VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

CSE 3020 - DATA VISUALISATION

PROJECT REPORT ON

PEER TO PEER CREDIT LENDING SYSTEM

SUBMITTED TO NALINI N

Submitted By

VIVEK REDDY M 18BCE0718

SRI TEJA ALURI 18BCE0759

VISHNU MALLELA 18BCE2131

Abstract

Peer-to-peer credit loaning framework is the act of loaning cash to people or organizations through online administrations that match banks with borrowers. "Loaning Club" is the world's biggest shared loaning stage. The loan fees can be set by banks who go after the most reduced rate on the converse closeout model or fixed by the mediator organization based on an investigation of the borrower's credit. There is the threat of the borrower defaulting on the credits taken out from peer-advancing destinations.

How to stay away from this danger?

At the point when a loan specialist is giving cash he/she needs to comprehend which stage he is utilizing for loaning cash and to take fundamental arrangement whether the stage deals with the cash they have contributed. Consequently stage needs to deal with borrowers character for security purposes.

Thus all the loaning stages needs to consider numerous properties before they award credit to the borrower from a particular moneylender. Every single loaning stage has encountered a few chargeoffs and we need to comprehend the reason for charge offs and work in like manner. In spite of the fact that there exists a few expectation frameworks for a portion of the loaning stages we need to have an example for all the current loaning frameworks for charge off conditions.

KEYWORDS: Credit, risk, loan, evaluation, neural networks.

Introduction

Distributed loaning stage industry is flourishing lately. A great many financial backers are making benefit through these stages; a huge number of borrowers are getting cash all the more without any problem. Albeit these stages give FICO assessment and fundamental data of borrowers to guarantee loaning tradings are in a wellbeing climate, there are still great many individuals under hazard of losing cash.

Indeed, even Prosper — a main monetary distributed loaning stage organization, still experienced credit hazard issues previously. Be that as it may, after their recreation and new credit framework dispatched, the credit hazard has been improved. It drives me need to discover stories in the background. Here I will investigate this Prosper informational index and attempt to discover a few designs behind borrowers properties, distinctive Rating type, and how they connect to default credit or finished advance.

Plus, we will likewise share a few thoughts regarding inspirations of variable investigation. All things considered, in the a particularly enormous informational collection, on the off chance that we start from knowing some essential space information, it's simpler to uncover significant highlights among this large information.

Distributed loaning is another decentralized model for financial backers loan capital and possible borrowers to profit credit. Different financial backers and expected borrowers work together on normal online stage that doesn't include any monetary establishments or agents eventually to-end measure. The borrower demands for advance sum, which reach is ordered on premise of loan fees. The financial backer has the freedom to pick the measure of capital, the interest cap and surprisingly the borrower. These credits are not totally secure as they imply generous danger of default and subsequently require added exertion to distinguish and decide a borrower from a pool on obscure clients. Distinguishing potential credit defaults is required in shared loaning stages, however deciding the great credits and financing them has higher need, as the P2P income model is reliant upon the quantity of advances, volume of credit and so on, yet causing misfortunes because of defaults will make the business unviable. The handling of borrower's advance application includes assortment of client information like yearly pay, record, bank balance, different credits, and so forth The premier need is to recognize the subset of these properties that is equipped for grouping the credit application as one with potential default hazard or not tree based classifiers are utilized, as the run-time intricacy of building the tree is directly relative to the quantity of highlights. They are additionally fit for dealing with the skewness of the information, which is beyond the realm of imagination on account of "Calculated Regression", where oversampling is required. Tree based grouping is an information mining strategy which recursively fabricates a bunch of rules dependent on various information factors that are either mathematical or straight out and yields a class. Tree based classifiers like Decision Tree [2], Random Forest [3], Bagging [4] and Extra Trees [5] are utilized to prepare forecast models for the shared moneylenders information. The profit from speculation for a financial backer on any advance is high if the advance has a higher loan cost, however the danger of defaulting is additionally relatively high. Financial backer can likewise decide to play protected by putting resources into low interest advances which has higher odds of being reimbursed. Consequently the tree based models are upgraded to improve exactness in front of precision. The subtleties and split up of the advance's loan costs, is clarified in the forthcoming areas.

Distributed loaning is a type of direct loaning of cash to people or organizations without an authority monetary foundation partaking as a mediator in the arrangement. P2P loaning is by and large done through online stages that match banks with the possible borrowers.

P2P loaning offers both got and unstable credits. Notwithstanding, the vast majority of the advances in P2P loaning are unstable individual credits. Gotten advances are uncommon for the business and are generally supported by extravagance merchandise. Because of some novel qualities, shared loaning is considered as an elective wellspring of financing.

A few destinations spend significant time specifically kinds of borrowers. StreetShares, for instance, is intended for private ventures. Also, Lending Club has a "Patient Solutions" class that connections specialists who offer financing programs with planned patients.

The primary thought is savers getting higher premium by loaning out their cash as opposed to saving it, and borrowers getting assets at relatively low financing costs.

It normally utilizes an online stage where the borrowers and loan specialists register themselves. Due determination is done prior to permitting the gatherings to take an interest in any loaning or getting action. All P2P stages will presently be viewed as non-banking monetary organizations and directed by the RBI.

How does peer-to-peer lending function?

Peer-to-peer lending is a genuinely clear cycle. Every one of the exchanges are helped out through a specific online stage. The means beneath portray the overall P2P loaning measure:
A potential borrower keen on getting a credit finishes an online application on the shared loaning stage. The stage surveys the application and decides the danger and FICO assessment of the candidate. At that point, the candidate is allocated with the suitable loan cost.

At the point when the application is endorsed, the candidate gets the accessible alternatives from the financial backers dependent on his FICO score and relegated loan costs. The candidate can assess the recommended alternatives and pick one of them. The candidate is answerable for paying intermittent (typically month to month) interest installments and reimbursing the chief sum at development.

The organization that keeps up the online stage charges an expense for the two borrowers and financial backers for the offered types of assistance.

The proposed dynamic model joins social issues, gauging them up with monetary issues for dynamic by socially mindful moneylenders. These organizations have various missions; for instance, some focus on the climate, though others focus on ladies strengthening. The model joins the significance of every angle, in a rational route with the institutional mission. This is finished through Scientific Pecking order Cycle (AHP) procedure by Saaty (1980) a strategy that works on a diverse issue through progressive investigation approach. AHP permits joining the information on experts in various fields inside a specialist framework and empowers abstract decisions between various standards. AHP has been applied in friendly issues to total proportions of corporate social execution; see Rufet al (1998).The model surveys the record of loan repayment of the candidate (past), bookkeeping data and elusive resources from the actual candidate (present), and the undertaking to be financed, from the monetary and from the social perspective (future). These standards are reflected in various quantifiable markers, which are assessed by credit experts. Past a score, the model permits distinguishing the qualities and shortcomings of the undertaking to be financed. The most difficult part of the model is the way to esteem social effects identified with hierarchical points (Forbes, 1998; Munda, 2004; Casing and O'Connor, 2011). Among every one of the distinctive accessible methodologies, the Social Profit from Venture (SROI) by REDF (2001) has been picked. SROI attempts to change social points into monetary measures by utilizing intermediaries. This is particularly valuable for scoring purposes.

Benefits and drawbacks of shared loaning

Distributed loaning gives some huge benefits to the two borrowers and moneylenders:

Better yields to the financial backers: P2P loaning by and large gives more significant yields to the financial backers comparative with different sorts of speculations.
More open wellspring of subsidizing: For certain borrowers, shared loaning is a more available wellspring of subsidizing than standard mortgages from monetary foundations. This might be brought about by the low credit score of the borrower or abnormal reason for the advance.

Lower financing costs: P2P credits normally accompany lower loan fees as a result of the more noteworthy rivalry among banks and lower start charges.

By shared loaning accompanies a couple of disservices:

Credit hazard: Distributed advances are presented to high credit chances. Numerous borrowers who apply for P2P advances have low FICO assessments that don't permit them to get a typical mortgage from a bank. Consequently, a bank ought to know about the default likelihood of his/her counterparty.
No protection/government assurance: The public authority doesn't give protection or any type of security to the loan specialists in the event of the borrower's default.

Enactment: A few locales don't permit shared loaning or require the organizations that offer such types of assistance to conform to venture guidelines. Consequently, shared loaning may not be accessible to certain borrowers or moneylenders.

Literature Review

1. Loan Credibility Prediction System Based on Decision Tree Algorithm

Sivasree, M. S., & Rekha Sunny, T. (2015) *International Journal of Engineering Research & Technology (IJERT).*

Information mining strategies are turning out to be mainstream these days as a result of the wide accessibility of enormous amounts of information and the requirement for changing such information into information. Procedures of information mining are executed in different spaces, for example, retail industry, media transmission industry, organic information investigation, interruption discovery and other logical applications. Information mining methods can likewise be utilized in the financial business which assist them with contending the market exceptional. In this paper the creators present a viable forecast model for the investors that assist them with anticipating the sound clients who have applied for a credit. Choice Tree Induction Data Mining Algorithm is applied to foresee the traits significant for believability. A model of the model is portrayed in this paper which can be utilized by the associations in settling on the correct choice to support or reject the advance solicitation of the clients.

2. Credit Risk Evaluating System Using Decision Tree – Neuro BasedModel

Kabari, L. G., & Nwachukwu, E. O. (2013). Credit risk evaluating system using decisiontree–Neuro based model. *Int J Eng Res & Technol (IJERT), 2*(6).

Dynamic in a complex and powerfully changing climate of the current day requests new procedures of computational knowledge for building a similarly versatile, crossover clever choice emotionally supportive network. In this paper, a Decision TreeNeuro Based model was created to deal with credit giving choice emotionally supportive networks. The framework utilizes an incorporation of Decision Tree and Artificial Neural Networks with a crossover of Decision Tree calculation and Multilayer Feed-forward Neural Network with backpropagation learning calculation to develop the proposed model. Diverse delegate instances of credit applications were viewed as dependent on the rules of various banks in Nigeria, to approve the framework. Item Oriented Analysis and Design (OO-AD) strategy was utilized in the advancement of the framework, and an article situated programming language was utilized with a MATLAB motor to execute the models and classes planned in the framework. The outcome shows that Decision Tree-Neuro Based Models with its 88% achievement rate and great illustrative foundation are effective innovation that can adjust to the current day credit application assessment in business banks.

3. Determinants of Default in P2P Lending

Serrano-Cinca, C., Gutiérrez-Nieto, B., & López-Palacios, L. (2015). Determinants of defaultin P2P lending. *PloS one*, *10*(10), e0139427.

The exact investigation depends on advances' information gathered from Lending Club (N = 24,449) from 2008 to 2014 that are first examined by utilizing univariate implies tests and endurance examination. Elements clarifying default are advance reason, yearly pay, current lodging circumstance, financial record and obligation. Also, a strategic relapse model is created to anticipate defaults. The evaluation appointed by the P2P loaning site is the most prescient factor of default, however the precision of the model is improved by adding other data, particularly the borrower's obligation level.

4. Risk Assessment in Social Lending via Random Forests

Malekipirbazari, M., & Aksakalli, V. (2015). Risk assessment in social lending via randomforests. *Expert Systems with Applications*, *42*(10), 4621-4631.

With the development of electronic trade and social stages, social loaning (otherwise called distributed loaning) has arisen as a feasible stage where moneylenders and borrowers can work together without the assistance of institutional middle people like banks. Social loaning has acquired critical force as of late, for certain stages coming to multi-billion dollar advance course in a short measure of time. Then again, supportability and conceivable inescapable reception of such stages rely intensely upon dependable danger attribution to singular borrowers. For this reason, the authors proposed an irregular woods(RF) based grouping technique for foreseeing borrower status. Our outcomes on information from the famous social loaning stage Lending Club (LC) demonstrate the RF-based strategy beats the FICO financial assessments just as LC grades in recognizable proof of good borrowers.

5. Neuro-Based Artificial Intelligence Model for Loan Decisions

Eletter, S. F., Yaseen, S. G., & Elrefae, G. A. (2010). Neuro-based artificial intelligence modelfor loan decisions. *American Journal of Economics and Business Administration*, *2*(1), 27.

Notwithstanding the expansion in shopper advances defaults and contest in the financial market, the greater part of the Jordanian business banks are hesitant to utilize computerized reasoning programming frameworks for supporting credit choices. This examination fostered a proposed model that recognizes fake neural organizations as an empowering device for assessing credit applications to help advance choices in the Jordanian Commercial banks. A multi-facet feed-forward neural organization with backpropagation learning calculation was utilized to develop the proposed model.

6. Ensemble Neural Network Strategy for Predicting Credit Default Evaluation

Ghatge, A. R., & Halkarnikar, P. P. (2013). Ensemble neural network strategy for predicting credit default evaluation. International Journal of Engineering and Innovative Technology (IJEIT), 2(7), 223-225.

In this paper, the engineering utilized the hypothesis of counterfeit neural organizations and business rules to accurately decide if a client is default or not. The Feed-forward back proliferation neural organization is utilized to foresee the credit default. The aftereffects of applying the counterfeit neural organizations strategy to order credit hazard dependent on chose boundaries show capacities of the organization to get familiar with the examples.

7. Credit Risk Modeling using Multiple Regressions

Dimitriu, M., Oprea, I. A., & Scrieciu, M. A. (2012). Credit Risk Modeling using Multiple Regressions. International Journal of Advances in Management and Economics, 1(5), 125.

In traditional hypothesis, the danger is restricted to numerical assumption for misfortunes that can happen while picking one of the potential variations. For banks, hazard is addressed as misfortunes emerging from the fulfillment of some choice. Bank hazard is a marvel that happens during the movement of banking tasks and that causes adverse consequences for those exercises: crumbling of business or record bank misfortunes influencing usefulness. It very well may be brought about by inside or outer causes, created by the serious climate.

8. The legal protection of lenders in peer to peer lending system

*Budiharto Budiharto - Fakultas Hukum, Universitas Diponegoro, Indonesia
Sartika Nanda Lestari - Fakultas Hukum, Universitas Diponegoro, Indonesia
Gusto Hartanto - Fakultas Hukum, Universitas Diponegoro, Indonesia

Monetary innovation dependent on Peer to Peer Lending is one of the new forward leaps in monetary administrations organizations in Indonesia. The shared loaning stages are basically online business sectors that match the market interest of assets as one of the elective financing systems for individual or business. In any case, there is as yet not many of guideline in regards to distributed loaning. The creators address two inquiries by hypothetical lawful examination by inspecting optional information through writing contemplates. In the first place, the component of acknowledge arrangements for a distributed loaning; second, break down moneylender's lawful insurance in credit arrangements in shared loaning. In view of the examination, the creators tracked down that the component of loaning through a shared loaning credit understanding is in accordance with Financial Services Authority (Otoritas Jasa Keuangan - OJK) Regulation No. 77/POJK.01/2016 concerning Information Technology-Based Lending and Borrowing Services. Futhermore, the assurance of legitimate banks shared loaning from the part of law public has been adequate however in private law, OJK has not had the option to give greatest insurance.

9. A Credit Score System for Socially Responsible Lending

Begoña Gutiérrez-Nieto, Carlos Serrano-Cinca, Juan Camón-Cala

Moral banking, microfinance foundations or certain credit cooperatives, among others, award socially capable advances. This paper presents a FICO assessment framework for them. The model assesses social and monetary parts of the borrower. The monetary perspectives are assessed under the customary financial system, by examining bookkeeping proclamations and monetary projections. The social perspectives attempt to evaluate the advance effect on the accomplishment of Millennium Development Goals like business, schooling, climate, wellbeing or local area sway. The social financial assessment model should fuse the bank's expertise and ought to likewise be reasonable with its central goal. This is finished utilizing Multi-Criteria Decision Making (MCDM). The paper shows a genuine case: a credit application by a social business person introduced to a socially capable loan specialist. The choice emotionally supportive network delivers a score, yet additionally uncovers qualities and shortcomings of the application.

10. Process model on P2P lending

Huaiqing Wang, Kun Chen, Wei Zhu & Zhenxia Song

Online shared loaning (P2P loaning) is blasting as the prevalence of e-money. To foster a reasonable model for the P2P loaning measure is incredible significant for directors to tack the issues of promoting, the board and activity. In this paper, the creators center around the P2P loaning measure display and furnish a relative investigation contrasting and conventional bank credit measure. Initially, our model shows that the data stream in P2P loaning is more incessant and straightforward. Furthermore, the model uncovers that P2P loaning utilizes a very unique credit tryout strategy, which depends on data and the choice model in the P2P frameworks. Thirdly, the advance administration isn't finished regularly in P2P loaning, in light of the fact that most P2P organizations don't have the post-credit records of borrowers.

11. Online Peer-to-Peer Lending

Alexander Bachmann (Alexander.Bachmann@stud.leuphana.de),
Alexander Becker (Alexander.Becker@stud.leuphana.de),
Daniel Buerckner (Daniel.Buerckner@stud.leuphana.de),
Michel Hilker (Michel.b.Hilker@stud.leuphana.de),
Frank Kock (Frank.Kock@stud.leuphana.de),
Mark Lehmann (Mark.Lehmann@stud.leuphana.de)
Phillip Tiburtius (phillip.tiburtius@uni.leuphana.de)

With the impending ubiquity of online networks in the previous decade another method of advance beginning has entered the credit market: online shared (P2P) loaning. It moves the old thought of individual credits into the World Wide Web. In this sort of loaning model the intercession of monetary establishments isn't needed (Herzenstein et al., 2008;Galloway, 2009). The choice cycle of credit beginning is surrendered to the hand of private banks and borrowers, and sites like Prosper.com offer them a stage to draw in with one another. Inside these stages borrowers for the most part depict the reason for their advance ask for and give data about their present monetary circumstance, similar to pay or open credit lines. Moneylenders at that point have the chance to bring to the table an advance with a financing cost inferred upon this data. For borrowers, online P2P loaning is an approach to get an advance without a monetary organization associated with the choice cycle and may likewise be a likelihood to get preferable conditions over in the customary financial framework. For moneylenders it very well may be viewed as a venture model where the speculation hazard is coupled to the FICO score of the financed credits. The actual stages frequently advantage by raising charges for effective acknowledged exchanges (Galloway,2009). Albeit online P2P loaning is a moderately youthful field of exploration an expanding measure of logical commitments has been distributed lately (Iyer, Khwaja, Luttmer, and Shue, 2009; Pope and Sydnor, 2008; Ravina, 2007). With the development of the first online P2P loaning stage "Zopa" the new loaning model raised consideration without precedent for the year 2006 (Hulme and Wright, 2006). Anyway it was Prosper.com, who caused a flood of logical commitments by unveiling the whole stage's information in 2007. From that point forward, the subject has drawn in specialists from the fields of financial matters, data innovation and sociologies to research the connections among banks and borrowers in online P2P loaning stages.

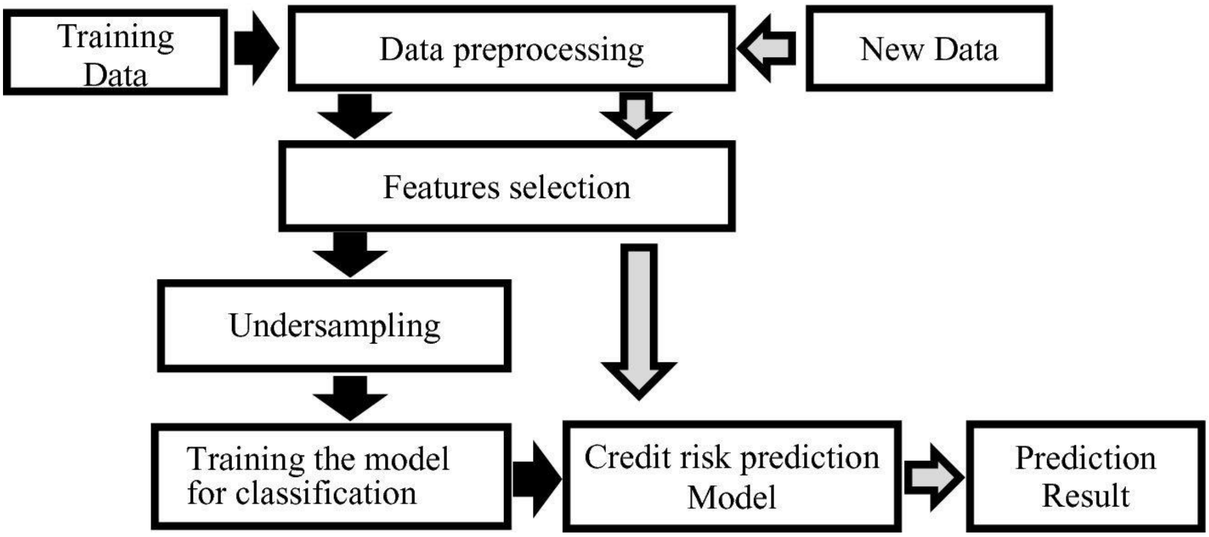
Following this new exploration region, the authors saw the requirement for a thorough writing survey that gives an outline of the present status of logical examination in P2P loaning. It is our objective to incorporate distinctive delegate research commitments and give a rundown of the determinants in P2P-loaning. Our primary objective gatherings are per users from the data framework and monetary administration networks.

In the wake of portraying the paper choice methodology a short outline of the online P2P loaning business sector and its partners is given. The principle some portion of this survey covers the determinants in P2P loaning. The authors partition them into monetary qualities, additionally called hard-factors, and delicate variables like segment attributes and gathering intermediation. The authors finish up with thoughts for future examination in online P2P loaning.

Proposed Methodology

Framework

Predicting whether the borrower would “charge-off” or will fall in the low,medium , high interest rate range.



Data Collection

Attributes	Description
acc_now_delinq	The number of accounts on which the borrower is now delinquent.
acc_open_past_24mths	Number of trades opened in the 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 excludingmedical 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 requestedLC 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.

Libraries used:

- **NumPy**: NumPy is a universally useful exhibit preparing bundle. It gives an elite multidimensional exhibit article, and apparatuses for working with these clusters.
- **Pandas**: Pandas is a Python bundle giving quick, adaptable, and expressive information structures intended to make working with organized (plain, multidimensional, conceivably heterogeneous) and time arrangement information both simple and natural. It intends to be the principal significant level structure block for doing down to earth, true information examination in Python.
- **Matplotlib**: Matplotlib is a Python 2D plotting library which produces distribution quality figures in an assortment of printed copy designs and intuitive conditions across stages.
- **Seaborn**: Seaborn is a Python information perception library dependent on matplotlib. It gives an undeniable level interface to drawing appealing and educational measurable designs.
- **Plotly**: Plotly gives internet diagramming, investigation, and insights apparatuses for people and cooperation, just as logical charting libraries for Python, R, MATLAB, Perl, Julia, Arduino, and REST.

Algorithms Used:

Logistic Regression

Logistic regression is a measurable model that in its essential structure utilizes a calculated capacity to demonstrate a double reliant variable, albeit a lot more mind boggling expansions exist. In relapse investigation, calculated relapse (or logit relapse) is assessing the boundaries of a Logistic regression (a type of twofold relapse). Numerically, a paired calculated model has a reliant variable with two potential qualities, like pass/bomb which is addressed by a marker variable, where the two qualities are named "0" and "1". In the strategic model, the log-chances (the logarithm of the chances) for the worth marked "1" is a straight mix of at least one free factors ("indicators"); the autonomous factors can each be a double factor (two classes, coded by a marker variable) or a persistent variable (any genuine worth). The comparing likelihood of the worth named "1" can shift between 0 (unquestionably the worth "0") and 1 (absolutely the worth "1"), thus the marking; the capacity that converts log-chances to likelihood is the calculated capacity, consequently the name. The unit of estimation for the log-chances scale is known as a logit, from strategic unit, henceforth the elective names. Similar to models with an alternate sigmoid capacity rather than the calculated capacity can likewise be utilized, for example, the probit model; the characterizing normal for the strategic model is that expanding one of the free factors multiplicatively scales the chances of the given result at a steady rate, with every autonomous variable having its own boundary; for a double reliant variable this sums up the chances proportion.

Random Forest

Random forests or random decision forests are a troupe learning technique for order, relapse and different errands that works by developing a large number of choice trees at preparing time and yielding the class that is the method of the classes (characterization) or mean expectation (relapse) of the individual trees. Irregular choice backwoods right for choice trees' propensity for overfitting to their preparation set. The principal calculation for arbitrary choice backwoods was made by Tin Kam Housing the irregular subspace strategy, which, in Ho's plan, is an approach to carry out the "stochastic segregation" way to deal with characterization proposed by Eugene Kleinberg.

Imbalance Data

Anomaly Detection or Outlier Analysis:

In data mining, anomaly detection is the identification of rare items, events or observations which raise suspicions by differing significantly from the majority of the data.

Over-sampling:

In signal processing, oversampling is the process of sampling a signal at asampling frequency significantly higher than the Nyquist rate. Theoretically, a bandwidth-limited signal can be perfectly reconstructed if sampled at the Nyquist rate or above it.

Under-sampling:

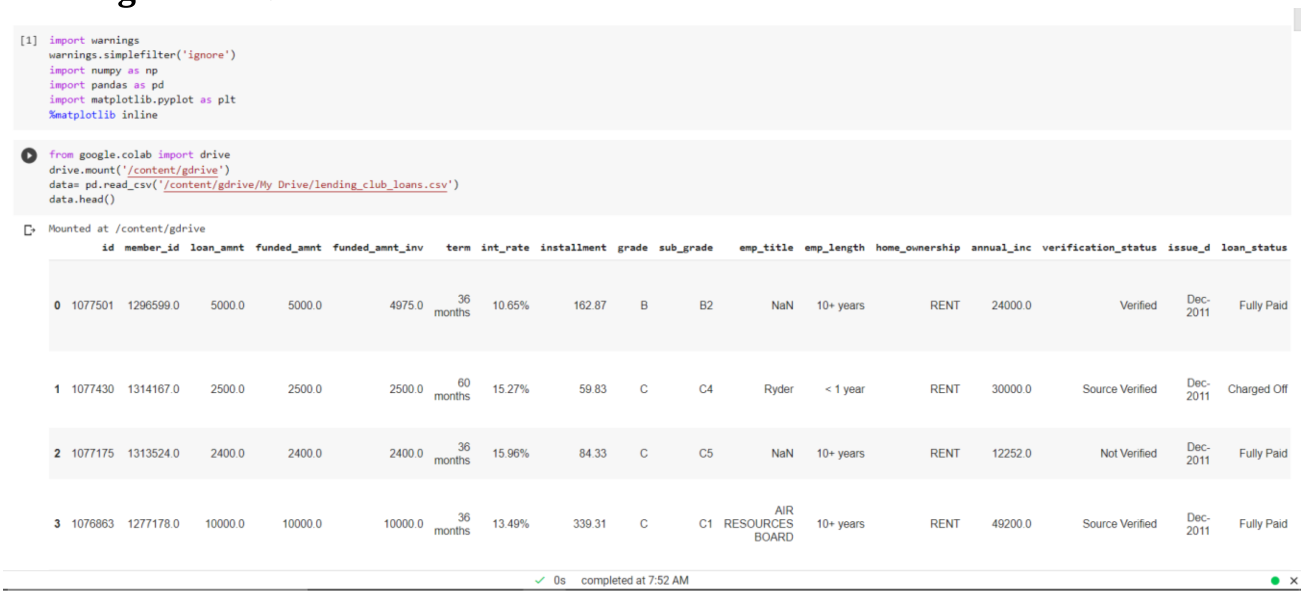
In signal processing, under-sampling of bandpass sampling is a techniquewhere one samples a bandpass-filtered signal at a sample rate below its Nyquist rate, but is still able to reconstruct the signal.

SMOTE and ADASYN:

This is a statistical technique for increasing the number of cases in your dataset in a balanced way.

Code And Results

Loading Dataset:



```
[1] import warnings
    warnings.simplefilter('ignore')
    import numpy as np
    import pandas as pd
    import matplotlib.pyplot as plt
    %matplotlib inline
```

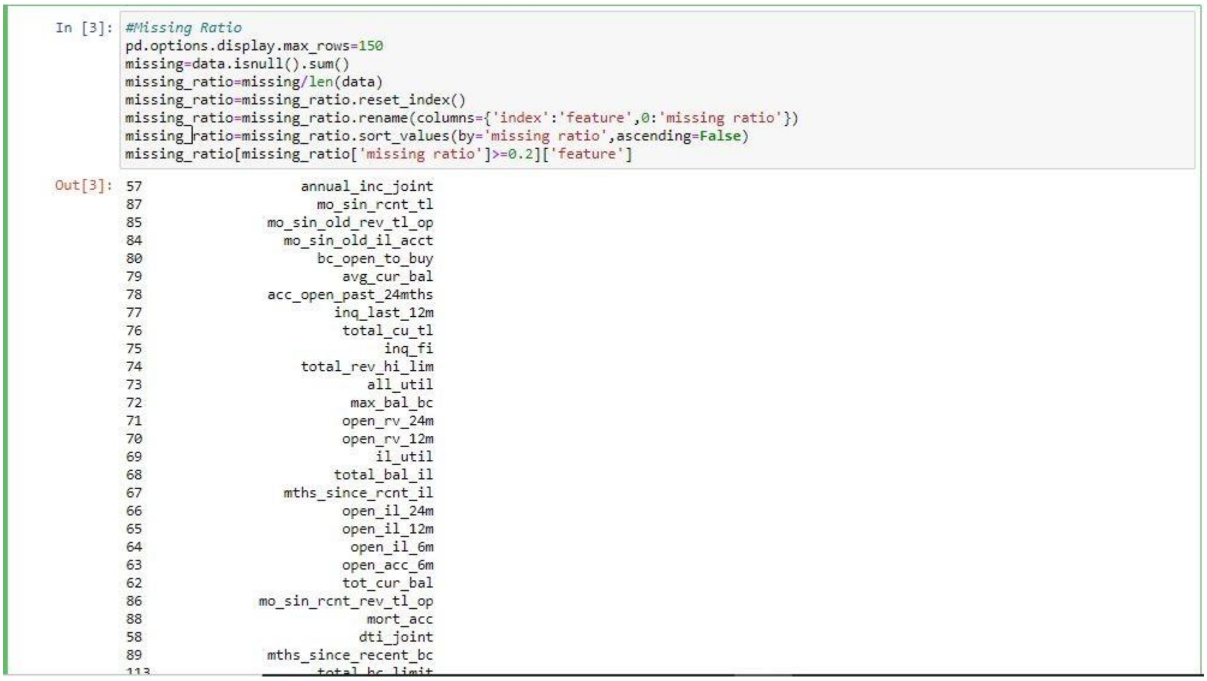
```
from google.colab import drive
drive.mount('/content/gdrive')
data= pd.read_csv('/content/gdrive/My Drive/lending_club_loans.csv')
data.head()
```

```
Mounted at /content/gdrive
```

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	emp_title	emp_length	home_ownership	annual_inc	verification_status	issue_d	loan_status
0	1077501	1296599.0	5000.0	5000.0	4975.0	36 months	10.65%	162.87	B	B2	NaN	10+ years	RENT	24000.0	Verified	Dec-2011	Fully Paid
1	1077430	1314167.0	2500.0	2500.0	2500.0	60 months	15.27%	59.83	C	C4	Ryder	< 1 year	RENT	30000.0	Source Verified	Dec-2011	Charged Off
2	1077175	1313524.0	2400.0	2400.0	2400.0	36 months	15.96%	84.33	C	C5	NaN	10+ years	RENT	12252.0	Not Verified	Dec-2011	Fully Paid
3	1076863	1277178.0	10000.0	10000.0	10000.0	36 months	13.49%	339.31	C	C1	AIR RESOURCES BOARD	10+ years	RENT	49200.0	Source Verified	Dec-2011	Fully Paid

✓ 0s completed at 7:52 AM

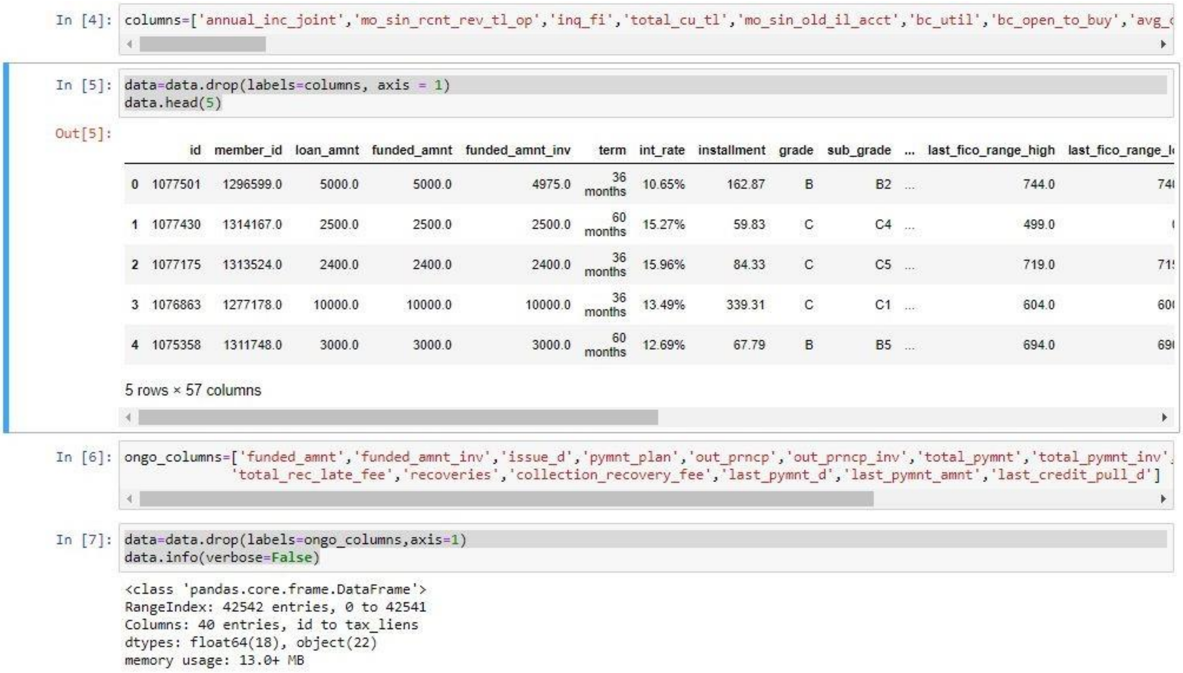
Finding Columns with missing ratio > 0.2:



```
In [3]: #Missing Ratio
        pd.options.display.max_rows=150
        missing=data.isnull().sum()
        missing_ratio=missing/len(data)
        missing_ratio=missing_ratio.reset_index()
        missing_ratio=missing_ratio.rename(columns={'index':'feature',0:'missing ratio'})
        missing_ratio=missing_ratio.sort_values(by='missing ratio',ascending=False)
        missing_ratio[missing_ratio['missing ratio']>=0.2]['feature']
```

```
Out[3]: 57                annual_inc_joint
        87                 mo_sin_rcnt_tl
        85            mo_sin_old_rev_tl_op
        84             mo_sin_old_il_acct
        80                 bc_open_to_buy
        79                    avg_cur_bal
        78           acc_open_past_24mths
        77                   inq_last_12m
        76                    total_cu_tl
        75                         inq_fi
        74                total_rev_hi_lim
        73                       all_util
        72                     max_bal_bc
        71                    open_rv_24m
        70                    open_rv_12m
        69                        il_util
        68                   total_bal_il
        67             mths_since_rcnt_il
        66                    open_il_24m
        65                    open_il_12m
        64                     open_il_6m
        63                    open_acc_6m
        62                    tot_cur_bal
        86           mo_sin_rcnt_rev_tl_op
        88                       mort_acc
        58                      dti_joint
        89            mths_since_recent_bc
        113                 total_bc_limit
```

Dropping the Columns



```
In [4]: columns=['annual_inc_joint','mo_sin_rcnt_rev_tl_op','inq_fi','total_cu_tl','mo_sin_old_il_acct','bc_util','bc_open_to_buy','avg_
```

```
In [5]: data=data.drop(labels=columns, axis = 1)
        data.head(5)
```

Out[5]:

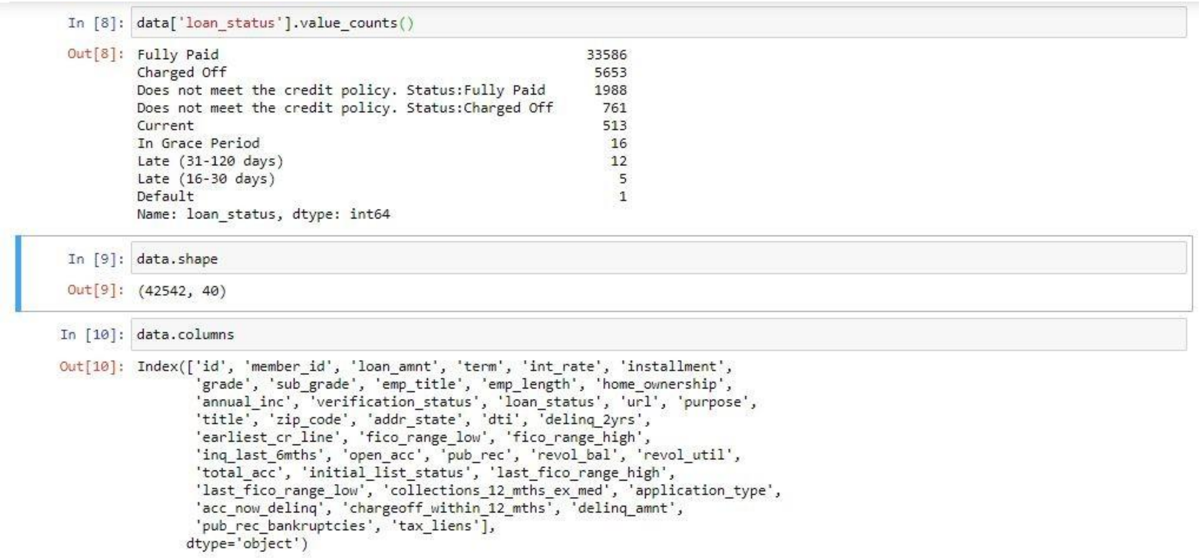
	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	...	last_fico_range_high	last_fico_range_l
0	1077501	1296599.0	5000.0	5000.0	4975.0	36 months	10.65%	162.87	B	B2	...	744.0	74
1	1077430	1314167.0	2500.0	2500.0	2500.0	60 months	15.27%	59.83	C	C4	...	499.0	
2	1077175	1313524.0	2400.0	2400.0	2400.0	36 months	15.96%	84.33	C	C5	...	719.0	71
3	1076863	1277178.0	10000.0	10000.0	10000.0	36 months	13.49%	339.31	C	C1	...	604.0	60
4	1075358	1311748.0	3000.0	3000.0	3000.0	60 months	12.69%	67.79	B	B5	...	694.0	69

5 rows × 57 columns

```
In [6]: ongo_columns=['funded_amnt','funded_amnt_inv','issue_d','pymnt_plan','out_prncp','out_prncp_inv','total_pymnt','total_pymnt_inv',
                      'total_rec_late_fee','recoveries','collection_recovery_fee','last_pymnt_d','last_pymnt_amnt','last_credit_pull_d']
```

```
In [7]: data=data.drop(labels=ongo_columns,axis=1)
        data.info(verbose=False)

        <class 'pandas.core.frame.DataFrame'>
        RangeIndex: 42542 entries, 0 to 42541
        Columns: 40 entries, id to tax_liens
        dtypes: float64(18), object(22)
        memory usage: 13.0+ MB
```

Loan Status:

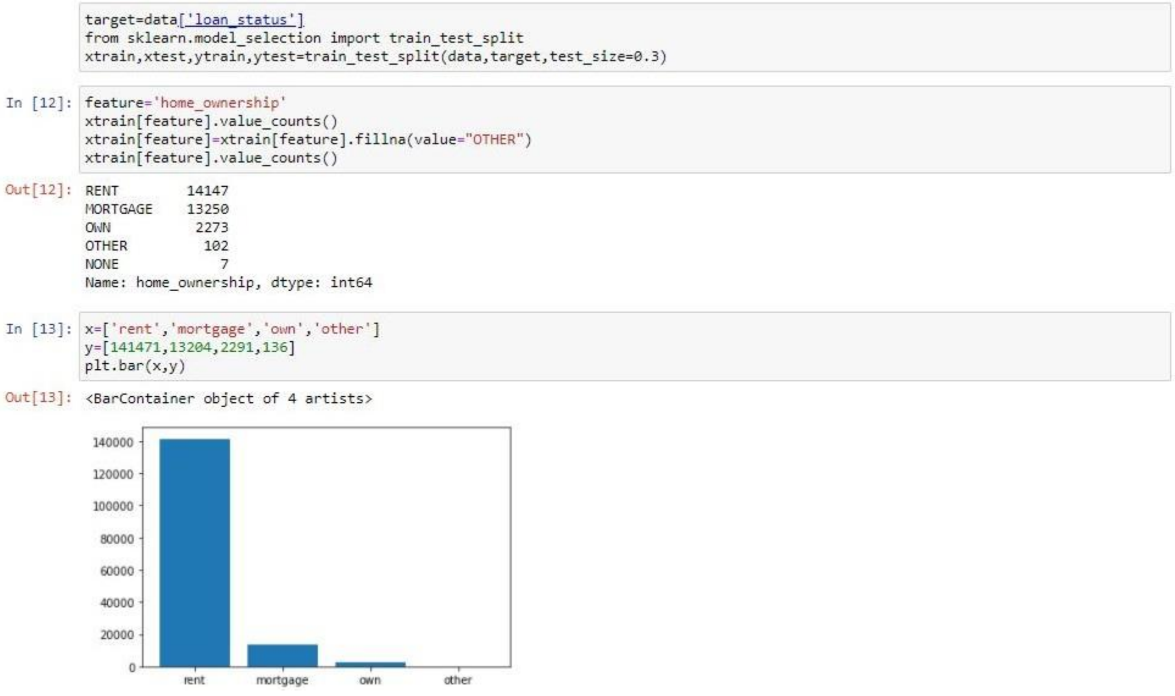


```
In [8]: data['loan_status'].value_counts()

Out[8]: Fully Paid                                             33586
        Charged Off                                             5653
        Does not meet the credit policy. Status:Fully Paid      1988
        Does not meet the credit policy. Status:Charged Off      761
        Current                                                  513
        In Grace Period                                           16
        Late (31-120 days)                                        12
        Late (16-30 days)                                          5
        Default                                                    1
        Name: loan_status, dtype: int64
```

```
In [9]: data.shape

Out[9]: (42542, 40)
```

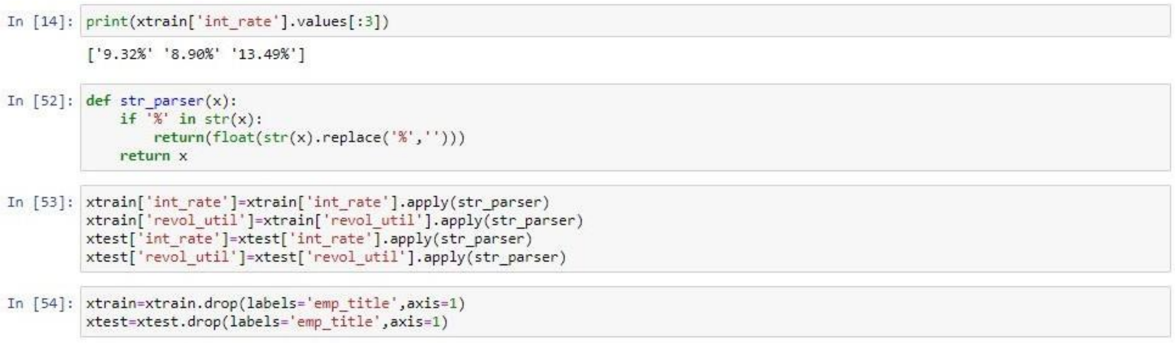
```
In [10]: data.columns

Out[10]: Index(['id', 'member_id', 'loan_amnt', 'term', 'int_rate', 'installment',
                'grade', 'sub_grade', 'emp_title', 'emp_length', 'home_ownership',
                'annual_inc', 'verification_status', 'loan_status', 'url', 'purpose',
                'title', 'zip_code', 'addr_state', 'dti', 'delinq_2yrs',
                'earliest_cr_line', 'fico_range_low', 'fico_range_high',
                'inq_last_6mths', 'open_acc', 'pub_rec', 'revol_bal', 'revol_util',
                'total_acc', 'initial_list_status', 'last_fico_range_high',
                'last_fico_range_low', 'collections_12_mths_ex_med', 'application_type',
                'acc_now_delinq', 'chargeoff_within_12_mths', 'delinq_amnt',
                'pub_rec_bankruptcies', 'tax_liens'],
               dtype='object')
```

Preparing the train and test dataset:

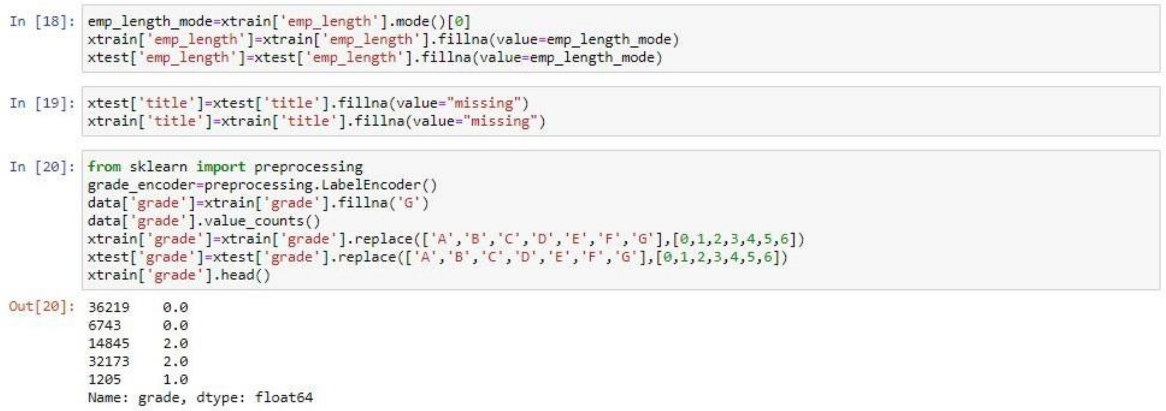


```
        target=data['loan_status']
        from sklearn.model_selection import train_test_split
        xtrain,xtest,ytrain,ytest=train_test_split(data,target,test_size=0.3)

In [12]: feature='home_ownership'
         xtrain[feature].value_counts()
         xtrain[feature]=xtrain[feature].fillna(value="OTHER")
         xtrain[feature].value_counts()

Out[12]: RENT        14147
         MORTGAGE    13250
         OWN          2273
         OTHER         102
         NONE            7
         Name: home_ownership, dtype: int64

In [13]: x=['rent','mortgage','own','other']
         y=[141471,13204,2291,136]
         plt.bar(x,y)

Out[13]: <BarContainer object of 4 artists>
```

Converting Invalid data:

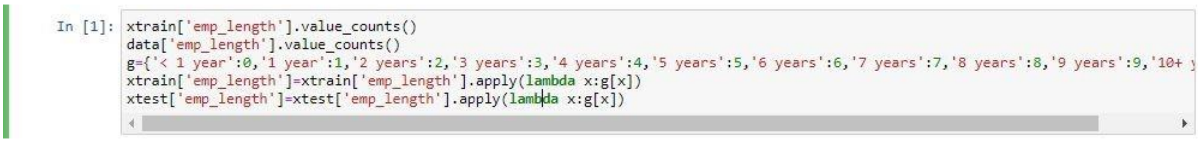


```
In [14]: print(xtrain['int_rate'].values[:3])

         ['9.32%' '8.90%' '13.49%']

In [52]: def str_parser(x):
             if '%' in str(x):
                 return(float(str(x).replace('%','')))
             return x

In [53]: xtrain['int_rate']=xtrain['int_rate'].apply(str_parser)
         xtrain['revol_util']=xtrain['revol_util'].apply(str_parser)
         xtest['int_rate']=xtest['int_rate'].apply(str_parser)
         xtest['revol_util']=xtest['revol_util'].apply(str_parser)

In [54]: xtrain=xtrain.drop(labels='emp_title',axis=1)
         xtest=xtest.drop(labels='emp_title',axis=1)
```



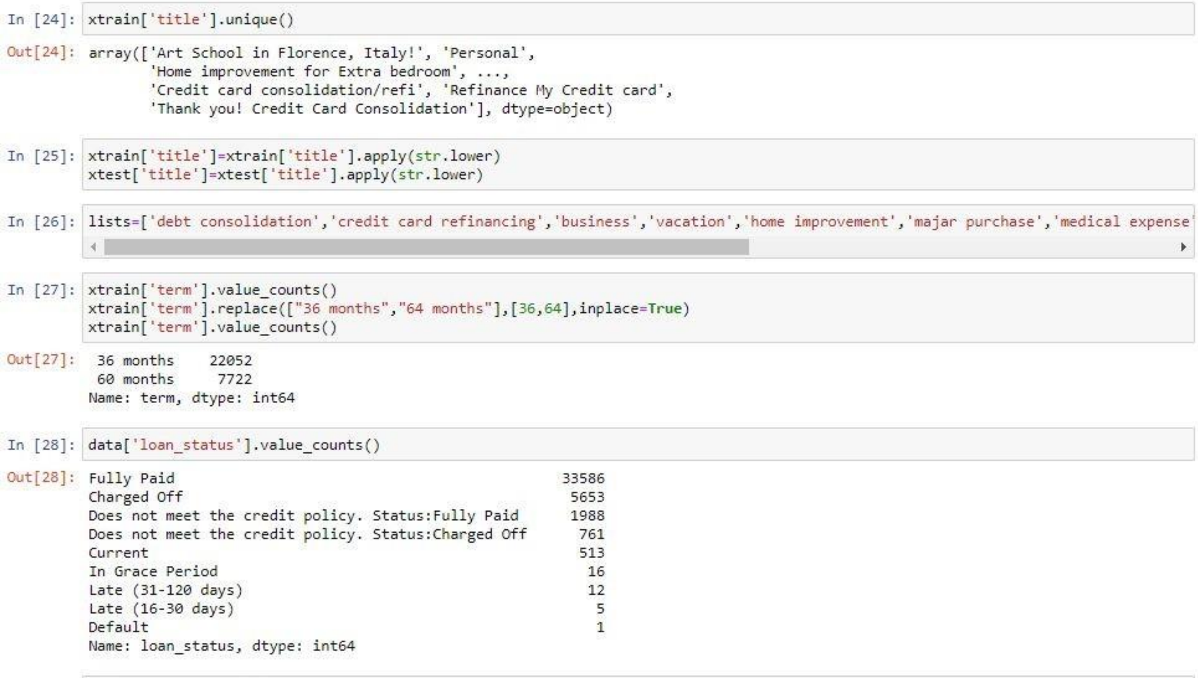
```
In [18]: emp_length_mode=xtrain['emp_length'].mode()[0]
         xtrain['emp_length']=xtrain['emp_length'].fillna(value=emp_length_mode)
         xtest['emp_length']=xtest['emp_length'].fillna(value=emp_length_mode)

In [19]: xtest['title']=xtest['title'].fillna(value="missing")
         xtrain['title']=xtrain['title'].fillna(value="missing")

In [20]: from sklearn import preprocessing
         grade_encoder=preprocessing.LabelEncoder()
         data['grade']=xtrain['grade'].fillna('G')
         data['grade'].value_counts()
         xtrain['grade']=xtrain['grade'].replace(['A','B','C','D','E','F','G'],[0,1,2,3,4,5,6])
         xtest['grade']=xtest['grade'].replace(['A','B','C','D','E','F','G'],[0,1,2,3,4,5,6])
         xtrain['grade'].head()

Out[20]: 36219    0.0
         6743     0.0
         14845    2.0
         32173    2.0
         1205     1.0
         Name: grade, dtype: float64
```

Training and Testing using emp_length:



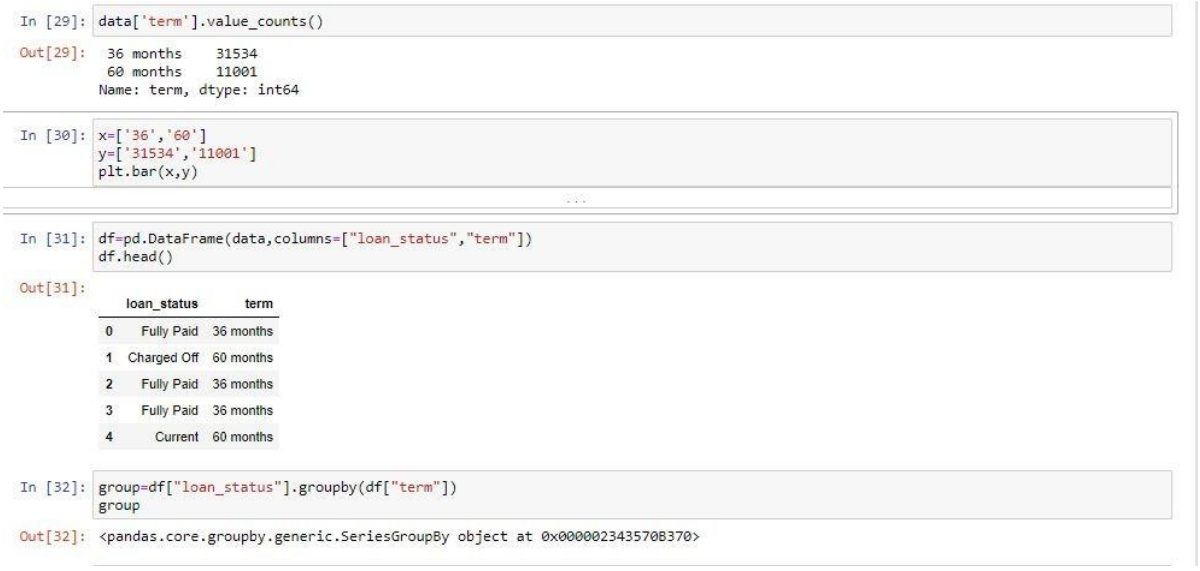
```
In [1]: xtrain['emp_length'].value_counts()
        data['emp_length'].value_counts()
        g={'< 1 year':0,'1 year':1,'2 years':2,'3 years':3,'4 years':4,'5 years':5,'6 years':6,'7 years':7,'8 years':8,'9 years':9,'10+ y
        xtrain['emp_length']=xtrain['emp_length'].apply(lambda x:g[x])
        xtest['emp_length']=xtest['emp_length'].apply(lambda x:g[x])
```

Training and Testing with different elements:



```
In [24]: xtrain['title'].unique()

Out[24]: array(['Art School in Florence, Italy!', 'Personal',
                'Home improvement for Extra bedroom', ...,
                'Credit card consolidation/refi', 'Refinance My Credit card',
                'Thank you! Credit Card Consolidation'], dtype=object)

In [25]: xtrain['title']=xtrain['title'].apply(str.lower)
         xtest['title']=xtest['title'].apply(str.lower)

In [26]: lists=['debt consolidation','credit card refinancing','business','vacation','home improvement','majar purchase','medical expense'

In [27]: xtrain['term'].value_counts()
         xtrain['term'].replace(["36 months","64 months"],[36,64],inplace=True)
         xtrain['term'].value_counts()

Out[27]:  36 months    22052
          60 months     7722
         Name: term, dtype: int64

In [28]: data['loan_status'].value_counts()

Out[28]: Fully Paid                                             33586
         Charged Off                                             5653
         Does not meet the credit policy. Status:Fully Paid      1988
         Does not meet the credit policy. Status:Charged Off      761
         Current                                                  513
         In Grace Period                                           16
         Late (31-120 days)                                        12
         Late (16-30 days)                                          5
         Default                                                    1
         Name: loan_status, dtype: int64
```

Grouping the dataset:

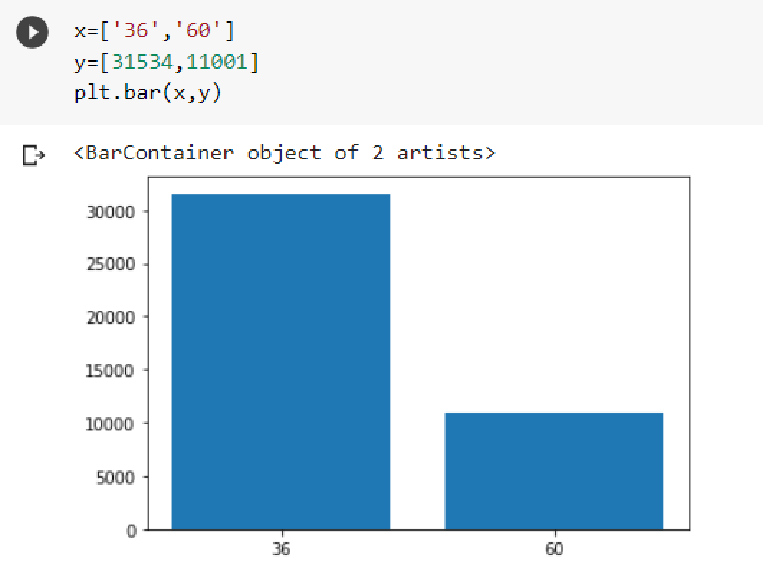


```
In [29]: data['term'].value_counts()

Out[29]:  36 months    31534
          60 months    11001
         Name: term, dtype: int64

In [30]: x=['36','60']
         y=['31534','11001']
         plt.bar(x,y)

In [31]: df=pd.DataFrame(data,columns=["loan_status","term"])
         df.head()

Out[31]:
```

	loan_status	term
0	Fully Paid	36 months
1	Charged Off	60 months
2	Fully Paid	36 months
3	Fully Paid	36 months
4	Current	60 months

```
In [32]: group=df["loan_status"].groupby(df["term"])
         group

Out[32]: <pandas.core.groupby.generic.SeriesGroupBy object at 0x000002343570B370>
```

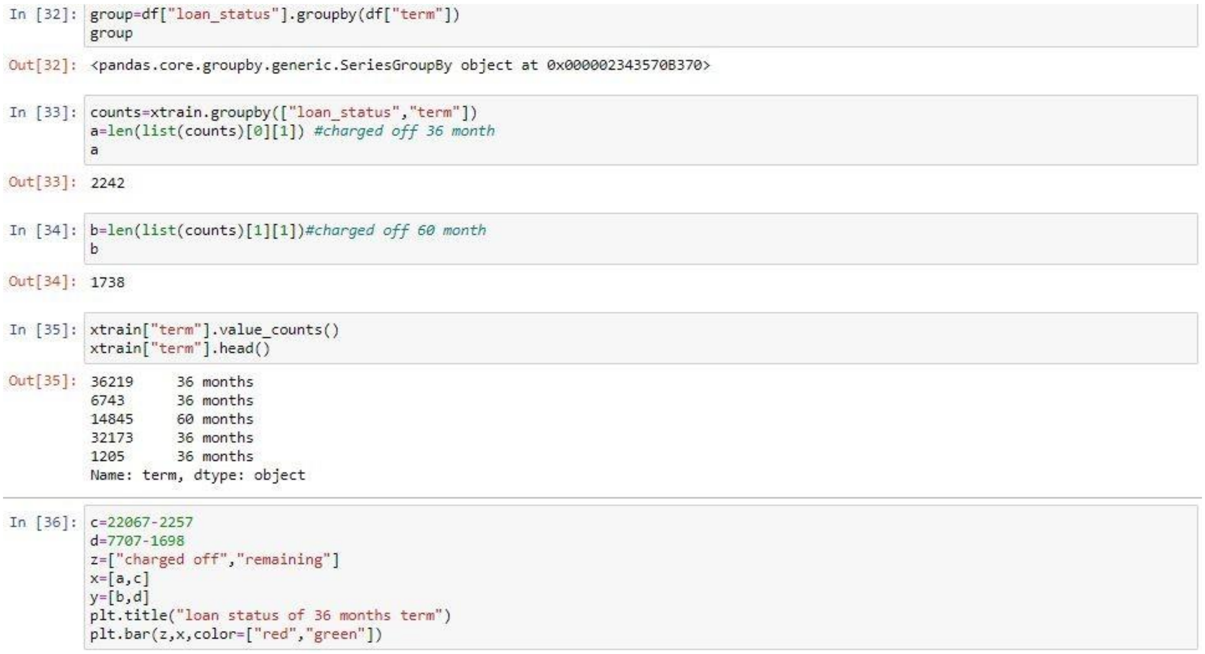
Value Count Of 36 months and 60 months:



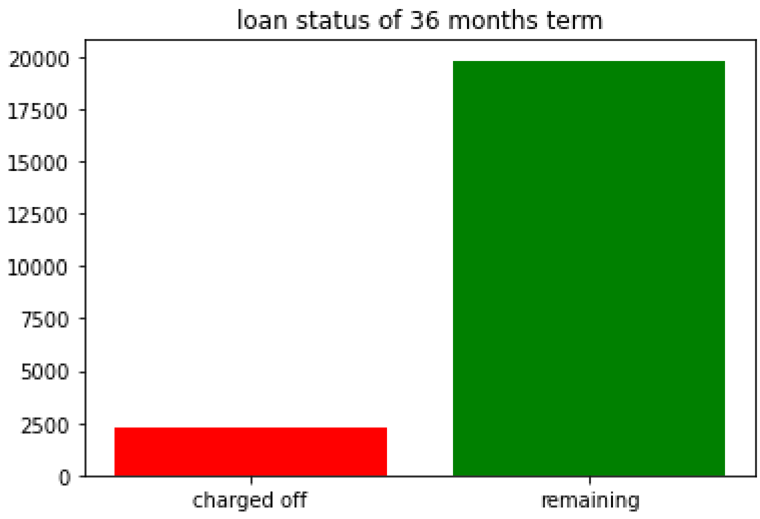
```
x=['36','60']
y=[31534,11001]
plt.bar(x,y)
```

```
<BarContainer object of 2 artists>
```

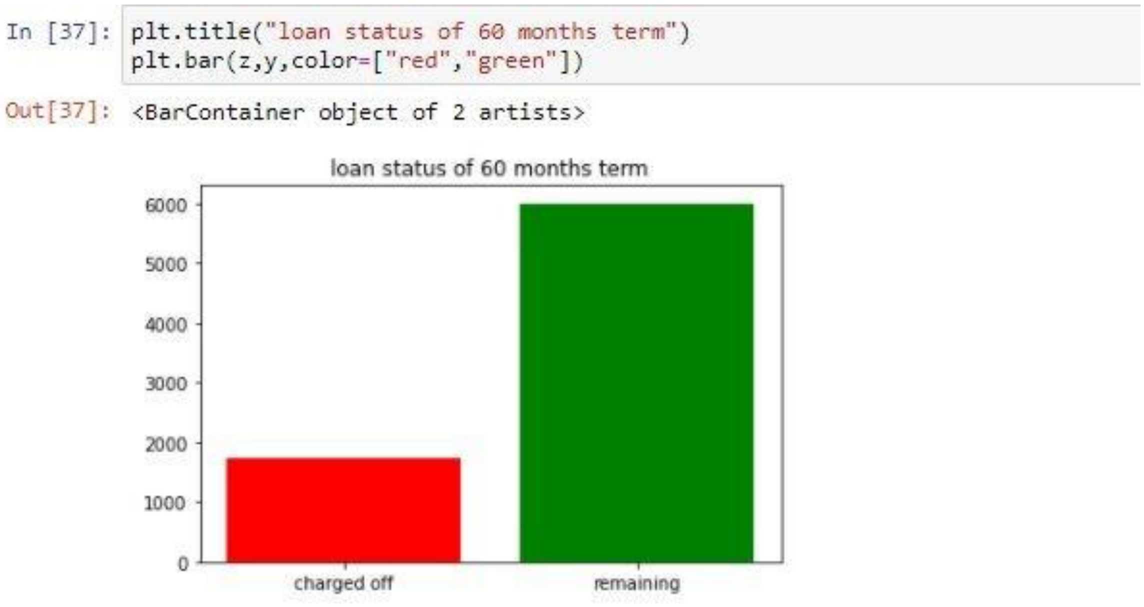
Loan Status (For 36 months):



```
In [32]: group=df["loan_status"].groupby(df["term"])
         group

Out[32]: <pandas.core.groupby.generic.SeriesGroupBy object at 0x000002343570B370>

In [33]: counts=xtrain.groupby(["loan_status","term"])
         a=len(list(counts)[0][1]) #charged off 36 month
         a

Out[33]: 2242

In [34]: b=len(list(counts)[1][1])#charged off 60 month
         b

Out[34]: 1738

In [35]: xtrain["term"].value_counts()
         xtrain["term"].head()

Out[35]: 36219     36 months
         6743      36 months
         14845     60 months
         32173     36 months
         1205      36 months
         Name: term, dtype: object

In [36]: c=22067-2257
         d=7707-1698
         z=["charged off","remaining"]
         x=[a,c]
         y=[b,d]
         plt.title("loan status of 36 months term")
         plt.bar(z,x,color=["red","green"])
```

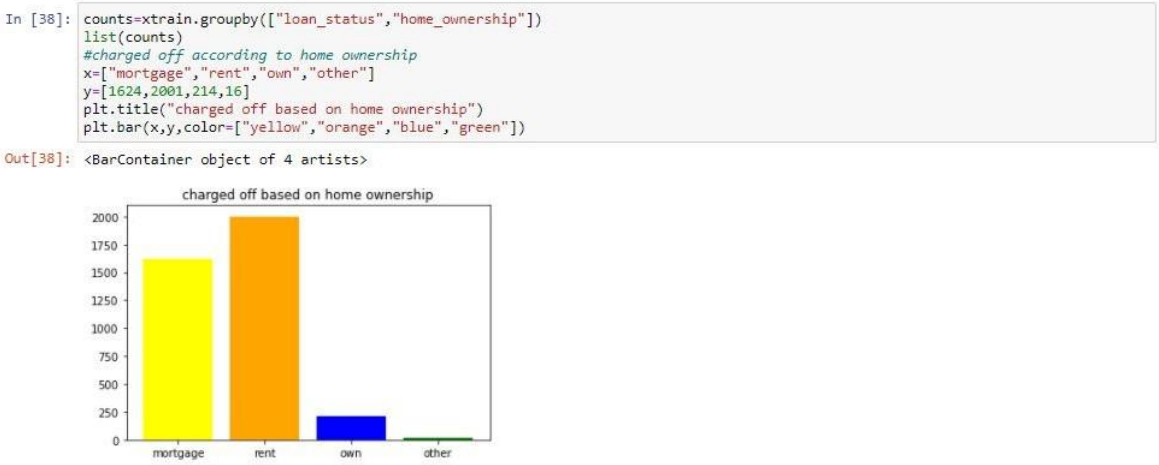


Loan Status (For 60 months):



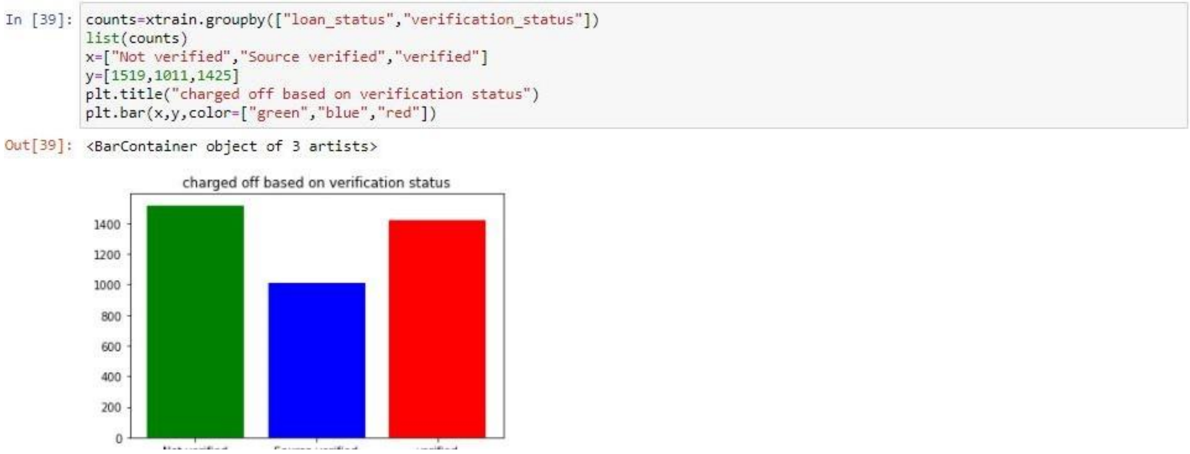
```
In [37]: plt.title("loan status of 60 months term")
         plt.bar(z,y,color=["red","green"])

Out[37]: <BarContainer object of 2 artists>
```

Charge off based on Home Ownership:



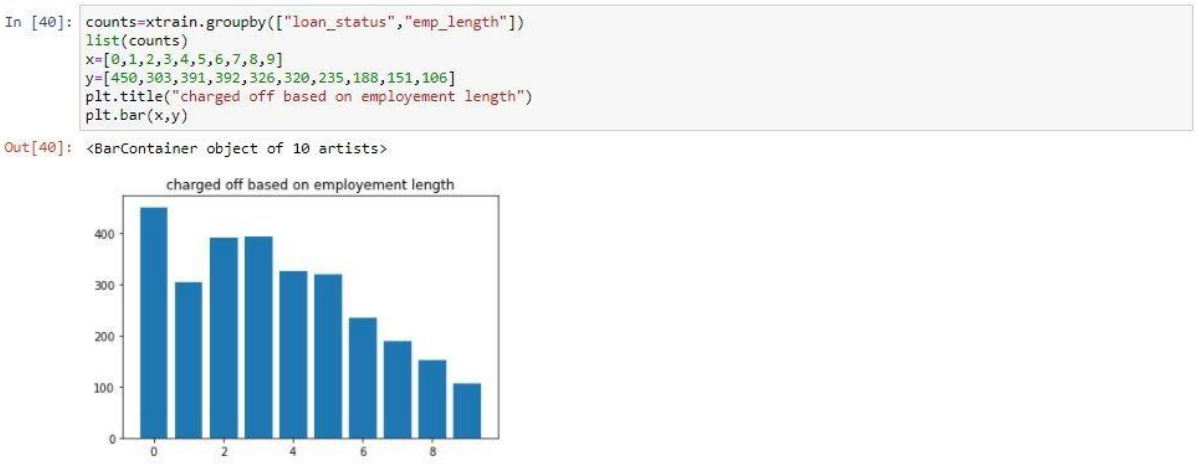
```
In [38]: counts=xtrain.groupby(["loan_status","home_ownership"])
         list(counts)
         #charged off according to home ownership
         x=["mortgage","rent","own","other"]
         y=[1624,2001,214,16]
         plt.title("charged off based on home ownership")
         plt.bar(x,y,color=["yellow","orange","blue","green"])

Out[38]: <BarContainer object of 4 artists>
```

Charge off based on Verification Status:

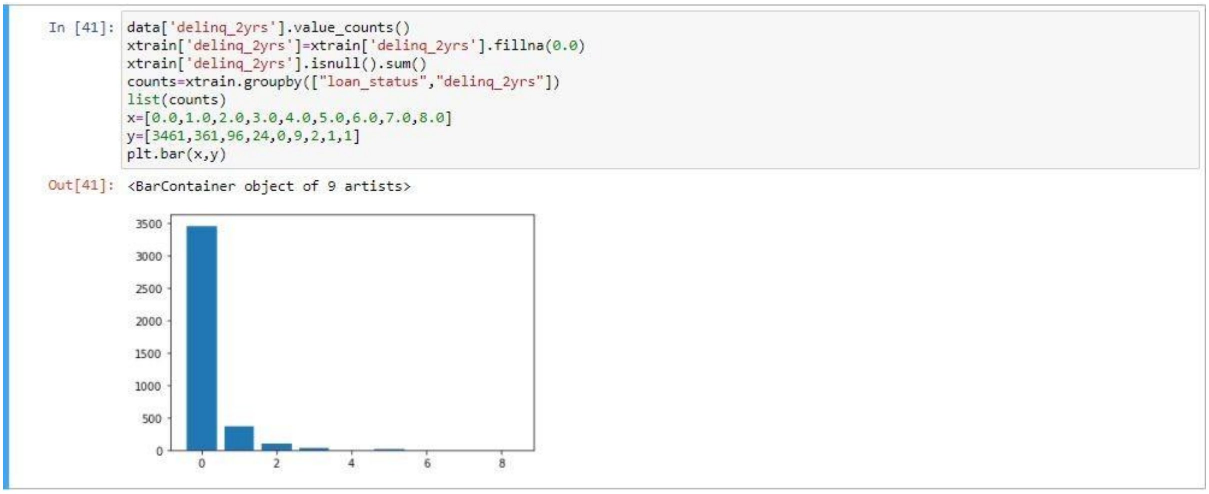


```
In [39]: counts=xtrain.groupby(["loan_status","verification_status"])
         list(counts)
         x=["Not verified","Source verified","verified"]
         y=[1519,1011,1425]
         plt.title("charged off based on verification status")
         plt.bar(x,y,color=["green","blue","red"])

Out[39]: <BarContainer object of 3 artists>
```

Charge off based on Employment Length:

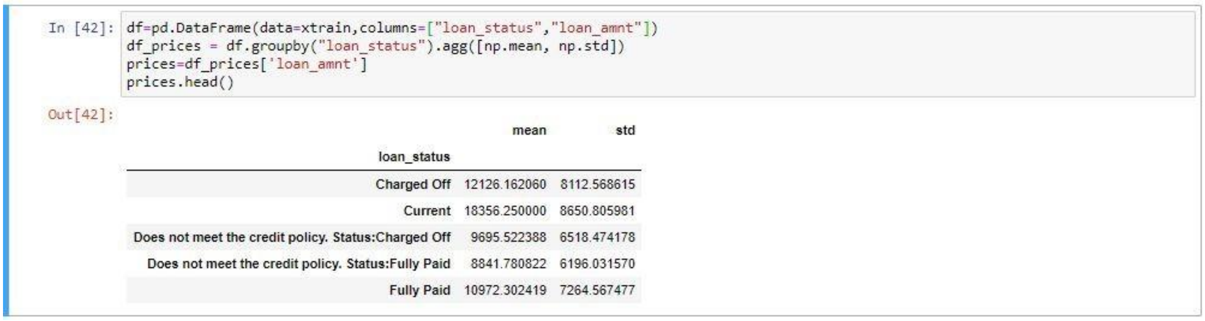


```
In [40]: counts=xtrain.groupby(["loan_status","emp_length"])
         list(counts)
         x=[0,1,2,3,4,5,6,7,8,9]
         y=[450,303,391,392,326,320,235,188,151,106]
         plt.title("charged off based on employement length")
         plt.bar(x,y)

Out[40]: <BarContainer object of 10 artists>
```



```
In [41]: data['delinq_2yrs'].value_counts()
         xtrain['delinq_2yrs']=xtrain['delinq_2yrs'].fillna(0.0)
         xtrain['delinq_2yrs'].isnull().sum()
         counts=xtrain.groupby(["loan_status","delinq_2yrs"])
         list(counts)
         x=[0.0,1.0,2.0,3.0,4.0,5.0,6.0,7.0,8.0]
         y=[3461,361,96,24,0,9,2,1,1]
         plt.bar(x,y)

Out[41]: <BarContainer object of 9 artists>
```

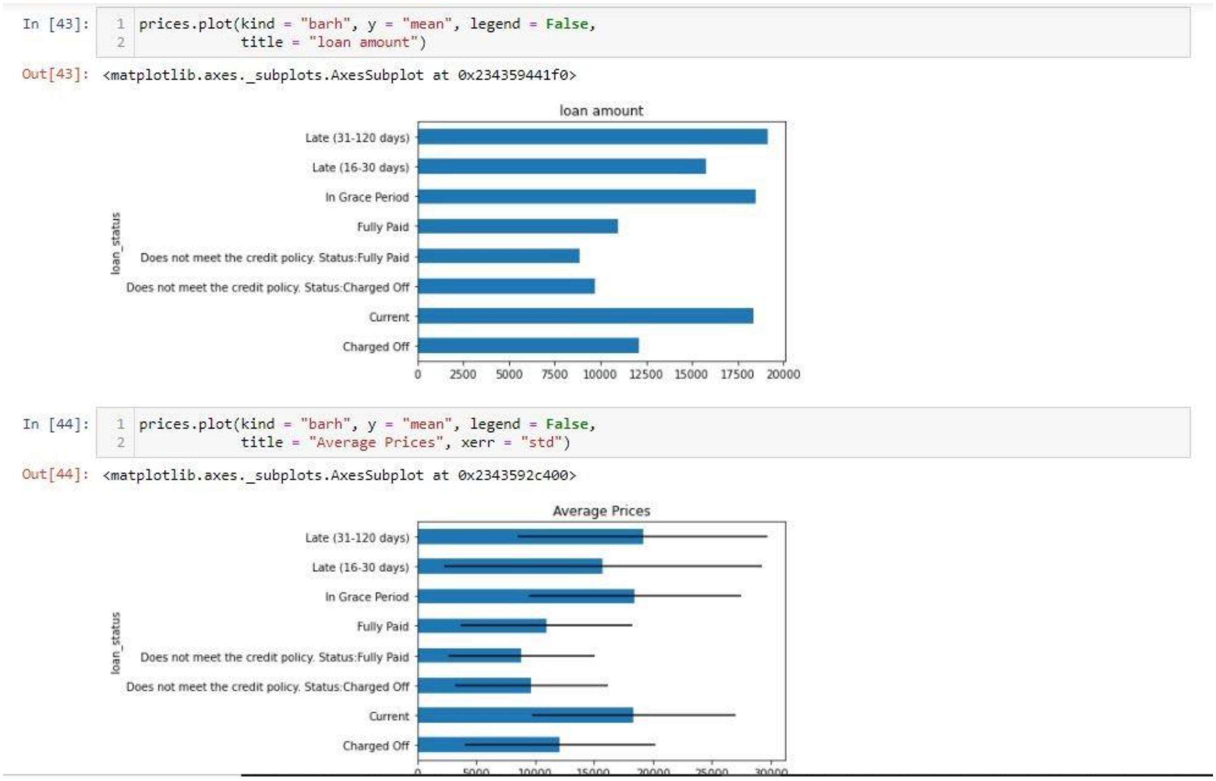
Mean and Standard deviation:



```
In [42]: df=pd.DataFrame(data=xtrain,columns=["loan_status","loan_amnt"])
         df_prices = df.groupby("loan_status").agg([np.mean, np.std])
         prices=df_prices['loan_amnt']
         prices.head()
```

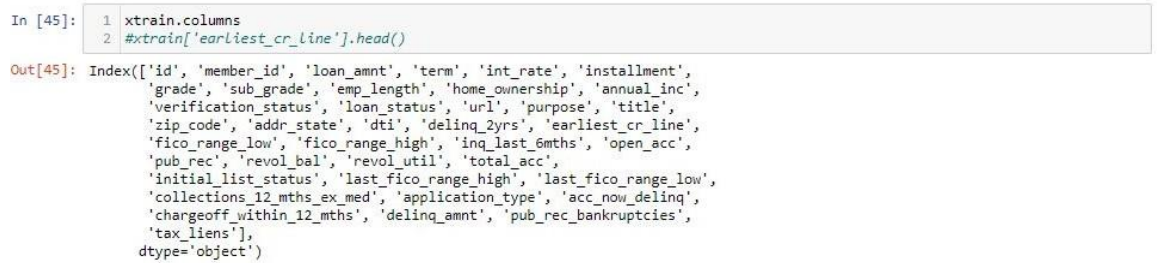
Out[42]:

loan_status	mean	std
Charged Off	12126.162060	8112.568615
Current	18356.250000	8650.805981
Does not meet the credit policy. Status:Charged Off	9695.522388	6518.474178
Does not meet the credit policy. Status:Fully Paid	8841.780822	6196.031570
Fully Paid	10972.302419	7264.567477

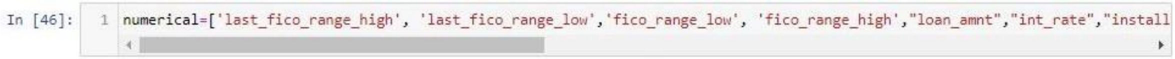


```
In [43]: prices.plot(kind = "barh", y = "mean", legend = False,
                     title = "loan amount")
Out[43]: <matplotlib.axes._subplots.AxesSubplot at 0x234359441f0>
```

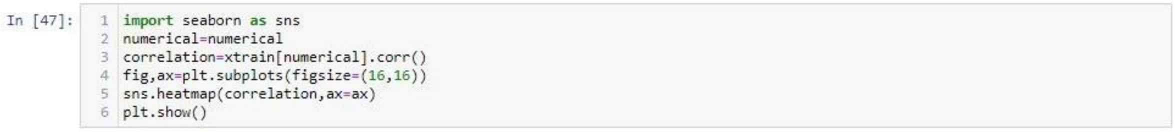
```
In [44]: prices.plot(kind = "barh", y = "mean", legend = False,
                     title = "Average Prices", xerr = "std")
Out[44]: <matplotlib.axes._subplots.AxesSubplot at 0x2343592c400>
```



```
In [45]: xtrain.columns
         #xtrain['earliest_cr_line'].head()
Out[45]: Index(['id', 'member_id', 'loan_amnt', 'term', 'int_rate', 'installment',
                'grade', 'sub_grade', 'emp_length', 'home_ownership', 'annual_inc',
                'verification_status', 'loan_status', 'url', 'purpose', 'title',
                'zip_code', 'addr_state', 'dti', 'delinq_2yrs', 'earliest_cr_line',
                'fico_range_low', 'fico_range_high', 'inq_last_6mths', 'open_acc',
                'pub_rec', 'revol_bal', 'revol_util', 'total_acc',
                'initial_list_status', 'last_fico_range_high', 'last_fico_range_low',
                'collections_12_mths_ex_med', 'application_type', 'acc_now_delinq',
                'chargeoff_within_12_mths', 'delinq_amnt', 'pub_rec_bankruptcies',
                'tax_liens'],
               dtype='object')
```

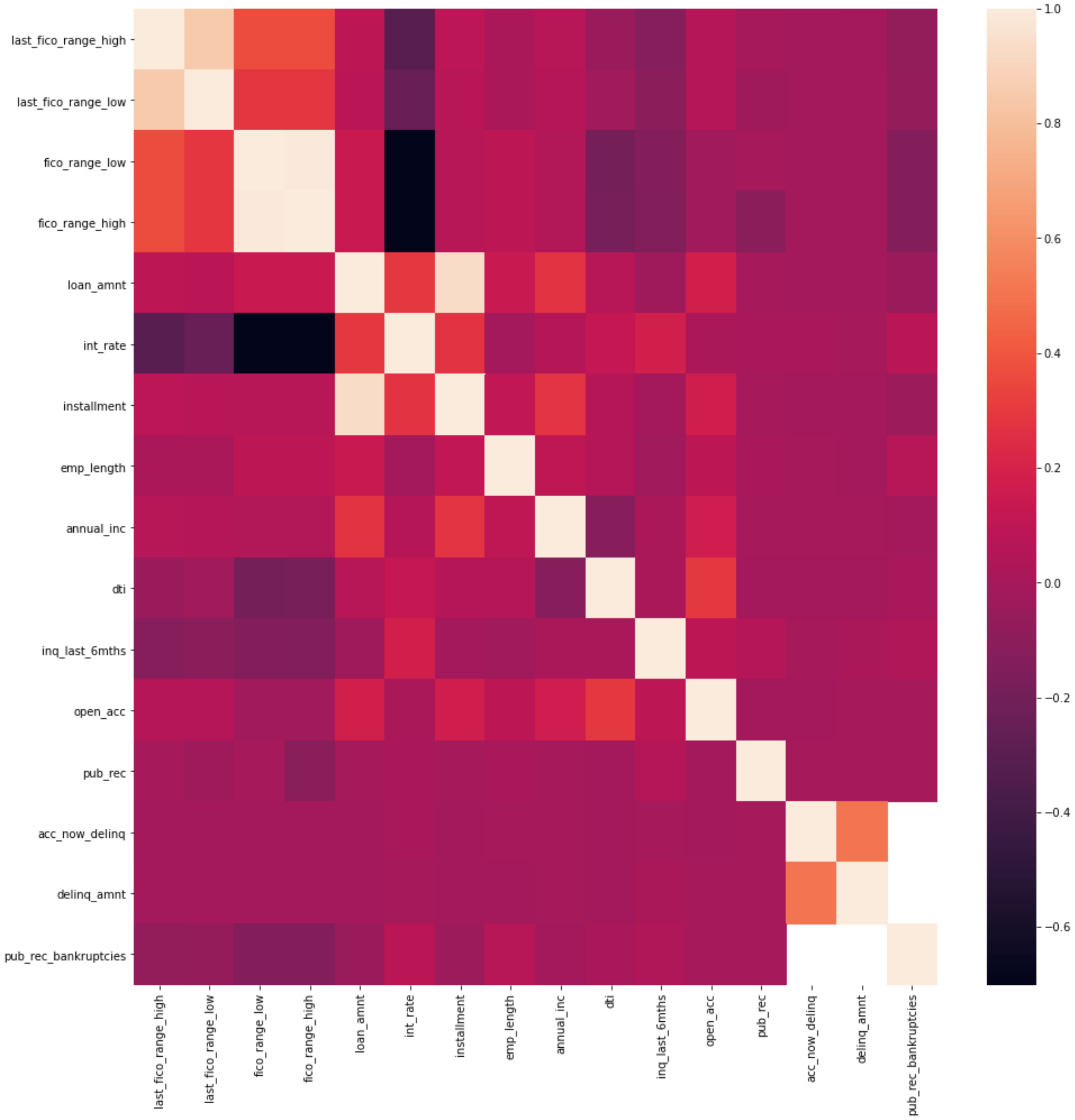


```
In [46]: numerical=['last_fico_range_high', 'last_fico_range_low','fico_range_low', 'fico_range_high',"loan_amnt","int_rate","install
```

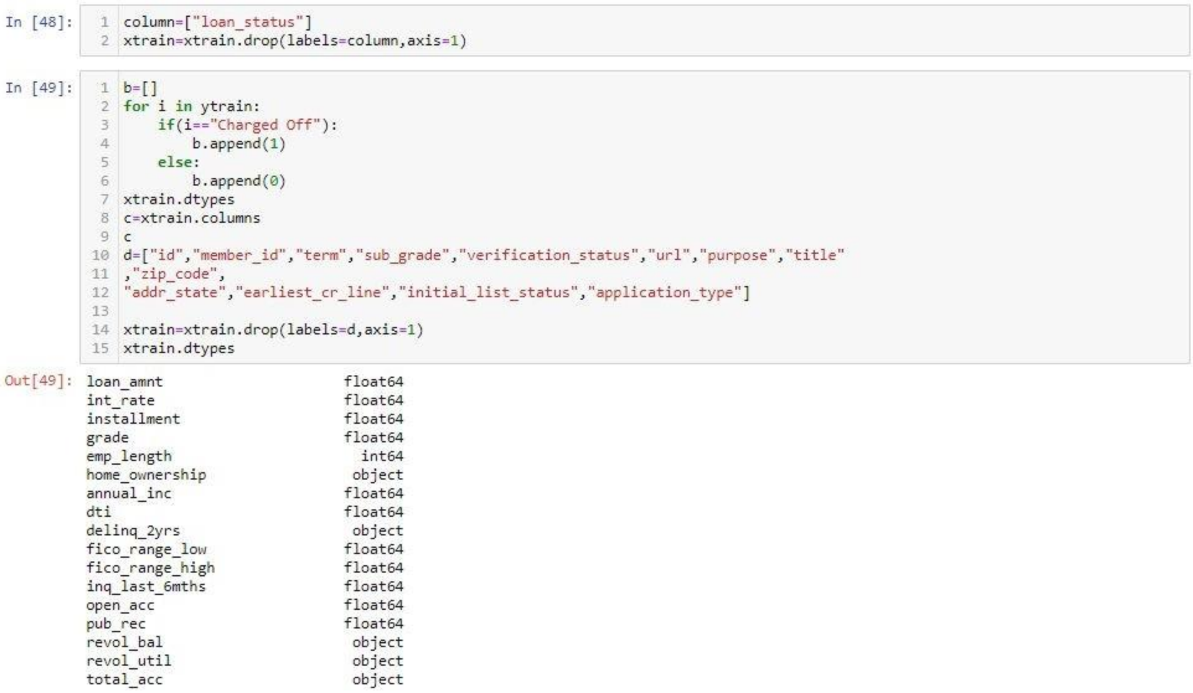


```
In [47]: import seaborn as sns
         numerical=numerical
         correlation=xtrain[numerical].corr()
         fig,ax=plt.subplots(figsize=(16,16))
         sns.heatmap(correlation,ax=ax)
         plt.show()
```

Heat map



last_fico_range_high
last_fico_range_low
fico_range_low
fico_range_high
loan_amnt
int_rate
installment
emp_length
annual_inc
dti
inq_last_6mths
open_acc
pub_rec
acc_now_delinq
delinq_amnt
pub_rec_bankruptcies
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6



In [48]:

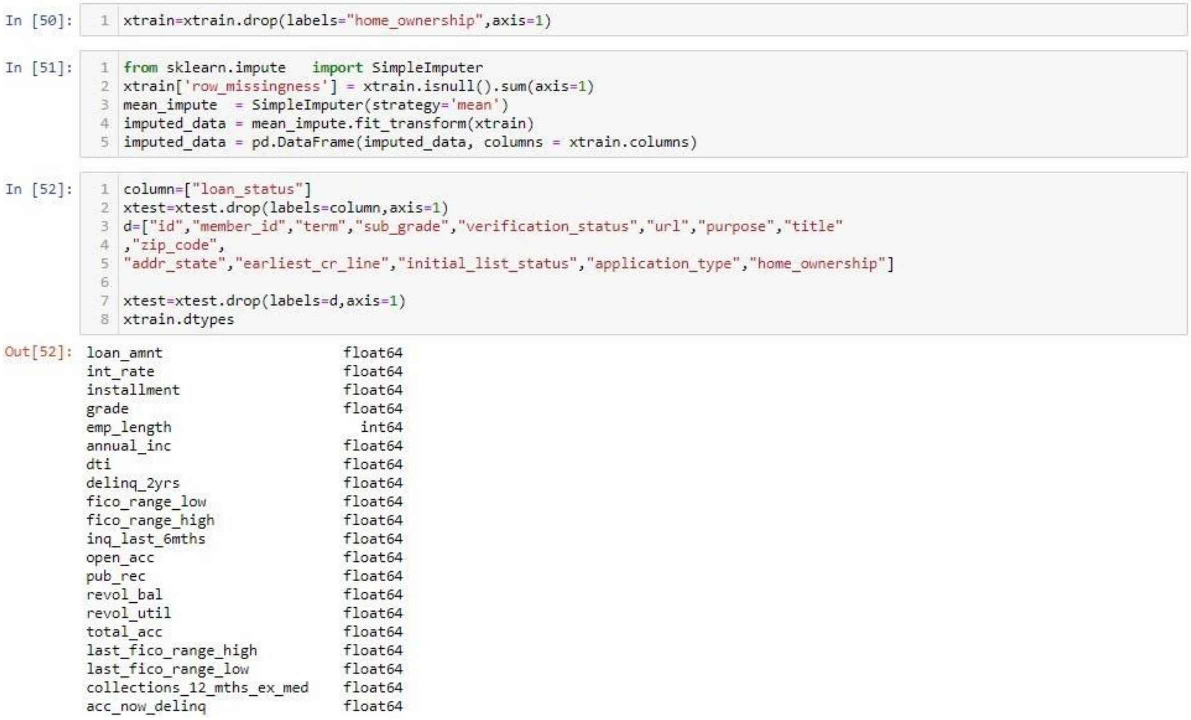
```
column=["loan_status"]
xtrain=xtrain.drop(labels=column,axis=1)
```

In [49]:

```
b=[]
for i in ytrain:
    if(i=="Charged Off"):
        b.append(1)
    else:
        b.append(0)
xtrain.dtypes
c=xtrain.columns
c
d=["id","member_id","term","sub_grade","verification_status","url","purpose","title"
,"zip_code",
"addr_state","earliest_cr_line","initial_list_status","application_type"]

xtrain=xtrain.drop(labels=d,axis=1)
xtrain.dtypes
```

Out[49]:

```
loan_amnt                  float64
int_rate                   float64
installment                float64
grade                      float64
emp_length                   int64
home_ownership              object
annual_inc                 float64
dti                        float64
delinq_2yrs                 object
fico_range_low             float64
fico_range_high            float64
inq_last_6mths             float64
open_acc                   float64
pub_rec                    float64
revol_bal                   object
revol_util                  object
total_acc                   object
```



In [50]:

```
xtrain=xtrain.drop(labels="home_ownership",axis=1)
```

In [51]:

```
from sklearn.impute   import SimpleImputer
xtrain['row_missingness'] = xtrain.isnull().sum(axis=1)
mean_impute  = SimpleImputer(strategy='mean')
imputed_data = mean_impute.fit_transform(xtrain)
imputed_data = pd.DataFrame(imputed_data, columns = xtrain.columns)
```

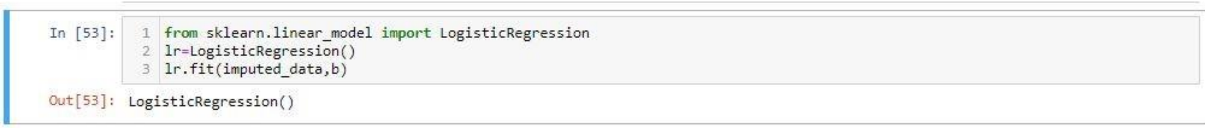
In [52]:

```
column=["loan_status"]
xtest=xtest.drop(labels=column,axis=1)
d=["id","member_id","term","sub_grade","verification_status","url","purpose","title"
,"zip_code",
"addr_state","earliest_cr_line","initial_list_status","application_type","home_ownership"]

xtest=xtest.drop(labels=d,axis=1)
xtrain.dtypes
```

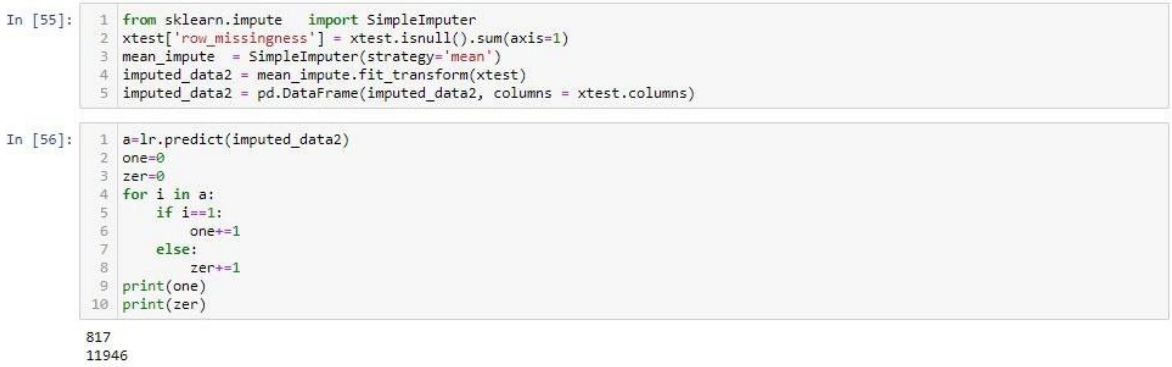
Out[52]:

```
loan_amnt                  float64
int_rate                   float64
installment                float64
grade                      float64
emp_length                   int64
annual_inc                 float64
dti                        float64
delinq_2yrs                float64
fico_range_low             float64
fico_range_high            float64
inq_last_6mths             float64
open_acc                   float64
pub_rec                    float64
revol_bal                  float64
revol_util                 float64
total_acc                  float64
last_fico_range_high       float64
last_fico_range_low        float64
collections_12_mths_ex_med float64
acc_now_delinq             float64
```

Logistic Regression:

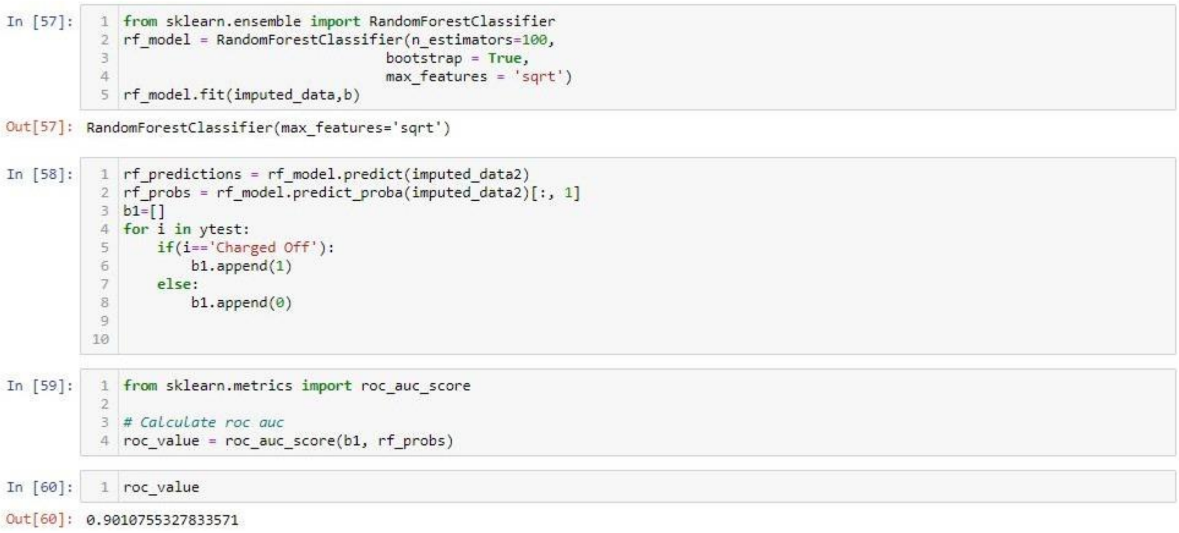


```
In [53]: from sklearn.linear_model import LogisticRegression
         lr=LogisticRegression()
         lr.fit(imputed_data,b)

Out[53]: LogisticRegression()
```

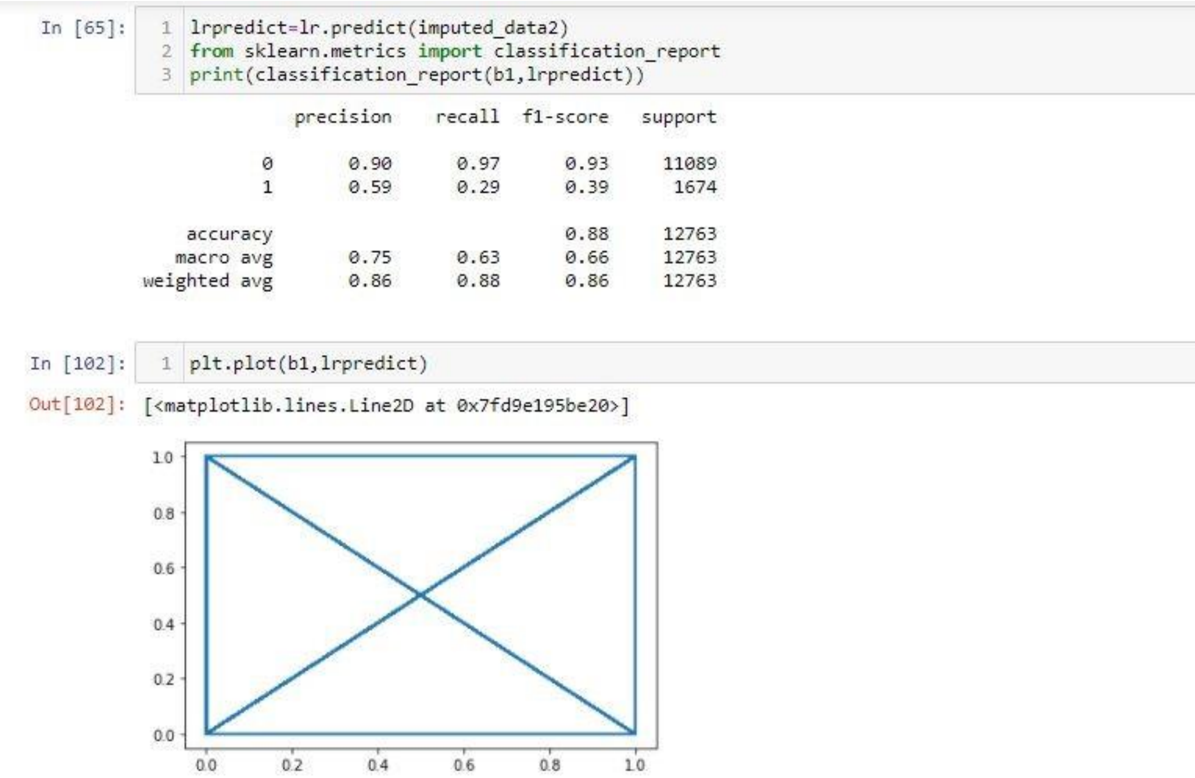


```
In [55]: from sklearn.impute   import SimpleImputer
         xtest['row_missingness'] = xtest.isnull().sum(axis=1)
         mean_impute  = SimpleImputer(strategy='mean')
         imputed_data2 = mean_impute.fit_transform(xtest)
         imputed_data2 = pd.DataFrame(imputed_data2, columns = xtest.columns)

In [56]: a=lr.predict(imputed_data2)
         one=0
         zer=0
         for i in a:
             if i==1:
                 one+=1
             else:
                 zer+=1
         print(one)
         print(zer)

817
11946
```

Random forest Classifier:



```
In [57]: from sklearn.ensemble import RandomForestClassifier
         rf_model = RandomForestClassifier(n_estimators=100,
                                           bootstrap = True,
                                           max_features = 'sqrt')
         rf_model.fit(imputed_data,b)

Out[57]: RandomForestClassifier(max_features='sqrt')

In [58]: rf_predictions = rf_model.predict(imputed_data2)
         rf_probs = rf_model.predict_proba(imputed_data2)[:, 1]
         b1=[]
         for i in ytest:
             if(i=='Charged Off'):
                 b1.append(1)
             else:
                 b1.append(0)



In [59]: from sklearn.metrics import roc_auc_score

         # Calculate roc auc
         roc_value = roc_auc_score(b1, rf_probs)

In [60]: roc_value

Out[60]: 0.9010755327833571
```



In [65]:

```
lrpredict=lr.predict(imputed_data2)
from sklearn.metrics import classification_report
print(classification_report(b1,lrpredict))
```

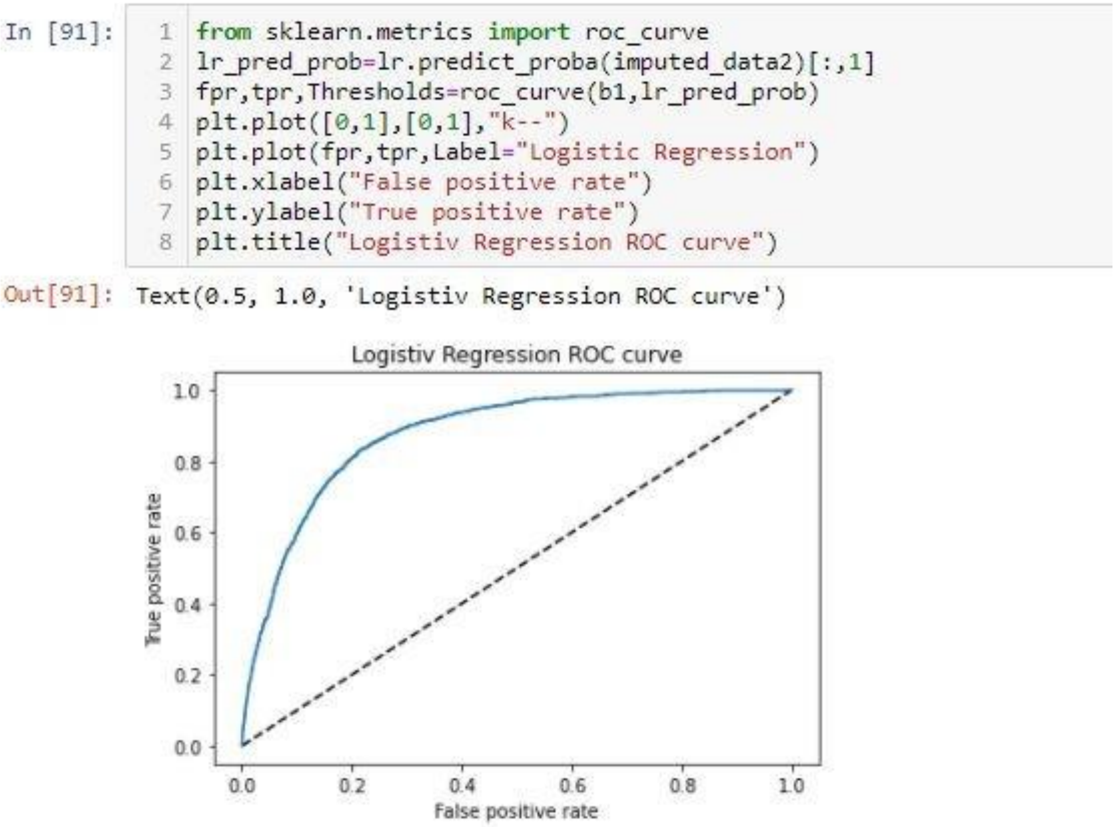
```
              precision    recall  f1-score   support

           0       0.90      0.97      0.93     11089
           1       0.59      0.29      0.39      1674

    accuracy                           0.88     12763
   macro avg       0.75      0.63      0.66     12763
weighted avg       0.86      0.88      0.86     12763
```

In [102]:

```
plt.plot(b1,lrpredict)
```

Out[102]:

```
[<matplotlib.lines.Line2D at 0x7fd9e195be20>]
```

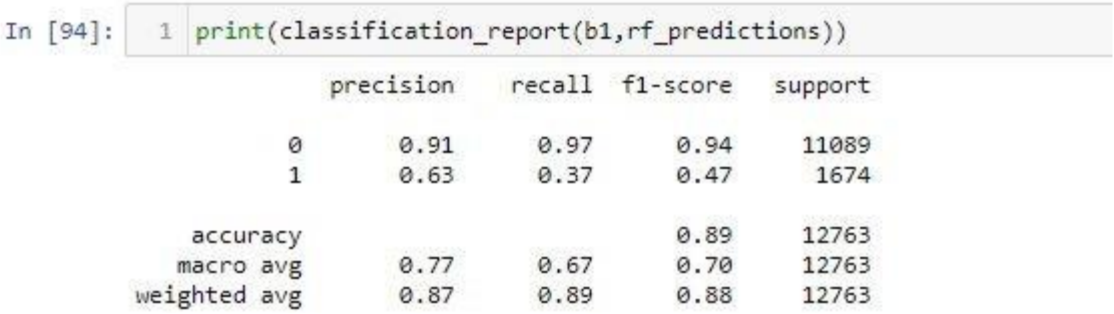


In [91]:

```
from sklearn.metrics import roc_curve
lr_pred_prob=lr.predict_proba(imputed_data2)[:,1]
fpr,tpr,Thresholds=roc_curve(b1,lr_pred_prob)
plt.plot([0,1],[0,1],"k--")
plt.plot(fpr,tpr,Label="Logistic Regression")
plt.xlabel("False positive rate")
plt.ylabel("True positive rate")
plt.title("Logistiv Regression ROC curve")
```

Out[91]:

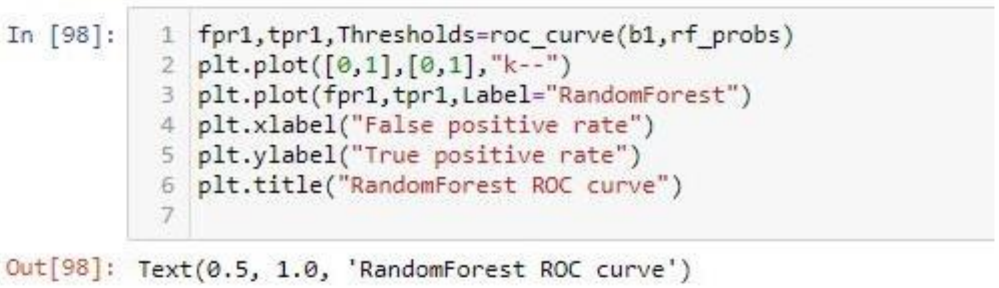
```
Text(0.5, 1.0, 'Logistiv Regression ROC curve')
```



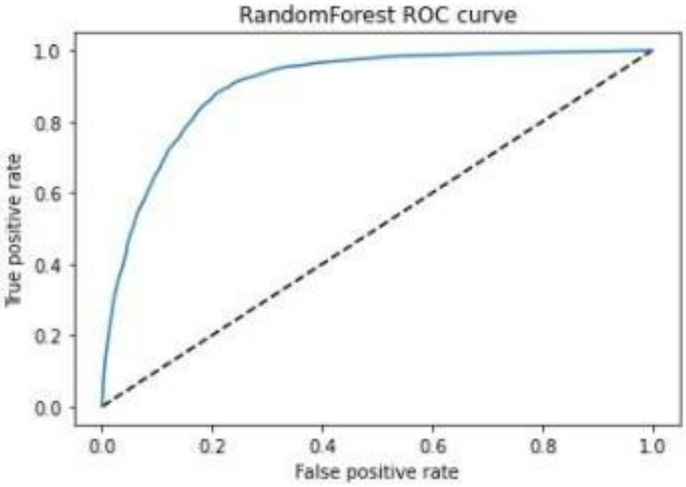
In [94]:

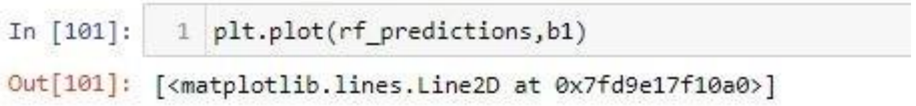
```
print(classification_report(b1,rf_predictions))
```

```
              precision    recall  f1-score   support

           0       0.91      0.97      0.94     11089
           1       0.63      0.37      0.47      1674

    accuracy                           0.89     12763
   macro avg       0.77      0.67      0.70     12763
weighted avg       0.87      0.89      0.88     12763
```

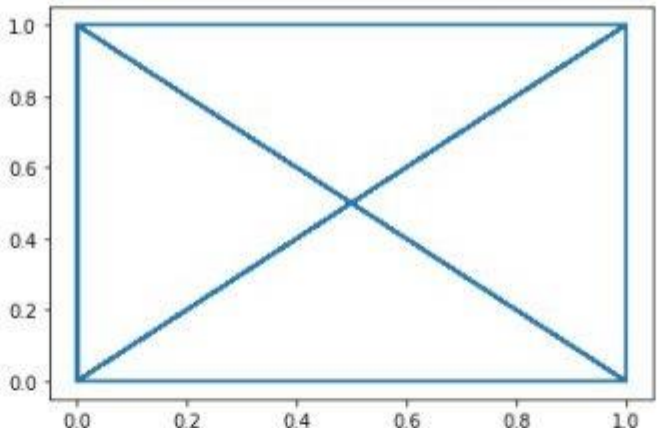


```
In [98]:  fpr1,tpr1,Thresholds=roc_curve(b1,rf_probs)
          plt.plot([0,1],[0,1],"k--")
          plt.plot(fpr1,tpr1,Label="RandomForest")
          plt.xlabel("False positive rate")
          plt.ylabel("True positive rate")
          plt.title("RandomForest ROC curve")

Out[98]: Text(0.5, 1.0, 'RandomForest ROC curve')
```





```
In [101]:  plt.plot(rf_predictions,b1)

Out[101]: [<matplotlib.lines.Line2D at 0x7fd9e17f10a0>]
```



Conclusion:

Online P2P loaning has acquired logical pertinence over the previous years. The accessibility of information about business sectors and exchanges permits specialists from various orders to examine the different determinants that assume a part during the time spent financing. A large part of the examination that has been essential for this writing survey zeroed in on factors that impact financing achievement and loan fees of credit demand. Albeit this field offers further exploration potential and experiences for sociology by and large, we see a bunch of extra perspectives that merit being examined in more detail later on. As we recognized in 0 particularly outer partners do assume a significant part in connection with loaning sites. However, the interior point of view merits further examination in the writing. In this regard for instance an assessment of plans of action, authoritative plan, and those components that produce achievement of P2P stages is a promising exploration theme. There is additionally the part of moneylender execution. We distinguished an absence of examination commitments that emphasis on factors that decide the accomplishment of banks concerning profit from speculation and low default rates. Besides, the impact of the borrowers' advance depictions on subsidizing achievement merits further examination. While a few papers researched the impacts of individual content attributes (M. Greiner and Wang, 2007; Larrimore et al., 2009; Lin, 2009), little exploration has been finished concerning the gathering of moneylenders' assumptions; the overall tone of the portrayal, and the decrease of data imbalances. At last, further assessments are important to distinguish similitudes and contrasts between the customary banking and the P2P loaning market. The two business sectors contrast particularly in the normal size of the financed advance, the screening cycle, just as the information and assets to assess and oversee chances. That sort of examination may explain whether results from previous exploration in the conventional financial market are material to the P2P loaning business sector and the other way around.

We have examined loan specialist conduct in the distributed (P2P) loaning market, where people bid on unstable microloans mentioned by other individual borrowers. Online P2P trades are developing, however loan specialists in this market are not expert financial backers. Likewise, moneylenders need to face enormous challenges since credits in P2P loaning are allowed without security. While the P2P loaning portions of the overall industry a few qualities of online business sectors regarding crowding conduct, it additionally has attributes that may debilitate it. This examination exactly researches crowding conduct in the P2P loaning market where apparently clashing conditions and highlights of grouping are available. Utilizing a huge example of day by day information from one of the biggest P2P loaning stages in Korea, we discover solid proof of grouping and its reducing minimal impact as offering progresses. We utilize a multinomial logit piece of the pie model in which pertinent factors from earlier investigations on P2P loaning are evaluated. With the danger virus delay considered, the threat of hazard infection in the entire organization increments under a similar disease media conditions. Albeit the correspondence of data is by all accounts sufficient in the organization loaning measure, both the honesty and genuine straightforwardness of data gave are moderately low in China's present organization loaning measure, because of the country's deficient credit framework and banks' undisclosed credit data framework. Additionally, data imbalance is likewise extreme, and the absence of data straightforwardness can straightforwardly prompt delays in hazard openness. Since both P2P loaning stages and borrowers can shroud reality purposely or inadvertently, for other P2P loaning stages, financial backers, and financing foundations included it turns out to be more hard to react when chance data is unveiled. Hence, administrative specialists should develop and scatter more itemized data divulgence rules to the two borrowers and stages in the organization loaning interaction to comply with and in this manner improve data straightforwardness. As such, they should take proper measures not exclusively to decrease the likelihood of hazard ahead of time yet in addition to permit other market members to get hazard data as right on time as could really be expected and subsequently have the option to attempt avoidance measures and react to chances in an ideal way.

We effectively planned and fostered a model for anticipating whether a forthcoming borrower would "charge-off utilizing arbitrary woodland, choice tree and boosting calculation. We had the option to foresee whether a borrower will fall in the low, medium, high loan fee range. Foreseeing whether a borrower would get the total applied advance sum from the financial backer was accomplished. Our normal precision was 74.6% while the best exactness accomplished was 81.4%.

References:

1. Ashta, A., & Assadi, D. (2009). An Analysis of European Online micro-lending Websites. EMN 6th Annual Conference (Vol. 33, pp. 4-28). Milan: Fundación Nantik Lum. Retrieved from http://www.european-microfinance.org/data/file /microlendingwebsites.doc
2. Barasinska, N. (2009). The Role of Gender in Lending Business : Evidence from an Online Market for Peer-to-Peer Lending. The New York Times. Berlin.
3. Berger, S. C., & Gleisner, F. (2009). Emergence of Financial Intermediaries in Electronic Markets : The Case of Online P2P Lending. BuR - Business Research, Official Open Access Journal of VHB, 2(1), 39-65.
4. Böhme, R., & Pötzsch, S. (2010). Privacy in online social lending. AAAI 2010 Spring Symposium on Intelligent Privacy Management (pp. 23–28). Palo Alto: Stanford University. Retrieved from http://www.aaai.org/ocs/index.php/SSS /S S S 10/paper/viewPDFInterstitial/1048/1472
5. Chemin, M., & De Laat, J. (2009). Can Warm Glow Alleviate Credit Market Failure? Evidence from Online Peer-to-Peer Lenders. papers.ssrn.com. Montreal. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1461438
6. Chen, K. Y., Golder, S., Hogg, T., & Zenteno, C. (2008). How Do People Respond to Reputation: Ostracize, Price Discriminate or Punish? 2nd Intl. Workshop on Hot Topics in Web Systems and Technologies (p. 6). Palo Alto, CA: Hewlett-Packard Labs. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.157.9701&amp;rep=rep1&amp;type=pdf

7. Collier, B., & Hampshire, R. (2010). Sending Mixed Signals: Multilevel Reputation Effects in Peer-to-Peer Lending Markets. ACM Conference on Computer Supported Cooperative Work (pp. 1-10). Savannah, Georgia: ACM.
8. Dhand, H., Mehn, G., Dickens, D., Patel, A., Lakra, D., & McGrath, A. (2008). Internet Based Social Lending. Communications of the IBIMA, 2, 109-114. Retrieved from http://www.doaj.org/doaj?func=abstract&amp;id=564232
9. Everett, C. R. (2010). Group membership , relationship banking and loan default risk: the case of online social lending. Group. West Lafayette, IN. Retrieved from Available at SSRN: http://ssrn.com/abstract=1114428
10. Freedman, S., & Jin, G.Z. (2008). http://en.ccer.edu.cn/download/6641-1.pdf
11. https://www.kaggle.com/wendykan/lending-club-loan-data
12. https://www.lendingclub.com/auth/login?login_url=%2Finfo%2Fdownload-data.action
13. https://ieeexplore.ieee.org/document/7803017/
14. https://www.researchgate.net/publication/327836640_Peer-to-Peer_Lending_Business_Model_Analysis_and_the_Platform_Dilemma_in_International_Journal_of_Finance_Economics_and_Trade_IJFET_submitted_August_1st_2018_Accepted_Sept_24th
15. Freeman, R. E. (2010). Strategic management: A stakeholder approach (p. 276). Boston: Cambrigde University Press.
16. https://www.semanticscholar.org/paper/Credit-RiskAnalysis-in-Peerto-Peer-Lending-SystemVinodNatarajan/c0d182ab22921c2f914770dd83a195c8c2202d6f
17. http://iosrjournals.org/iosrjbm/papers/IESMCRC/Volume%201/79.85.pdf
18. https://www.adb.org/sites/default/files/publication/478611/adbi-wp912.pdf
19. https://ieeexplore.ieee.org/document/7803017
20. https://www.lendingclub.com/

Appendix A -Coding

```
import warnings
warnings.simplefilter('ignore')
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline

from google.colab import drive
drive.mount('/content/gdrive')
data= pd.read_csv('/content/gdrive/My Drive/lending_club_loans.csv')
data.head()

pd.options.display.max_rows=150
missing=data.isnull().sum()
missing_ratio=missing/len(data)
missing_ratio=missing_ratio.reset_index()
missing_ratio=missing_ratio.rename(columns={'index':'feature',0:'missing rati
o'})
missing_ratio=missing_ratio.sort_values(by='missing ratio',ascending=False)
missing_ratio[missing_ratio['missing ratio']>=0.2]['feature']

columns=['annual_inc_joint','mo_sin_rcnt_rev_tl_op','inq_fi','total_cu_tl','m
o_sin_old_il_acct','bc_util','bc_open_to_buy','avg_cur_bal','acc_open_past_24
mths','inq_last_12m','total_rev_hi_lim','all_util','max_bal_bc','open_rv_24m'
,'open_rv_12m','il_util','total_bal_il','mths_since_rcnt_il','open_il_24m','o
pen_il_12m','open_il_6m','open_acc_6m','tot_cur_bal','tot_coll_amt','verifica
tion_status_joint','mo_sin_old_rev_tl_op','mo_sin_rcnt_tl','mths_since_last_m
ajor_derog','mort_acc','total_bc_limit','total_bal_ex_mort','tot_hi_cred_lim'
,'percent_bc_gt_75','pct_tl_nvr_dlq','num_tl_op_past_12m','num_tl_90g_dpd_24m
','num_tl_30dpd','num_tl_120dpd_2m','num_sats','num_rev_tl_bal_gt_0','num_rev
_accts','num_op_rev_tl','num_il_tl','num_bc_tl','num_bc_sats','num_actv_rev_t
l','num_actv_bc_tl','num_accts_ever_120_pd','mths_since_recent_revol_delinq',
'mths_since_recent_inq','mths_since_recent_bc_dlq','mths_since_recent_bc','dt
i_joint','total_il_high_credit_limit','next_pymnt_d','mths_since_last_record'
,'mths_since_last_delinq','desc']

data=data.drop(labels=columns,axis=1)
data.head(5)


ongo_columns=['funded_amnt','funded_amnt_inv','issue_d','pymnt_plan','out_prn
cp','out_prncp_inv','total_pymnt','total_pymnt_inv','total_rec_prncp','policy
_code','total_rec_int',
              'total_rec_late_fee','recoveries','collection_recovery_fee','las
t_pymnt_d','last_pymnt_amnt','last_credit_pull_d']


data=data.drop(labels=ongo_columns,axis=1)
data.info(verbose=False)
```

```
data['loan_status'].value_counts()

data.shape

data.columns

target=data['loan_status']
from sklearn.model_selection import train_test_split
xtrain,xtest,ytrain,ytest=train_test_split(data,target,test_size=0.3)

feature='home_ownership'
xtrain[feature].value_counts()
xtrain[feature]=xtrain[feature].fillna(value="OTHER")
xtrain[feature].value_counts()

x=['rent','mortgage','own','other']
y=[141471,13204,2291,136]
plt.bar(x,y)

print(xtrain['int_rate'].values[:3])

def str_parser(x):
    if '%' in str(x):
        return(float(str(x).replace('%','')))
    return x

xtrain['int_rate']=xtrain['int_rate'].apply(str_parser)
xtrain['revol_util']=xtrain['revol_util'].apply(str_parser)
xtest['int_rate']=xtest['int_rate'].apply(str_parser)
xtest['revol_util']=xtest['revol_util'].apply(str_parser)

xtrain=xtrain.drop(labels='emp_title',axis=1)
xtest=xtest.drop(labels='emp_title',axis=1)

emp_length_mode=xtrain['emp_length'].mode()[0]
xtrain['emp_length']=xtrain['emp_length'].fillna(value=emp_length_mode)
xtest['emp_length']=xtest['emp_length'].fillna(value=emp_length_mode)

xtest['title']=xtest['title'].fillna(value="missing")
xtrain['title']=xtrain['title'].fillna(value="missing")

from sklearn import preprocessing
grade_encoder=preprocessing.LabelEncoder()
data['grade']=xtrain['grade'].fillna('G')
data['grade'].value_counts()
xtrain['grade']=xtrain['grade'].replace(['A','B','C','D','E','F','G'],[0,1,2,
3,4,5,6])
xtest['grade']=xtest['grade'].replace(['A','B','C','D','E','F','G'],[0,1,2,3,
4,5,6])
xtrain['grade'].head()

xtrain['emp_length'].value_counts()
data['emp_length'].value_counts()
```

```
g={'< 1 year':0,'1 year':1,'2 years':2,'3 years':3,'4 years':4,'5 years':5,'6 years':6,'7 years':7,'8 years':8,'9 years':9,'10+ years':10}

xtrain['emp_length']=xtrain['emp_length'].apply(lambda x:g[x])

xtest['emp_length']=xtest['emp_length'].apply(lambda x:g[x])

xtrain['title']=xtrain['title'].apply(str.lower)
xtest['title']=xtest['title'].apply(str.lower)

lists=['debt consolidation','credit card refinancing','business','vacation','home improvement','majar purchase','medical expense','car financing','moving and relocation','home buying','green loan','consolidation']

xtrain['term'].value_counts()
xtrain['term'].replace(["36 months","64 months"],[36,64],inplace=True)
xtrain['term'].value_counts()

data['loan_status'].value_counts()

data['term'].value_counts()

x=['36','60']
y=[31534,11001]
plt.bar(x,y)

df=pd.DataFrame(data,columns=["loan_status","term"])
df.head()

group=df["loan_status"].groupby(df["term"])
group

counts=xtrain.groupby(["loan_status","term"])
a=len(list(counts)[0][1]) #charged off 36 month
a

b=len(list(counts)[1][1])#charged off 60 month
b

xtrain["term"].value_counts()
xtrain["term"].head()

c=22067-2257
d=7707-1698
z=["charged off","remaining"]
x=[a,c]
y=[b,d]
plt.title("loan status of 36 months term")
plt.bar(z,x,color=["red","green"])

plt.title("loan status of 60 months term")
plt.bar(z,y,color=["red","green"])
```

```
counts=xtrain.groupby(["loan_status","home_ownership"])
list(counts)
#charged off according to home ownership
x=["mortgage","rent","own","other"]
y=[1624,2001,214,16]
plt.title("charged off based on home ownership")
plt.bar(x,y,color=["yellow","orange","blue","green"])

counts=xtrain.groupby(["loan_status","verification_status"])
list(counts)
x=["Not verified","Source verified","verified"]
y=[1519,1011,1425]
plt.title("charged off based on verification status")
plt.bar(x,y,color=["green","blue","red"])

counts=xtrain.groupby(["loan_status","emp_length"])
list(counts)
x=[0,1,2,3,4,5,6,7,8,9]
y=[450,303,391,392,326,320,235,188,151,106]
plt.title("charged off based on employement length")
plt.bar(x,y)

data['delinq_2yrs'].value_counts()
xtrain['delinq_2yrs']=xtrain['delinq_2yrs'].fillna(0.0)
xtrain['delinq_2yrs'].isnull().sum()
counts=xtrain.groupby(["loan_status","delinq_2yrs"])
list(counts)
x=[0.0,1.0,2.0,3.0,4.0,5.0,6.0,7.0,8.0]
y=[3461,361,96,24,0,9,2,1,1]
plt.bar(x,y)

df=pd.DataFrame(data=xtrain,columns=["loan_status","loan_amnt"])
df_prices = df.groupby("loan_status").agg([np.mean, np.std])
prices=df_prices['loan_amnt']
prices.head()

prices.plot(kind = "barh", y = "mean", legend = False, title = "loan amount")

prices.plot(kind = "barh", y = "mean", legend = False,
            title = "Average Prices", xerr = "std")

xtrain.columns
#xtrain['earliest_cr_line'].head()

numerical=['last_fico_range_high', 'last_fico_range_low','fico_range_low', 'f
ico_range_high',"loan_amnt","int_rate","installment","emp_length","annual_inc
","dti","delinq_2yrs","inq_last_6mths","open_acc","pub_rec",'revol_bal', 'rev
ol_util', 'total_acc','acc_now_delinq', 'delinq_amnt', 'pub_rec_bankruptcies'
,]

import seaborn as sns
numerical=numerical
correlation=xtrain[numerical].corr()
```

```
fig,ax=plt.subplots(figsize=(16,16))
sns.heatmap(correlation,ax=ax)
plt.show()

column=["loan_status"]
xtrain=xtrain.drop(labels=column,axis=1)

b=[]
for i in ytrain:
    if(i=="Charged Off"):
        b.append(1)
    else:
        b.append(0)
xtrain.dtypes
c=xtrain.columns
c
d=["id","member_id","term","sub_grade","verification_status","url","purpose",
"title"
,"zip_code",
"addr_state","earliest_cr_line","initial_list_status","application_type"]
xtrain=xtrain.drop(labels=d,axis=1)
xtrain.dtypes

xtrain=xtrain.drop(labels="home_ownership",axis=1)

from sklearn.impute   import SimpleImputer
xtrain['row_missingness'] = xtrain.isnull().sum(axis=1)
mean_impute  = SimpleImputer(strategy='mean')
imputed_data = mean_impute.fit_transform(xtrain)
imputed_data = pd.DataFrame(imputed_data, columns = xtrain.columns)

column=["loan_status"]
xtest=xtest.drop(labels=column,axis=1)
d=["id","member_id","term","sub_grade","verification_status","url","purpose",
"title","zip_code","addr_state","earliest_cr_line","initial_list_status","app
lication_type","home_ownership"]

xtest=xtest.drop(labels=d,axis=1)
xtrain.dtypes

from sklearn.linear_model import LogisticRegression
lr=LogisticRegression()
lr.fit(imputed_data,b)

from sklearn.impute   import SimpleImputer
xtest['row_missingness'] = xtest.isnull().sum(axis=1)
mean_impute  = SimpleImputer(strategy='mean')
imputed_data2 = mean_impute.fit_transform(xtest)
imputed_data2 = pd.DataFrame(imputed_data2, columns = xtest.columns)
```

```
a=lr.predict(imputed_data2)
one=0
zer=0
for i in a:
    if i==1:
        one+=1
    else:
        zer+=1
print(one)
print(zer)

from sklearn.ensemble import RandomForestClassifier
rf_model = RandomForestClassifier (n_estimators=100,bootstrap = True, max_fea
tures = 'sqrt')
rf_model.fit(imputed_data,b)

rf_predictions = rf_model.predict(imputed_data2)
rf_probs = rf_model.predict_proba(imputed_data2)[:,1]
b1=[]
for i in ytest:
    if (i=='Charged off'):
        b1.append(1)
    else:
        b1.append(0)


from sklearn.metrics import roc_auc_score
#Calculate roc auc
#roc_value = roc_auc_score(b1, rf_probs)
try:
     roc_auc_score(b1, rf_probs)
except ValueError:
    pass


lrpredict=lr.predict(imputed_data2)
from sklearn.metrics import classification_report
# Accuracy Of Logistic regression
print(classification_report(b1, lrpredict))

plt.plot(b1,lrpredict)

from sklearn.metrics import roc_curve
lr_pred_prob=lr.predict_proba(imputed_data2)[:,1]
fpr, tpr, Thresholds=roc_curve(b1, lr_pred_prob)
plt.plot([0,1],[0,1],"k--")
plt.plot(fpr, tpr, Label="Logistic Regression")
plt.xlabel("False positive rate")
plt.ylabel("True positive rate")
plt.title("Logistiv Regressive ROC curve")

# Accuracy Of Random Forest
print(classification_report(b1,rf_predictions))
```

```
fpr1, tpr1,Thresholds =roc_curve(b1,rf_probs)
plt.plot([0,1],[0,1],"k--")
plt.plot(fpr1, tpr1,Label="RandomForest")
plt.xlabel("False positive rate")
plt.ylabel("True positive mate")
plt.title("RandomForest ROC curve")

print(rf_predictions,b1)

plt.plot(rf_predictions,b1)
```