Loan Application Status Prediction

PROBLEM-This is a classification problem in which we need to classify whether the loan will be approved or not.

The 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.

The Binary Classification problem in which we need to predict our Target label which is “Loan Status”.

Loan status can have two values: Yes or NO.

Yes: if the loan is approved

NO: if the loan is not approved

Data Analysis:-

First lets look at data set

Categorical Columns: Gender (Male/Female), Married (Yes/No), Number of dependents (Possible values:0,1,2,3+), Education (Graduate / Not Graduate), Self-Employed (No/Yes), credit history(Yes/No), Property Area (Rural/Semi-Urban/Urban) and Loan Status (Y/N)(i. e. Target variable)

Numerical Columns: Loan ID, Applicant Income, Co-applicant Income, Loan Amount, and Loan amount term.

NULL VALUES- We observe that there are “13” missing values in “Gender”, “3” in “Married”, “15” in “Dependents”, “32” in “Self\_Employed”, “22”  in “Loan\_Amount”, “14” in “Loan\_Amount\_Term” and “50” in “Credit\_History”. The missing values of the numerical and categorical features are “missing at random (MAR)” i.e. the data is not missing in all the observations but only within sub-samples of the data. So the missing values of the numerical features should be filled with “mean” and the categorical features with “mode” i.e. the most frequently occurring values. We use Pandas “fillna()” function for imputing the missing values as the estimate of “mean” and “mode” remains unbiased.

Now- No more null values are present

DATA VISUALISATION:-

To create a new attribute named “Total\_Income” we’ll add two columns “Coapplicant\_Income” and “Applicant\_Income” as we assume that “Coapplicant” is the person from the same family for eg. Spouse, Father etc. and display the first five rows of the “Total\_Income”.

We see that they're extreme values in the range from “0-`10,000” and the data is left skewed which might be possible that some people may have applied for high loans due to specific needs. Very few applicants are in the range of “40,000-80,000”. So we’ll apply log transformation on “Total\_Income” to make it closer to normal in the distributed data.

DATA CLEANING:-

As a part of the data cleaning process, let’s [drop unnecessary columns](https://www.mage.ai/blog/data-cleaning-variance) which are not affecting the “Loan-Status” as it helps in improving the accuracy of the model and we’ll display the first five rows of the dataframe.

ENCODING

This is done to make system understand language as system can only understand numerics and doesn’t understand category

 By using “[label encoding](https://www.mage.ai/blog/feature-engineering-category-value-mapping)” we’ll convert the categorical features to numerical features and display the first five rows of the dataframe.

MODEL TRAINING

Now, It’s time to train the model!! For this, we’ll split the data where we keep “33%” of the test data and the remaining for training data. We’ll perform “cross-validation” for better performance of the model and check the accuracy of each model in percent.

We’ll train the model using “Logistic Regression” and check the accuracy of the model. “Logistic Regression” is a popular classification algorithm that is used to predict a binary outcome i.e. “yes/No”.

After implementing the machine learning algorithm the accuracy obtained by “Decision Tree” is 64.97%. Let’s plot the confusion matrix in the testing model and get the summary of the predicted results.

Hence we tried other machine learning algorithms also like Random Forest, Linear Regression which increased accuracy to 86% approx

Conclusion

So now we’ve automated the process of loan approval for the company and provided a low code solution through Random Forest application.