Capstone Project

Loan Default Prediction

Let’s Catch The Defaulters

- Defining problem statement
- EDA and feature engineering
- Feature Selection
- Dataset preparation for modelling
- Fitting our model using ML algorithms
- Model interpretation using SHAP
- Conclusions and Plans to improve

Data Pipeline

- **Data processing-1: We had huge number of features to work with initially. In this section we removed all the features which were irrelevant.**
- **Data processing-2: In this section, we will focus on some advanced operations including label and one hot encoding of features.**
- **EDA: We will extract some very important information from our features using visualization plots in this section.**
- **Fitting our model: We will focus on using different ML algorithms to fit our model and test our model on newer datapoints.**
- **Model interpretation: We will try to interpret our black-box model using SHAP plots (Shapley Additive Explanations)**

Explaining our features:

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.

Explaining our features: (contd.)

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.

Explaining our features: (contd.)

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.

Explaining our Target feature:

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.

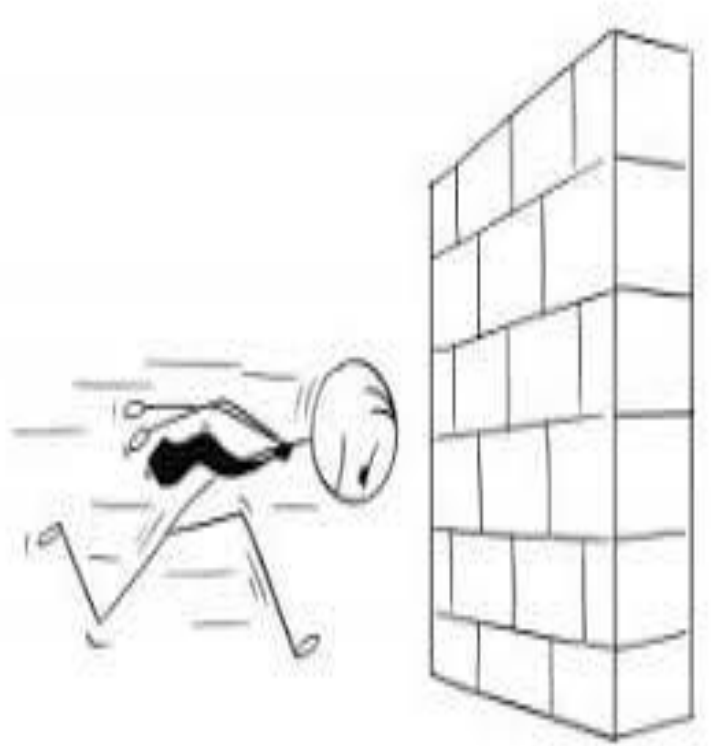
Current->Loan is up to date on current payments.,

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

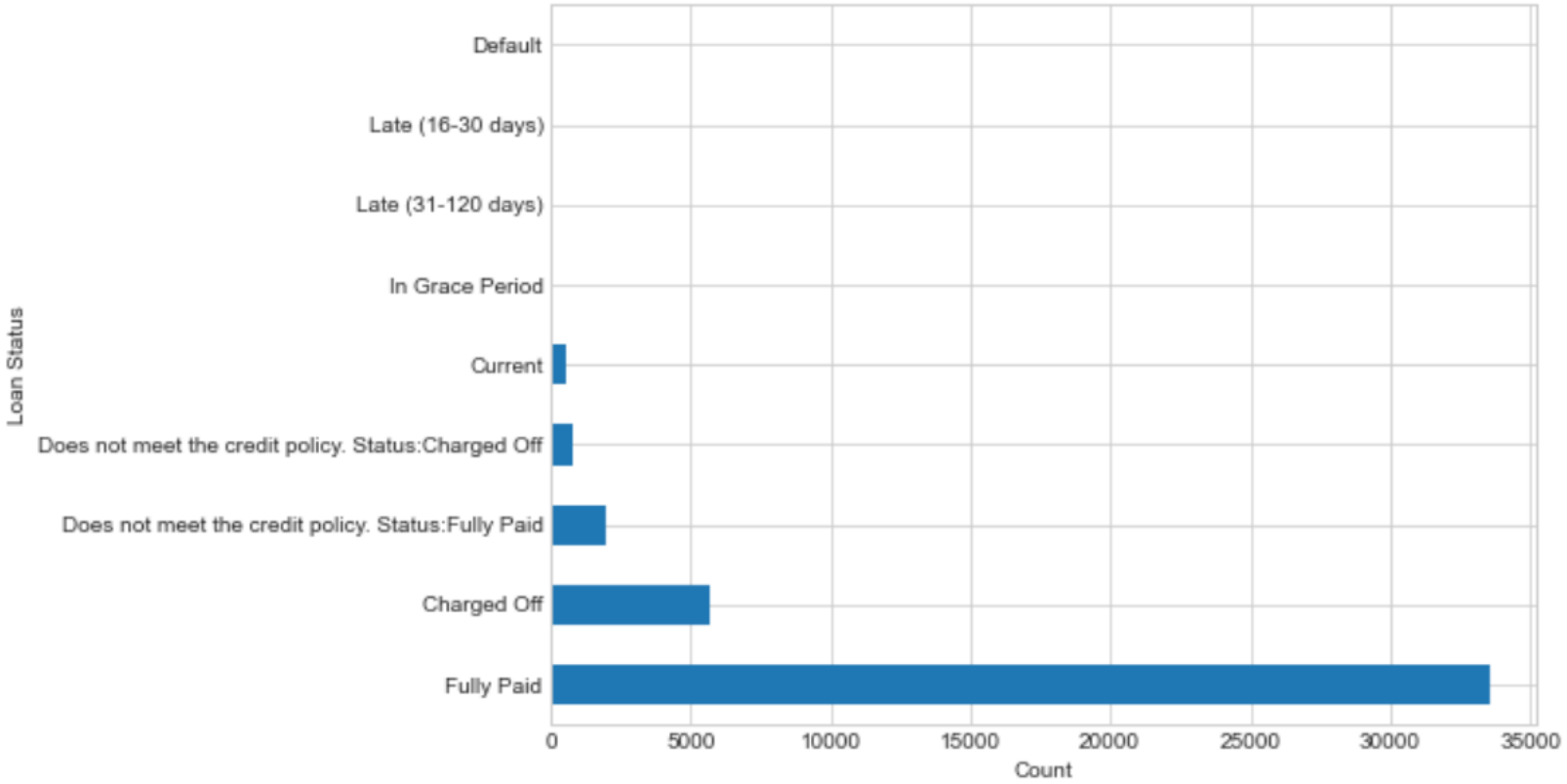
Late(30-120)->Loan hasn't been paid in 31 to 120 days (late on the current payment).

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.

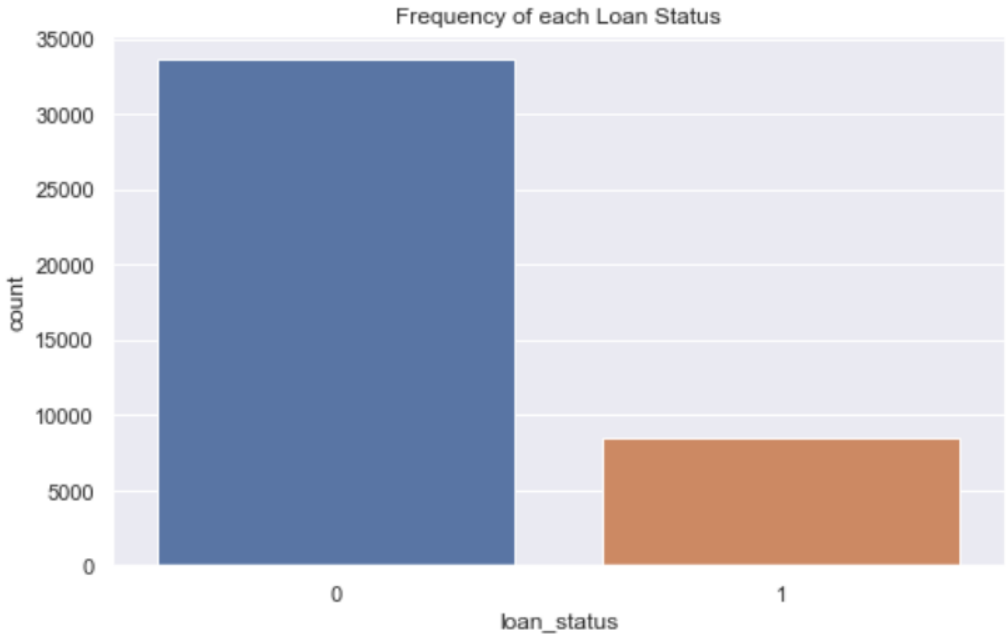


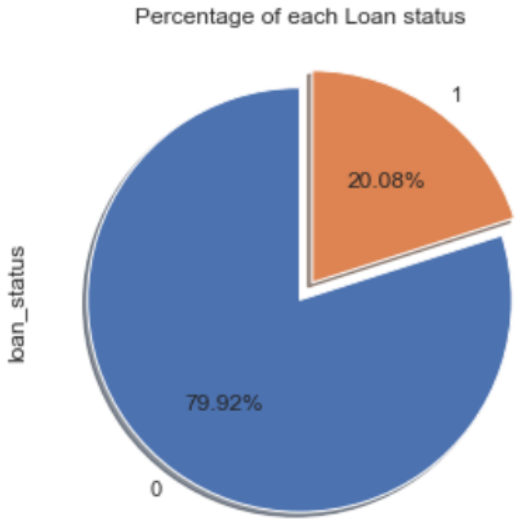
Explaining our Target feature: (contd.)



Explaining our Target feature: (contd.)

- **Distribution of our Target Feature**





Conclusion from the plots above:

Our dataset is clearly imbalanced with a ratio of 20:80 in favour of class 0 levels

Features that may lead to Data Leakage:

- **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.

Data Analysis and Feature Engineering:

Exploratory Data Analysis(EDA)

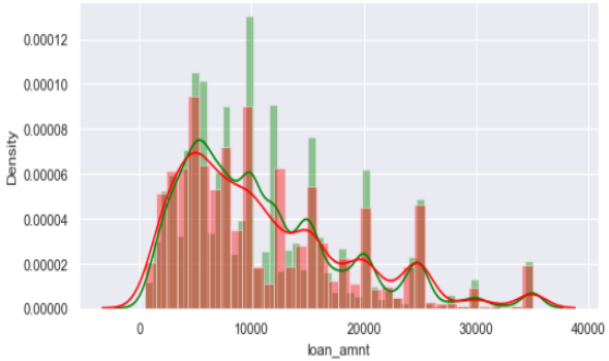
- Numerical Features
- Categorical Features

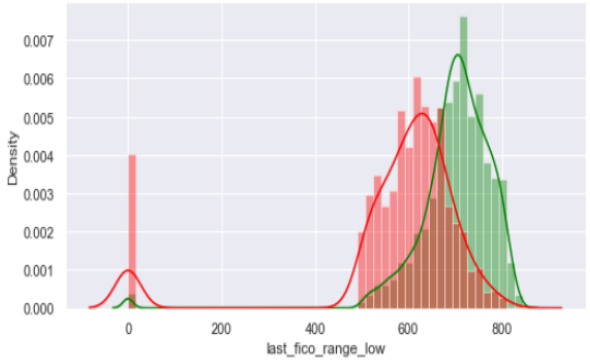
Missing value Imputation (KNN)

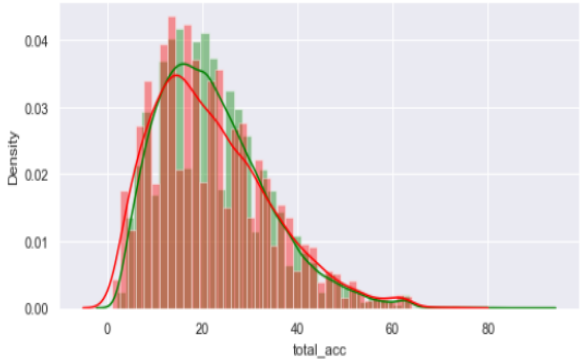
Dealing with class imbalance

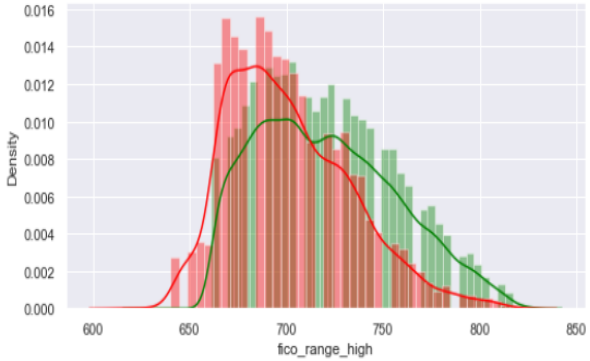
Exploratory Data Analysis(EDA)

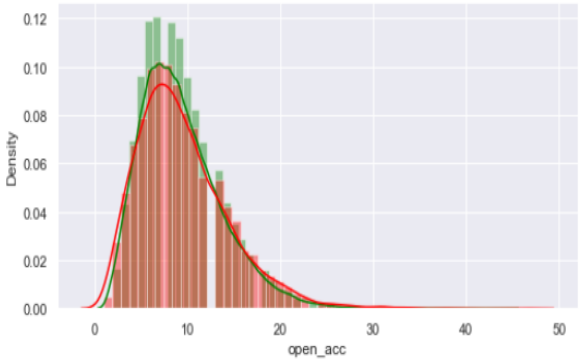
We have only chosen a specific number of features for EDA as we had a huge number of features to work with initially

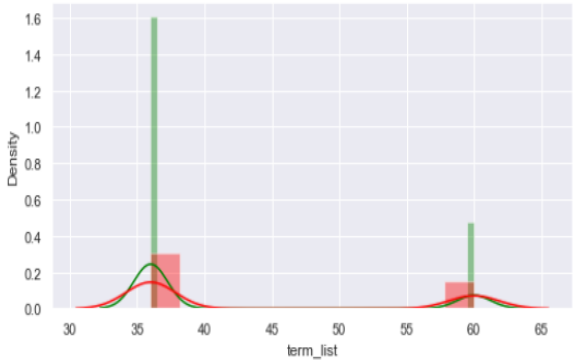






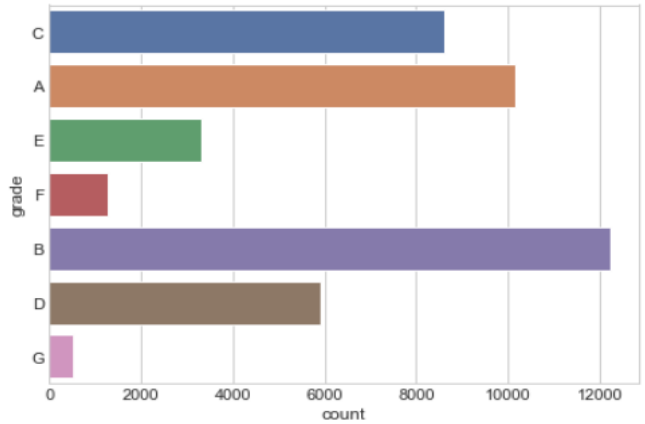


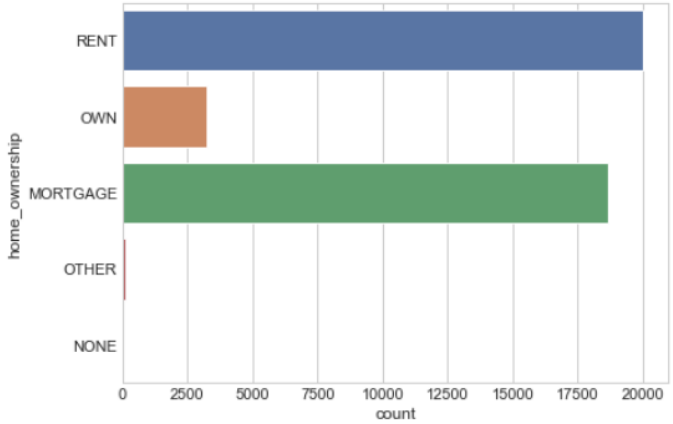


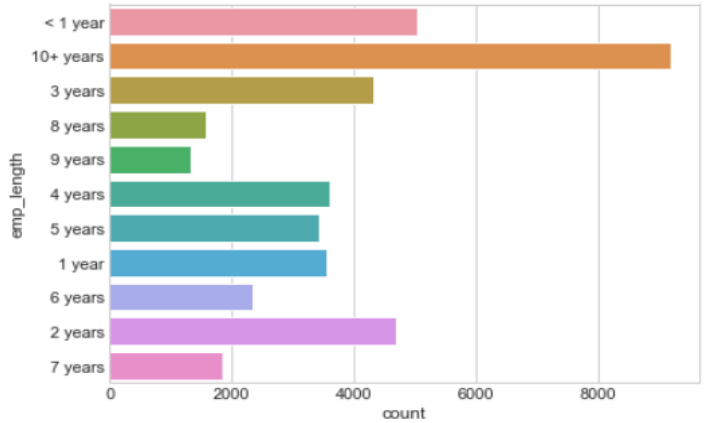


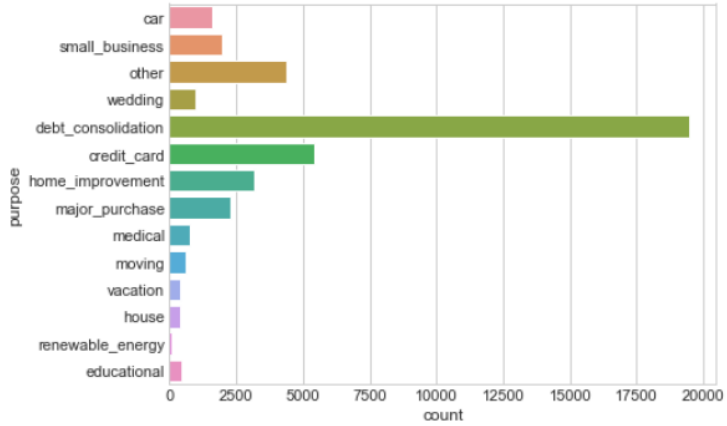
Exploratory Data Analysis (contd.)

EDA of categorical features:

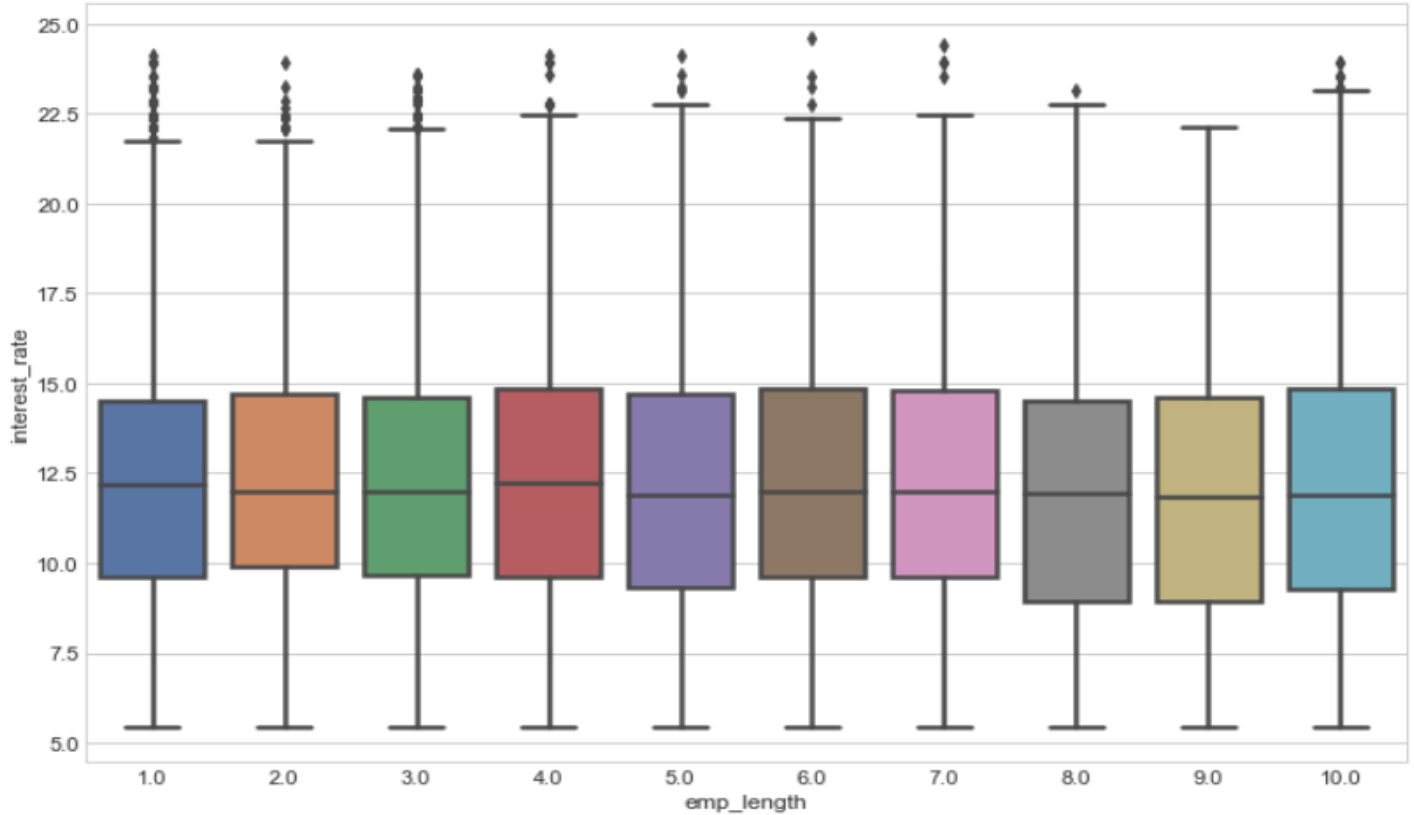




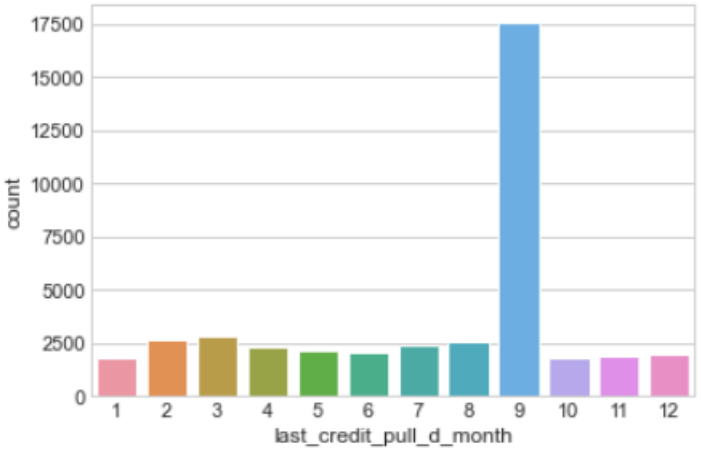


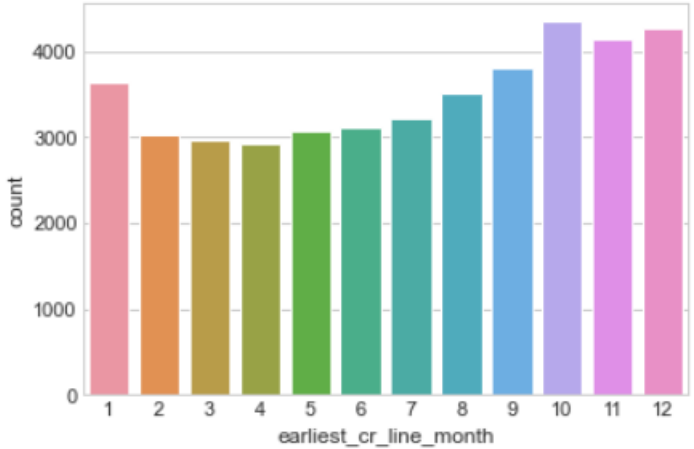


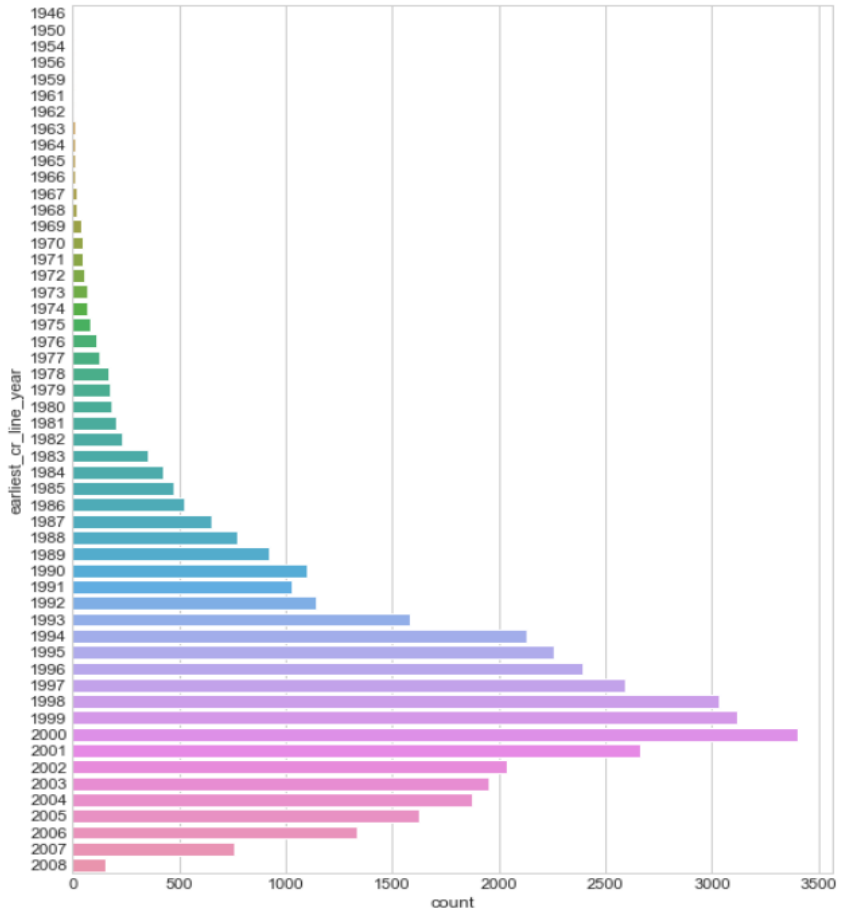
Exploratory Data Analysis(contd.)



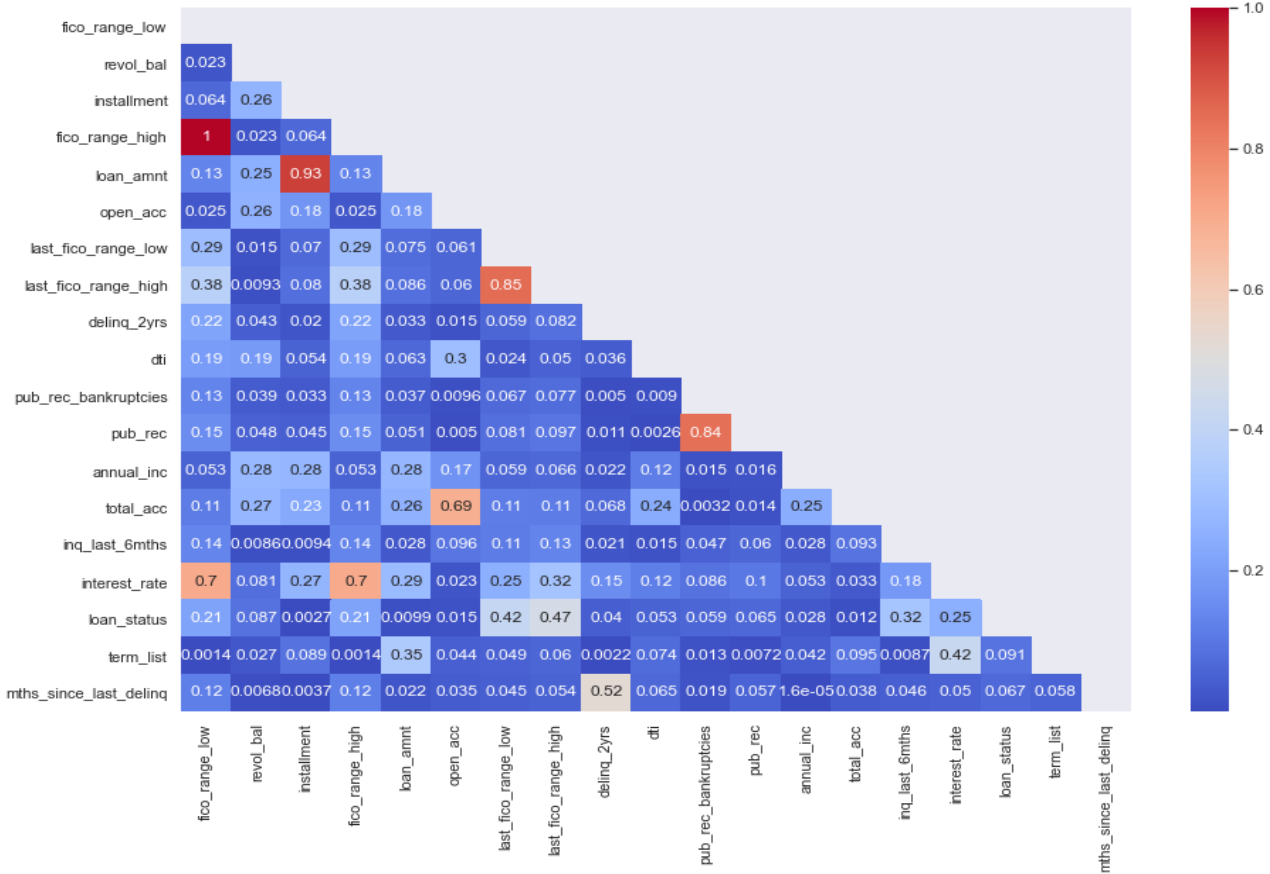
Exploratory Data Analysis(contd.)







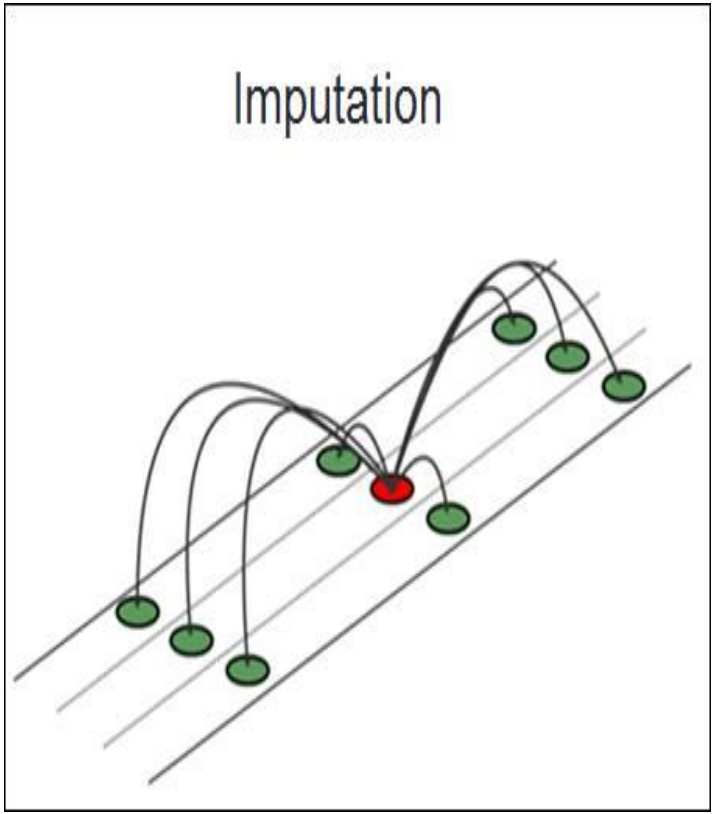
Exploratory Data Analysis(contd.)



Missing Value Imputation (KNN)

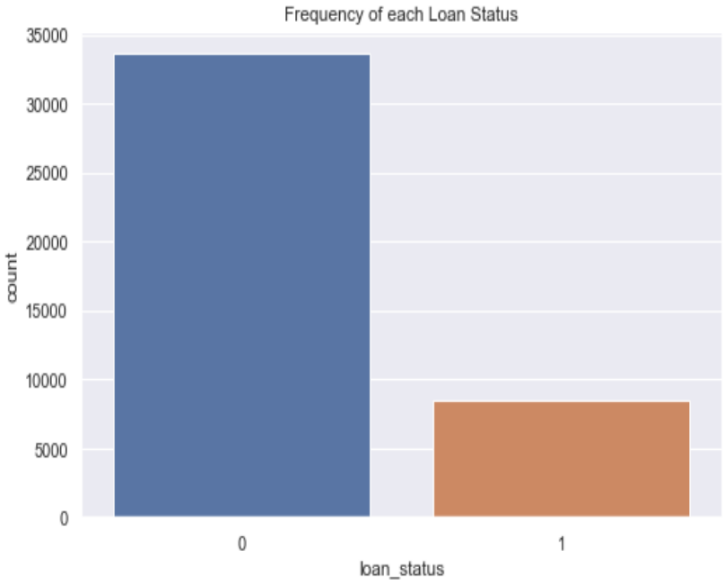
- Imputation for completing missing values using k-Nearest Neighbors. Each sample's missing values are imputed using the mean value from n_neighbors nearest neighbors found in the training set. Two samples are close if the features that neither is missing are close.

- The KNN Imputer class provides imputation for filling in missing values using the k-Nearest Neighbors approach. By default, an 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.



Dealing with class imbalance:

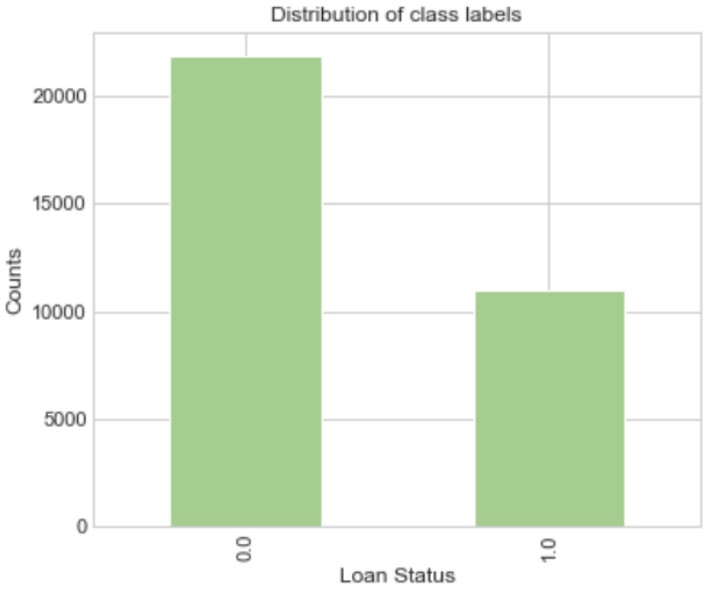
- Used SMOTEENN to handle class imbalance



Before applying SMOTEENN

Under sampling the majority class

Over sampling the minority class



After applying SMOTEENN (sampling_strategy=0.5)

Model Fitting and Evaluation:

DEFINING A BASELINE MODEL

MODEL SELECTION

HYPERPARAMETER TUNING

MODEL FITTING AND EVALUATION

Defining a Baseline Model:

- We have defined a baseline model such that it takes an ML model, the train and test datasets and the train and test labels as inputs

- The function returns us the following metric scores corresponding to a model including test accuracy, recall and precision score, f1 score, auc score, confusion matrix and all the corresponding scores for training set as well

- We have tested our model using a Logistic Regression model initially

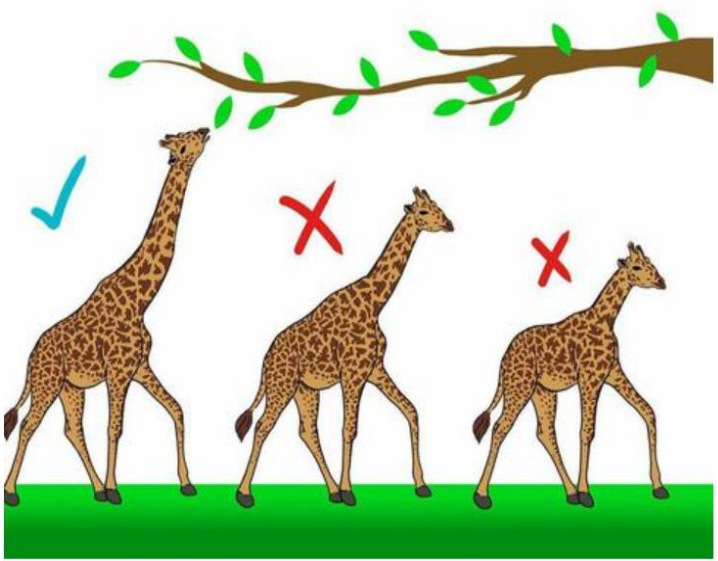
- Both the train and test scores are calculated to check whether our model is getting overfitted or not

	Logistic Regression
test_accuracy	0.903811
recall_test	0.839097
precision_test	0.868828
f1_test	0.853704
auc_test	0.887715
cm_test	[[5118, 348], [442, 2305]]
train_accuracy	0.90867
precision_train	0.876745
recall_train	0.845892
f1_train	0.861042
auc_train	0.893059
cm_train	[[15415, 980], [1270, 6971]]

Model Fitting using XG Boost Classifier

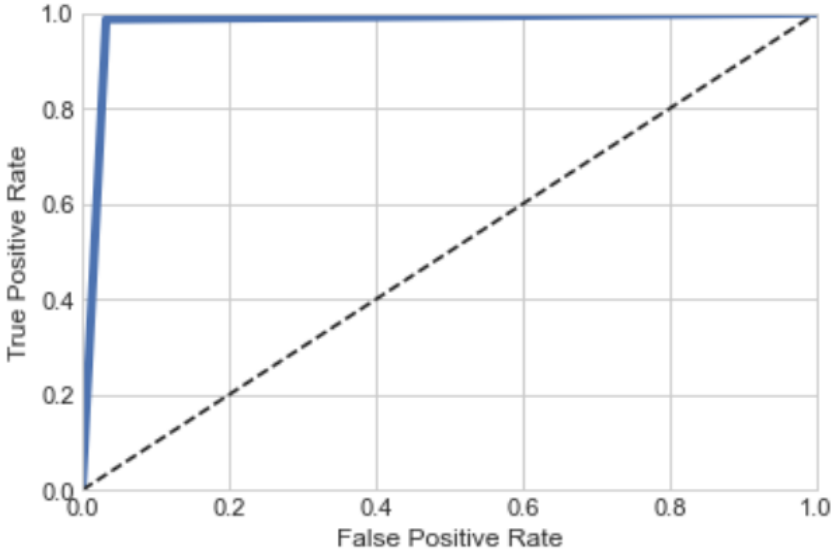
- We have used RandomizedSearchCV for hyperparameter tuning for our XGB algorithm

  - importance_type = 'gain'
  - max_depth= 13
  - learning rate = 0.05,
  - gamma = 0.2
  - min_child_weight = 1,
  - scale_pos_weight = 90,
  - n_estimators = 1000

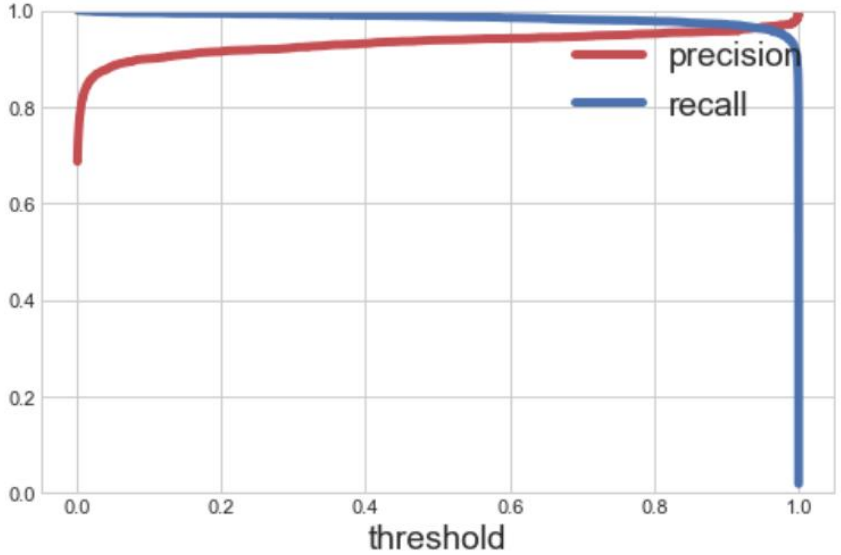


Model Evaluation:

AUC-ROC curve



Precision-Recall Curve



Model Interpretation using SHAP

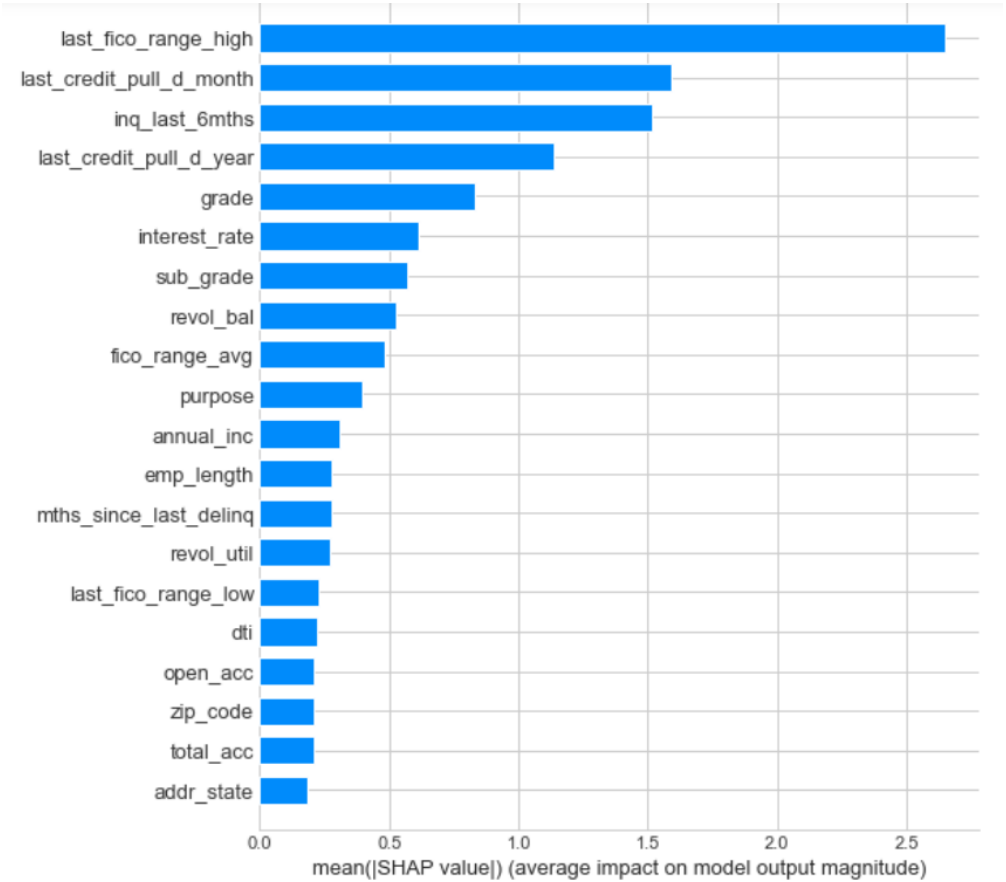
- Feature Importance Plot
- SHAP values (Impact on model plot)
- Force Plots

Feature Importance Plots (SHAP)

Conclusion from the SHAP value (feature importance) plot:

- The first plot gives us the feature importance corresponding to each feature. Our model says that '**last_fico_range_high**' is the most important feature in our dataset followed by '**last_credit_pull_month**'.

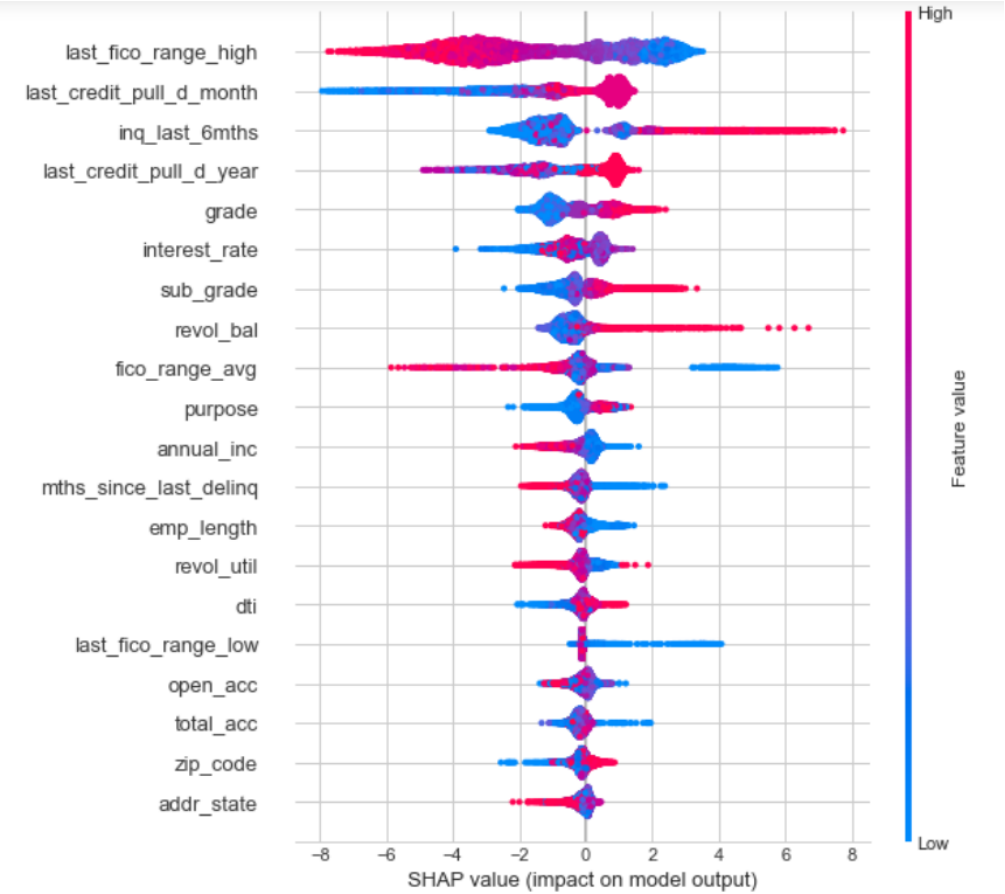
- '**addr_state**' and '**account balance**' are the features which are the least important.



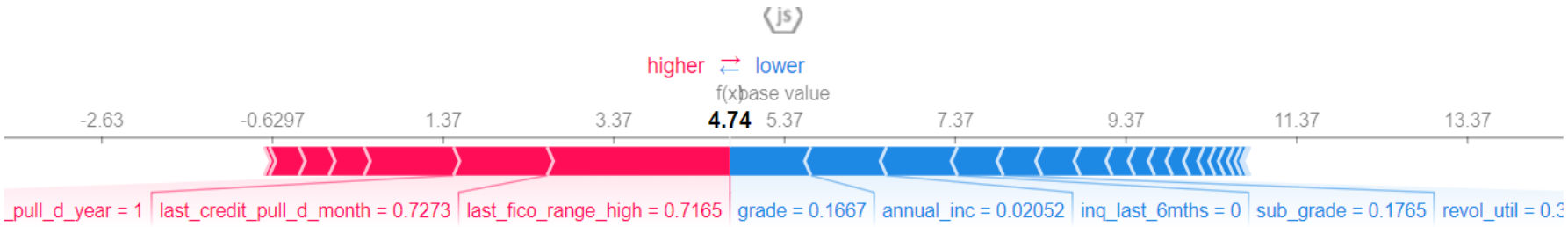
Impact of features on Target Feature(SHAP):

Conclusion from the SHAP value (impact) plot:

- This plot shows the impact of each feature in our model.
- The features are arranged in decreasing order of their importance.
- The red and blue color corresponds to higher and lower impact on the model output respectively
- For example, negative impact of lower_fico_range_high' and positive impact of 'inq_last_6_months' will lead to a person defaulting a loan.



Force Plots(SHAP):



Conclusion from the Force Plot (SHAP):

- This is a detailed plot which gives us information of the features on several grounds.
- Red impacts features moving the model to the class 1 level and vice-versa.
- We also have a base value which is simply the mean prediction
- The arrows represent the features pushing the model prediction to the higher(right) and lower(left) side
- We have also calculated the mean of each feature corresponding to the test set and the values of each feature in the force plot is compared to the mean value. If the value in force plot < mean_value, the prediction will be driven to the right and vice versa.

Conclusion and Plans to improve:

- The last_fico_range, grade, inq_last_6month features were found to be the most relevant for predic ting loan default in. The current model tries to predict default biased data from credit analysts' grade and assigned interest rate. The XGBC models provide substantial improvements on traditional credit screening. A recall score significantly and robustly above 95%, and ROC scores also above 95%. The fe atures provided to the model in our study generalize to any lending activity and institution, beyond P 2P lending. The present work could, therefore, be augmented in order to predict loan default risk wi thout the need for human credit screening.

- Our XGBoost model is working extremely well, however the complexity of the model may be further increased by applying Deep Neural Networks or by better hyperparameter tuning techniques.

- In the bank loan behavior 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 behaviors in the earlier stage and conduct the corresponding actions to reduce the possible loss.

Conclusion and Plans to improve:

- The last_fico_range, grade, inq_last_6month features were found to be the most relevant for predic ting loan default in. The current model tries to predict default biased data from credit analysts' grade and assigned interest rate. The XGBC models provide substantial improvements on traditional credit screening. A recall score significantly and robustly above 95%, and ROC scores also above 95%. The fe atures provided to the model in our study generalize to any lending activity and institution, beyond P 2P lending. The present work could, therefore, be augmented in order to predict loan default risk wi thout the need for human credit screening.

- Our XGBoost model is working extremely well, however the complexity of the model may be further increased by applying Deep Neural Networks or by better hyperparameter tuning techniques.

- In the bank loan behavior 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 behaviors in the earlier stage and conduct the corresponding actions to reduce the possible loss.

Challenges:

- A lot of research was needed to get a brief understanding of the features in the dataset
- Huge chunks of data were 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.