**Loan Application Status Prediction**

**Classification Problem**

**Submitted by:**

**Chandan Kumar Rai**

**Batch No. DS2312**

**Introduction:** Based on the information provided in the dataset like, credit history, loan amount, Income, dependents, property area, we have to predict whether the applicants loan request will be approved or not.

**Problem Definition**

Banks main source of income comes from the loan given by bank. Bank does not want to give loan to anyone who will be not be able to pay EMI’s. Before approving the loan, bank does its verification. Based on the information provide by the applicant, banks want’s to predicts whether to approve or reject the loan application.

**Data Analysis**

The dataset contains 13 variables. Out of 13 variables, 12 are features and 1 is label. Below are the features.

1. Loan\_ID
2. Gender
3. Married
4. Dependents
5. Education
6. Self\_Employed
7. Applicant Income
8. Co-applicant-income
9. Loan\_amount
10. Loan\_amount\_term
11. Credit History
12. Property Area

We have to predict **Loan\_Status (Yes/No)**

This was a classification problem. Datatype of the 12 variables are object -7, int -1, float -4. Following are the observation about dataset.

1. There was no duplicate record
2. No. of records in the dataset are 614
3. Data set was balanced
4. There were null values present in the data set
5. Following variables are having null values.
   * Gender
   * Married
   * Dependents
   * Self-Employed
   * Loan\_amount\_term
   * Credit\_history
   * Loan\_amount
6. We have deleted null values for **Loan\_amount** and **Credit\_history.** Dataset records reduced to 543 from 614 after deleting null values.
7. For rest null values are filled using mode.
8. We found outliers for:
   * Applicant Income
   * Co-Applicant Income
   * Loan Amount
   * Loan\_Amount\_term
   * Credit History
9. We have removed outliers using zscore. After removing outlier’s dataset records reduced to 512 from 543. There was data loss of 5.7% after removing outliers.
10. Skewness found in the below variables:
    * Applicant Income
    * Co-Applicant Income
    * Loan Amount
    * Loan\_Amount\_term
    * Credit History
11. We have removed skewness using **“boxcox”** and **“yeojohnson”** method.
12. We have converted categorical columns into numerical columns using **OrdinalEncoder.**
13. We do not found multicollinearity between variables.
14. We have checked vif score for all the features variables and found vif score less than 5 for all the variables.

**EDA Concluding Remarks**

In Exploratory data analysis of loan application dataset, information and properties of the dataset are found. We investigated correlation between variables, identifies outliers, identifies skewness in the dataset. The quantitative and qualitative features of the dataset are analyzed using various plots, including countplot, distplot and boxplot.

**Building Machine Learning Models**

This was a classification problem. We have to predict Loan status in yes or no. We have implemented below steps to build machine learning models.

1. Finding best random state for the model
2. Following 8 models were built.
   * RandomForestClassifier
   * ExtraTreesClassifier
   * SVC
   * GradientBoostingClassifier
   * AdaBoostClassifier
   * BaggingClassifier
   * DecisionTreeClassifier
   * KNeighborsClassifier
3. Accuracy score, confusion matrix and classification report were built for all the models.
4. SVC was our best model with the accuracy of **88.31%** among all the models
5. We do hyperparameter tunning for SVC
6. After hyperparameter tunning our accuracy increased to **89.61%** from **88.31%.**

**Concluding Remarks**

Problem statement was to predict whether the applicant is eligible for loan as per set criteria. EDA conducted on dataset to check whether dataset is balanced or not, treat null values, check multicollinearity, check skewness, etc. in the dataset. After cleaning the dataset, data visualization was done using countplot and distplot. We checked outliers using boxplot and removed them using zscore method.

We build 8 classification models and choose the best one based on the accuracy score. We saved the model and at last compared the actual and predicted values.