**Submitted by : Rahul jangra**

PGP IN DATA SCIENCE , Btach : 1832

**LOAN APPLICATION STATUS PREDICTION**

**Abstract :**

Lately people depend on bank loans to meet their wishes. The fee of loan packages will increase with a very rapid speed in current years. Risk is constantly involved in approval of loans. The banking officials are very acutely aware of the price of the mortgage quantity by its customers. Even after taking lot of precautions and analyzing the mortgage applicant information, the mortgage approval choices are not continually correct. There is need of automation of this system so that loan approval is much less risky and incur less loss for banks. Since it is a major activity for the banks,to identify whether a loan of the desired amount should be approved to the applicant or not,the Computer Science is capable of making such a system using Artificial Intelligence,which can make this tough decision accurately and quickly. Using data science,which is responsible to deal with the large amount of data efficiently,and some algorithms of Machine Learning,a prediction system is made,which,on the basis of some training data set,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.

**INTRODUCTION**

The two most pressing issues in the banking sector are: 1) How risky is the borrower? 2) Should we lend to the borrower given the risk? The response to the first question dictates the borrower's interest rate. Interest rate, among other things (such as time value of money), tests the riskiness of the borrower, i.e. the higher the interest rate, the riskier the borrower. We will then decide whether the applicant is suitable for the loan based on the interest rate. Lenders (investors) make loans to creditors in return for the guarantee of interest-bearing repayment. That is, the lender only makes a return (interest) if the borrower repays the loan. However, whether he or she does not repay the loan, the lender loses money. Banks make loans to customers in exchange for the guarantee of repayment. Some would default on their debts, unable to repay them for a number of reasons. The bank retains insurance to minimize the possibility of failure in the case of a default. The insured sum can cover the whole loan amount or just a portion of it. Banking processes use manual procedures to determine whether or not a borrower is suitable for a loan based on results. Manual procedures were mostly effective, but they were insufficient when there were a large number of loan applications. At that time, making a decision would take a long time. As a result, the loan prediction machine learning model can be used to assess a customer's loan status and build strategies. This model extracts and introduces the essential features of a borrower that influence the customer's loan status. Finally, it produces the planned performance (loan status). These reports make a bank manager's job simpler and quicker.

**PROBLEM STATEMENT:**

Loan approval is a completely important procedure for banking businesses. The system approve or reject the mortgage applications. Recovery of loans is a first-rate contributing parameter in the economic statements of a bank. It may be very hard to are expecting the possibility of fee of loan through the purchaser. In current years many researchers worked on mortgage approval 5 prediction structures.

This dataset includes details of applicants who have applied for loan. The dataset includes details like credit history, loan amount, their income, dependents etc. It includes the factors affecting the the decision of loan approval. Below are the factors or features given in the dataset:

**Independent Variables:**

- Loan\_ID

- Gender

- Married

- Dependents

- Education

- Self\_Employed

- ApplicantIncome

- CoapplicantIncome

- Loan\_Amount

- Loan\_Amount\_Term

- Credit History

- Property\_Area

**Dependent Variable (Target Variable):**

- Loan\_Status

Independent variables are those variables which are not dependent on any other variable means the change in other variable will not affect these variables , whereas dependent variables are those which are dependent on others , here loan status is dependent on the features given in the dataset .

**OBJECTIVE**:

Lately people depend on bank loans to meet their wishes. The fee of loan packages will increase with a very rapid speed in current years. Risk is constantly involved in approval of loans. The banking officials are very acutely aware of the price of the mortgage quantity by its customers. Even after taking lot of precautions and analyzing the mortgage applicant information, the mortgage approval choices are not continually correct. There is need of automation of this system so that loan approval is much less risky and incur less loss for banks.

You have to build 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 FORMATS IN THE DATASET :**

There are 3 types of data i.e. Categorical, Ordinal, Numerical data.

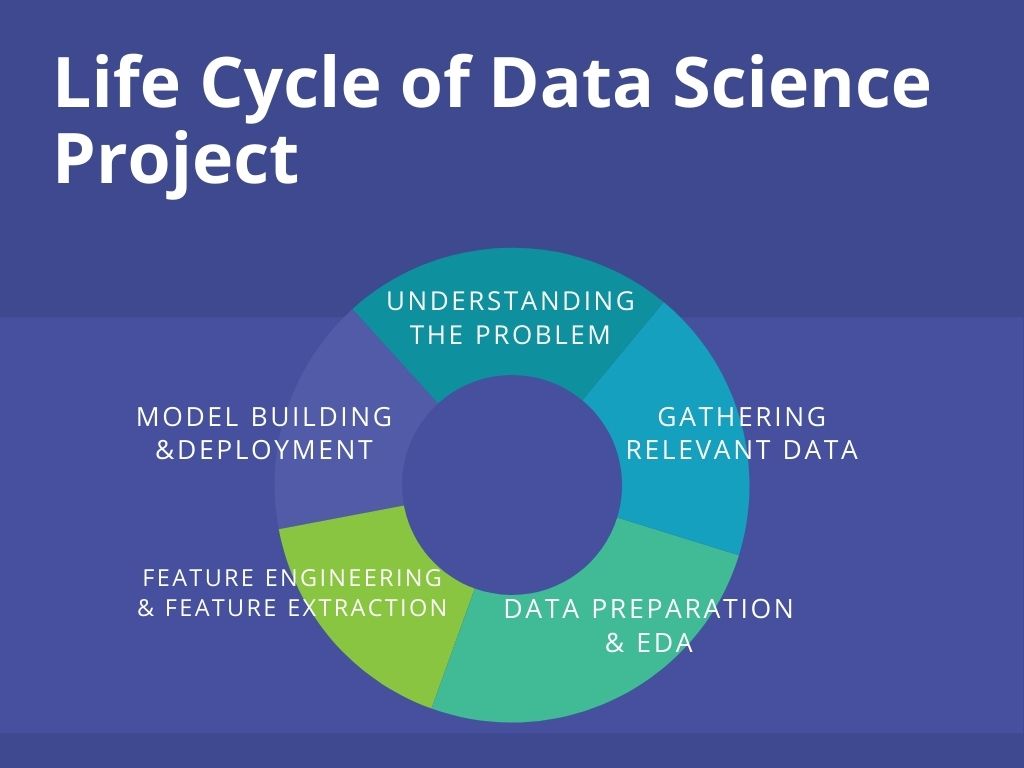
* **Data set Source :** Dataset Link

<https://github.com/dsrscientist/DSData/blob/master/loan_prediction.csv>

**METHODOLOGY**

Since it is a major activity for the banks,to identify whether a loan of the desired amount should be approved to the applicant or not,the Computer Science is capable of making such a system using Artificial Intelligence,which can make this tough decision accurately and quickly. Using data science,which is responsible to deal with the large amount of data efficiently,and some algorithms of Machine Learning,a prediction system is made,which,on the basis of some training data set,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 Science Project lifecycle :**

****

* **UNDERSTANDING THE PROBLEM :**

The problem statement stage is the first and most important step of solving an analytics problem. It can make or break the entire project. When a business approaches a data scientist with a problem they want to solve, they will always define the problem in layman’s terms. This means the problem will not be clear enough, from an analytics point of view, to begin solving it right away. The problem needs to be well framed.

* **GATHERING THE RELEVANT DATA :**

The 2nd step is to accumulate statistics.As already stated inside the project charter which records are needed and where to find that data. In this step it is ensured that you can use the statistics to your program, which means checking the existence of, quality, and get access to to the data. Data also can be added by means of third-party businesses and takes many forms starting from Excel spreadsheets to different sorts of databases.

* **DATA PREPARATION AND EDA :**

Data exploration is concerned with building a deeper expertise of records. You try to understand how variables engage with every other, the distribution of the records, and whether there are outliers. To attain this particularly use descriptive statistics, visual techniques, and simple modeling. This step is called Exploratory Data Analysis and regularly is going by the abbreviation EDA.

* **FEATURE ENGINEERING AND FEATURE EXTRACTION :**

Feature selection is nothing but a selection of required independent features. Selecting the important independent features which have more relation with the dependent feature will help to build a good model. Handing the outliers and skewness also cover in this .

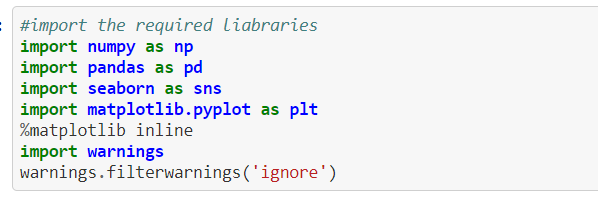
* **MODEL BUILDING AND DEPLOYMENT:**

Now , we’ll train our data by using various ML alogrithms and chose the best performing model and by using the best model we will predict the target .

**LANGUAGE AND IDE USED :**

Python , Jupyter notebook.

**IMPORTING THE REQUIRED LIABRARIES AND PACKAGES**

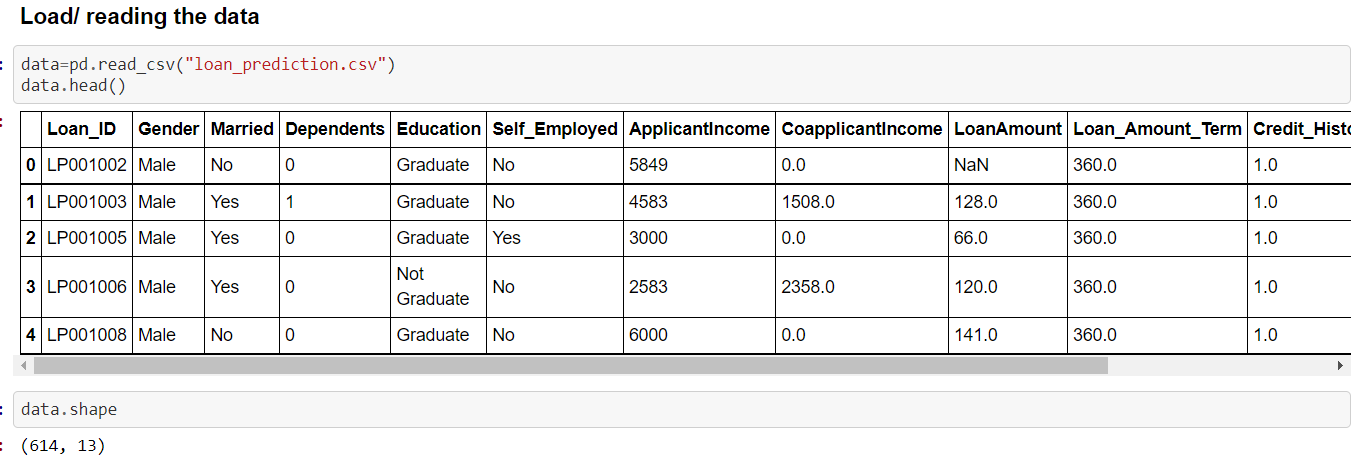
****

***NUMPY:***  NumPy, which stands **for Numerical Python**, is a library consisting of multidimensional array objects and a collection of routines for processing those arrays. Using NumPy, mathematical and logical operations on arrays can be performed.

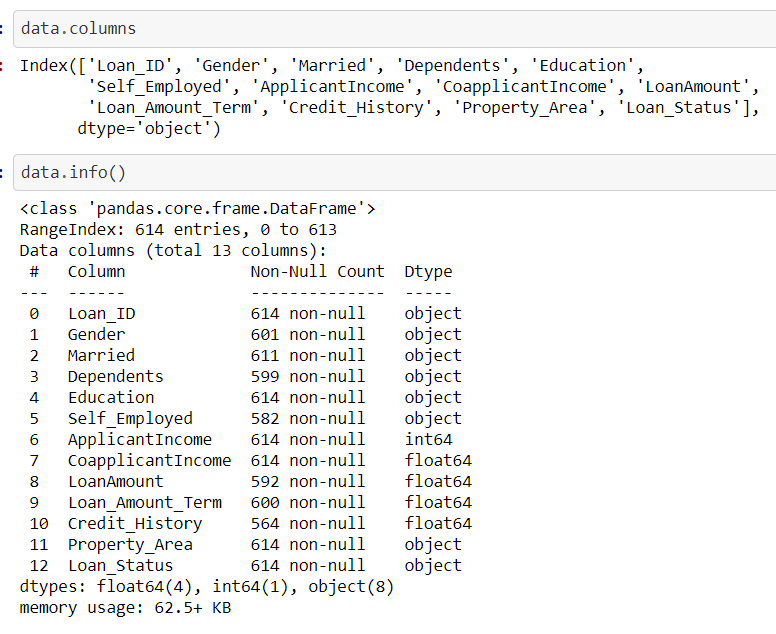
***PANDAS:*** Pandas is a Python package to work with structured and time series data.The data from various file formats such as csv, json, sql etc can be imported using Pandas. It is a powerful open source tool used for data analysis and data manipulation operations such as data cleaning, merging, selecting as well wrangling.

***SEABORN AND MATPLOTLIB:*** Seaborn and matplotlib is a python library for building graphs to visualise data. It provides integration with pandas. This open source tool helps in defining the data by mapping the data on the informative and interactive plots. Each element of the plots gives meaningful information about the data.

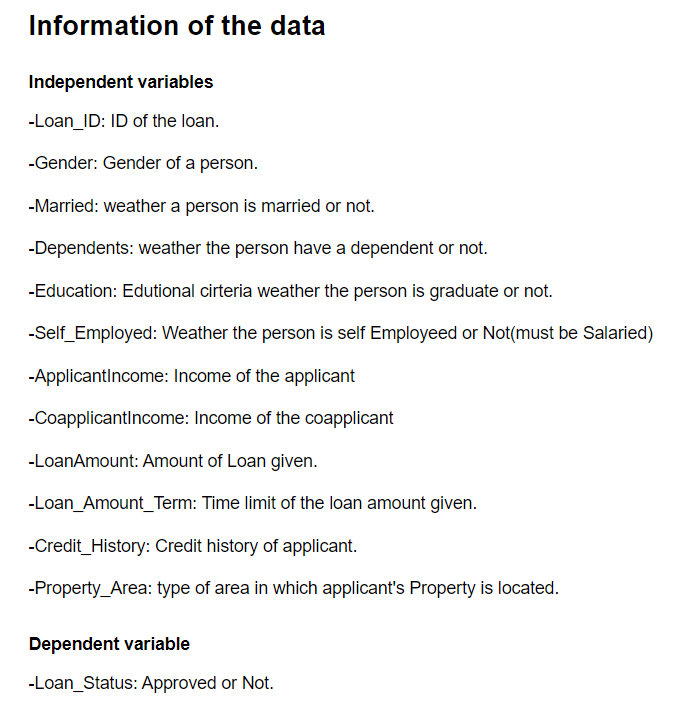
**IMPORTING / READING THE DATASET :**

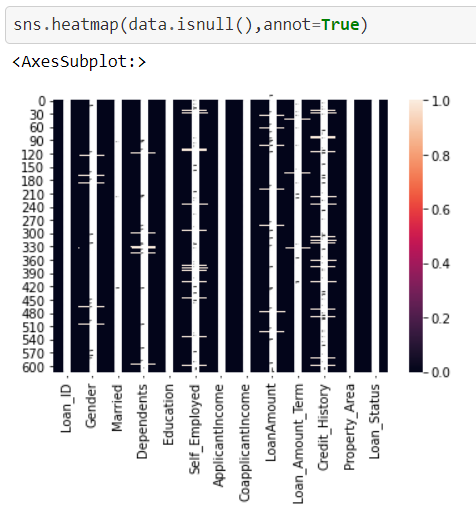
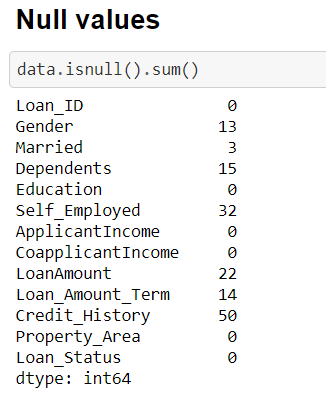
****

The machine learning model is trained using the training data set. Every new applicant details filled at the time of application form acts as a test data set. On the basis of the training data sets, the model will predict whether a loan would be approved or not. 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.

Below are the columns name and the datatypes of the columns :

Understanding the data :



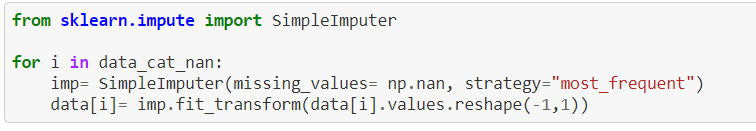


Null values are there in the datset , so now we’ll try to fill the null values.

First we make the copy of the categorical columns having null values so that we can pass them in a code in one go .



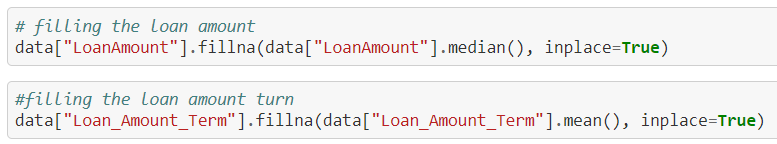
Through above code , we made a copy of columns to pass , as below :



Successfully we filled the null values in 4 categorical columns using simple imputer , mode technique .

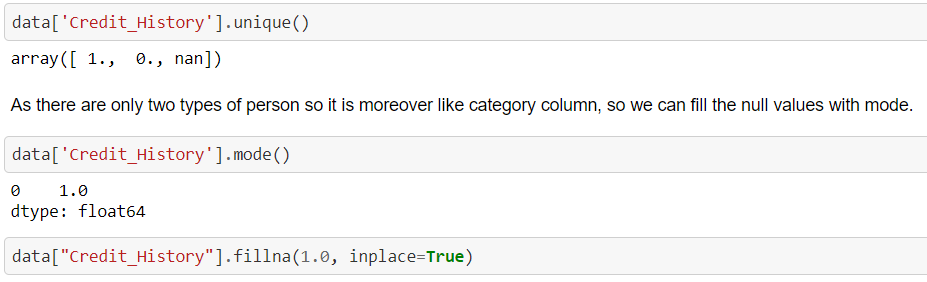
**SIMPLE IMPUTER :** The SimpleImputer class provides basic strategies for imputing missing values. Missing values can be imputed with a provided constant value, or using the statistics (mean, median or most frequent) of each column in which the missing values are located.

Now , we will fill the null values of int , float type columns, using fillna method :

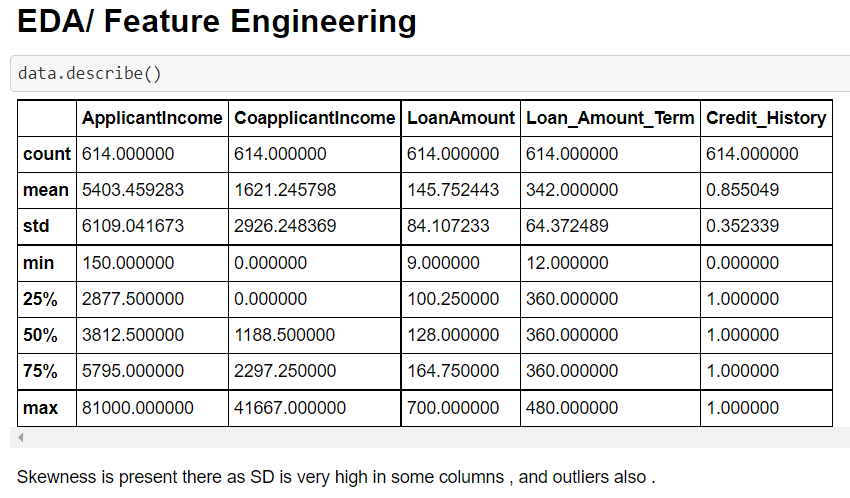


We filled the null values of “LoanAmount” with median and “LoanAmout Term” with mean .

Now , we’ll check the credit history column :

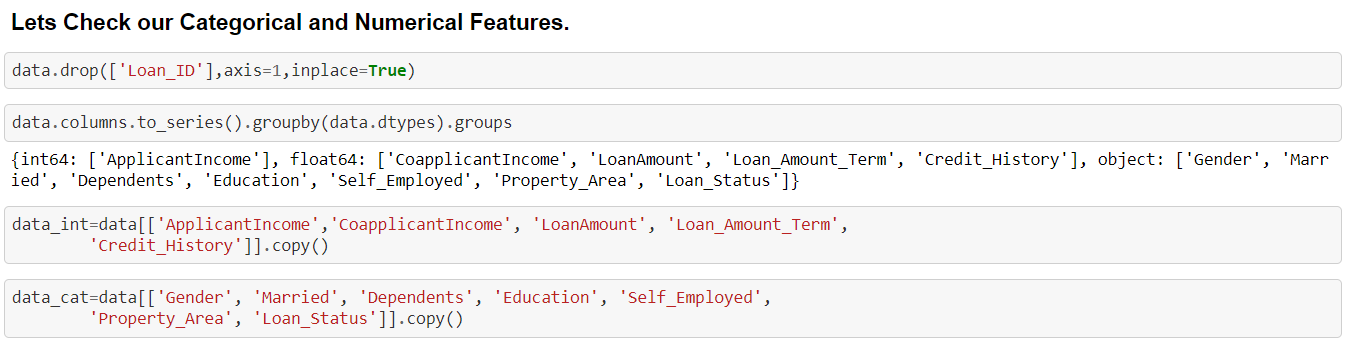


Now , we don’t have any null values in the dataset , so we can proceed further for Data Analysis and EDA.

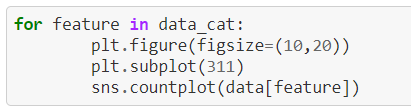


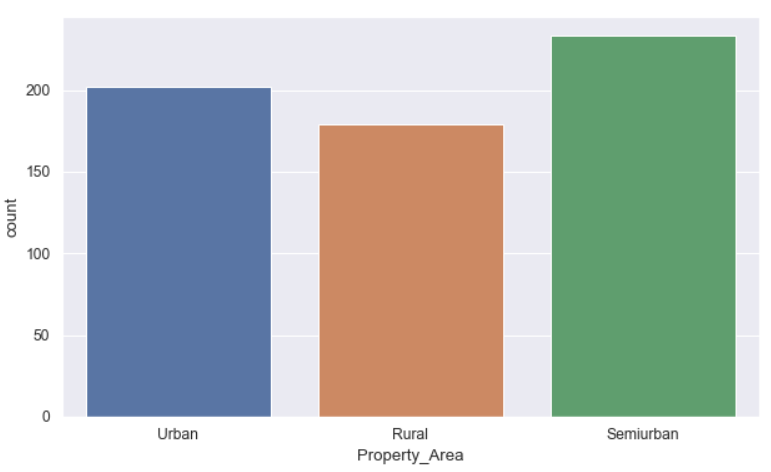
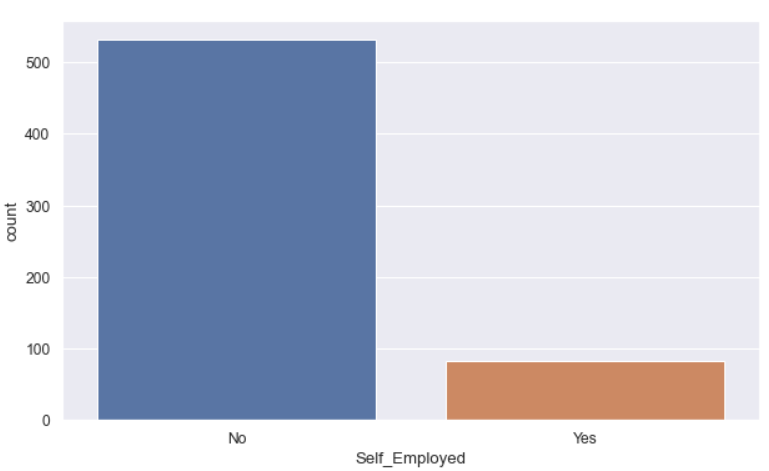
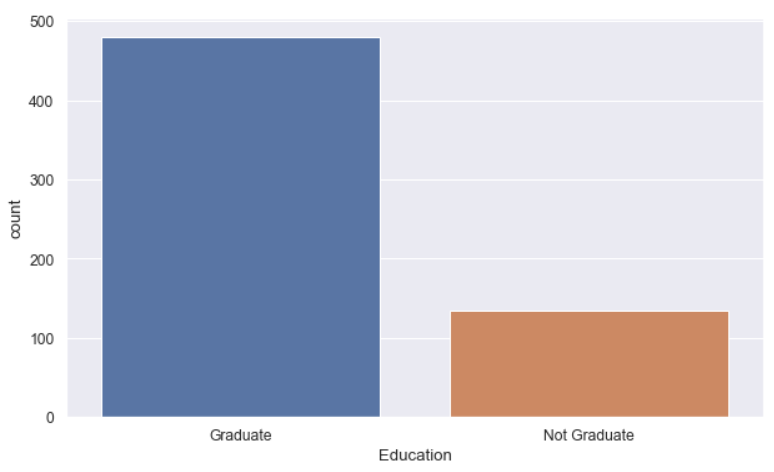
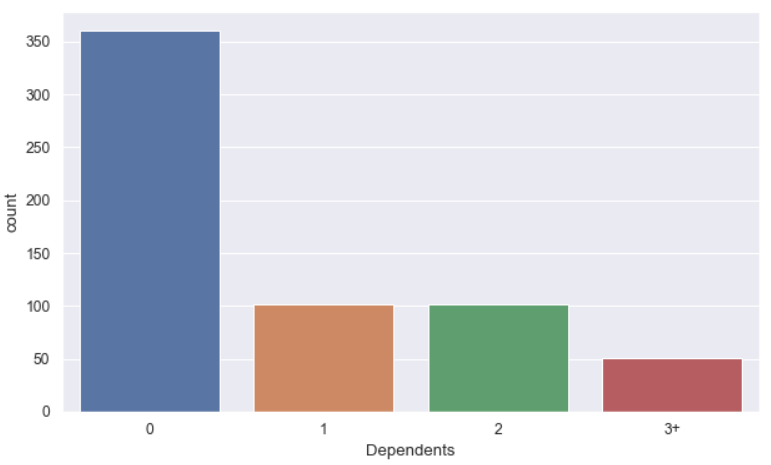
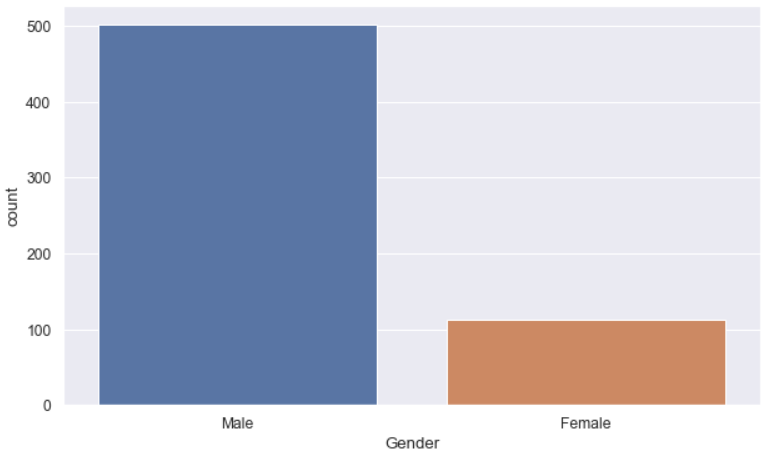
There is high skewness and outliers are present in Applicant income , Loan amount as well because SD is very high and there is huge difference in 75%ile and max value which indicates the outliers .

Now we’ll visualize the categorical and numerical data and will find some insights:



Removed the Loan id column , because it is of no use in decision of loan status . And we made a sepearte dataframe for categorical and int data column each .

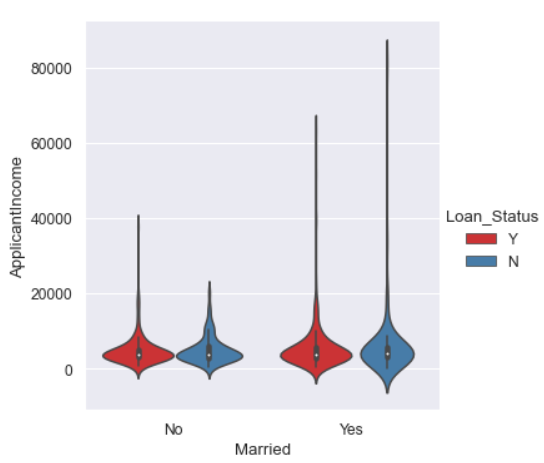
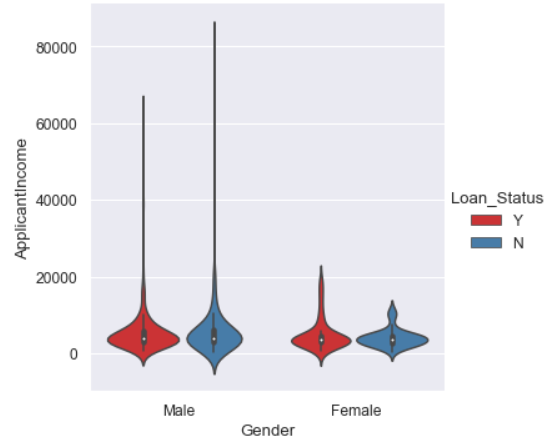
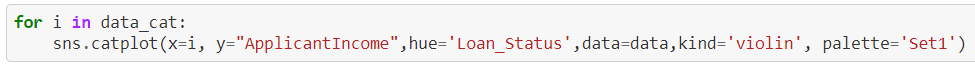


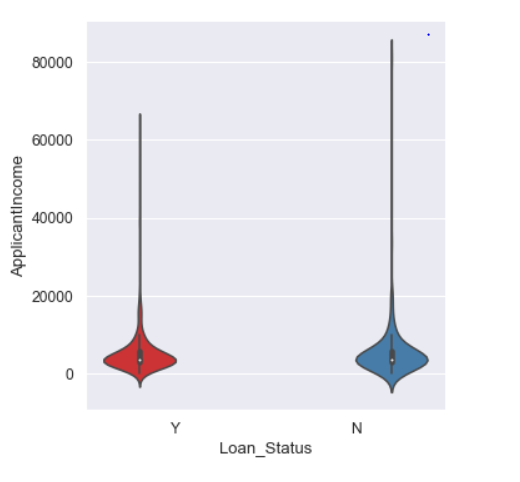
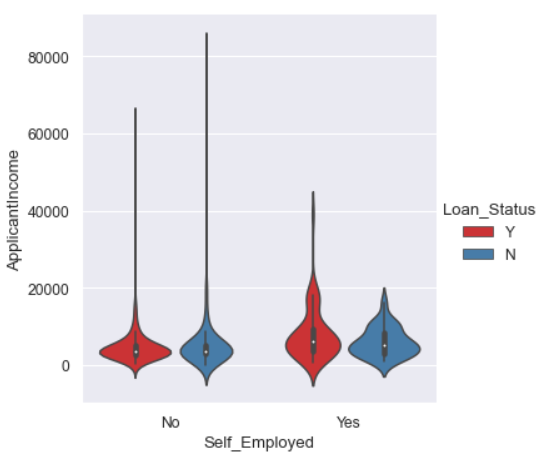
Using above code we plotted the count plot of each categorical column : 

Observations :

Most of the applicants are Male. Most of the applicants are Graduate, Most of the applicants are Married, Most of the applicants have no dependents, Most of the applicants are not self employed, Most of the applicants are from semi urban region.

Now , we’ll visualize the catergorical columns in comparison to applicant income , to know the loan status trends :

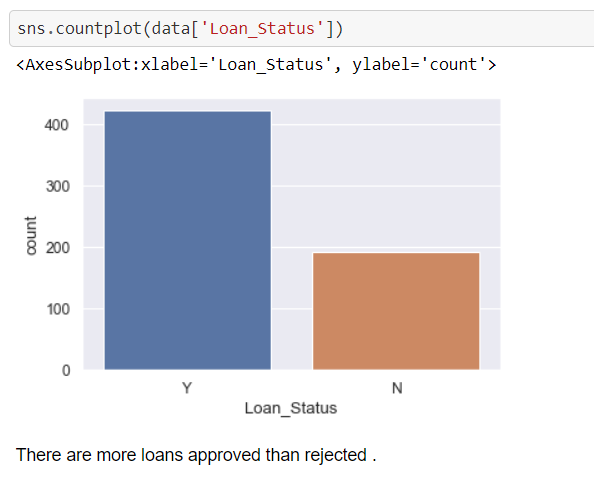




Observations :

From the above sets of graphs we can see that the males have silghtly higher chances of Loan approval, and Married people also have high chances for approval as total income is high compare to unmarried applicant,most of gradutaes have loan approvals althouth people with high income has loan status as no, might be because of credit score and as most of properties of applicants with high income are from rural areas.

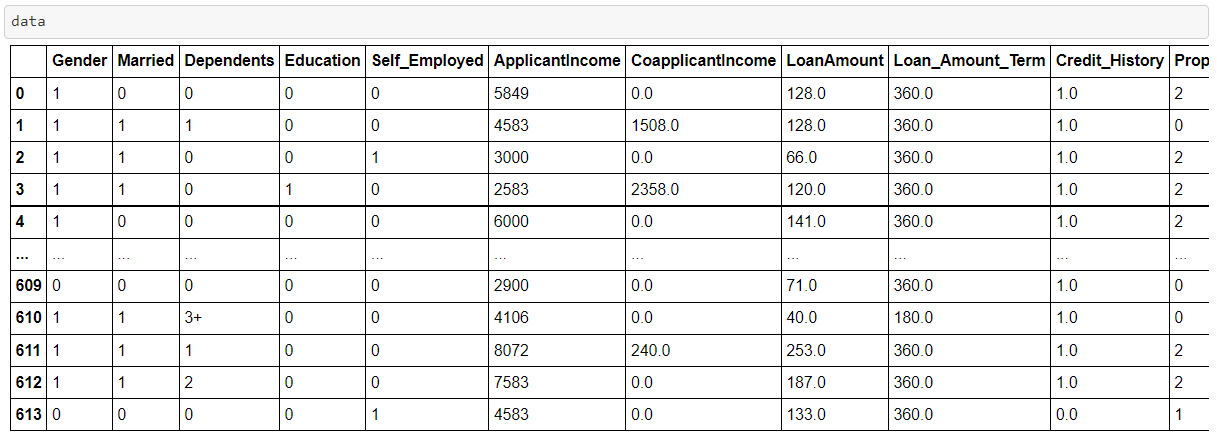
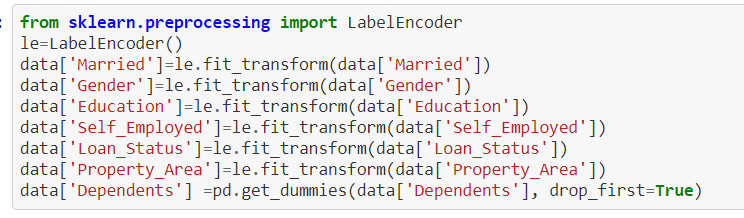
Similarly, We can perform the other visualizations .



In the provided dataset , there are more approved loans than rejected cases.

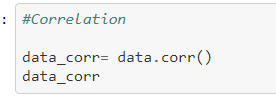
**Label Encoding :**

Label Encoding helps us **to converting the labels into a numeric form so as to convert them into the machine-readable form**. Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning.

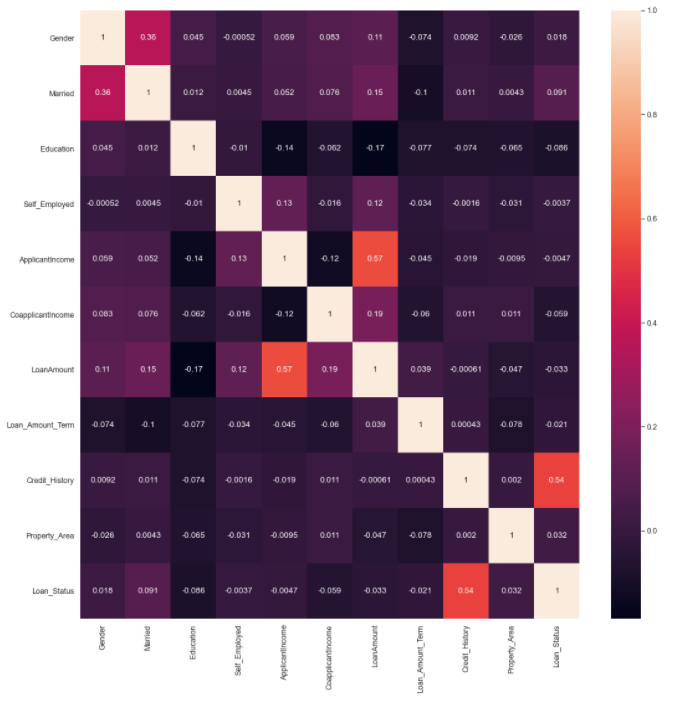
****

After performing the label encoding now we have all data in numeric form , because machine doesn’t understand the other data . Now we can check the correlation of features with target column , by using corr method.

**Correlation :**

****

****

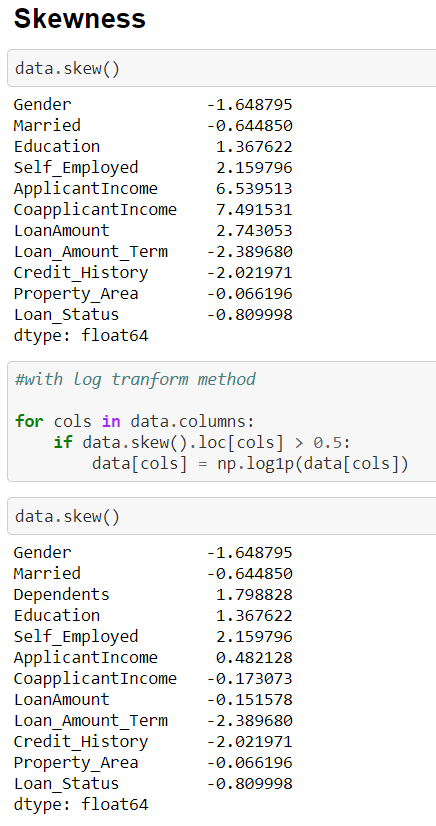
****

From above we can see , that credit history has positive correlation with loan status means if applicant has good credit history , his/her loan got approved .

Note : In this correlation heatmap . Darker the color – negative the correlation , Lighter the color – Positive the correlation .

**SKEWNESS :**

Skewness describes how much statistical data distribution is asymmetrical from the normal distribution, where distribution is equally divided on each side. If a distribution is not symmetrical or Normal, then it is skewed, i.e., it is either the frequency distribution skewed to the left side or to the right side*.*



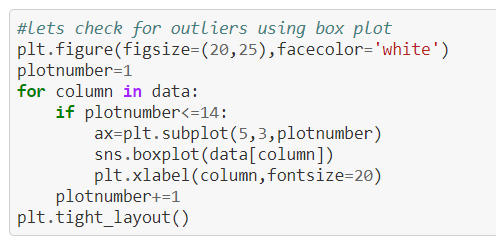
**Log Transform method**

When our original continuous data do not follow the bell curve, we can log transform this data to make it as “normal” as possible so that the statistical analysis results from this data become more valid . In other words, the log transformation **reduces or removes the skewness of our original data**.

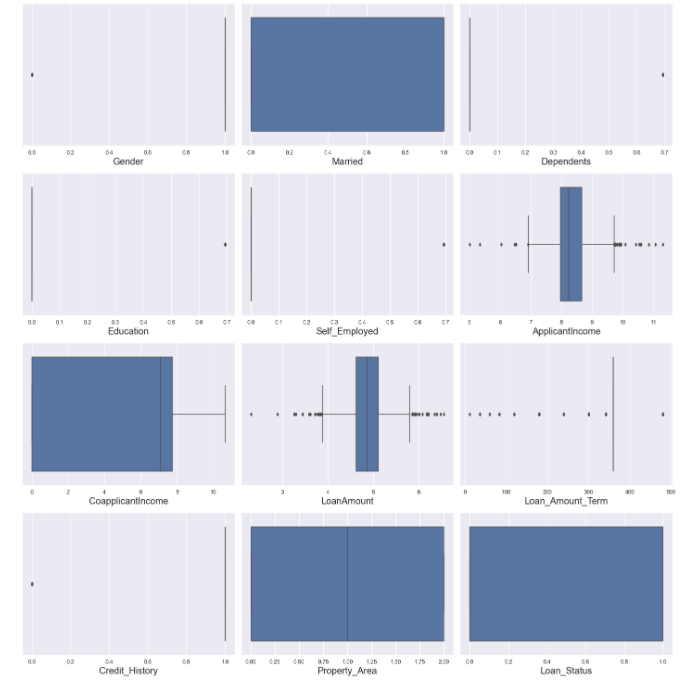
**Outliers :**

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal. Before abnormal observations can be singled out, it is necessary to characterize normal observations. We can detect and remove the outliers by using z-score method .

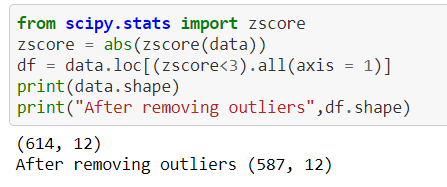
**Z-score :** Simply put, a z-score (also called a standard score) gives you **an idea of how far from the mean a data point is**. But more technically it's a measure of how many standard deviations below or above the population mean a raw score is. A z-score can be placed on normal distribution curve. And we treat points below or above 3 standard deviations as outliers .

****

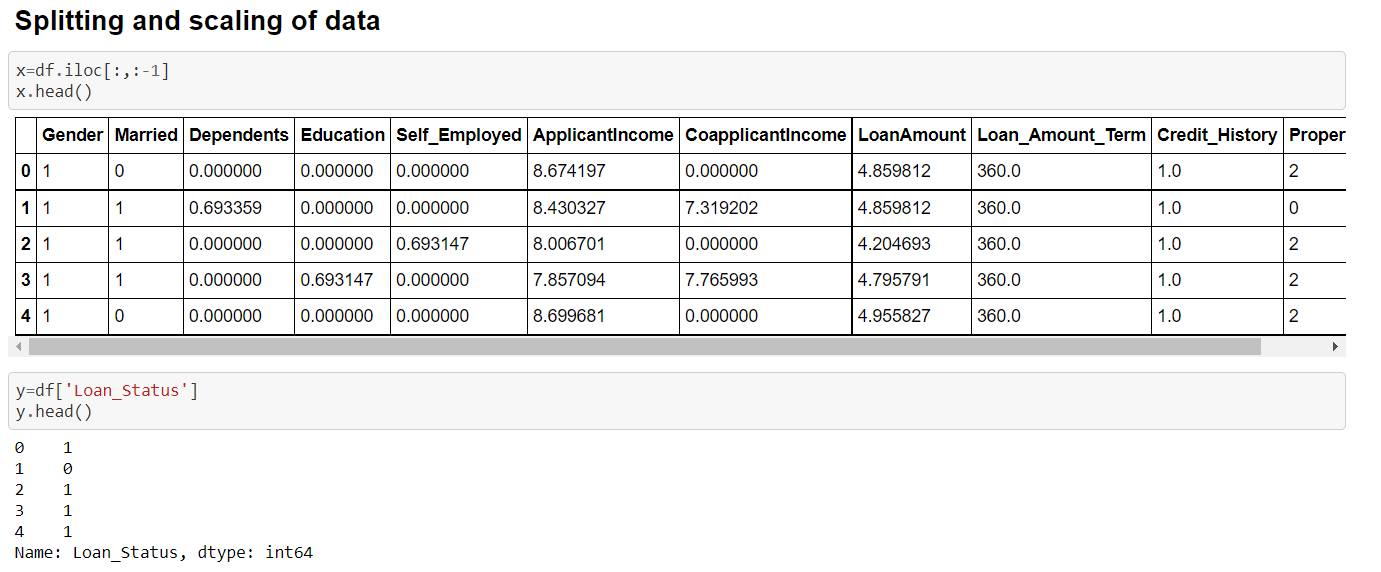
Through above code we can plot the box plot which gives us an idea about outliers as below :



Code for detecting and removing the outliers :



So , here after removing the ouliers we have 587 rows and 12 columns in our dataset . Now we can proceed further for building the model .



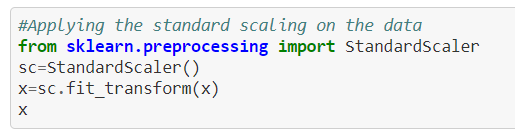
Through above code we splitted the data into two categories ,one : Feature columns and 2. Target columns , and stored them in different variables i.e. x and y .

**Scaling of Data :** When your data has different values, and even different measurement units, it can be difficult to compare them.Data scaling is a recommended pre-processing step when working with many machine learning algorithms.

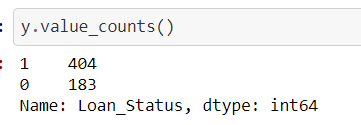
Data scaling can be achieved by normalizing or standardizing real-valued input and output variables. We can apply the StandardScaler to the dataset directly to standardize the input variables.

We will use the default configuration and scale values to subtract the mean to center them on 0.0 and divide by the standard deviation to give the standard deviation of 1.0. First, a StandardScaler instance is defined with default hyperparameters.

Once defined, we can call the fit\_transform() function and pass it to our dataset to create a transformed version of our dataset .

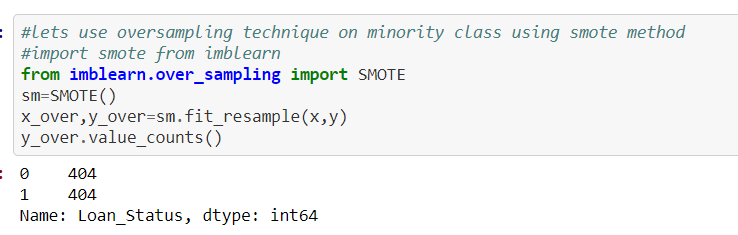


Now , we’ll check the values in the output/ Target column:



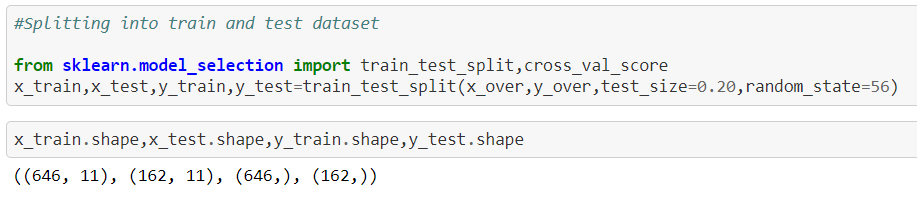
As , the values are imbalanced in the classes of y columns , we ‘ll apply the SMOTE technique to balance the classes .

**SMOTE:** Through SMOTE we can balance or equal the target column class which will help to learn the data better to build the model .

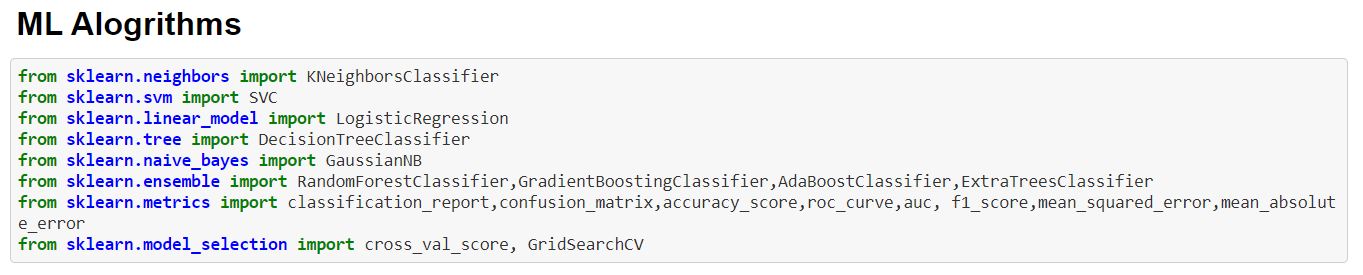


Through oversampling we increase the 183 to 404 and made it equal .

Now , our data is ready to go for training and testing , we’ll split the data in train and test . to fit the train data to models to learn , and test data to test the accuracy of the model .



Now , we’ll import the ML alogrithms liabraries:



Below are the description of models :

**Logistic Regression**

This is a classification algorithm which uses a logistic function to predict binary outcome (True/False, 0/1, Yes/No) given an independent variable. The aim of this model is to find a relationship between features and probability of particular outcome. The logistic function used is a logit function which is a log of odds in the favor of the event. Logit function develops a s-shaped curve with the probability estimate similar to a step function.

**Decision Tree classifier**

This is a supervised machine learning algorithm mostly used for classification problems. All features should be discretized in this model, so that the population can be split into two or more homogeneous sets or subsets. This model uses a different algorithm to split a node into two or more sub-nodes. With the creation of more sub-nodes, homogeneity and purity of the nodes increases with respect to the dependent variable.

**Random Forest Classifier**

This is a tree based ensemble model which helps in improving the accuracy of the model . It combines a large number of Decision trees to build a powerful predicting model. It takes a random sample of rows and features of each individual tree to prepare a decision tree model. Final prediction class is either the mode of all the predictors or the mean of all the predictors.

**AdaBoost Classifier**

This algorithm only works with the quantitative variable. It is a gradient boosting algorithm which forms strong rules for the model by boosting weak learners to a strong learner. It is a fast and efficient algorithm which recently dominated machine learning because of its high performance and speed. Below we are defining the alogrithms code:

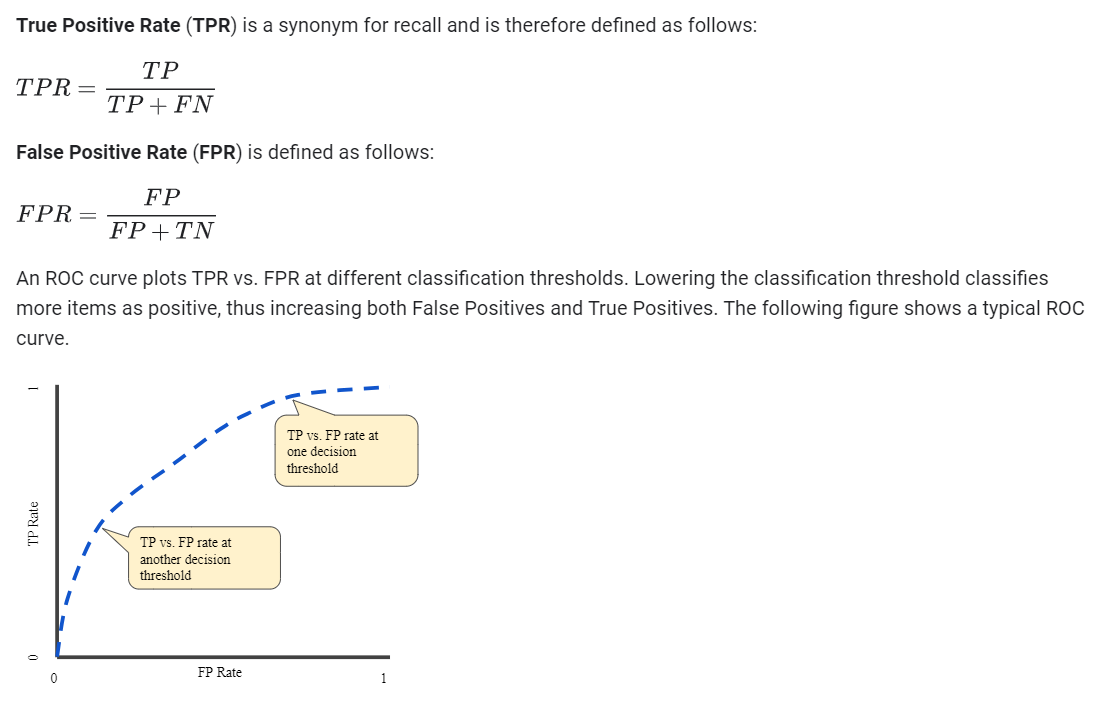


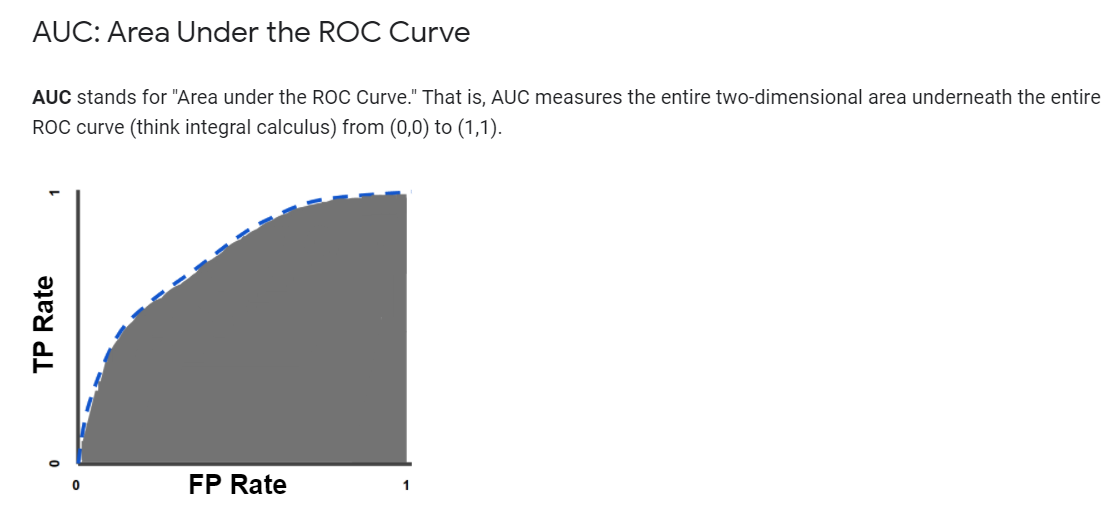


***Confusion Matrix :*** A confusion matrix is **a summary of prediction results on a classification problem**. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix.

***AUC/ROC :*** The **Area Under the Curve** (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes. An **ROC curve** (**receiver operating characteristic curve**) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:

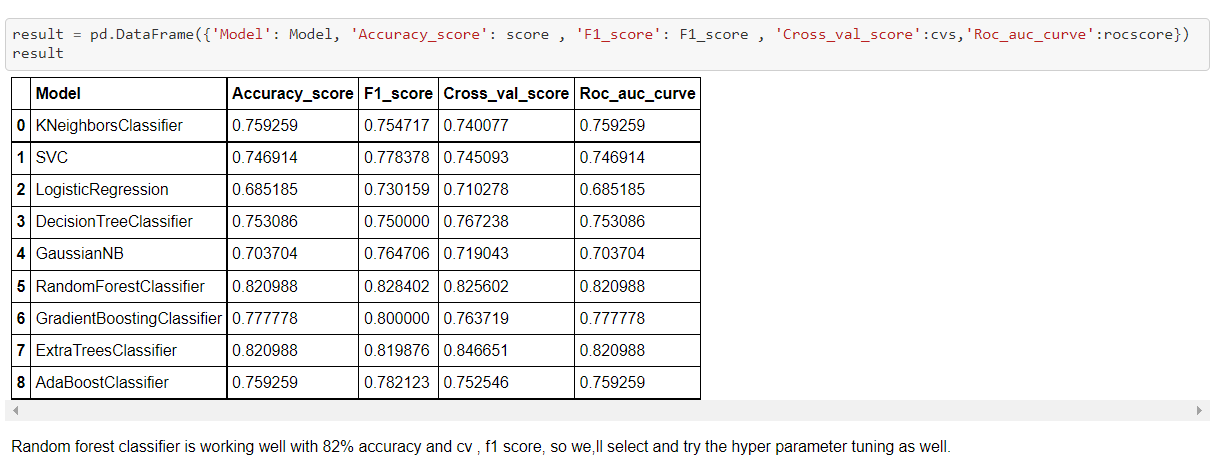
* True Positive Rate , False Positive Rate





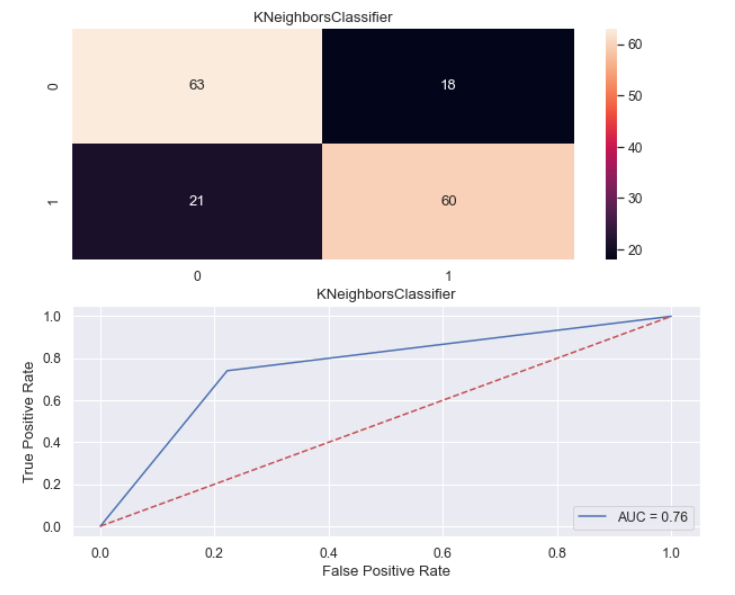
Below are the scores of different model Alogrithms we used to train our model:

**SCORES OF DIFFERENT MODELS**

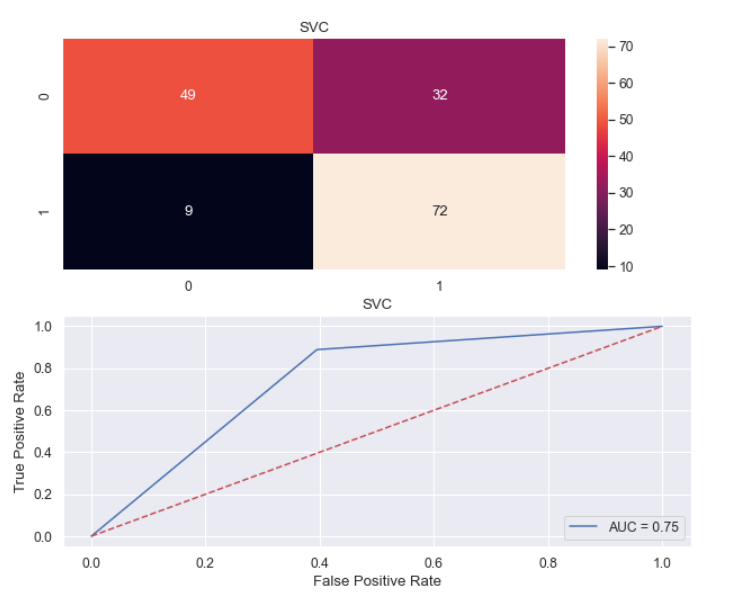


Below are the Confusion Matrix and AUC curve of all Models:

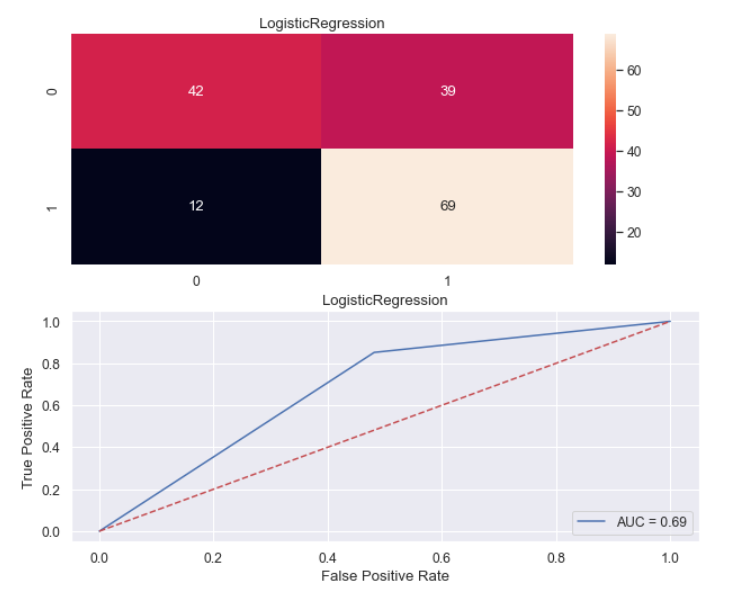
**KNN**



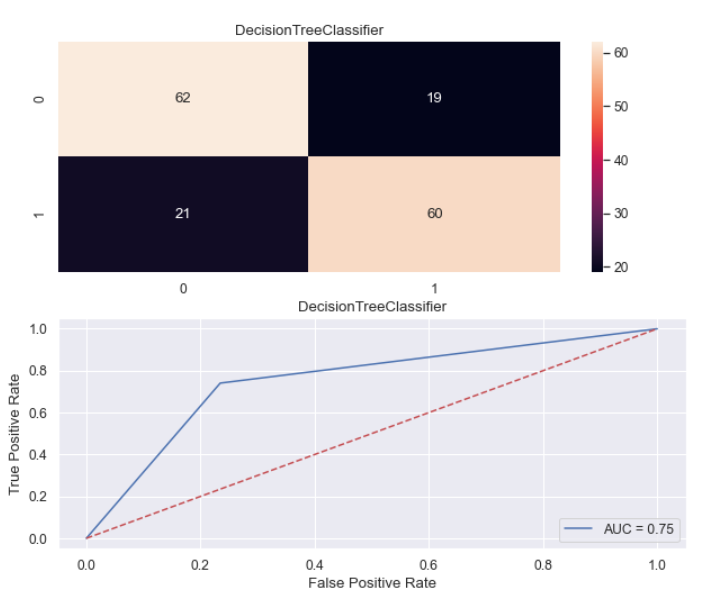
**SVC:**



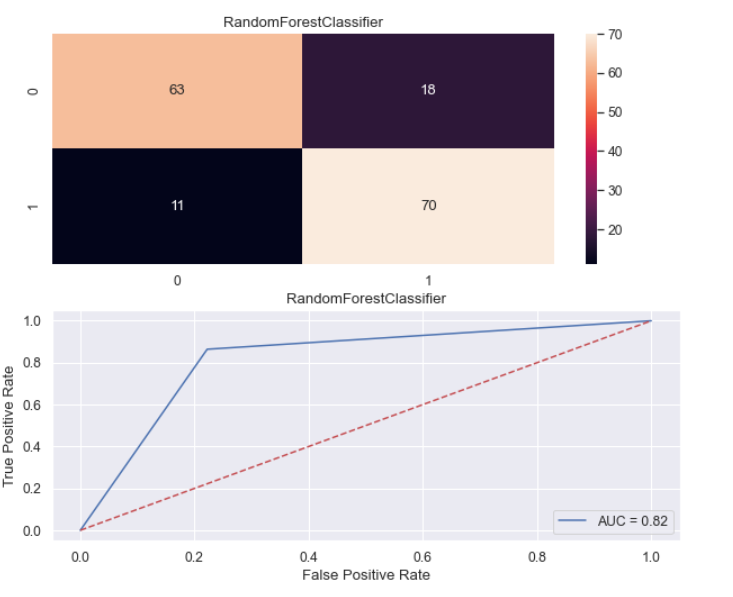
**Logistic Regression :**



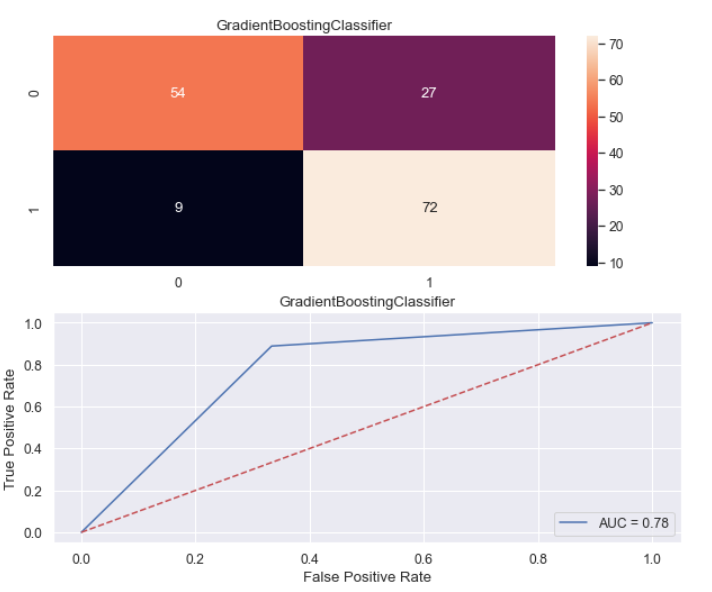
**Decision Tree Classifier :**



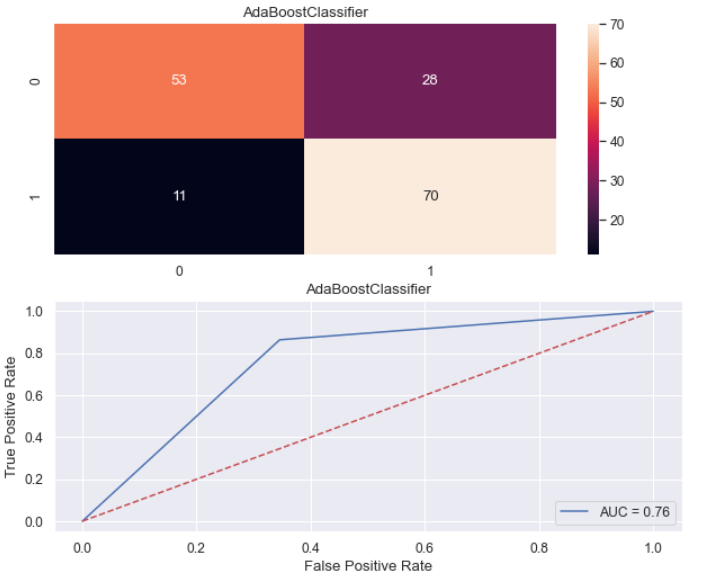
**Random Forest Classifier :**



**Gredient Boosting Classifier :**



**AdaBoost Classifier :**

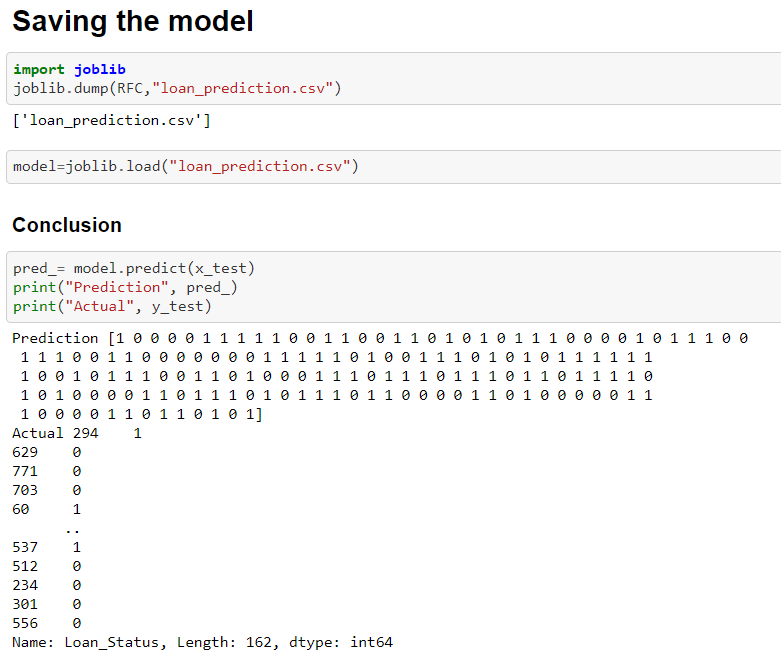


**From above , we can see that Random Forest Classifier is performing the best , with 82% of accuracy score and with best F1 score , AUC , ROC . We can do the hyperparameter tuning to increase the accuracy or performance of the model as below :**

**Hyperparameter Tuning :** In machine learning, hyperparameter optimization or tuning is the problem of choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a **parameter whose value is used to control the learning process.**



**SAVING THE MODEL AND LOADING IT TO PERFORM THE PREDICTION BY USING TEST DATA :**



**CONCLUSION OF PROJECT STUDY :**

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

**We did bivariate and multivariate analysis to see imapct 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 applicant income and loan amount have significant relation.**

**We constructed models taking different variables into account and found through that credit credit history is creating the most impact on loan giving decision**

**Finally, we got a model with coapplicant income and credit history as independent variable with highest accuracy.**

**We tested the data and got the accuracy of 82 % with Random Forest Classifier .**