***Loan Application Status Prediction Blog & Article Details.***

1. ***Problem Definition:***

Dream Housing Finance enterprise offers in all home loans. They have got a presence across all city, semi-urban and rural areas.

The client first applies for a home loan and after that, the enterprise validates the client's eligibility for the loan.

The enterprise desires to automate the loan eligibility system (real-time) based on client detail supplied even as filling out online application forms.

Those info are Gender, Marital status, education, number of Dependents, earnings, loan amount, credit score records, and others.

To automate this procedure, they have got provided a dataset to identify the client segments which might be eligible for loan amounts so that they can specially target those clients.

1. ***Data Analysis:***

Loan\_ID: Unique Loan ID

Gender: Male/ Female

Married: Applicant married (Y/N)

Dependents: Number of dependents

Education: Applicant Education (Graduate/ Under

Graduate)

Self\_Employed: Self-employed (Y/N)

ApplicantIncome: Applicant income

CoapplicantIncome: Coapplicant income

LoanAmount: Loan amount in thousands

Loan\_Amount\_Term: Term of loan in months

Credit\_History: credit history meets guidelines

Property\_Area: Urban/ Semi Urban/ Rural

Loan\_Status: Loan approved (Y/N)

As referred to above that is a Binary classification problem in which we want to predict our target label that is “loan status”.

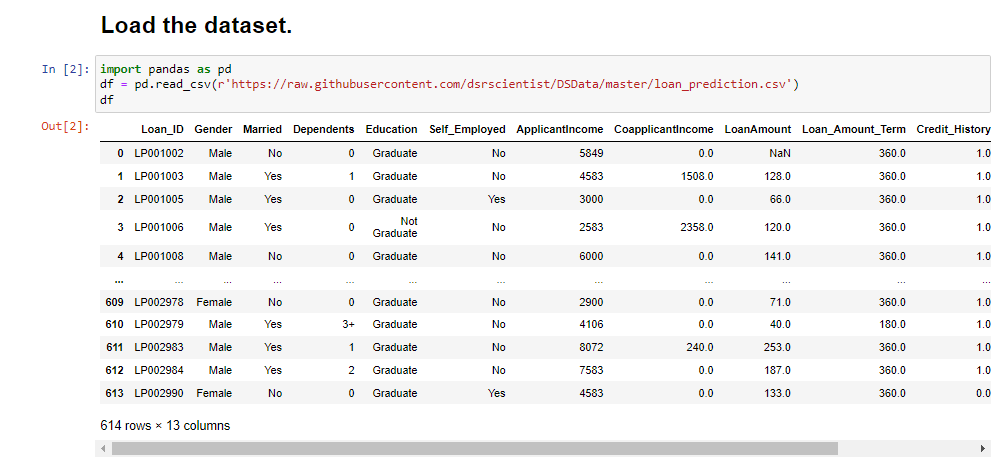
Yes: if the loan is approved

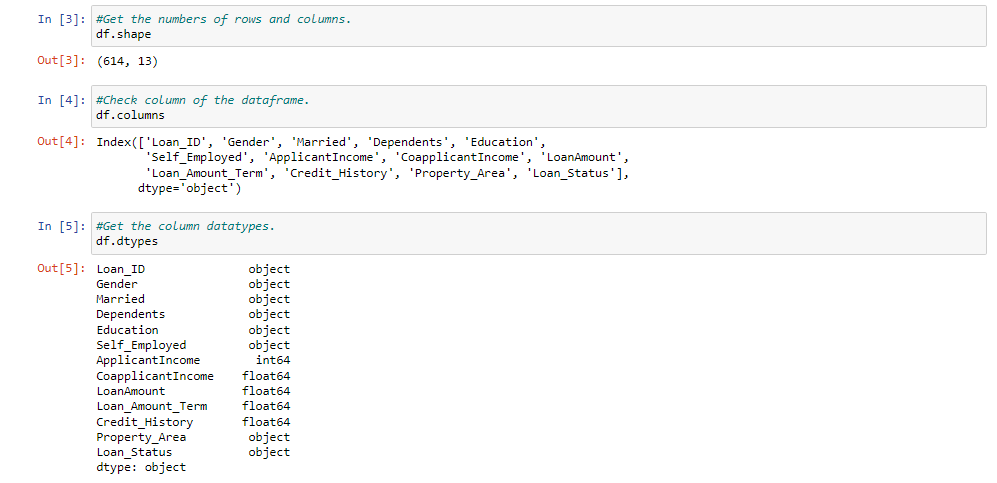
NO: if the loan is not approved

* ***Now import the packages /libraries to make it easier to write the program.***

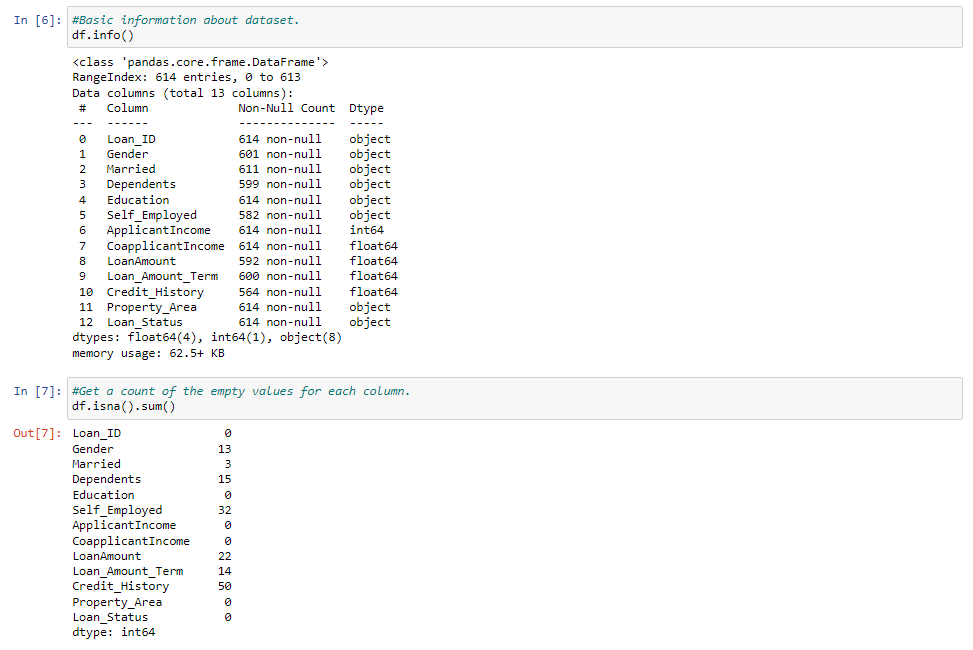


1. ***EDA Concluding Remark:***

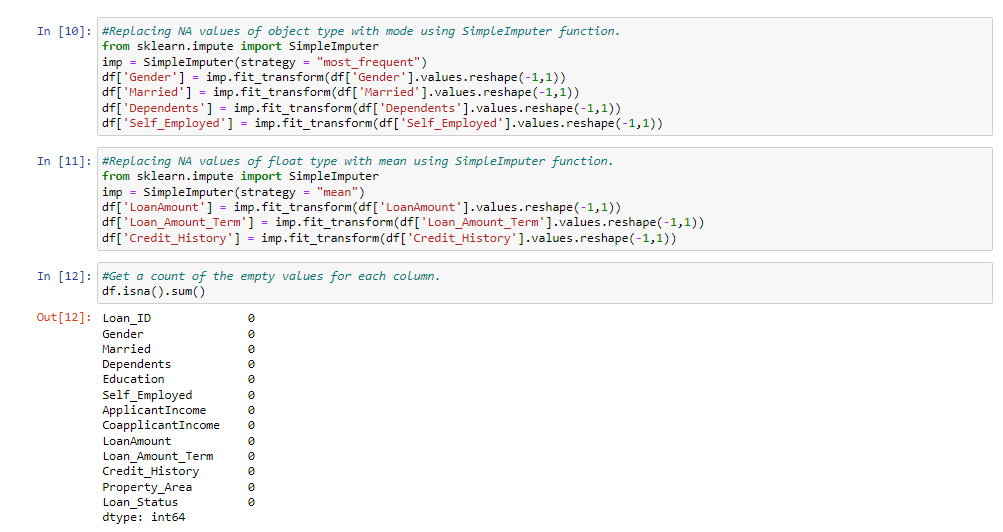


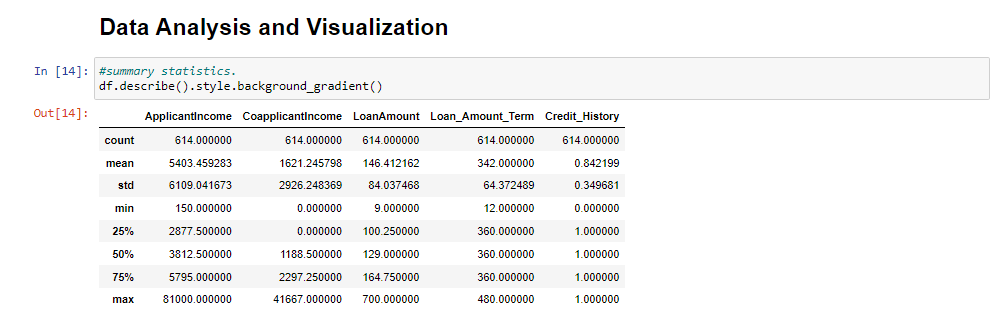


We have 13 features in total out of which we have 12 independent variables and 1 dependent variable i.e. Loan\_Status in train dataset and 12 independent variables in test dataset. The Loan\_ID, Gender, Married, Dependents, Education, Self\_Employed, Property\_Area, Loan\_Status are all categorical.



We also see some missing values, let’s take stock of missing columns and what are the possible replaced values for categorical and numerical columns.

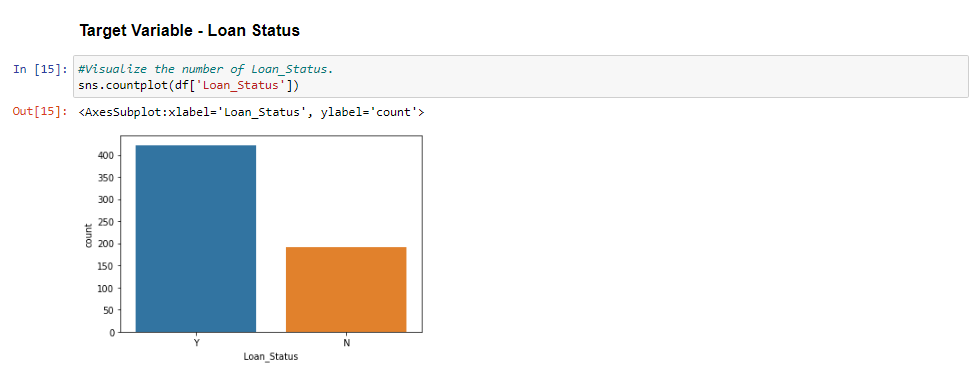




Observation from this statistical graph the mean of ApplicantIncome, CoapplicantIncome, LoanAmount columns are greater than median, hence it is right skewed.

Standard deviation is high in ApplicantIncome, CoapplicantIncome.

It means data spread is high. high gap between 75th percentile and max are present in ApplicantIncome and CoapplicantIncome so, few outliers are present.



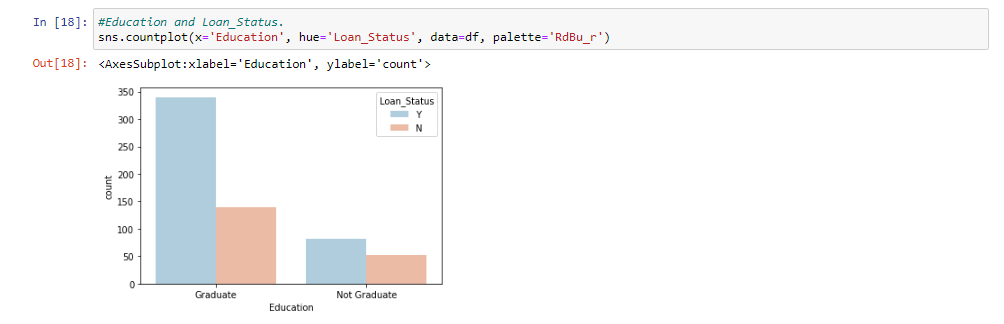
From this Univariate Analysis Observation More loans are approved then rejection.

Univariate Analysis:-

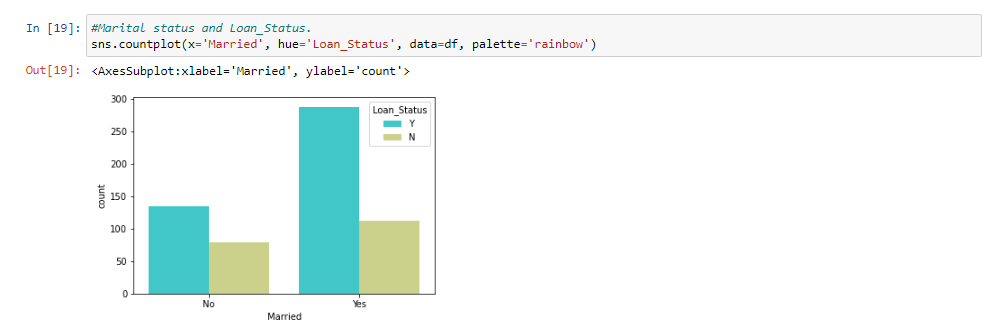


1. More Loans are approved Vs Rejected
2. Count of Male applicants is more than Female
3. Count of Married applicant is more than Non-married
4. Count of graduate is more than non-Graduate
5. Count of self-employed is less than that of Non-Self-employed
6. Maximum properties are located in Semiurban areas
7. Credit History is present for many applicants
8. The count of applicants with several dependents=0 is maximum

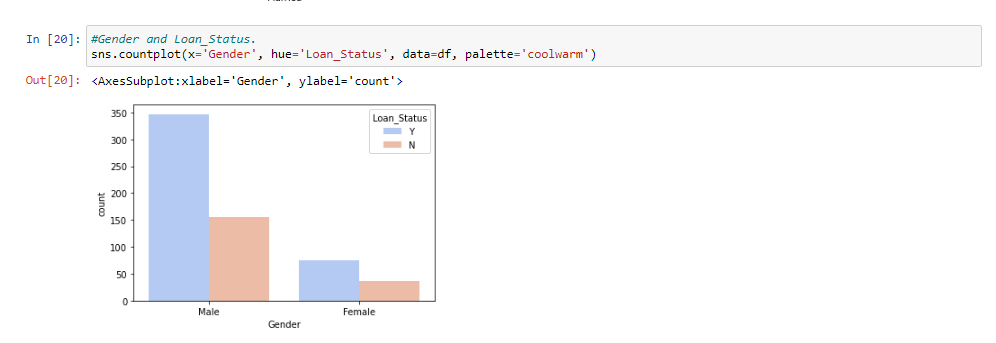
Bivariate Analysis:-



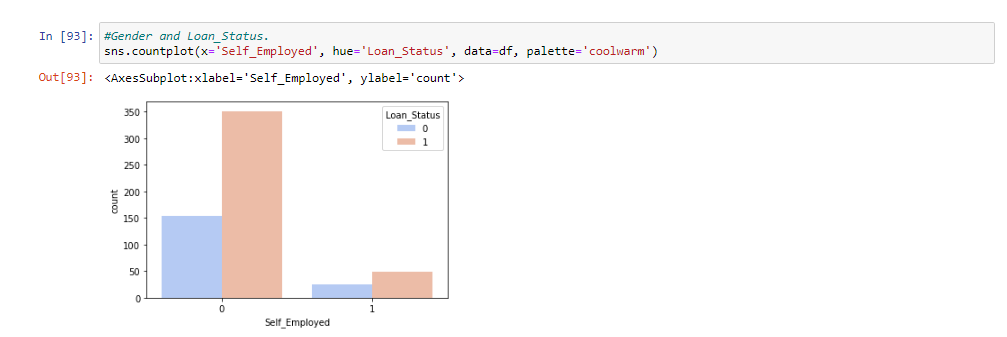
Graduates have higher chance of loan approval compared to non-graduates.



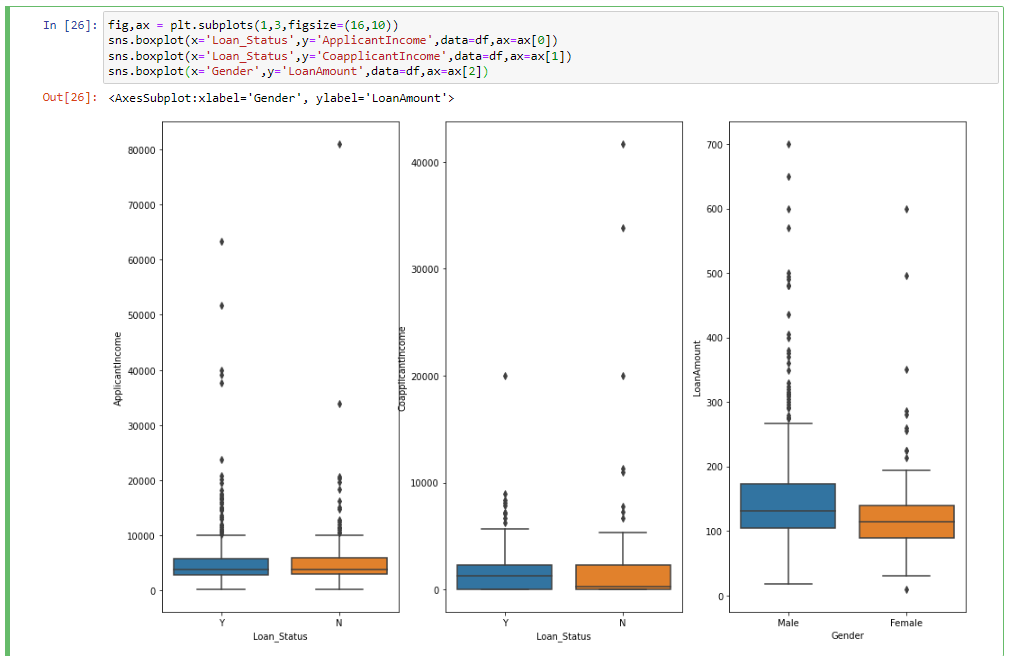
Married applicants have a slightly higher chances of loan approval.



here is a substantial difference between male and female approval rates.



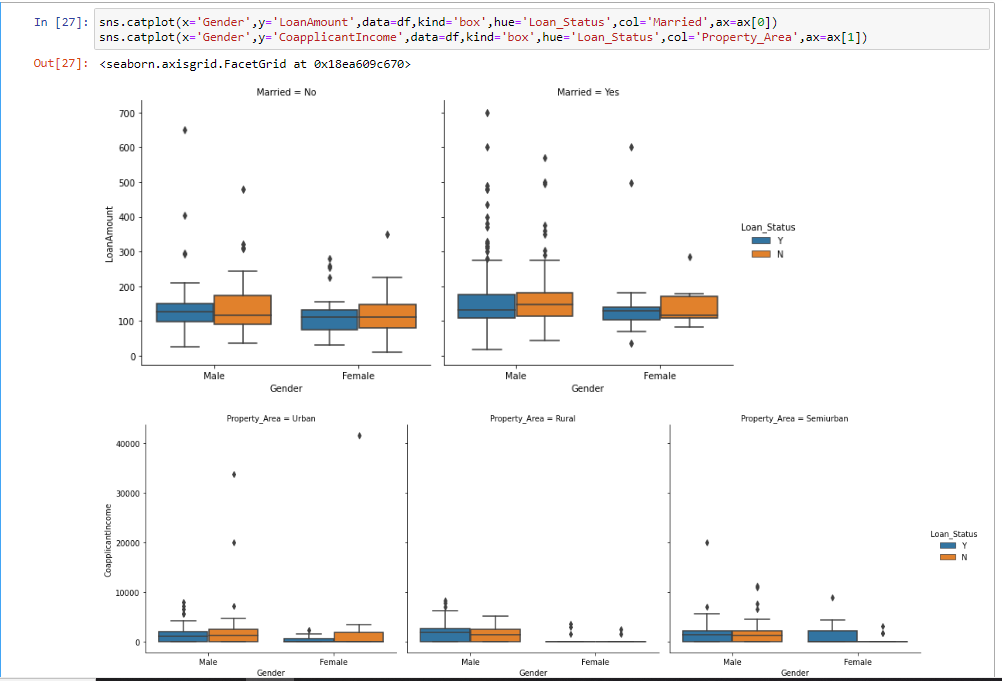
Self\_Employed employees have slightly lower chances of loan approval but the situation is not that bad.



Mean ApplicantIncome of 0 and 1 are almost the same (o: no,1: Yes)

Mean Co- ApplicantIncome of 1 is slightly more than 0 (o: no,1 Yes)

The mean value of Loan Amount applied by males (0) is slightly higher than Females(1).



If you are married then the loan amount requested is slightly higher than non-married.

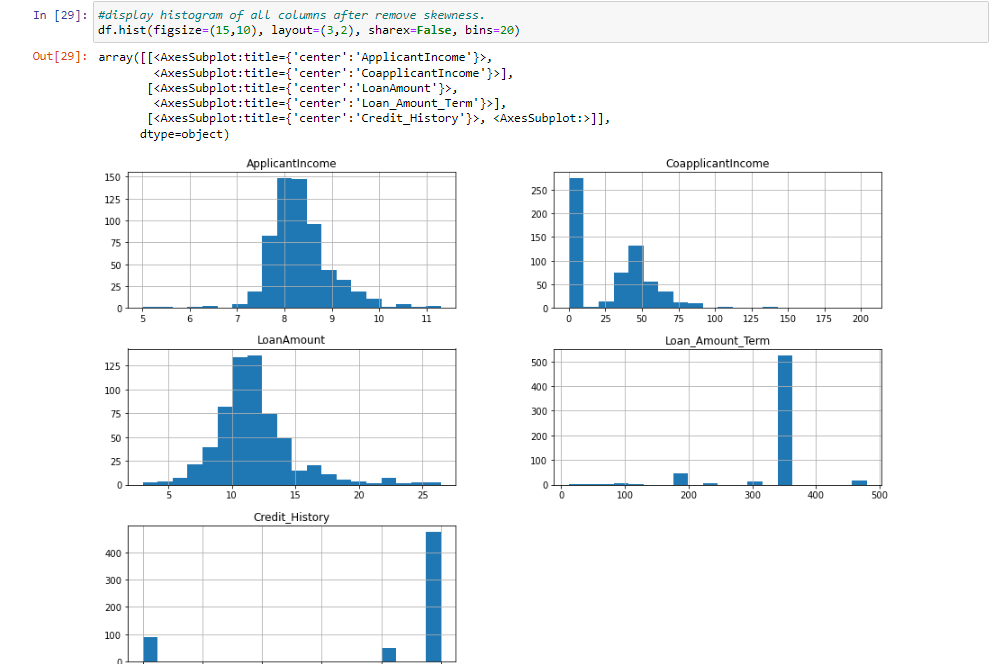
Male have higher Co-applicant income than females in all three property areas.



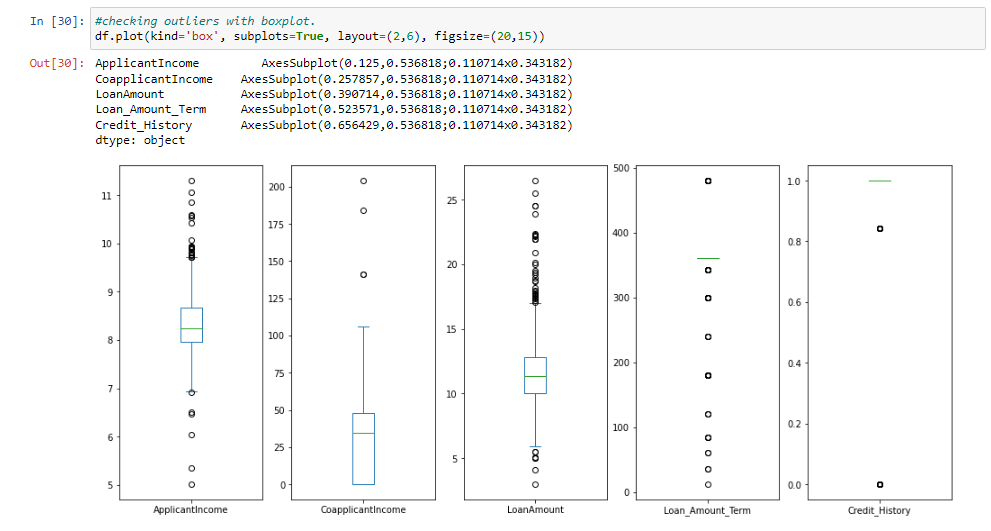
There are some numeric columns have skewness in the dataset i.e., ApplicationIncome, CoapplicationIncome, LoanAmount has right skewness.

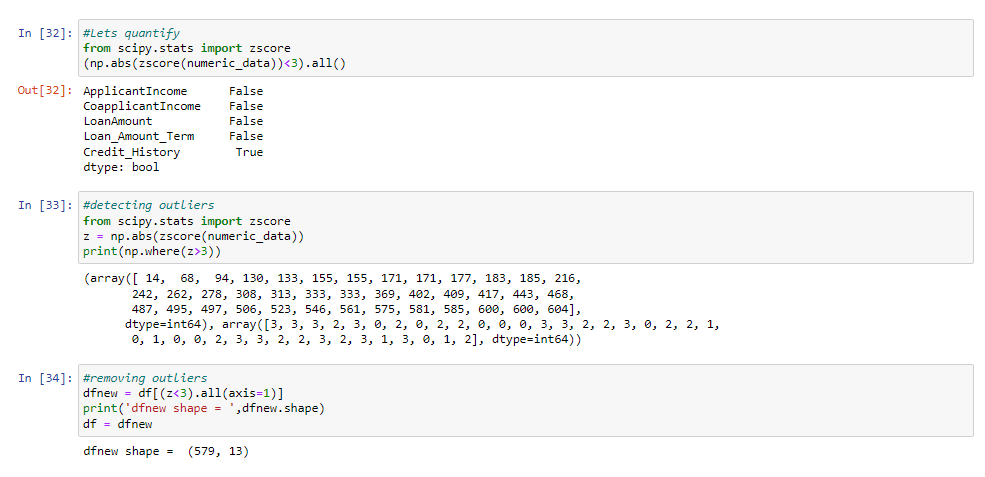
So, lets remove the skewness and outliers of this columns.





So, skewness is removed by log and sqrt methods. Now check the outliers present or not in numeric columns.



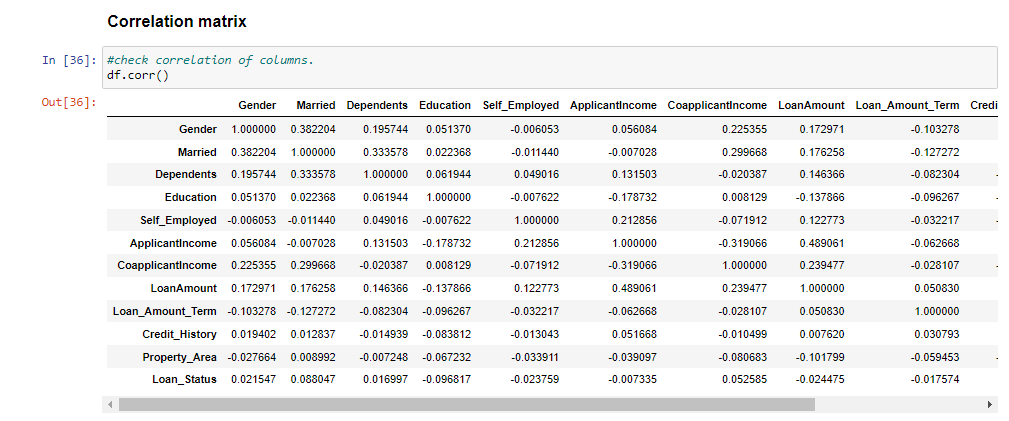


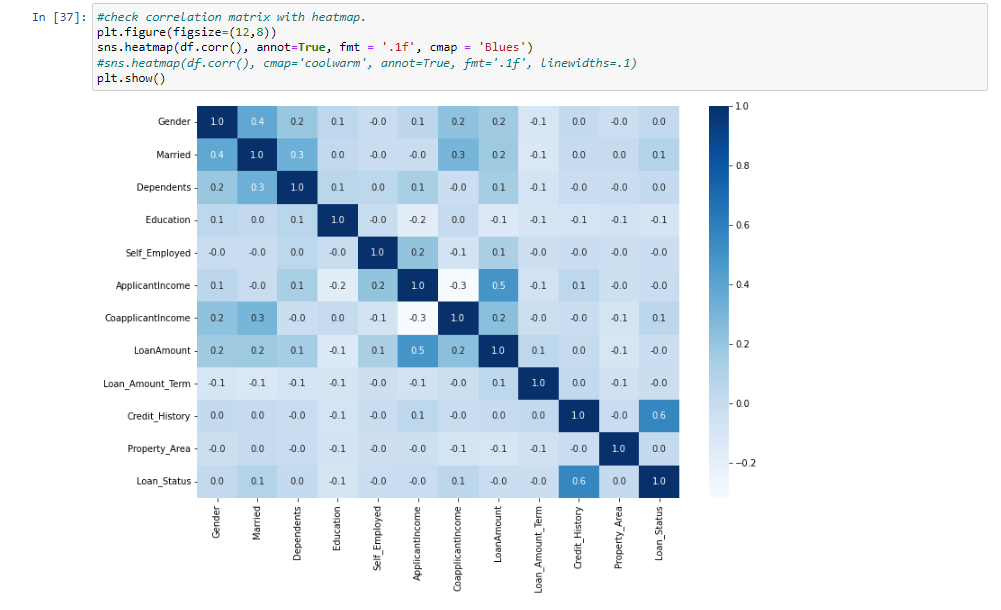
From above boxplot graph and zscore to check outliers present or not.

The threshold value is 3 if zscore values are greater than threshold value so they column have outliers.

ApplicantIncome, CoapplicantIncome, LoanAmount and Loan\_Amount\_Term columns have some outliers.

So remove the outliers and clean the dataset.

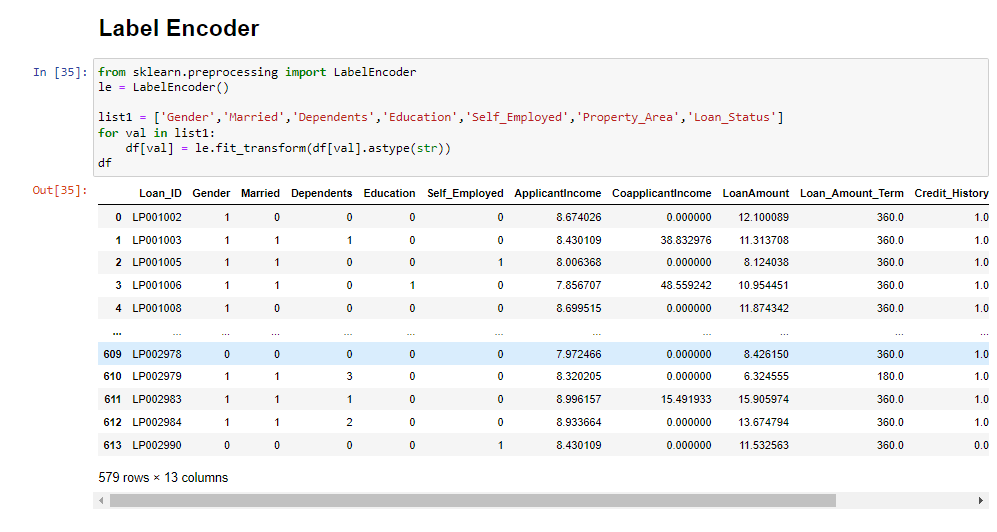




We see that the most correlated variables are (ApplicantIncome - LoanAmount) and (Credit\_History - Loan\_Status). LoanAmount is also correlated with CoapplicantIncome

Above, correlation matrix shaws that Credit\_History is highly correlate with Loan\_Status (target variable) means when credit history score is good then high chances to approval for loan.

1. ***Pre-processing Pipeline:***



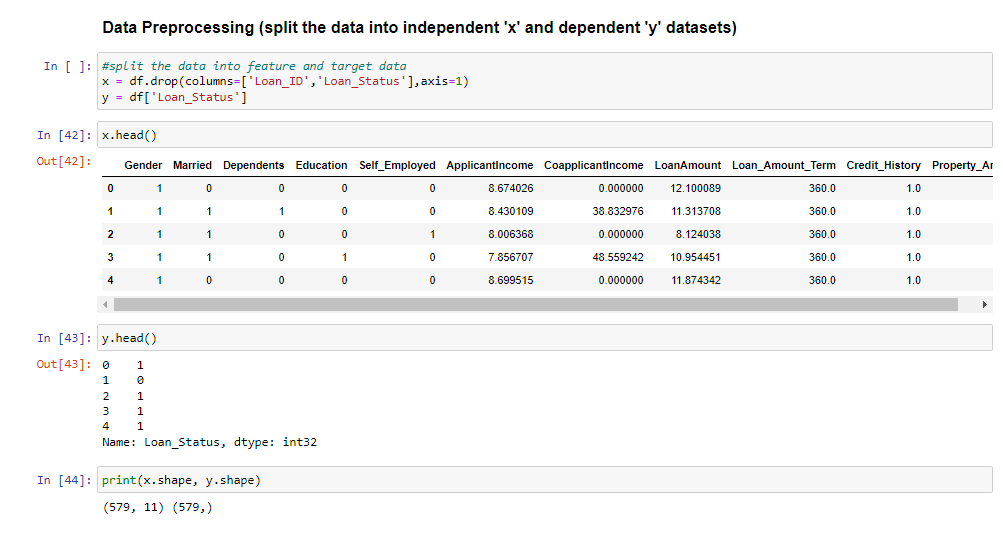
Now, we are apply label encoder for categorical features in dataset.

There are more than 5 columns are categorical in this dataset so, we are converting them using label encoder.

If binary column has yen or no, male or female then convert into 1 or 0 and they have more than two value then convert into 0,1,2,3 etc.

Now if we run the df.head() command again, we find that the values have been transformmed successfully.

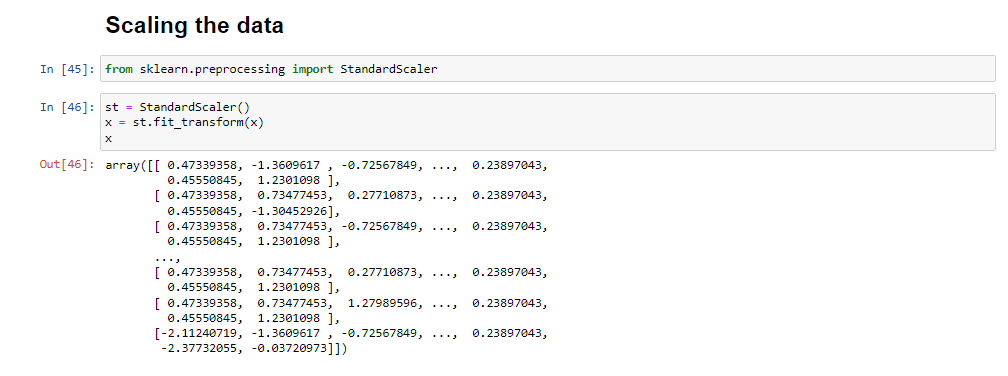
We also see, that there are few columns, which are not of much Importance in this process. Let us get rid of them.

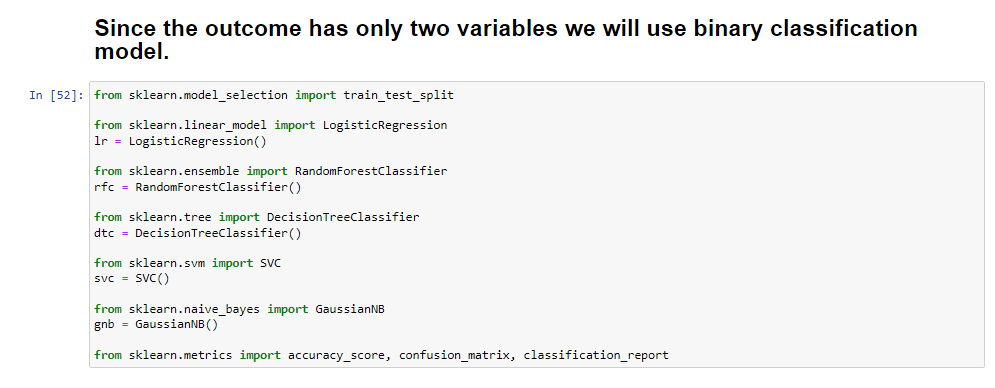


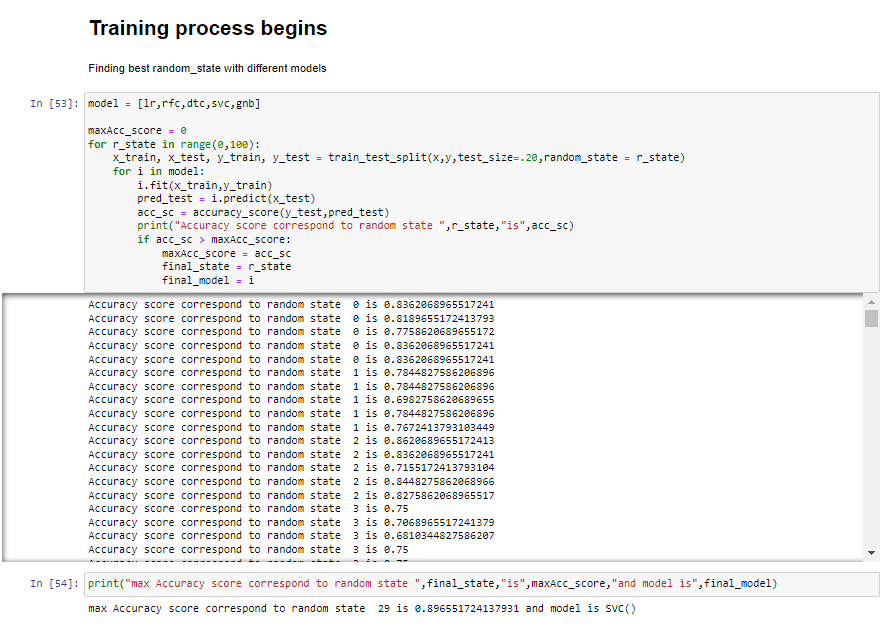
Here, x is the feature variable, containing all the features like Gender, Married, Dependents, Education, Self\_Employed etc. excluding the Loan\_Status column.

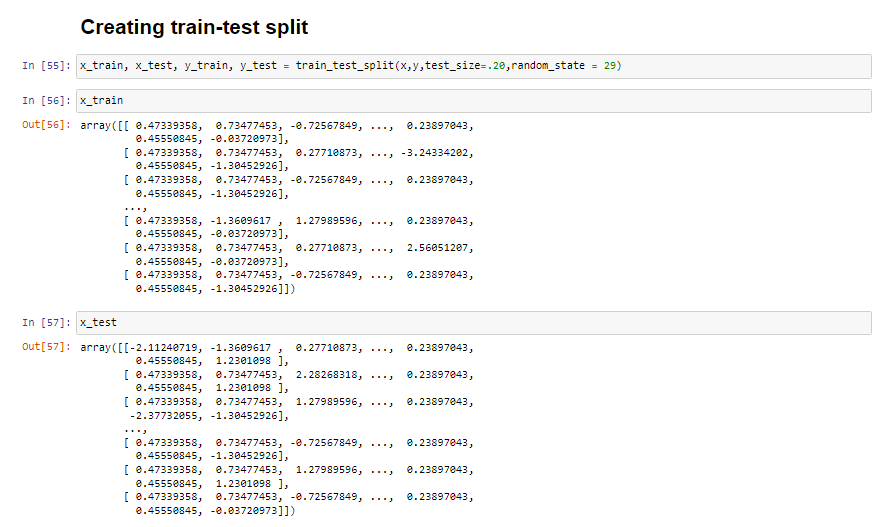
y, on the other hand, is the target variable, as that is the result that we want to determine, i.e, whether a customer has loan approve or not.

Now, we will be scaling the data and then find best random state for various models and then splitting the data into four variables, namely, x\_train, y\_train, x\_test, y\_test.











Let's understand the variables :-

x\_train: contains a set of values from variable ' x '

y\_train: contains the output (whether the customer loan status is approved or not) of the corresponding value of x\_train.

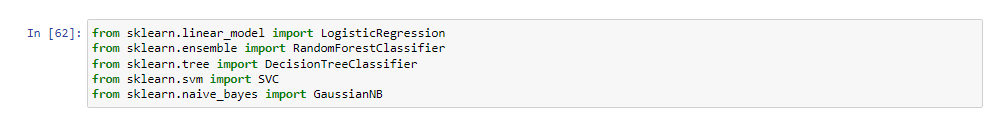
x\_test: contains a set of values from variable ' x ', excluding the ones from x\_train.

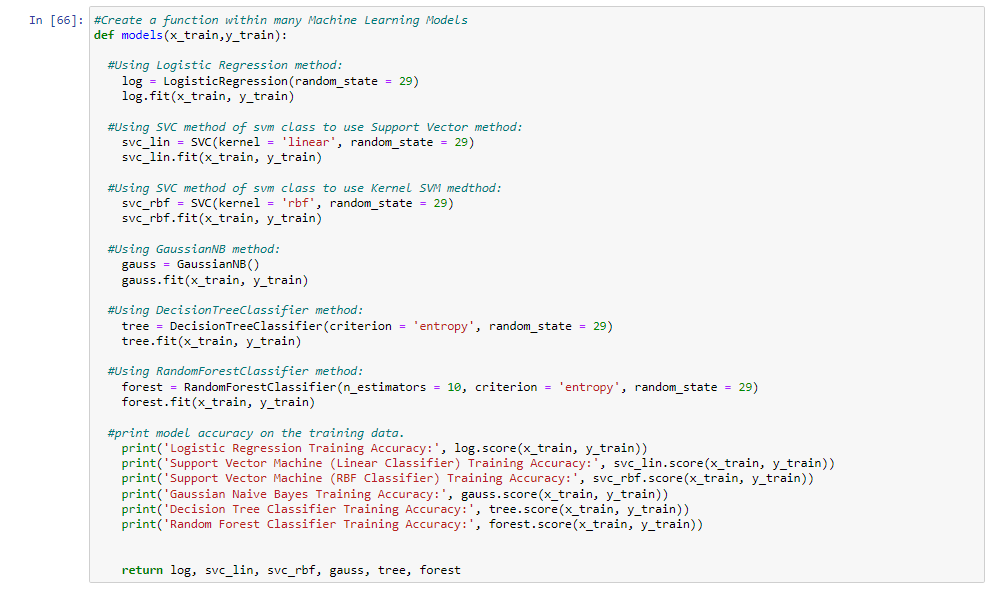
y\_test: contains the output (whether the customer loan status is approved or not) of the corresponding value of x\_test.

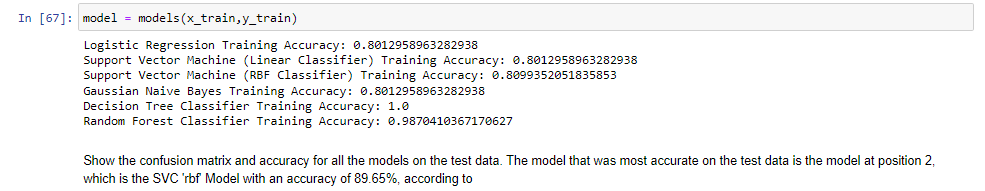
test size: represents the percentage ratio of x\_train:x\_test (Here 0.2 means that the data will be segregated in the x\_train and x\_test variables in a 80:20 ratio). You can use any value you want. A value <0.3 is preferred.

1. ***Building Machine Learning Models:***

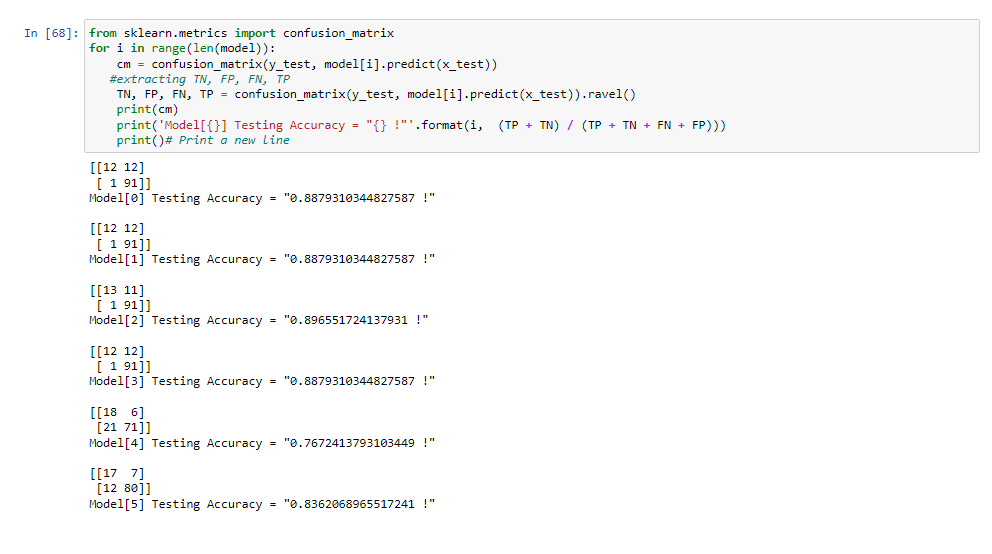
Create a function that has within it many different machine learning models that we can use to make our predictions.

******

******

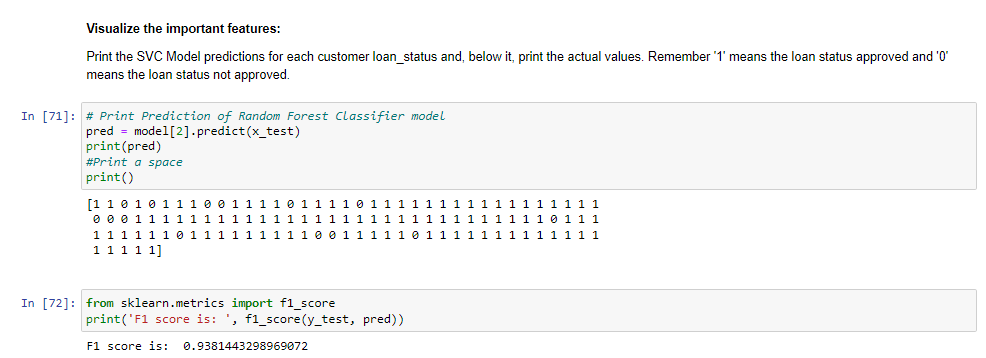
******

Show the confusion matrix and accuracy for all the models on the test data. The model that was most accurate on the test data is the model at position 2, which is the SVC 'rbf' Model with an accuracy of 89.65%, according to

******

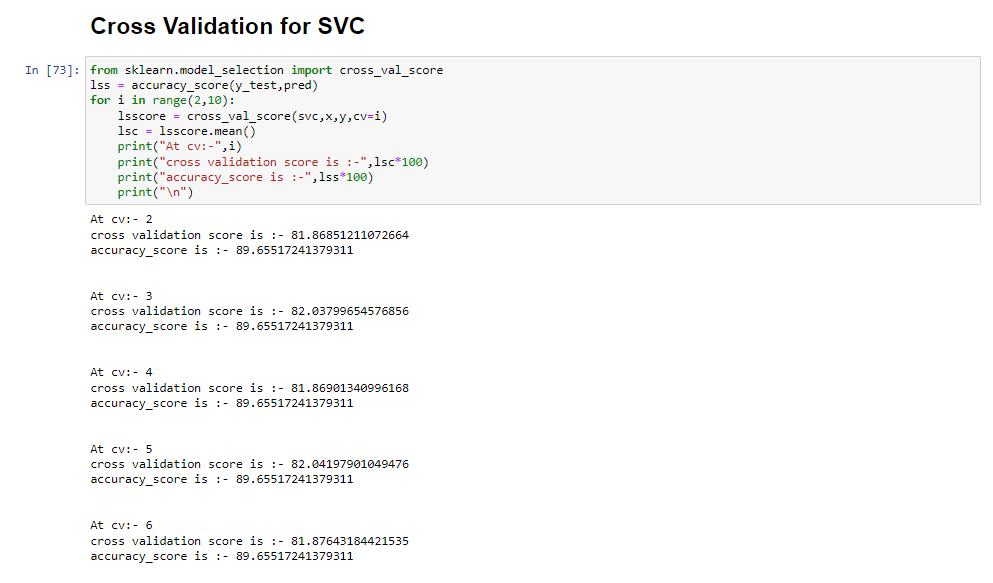
The model that I will use to predict loan status , will be the model at position 2, the SVC.

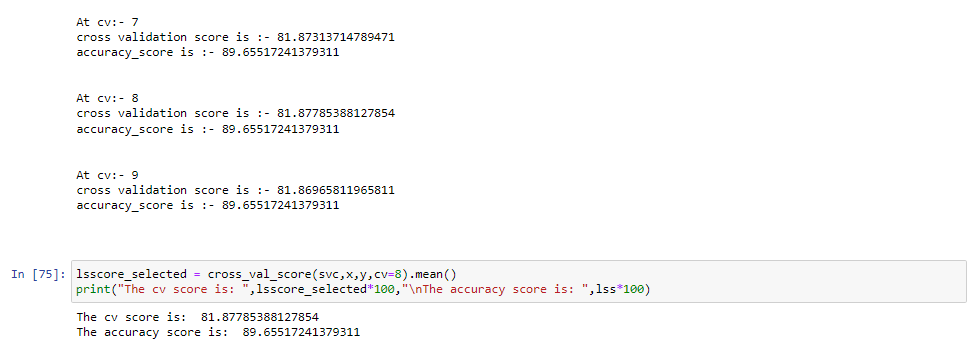
I choose that model because it did second-best on the training and testing data and has an accuracy of 89.65% on the testing data and 80.99% on the training data.



Print the SVC Model predictions for each customer loan status and, below it, print the actual values. Remember '1' means the loan status approved and '0' means the loan status not approved.

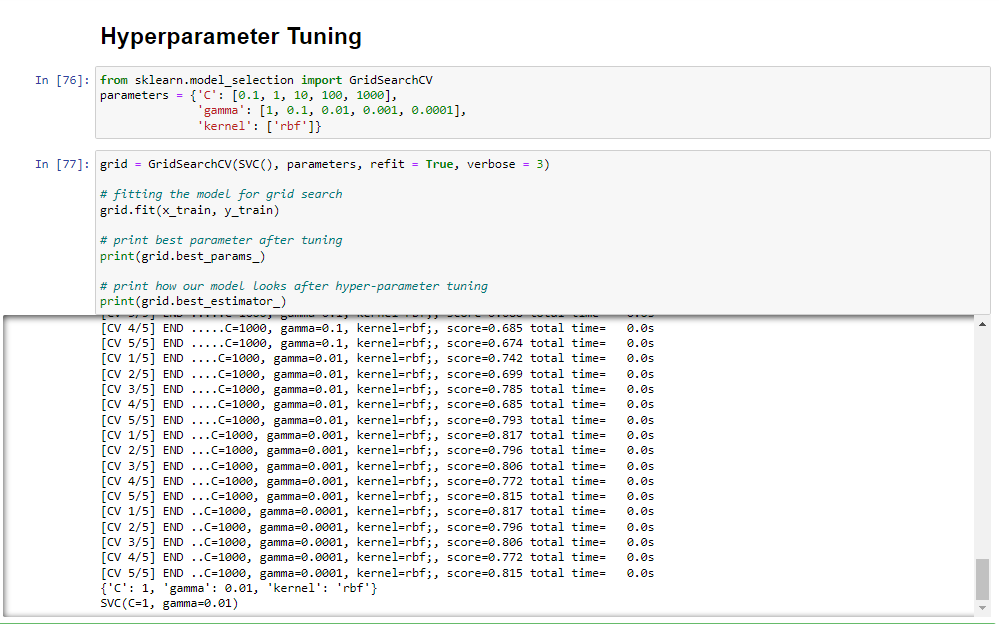
Now we see the cross validation for svc model.

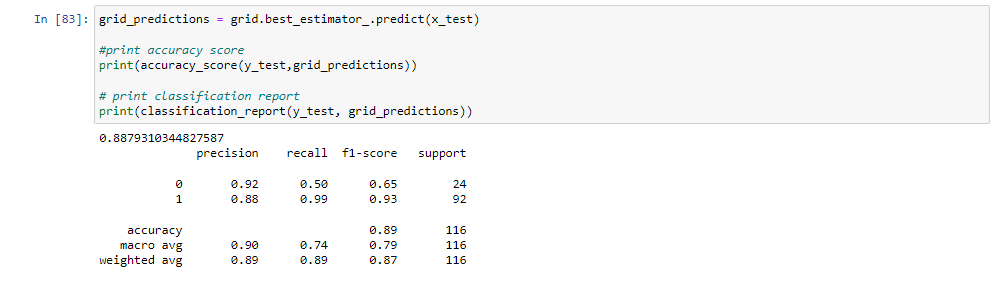




The best cv score is 81.77 at cv=8 and accuracy score of testing model is 89.65.

Now we see the hyperparameter tuning for the grid.best\_params\_.

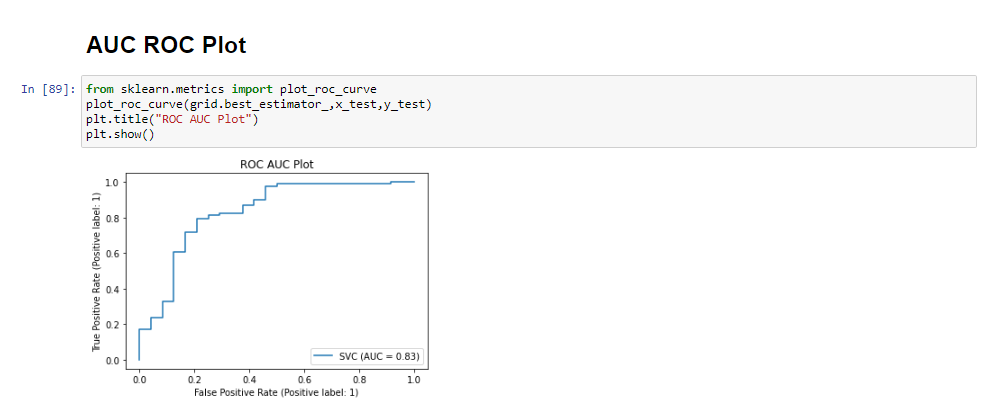




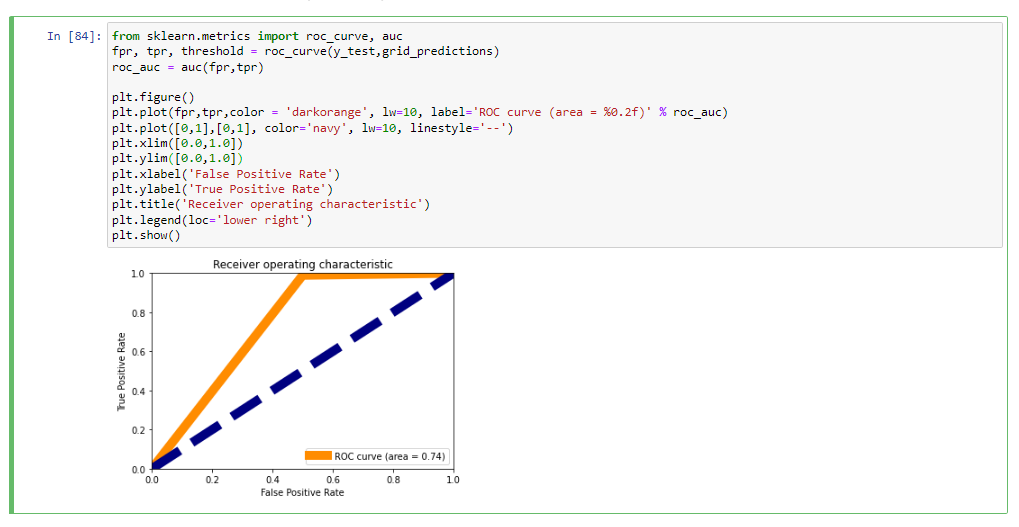
In hyperparameter tunning the grid.best\_estimators are SVC(C=1, gamma=0.01)

According to grid.best\_estimators accuracy score is 88.79 that is pretty much good.

Now we will see the AUC ROC plot and ROC curve according to,

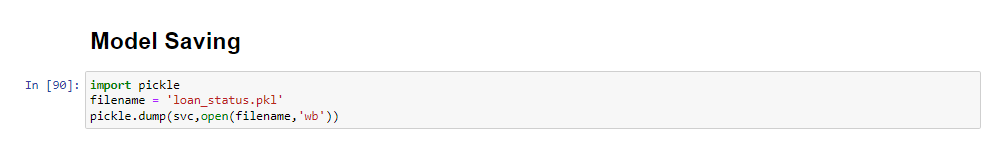


Here, the AUC score is 0.83 based on grid.best\_estimator\_.



Here, the ROC curve area is 0.74.

End the last we save the model as per over prediction of good model choice.



1. ***Concluding Remarks:***

We did Exploratory data Analysis on the features of this dataset and saw how each feature is distributed.

We did bivariate and univariate analysis to see impact of one another on their features using charts.

We analysed each variable to check if data is cleaned and normally distributed.

We cleaned the data and removed NA values

We also generated hypothesis to prove an association among the Independent variables and the Target variable. And based on the results, we assumed whether or not there is an association.

We calculated correaltion between independent variables and found that credit history score and loan status have significant relation.

We created model list for choose best random state for constructing the model.

Finally, we got a SVC model with highest accuracy.

We tested the data and got the accuracy of 88 %