## Evaluation Project 6 Loan Application Status Prediction

**Problem Statement:**

* This dataset includes details of applicants who have applied for loan. The dataset includes details like credit history, loan amount, their income, dependents etc.
* Independent Variables:
* Loan\_ID
* Gender
* Married
* Dependents
* Education
* Self\_Employed
* ApplicantIncome
* CoapplicantIncome
* Loan\_Amount
* Loan\_Amount\_Term
* Credit History
* Property\_Area
* Dependent Variable (Target Variable):
* Loan\_Status

Problem Definition-

We need to build the model that help to predict the loan application to be accepted or rejected so we have to see various inputs and check relation between each variable to target .

2. Data Analysis

* + In loan application status dataset we have 614 rows with 13 columns including target variable.
  + A Target Variable is 'Loan\_Status' having object datatype and It is categorical variable.
  + Gender, Married, Education,Self Employed, Credit History, Loan Status are categorical features.
  + There are three types of datatype dtypes: float64(4), int64(1), object(8)
  + 7 out 13 columns contains missing value.
  + As small amount of data is missing so we use mean amd mode to replace with NaN values.

**Lets explore categorical features before missing value imputation.**

* + Out of Total loan application 80 % applicants are Male. We can Explore loan amount for each gender applied and evaluate whether on the same basis loan is approved for each gender or not?
  + Only 20% applicants are self employed. So it will interesting to gain insight on relation between Applicant income and loan approval for non self employed category. We will look to find any benchmark range of Income for loan approval.Another benchmark we will try to find is about loan requirement for these two categories.
  + Nearly 70% are married and 75% of loan applicants are graduates
  + Almost 60% of the applicants have no dependents.
  + Most of applicants come from Semi Urban areas, followed by Urban and Rural areas.
  + 80% people previously have credit history. Normally people having credit history are seen more prone to get loan approval.
  + Nearly 70 % applicant gets loan approved.

**We have filled the Categorical Data by Mode Now we see Numerical data we will compare Mean and Mode for numerical column which ever is low we use to replace NULL values**

* The mean is greater than median loan amount.
* Clearly we can see outliers in boxplot and feature is strecth to far in distribution plot.

**An extreme outliers seen we use Median**

### Filling Missing Values

#### Imputation details :

* + .Missing values in Loan amount is impute with median value.
  + Maximum Loan term is 360 Months so Missing value in Loan amount term is replace with 360 Months.
  + Credit History, Self Employed, dependents, Gender and Married are replace with mode of repective features.
* In Applicant Income & Coapplicant Income Std deviation value is greater than median. So data is spread and skewed.
* Taking 75% and Max rows into consideration we can surely say that Outliers exist in Applicant Income, Coapplicant Income,Loan Amount.
* Since Credit History is Categorical variable there is no significance in different statstical parameter of it.
* Minimum Tenure for Loan is 12 Months and Maximum Loan tenure is 480 Months.
* Minimum Applicant income is 150 and maximum is 81000.

**Cheking Data**

**Comment:**

* 65.3 % loan applicants are married. We can see from Count plot that Married people have more possibility of getting loan approval.
* But Here comes Interesting observation from crosstab :
  + Unmarried Women are more chances of getting loan approval compare to married women.We will try to find which factor actually play deciding role here like education, employeement or Income.
  + Married Men are more chances of loan approval.

**Observation:**

Most of feature are poorly or moderately correlated with target variable expect Credit History.

* Maximum correlation of 0.561 exist between Credit History and Loan status.

## Checking Multicollinearity between features using variance\_inflation\_factor

All features VIF is within permissible limit of 10.

So No Need to Worry About Multicollinearity.

## 5. Balanceing Imbalanced target feature

As Target variable data is Imbalanced in nature we will need to balance target variable.

As Target variable data is Imbalanced in nature we will need to balance target variable.

#### We can see that RandomForestClassifier() gives us good Accuracy and maximum f1 score along with best Cross-validation score. we will apply Hyperparameter tuning on Random Forest model and Used it as final model.

Working on various ML models we use

* + - LogisticRegression()
    - ABC=AdaBoostClassifier()
    - ET=ExtraTreesClassifier()
* KNN=KNeighborsClassifier()

To check the Accuracy of score we use Cross Validation Score

Concluded Remarks

While working with Data we have learned various correlation with target variable and Inputs and implement multiple Algorithms to see the accuracy

We have taken ExtraTree Classifier because it have highest Accuracy Final accuracy score is 86.6%