LOAN DEFAULT PREDICTION

TABLE OF CONTENTS

PROBLEM STATEMENT

DATA COLLECTION

EXPLORATORY DATA ANALYSIS

DATA WRANGLING

MACHINE LEARNING

CONCLUSION

PROBLEM STATEMENT

According to the American Bankers Association survey, Banks with a Home equity line of credit and Home equity loans average a 1 percent return on assets (ROA) overall, but earn 1.5 percent on their HELOCs and 1.25 percent on their HELs, which means that these two products are more profitable for banks than their average loans. But the delinquency rate on Home equity loans is higher than all other types of consumer loans. It is imperative for banks to take calculated risks in this area for financial growth.

A Bank offers a home equity line of credit to its clients. As per the dataset, many of these accepted applicants (approximately 20%) have defaulted on their loans. By using geographic, demographic, and financial variables, the bank wants to build a model to predict whether an applicant will default.

DATA COLLECTION

The data set HMEQ reports characteristics and delinquency information for 5,960 home equity loans. A home equity loan is a loan where the obligor uses the equity of his or her home as the underlying collateral.

The data was obtained from the website http://www.creditriskanalytics.net in csv format.

The data set HMEQ reports characteristics and delinquency information for 5,960 home equity loans. The data set has the following attributes :

- BAD: 1 = applicant defaulted on loan or seriously delinquent; 0 = applicant paid loan
- LOAN: Amount of the loan request
- MORTDUE: Amount due on existing mortgage
- VALUE: Value of current property
- REASON: DebtCon = debt consolidation; HomeImp = home improvement
- JOB: Occupational categories
- YOJ: Years at present job
- DEROG: Number of major derogatory reports
- DELINQ: Number of delinquent credit lines
- CLAGE: Age of oldest credit line in months

- NINQ: Number of recent credit inquiries
- CLNO: Number of credit lines
- DEBTINC: Debt-to-income ratio

MISSING VALUE IMPUTATION

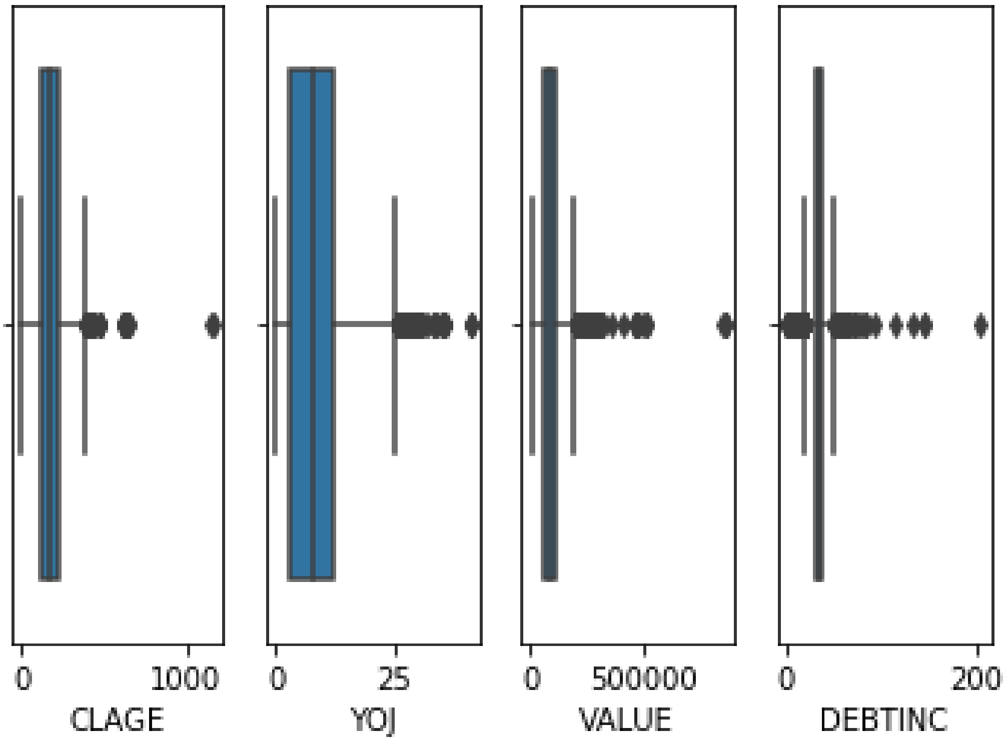
Almost 40% of the data had missing values and hence it needs to be imputed.

The most common category is replaced for missing values of categorical variable. The ‘Other’ value will be imputed to the missing values of the attribute JOB since it already accounts for 40% of the existing values.

Similarly each of the numerical variables DEROG, DELINQ,NINQ having missing values are imputed with most common value(mode) which is found using the value_counts() function on the dataframe.

IDENTIFYING AND HANDLING OUTLIERS

Box plots from the Seaborn package is utilized to identify and visulaize the outliers



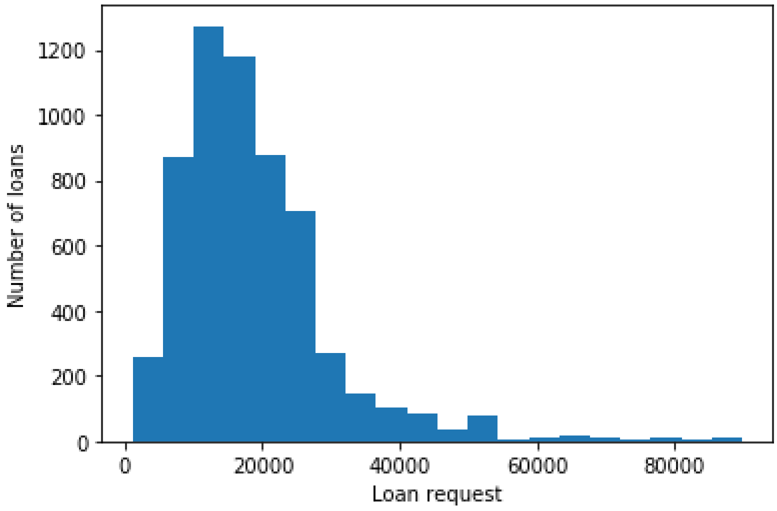
CLAGE :We found the outlier value 1168 months(or 97 years) which is logically incorrect for the age of oldest credit line.

YOJ: Though value 41 looks as an outlier, the value is still logical and hence will not be removed from data.

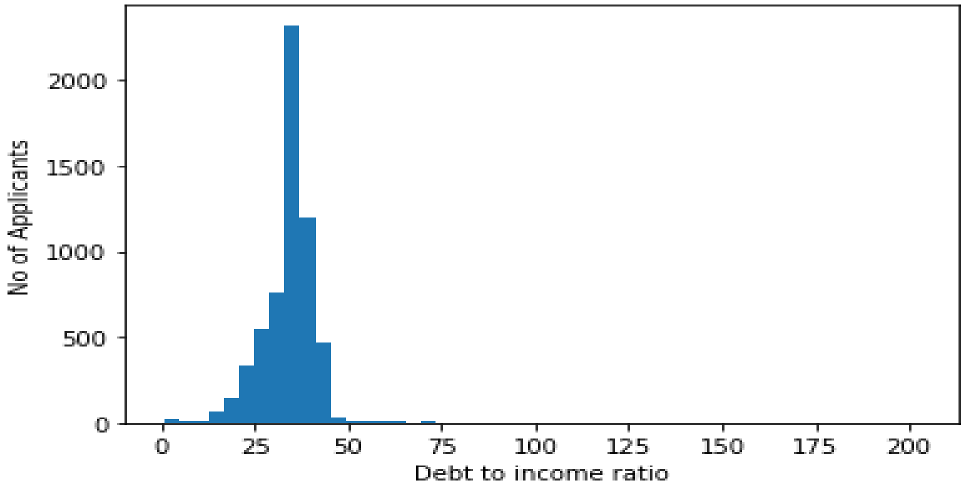
VALUE: The value 855909 is logical and hence will not be removed from data.

DEBTINC: The value 203 is definitely an incorrect value for the DEBTINC(debt to income ratio) and will be excluded.

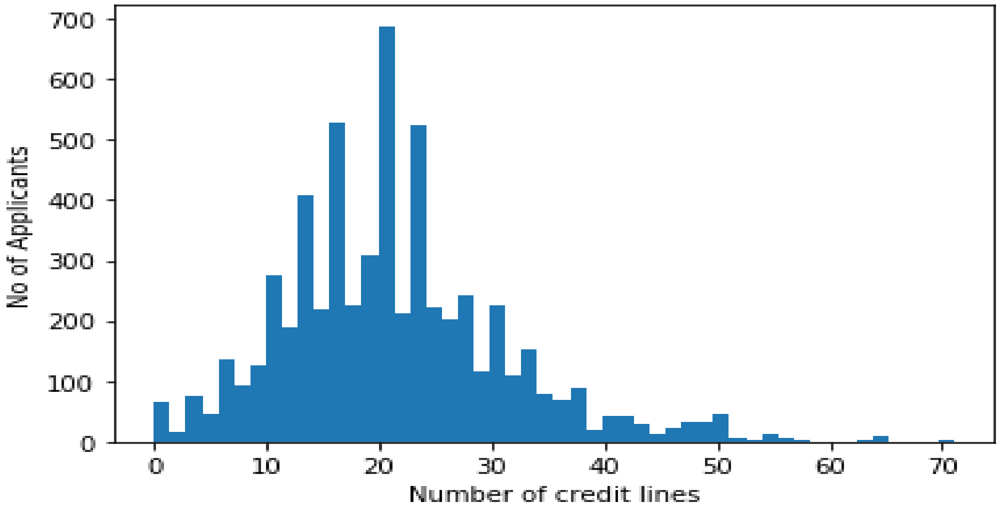
HISTOGRAM : LOANS ,DEBTINC,CLNO



Most of the loans were provided less than or equal to 20,000 and most of them lie between 10,000 and 20,000.



As per the consumer financial protection bureau , the debt to income ratio ideally should be less than 43 for a good applicant.We can infer that most of the applicants here fall within the criteria and are less than 43%

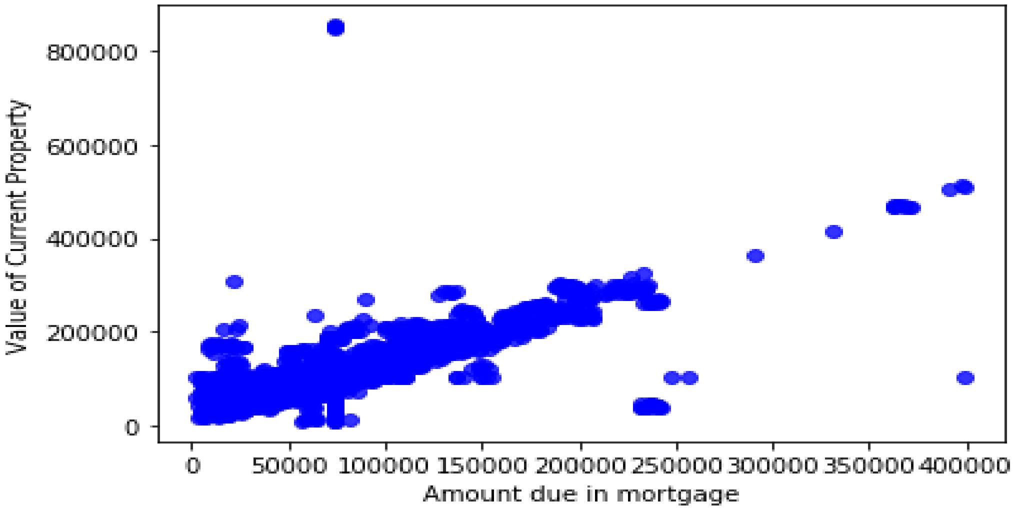


According to data compiled by TransUnion from the first quarter of 2017, the average number of credit cards per person is **2.69**. Most of the applicants in our data have less than or equal to 20 credit lines. Hence applicants having more than 50 credit lines, doesn't necessarily mean they will have other financial implication or lower credit score.

BIVARIATE VISUALIZATION

SCATTER PLOT:

Amount Due vs Value of Current property

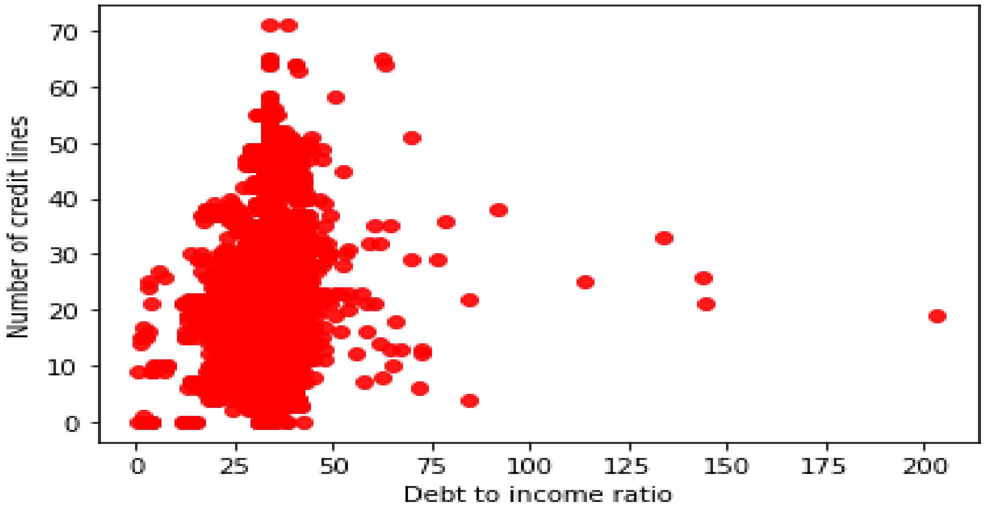


From the plot,we can clearly identify a potential BAD LOAN by following reason

1. In the bottom right corner, the amount due is Greater than value of property which indicates applicant has more probability to become delinquent.

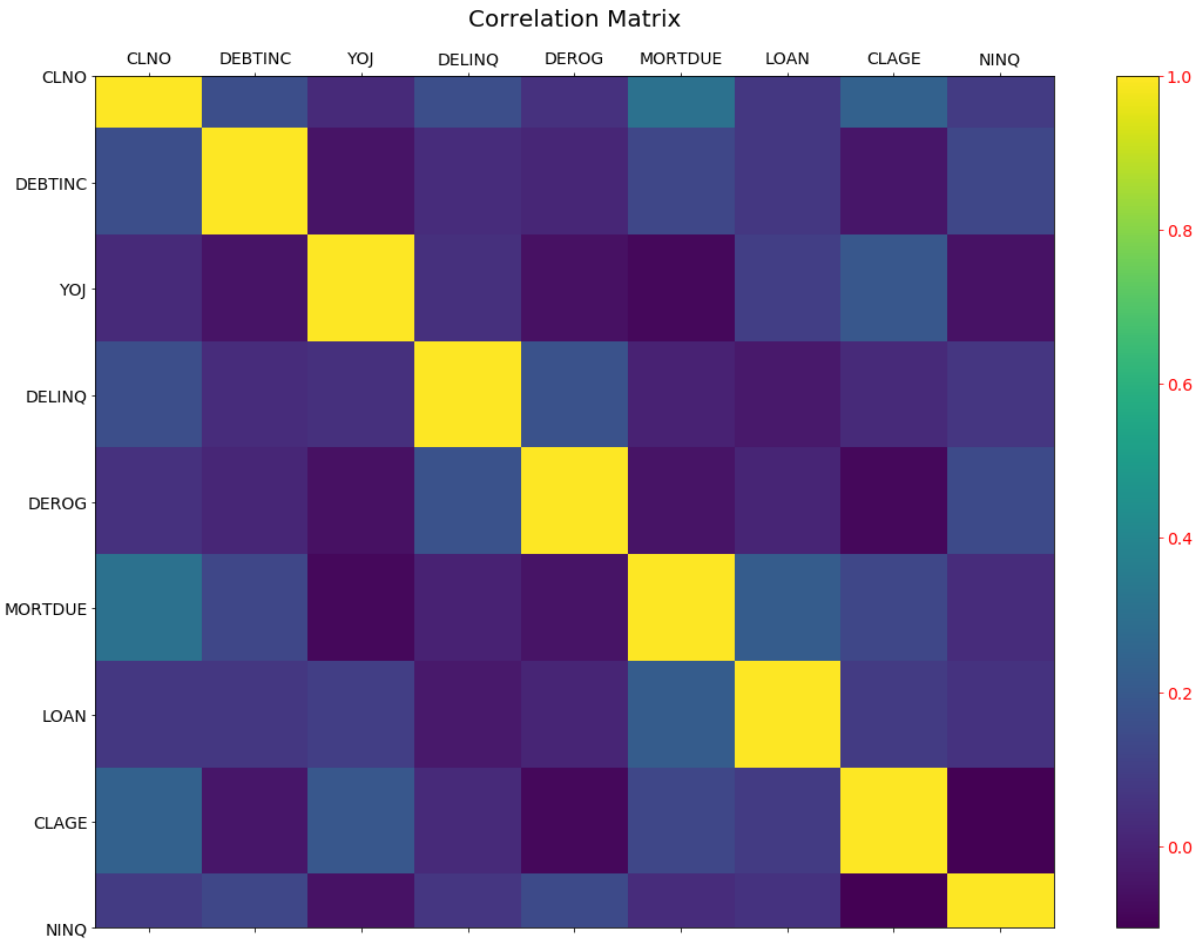
2. Whereas, at the top of the curve ,applicants have high value of current property and less amount due in mortgage which indicates they have less probability to become delinquent.

Debt to income ratio vs Number of credit lines



From this plot, we can infer that applicants who have more credit lines (more than 50) are having less than 43(favorable) in debt to income ratio.

CORRELATION BETWEEN THE VARIABLES



Findings from the Correlation matrix,

1. The pearson Coefficient of 0.305 between MORTDUE and CLNO is a moderate positive correlation. This can be translated as applicants who have more credit lines tend to have more amount due on their existing mortgage.

2. The pearson Coefficient of 0.216 between MORTDUE and LOAN is a moderate positive correlation. This can be translated as applicants who have requested more loans requested for more LOAN amount.

From these two findings, we can consider applicants with high MORTDUE are potentially risky applicants.

3. With coefficient of 0.238 ,we can infer the applicants who have more credit lines have age of oldest credit line as high.

Note: The other variables do not have a significant correlation between them .

ATTRIBUTE RELEVANCE ANALYSIS

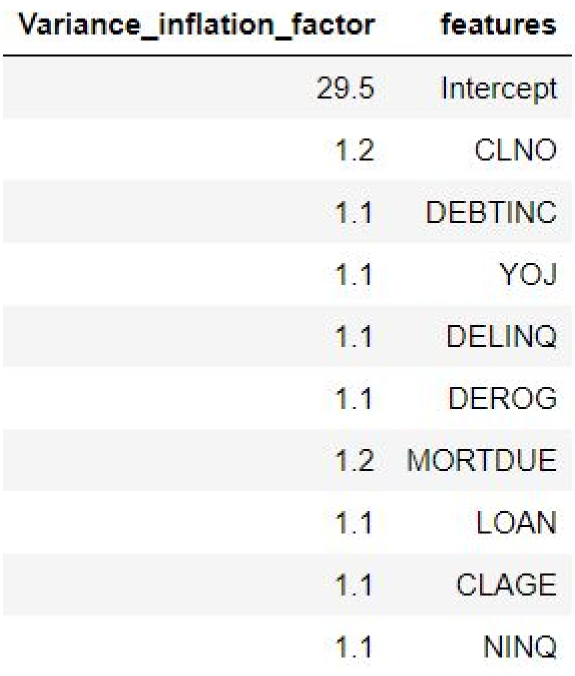
Created a function to calculate the weight of evidence and IV for each of the attribute. THe IV scores of the attributes are as follows ,

ATTRIBUTE	IV SCORE
LOAN	0.71
MORTDUE	0.29
VALUE	0.69
REASON	0.01
JOB	0.08
YOJ	0.11
DEROG	0.35
DELINQ	0.49
CLAGE	0.39
NINQ	0.17
CLNO	0.16
DEBTINC	1.06

The good predictor cut off value for the IV score is 0.10 or more.

The attribute REASON has an IV score: 0.01 and hence it is taken removed from the dataframe.

Calculating Variance Inflation Factor to detect Multicollinearity



Variance_inflation_factor	features
29.5	Intercept
1.2	CLNO
1.1	DEBTINC
1.1	YOJ
1.1	DELINQ
1.1	DEROG
1.2	MORTDUE
1.1	LOAN
1.1	CLAGE
1.1	NINQ

As per the above VIF values ,all these variables have values less than the threshold value of 5. Hence, MultiColinearity is NOT present and we are not dropping any other variables from our dataframe.

Data Split:

Target variable count: BAD LOAN : 1188 ; GOOD LOAN : 4771

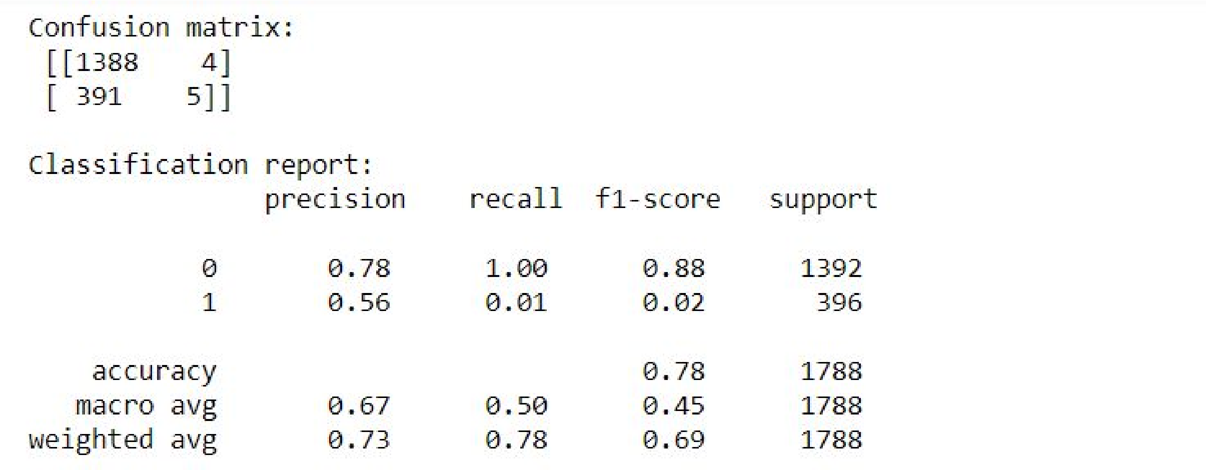
There is no class imbalance and the data is split with 70% of our data as training dataset and 30% as test dataset.

MACHINE LEARNING

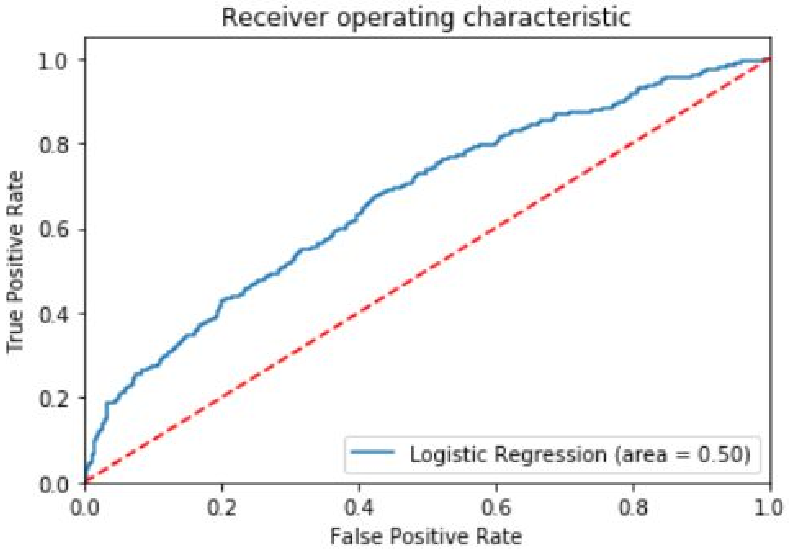
LOGISTIC REGRESSION

A Logistic regression classifier is initiated and the model is trained using the training dataset and validated for accuracy on the test data set.

Accuracy of logistic regression classifier on test set: **0.78**



```
Confusion matrix:
 [[1388    4]
 [ 391    5]]

Classification report:
               precision    recall  f1-score   support

           0       0.78      1.00      0.88      1392
           1       0.56      0.01      0.02       396

    accuracy                           0.78      1788
   macro avg       0.67      0.50      0.45      1788
weighted avg       0.73      0.78      0.69      1788
```



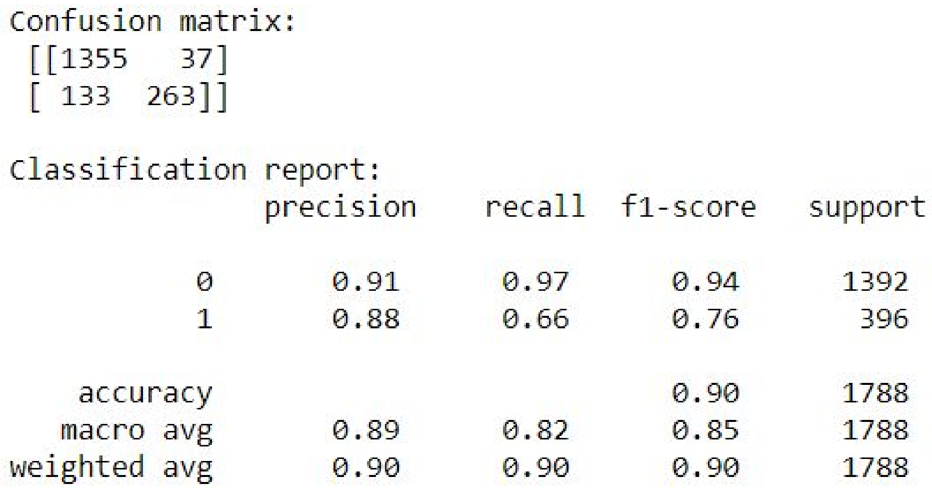
From the ROC curve , the resulting Area under the curve(AUC) value is 0.50

RANDOM FOREST

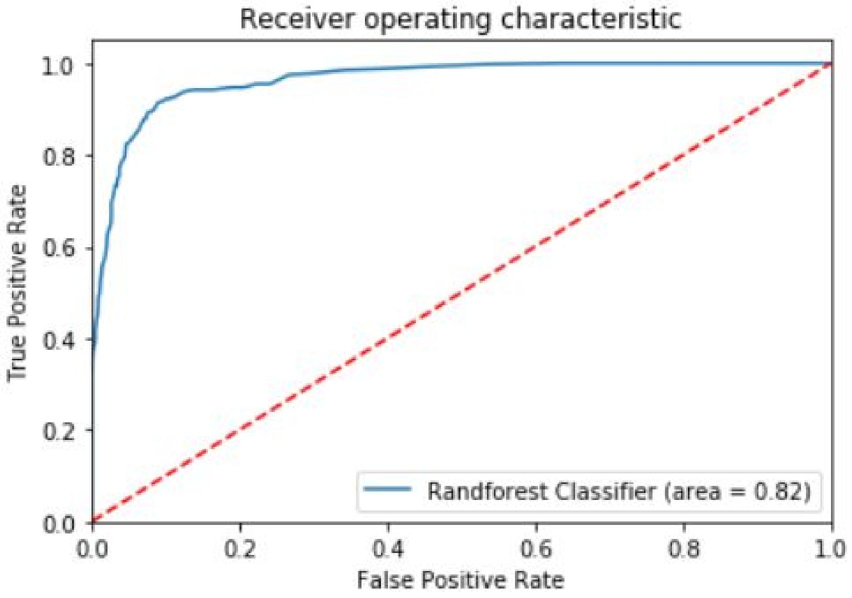
A Random Forest classifier is created using 100 trees(n_estimators=100)

Accuracy of Random Forest classifier on test set: **0.90.**

Confusion matrix,Classification report and ROC Curve on test data is created to evaluate the model performance.



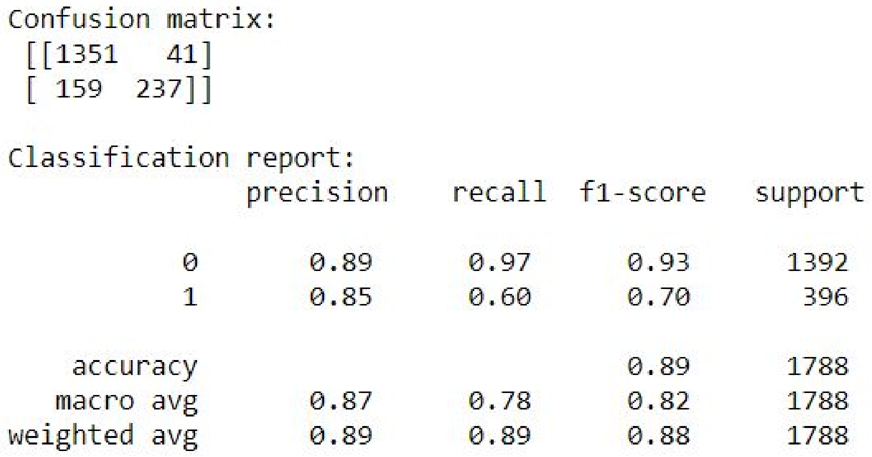
```
Confusion matrix:
 [[1355   37]
 [ 133  263]]

Classification report:
              precision    recall  f1-score   support

           0       0.91      0.97      0.94      1392
           1       0.88      0.66      0.76       396

    accuracy                           0.90      1788
   macro avg       0.89      0.82      0.85      1788
weighted avg       0.90      0.90      0.90      1788
```



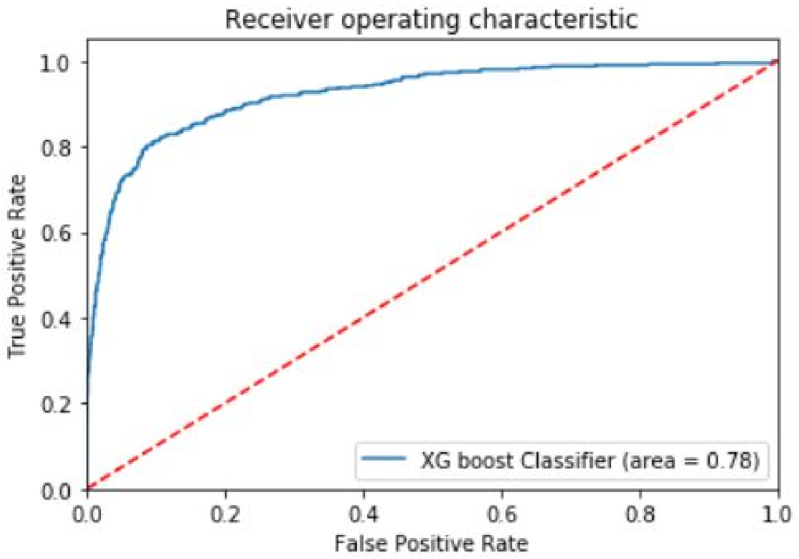
From the ROC curve , the resulting Area under the curve(AUC) value is 0.82. The random forest outperformed the logistic regression model.

GRADIENT BOOSTING

The Accuracy of XGBoost classifier on test set is **0.89 ,** which is marginally close to our randomforest model’s accuracy of 0.90.



```
Confusion matrix:
 [[1351   41]
 [ 159  237]]

Classification report:
               precision    recall  f1-score   support

           0       0.89      0.97      0.93      1392
           1       0.85      0.60      0.70       396

    accuracy                           0.89      1788
   macro avg       0.87      0.78      0.82      1788
weighted avg       0.89      0.89      0.88      1788
```



Also, the resulting Area under the curve(AUC) value of 0.78 is still lower than our RandomForest model value 0.82.

Comparing the 3 different classifier models, Random Forest classifier has the highest accuracy value and AUC value (0.82) and will be chosen for further hyperparameter tuning.

HYPERPARAMETER TUNING

The initial K fold cross validation (with roc_auc as scoring) is performed on the random forest model and the mean value is 0.9624 .

RANDOM SEARCH CV

The following grid is created for the random search cv method,

```
{'bootstrap': [True],
 'max_depth': [10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, None],
 'max_features': ['auto', 'sqrt'],
 'min_samples_leaf': [1, 2, 4],
 'min_samples_split': [2, 5, 10],
 'n_estimators': [200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000]}
```

A new random forest classifier(ranfor_rand) is created and used in the RandomizedSearchCV along with updated parameter values with above grid and fitted to our model.

This method yields a cross validation score of 0.96308

Accuracy with Random search cv on test set: **0.96**

GRID SEARCH CV

The following grid is created for the Grid search cv method,

```
param_grid = {
    'bootstrap': [True],
    'max_depth': [80, 90, 100, 110],
    'max_features': [2, 3],
    'min_samples_leaf': [3, 4, 5],
    'min_samples_split': [8, 10, 12],
    'n_estimators': [100, 200, 300, 1000] }
```

A new random forest classifier(ranfor_grid) is created and used in the GridSearchCV along with updated parameter values with above grid and fitted to our model.

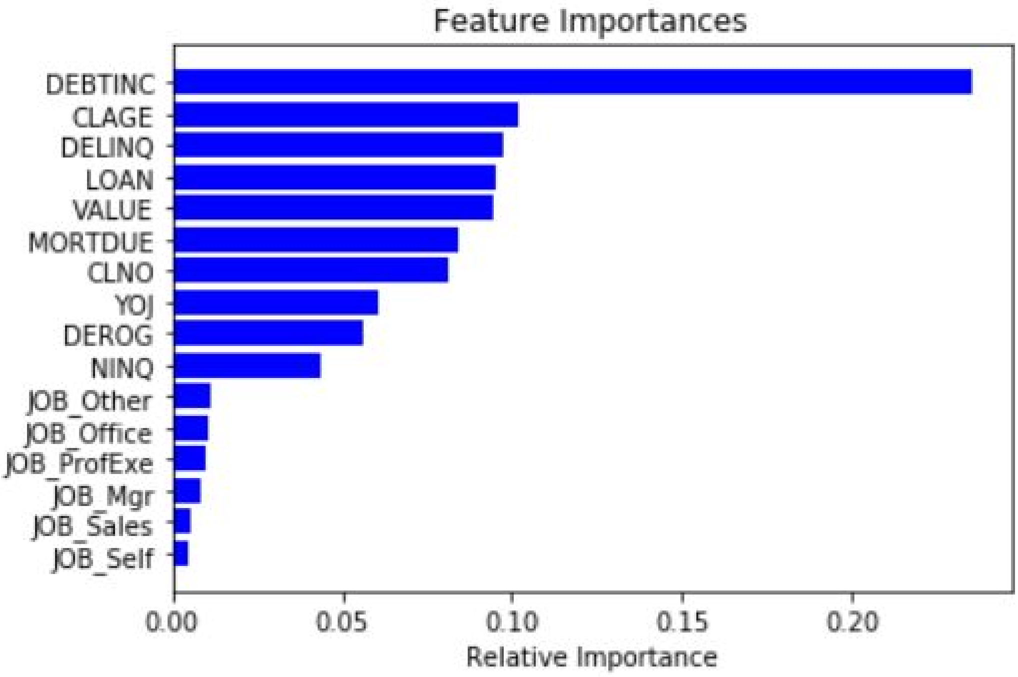
This method yields a cross validation score of 0.9544 and Accuracy with Random search cv on test set: **0.90**

The RANDOMSEARCH classifier model yielded better accuracy score and higher cross validation score compared to the GRIDSEARCH.

FEATURE IMPORTANCE

The feature_importances_ parameter is used to find the most important attributes for the model.

The relative importance values are calculated and visualized using matplotlib



The Debt to income ratio(DEBTINC) turns out to be the most important feature with value of 0.24.

CONCLUSION:

The Bank can apply the RandomForest classifier model with updated hyperparameter values to predict the potential loan defaulters. The debt-to-income is a ratio that compares your monthly debt expenses to your monthly gross income. From our feature importance analysis, we learn that it contributes most among all the features . Evidence from studies of loans also suggest that borrowers with a higher debt-to-income ratio are more likely to run into trouble making monthly payments.