Capstone Project

Team 2 : Loan Default Prediction

Suraj Pandey

Let’s Catch The Defaulters

1. Defining problem statement
2. EDA and feature engineering
3. Feature Selection
4. Preparing dataset for modeling
5. Model Validation and Selection

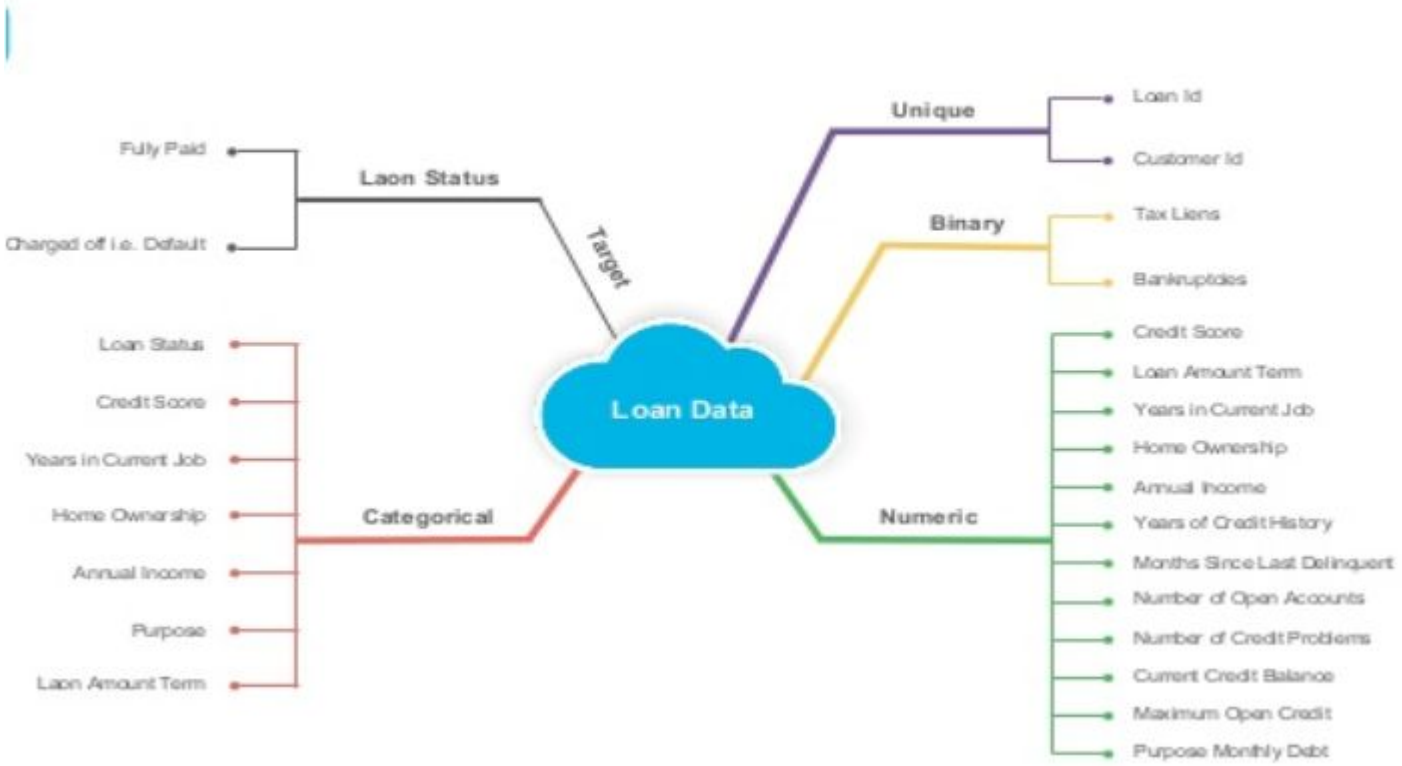
The Dilemma

Lending Club is the world's largest online marketplace connecting borrowers and investors. An inevitable outcome of lending is default by borrowers. The idea of this tutorial is to create a predictive model that identifies applicants who are relatively risky for a loan. In order to accomplish this, I organized the whole series into four parts as follows:

Data Pipeline

- **Data processing-1**: In this first part we've removed unnecessary features. There were nearly Many columns with all Null values.
- **Data processing-2**: In this part, we manually go through each features selected from part 1. This is the most time-consuming part, but worth it for a better model.
- **EDA**: In in this part, we do some exploratory data analysis (EDA) on the features selected in part-1 and 2.
- **Create a model**: Finally, In this last but not the last part, we create models. Creating a model is also not an easy task. It's also an iterative process. we show how to start with a with a simple model, then slowly add complexity for better performance.

Data Summary



Data Summary

loan_amnt: The amount the borrower promises to repay, as set forth in the loan contract.The loan amount may exceed the original amount requested by the borrower if he or she elects to include points and other upfront costs in the loan.

Purpose:The purpose of the loan is used by the lender to make decisions on the risk and may even impact the interest rate that is offered. For example, if an applicant is refinancing a mortgage after having taken some cash out, the lender might consider that an increase in risk and increase the interest rate that is offered or add additional conditions. Loan purpose is important to the process of obtaining mortgages or business loans that are connected with specific types of business activities.

Data Summary

Dti: Your debt-to-income ratio is all your monthly debt payments divided by your gross monthly income. This number is one way lenders measure your ability to manage the monthly payments to repay the money you plan to borrow.

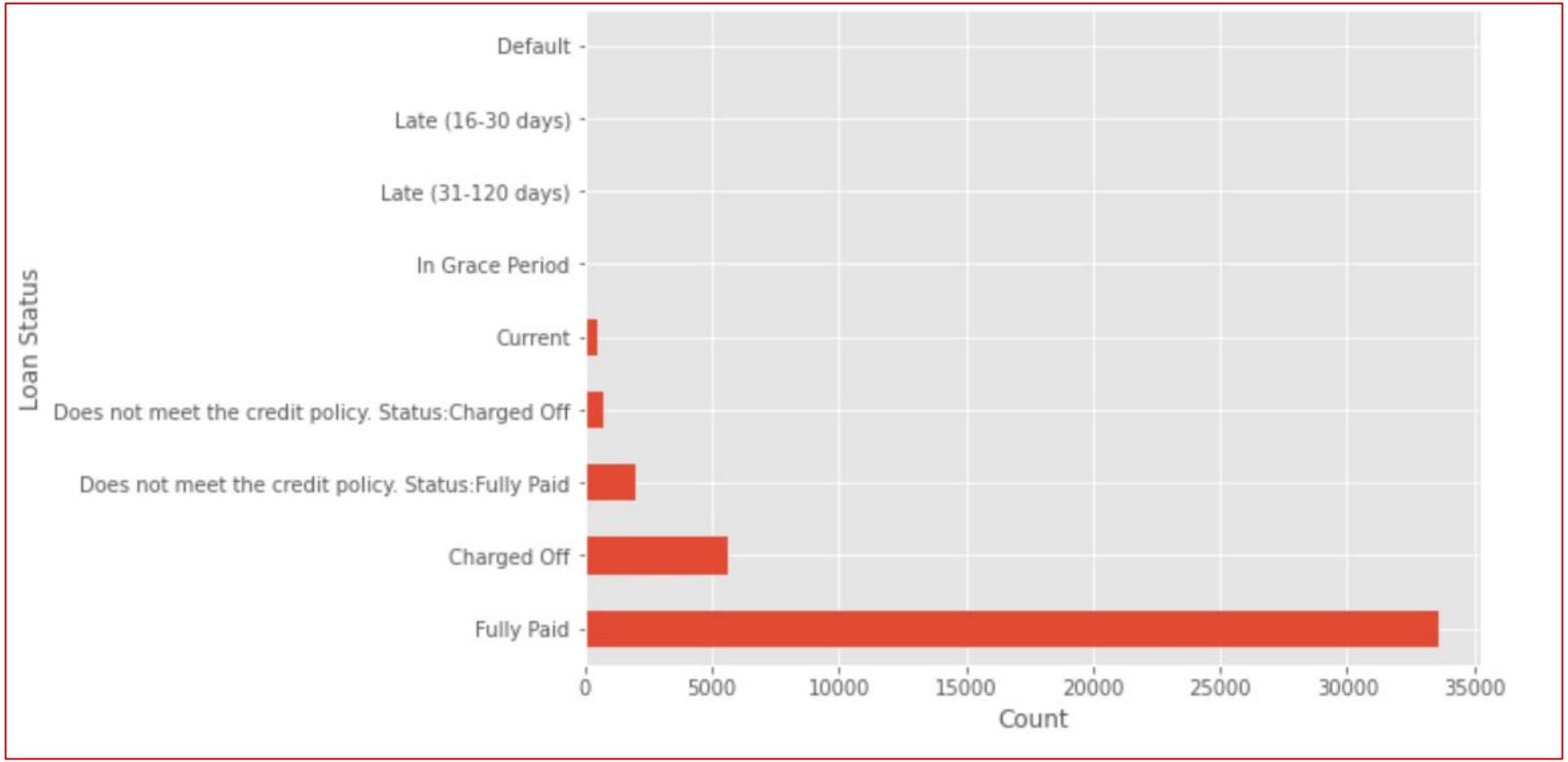
Fico_range: A FICO score is a credit score created by the Fair Isaac Corporation (FICO).[1] Lenders use borrowers' FICO scores along with other details on borrowers' credit reports to assess credit risk and determine whether to extend credit. FICO scores take into account data in five areas to determine creditworthiness: payment history, current level of indebtedness, types of credit used, length of credit history, and new credit accounts.

Data Summary

Delinq_amnt : Delinquent describes something or someone who fails to accomplish that which is required by law, duty, or contractual agreement, such as the failure to make a required payment or perform a particular action.

Int_rate: An interest rate is defined as the proportion of an amount loaned which a lender charges as interest to the borrower, normally expressed as an annual percentage. It is the rate a bank or other lender charges to borrow its money, or the rate a bank pays its savers for keeping money in an account.

Define Dependent Variable



Define Dependent Variable

Fully Paid->Loan has been fully paid off.

Charged off->Loan for which there is no longer a reasonable expectation of further payments.

Does not meet the credit policy. Status:Fully Paid--> While the loan was paid off, the loan application today would no longer meet the credit policy and wouldn't be approved on to the marketplace.

Does not meet the credit policy. Status:charged off->While the loan was charged off, the loan application today would no longer meet the credit policy and wouldn't be approved on to the marketplace.

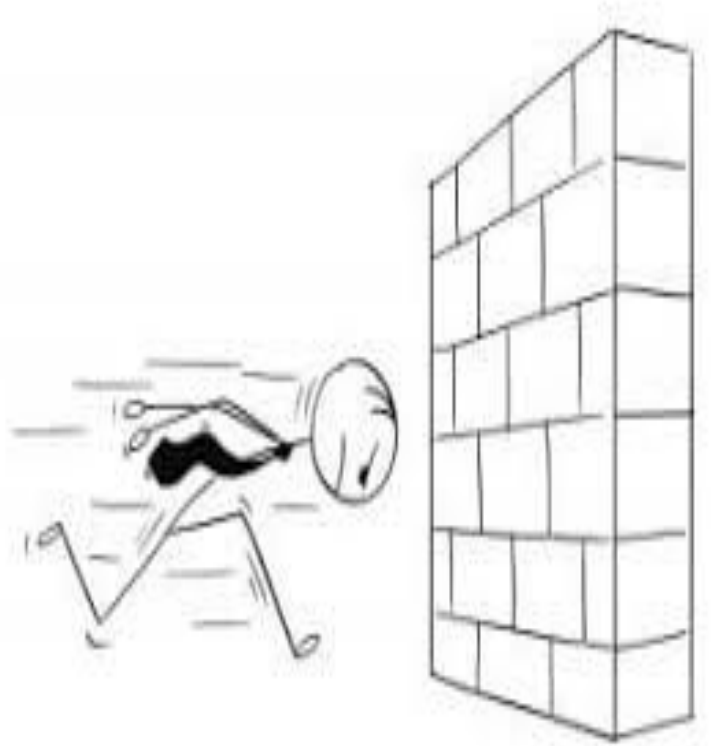
Current->Loan is up to date on current payments.,

In Grace period->The loan is past due but still in the grace period of 15 days.

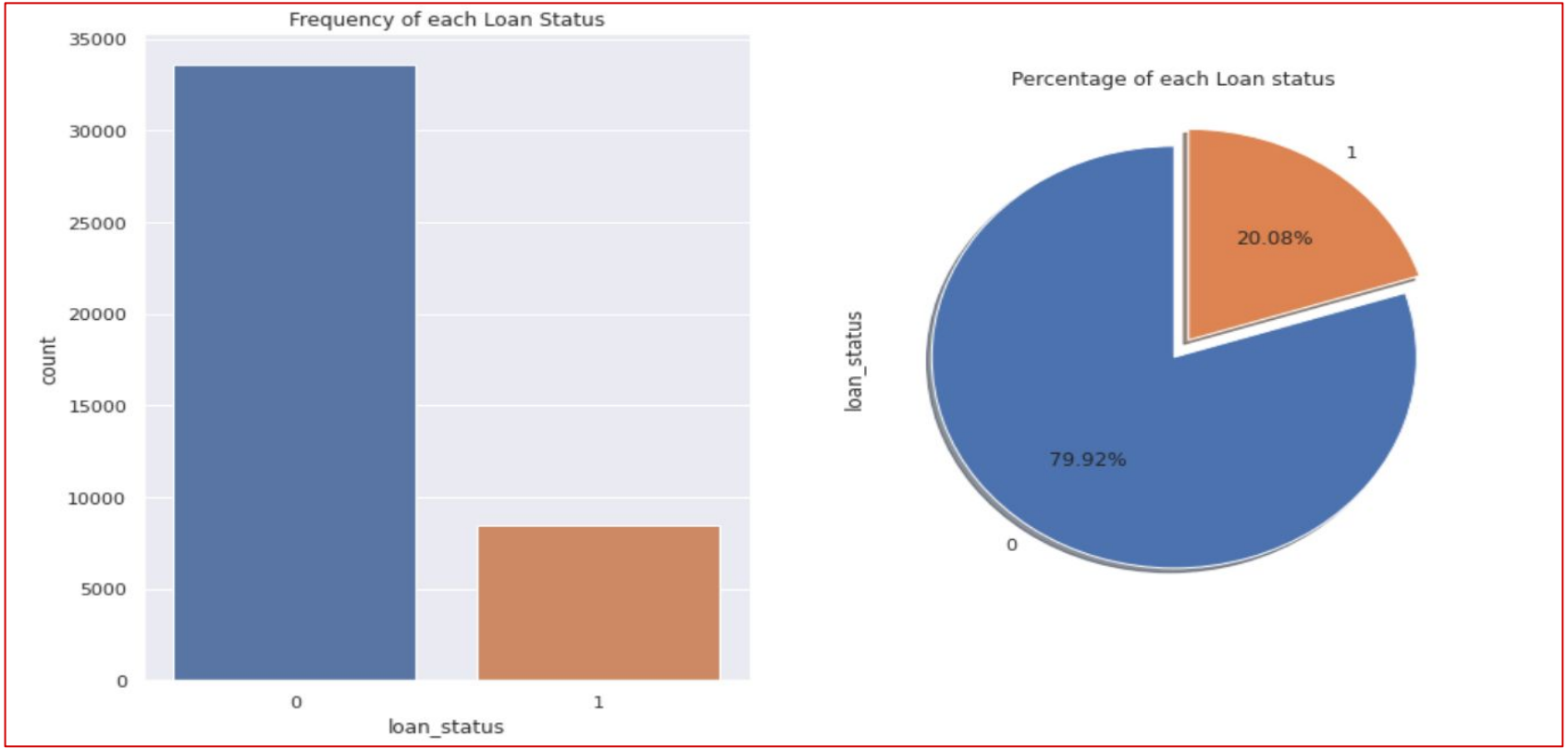
Late(30-120)->Loan hasn't been paid in 31 to 120 days (late on the current payment).

Late(16-30)->Loan hasn't been paid in 16 to 30 days (late on the current payment).

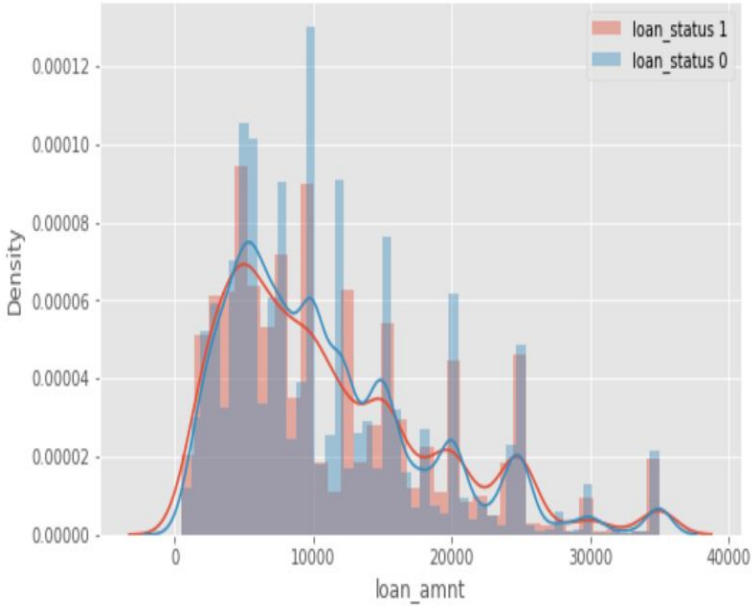
Default->Loan is defaulted on and no payment has been made for more than 121 days.

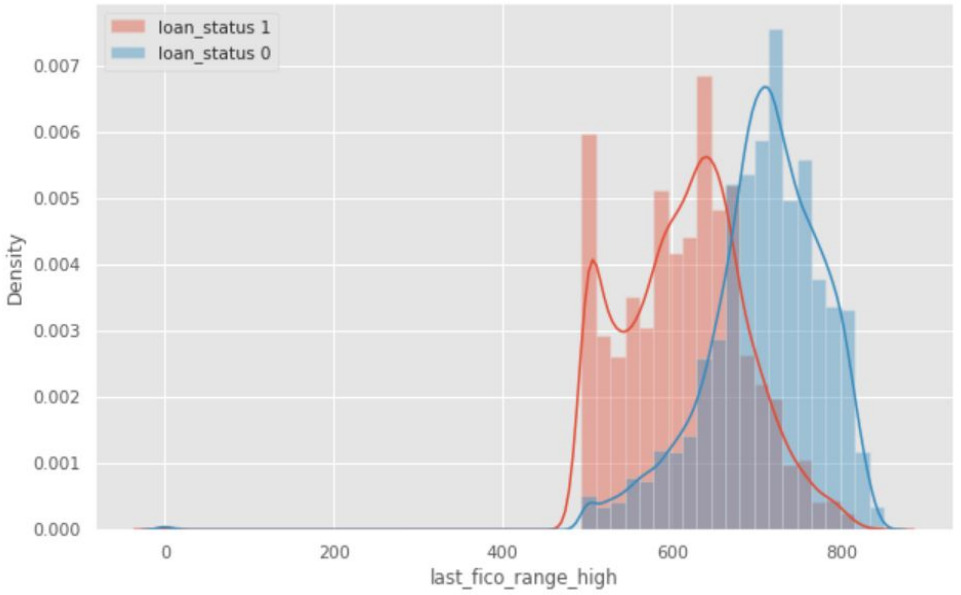


Define Dependent Variable

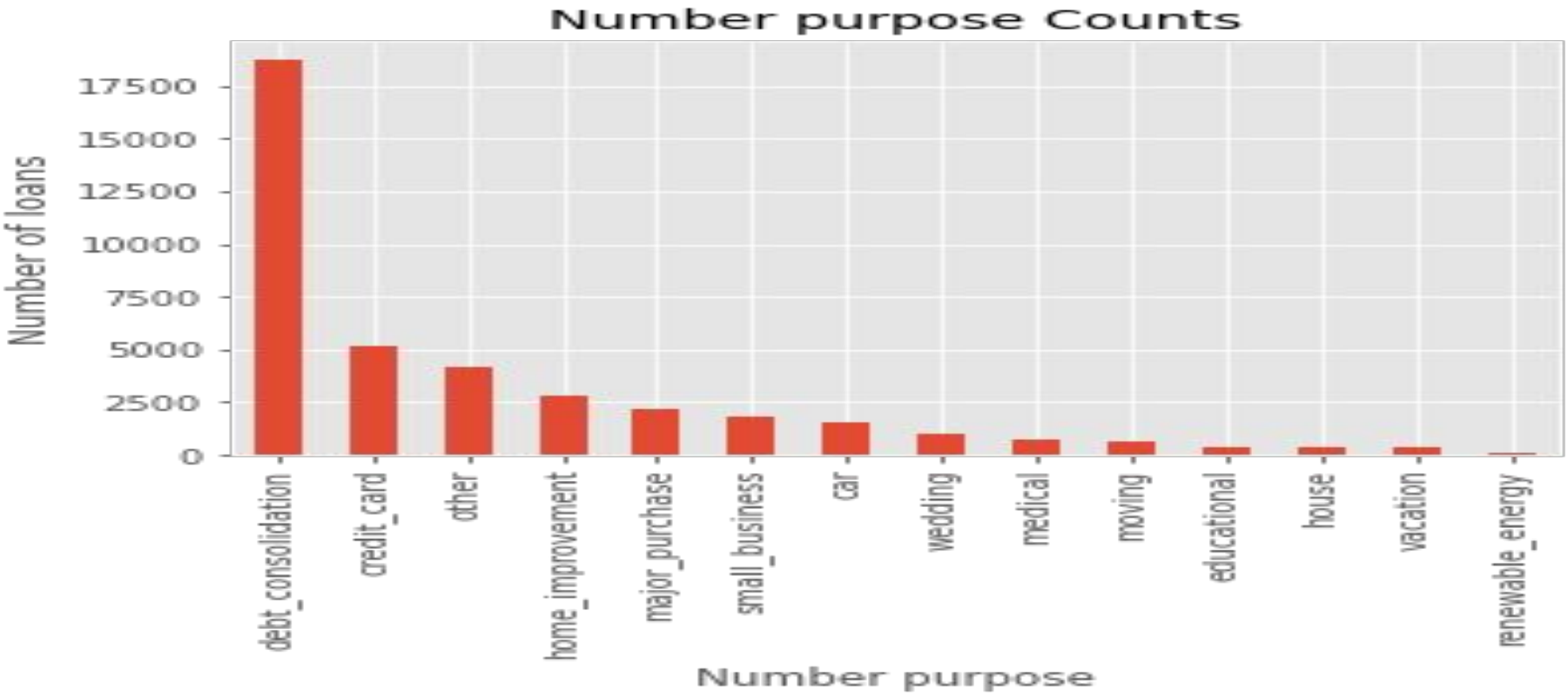


EDA

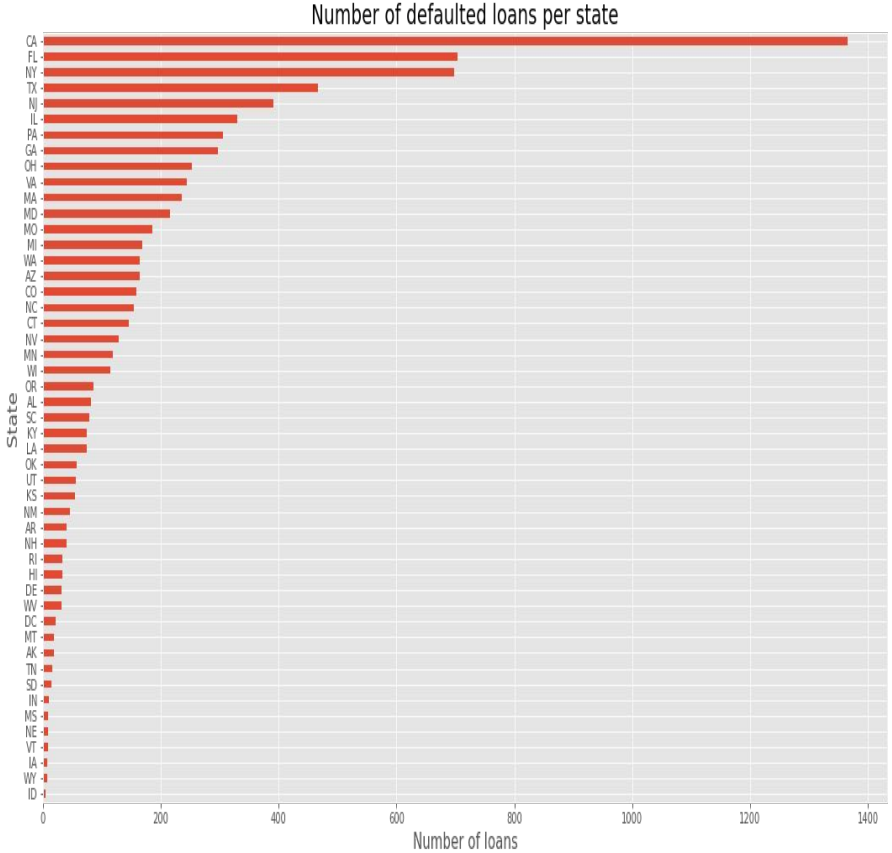

loan_status 1
loan_status 0
Density
0.00012
0.00010
0.00008
0.00006
0.00004
0.00002
0.00000
0
10000
20000
30000
40000
loan_amnt

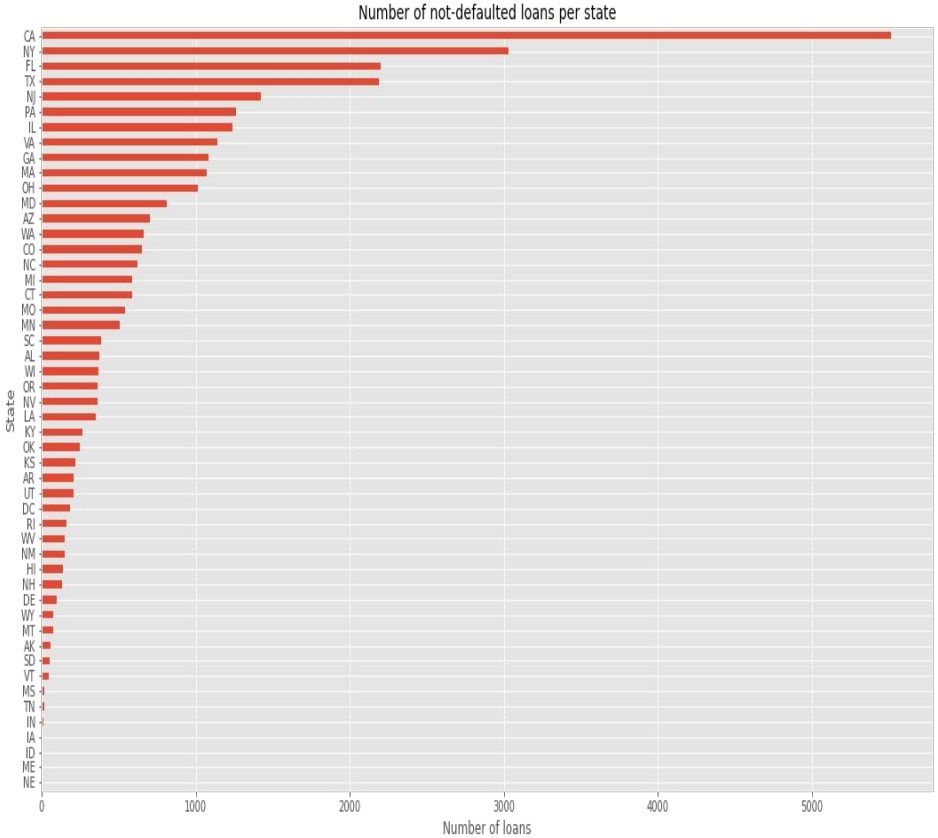

loan_status 1
loan_status 0
Density
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0.000
0
200
400
600
800
last_fico_range_high

EDA Cont..

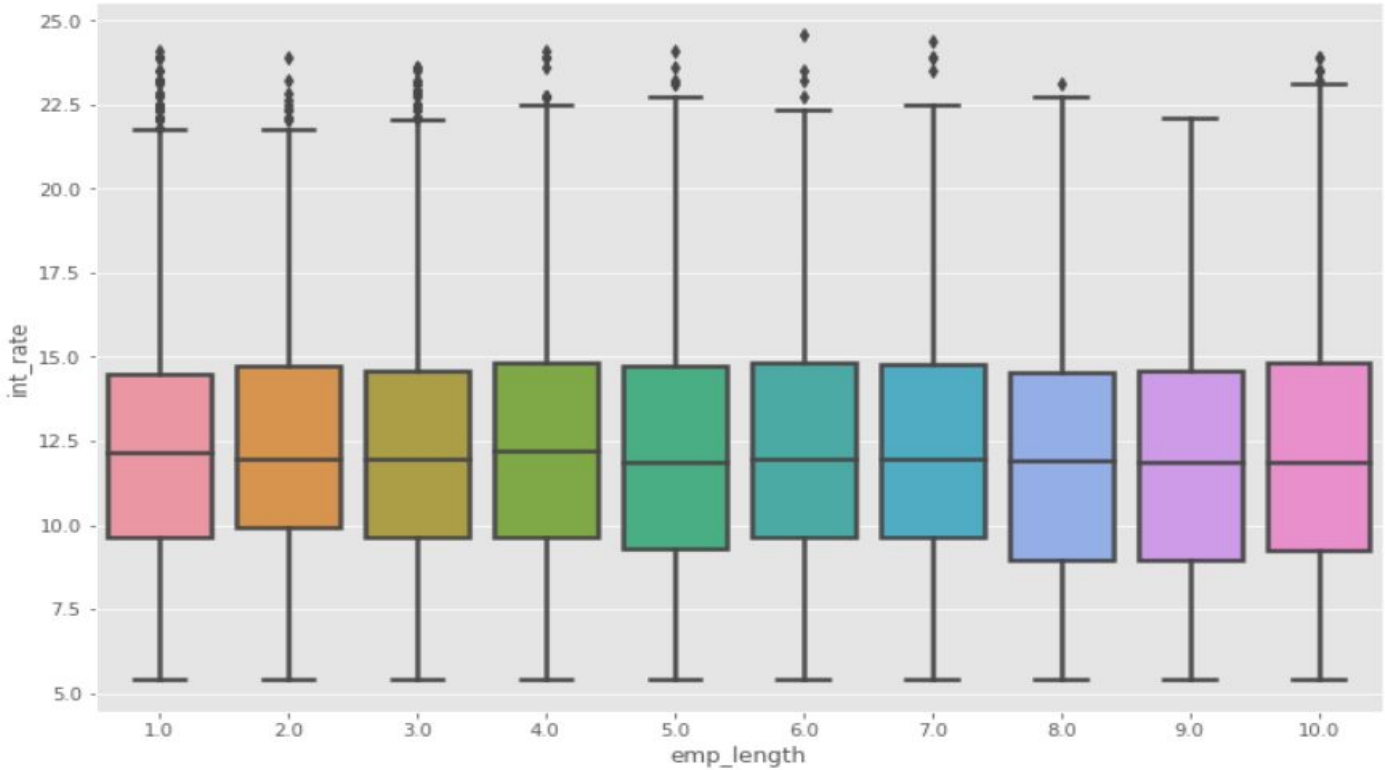

Number purpose Counts
Number of loans
17500
15000
12500
10000
7500
5000
2500
0
debt_consolidation
credit_card
other
home_improvement
major_purchase
small_business
car
wedding
medical
moving
educational
house
vacation
renewable_energy
Number purpose

EDA cont..

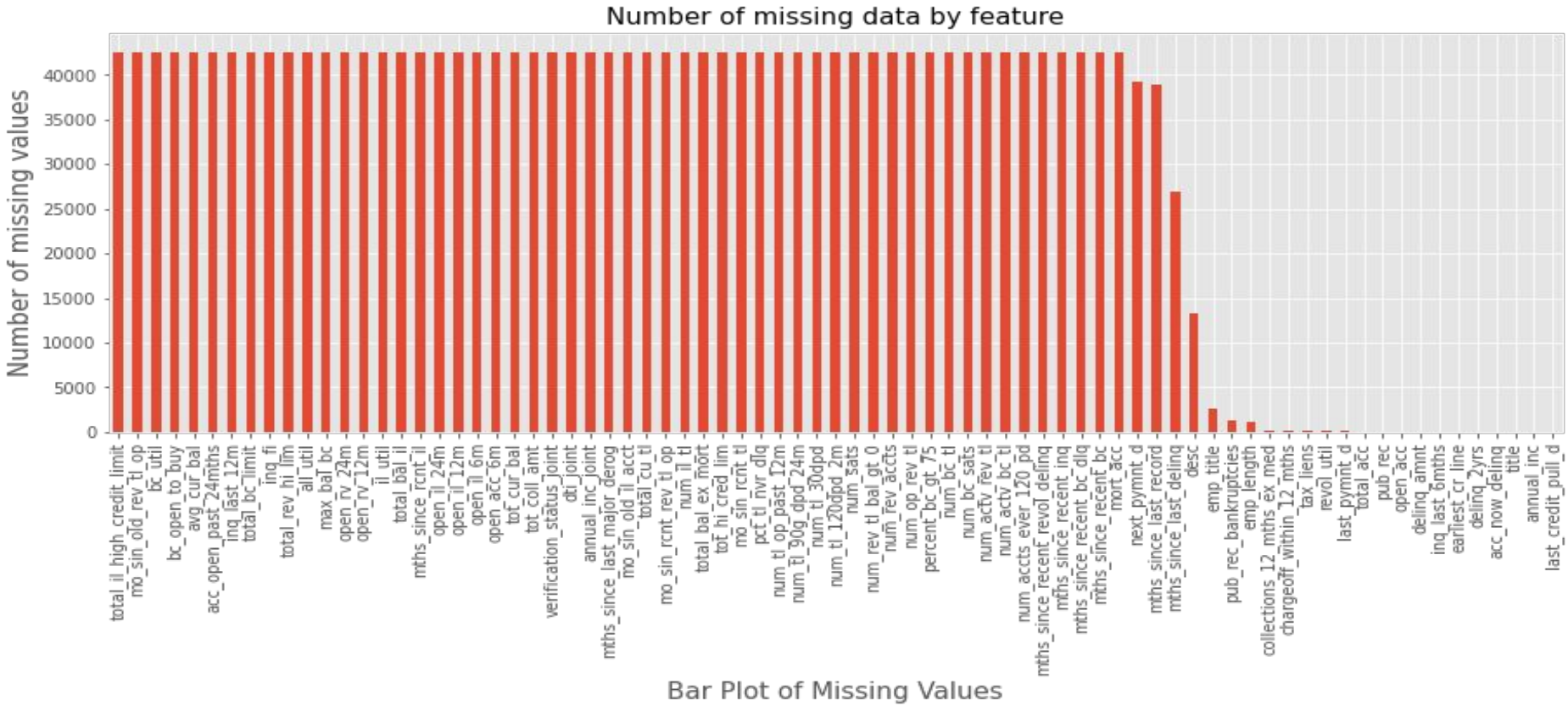

Number of defaulted loans per state
State
CA
FL
NY
TX
NJ
IL
PA
GA
OH
VA
MA
MD
MO
MI
WA
AZ
CO
NC
CT
NV
MN
WI
OR
AL
SC
NM
LA
OK
UT
KS
AR
NH
RI
HI
DE
WV
DC
MT
AK
TN
SD
IN
MS
NE
VT
IA
WY
ID
0
200
400
600
800
1000
1200
1400
Number of loans


Number of not-defaulted loans per state
State
CA
NY
FL
TX
NJ
PA
IL
VA
GA
MA
OH
MD
AZ
WA
CO
NC
MI
CT
MO
MN
SC
AL
WI
OR
NV
LA
KY
OK
KS
AR
UT
DC
RI
WV
NM
HI
NH
DE
WY
MT
AK
SD
VT
MS
TN
IN
IA
ID
ME
NE
0
1000
2000
3000
4000
5000
Number of loans

EDA cont..

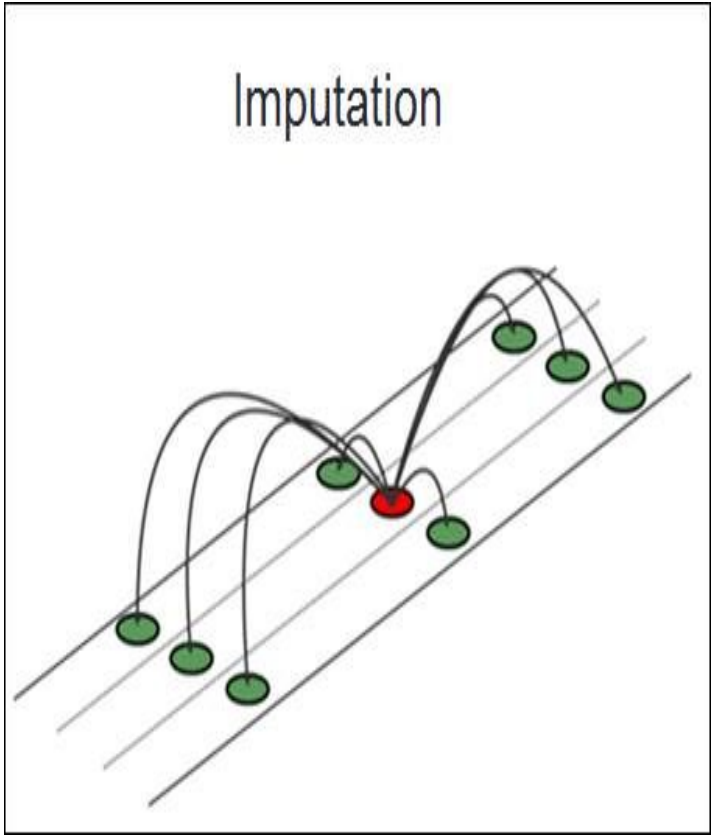

25.0
22.5
20.0
17.5
15.0
12.5
10.0
7.5
5.0
int_rate
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
emp_length

EDA cont..


Number of missing data by feature
Number of missing values
0
5000
10000
15000
20000
25000
30000
35000
40000
total_il_high_credit_limit
mo_sin_old_rev_tl_op
bc_util
bc_open_to_buy
avg_cur_bal
acc_open_past_24mths
inq_last_12m
total_bc_limit
inq_fi
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
open_il_12m
open_il_6m
open_acc_6m
tot_cur_bal
tot_coll_amt
verification_status_joint
dti_joint
annual_inc_joint
mths_since_last_major_derog
mo_sin_old_il_acct
total_cu_tl
mo_sin_rcnt_rev_tl_op
num_il_tl
total_bal_ex_mort
tot_hi_cred_lim
mo_sin_rcnt_tl
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
percent_bc_gt_75
num_bc_tl
num_bc_sats
num_actv_rev_tl
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
mort_acc
next_pymnt_d
mths_since_last_record
mths_since_last_delinq
desc
emp_title
pub_rec_bankruptcies
emp_length
collections_12_mths_ex_med
chargeoff_within_12_mths
tax_liens
revol_util
last_pymnt_d
total_acc
pub_rec
open_acc
delinq_amnt
inq_last_6mths
earliest_cr_line
delinq_2yrs
acc_now_delinq
title
annual_inc
last_credit_pull_d
Bar Plot of Missing Values

EDA cont.. (**Implement KNN Imputer for missing value**)

- Imputation for completing missing values using k-Nearest Neighbors.Each sample's missing values are imputed using the mean value from n_neighbors nearest neighbors found in the training set. Two samples are close if the features that neither is missing are close.
- The KNN Imputer class provides imputation for filling in missing values using the k-Nearest Neighbors approach. By default, a euclidean distance metric that supports missing values, nan_euclidean_distances, is used to find the nearest neighbors. Each missing feature is imputed using values from n_neighbors nearest neighbors t hat have a value for the feature. The feature of the neighbors are averaged uniformly or weighted by distance to each neighbor.



Preparing dataset for modeling

	loan_amnt	term	int_rate	installment	grade	emp_length	annual_inc	purpose	addr_state	dti	delinq_2yr
	5000.0	0.0	10.65	162.87	1.0	10.0	24000.0	1.0	3.0	27.65	0.
	2500.0	1.0	15.27	59.83	2.0	1.0	30000.0	0.0	10.0	1.00	0.
	2400.0	0.0	15.96	84.33	2.0	10.0	12252.0	11.0	14.0	8.72	0.
	10000.0	0.0	13.49	339.31	2.0	10.0	49200.0	9.0	4.0	20.00	0.
	5000.0	0.0	7.90	156.46	0.0	3.0	36000.0	13.0	3.0	11.20	0.
	7000.0	1.0	15.96	170.08	2.0	8.0	47004.0	2.0	27.0	23.51	0.
	3000.0	0.0	18.64	109.43	4.0	9.0	48000.0	0.0	4.0	5.35	0.
	5600.0	1.0	21.28	152.39	5.0	4.0	40000.0	11.0	4.0	5.55	0.
	5375.0	1.0	12.69	121.45	1.0	1.0	15000.0	9.0	42.0	18.08	0.
	6500.0	1.0	14.65	153.45	2.0	5.0	72000.0	2.0	3.0	16.12	0.
0	12000.0	0.0	12.69	402.54	1.0	10.0	75000.0	2.0	4.0	10.78	0.
1	9000.0	0.0	13.49	305.38	2.0	1.0	30000.0	2.0	44.0	10.08	0.
2	3000.0	0.0	9.91	96.68	1.0	3.0	15000.0	1.0	14.0	12.56	0.
3	10000.0	0.0	10.65	325.74	1.0	3.0	100000.0	9.0	4.0	7.06	0.
4	1000.0	0.0	16.29	35.31	3.0	1.0	28000.0	2.0	24.0	20.31	0.

Task:-

Train Set:-

Test set:-

Response:-

Model Validation & Selection(Baseline Model)

	Logistic Regression
test_accuracy	0.832096
recall_test	0.330956
precision_test	0.660969
f1_test	0.441065
auc_test	0.644235
cm_test	[[8046, 357], [1407, 696]]
train_accuracy	0.831165
precision_train	0.662284
recall_train	0.32607
f1_train	0.436991
auc_train	0.642128
cm_train	[[24130, 1053], [4268, 2065]]

	GRad_BOOst
test_accuracy	0.868646
recall_test	0.904898
precision_test	0.617256
f1_test	0.733899
auc_test	0.882236
cm_test	[[7223, 1180], [200, 1903]]
train_accuracy	0.890849
precision_train	0.659499
recall_train	0.944418
f1_train	0.776652
auc_train	0.5
cm_train	[[22095, 3088], [352, 5981]]

Model Validation & Selection(SMOTE)

	Model Name	Train Accuracy	Test Accuracy	Precision Train	Precision Test	Recall Train	Recall test	ROC AUC Train	ROC AUC Test	AUC Train	AUC Test
0	LogisticRegression	0.84	0.83	0.83	0.54	0.85	0.81	0.84	0.82	0.838939	0.819453
1	DecisionTreeClassifier	1.00	0.88	1.00	0.68	1.00	0.72	1.00	0.82	1.000000	0.817953
2	RandomForestClassifier	1.00	0.92	1.00	0.77	1.00	0.83	1.00	0.89	1.000000	0.885250
3	GradientBoostingClassifier	0.94	0.91	0.93	0.76	0.95	0.84	0.94	0.89	0.940555	0.885723
4	XGBClassifier	0.94	0.91	0.93	0.76	0.95	0.84	0.94	0.89	0.938768	0.885544
5	KNeighborsClassifier	0.84	0.60	0.78	0.23	0.95	0.44	0.84	0.54	0.843744	0.539154
6	SVC	0.58	0.59	0.58	0.26	0.56	0.55	0.58	0.58	0.579597	0.575214

Model Validation & Selection(Classification Matrix)

Test Data

```
For LogisticRegression Confusion matrix is :
[[6983 1420]
 [ 404 1699]]
For DecisionTreeClassifier Confusion matrix is :
[[7701  702]
 [ 590 1513]]
For RandomForestClassifier Confusion matrix is :
[[7893  510]
 [ 355 1748]]
For GradientBoostingClassifier Confusion matrix is :
[[7849  554]
 [ 342 1761]]
For XGBClassifier Confusion matrix is :
[[7842  561]
 [ 341 1762]]
For KNeighborsClassifier Confusion matrix is :
[[5353 3050]
 [1175  928]]
For SVC Confusion matrix is :
[[5040 3363]
 [ 945 1158]]
```

Train Data

```
For LogisticRegression Confusion matrix is :
[[20834  4349]
 [ 3763 21420]]
For DecisionTreeClassifier Confusion matrix is :
[[25183     0]
 [    0 25183]]
For RandomForestClassifier Confusion matrix is :
[[25183     0]
 [    0 25183]]
For GradientBoostingClassifier Confusion matrix is :
[[23481  1702]
 [ 1292 23891]]
For XGBClassifier Confusion matrix is :
[[23438  1745]
 [ 1339 23844]]
For KNeighborsClassifier Confusion matrix is :
[[18531  6652]
 [ 1218 23965]]
For SVC Confusion matrix is :
[[15080 10103]
 [11071 14112]]
```

Model Validation & Selection(Hyperparameter tuned)

Accuracy Score 0.9237578526556254

	precision	recall	f1-score	support
0.0	0.95	0.95	0.95	8403
1.0	0.81	0.81	0.81	2103
accuracy			0.92	10506
macro avg	0.88	0.88	0.88	10506
weighted avg	0.92	0.92	0.92	10506

Random Forest Classifier

XG Boost Classifier

Accuracy Score 0.73043974871502

	precision	recall	f1-score	support
0.0	0.99	0.67	0.80	8403
1.0	0.43	0.99	0.59	2103
accuracy			0.73	10506
macro avg	0.71	0.83	0.70	10506
weighted avg	0.88	0.73	0.76	10506

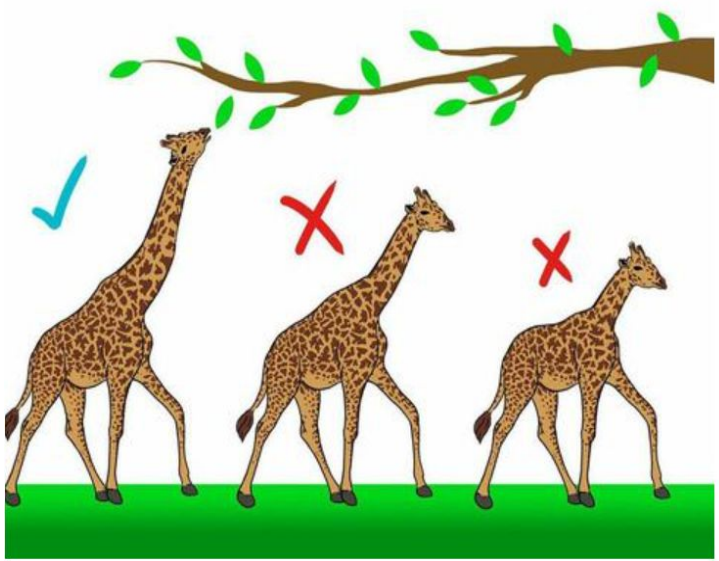
Model Validation & Selection(continued)

Observation 1: As seen in the table above, DT classifier gives us a better precision value whereas Random Forest classifier is lagging behind in terms of precision,but from the business perspective we are looking for a better recall value, which is given to us by Random Forest classifier.

Observation 2: Gaussian NB has performed poorly in terms of precision and AUC value, whereas we see that DT Classifier,RT Classifier,GB Classifier & XBG Classifier have performed equally good in terms of F-1 Score and Test accuracy.

Model Validation & Selection(continued)

Observation 3: From the above observation we have come to the conclusion that we would choose our model from RF Classifier,GB Classifier or XGB Classifier.
Although, we see DT Classifier also giving great results but that may lead of overfitting of the data.

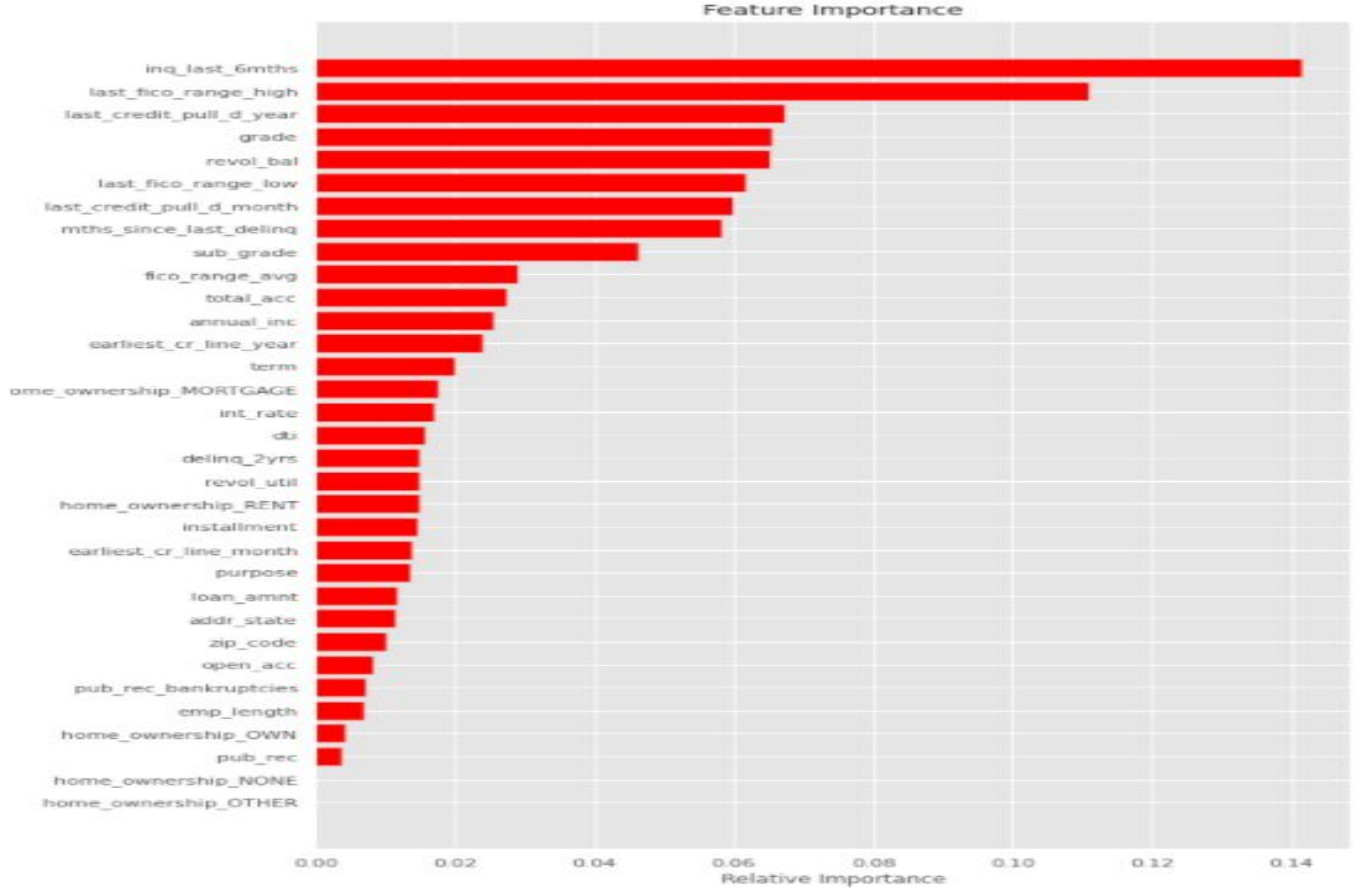


Model Validation & Selection(continued)

We had chosen XG Boost Classifier for our prediction and the best hyperparameters obtained are as below.

Best hyperparameters:
colsample_bylevel=1,
colsample_bynode=1
colsample_bytree=0.9,eta=0.3,gamma=0,
learning_rate=0.1, max_delta_step=0, max_depth=5,
min_child_weight=1, missing=None, n_estimators=100,
n_jobs=1,nthread=4,num_boost_round=10
objective='binary:logistic'
random_state=0, reg_alpha=0,
reg_lambda=1,scale_pos_weight=90

Model Validation & Selection(continued)



Challenges

- Huge chunk of data was to be handled keeping in mind not to miss anything which is even of little relevance.
- Feature selection was quite a challenge as our dataset had many futuristic features which had no relevance for initial detection of loan defaulter.

Futuristic Features

total_pymnt:- Payments received to date for total amount funded.
total_pymnt_inv:-Payments received on date for portion of total amount funded by investors.
total_rec_prncp:-Principal received to date.
total_rec_int:-Interest received to date.
total_rec_late_fee:-Late fees received to date.
recoveries:-post charge off gross recovery.
collection_recovery_fee:-post charge off collection fee.
last_pymnt_d:-Last month payment was received.
last_pymnt_amnt:-Last total payment amount received.

funded_amnt:-The total amount committed to that loan at that point in time.

funded_amnt_inv:-The total amount committed by investors for that loan at that time.

Q & A