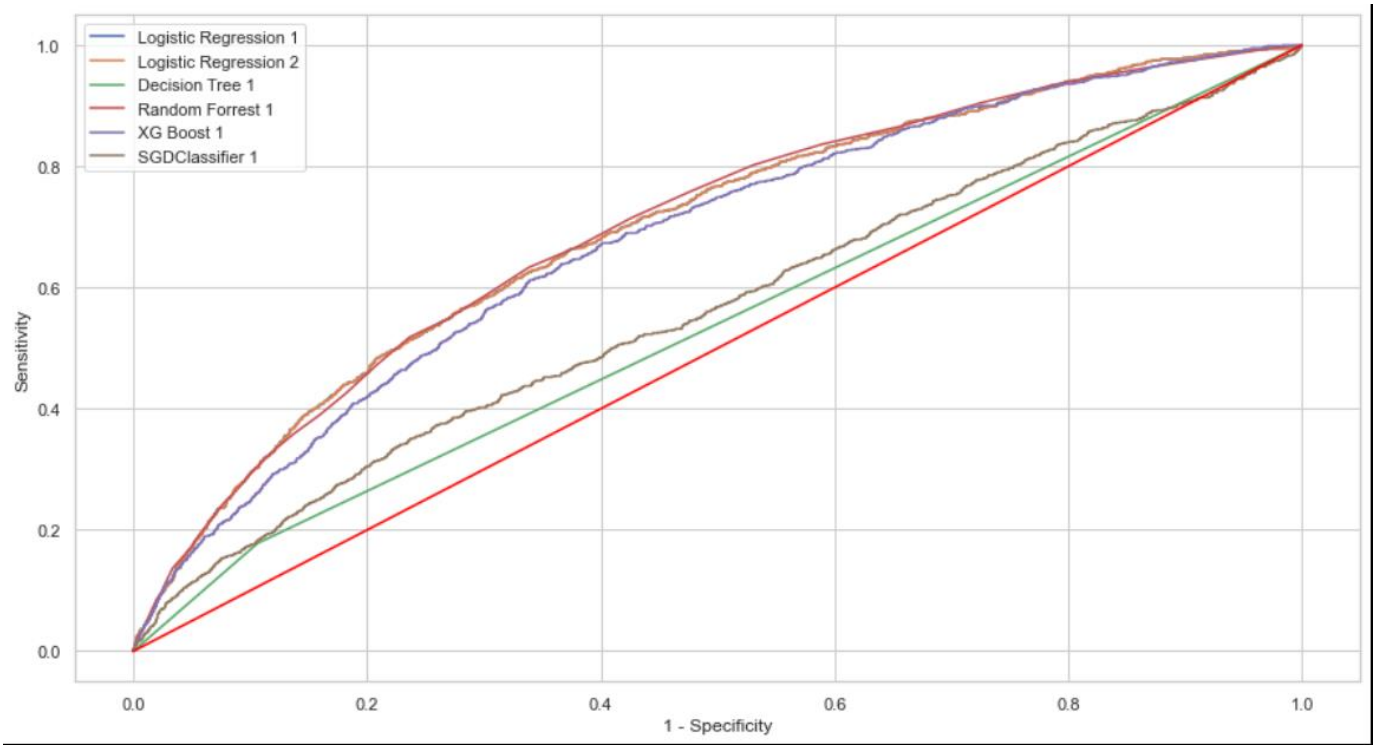
We can see that, XGBOOST classfier is doing better as compared to every other classifiers from previous operations.

In below graph we can see that, Classifier (XGBOOST 1) is far from red line in better way as compared to prvious classifiers.



We still thrive for further operations to get better results.

SGDClassifier:

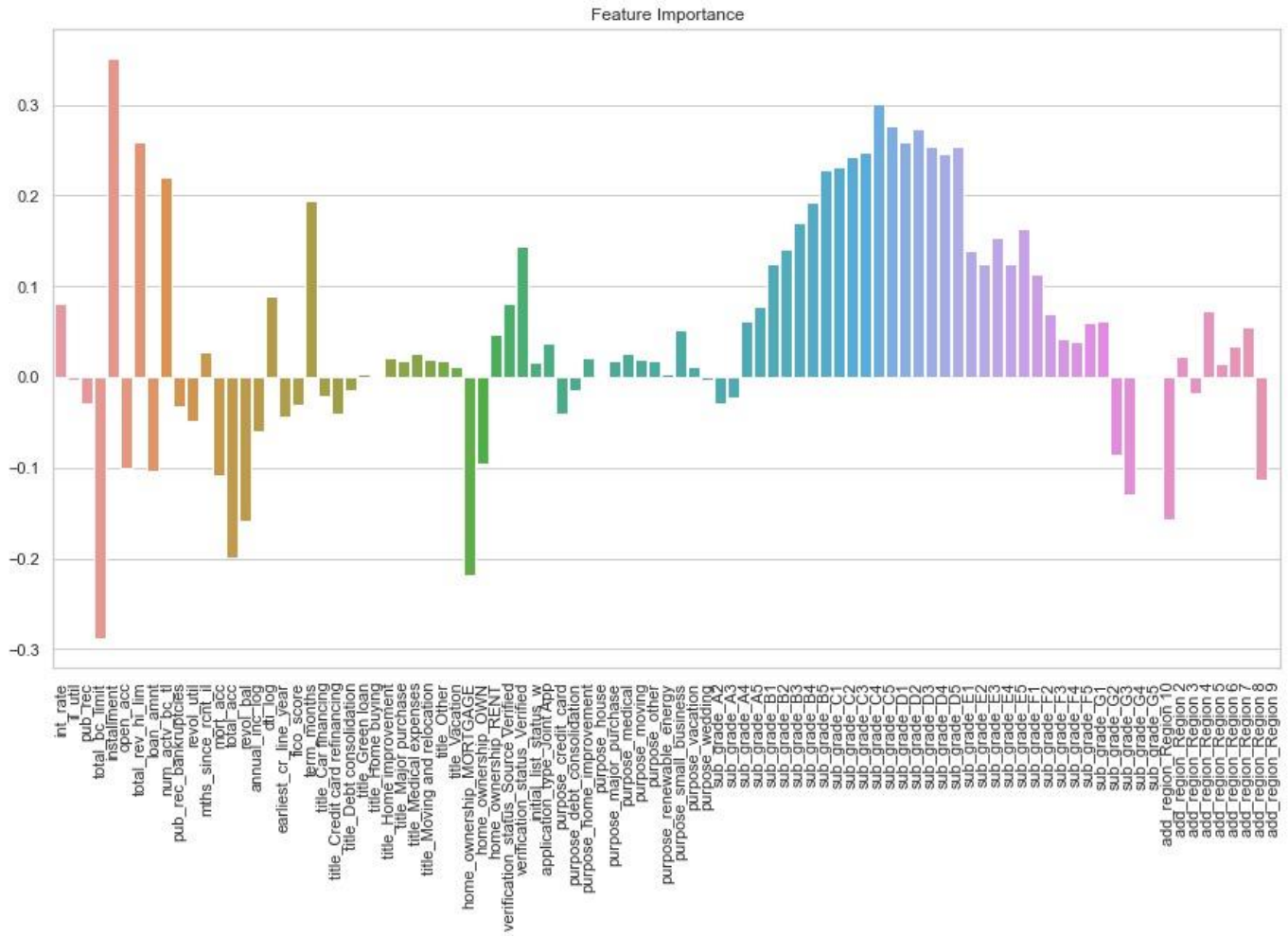


SGDClassifier is giving good results but the results are not making big difference compared to others.

CONCLUSION:

- We have done all the Univariate and Bivariate Analysis on the dataset.
- We used Logistic Regression, Decision Tree, Random Forrest and XGBoost for classification.
- Our aim was to improve Recall.
- After doing SMOTE we found that the LogisticRegression was doing good job out of all the alorithms.
- So for further improvement we did Hyperparameter tuning on LogisticRegression, in which we got max_iter as 50, which is reduced from 1000 iterations.
- This tuning improved the Recall and at the same time, it reduced the time complexity as max iters are reduced.

BUSINESS PROBLEM SOLUTION:



- Total bank card limit and installment play a major role in predicting if some one is going to default or not, which means that a person with higher installment is more likely to default which can be quite obvious that that applicant is not able to manage his income in order to pay back the loan.
- Similarly a person having higher bank card limit is less likely to default.
- Home owenership also plays a major role in predicting default.
- Total credit accounts, which as we stated in EDA, means that person is getting thismany number of accounts because he/she was in good terms with the previous creditor.
- It is interesting to see that some regions have major play in default, which means that company have to take region into consideration while giving out loan.

REFERENCES:

- LendingClub Websit: (https://www.lendingclub.com)
- LendingClub Wikipedia page: (https://en.wikipedia.org/wiki/Lending_Club)
- Wiki: https://en.wikipedia.org/wiki/List_of_regions_of_the_United_States
- Usps: https://pe.usps.com/text/pub28/28apb.htm