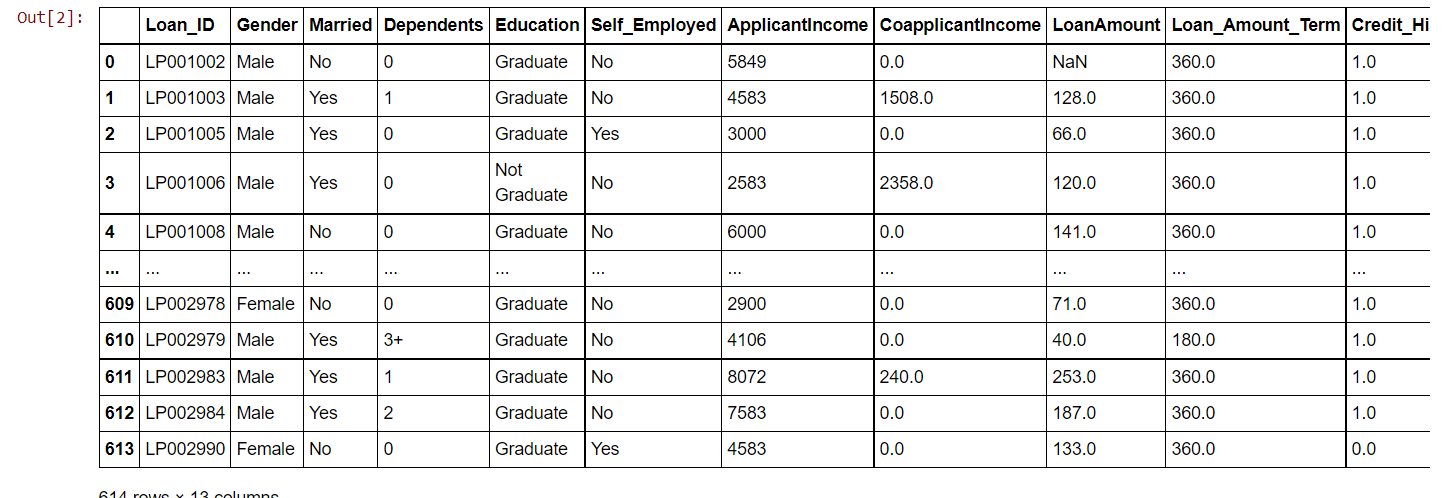
**LOAN APPLICATION STATUS PREDICTION**

1. **Problem Defination :-**

This dataset includes details of applicants who have applied for loan. The dataset includes details like credit history, loan amount, their income, dependents etc.

****

The above image shows us the dataset of the loan application status. The dataset contains 613 rows and 13 columns.

**Independent Variables:-**

- Loan ID – Identity of the loan application

- Gender- gender of the applicant

- Married – martial details of the applicant

- Dependents – dependent members on the applicant

- Education – qualtification and education level of the applicant

- Self Employed – knowing whether they are self employed

- Applicant Income – income details

- Co-applicant Income

- Loan Amount – amount of money fo which the loan is applied

- Loan Amount Term

- Credit History – credit history details of the applicant

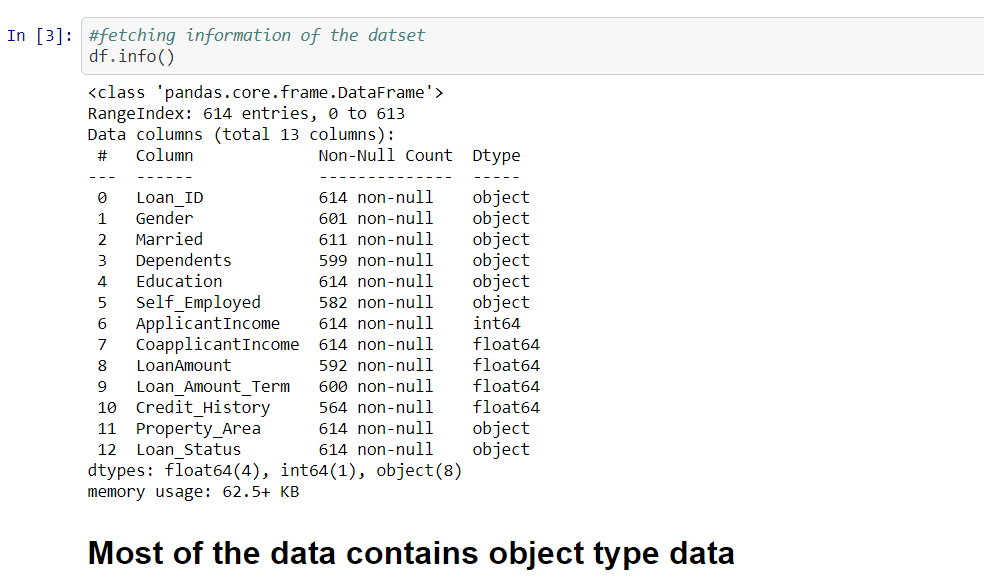
- Property Area

**Dependent Variable (Target Variable):**

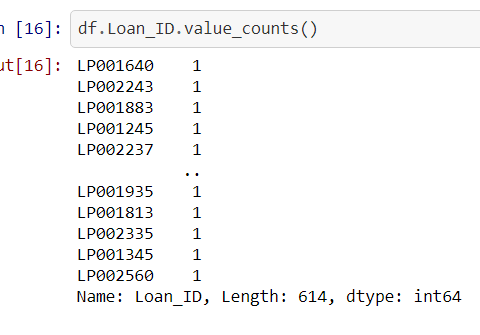
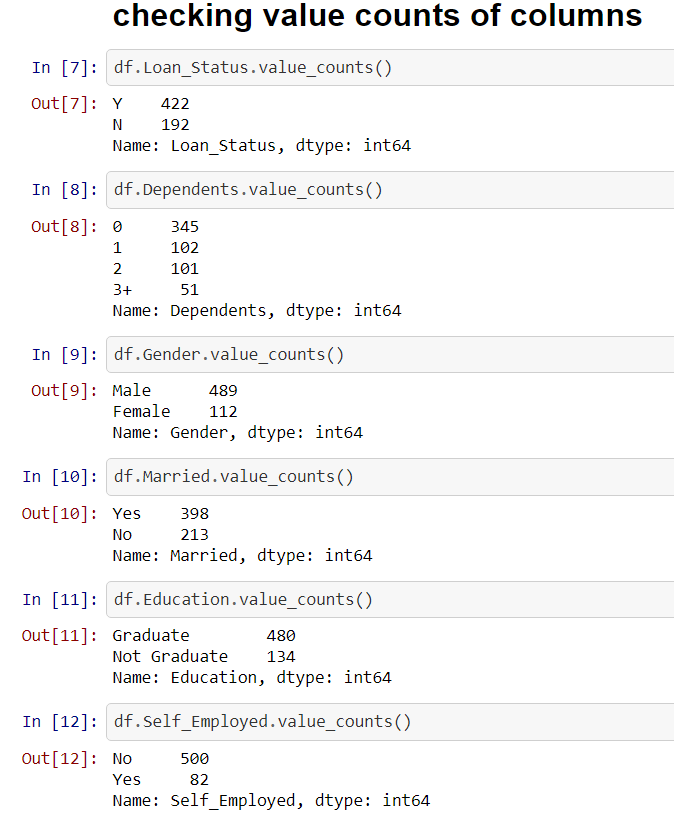
- Loan\_Status – if the loan is approved or not

Building a model that can predict whether the loan of the applicant will be approved or not on the basis of the details provided in the dataset.

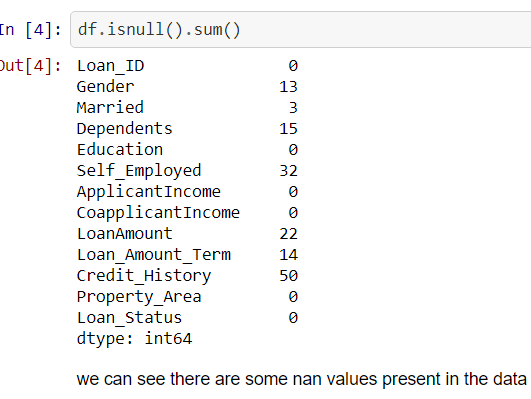
**Data Analysis :-**

****

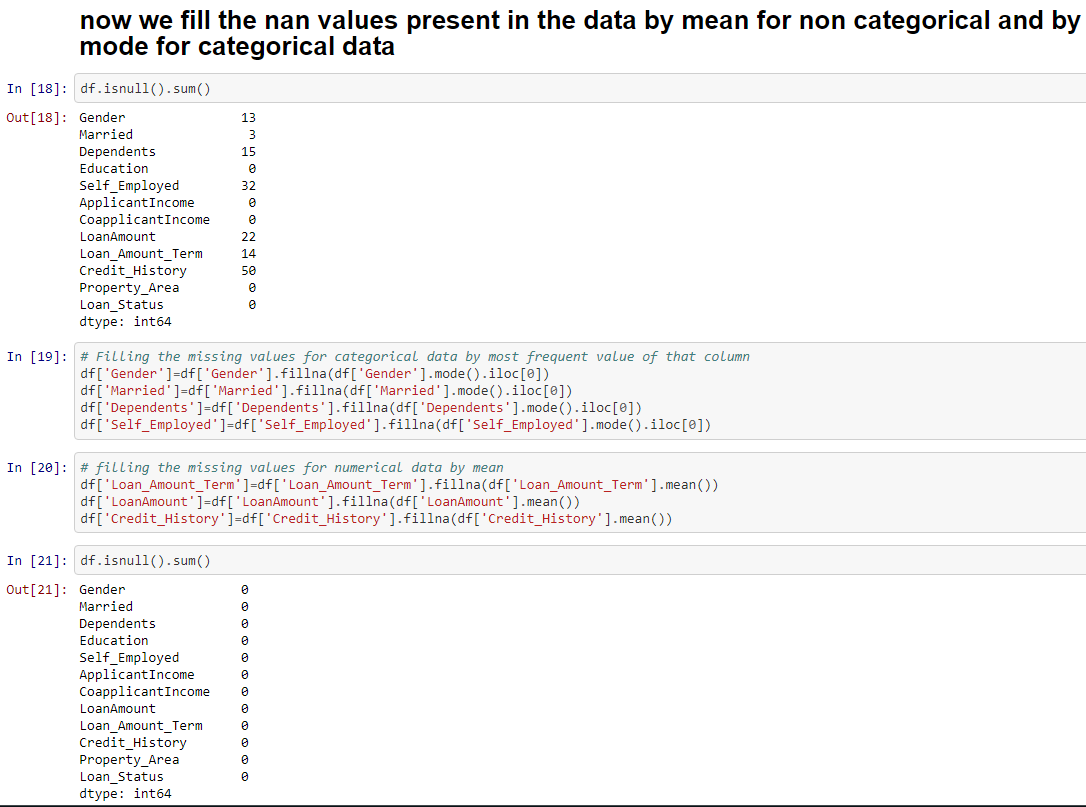
What we can see that most of the data is in object type which will need to process into numerical while building prediction model.



We checked value counts and what we found out that calue counts for ‘loan ID’ column was unique for every row so it won’t help us in predicting data. So we can drop the column.



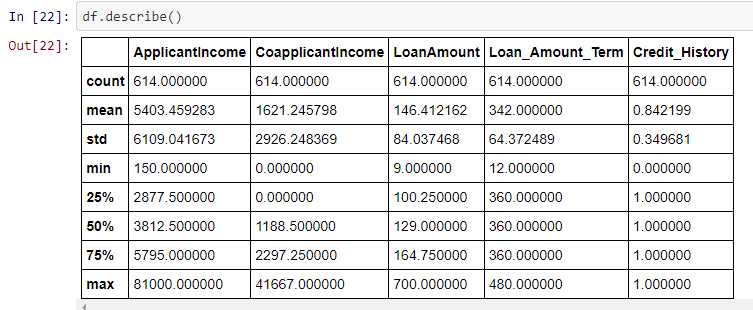
There are some Nan values present in the data which we need to fill and clean the data.



We filled the nan values with mean for non categorical data and we for categorical we used mode of that specific columns.

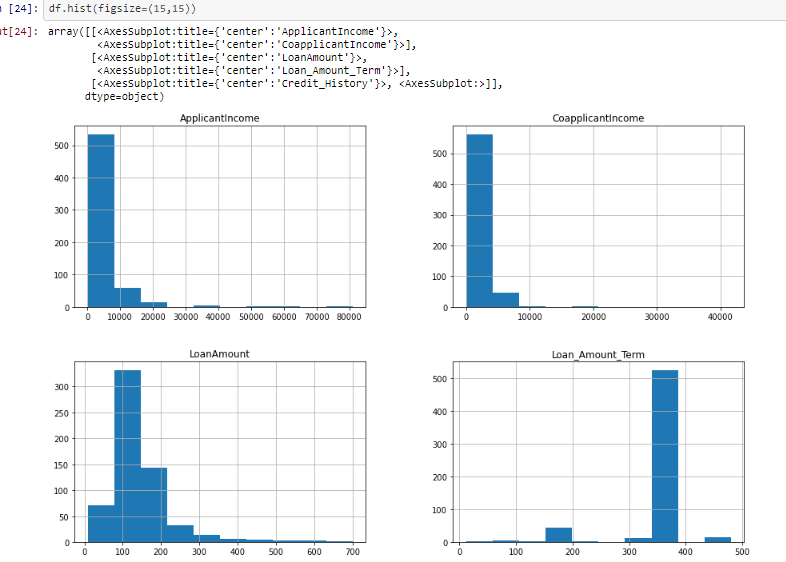
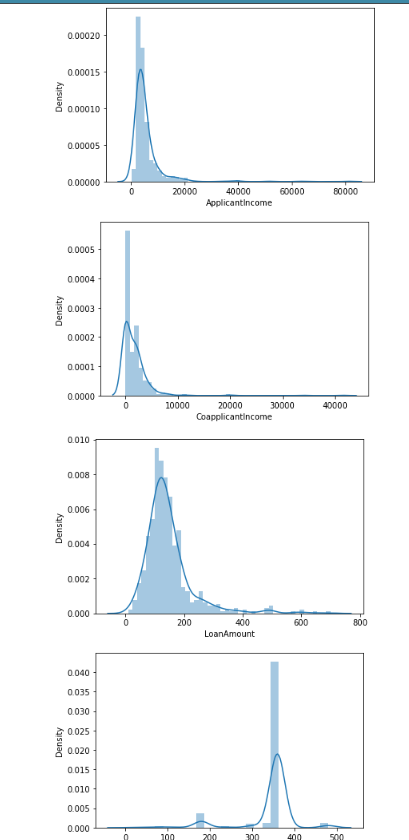


Also dropping the ‘loan id’ column because we don’t need the column for prediction of data.



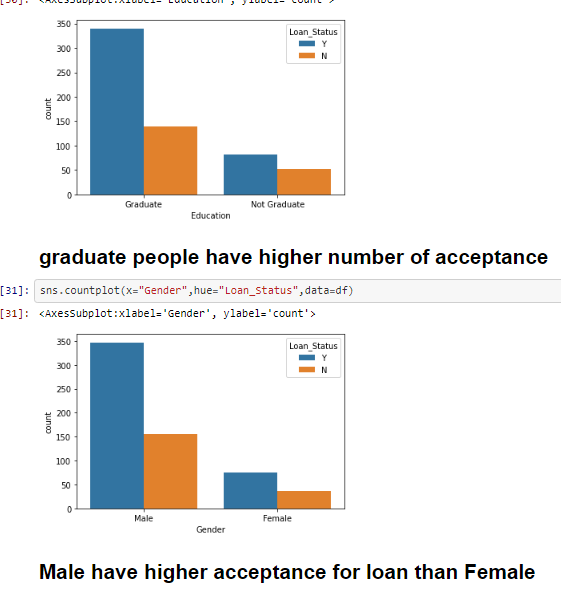
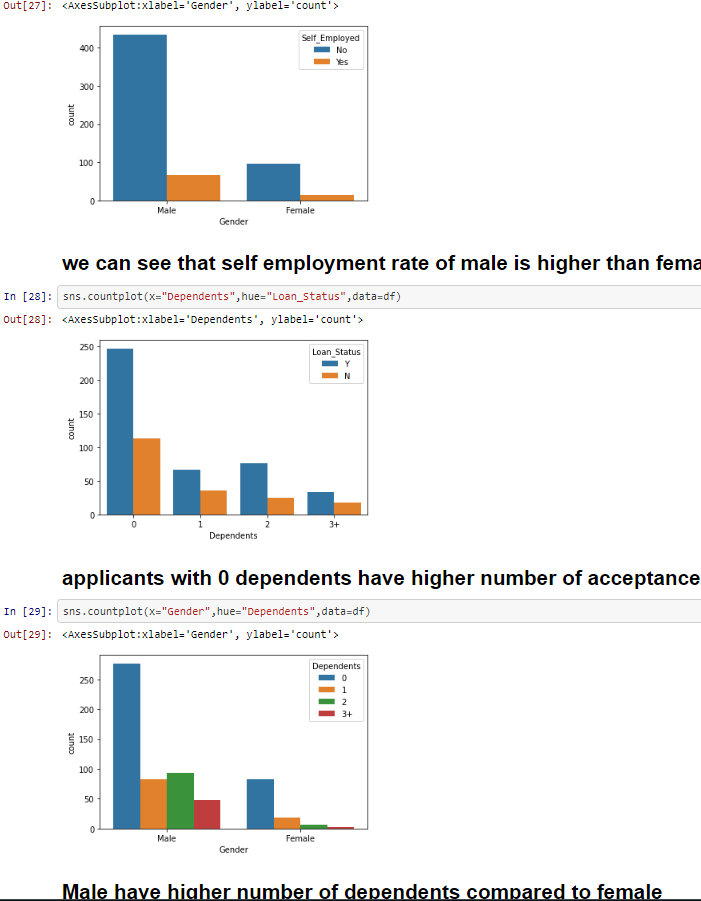
What we can analyze is that the average of loan amount term is 342 and credit history is 0.82 also everything looks fine in the dataset.

**EDA Concluding Remarks :-**

****

We can definitely see some skewness present in the above data which will need to process before prediction of data.

Now will check counts of the data



what we can conclude from the above graphs is that :-

Male have higher acceptance for loan as well as more qualification compared to female.

Female have less dependents compared to men.

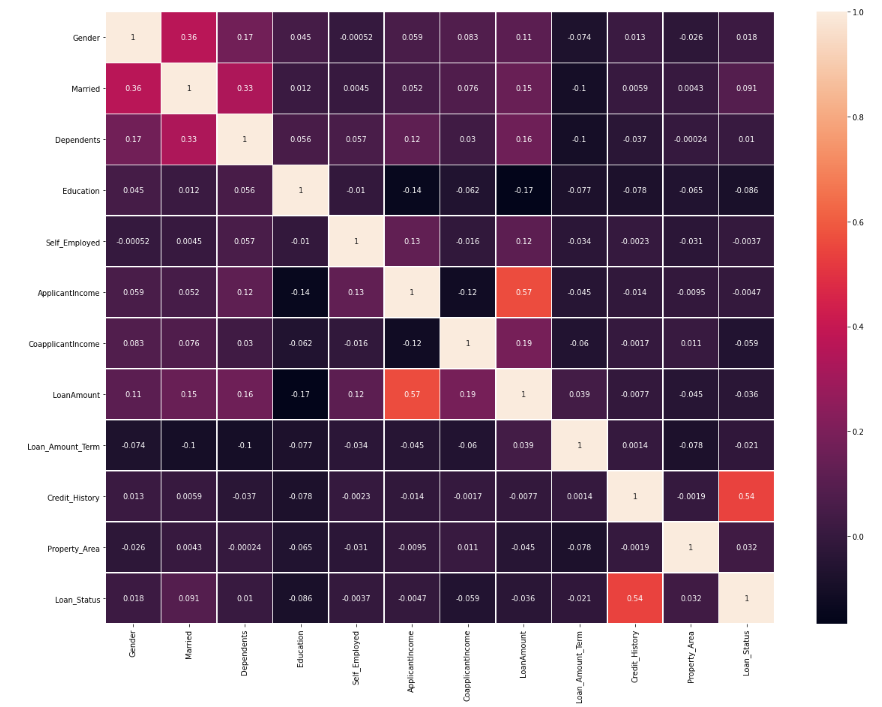
Applicants with 0 dependents have higher number of acceptance for loan.

Graduate or highly qualified people are more likely to get accepted for loan.

Employment rate for men is higher compared to female.

All these things come into consideration for loan approval of an individual as people with good education, 0 dependents have less chance of fraud according to above dataset.

Checking for correlation in the data :-

we can see that **‘**self employed**’** have the lowest correlation with the loan Status with corr of -0.0037.

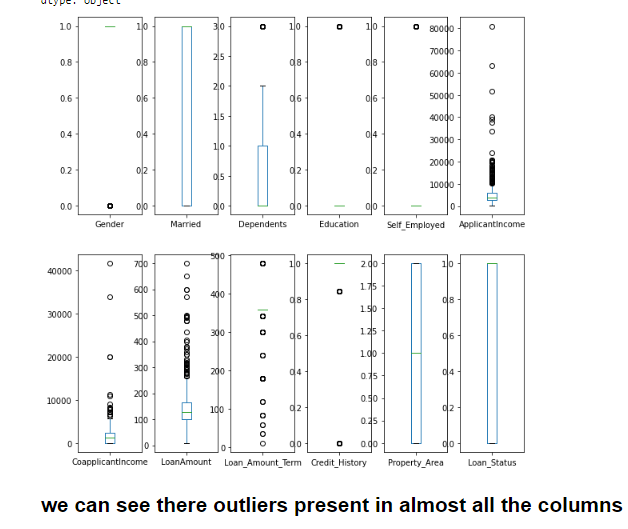
Also column Applicant income with -0.0046 correlation with loan status.

Property area have 0.032 correlation with the loan status.

There are also some more columns present with low correlation like gender and co-applicant income but that’s ok.

Credit history have the highest correlation with loan status compared to all others.

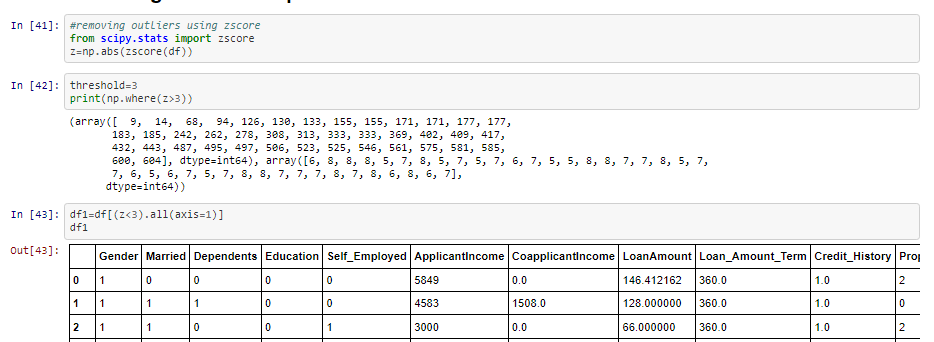
Checking outliers in the data :-



What we can see that there are outliers present in the data set which we need to clean.

Columns like Loan amount, co-applicant income, loan\_amount\_term and applicant income have the highest outliers compared to other input variables.

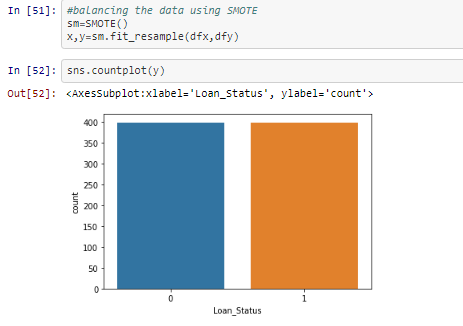
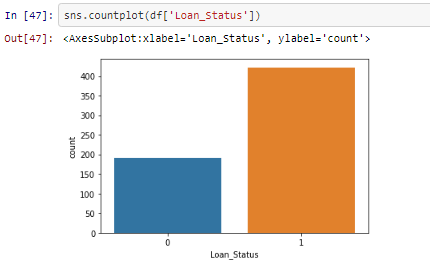
we need to clean the data for proper processing and buildng prediction models for the data.



We used Zscore method to remove outliers from the data with the threshold set to 3.

So the cleansing is done.

Balancing the unbalanced data:-

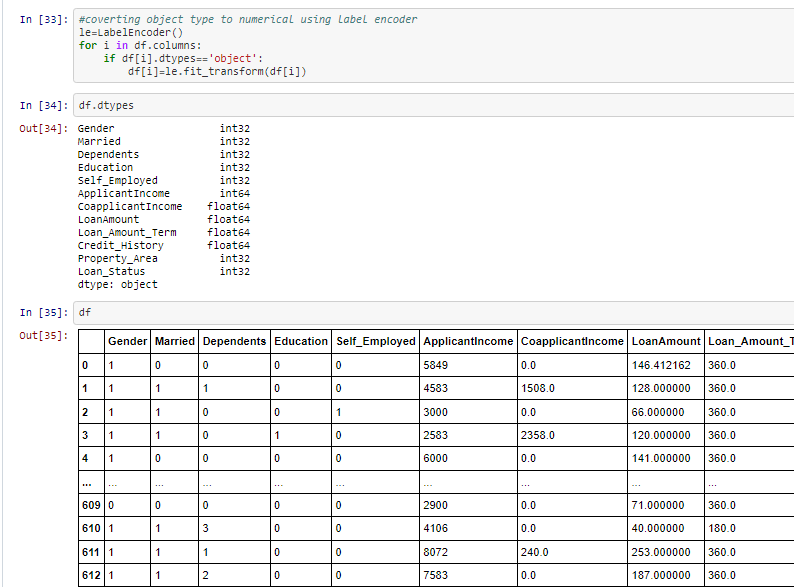


We saw that the data was unbalanced for loan status or we can say the dependent variable of our dataset.

So we used Smote to use over sampling technique to balance the dataset so that we can predict good results for our model and help it make better predictions.

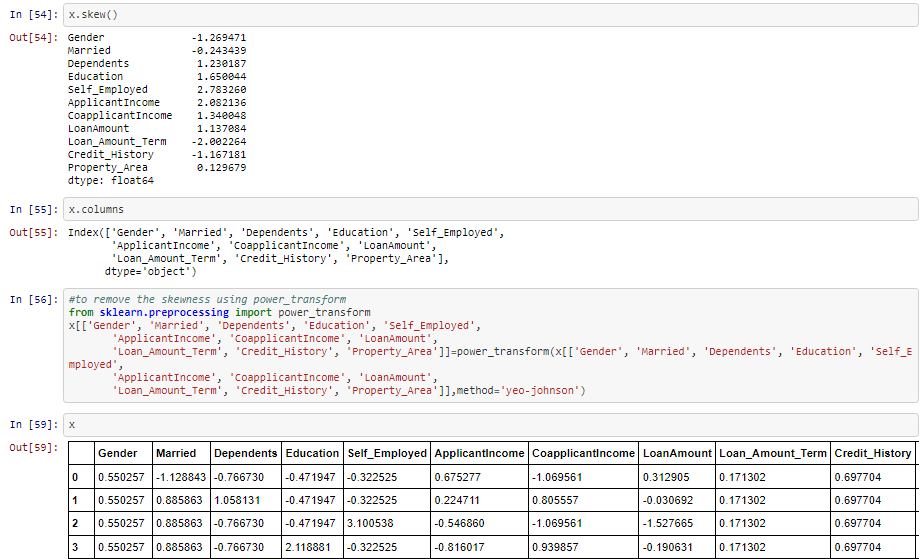
**Pre-processing Pipeline :-**

Before we moving on to prediction of models we did some processing of data like we converted object type data to numerical data so that we can use it for prediction of data.

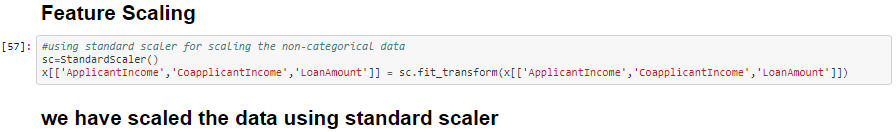


We used Label Encoder to convert all the object data to numerical.

Also we saw that there was some skewness present in the data and that could lead us to biasing of the data. So we need to remove the biasing so that our model can work better.



We used power transform method to remove the biasing so that it won’t affect our prediction model.



Also we used standard scaler to do some scaling to the data before moving on and building prediction models.

Pre-processiong is an important step as it directly affect the ablitiy of our model to learn, therefore it is extremely important that we pre-process our data before feeding it to the model.

**Building Machine Learning Models :-**

Before predicting we split our data where ‘X’ hold all the input variables and ‘Y’ holds the output variable which is loan status. Also it contains only 2 unique values which is either yes or no.

So it’s a classification problem so we need to use classification models to predict the data.

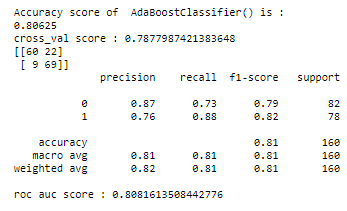
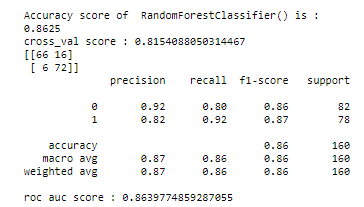
We need to first train the data with 80% of the data set and then the model will learn from the patterns of the dataset and it will help the model to predict the data on test data which is other 20% of the dataset.

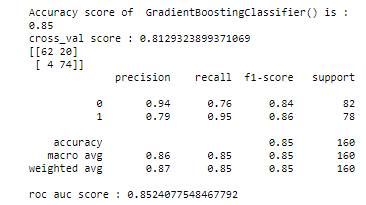
We will be checking accuracy score, cross validation score, and roc-auc score as well.

Machine Learning analyzes large chunks of data automatically. Machine Learning basically automates the process of Data Analysis and makes data-informed predictions in real-time without any human intervention. A Data Model is built automatically and further trained to make real-time predictions. This is where the Machine Learning Algorithms are used in the Data Science Lifecycle.

****

After predicting of data we found the results of our best performing models which are shown below

****

****

We choose Random Forest Classifier to hyper tune.

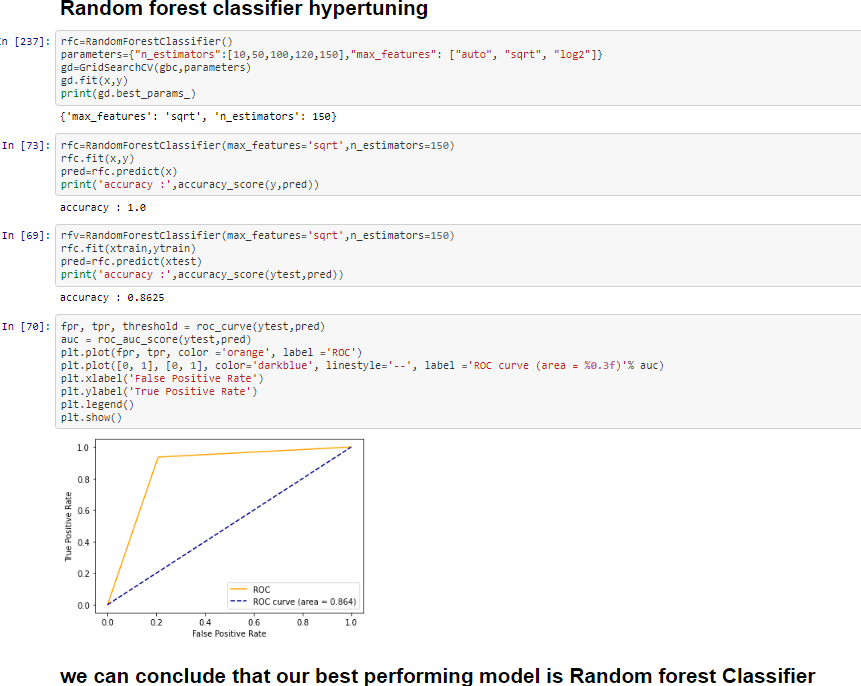
**Random forest classifier:-**

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

Random forests has a variety of applications, such as recommendation engines, image classification and feature selection. It can be used to classify loyal loan applicants, identify fraudulent activity and predict diseases. It lies at the base of the Boruta algorithm, which selects important features in a dataset.

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model’s prediction.

Hyper tuning the model:

****

we can see from the above image that Random Forest Classifier is our best performing model for this dataset when it comes to predicting data for loan status of the applicants.

Also check the Roc-auc curve

****

Saving the best model.

**Concluding Remarks :-**

**After testing all our classification models we found out that**

**Random forest classifier is the best performing model with close 86% ROC score after hypertuning**

**after Random forest ADAboost and gradient boosting classifier is the best performing model with close to 83 % roc score.**

**We choose to save Random forest as our best performing model for this data set when it comes to predicting Loan Status.**