LOAN PREDICTION

Tinu Shiby

INTRODUCTION

Loan approval is a completely important procedure for banking businesses. The system of approval or rejection of mortgage applications. Recovery of loans is a first-rate contributing parameter in the economic statements of a bank. It may be very hard to expect the possibility of a fee of loan through the purchaser. In recent years many researchers worked on mortgage approval 5 prediction structures.ML techniques are very useful in predicting consequences for big quantities of information. In this project some ML algorithms like Logistic Regression, Decision Tree,Random Forest,etc are implemented and are expecting loan approval for customers. The experimental results conclude that the accuracy of XGBoost classifier is better in comparison to other algorithms.

Using data science, a ML model is made. On the basis of some training data set,it is capable of identifying if the loan applicant is ideal for the loan approval or not. Machine Learning algorithms like Decision Tree,Logistic Regression,Random Forest,etc. are used for the analysis.These are efficient algorithms that are followed for data analysis and prediction making. The system will look into some basic information of the applicant such as his/her profession,age,gender,marital status,etc.,and after analyzing all this information,using visualization and machine learning algorithms,it will come to a decision.

DATA ANALYSIS

The dataset contains various feature variables which are responsible for approval or rejection of a loan. Loan_Status is the target variable and it has two classes Yes and No,so clearly the statement is a classification problem. To start with the data analysis, i have imported the following libraries:

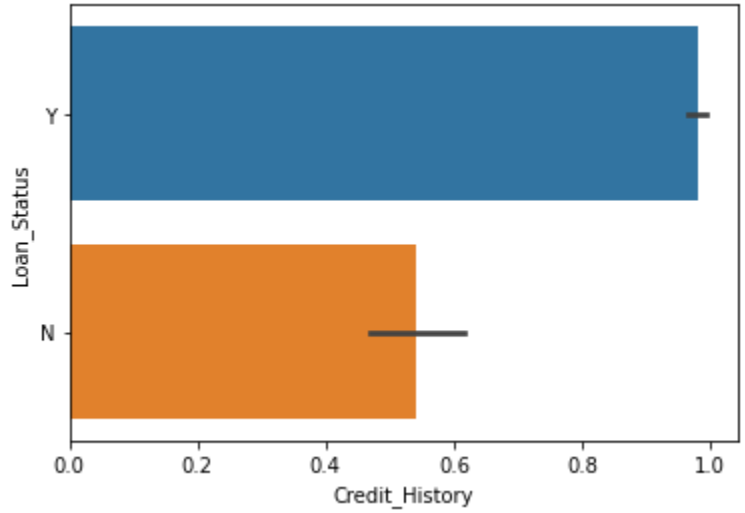
- Pandas
- Matplotlib
- Seaborn
- Standard Scalar
- Smote(for upsampling)
- Principal Component Analysis
- Train test split
- GridSearchCV
- Cross-validation score
- Roc-curve
- Roc-auc score
- Classification report

EDA

Exploratory data analysis is one of the most crucial steps in data analysis. It provides a better understanding of data set variables and the relationships between them. It can also help in determining if the statistical techniques you are considering for data analysis are appropriate.

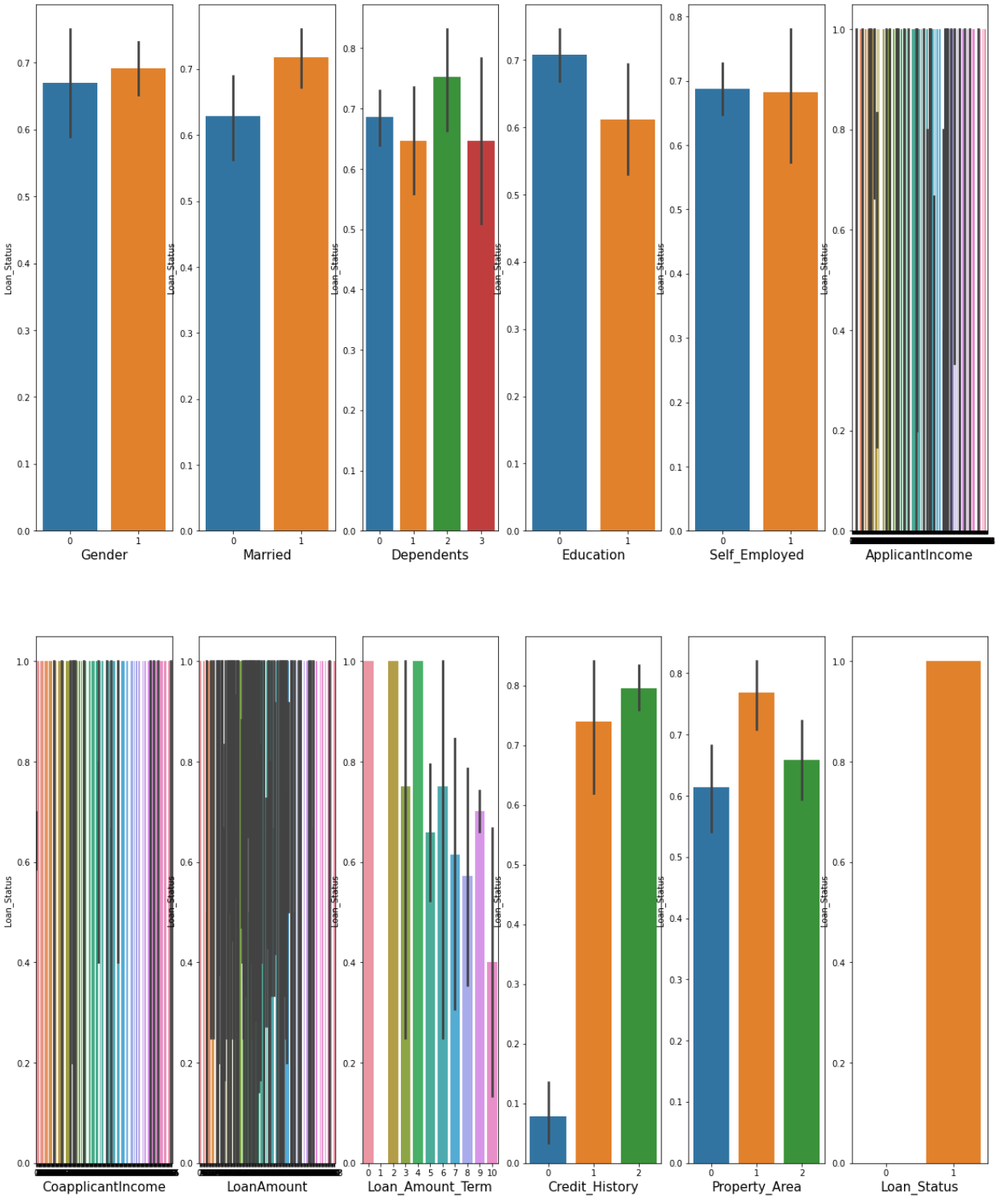
The dataset has 13 variables and 614 rows. In the dataset 7 variables have null values and it has both categorical variables as well as numerical variables.

Before replacing the null values with mean or mode, a bivariate analysis is done,



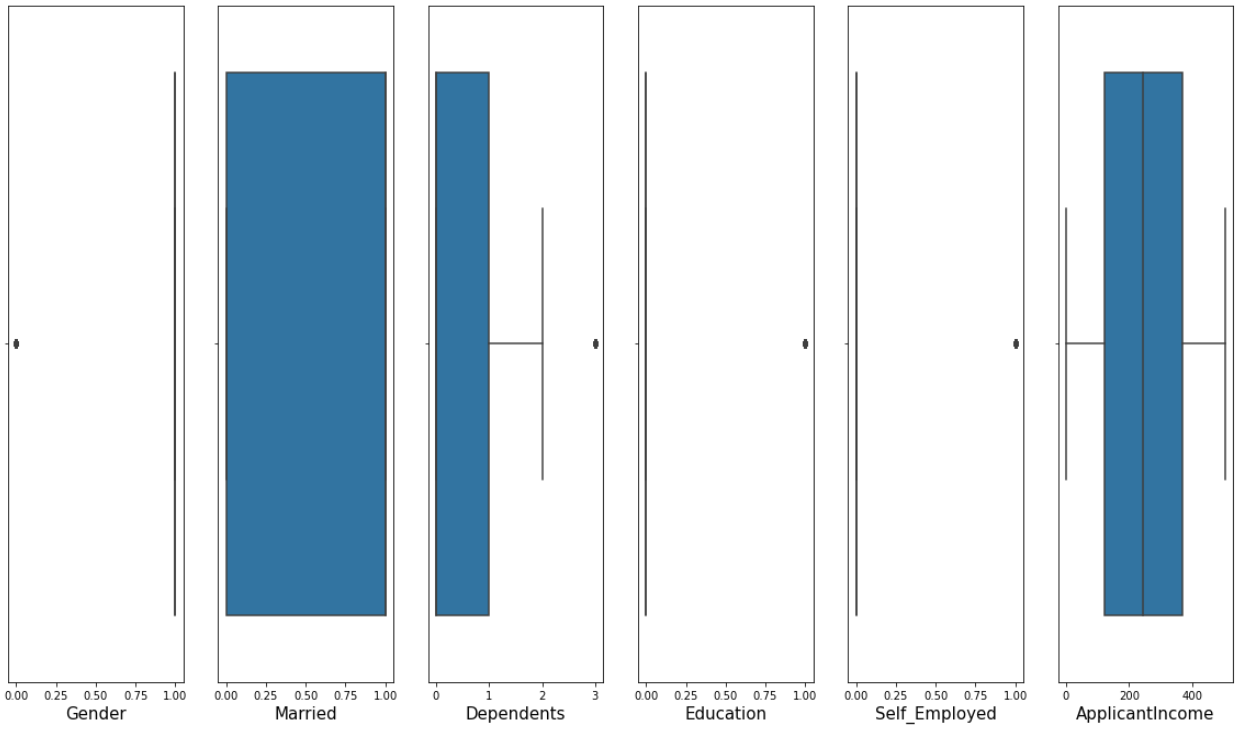
Then the null values on columns ‘Gender’,’Married’,’Dependents’,’Self_employed’ are replaced with the mode and ‘LoanAmount’,’LoanAmountTerm’ and ‘Credit History’ are replaced with mean. Again a bivariate analysis is done and the trend remains the same.

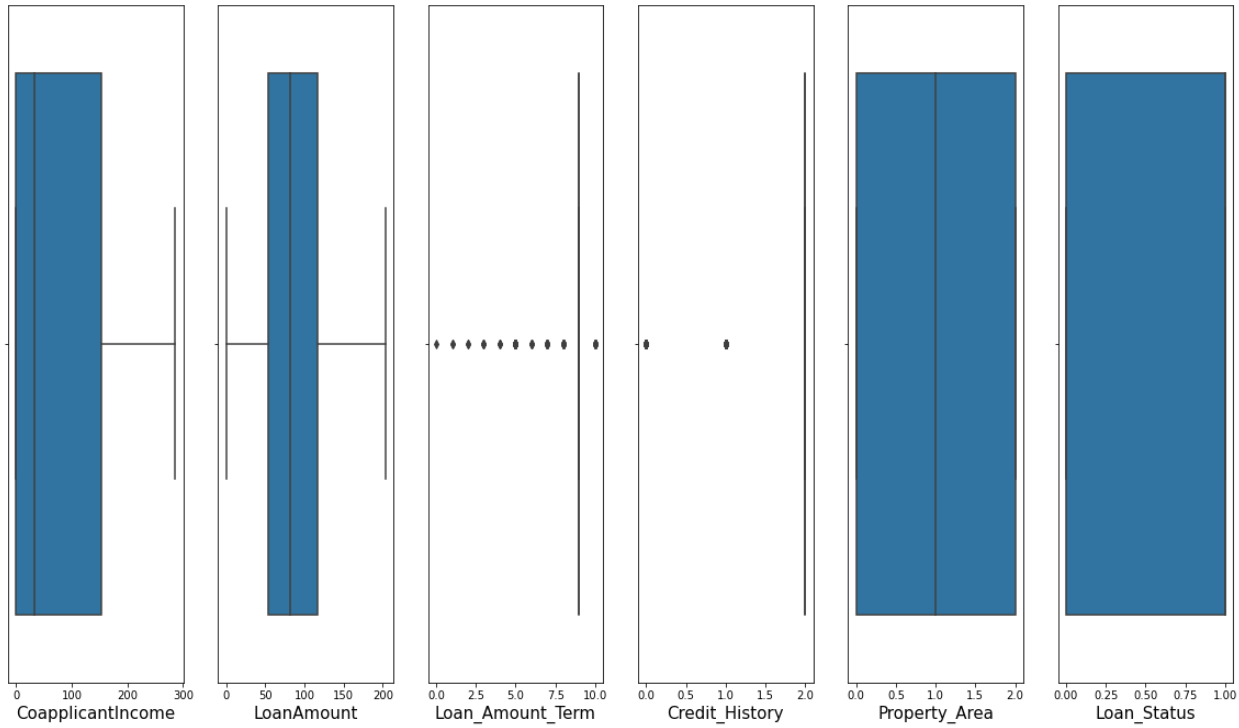
The following graph is a pictorial representation of the relation between target variables and feature variables;



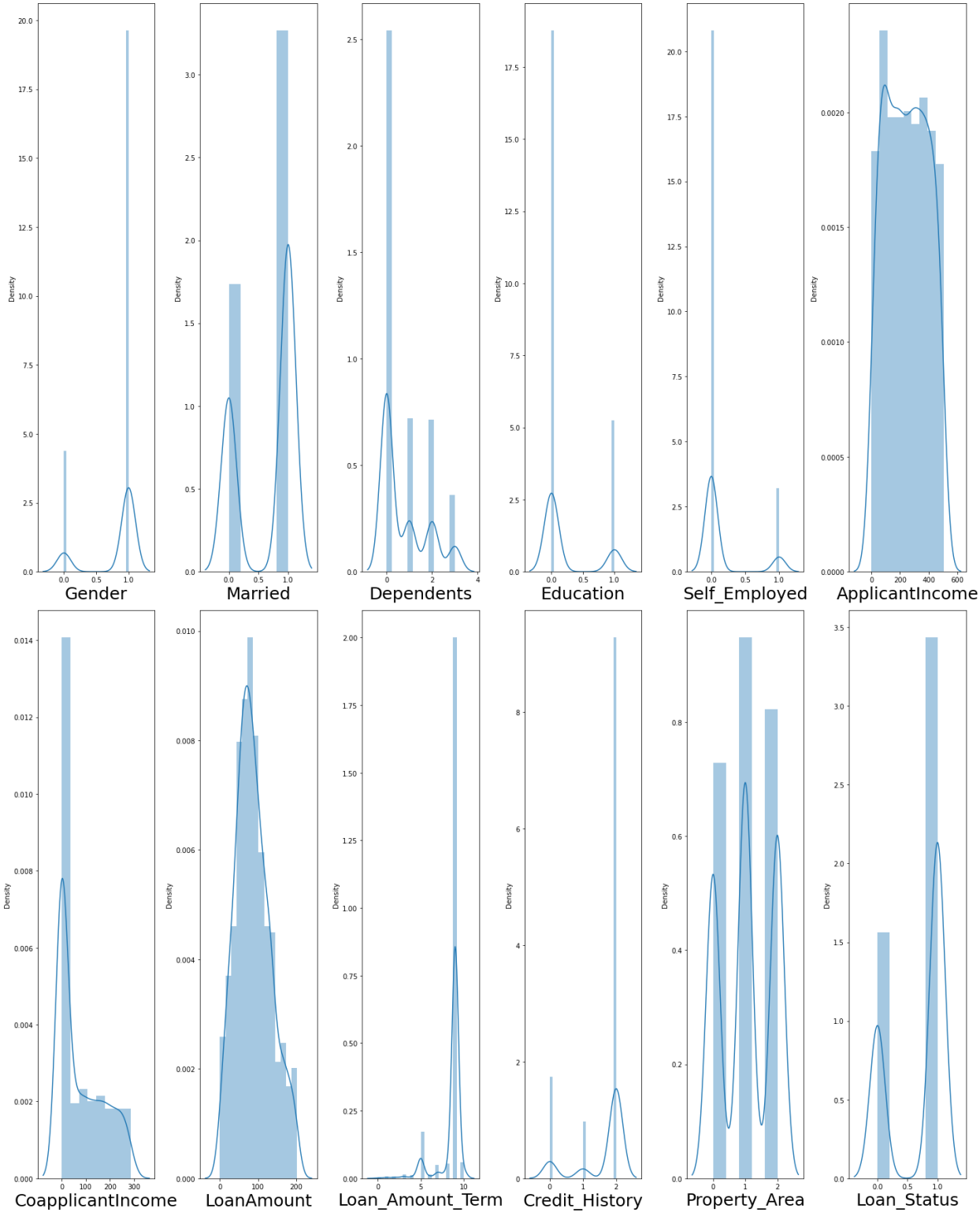
From the graph it's clear that Credit History is playing a major role in the Loan Status. And ApplicantIncome is least correlated with the target variable

After bivariate analysis I have checked the correlation and it was found out that Credit_History is having the highest correlation with target variable and ApplicantIncome is having the least correlation with LoanStatus. Multicollinearity exists between ApplicantIncome and LoanAmount.
In order to check for the outliers a statistical analysis is done using the describe method. Also, I have plotted the boxplot to get a better clarity on outliers.
From the boxplot it is clear that outliers exist in Loan Amount Term, Credit History etc.


0.00 0.25 0.50 0.75 1.00
Gender
0.00 0.25 0.50 0.75 1.00
Married
0 1 2 3
Dependents
0.00 0.25 0.50 0.75 1.00
Education
0.00 0.25 0.50 0.75 1.00
Self_Employed
0 200 400
ApplicantIncome


0 100 200 300
CoapplicantIncome
0 50 100 150 200
LoanAmount
0.0 2.5 5.0 7.5 10.0
Loan_Amount_Term
0.0 0.5 1.0 1.5 2.0
Credit_History
0.0 0.5 1.0 1.5 2.0
Property_Area
0.00 0.25 0.50 0.75 1.00
Loan_Status

Then I checked the distribution and found out that except ApplicantIncome,Loan Amount and Property Area all other feature variables are skewed. Below figure explains that:



DATA PREPROCESSING

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format. Categorical columns with null values are replaced with mode and numerical columns are replaced with mean.
The feature columns 'Loan_Amount_Term','Self_Employed','Dependents','Gender' are dropped as they have negligible correlation with target variable.
Class imbalance was there in the target variable. To overcome that issue, I have applied the SMOTE method.
Skewness of the feature columns are removed by the Power Transform method.
To standardize the data I have used StandardScalar, the whole idea of using Standard Scalar is that it will transform the data such that its distribution will have a mean value of 0 and standard deviation of 1.
Then before doing the train_test_split I made an attempt to find the best random state and got 9 as the best random state. Then splitted the dataset into train and test wherein 75% of data comes under training and 25% was reserved for testing purpose.

BUILDING MACHINE LEARNING MODELS

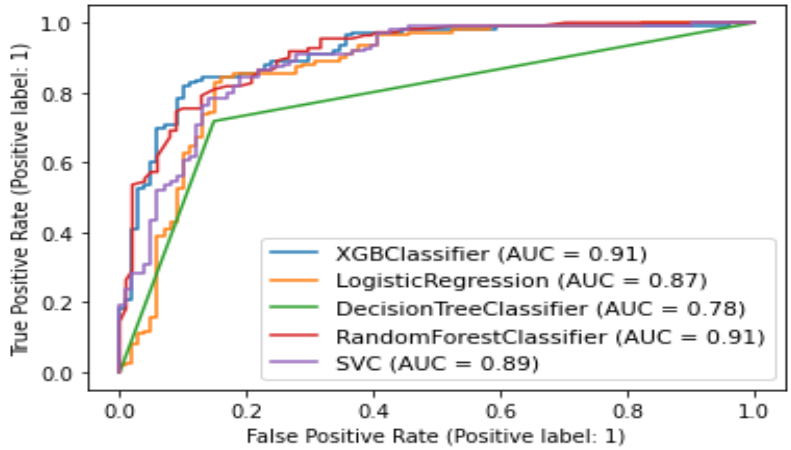
As the problem is a classification problem I have used the following models:

- Logistic Regression
- Decision tree Classifier
- Random forest classifier
- SVC
- XGBoost Classifier

Accuracy of Logistic regression model Decision tree, Random forest,SVC and XGBoost were 83.89%,78.20%,82.46%,82% and 81.52% respectively. The high accuracy can be due to overfitting so I have found out the cross validation score also.
Cross-validation scores of Logistic Regression, Decision Tree, Random Forest, SVC and XGBoost were 76.31%,74.17%,79.39%,76.66% and 81.52% respectively.
The below table explains the accuracy of each model:

MODEL	ACCURACY	CROSS-VALIDATION SCORE	DIFFERENCE
Logistic Regression	83.89	76.31	7.58
Decision tree	78.20	74.17	4.03
Random forest	82.46	79.39	3.-07
SVC	82	76.66	5.34
XGBoost	81.52	81.52	0.00

Since the difference is least for XGBoost Classifier that is taken as the best model and applied Hyper parameter tuning using GridsearchCV. The best parameter values are as follows:
Learning rate: 0.05
Max_depth : 12
Min_child_weight : 1
n_estimators : 200
After hyperparameter tuning the score increases to 85.31% .
ROC-AUC Curve is plotted;



CONCLUSION

For a banking firm loan approval or denial is very crucial.The dataset gives different factors which are potential enough to affect the loan status in the banking sector.. After careful analysis, it was found out that Credit History is the major factor.
Data was trained on different models and XGBoost Classifier turns out to be the best model with an accuracy score of 85.31%