**Loan Approval Prediction Using M.L**

Introduction

In this article, we are going to solve the Loan Approval Prediction. This is a classification problem in which we need to classify whether the loan will be approved or not. Classification refers to a predictive modeling problem. In which we predict the given input data using Machine Learning Algorithms.

Table Of Content

1.      Problem Definition  
2.      Data Analysis  
3.      EDA Concluding Remarks  
4.      Pre-processing Pipeline  
5.      Building Machine Learning Models  
6.     Concluding Remarks

Problem Definition

A Housing Finance company deals in all kinds of home loans. They have a presence across all urban, semi-urban and rural areas. The customer first applies for a home loan and after that, the company validates the customer eligibility for the loan. The company wants to automate the loan eligibility process (real-time) based on customer detail provided while filling out online application forms. These details are Gender, Marital Status, Education, number of Dependents, Income, Loan Amount, Credit History, and others.

To automate this process, they have provided a dataset to identify the customer segments that are eligible for loan amounts so that they can specifically target these customers.

Data Analysis

1. Import necessary libraries for reading dataset like PANDAS, NUMPY.
2. Read data.
3. Check for null values in dataset either present or not.
4. Some null values are present in dataset, we can fill nulls by using Mode, Mean according to situation.
5. Now check for datatype.
6. If any columns are present as object datatype, then we have to encode that column, let’s check for it.
7. We will do encoding with Ordinal And Label Encoder wherever required according to data.
8. “Gender” column is having some null values, filled null values with the help of mode method because column is present as object datatype. And also encode it by using Label Encoder.
9. “Married” column is having some null values, filled null values by using mode method, and also encode it by using Label Encoder.
10. “Dependents” column is having some null values, filled null values by using mode method because we cannot use mean or median here.
11. “Education” column, no null values are present here. This column is in object datatype so we have to encode it by using Ordinal Encoder because I want to give some order here.
12. “Self-Employed” column, having some null values, filling null by using mode method, because column present as object datatype, so we have to encode it by using Label Encoder.
13. “Applicant-Income” column, no nulls are present. Column present in int type so we don’t need here to encoding.
14. “Co-Applicant Income” column, no nulls are present. Column is in int type.
15. “Loan Amount” column, some null values are present, filled by using mode method, No need to encode because column is in int type.
16. “Loan Amount Term” column, some null values are present, filled by using mode method. Column is in int type.
17. “Credit History” column, some null values are present, filled by using mode method. Column is in int type.
18. “Property Area” column, no null values are present. Because column is in object datatype, so we have to encode it by using Ordinal Encoder to give some order to property area.
19. “Loan Status” column, we will only encode this column because it is in object datatype.
20. Drop “Loan-ID” column, because it is just a nominal data column.

EDA

1. More applicants are male than female.
2. More applicants are married than unmarried.
3. Count of dependents with 0 is maximum.
4. Graduate are more than Not Graduate.
5. Self-Employed are less in count.
6. More properties are in Semiurban Area.
7. Credit-History is present of more applicants.

This process is done by plotting Distribution Plot using necessary library.

Now, plot boxplot having a look either outliers are shown or not.

Some outliers are shown by boxplot.

Now, treat with all outliers by help of Quantiles and Inter Quantile Range.

A lot of rows have deleted during removal of outliers.

After removing outliers, when we plot again Distribution Plot, now data looks into some type of shape like bell-curve shape.

Preprocessing Pipeline

Now, split dataset into two variables.

Import libraries for standardize the data and for train test split.

Because it is classification problem (loan will be approved or not).

So, we also import some library for checking the model accuracy and model error.

Building M.L Models

First Model -- Logistic Regression

When we train our data by using first model, and predict.

Our first model giving accuracy 83.33%.

Second Model -- Decision Tree

When we train our data by using second model, and predict.

Our second model giving accuracy 75.75%.

Third Model – Random Forest

When we train our data by using third model, and predict.

Our third model giving accuracy 84.84%.

Fourth Model – SVM

When we train our data by using fourth model, and predict.

Our fourth model giving accuracy 81%.

Concluding Remarks

Now compare all four models on training and test data, how much our models understand the data and how perfect all four models predict.

By using roc curve, we plot for all four models together.

First, when we plot roc curve on training data ---

First model – 76%

Second model – 100%

Third model – 100%

Fourth model – 93%

Second, when we plot roc curve on test data ---

First model -- 56%

Second model -- 68%

Third model -- 76%

Fourth model -- 77%

From above all observations, we only consider our third model – “Random Forest”.

Saving model by using Pickle library.