LOAN APPLICATION STATUS PREDICTION

With the world moving in its faster pace, the unexpected emergencies and needs arises and that can stress us much. To overcome this situation and to solve the needs which require the adequate financial aid getting loan from a bank. In our banking system, banks have many products to sell but main source of income of any banks is on its credit line. So, they can earn from interest of those loans which they credit. A bank’s profit or a loss depends to a large extent on loans i.e., whether the customers are paying back the loan or defaulting. By predicting the loan defaulters, the bank can reduce its Non-performing Assets. This makes the study of this phenomenon very important. Previous research in this era has shown that there are so many methods to study the problem of controlling loan default. But as the right predictions are very important for the maximization of profits, it is essential to study the nature of the different methods and their comparison. A very important approach in predictive analytics is used to study the problem of predicting loan defaulters (i) Collection of Data, (ii) Data Cleaning and (iii) Performance Evaluation.

INTRODUCTION

Loan Prediction is very helpful for employee of banks as well as for the applicant also. The aim of this report is to provide quick, immediate and easy way to choose the deserving applicants. Customer first apply for loan after that company or bank validates the customer eligibility for loan. Company or bank wants to automate the loan eligibility process (real time) based on customer details provided while filling application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and other. This project has taken the data of previous customers of various banks to whom on a set of parameters loan were approved. So, the machine learning model is trained on that record to get accurate results. Our main objective of this project is to predict the safety of loan.

PROBLEM DEFINITION:

This dataset includes details of applicants who have applied for loan. The dataset includes details like credit history, loan amount, their income, dependents etc. Dependent Variable (Target Variable) - Loan\_Status. 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.

MOTIVATION:

Loan approval is a very important process for banking organizations. The system approved or reject the loan applications. Recovery of loans is a major contributing parameter in the financial statements of a bank. It is very difficult to predict the possibility of payment of loan by the customer. Using Machine learning we predict the loan approval.

PROPOSED MODEL

This system predict whether the loan is approve or reject . This System refers the following things.

* Data Collection
* Data Pre-processing (Data Cleaning)
* Model Selection
* Model Evaluation
* Classification
* Result (output)

IMPLEMENTATION:

Dataset :

Loan Dataset is very useful in our system for prediction of more accurate result. Using the loan Dataset the system will automatically predict which costumer’s loan it should approve and which to reject. System will accept loan application form as an input. Justified format of application form should be given as an input to get processed.

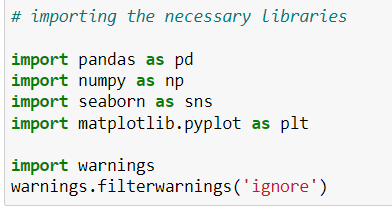
Determine the training and testing data: Typically, Here the system separate a dataset into a training set and testing set ,most of the data use for training ,and a smaller portions of data is use for testing. after a system has been processed by using the training set, it makes the prediction against the test set.

Data cleaning and processing: In Data Cleaning the system detect and correct corrupt or inaccurate records from database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing , modifying or detecting the dirty or coarse data. In Data processing the system convert data from a given form to a much more usable and desired form i.e. make it more meaningful and informative.

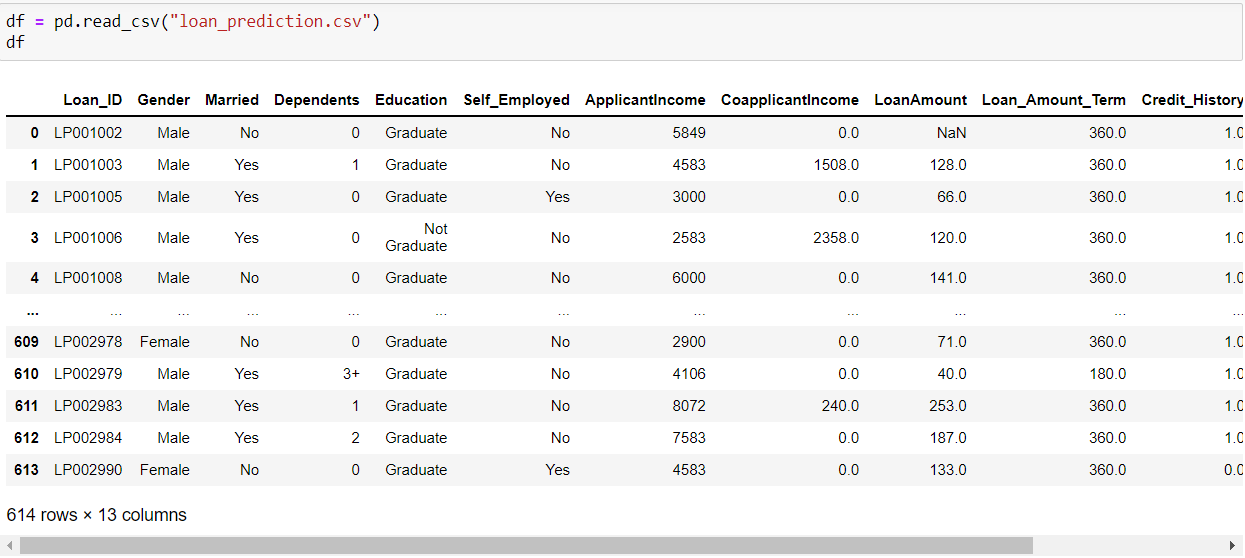
DATA ANALYSIS:

The process of cleaning, transforming and extracting data to discover the useful information for business decision making is called data analysis.

**Importing necessary libraries**



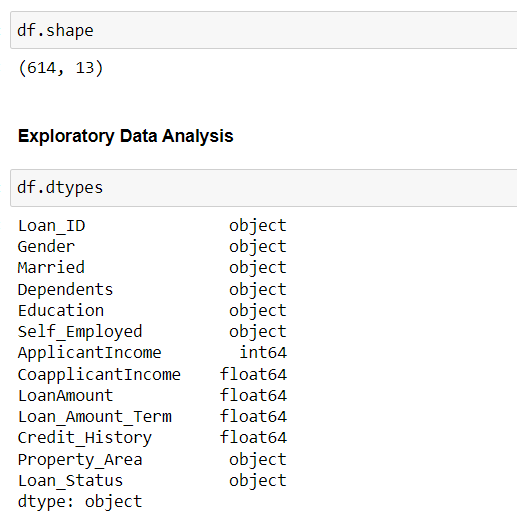
**Importing Dataset:**



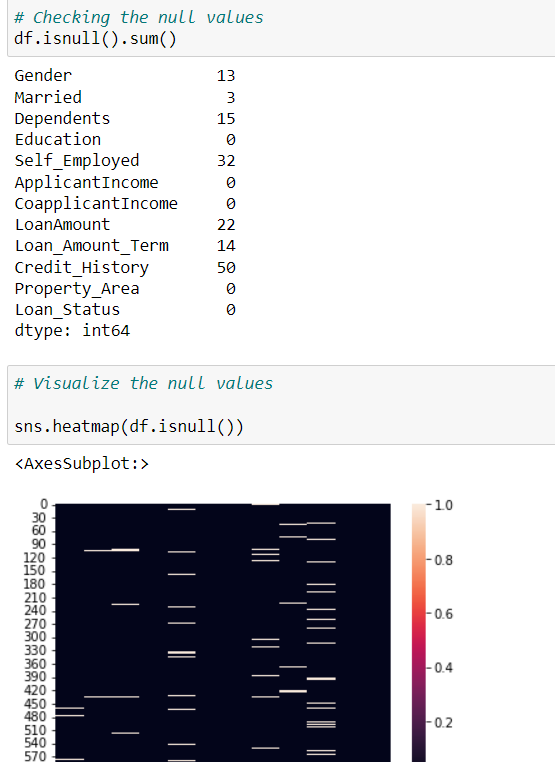
In this problem the target variable is ‘Loan\_Status’. With the help of that we need to predit whether to provide loan or not.

EXPLORATORY DATA ANALYSIS:

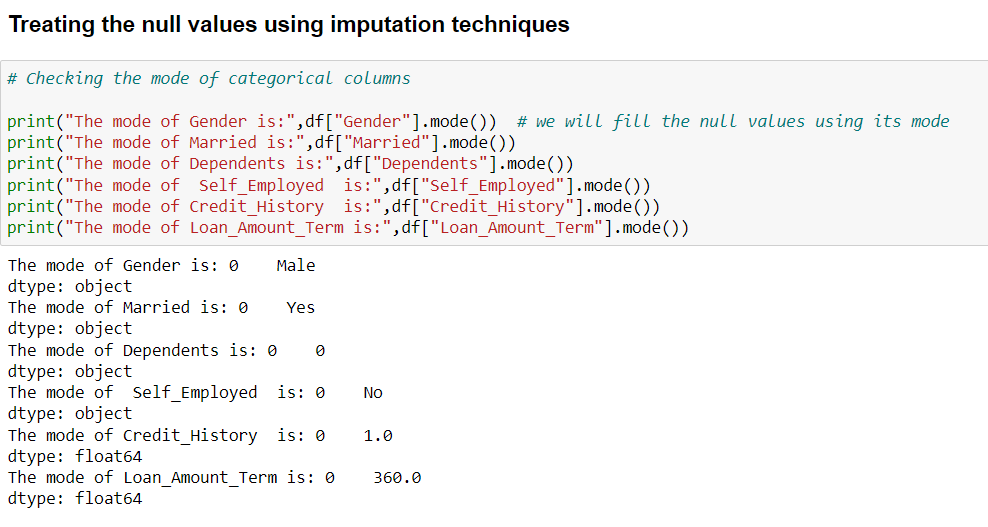
In this we need to analyse the data. Using df.shape gives you the number of rows and columns in the dataset. Here we have 614 rows with 13 columns. And followed by finding the type of data each column is. Here we have object, int and float data type.



CHECKING THE NULL VALUES:  
We have null values present and visualizing it using heatmap.

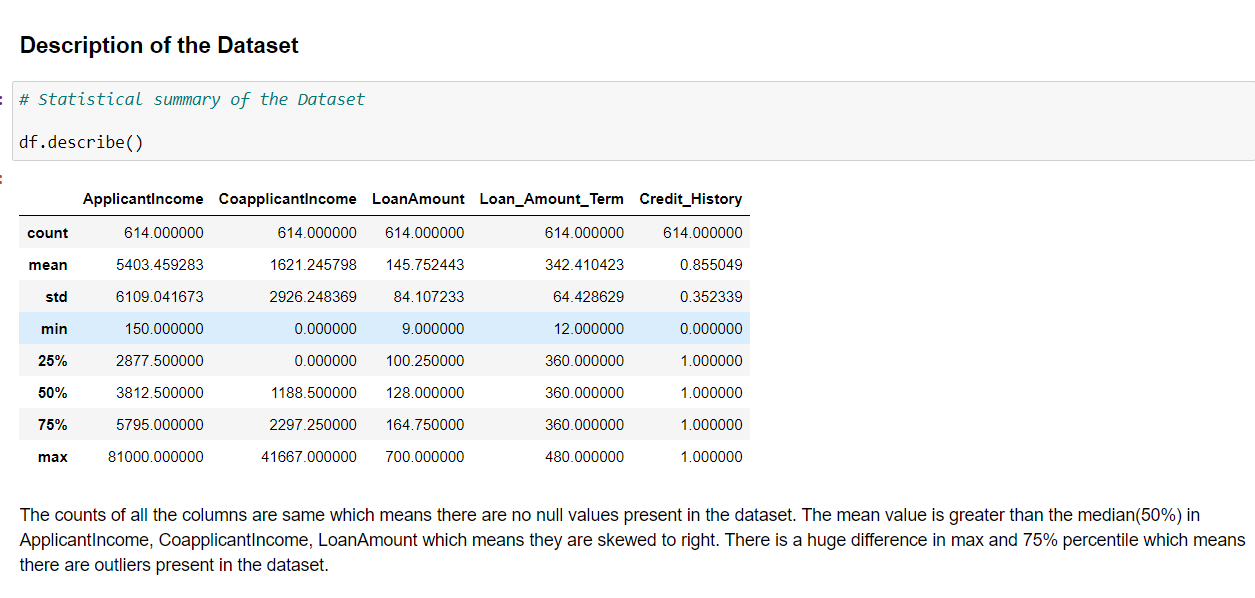


TREATING THE NULL VALUES:



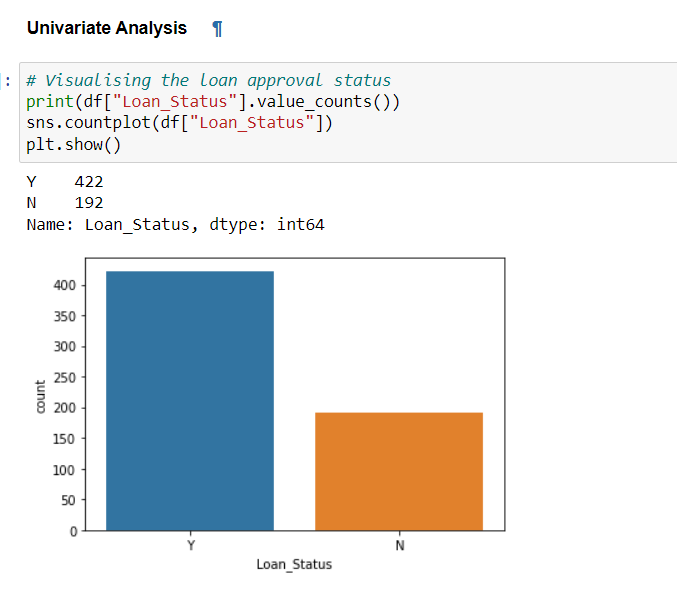
Now the null values are treated using imputation techniques. For the object data type mode is used. And for int and float type mean is used.

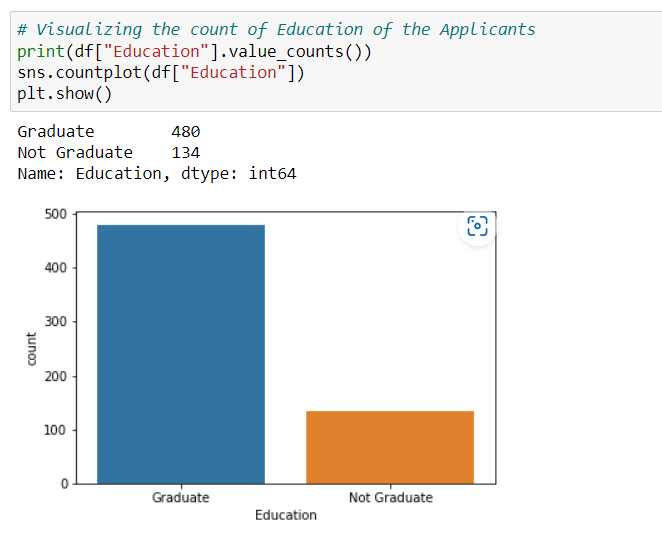
DATA DESCRIPTION:



Df.describe gives the statistical data of the dataset.

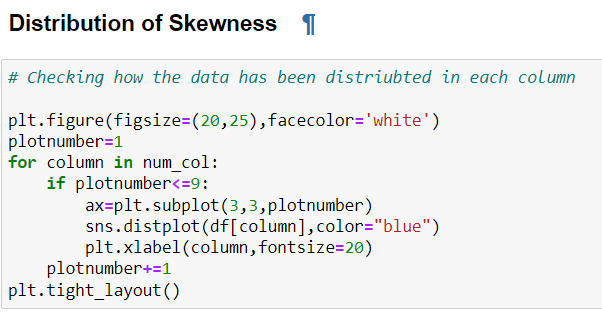
DATA VISUALIZATION:

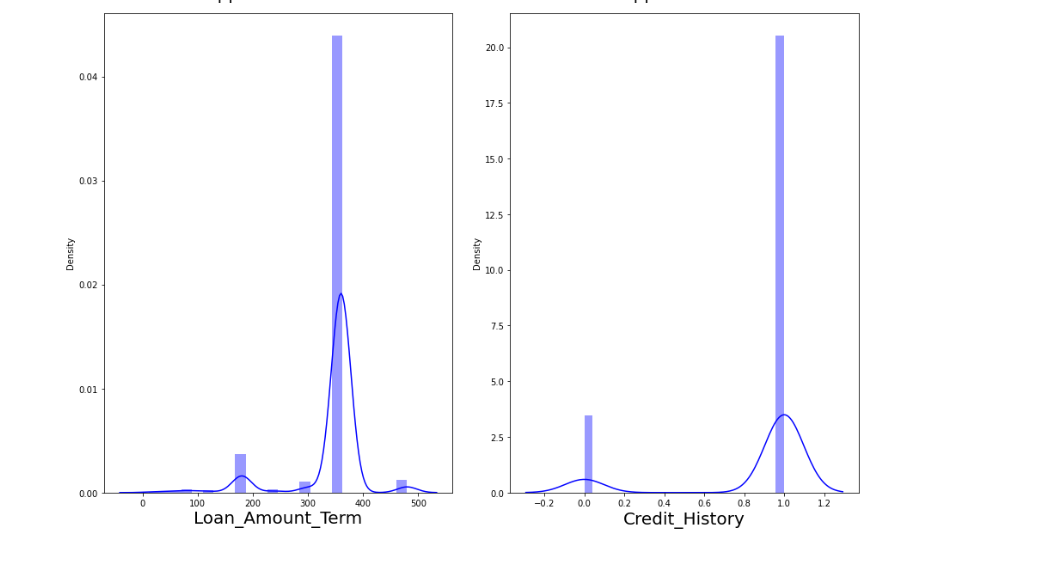
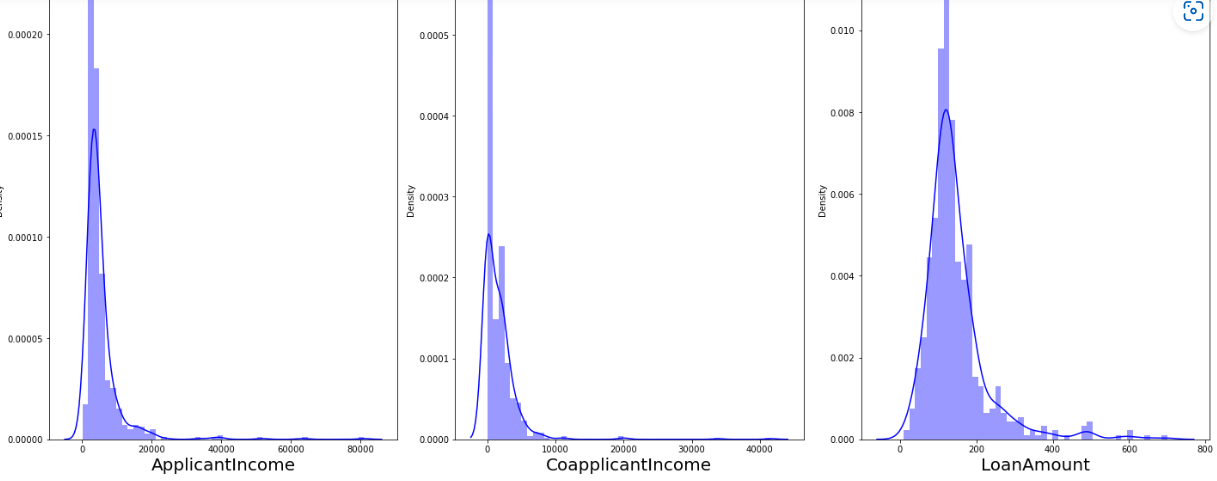




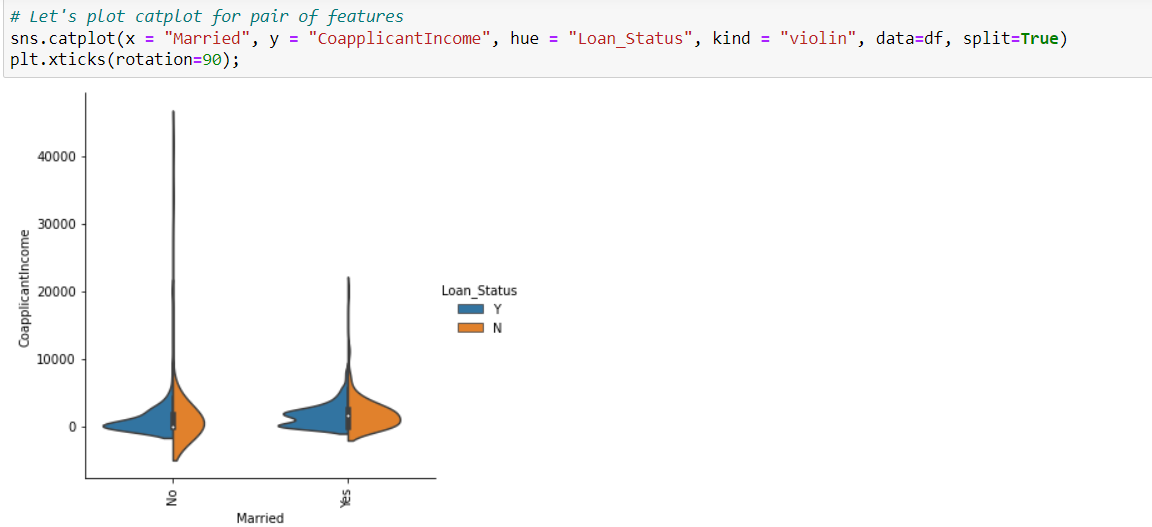
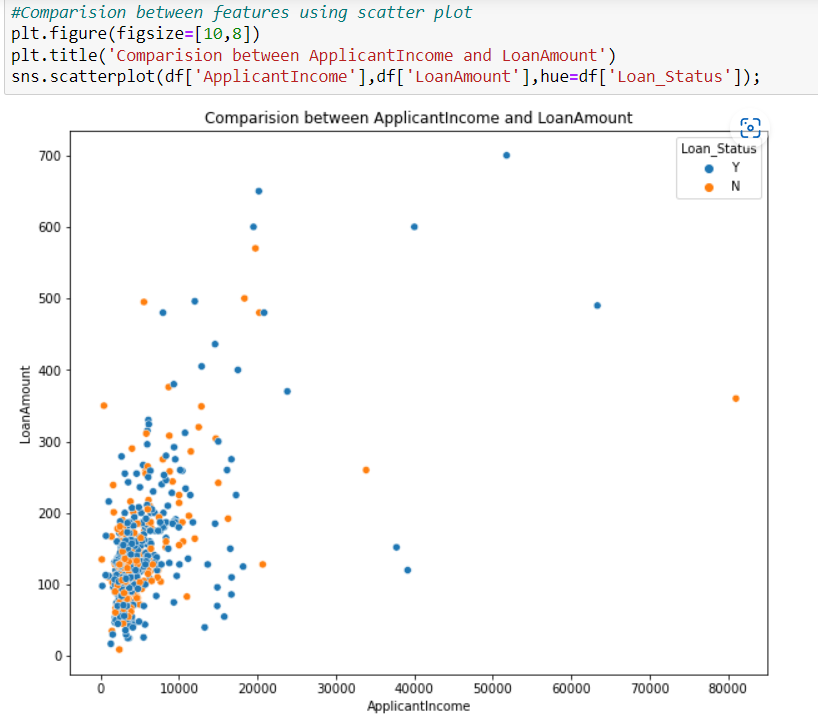


CHECKING FOR SKEWNESS:

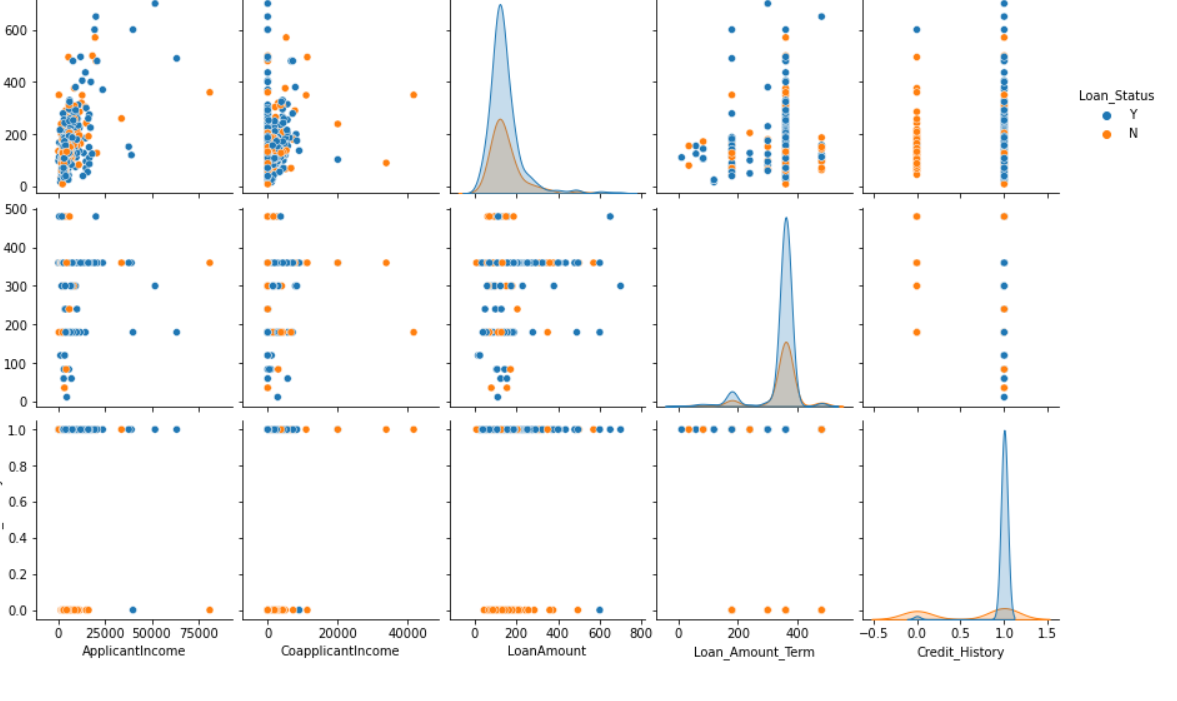
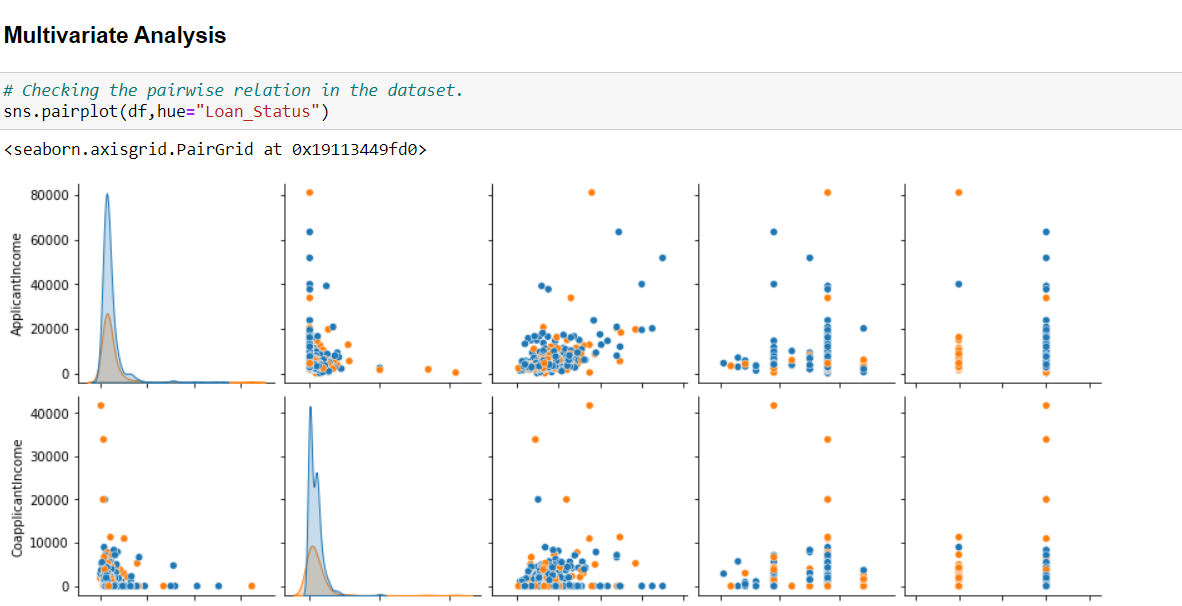




BIVARIATE ANALYSIS:

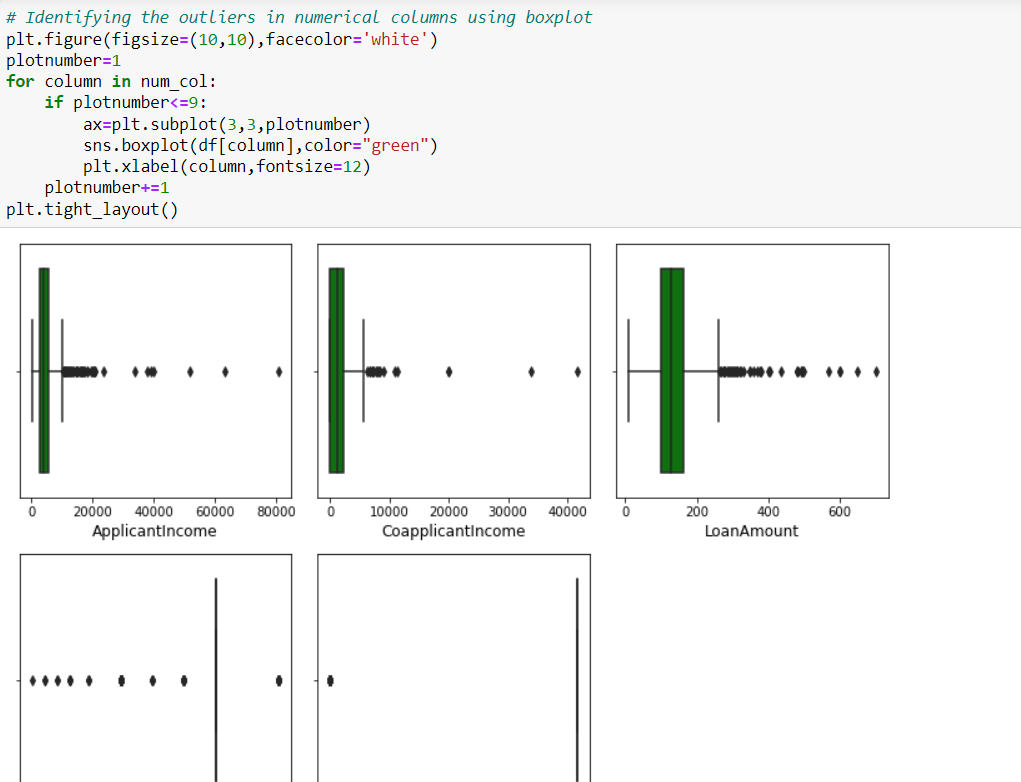


MULTIVARIATE ANALYSIS:

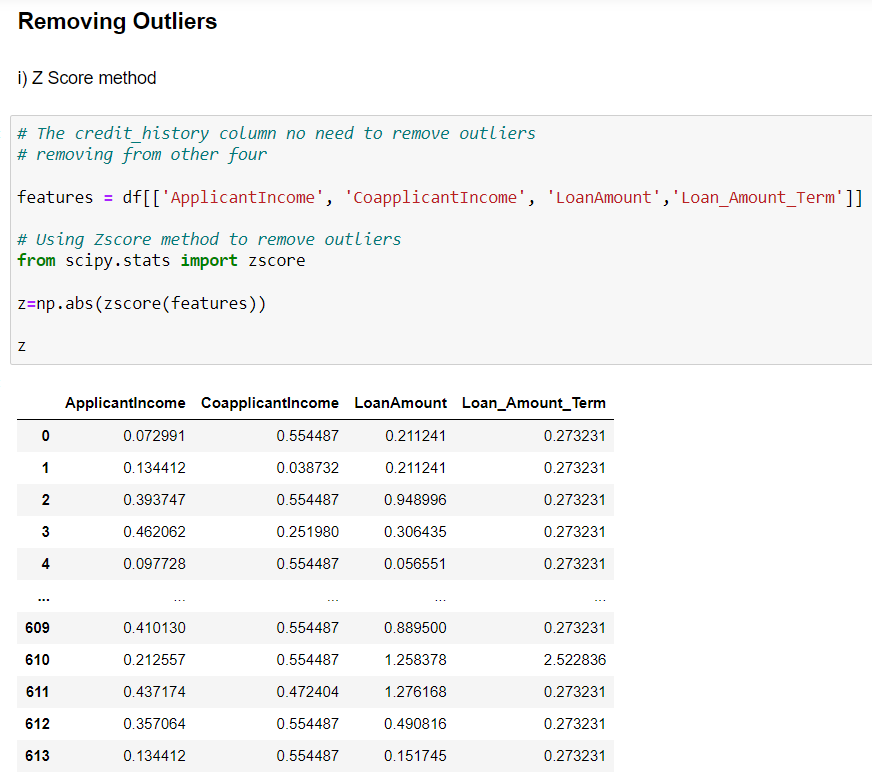


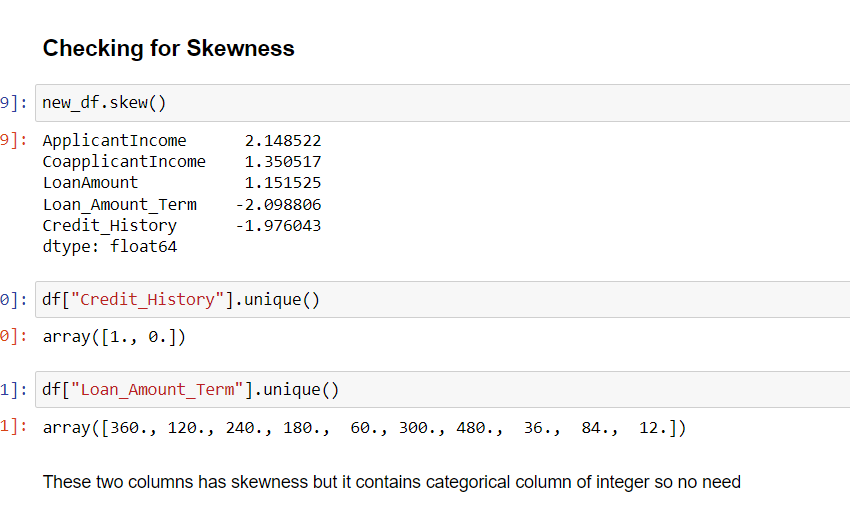
Here with the target variable we are comparing the other variables.

OUTLIERS:

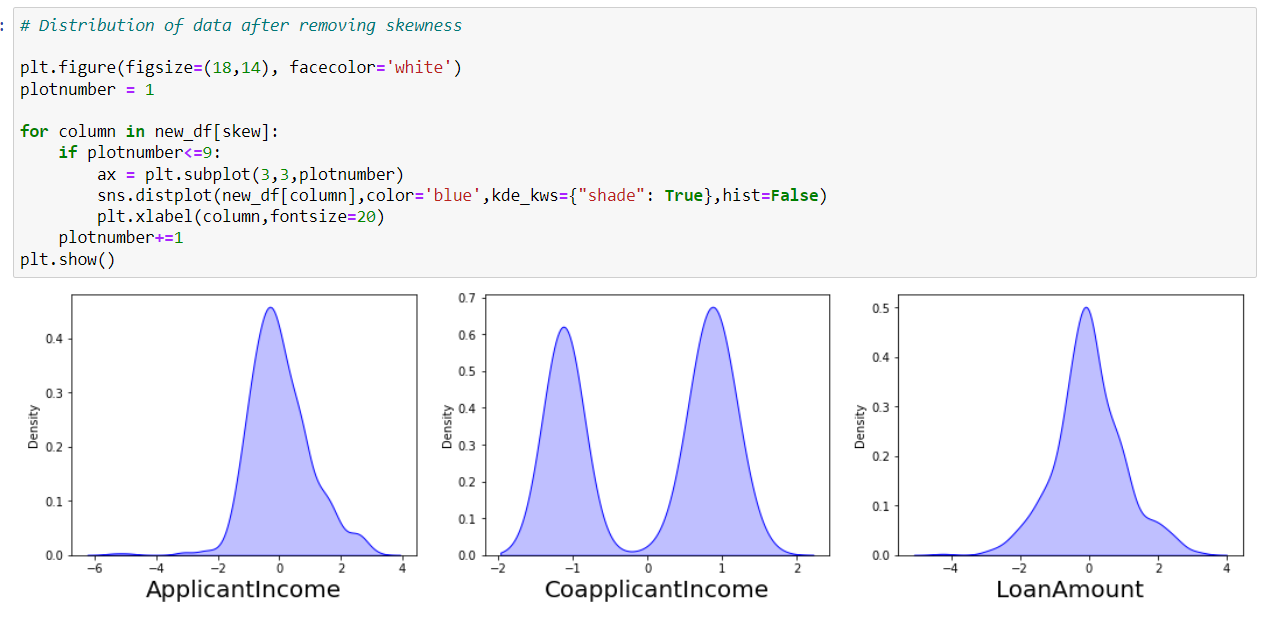


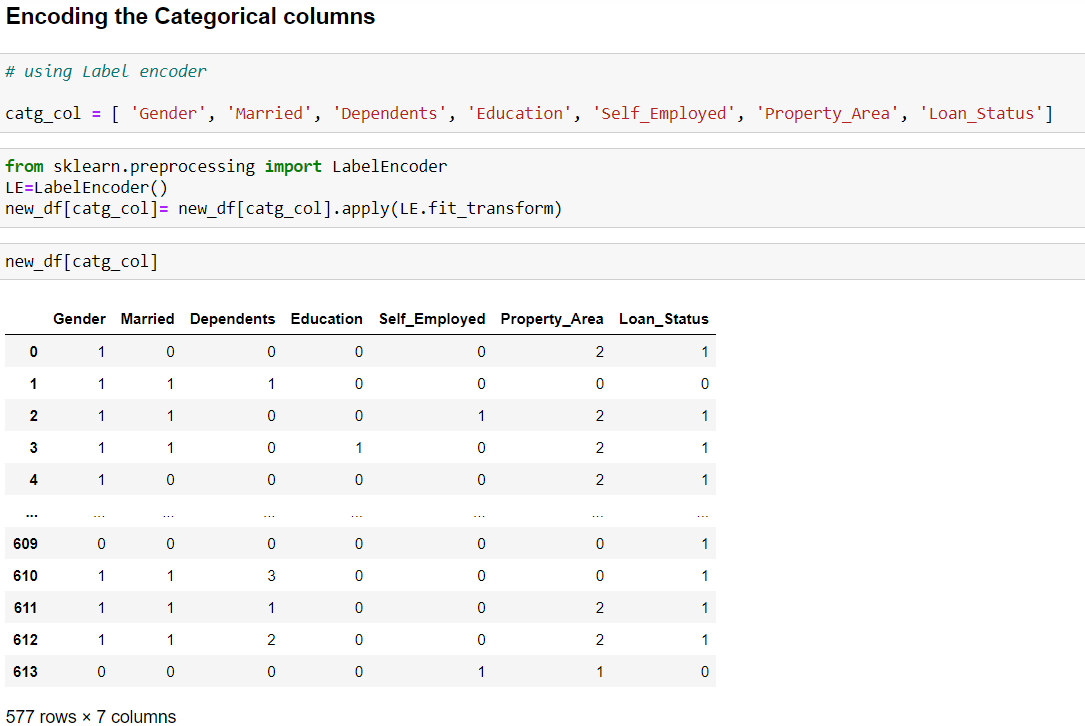
From the above plot we came to know that there are some outliers present and it can be treated in the following ways.





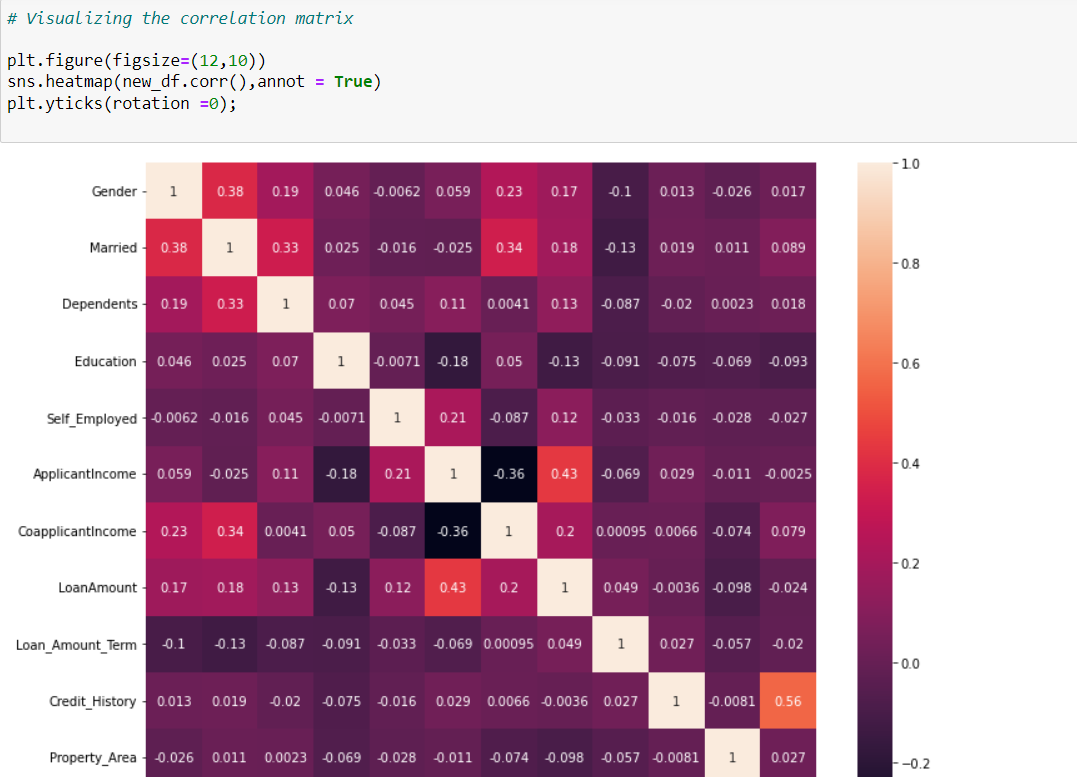
There is skewness present in the columns so that can be removed by yeo-johnson method.



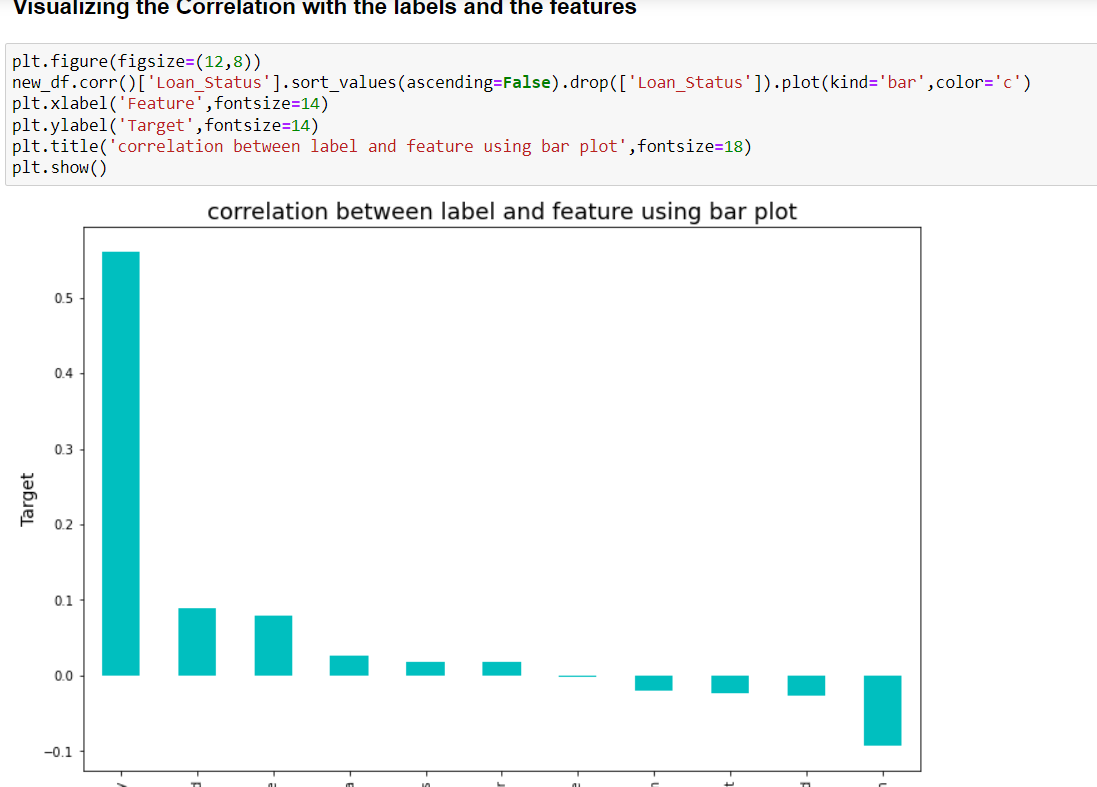


A process of converting categorical data into integer format so that the data with converted categorical values can be provided to the different models.

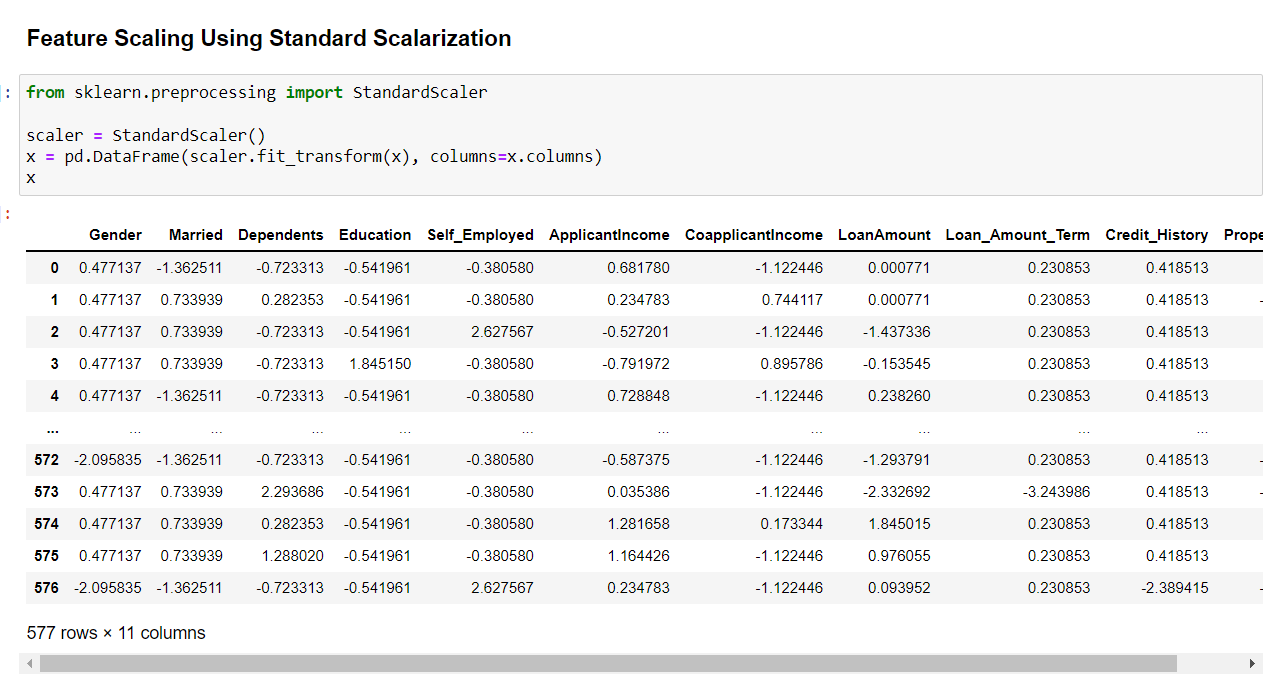
Now finding the correlation between the target variable and the independent variable using heat map.



VISUALIZING THE CORRELATION:



FEATURE SCALING:

**scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.**

**BUILDING MACHINE LEARNING MODELS:**

Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed.

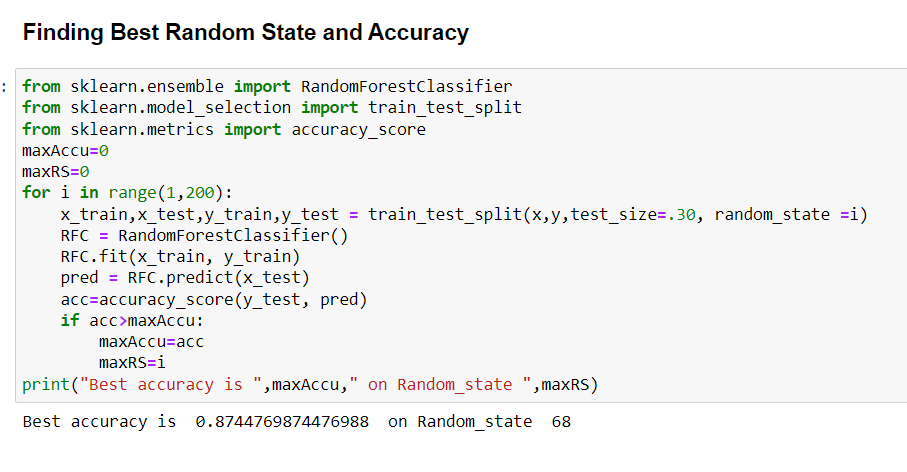
-Arthur Samuel

The main goal in this step is to develop a benchmark model serves us as a baseline, upon which we will measure the performance of a better and more tuned algorithm. We are using different Classification Technique and comparing them to see which algorithm is giving better performance.

All the required machine learning algorithms are imported from sklearn library.

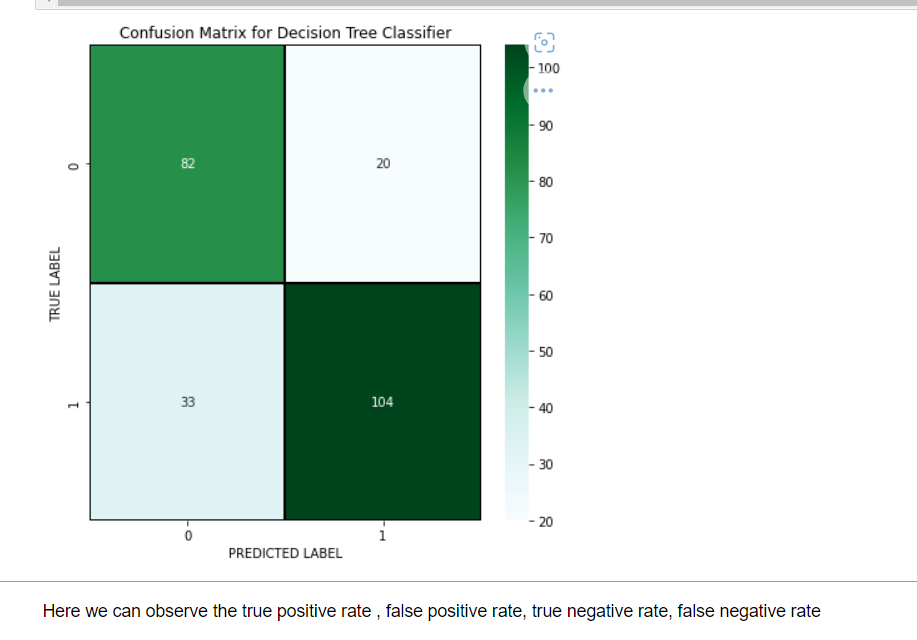
In this project I have used 7 different algorithms. The model which gives the best performance amongst them, we will be using that as best model for prediction. Before creating the model, let us check the best random state and accuracy using any of the DecisionTreeClassifier algorithm.

**FINDING RANDOM STATE AND ACCURACY:**



**The accuracy is about 87% on the random state 68 is found.**

**CLASSIFICATION ALGORITHMS:**



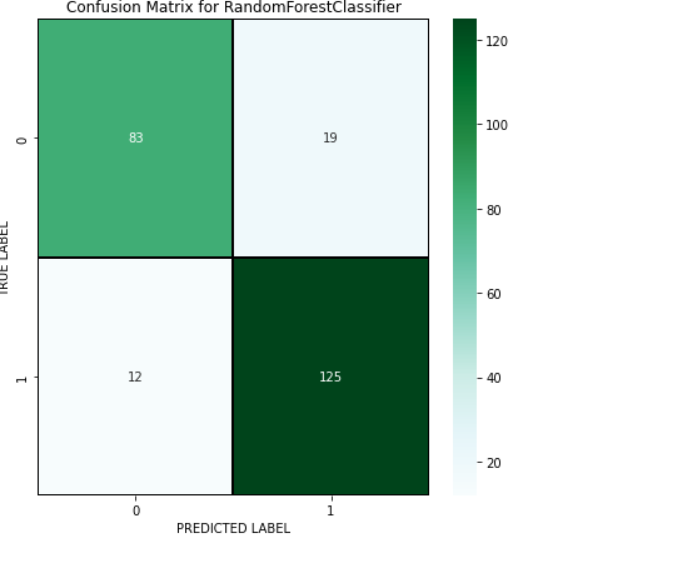
**DecisionTreeClassifier**

Decision Tree are versatile Machine learning algorithms that can perform both classification and regression tasks. They are powerful algorithms, capable of fitting complex datasets.

Decision Tree Classifier giving 81.87% accuracy and also we can observe the confusion matrix also known as an error matrix, is a specific table layout that allows visualization of the performance of an algorithm. This gives the values like True Positive, False Positive, True Negative and False Negatives.

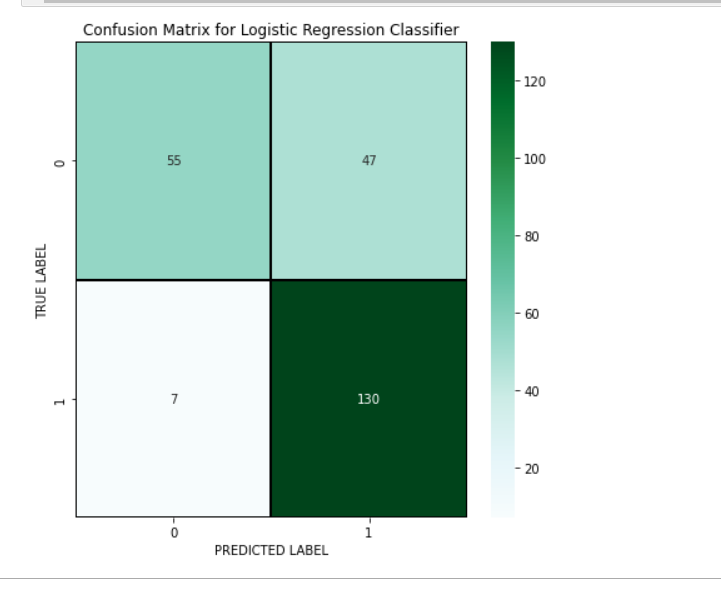
**Random Forest Classifier**

Random Forest is an ensemble of Decision Tree and capable of performing both classification and regression task using multiple decision tree and a statical technique called “Bagging”.



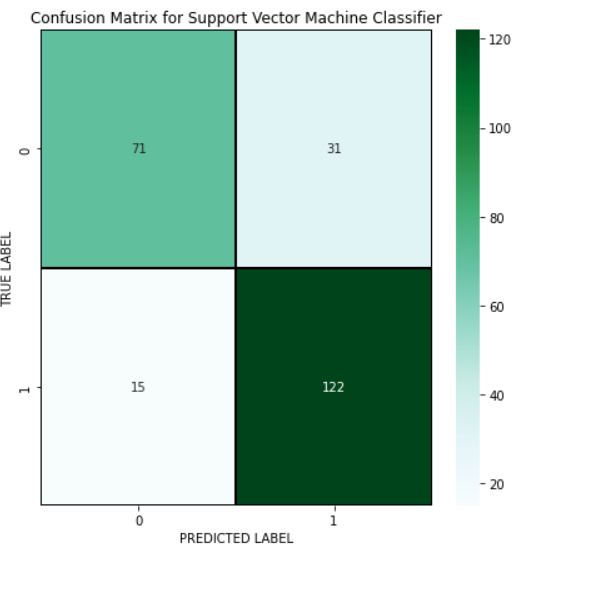
**Logistic Regression**

Logistic regression commonly used to estimate the probability that an instance belongs to a particular class.



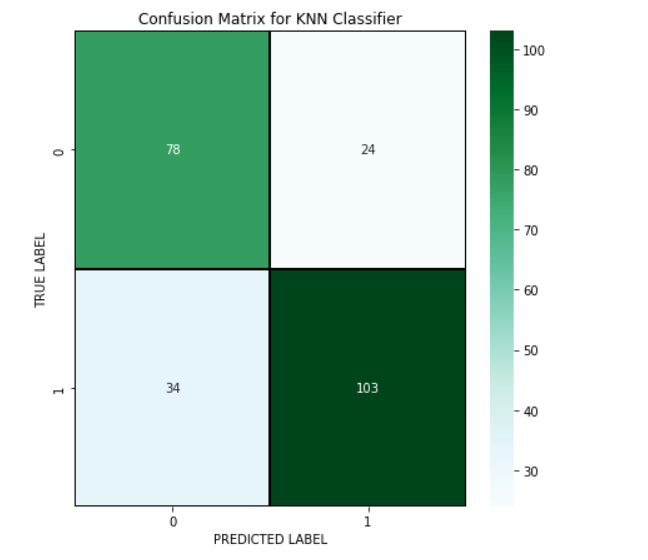
**Support Vector Machine Classifier**

Support vector machine learning is capable of performing linear or non linear classification, regression and even outlier detection. It is one of the most popular models in machine learning and are particularly well suited for classification of complex small or medium sized datasets.



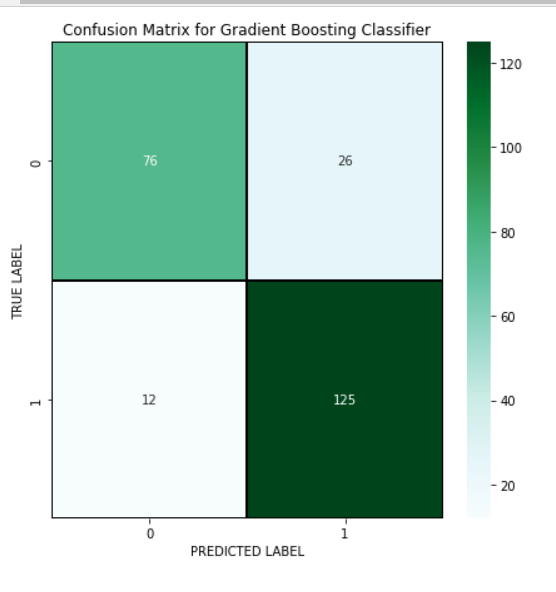
**KNeighbours Classifier**

K Nearest Neighbor algorithm falls under supervised learning and it is used most commonly in classification.

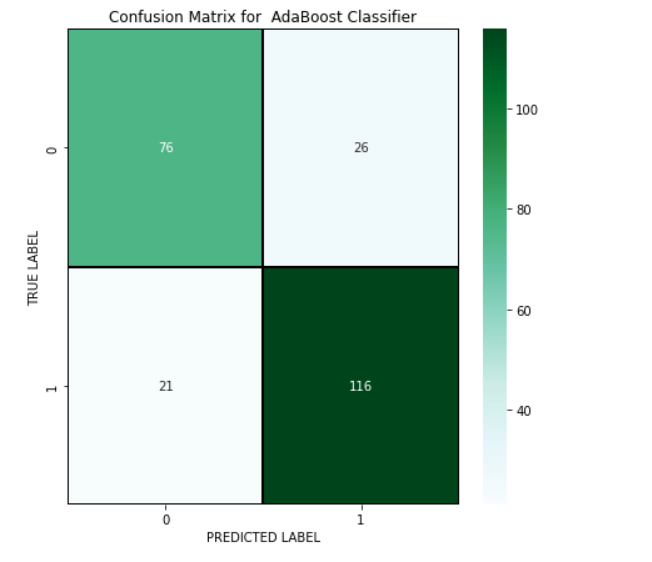


**Gradient Boosting Classifier**

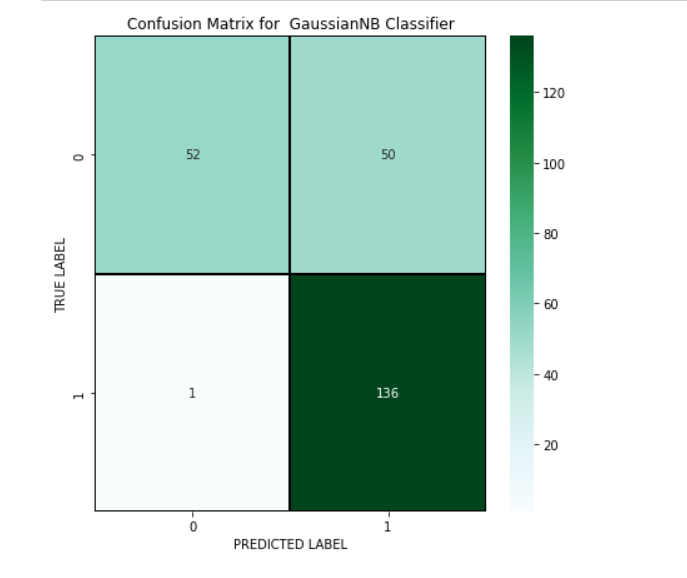
Gradient boosting works by sequentially adding predictors to an ensemble, each one correcting its predecessor.



**AdaBoost Classifier**

One way for a new predictor to correct its predecessor is to pay a bit more attention to the training instances that the predecessor underfitted. This results in new predictors focusing more and more on the hard cases. This is the technique used by AdaBoost.

GAUSSIAN NB CLASSIFIER:



CROSS VALIDATION SCORE:



**Difference between Accuracy and Cross validation Score**

|  |  |  |  |
| --- | --- | --- | --- |
| **Models** | **Accuracy** | **CV Score** | **Difference** |
| DecisionTreeClasssifier | 78 | 76.52 | 1.48 |
| Random Forest Classifier | 87 | 82.92 | 4.08 |
| Logistic Regression | 77 | 71.98 | 5.02 |
| SVM Classifier | 81 | 73.37 | 7.63 |
| KNN Classifier | 76 | 73.25 | 2.75 |
| Gradient Boosting Classifier | 84 | 80.16 | 3.84 |

The model Decision tree Classifier giving the very less difference compared to other models.

Since Decision Tree Classifier is giving best Accuracy and CV score difference, I choose Decision tree Classifier as best fitting model. Let’s check whether we can increase the Accuracy score by using Hyper parameter tuning.

HYPER PARAMETER TUNING:

In GridSearchCV approach, machine learning is evaluated for a range of hyperparameter values. This approach is called “GridSearchCV” , because it searches for best set of hyperparameters from a grid of hyperparameters values.

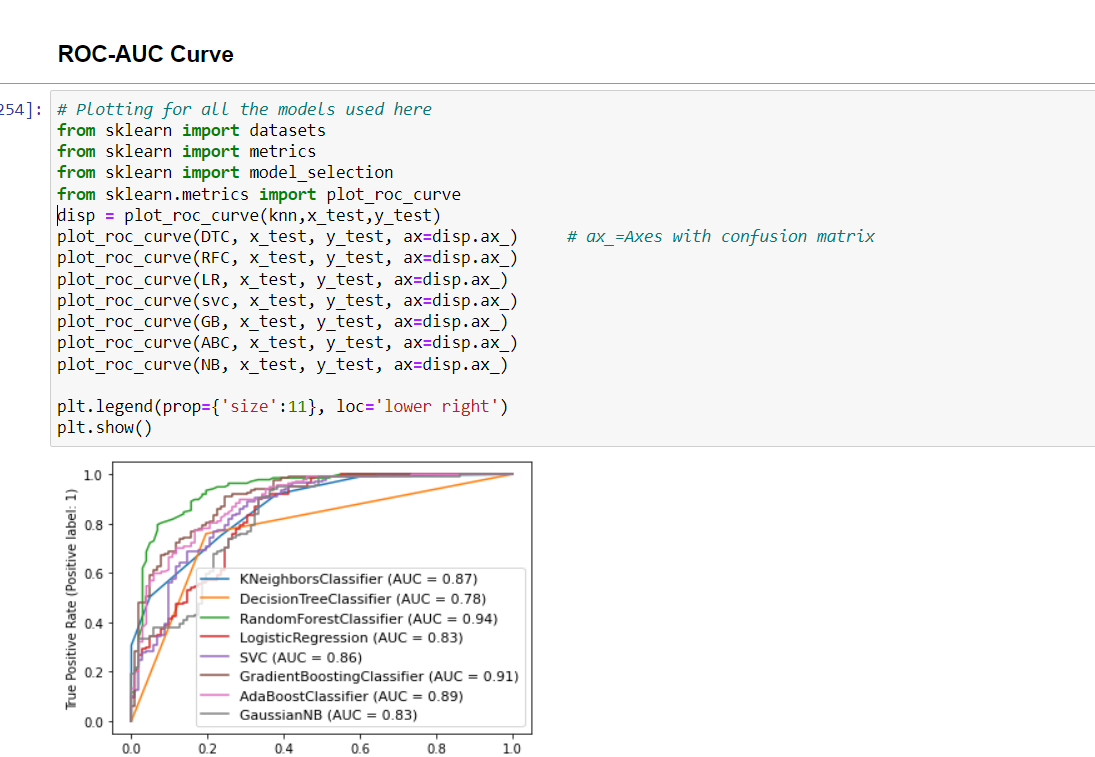


After tuning the best fitting model (Decision Tree Classifier) using GridSearchCV, the accuracy of the model increased to 79.91%**.**

**AUC-ROC Curve:**

It is a graph that shows the performance of a classification model at all possible thresholds. The curve is plotted between two parameters

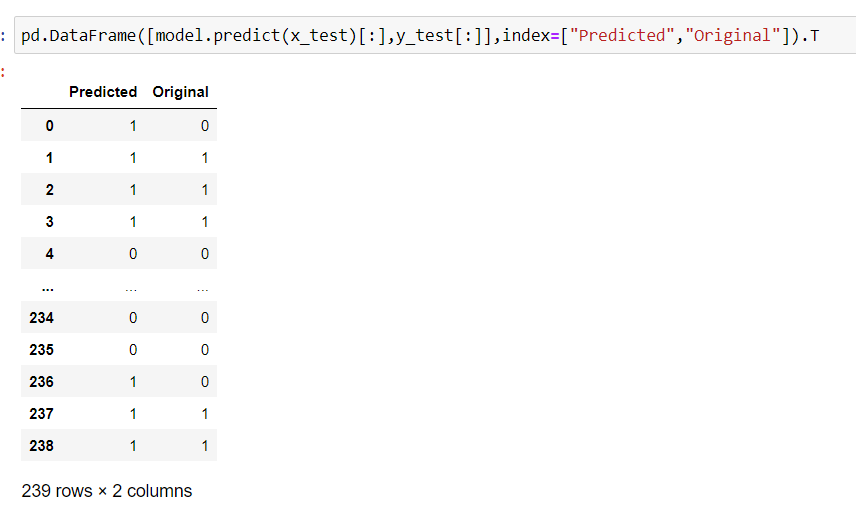
* True Positive Rate
* False Positive Rate



We have built the model and performed the hyper parameter tuning, now we will save the model and load the model to compare the actual and predicted values

SAVING THE MODEL:





We can see from the above observation, Predicted and the Original values are matching each other, which means model performance is good.

CONCLUSION:

So here, it can be concluded with confidence that the model built is efficient and gives a better result when compared to other models. It works correctly and fulfils all requirements of bankers. This system properly and accurately calculates the result. It predicts the loan is approve or reject to loan applicant or customer very accurately.

In this project we have gone through the feature engineering which is the most important thing to get the better performance models, we have removed the outliers, skewness and also handled the categorical columns by encoding the data, scaled the data, handled the data imbalance and at last, we built the different classification models to predict the attrition and perform the hyper parameter tuning to improve the model accuracy by using different parameters.