**Loan Application Status Prediction**

We will be predicting Loan Application Status though Machine Learning model in this blog.

**Load Application Status Prediction**is a task that can be done based on historical information of the customer and bank,the relationship of customer with bank and previous loans by checking the credit history of customer.

First of all,we should know about what is machine learning

**Machine Learning**

**Machine Learning** is the field of study that gives computers the capability to learn without being explicitly programmed. ML is one of the most exciting technologies that one would have ever come across.

As Loan Application Status Prediction is a classification problem,so we will be using Classification models to figure out a working model for predicting loan app status.

**But , What are the classification models ?**

Classification model: A classification model tries to draw some conclusion from the input values given for training. It will predict the class labels/categories for the new data.

* **Logistic Regression**: Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable.
* Naive Bayes
* K-Nearest Neighbours
* Decision Tree
* Random Forest Classifier
* Support Vector Machine

Dataset used for Loan Application Status Prediction includes details of applicants who have applied for loan. The dataset includes details like credit history, loan amount, their income, dependents etc.

We will 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.

We know that any bank approves the loan of customer based on his previous credit history ,customer risk factor .

Now ,lets start creating the model

**Importing Modules and Libraries**

Importing all the modules and libraries required for the project

import pandas as pd

import numpy as np

from sklearn.preprocessing import LabelEncoder

import matplotlib.pyplot as plt

import seaborn as sns

# importing libraries

from sklearn.linear\_model import LogisticRegression

from sklearn.tree import DecisionTreeClassifier

from sklearn.svm import SVC

from sklearn.neighbors import KNeighborsClassifier

from sklearn.metrics import accuracy\_score,classification\_report,confusion\_matrix,roc\_auc\_score,roc\_curve

from sklearn.model\_selection import cross\_val\_score,GridSearchCV

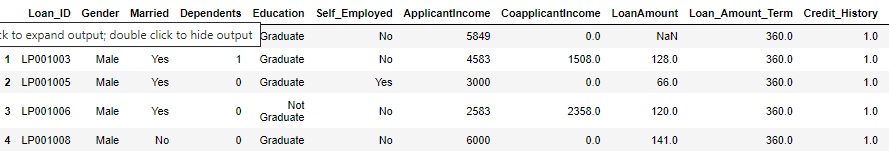
**Importing the csv File**

The dataset will be in the format of a CSV. Thus to read it, we will be taking the help of the pandas method, called read\_csv()

df=pd.read\_csv(r'C:\ProgramData\loan\_prediction.csv')

Now, head() method will print first 5 rows of the dataset

df.head() will look like



For checking the number of rows and columns in the dataset,we use

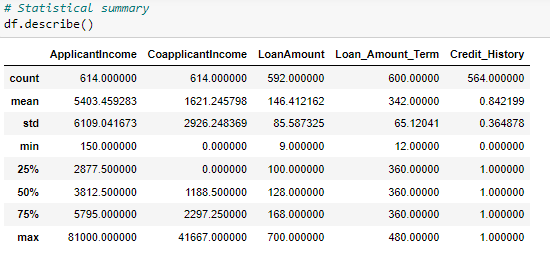
**df.shape**

head1.png

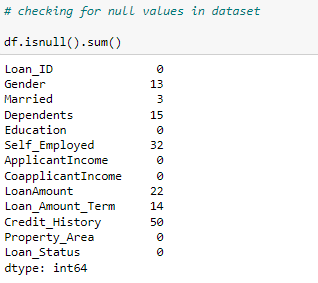
which returns that dataset has 614 rows and 13 columns

After loading dataset ,we will check the description of the dataset

df.describe()



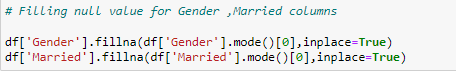
After checking for mean,std ,min ,max of dataset.We will check for null values.We can find many null values in dataset



**DATA PREPROCESSING**

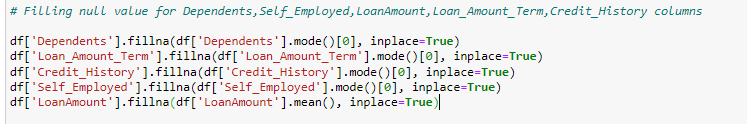
Now we will fill null values using various methods

Filling null value for Gender ,Married columns using mode



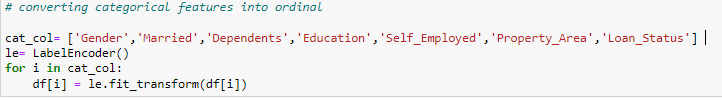
Filling null value for Dependents,Self\_Employed, Loan\_Amount\_Term,

Credit\_History columns with mode and LoanAmount with mean



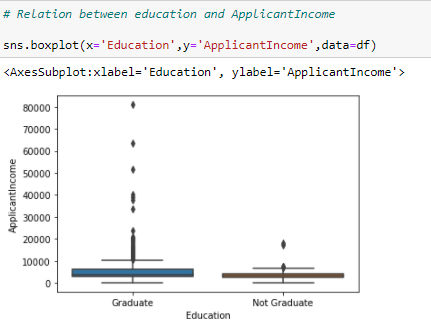
**LABEL ENCODING**

We will now convert all categorical features into numeric



**DATA VISUALIZATION**

Showing relation between ‘education’ and ‘applicant income’



# Clearly Graduate have better income than non-graduate and has outliers also.will check further that also

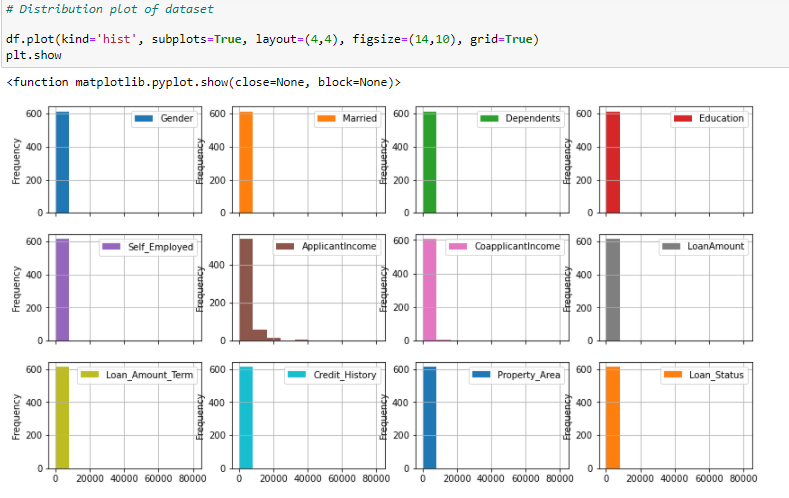
**CORRELATION**

Now,we will check correlation of independent variables with 'Loan\_Status ' variable

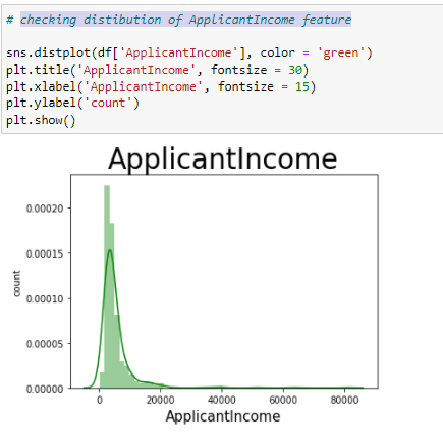


# Clearly from above correlation matrix and data , 'Loan\_Status ' has positive correlation maximum with Credit\_History and negatively correlated with Education ,CoapplicantIncome

**Distribution Plot of Features**

****

Now we will check distibution of ‘ApplicantIncome’ feature

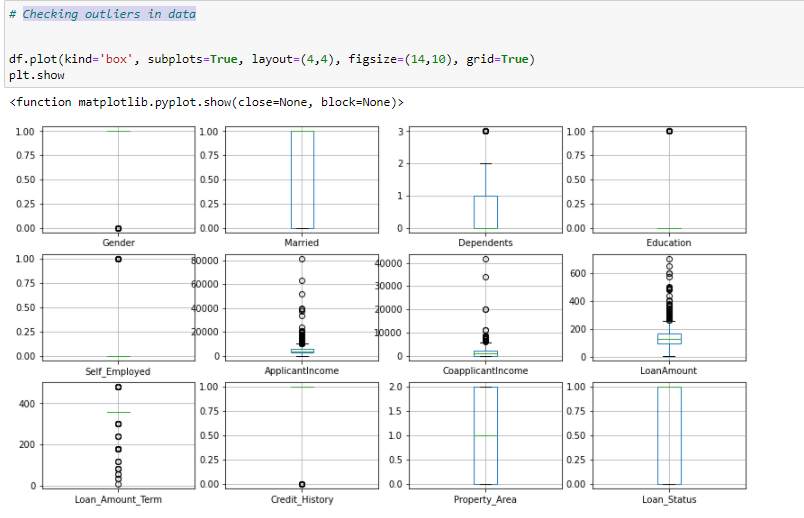


**OUTLIER DETECTION**

The outlier is a data point that diverges too much from the rest of the dataset. The majority of the real-world datasets have an outlier. Outlier detection plays a significant role in the data mining field. Outlier detection is valuable in numerous fields like network interruption identification, credit or debit card fraud detection, detecting outlying in wireless sensor network data, etc.

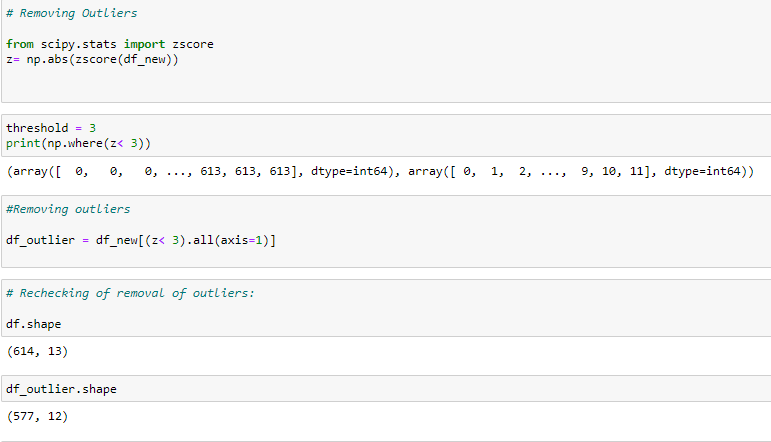
**Now we will be Checking outliers in data**

Visualizing Outliers through Distribution plot



# Many outliers are present in different features in dataset

Now we will remove outliers from dataset using Z-score method



Splitting Data :

Splitting the data into the X and Y variables :

CR_1.png

X variable will store all the features on which the loan status depends, excluding the loan status itself. Y variable will store only the Loan Status.

Now we will check for skewness in dataset

X.skew().sort\_values()

Loan\_Amount\_Term -2.098806

Credit\_History -1.976043

Gender -1.622920

Married -0.630211

Property\_Area -0.055332

Dependents 1.052106

LoanAmount 1.113132

Education 1.306588

CoapplicantIncome 1.350517

ApplicantIncome 2.148522

Self\_Employed 2.252848

dtype: float64

We will now remove skewness from different features having skewness using

Power transform

from sklearn.preprocessing import power\_transform

z = power\_transform(X[0:])

data\_new= pd.DataFrame(z,columns=X.columns)

X = data\_new

**Scaling the dataset**

from sklearn.preprocessing import MinMaxScaler

X=MinMaxScaler().fit\_transform(X)

#### Splitting X and Y into Training and Testing Variables

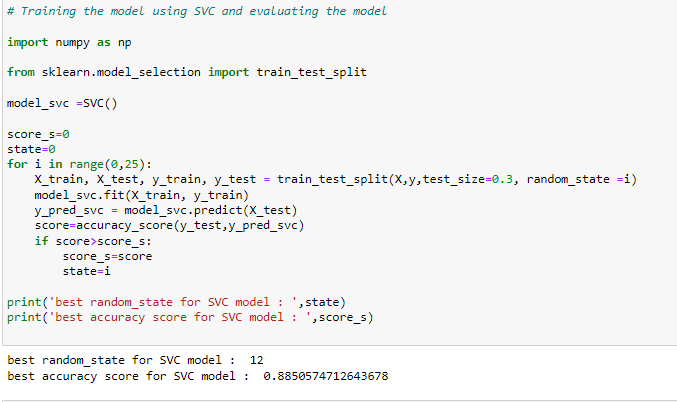
#### splitting the data into four variables,i.e X\_train, Y\_train, X\_test, Y\_test

X\_train,X\_test,Y\_train,Y\_test= train\_test\_split(X,Y,test\_size=0.25,random\_state=12)

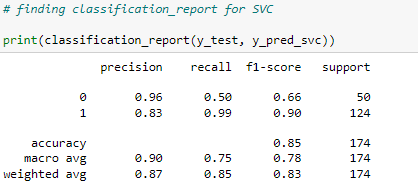
**Evaluating Models**

Now we will Train the different models and evaluate the models.

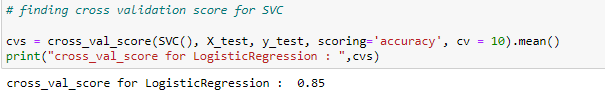
Here training the model using SVC and evaluating it



Evaluating SVC model using classification report , cross validation score



Cross Val Score for SVC



Comparing accuracy score and cross val score of different models,SVC model seems to be the best model for predicting Loan Applciation Status

Now we will minimize **overfitting** using **Ridge Technique**

What is overfitting and how do we remove overifitting through regularization

**Overfitting** happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data

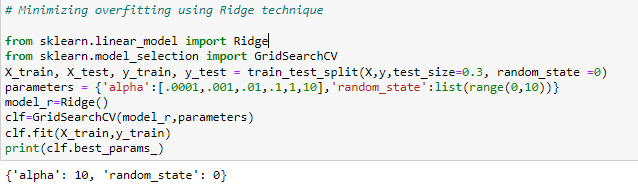
Regularization is a technique used to reduce the errors by fitting the function appropriately on the given training set and avoid overfitting.

The commonly used regularization techniques are :

1. L1 regularization(LASSO)
2. L2 regularization(RIDGE)
3. Dropout regularization

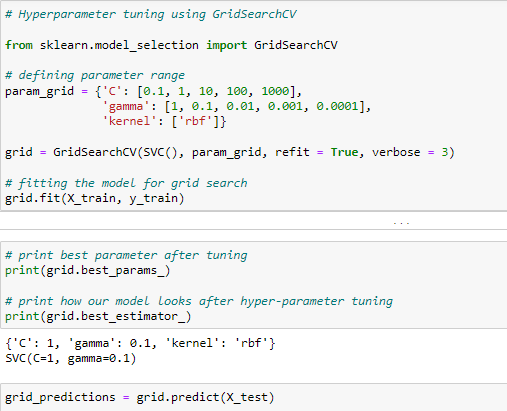
A regression model which uses **L1 Regularization**technique is called **LASSO** regression.   
A regression model that uses **L2 regularization** technique is called **Ridge regression**.

Minimizing overfitting using Ridge technique

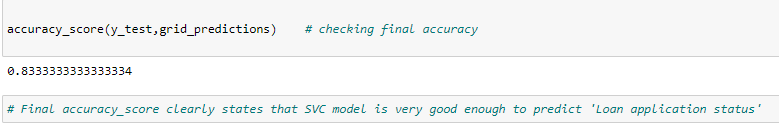


Now we will perform Hyperparameter tuning on SVC model through GridSearchCV

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**.



After finding the final accuracy score, we will plot AUC\_ROC Curve



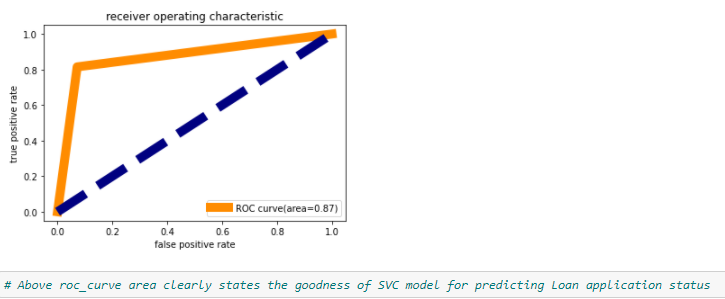
AUC\_ROC Curve

## What is the AUC-ROC curve?

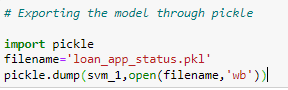
The **Receiver Operator Characteristic (ROC)** curve is an evaluation metric for binary classification problems. It is a probability curve that plots the **TPR**against **FPR**at various threshold values and essentially **separates the ‘signal’ from the ‘noise’**. The **Area Under the Curve (AUC)**is 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.





Now we will export the model through pickle



**Conclusion**

Clearly , Final accuracy\_score and other metrics ,aur\_roc curver clearly states that SVC model is very good enough to predict 'Loan application status'.