**Loan Application Status Prediction**

**Vinod Adep Satyanarayana**

**1.Problem Definition :-**

I Have a Dataset of Loan application, The dataset contains 12 features and 1 label total 13 columns present, Features are Loan\_ID , Gender, Married , Dependents , Education, Self\_Employed,  ApplicantIncome, CoapplicantIncome, Loan\_Amount, Loan\_Amount\_Term , Credit History ,Property\_Area, etc. Label is Loan\_Status

So Above Information is states that, the loan applicant will be approved or not, We have to build a model that can predict whether the loan applicant will be approved or not, Whenerver we build a model for this project it could be very usefull for loan approve process.

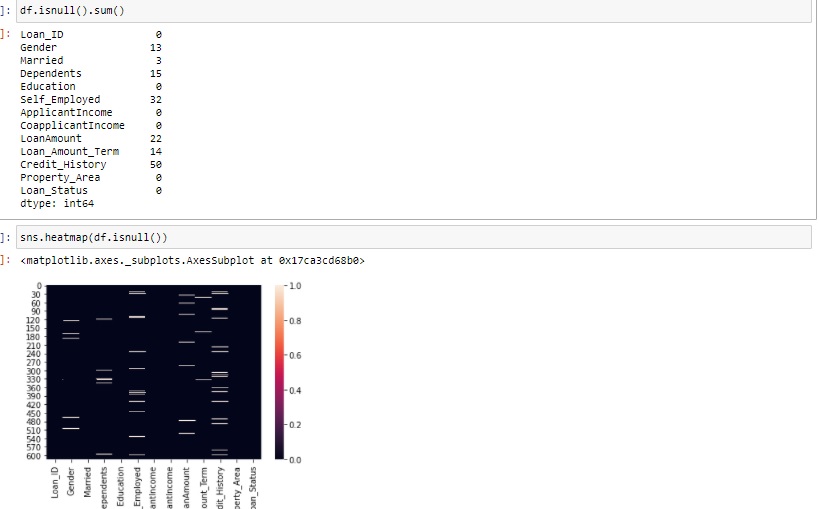
**2.Data Analysis :-**

The Loan Id is different for all applicants ,When we applying a loan the bank will investigate on credit history and property area ,So in this case credit history is very less for all applicants but Property area is in 3 different categories which are Urban, Rural, Semiurban .

Value Count for Property area is Urban is more than others 2, Rural is more than Semiurban, Loan Status columns is in Y or N ,Also loan status is depend upon the customer is educated or not, self employed or not so these two columns are listed in dataset ,Another category is for loan is Dependents or not ,The Main part for loan status is how much applicant income per/month or per/year also this is listed in dataset .

Shape of Dataset is [614r \* 13c] ,Null values have more in data ,I think it would be hard to handle with missing values ,but as data scientist we have to ready for any problems Null values of columns are Gender[13], Married[3], Dependents[15], Self-Employed[32], Loan Amount[22], Loan amount Term [14], Credit History [50].

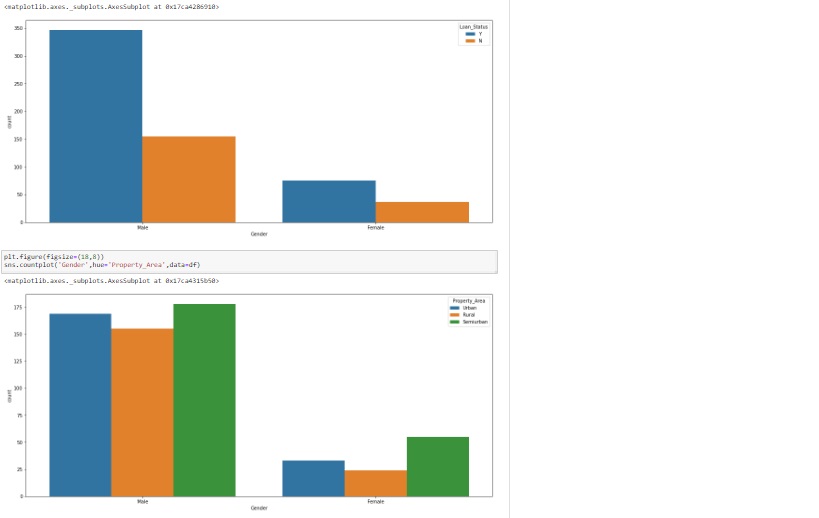
6 Objects data types are present in dataset which are Loan ID, Gender, Married, Education, Self-Employed, Loan Status, etc.



**3.Explotary Data Analysis :-**

In EDA Loan Status is two category M or F, Male applicants are more than Female applicants and also Male applicants are higher in acceptance of loan process than Female applicants, So in data visualization Males are more than Female applicants ,This is because value counts in [Gender] column Males are more than Females .Next coming to part of Property Area ,it have 3 categories Urban, Rural, Semiurban ,In data visualization Semiurban area applicants are more than other 2 and Urban area is more than Rural ,Semiurban is higher than others because, the loan will be needed for middle class and self employed peoples ,Rural applicants are lives in villages so they don’t need for much more money to survive or home loan, they have already own house and property ,And the other part is Urban area is considered as a Rich peoples are living in that area.

Loan Status for Self employed is Y have less and N have more, this is because Self Employed applicants are less as per dataset ,Here imbalance data for Loan Status ,whenever label data is imbalance we have to balance data or directly apply training data to model but recall and f1score decides the training data is good fit for model or not, If training data is good fit for model we don’t need to imbalance data directly apply to model .In this case I do not tried balance the data ,it would be decided after model prediction.





Now the Data Cleaning part, I had removed all null values through Simple Imputer on [Gender, Married, Self-Employed] with “most-frequent” parameter and [Dependents, Loan Amount, Loan Amount Term, Credit History] with most “most-frequent” parameter on nan .

So Now dataset does not have any null values but remaining part I have to convert all features and label in “int” or “float” datatype ,Data consist so many object data types ,I’m applying Label Encoding for coverting object to int ,This is all for done Data Cleaning part .

**4.Building Machine Learning Models :-**

Splitting of Data into train and test for the model prediction ,we cannot model prediction without splitting of data ,I given values of test size=.30 and random state = 1

Why we have to use machine learning model for prediction ,The answer is machine learning have a capability of read a binary data and gives a accuracy for that particular model. With machine learning we can solve the testical part of AI(Artificial Intelligence).This part is very important for outcome of project or dataset ,with this result anyone can build a project or making business decisions.

To Build a Model what we want, following are requirements splitting of data into train & test ,we have to split data on features and label ,test size which is given to model for prediction results it will depend upon you, how much you want to give but there is certain limit for test size 30 to 25 ,because we have to give more than 70% data to training ,In Model generation training and test split is very important role, So I could consider Machine Learning modeling is big role for any predictions.

I applied classification algorithms for training data ,I had given this models for training data Random Forest Classifier with 78% accuracy ,Logistic Regression with 78% accuracy, Support Vector Classifier with 78 %accuracy ,AdaBoost Classifier with 75 % accuracy, Decision Tree Classifier with 68% accuracy, KNeighbours Classifiers with 77 % accuracy. These all algorithms are given different accuracy because they have different functions and working ,but we need to try all algorithms with different parameters, The problem is based on Classification problem ,it contains metrics such as accuracy score, confusion matrix, classification report ,auc score and so may.

After model applying ,we have to analyse on confusion matrix and classification report this will gives the outcome is good fit to model or not. Also I had done Hyper Parameter Tuning with different parameters ,I applied SVC to “rbf” , “linear” with C[1,10] then model gives the best parameter “linear” with C[1] with accuracy 81 %.Also I had tried Cross Validation score with different algorithms ,Finally The Best Model for this Dataset is Support Vector CLassifier with 82% accuracy.

**6.Conclusion :-**

The Best Model for this dataset is Support Vector Classifier with 82 % accuracy. I think this model prediction is used for whenever any applicant apply for loan it will gives that particular applicant eligible or not for loan application .As of my best I tried to give for solving this dataset ,This type of techniques or model generation is used in bank industry for loan application status, In Pandemic everyone have lost their job and business is going down ,to survive take a loan from bank then customers will pay interest every month ,I think in this situation this project is very usefull in present and future ,because loan option is very high priority for middle class to survive or investment or business.