Loan Application

Problem Definition:

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
* Applicant Income
* Co-applicant Income
* Loan Amount
* Loan Amount Term
* Credit History
* Property Area

Dependent Variable (Target Variable):

* Loan Status

We 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.

Data Analysis:

First of all, we need to upload our csv file in our jupiter note book and then we need to describe our data like count, mean, std, min,25%,50%,75% and max with the numerical values given in the data frame

Then we need to check the null values present in the date frame with info command, the output will be like floating, integer or object in the data structure, it shows the data structure of each row present in it

And then we need to check the null values present in the data frame using is null command, then we need to fill the null values using mean and mode function

Finally, we need to check again with the null values present in it or not using is null command.

EDA Concluding Remarks:

The next step of action is Exploratory data analysis, in that we are using count plot method. Which use to visualize the categorical variables likes gender, education, married, dependents, self-employed, property area and loan status.

And we use dist. plot, numerical values to visualize the applicant income, loan amount in the data frame, it shows that it is left skewed so that we are using log transformation to normalization

Pre-Processing Pipeline:

Combining the applicant income and co-applicant income into a single column and named as total income.

Then we need to find the correlation in matrix using heatmap, we found that the values with high color contrast is highly correlated and the color with less contrast is less correlated values.

Further, we need to drop the unwanted columns in the extracted data frame,

Finally, we use label encoding technique to convert categorical values into numerical data frame by importing Label Encoder.

Building Machine Learning Models: Before building the machine learning model we need to separate dependent and independent variable and named in as x as independent variable and y as dependent variable.

We need to call the train test split dictionary, and we use test size is 0.25 and random state is 42 and further we move towards model training.

By importing cross validation score to see the cross-validation score of the model then we use logistic regression, Decision tree classifier, Random Forest classifier and Extra tree classifier and we need to run with the codes and analysis which mode is the best by comparing the less difference between cross validation and accuracy.

Finally, we need to choose the best model which is random forest classifier and further process with the hyper parameter turning to boost the model accuracy and cross validation and that’s the final prediction score of the model.

Concluding Remarks:

We choose random forest classifier as the best model because the accuracy values are 77.27 and the cross validation are 77.52, the difference between the values are much lesser when compared to the other model and finally we save the model.