**Loan Application Status Prediction**

**Problem definition**

Along with the advancements in the banking sector getting loan is not at all tough for the customer, but it is a difficult decision for the bankers (bank) whether to approve the loan or not, As the most relevant question that strikes on the loan approver is whether the applicant will be able to repay or not or whether he is a defaulter or not

So to make a quick and accurate decision about granting the loan to the applicant, The Machine Learning in the Data Science has come with the solution of making a model that learns from the old records or some kind of information of old Loan Applicants that whether they were granted Loan or not, The Model tries to predict that the person with the same kind of Information should be granted Loan or Not. Previously it was done with variety of paper works, with the signature of guarantors, the bank tries to assure about the recovery. But there were also other factors that were not taken in the consideration, like the applicant himself/herself is dependent or not, is there a family to look after or he has to look after them ,

**About Dataset**

The Loan Applicant dataset consist of some information about the Loan Applicants who applied for the Loan and they got the loan or not depending on the information provided .

To help the company or bank to take such dynamic decision instantly we will Analyse the data, then we will do Exploratory Data Analysis and write its remark, then we will Train the model and conclude it

**Data Analysis**

The Loan Application Status Prediction is a dataset that contains 613 rows and 13 Columns the data in this table is as follows

1. **Loan\_ID** - The Loan\_ID is the Unique Loan Identification Number of every Loan Applicant , just to keep record about how many applicants applied for the Loan, It is having common prefix of ‘LP00’ and then numbers start from 1002 to 2990, there are not continuous numbers between them the numbers are randomly given in the dataset, it is showing as of Object datatype .
2. **Gender** – This columns specifies the gender of applicant is Male or Female, and in the dataset there were 489 Males and 112 Females,
3. **Married** – The column specifies the Marital status of the Applicant whether he married or not, 398 Applicants are married and 213 Applicants are unmarried
4. **Dependents-**  It states like how many dependents are there on the the Applicants, like 345 Applicant have 0 dependents on them ,102 Applicants have 1 Applicant on them, 101 Applicants have 2 dependents on them and 51 Applicant have 3+ dependent on them.
5. **Education** – It states the Qualification of the Applicant, 480 Applicants are Graduated and 134 Applicants are Not-Graduated
6. **Self-Employed** – it means that the dependent is self employed ( entrepreneur or businessman ) or whether the applicant is on job working under some other employer. 500 Applicants are Self-Employed and 82 Applicants are Not Self Employed.
7. **Applicant-Income** – it tells about the income of the Applicant ,
8. **Co–Applicant Income** – Co-Applicant Income is the income when the bank finds that the Applicant’s income is not enough, then they add co-applicant’s Income to recover their loan amount.
9. **Loan-Amount** – loan amount is the amount for which the applicant is applying for.
10. **Loan Amount Term** is the tenure for which the time the applicant wants loan , it is generally count in number of month
11. **Credit History** – Credit history is whether the person has already taken loan or not.
12. **Property Area** – Property area is the location of the property which was to be mortgaged for the loan. Like 233 people offered their SemiUrban’s property area, 202 Applicant offered their Urban’s property area and 179 Applicant’s offered Rural’s property area.
13. **Loan\_status** – It is the final status that whether the loan was approved or not. 422 Applicant’s Loan was approved and 192 Applicant’s Loan was not approved.

**Exploratory Data Analysis**

Exploratory Data Analysis(EDA) is an approach of analysing data sets to summarize their main characteristics, often with visual methods, a statistical model can be used or not , but primarily EDA is for seeing what the data can tell us beyond the formal modelling or hypothesis testing task. We can say that EDA is statisticians way of telling story where you explore data , find patterns and tell insights. EDA is a phenomenon under data analysis used for gaining a better understanding of data aspects like :- main features of data - variables and relationships that hold between them - identifying which varaibles are important for our problem.

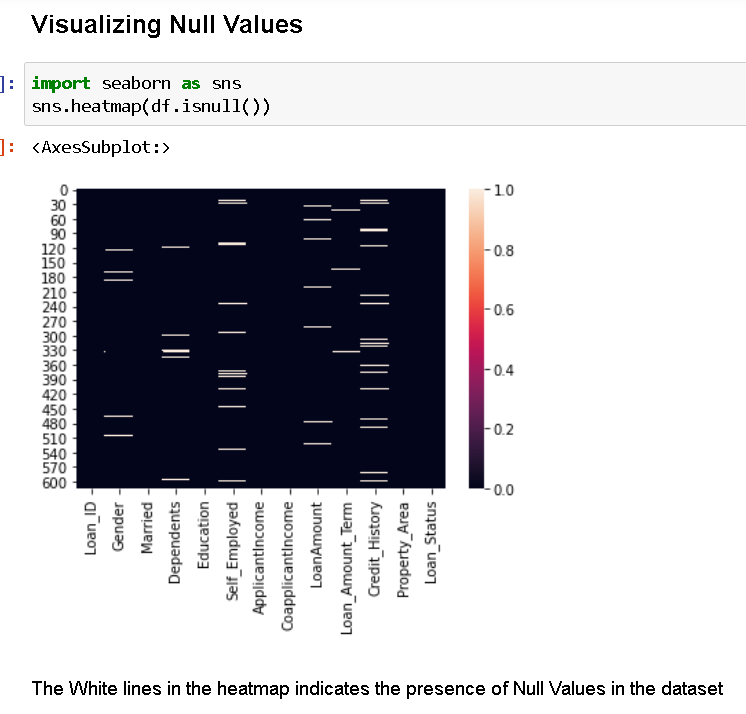
we shall look at various exploratory data analysis methods like –

1. **Variable Identification**-

Looking the input data, what will be the output variable

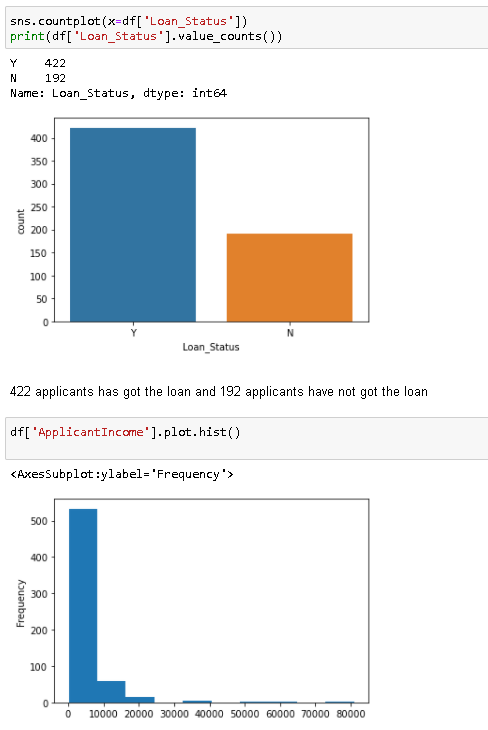
1. **Handling/ Removing null values**

It is possible that there may be null values in the dataset. It can be checked with isnull() function. Missing values usually occurs when an entry is left empty .It can be filled with the mean ,median or mode of the column through fillna() function. Or sometimes SimpleImputer() function is used to replace the Nulll values along with label encoding them ,as in Gender , Married, Dependents ,Self-Employeed ,Loan\_Amount ,Loan Amount Term and Credit\_History have Null values we can also visualize it as you can see in the following image



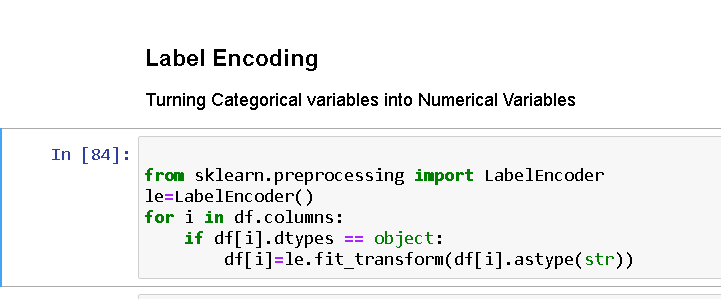
1. **Visualization of dataset**

It consist of exploring the dataset visually through histogram, bar chart usually for single variables, Like its better to use countplot for the categorical value and histogram or distplot for continuous variable like for all the categorical values like Gender, Married, Self-Employed, Education, Dependents, Credit History, Property area, and Loan\_status we used countplot and for the continuous values we used histogram as you can see in following image



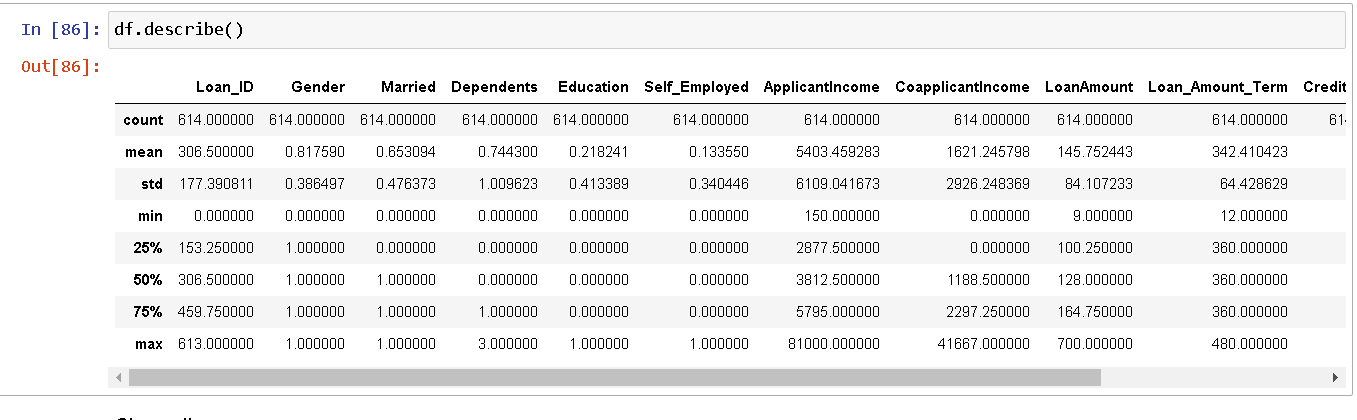
1. **Handling Categorical variables** –

Most statistical model cannot take Objects / Strings as input they only takes numbers as inputs , with LabelEncoder() it is possible to categorize the string into Numbers as 1,2,3 and so on like Gender, Married, Self-Employed, Education, Property area, and Loan\_status have string input they all are converted through LabelEncoder().fit\_transform function, as you can see in the image



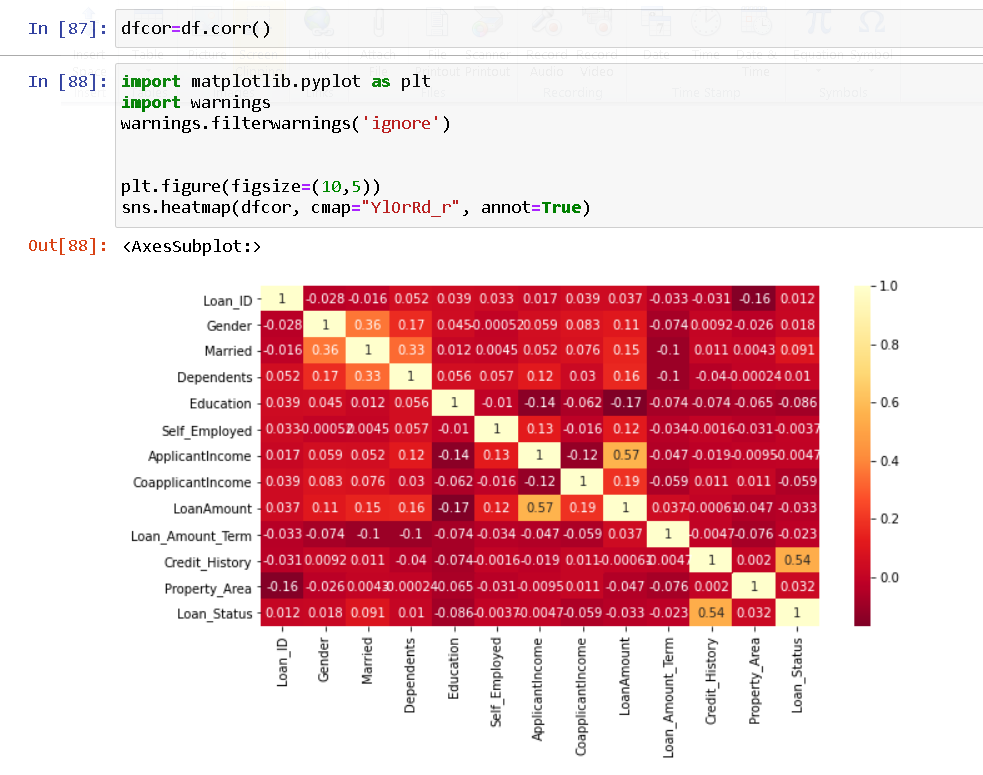
1. **Descriptive Statistics** :

It is a way of giving a brief overview of the dataset we are dealing with , including some measures and features of the sample.   
 For this we can use the describe() function the describe() function tells the mean() , min(), max(), 25th percentile , median 50th percentile and 75th percentile of the data.

  
 If there is significant difference between 75th percentile and max, it means Outliers are present just like in ApplicantIncome, CoApplicantIncome , and Loan Amount.  
 The Standard Deviation std() range is high it means the data is highly spread and skewness is present in the data just like in ApplicantIncome, CoApplicantIncome , and Loan Amount.

1. **Correlation**

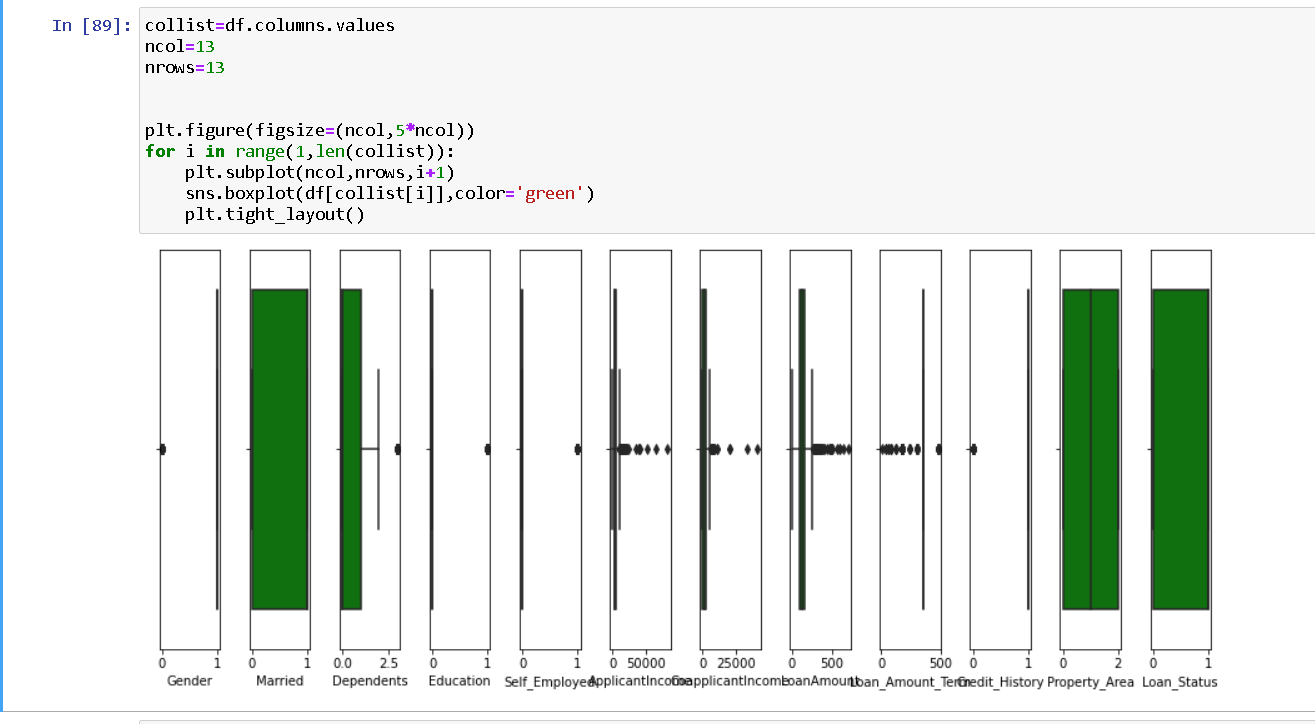
Correlation is the statistical metric for measuring to what extent Different variables are interdependent, like if one variable changes how it affects the change in other variables corr() function is used to see the correlation among the dependent variable and independent variable.



Like Loan\_Status is positively correlated with Credit-History and Negatively correlated with Education all other features have week relation with Loan\_Status.

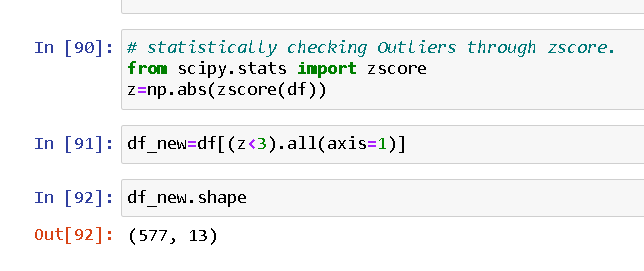
1. **Outliers**

Outliers are the value that lies outside the range of the data, Outliers can be visualized using boxplot as shown in the figure



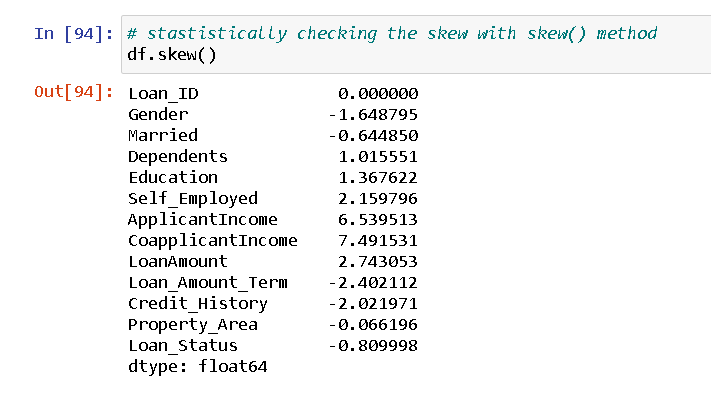
In our case ApplicantIncome, CoApplicantIncome , and Loan Amount and Loan\_Amount\_Tem columns has outliers .

Outliers can also be identified through zscore() function from scipy.stats library or sometimes through IQR method, To remove the Outliers we have to select the data that has zscore less than 3 ,

  
96 rows contains outliers which is 15% of the total data, Loosing 15% is not at all affordable so dropping the idea of removing outlier.

1. **Skewness**

The data can be right skewed or left skewed if the median or mean is high and data is highly spread it can be observed through the skew() method, if the skew score is negative and greater than 5 it means data is negatively skewed on left side and if the data is more than +5 it means the data is skewed on right side. In our dataset ApplicantIncome CoApplicantIncome is the skew score is greater than +5 and Loan\_status and Married is negatively scored .It is necessary to cure skewness for this Power Tranform function is useful but it should be done after separating features and target

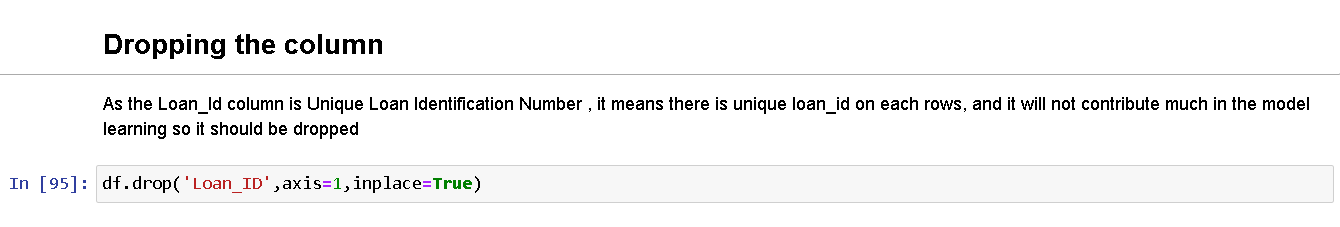


1. **Preprocessing Pipeline**

Pipelines are the special way to simplify the code , Pipeline is generally used if we have to perfrom the code repeatedly and it is generally used in the case where there are separate training and testing data

1. **Dropping the Column**

The columns which do not contribute much to the dataset can be dropped like in this example Loan\_ID its unique identification number. it can confuse the model because every record is unique so we are dropping the column



**Building Machine Learning Models**

A machine Learning Model can be a thought of a mathematical equation used to predict a given one or more other values

1. **Separating Features and Target column**

It is necessary to separate the independent/Features column into a variable ( x ) and target column into a variable ( y ).

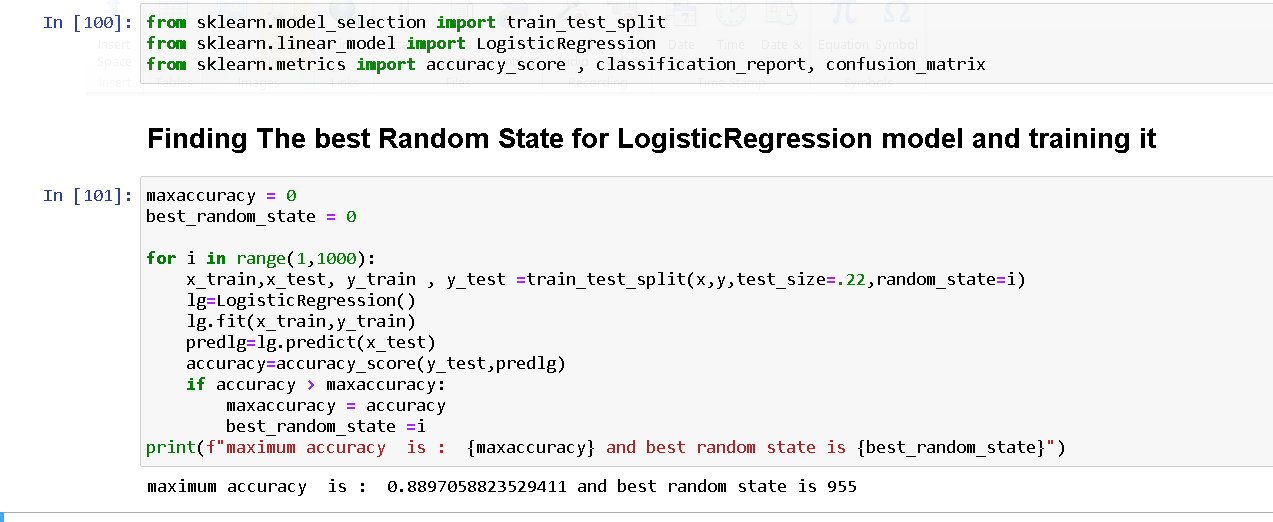
1. **Splitting the Data for Training and Testing -**

In ML the separated data is splitted into 4 parts for Training and Testing of features ( x ) and for Training and Testing of Target ( y ) like x\_train , x\_test , y\_train , y\_test.

It is possible through a inbuilt library of sklearn’s train\_test\_model , it separated the data into test\_size suggested by the DataScientist

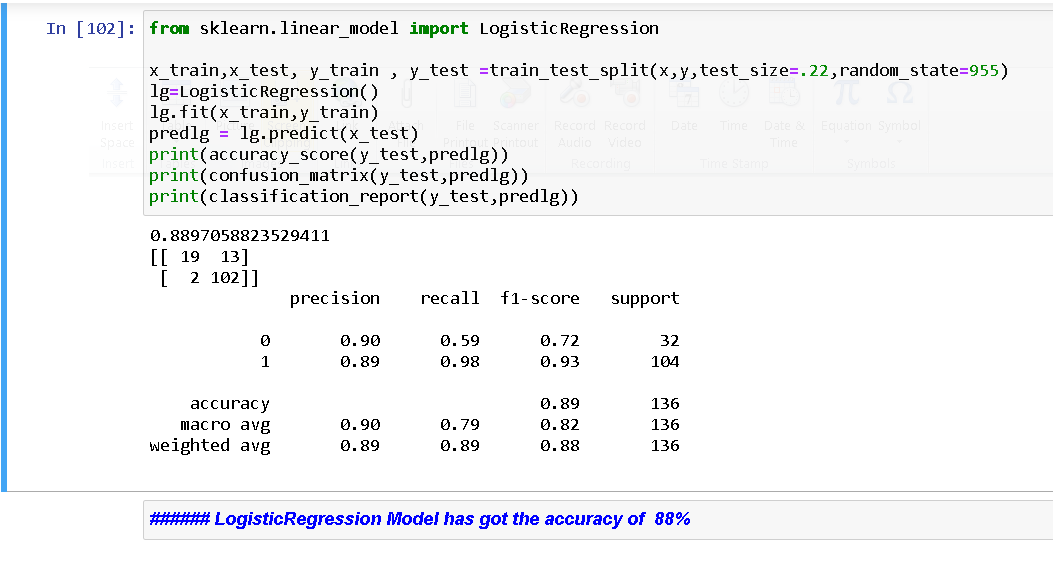
1. **Finding the Best Random State –**

The random\_state is one of the important parameter to be submitted and inserting best random state gives the best accuracy. It can be done through the for loop , In it we can compare the accuracy at each random state and find the best accuracy.



1. **Training the Models –**

To find the best model it is necessary to train 3-4 models,



In the same way I have trained LogisticRegressionModel , DecisionTreeClassifier, RandomForestClassifier, and SupportVectorClass to check which model is giving the best accuracy

i) LogisticRegression Model is giving the accuracy of 88%

ii) DecisionTreeClassifier Model is giving the accuracy of 81%

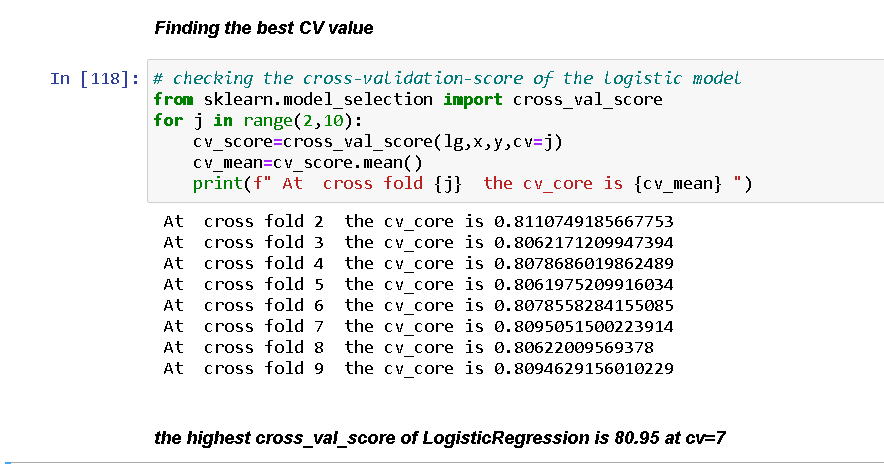
iii) RandomForestClassifier Model is giving the accuracy of 85%

iv) SupportVectorClass Model is giving the accuracy of 88%

Here the above models are giving quite good accuracy, so it is necessary to recheck the model’s learning

5)Cross-Validation of the Model

Sometimes it is possible that model may overfit for that cross validation is used to check model’s learning ,so we have done cross-validation of all the models along with 9 cv values and compared the difference to check for best model and select it, you can go through the image below



Like this we find the cross\_val\_score of all four models

i) cross\_val\_score of LogisticRegression Model is 80.95 % at cv= 7

ii) cross\_val\_score of DecisionTreeClassifier Model is 71.99 % at cv= 5

iii) cross\_val\_score of RandomForestClassifier Model is 79.15 % at cv= 6

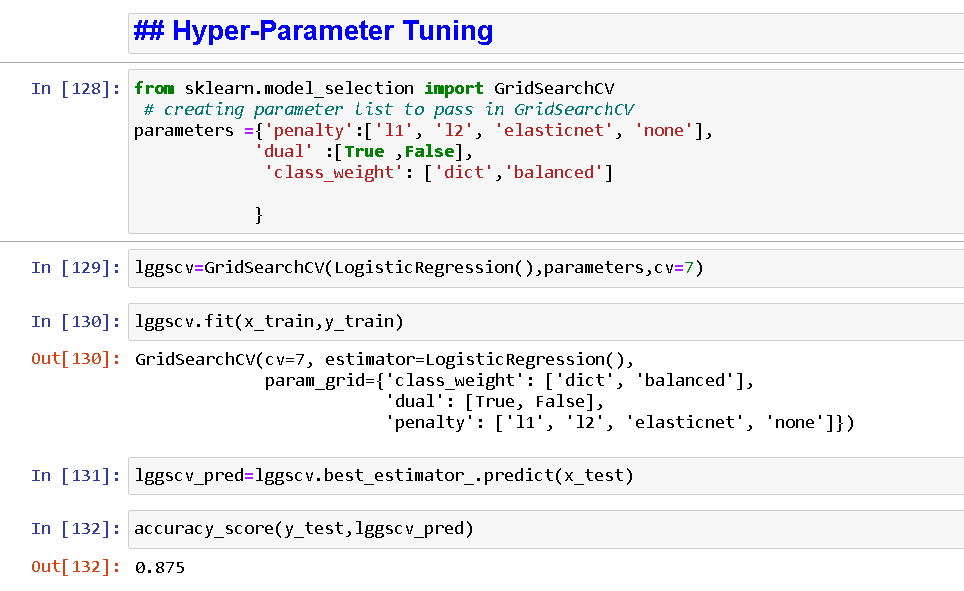
iv) i) cross\_val\_score of SupportVectorClass Model is 80.95 % at cv= 4

6) **Selecting the Best Model -**

After getting the accuracy and cross\_val\_score of all the models the comparision is made and the concusion can be made that LogisticRegression and SupportVectorClass has got the same accuracy and same cross\_val\_score , so we can go with any of this model, but as the classification is of two variables , so I will go with Logistic Regression, as Logistic regression is mostly used for binary\_classification and SVC supports multiple classification. so finally trainig our model further with Logistic Regression

7) **Hyper Parameter Tuning Using GridSearchCV –**

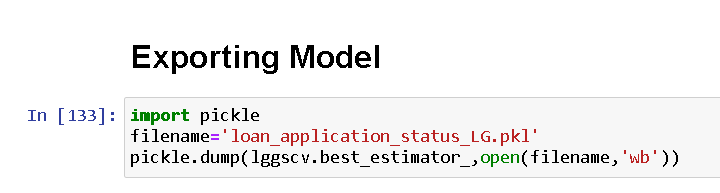
After selection of model it is necessary to Give the best parameters to the model and it is possible with GridSearchCV method which chooses the best parameters and makes the model fit for exporting



As you can see that we have Trained the LogisticRegression Model and we have got the accuracy of 87 %

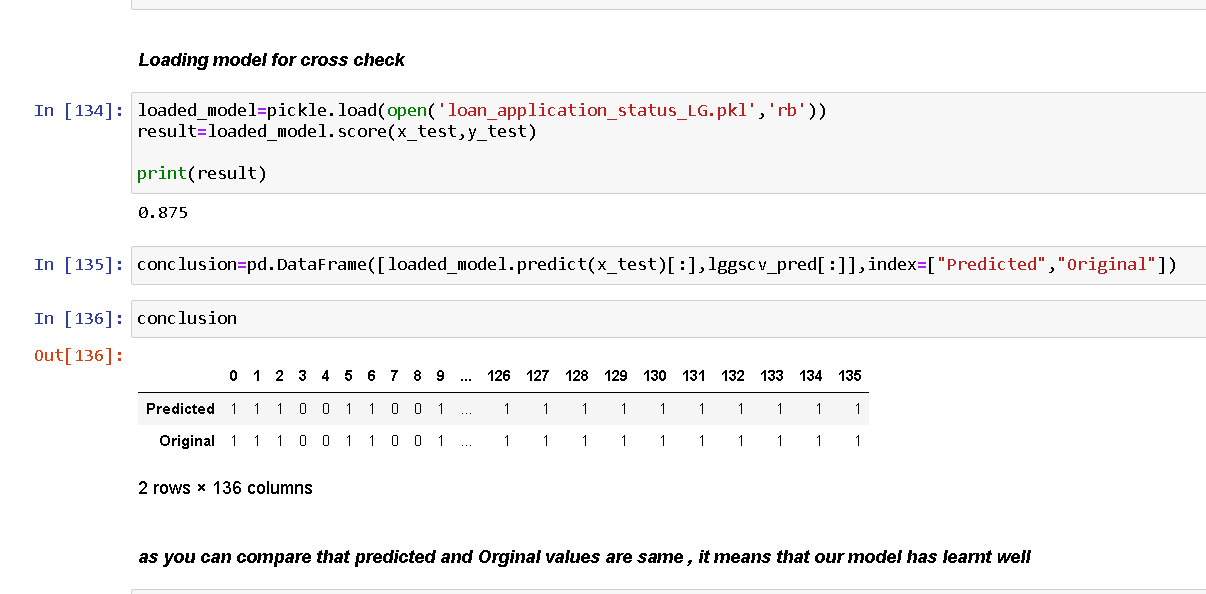
7)**Exporting the Model**

After GridSearchCV the model is exported into .obj file or .pkl file it can be done through Joblib Library or through pickle library, you can see the code



**Conclusion**

As we have gone through a good portion of the problem, we have done data analysis, data cleaning, EDA, we have built the model and exported it too. After we have exported the model we can cross check the performance of our model by making DataFrame of predicted and original values. We can also do a cross check to test our model’s performance you can see below in the image.



Now we can easily predict whether the applicant is defaulter or not or should bank give him/her loan or not by loading the model and providing the value.

In this way we have successfully created the model and exported it .