Capstone Project - 1
Loan Default Prediction

Let's Catch The Defaulters

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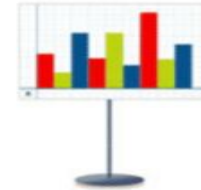5. **Applying Model**
6. **Model Validation and Selection**

The Dilemma

LendingClub

How Lending Club Works

Borrowers apply for loans.
Investors open an account.

Borrowers get funded.
**Investors** build a portfolio.

Borrowers repay automatically.
Investors earn & reinvest.

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 project is to create a predictive model that identifies applicants who are relatively risky for a loan. In order to accomplish this, we organized the whole series into four parts as follows:

Data Pipeline

- **Data processing-1**: In this first part we've removed unnecessary features. Since 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, And encoded the categorical features ,changed the columns containing date time values .
- **EDA**: In in this part, we do some exploratory data analysis (EDA) on the features selected in part-1 and 2 to see the trend.
- **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
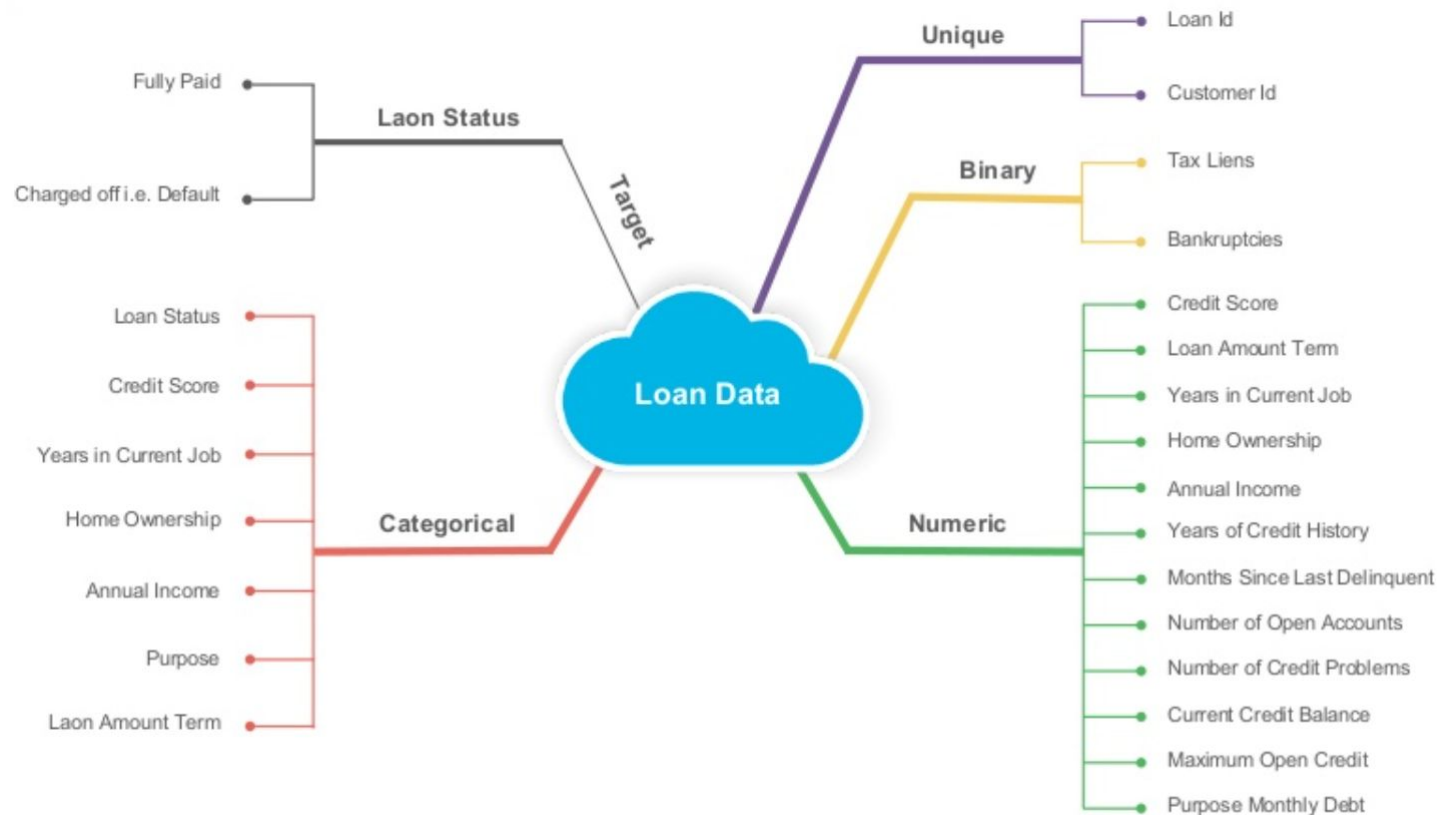
Data Summary

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 : A loan is considered "delinquent" when a borrower doesn't make a loan payment on time. Most lenders allow consumers a grace period to make up a missed payment and get their loan out of delinquency. However, once a loan is delinquent for a certain period of time, it becomes at risk of going into default.

inq_last_6mths: The fundamental problem with multiple hard inquiries is that they influence lenders to form a negative impression about you credit behaviour. Too many hard inquiries over a short time-period showcase you as a credit-hungry customer with a consequently high risk-quotient.

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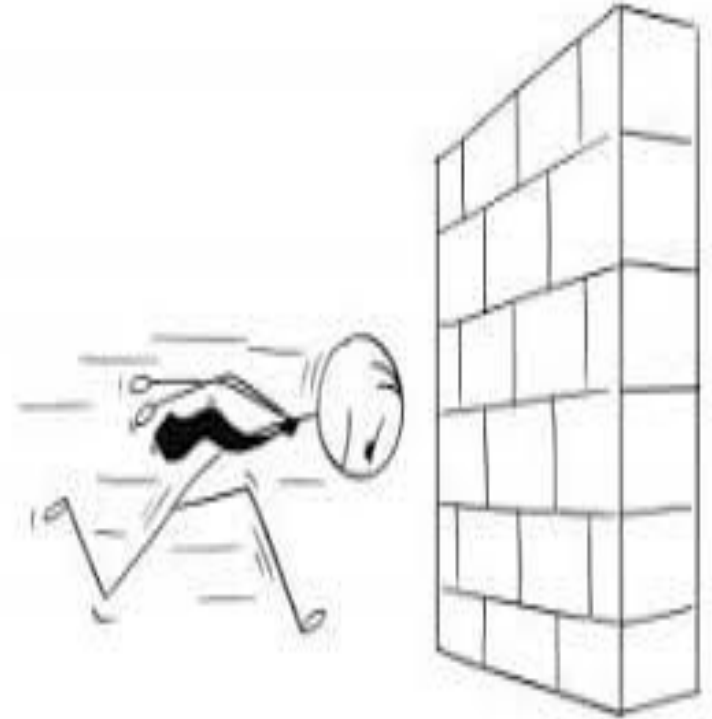
Current->Loan is up to date on current payments.,

In Grace period->The loan is past due but still in the grace period of 15 days.
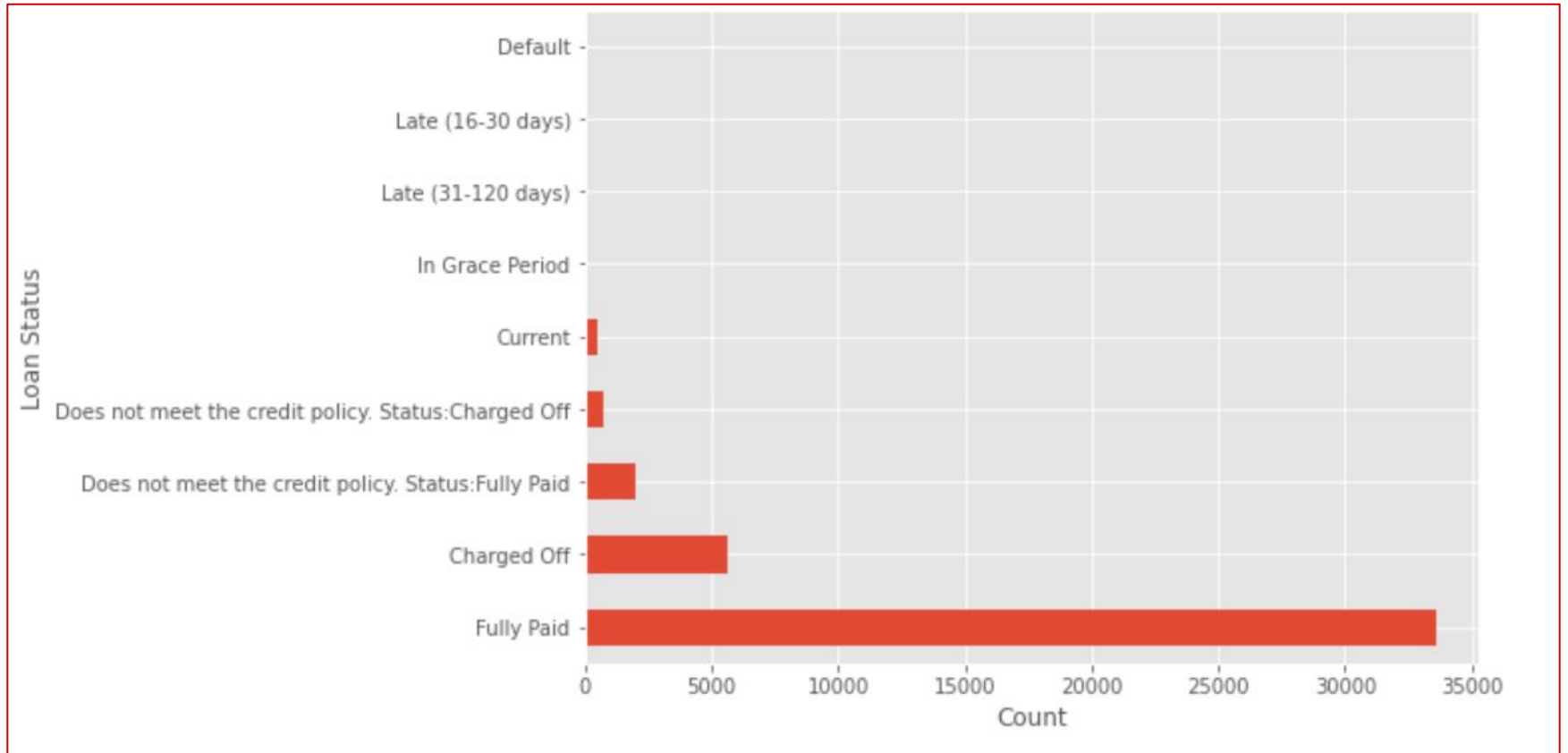
Late(30-120)->Loan hasn't been paid in 31 to 120 days (late on the current payment).

Late(16-30)->Loan hasn't been paid in 16 to 30 days (late on the current payment).
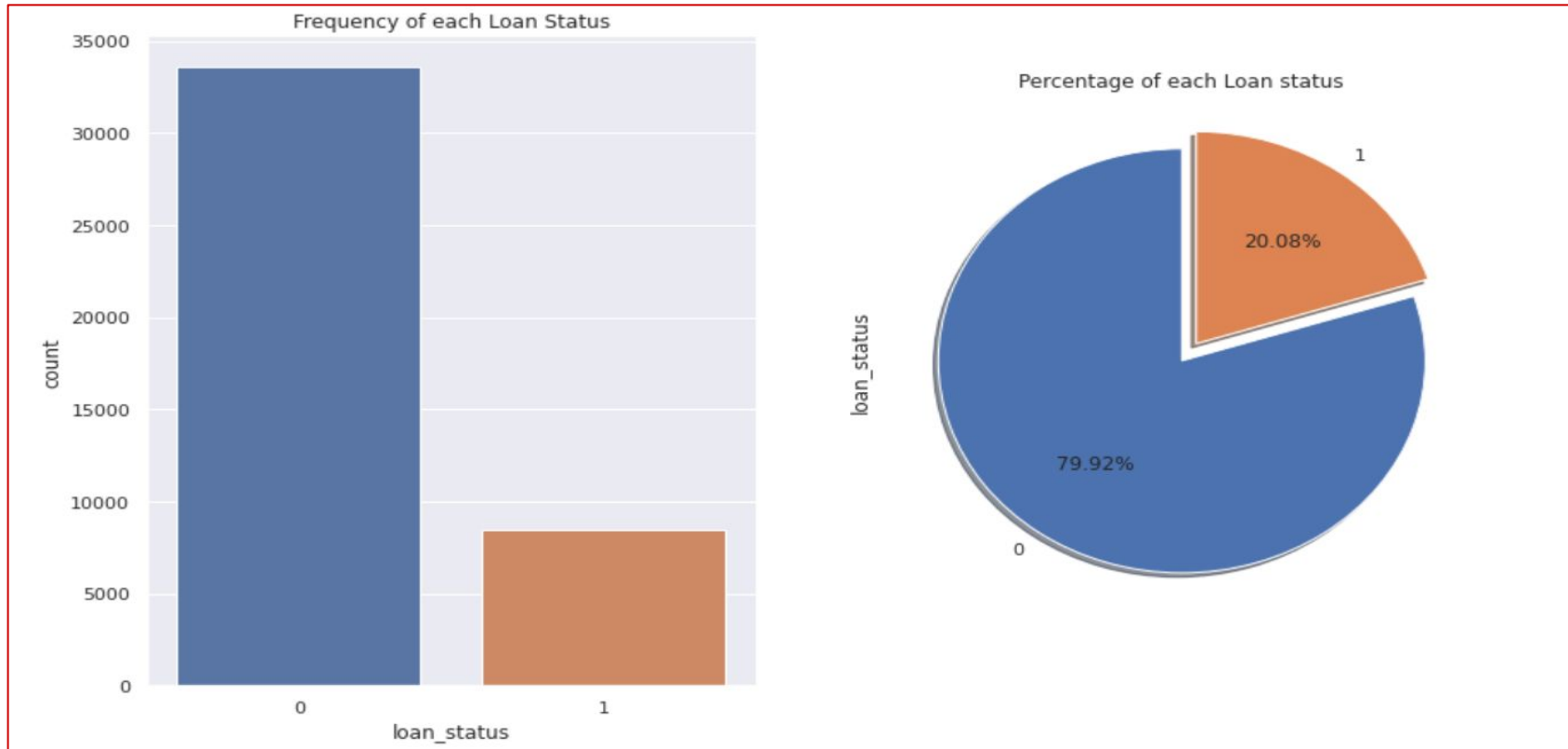
Default->Loan is defaulted on and no payment has been made for more than 121 days.

Define Dependent Variable

Define Dependent Variable

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.

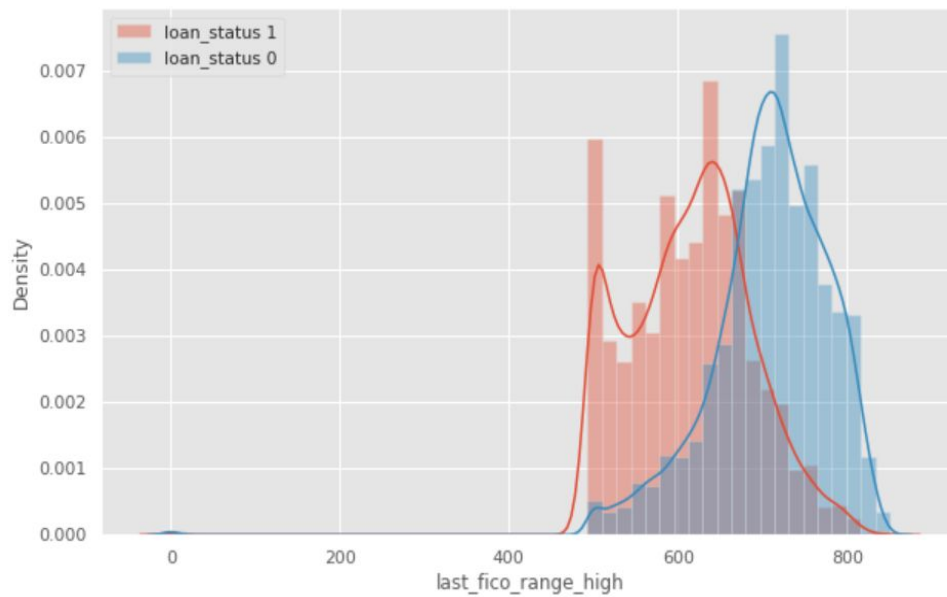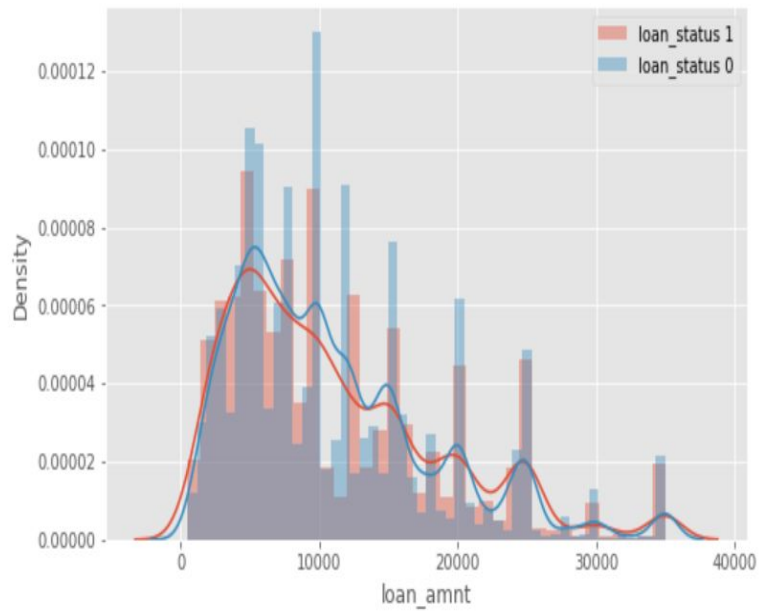collection_recovery_fee:-post charge off collection fee.

last_pymnt_d:-Last month payment was received.

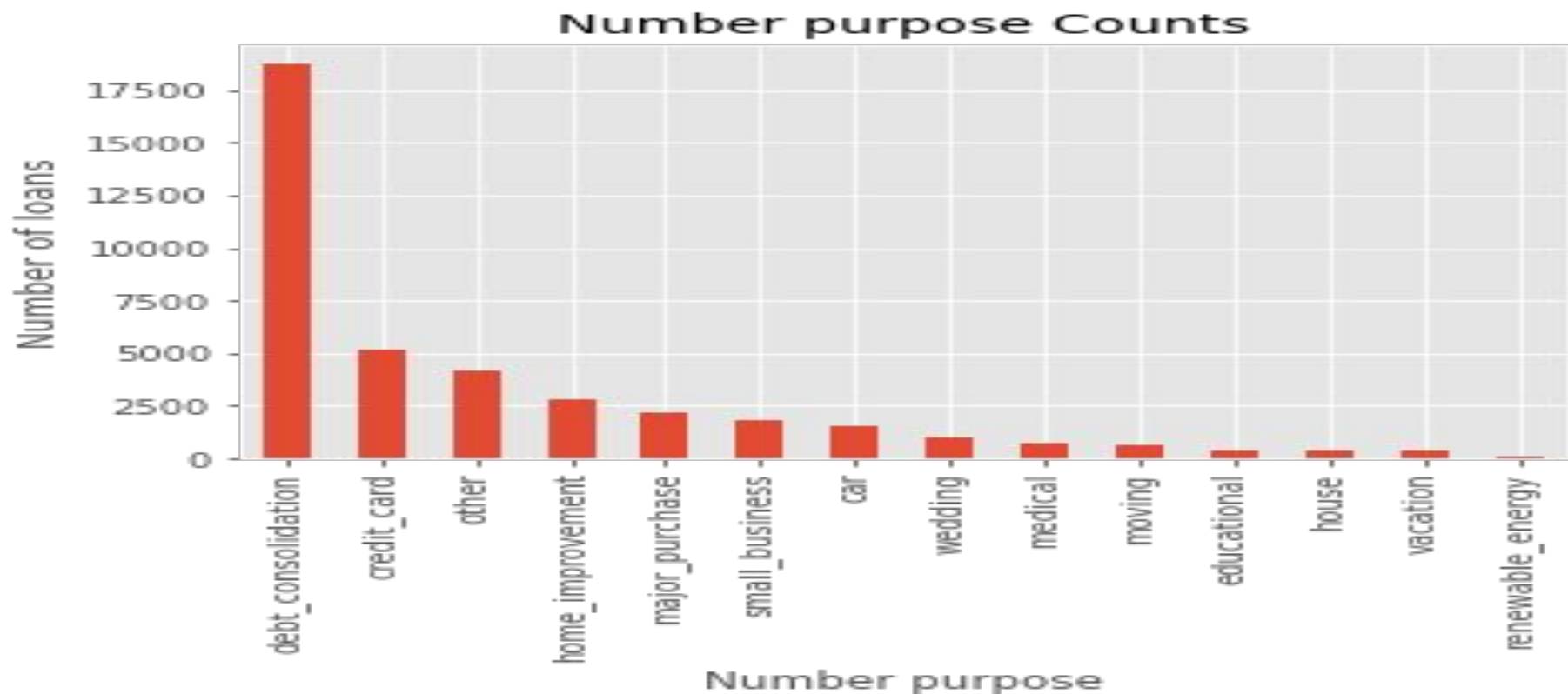last_pymnt_amnt:-Last total payment amount received.

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.
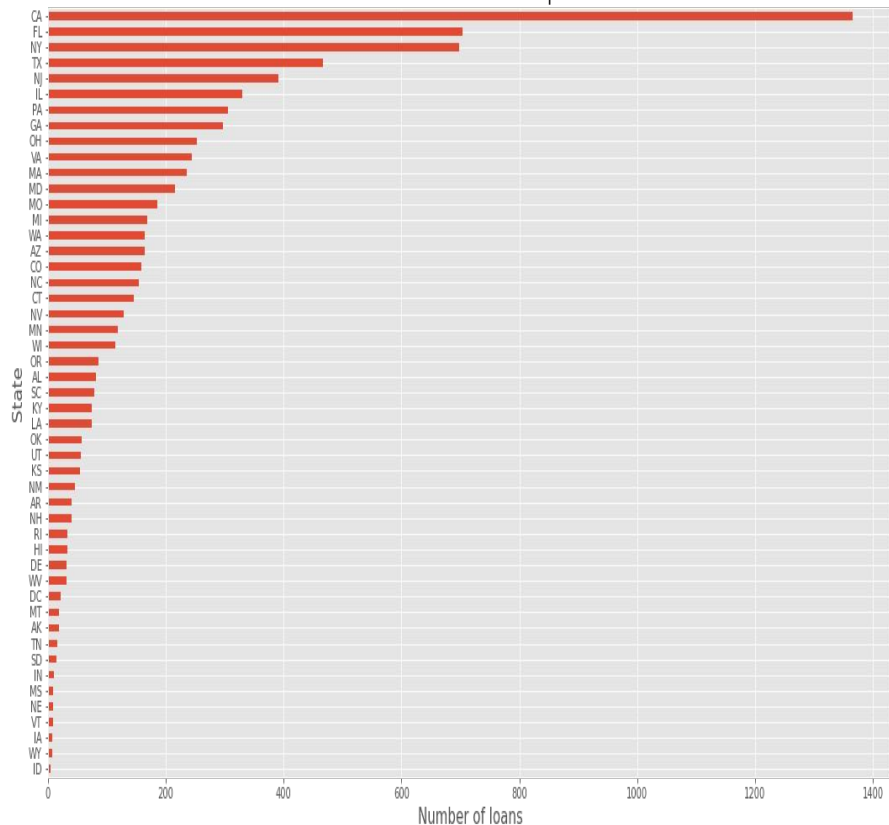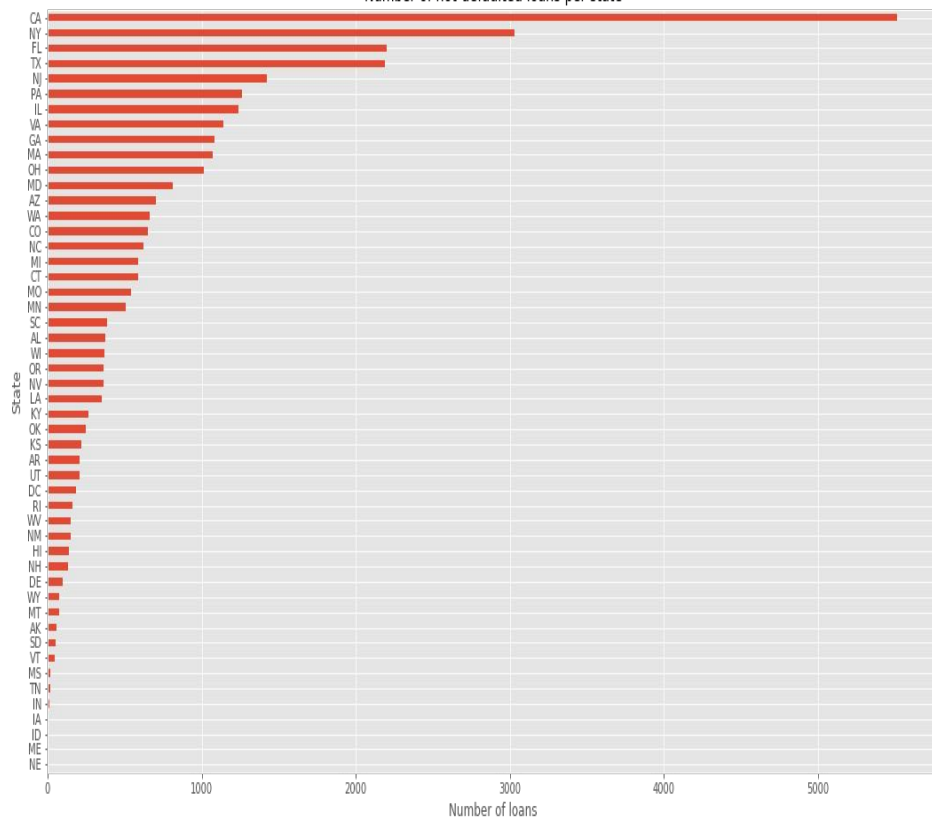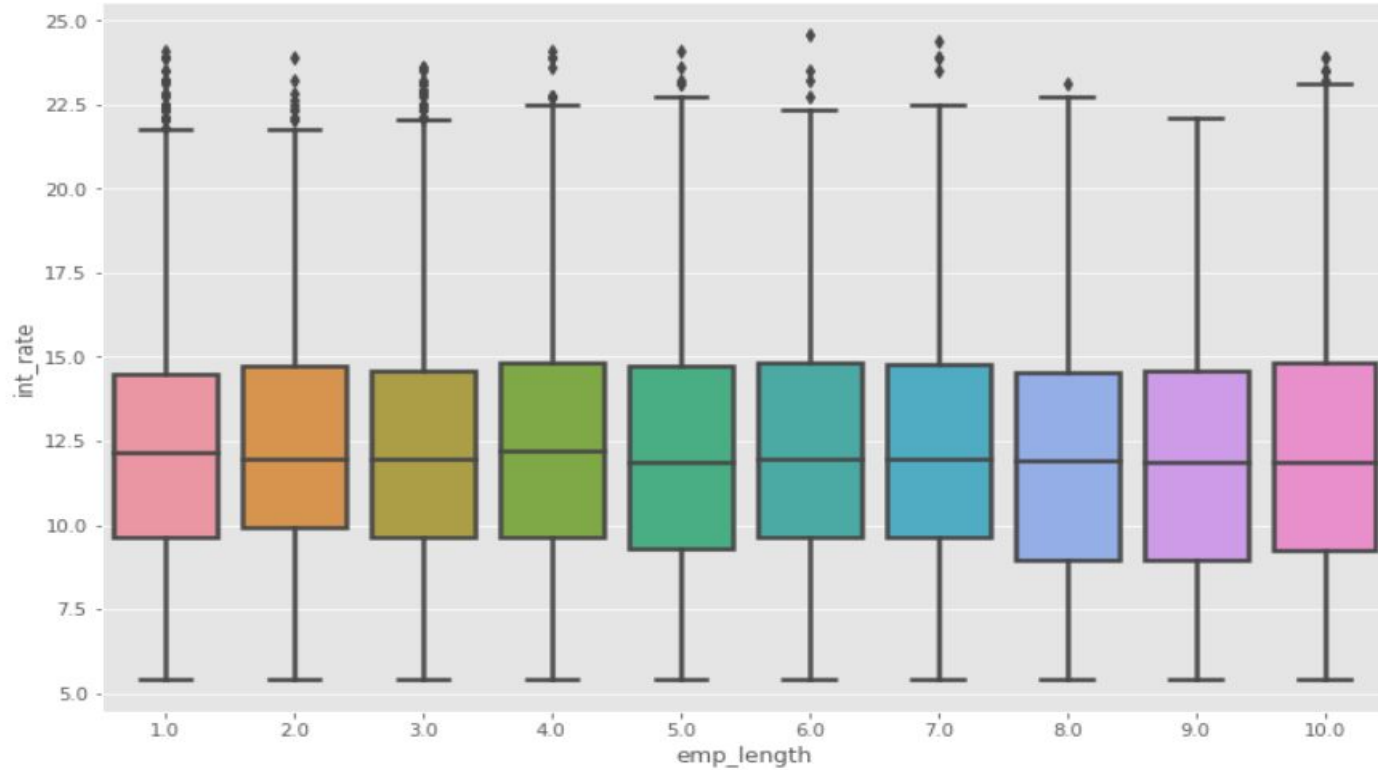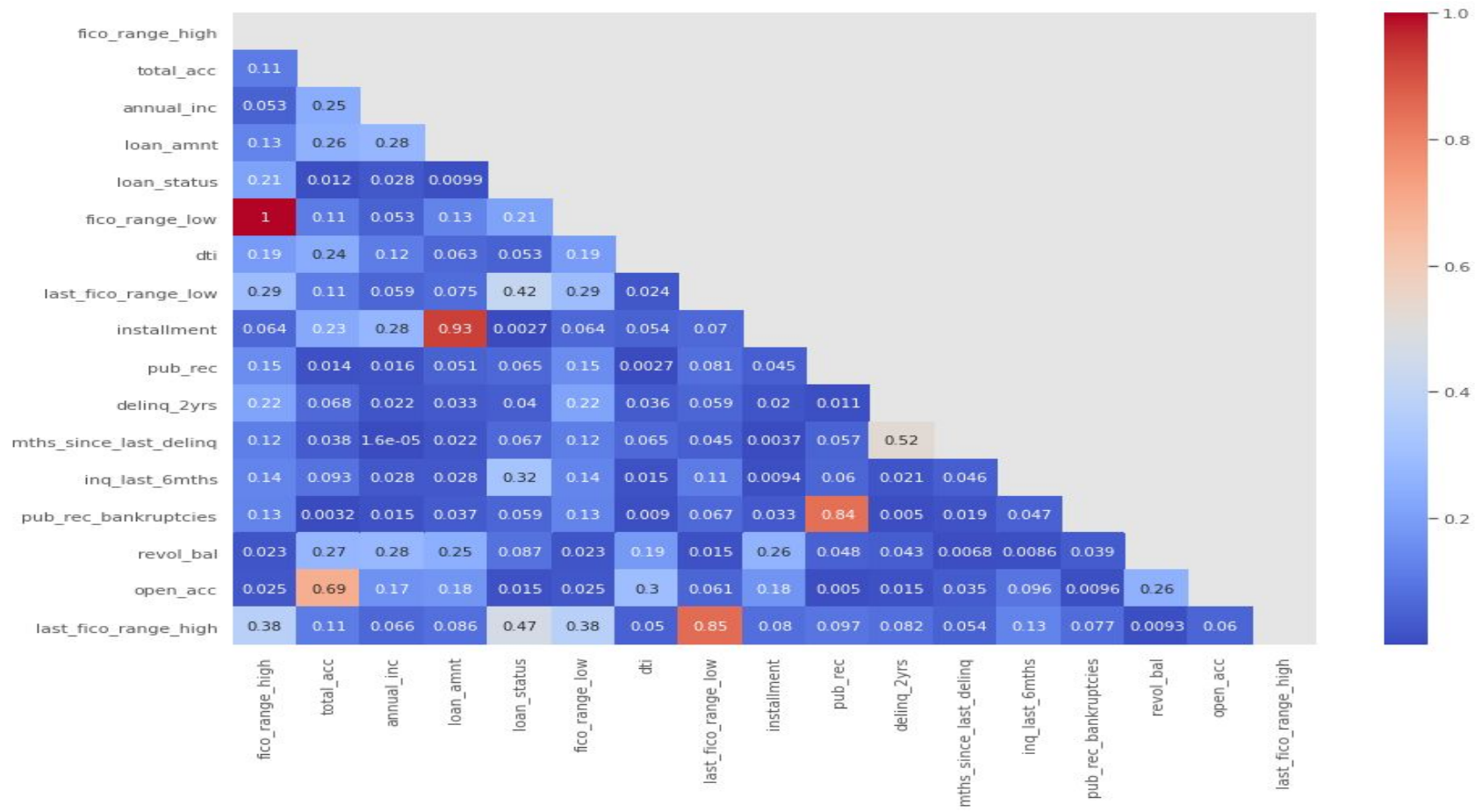
EDA

EDA (continued)

Number purpose Counts

EDA (continued)

EDA **(continued)**

EDA (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.

Imputation

Preparing dataset for modeling

AI

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	0.0
2500.0	1.0	15.27	59.83	2.0	1.0	30000.0	0.0	10.0	1.00	0.0
2400.0	0.0	15.96	84.33	2.0	10.0	12252.0	11.0	14.0	8.72	0.0
10000.0	0.0	13.49	339.31	2.0	10.0	49200.0	9.0	4.0	20.00	0.0
5000.0	0.0	7.90	156.46	0.0	3.0	36000.0	13.0	3.0	11.20	0.0
7000.0	1.0	15.96	170.08	2.0	8.0	47004.0	2.0	27.0	23.51	0.0
3000.0	0.0	18.64	109.43	4.0	9.0	48000.0	0.0	4.0	5.35	0.0
5600.0	1.0	21.28	152.39	5.0	4.0	40000.0	11.0	4.0	5.55	0.0
5375.0	1.0	12.69	121.45	1.0	1.0	15000.0	9.0	42.0	18.08	0.0
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	0.0
9000.0	0.0	13.49	305.38	2.0	1.0	30000.0	2.0	44.0	10.08	0.0
3000.0	0.0	9.91	96.68	1.0	3.0	15000.0	1.0	14.0	12.56	0.0
10000.0	0.0	10.65	325.74	1.0	3.0	100000.0	9.0	4.0	7.06	0.0
1000.0	0.0	16.29	35.31	3.0	1.0	28000.0	2.0	24.0	20.31	0.0

Task:-Binary Classification

Train Set:-(31516, 33)

Test set:-(10506, 33)

Response:- 0/1

Applying Model (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]]

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(continued)

**Observation 1**: As seen in the table above, Logistic Regression is not giving great results,Decision Trees are had bit better than Logistic Regression,but having lesser recall value on test data.

**Observation 2**:GB Classifier & XBG Classifier have performed equally good in terms of recall  Score and Test accuracy.

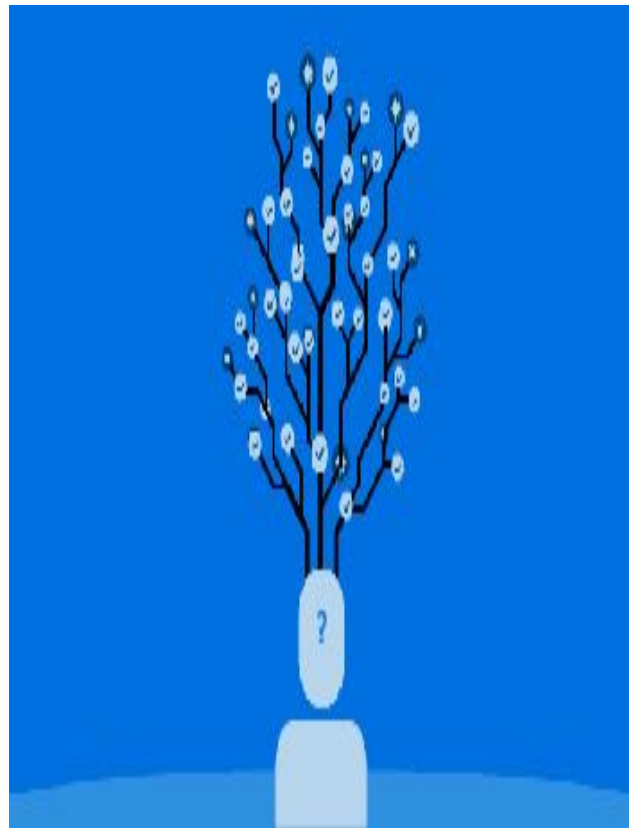# Model Validation & Selection (continued)

Observation 3: From the above observation we have come to the conclusion that we would choose our model from RF Classifier or XGB Classifier.
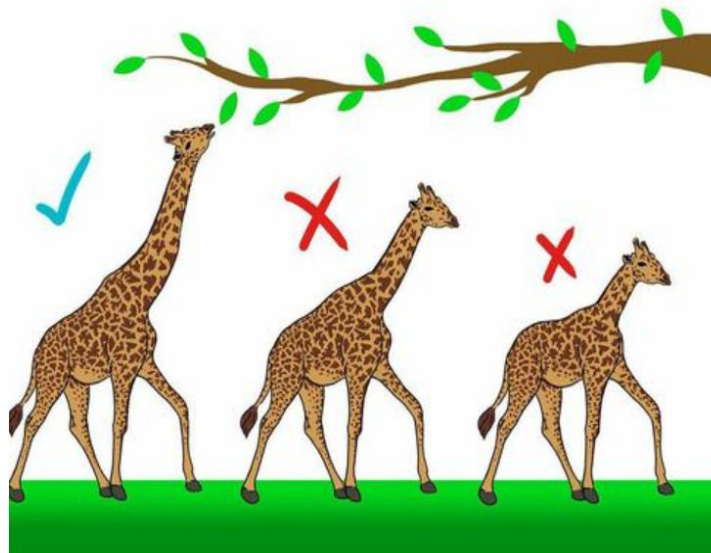
# Model Validation & Selection(continued)

We had chosen Random Forest Classifier for our prediction and the best hyperparameters obtained are as below.
criterion='gini'
max_depth=25
max_features='auto',
max_leaf_nodes=None
max_samples=None,
min_impurity_decrease=0.0
min_impurity_split=None,
min_samples_leaf=2
min_samples_split=10,
min_weight_fraction_leaf=0.0
n_estimators=800

# 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'
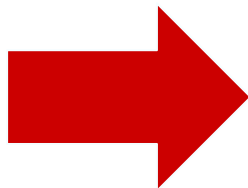random_state=0, reg_alpha=0,
reg_lambda=1,scale_pos_weight=90

# Model Validation & Selection (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
```
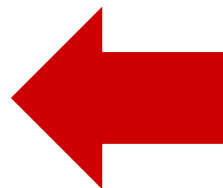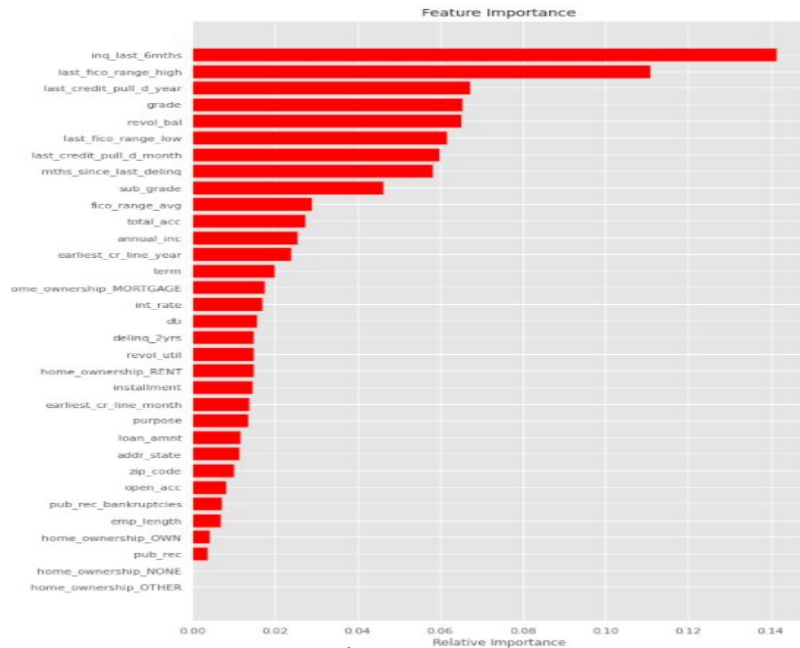
**Random Forest Classifier**

**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
```
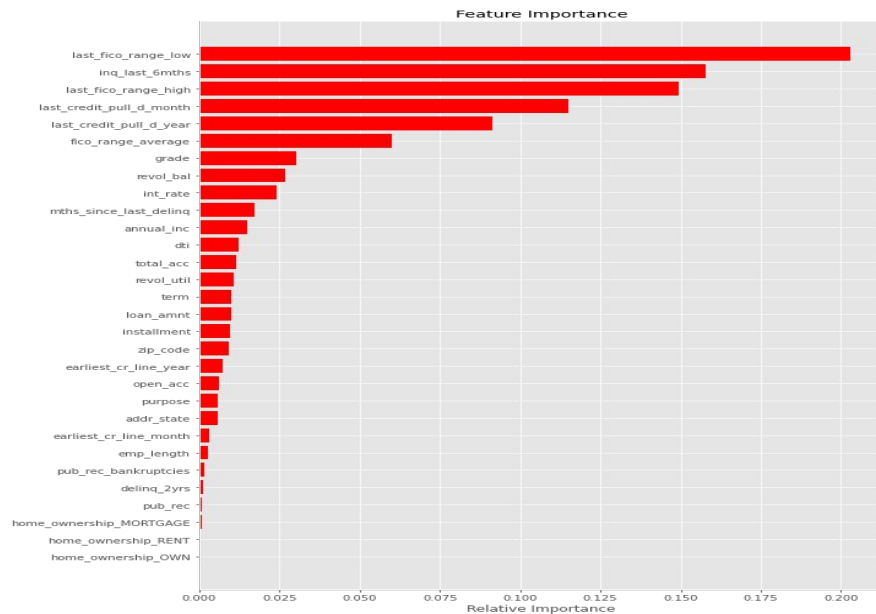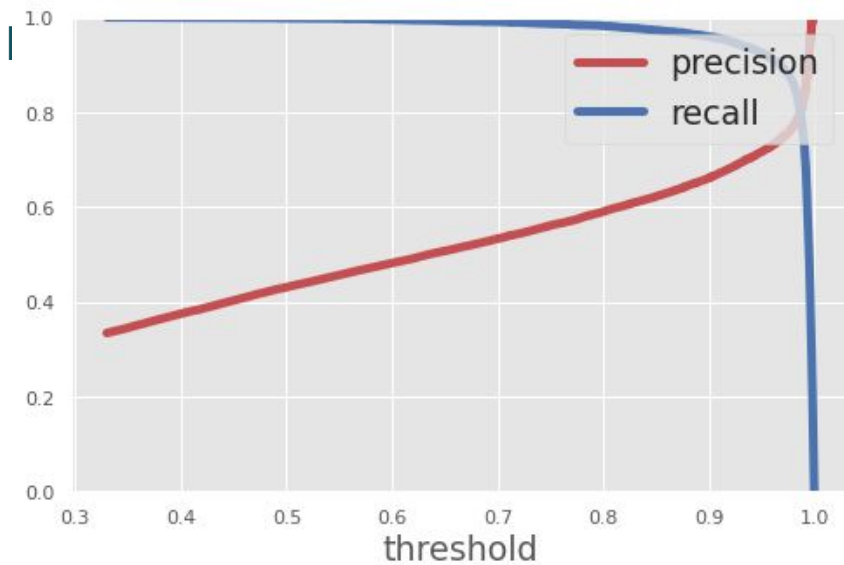
# Model Validation & Selection (Feature Importance)



XG Boost Classifier

Random Forest Classifier

# Model Validation & Selection



**Precision vs Recall**

**XG Boost Classifier**(scale_pos_weight)

# Conclusion

- The last_fico_range, grade, inq_last_6month features were found to be the most relevant for predicting loan default in. The current model tries to predict default biased data from credit analysts grade and assigned interest rate. . The XGB and Rf models provide substantial improvements on traditional credit screening. A recall score significantly and robustly above 90%, with AUC-ROC scores ≈74%. The features provided to the model in our study generalize to any lending activity and institution, beyond P2P lending. The present work could, therefore, be augmented in order to predict loan default risk without the need for human credit screening.
- Only the Random Forest., XGBoost, model is used, but there are many good ones out there even neural networks. The models can also be improved further by finer tunings on hyperparameter.
- In the bank loan behaviour prediction, for example, banks want to control the loss to a acceptable level, so they may use a relatively low threshold. This means more customers will be grouped as "potential bad customers" and their profiles will be checked carefully later by the credit risk management team. In this way, banks can detect the default behaviours in the earlier stage and conduct the corresponding actions to reduce the possible loss.

# 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.

- Computation time.