LOAN PREDICTION

This project is basically for the prediction of loan, that is whether for individual loan will be passed or not. Supervised machine learning model is prepared for prediction of passing of loan in this project.

In this dataset many features and a label is given for making the prediction.

For building the model Logistic Regression algorithm is used

Starting with importing the necessary library that is are used and other are import when requirement is needed. The required libraries are

import pandas as pd

import numpy as np

from sklearn.preprocessing import StandardScaler

from sklearn.linear\_model import LogisticRegression

from sklearn.model\_selection import train\_test\_split

import matplotlib.pyplot as plt

import seaborn as sns

import pickle

import warnings

warnings.filterwarnings('ignore')

For using pandas pandas has been imported as pd

Numpy has been imported ,standard scaler has been imported from sklearn

Since we are using logistic Regression for our project so logistic Regression has been imported from sklearn

Train test split has been imported

Matplotlib is used for visualisation process so it has been imported

Many graphs are drawn from using seaborn library.so for plotting the graph seaborn has been imported

Pickle is used to save the complete model so it has been imported

Many times unnecessary warning are seen while building the model, so warnings has been imported

DATA ANALYSIS

After importing the necessary libraries, first thing is to read the data from the given location through read function.here we find that there are 13 columns, these represent 12 features and 1 label,

Features are

Loan\_ID

Gender

Married

Dependents

Education

Self\_Employed

ApplicantIncome

CoapplicantIncome

LoanAmount

Loan\_Amount\_Term

Credit\_History

Property\_Area

And there is one label that is Loan\_status

Checking the shape of dataset given through shape function. we find there are 614 row and 13 columns here columns denotes 12 features and 1 label

Now Nan values present in the data are found through isna() function, summing all the nan values by applying sum function. we can clearly see there are many nan values present in the data

Loan\_ID 0

Gender 13

Married 3

Dependents 15

Education 0

Self\_Employed 32

ApplicantIncome 0

CoapplicantIncome 0

LoanAmount 22

Loan\_Amount\_Term 14

Credit\_History 50

Property\_Area 0

Loan\_Status 0

These nan values are removed by fillna function. these nan values are filled by mean values for the given columns

Dropping some features Loan id and gender that are not much corelated from label

Encoding

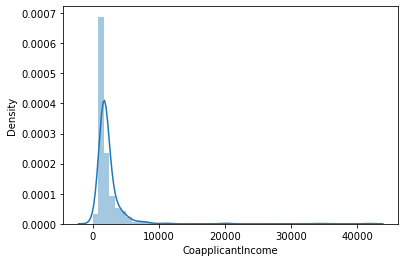
There are some features which are in object form these are converted into integer form because our machine learning algorithm can understand numeric format, For this we are applying get dummies method it is similar to one hot encoding.

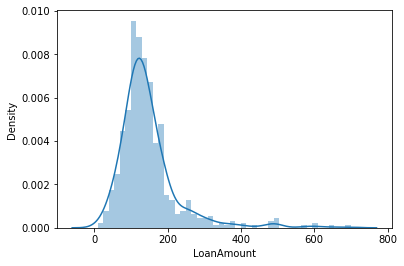
EDA

**Describe function is used for description of continuous data. after describe we can see there is much difference in std and mean, and data variation between quantile range is also large, so we can predict that there are outliers and data is not normalised(bell-shape)**

For analysing the data Distplot has been used.

Plotting the distplot for continuous features that are Applicant income, Co-applicant Income and Loan amount





From distplot it is observed that though data is normalised but there is some skewness in the features

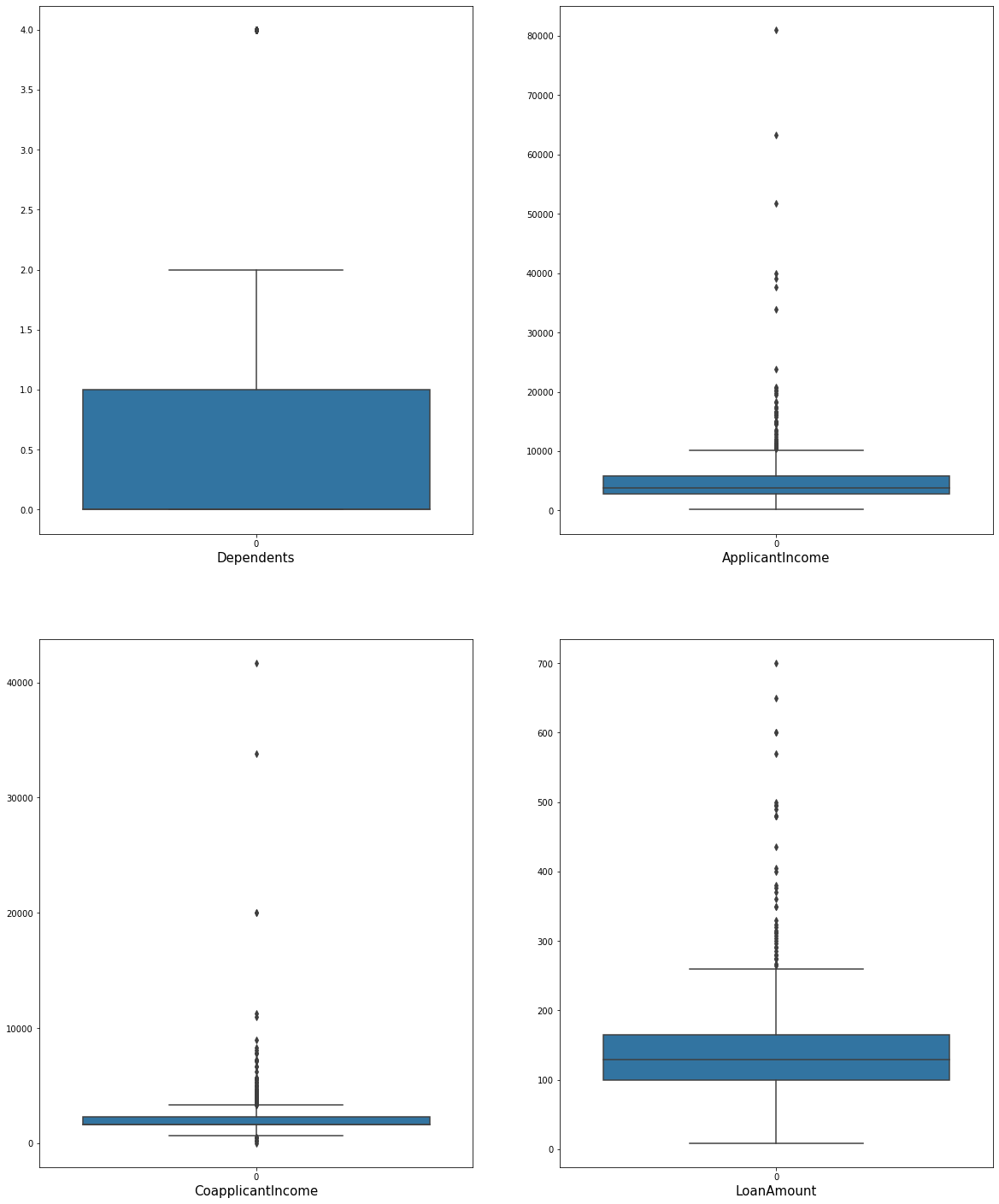
From graph one more thing can be analysed that features are showing values zero or less than zero.so replacing these zero values from mean of the features

**BOXPLOT**

Next Boxplot are drawn to check the outliers

From the graph given below we can analysis that there are many outliers present in continuous features

For removing the outliers IQR PROXIMITY RULE has been followed, interquantile range and upper side and lower side of box plot has been found. All the values above the upper quantile has been droped because they are outliers and some values are below lower quantile are also removed.



**Standard Scaler transformation**

Applying the **Standard Scaler transformation** to convert the values of the features to similar and smaller scale so that it may easy for machine learning algorithm.

**variance inflation factor**

Now applying **variance inflation factor** to find the multicollinearity in the features. From the result it can be seen that all the vif values are less than 5 so there is no multicollinearity present in the features. (considering 5 as standard value for vif the vif values less than 5 have no multicollinearity and vif values above 5 have multicollinearity)

Next splitting the features and label in X and y dataframe. Sending the data for train test split. Test data is about 25 % and train data is 75 % of whole data.

**Logistic regression**

Now our data is ready to pass through the model, applying **logistic regression** because label have categorical data and there is only two category in label either yes or no, applying the logistic regression and fitting the train data for transformation.

Predicting the y values through x-test data that is features data which was kept for testing.

Finding the accuracy by accuracy score comparing y\_pred and y\_test values accuracy score is nearly 87% .

After this importing confusion matrix, roc\_auc curve from sklearn.matrics Applying confusion matrix, it is showing result

array([[22, 14],

[ 1, 82]], dtype=int64)

We can see from the confusion matrix that type 1 error have one one value and type 2 error have 14 values

Producing the classification report, this value is showing accuracy of 87%

Drawing the roc\_auv curve the above result

Saving the model using pickle library.