LOAN APPLICATION STATUS PREDICTION-A CLASSIFICATION MODEL BUILD USING MACHINE LEARNING

Seep Bansal

Department of I.T, Baba Banda Singh Bahadur Engineering College, PTU
Post Graduate in Banking Services, Amity University, Noida.

ABSTRACT

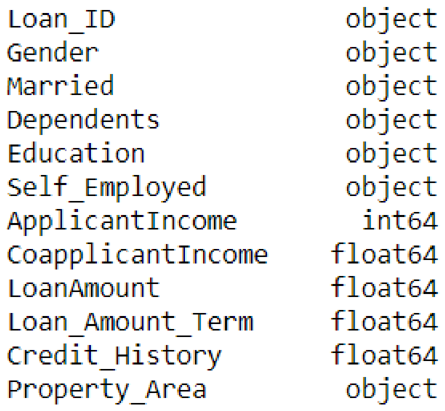
Banking system plays a very important role in boosting Indian economy. It is considered as one of the crucial pillars of our financial sector. In fact, it reflects the health of our country. The Indian banking system has witnessed the major structural and policy changes since 1991. Banks play a very important role in the mobilization of deposits and the disbursement of the credit to the various sectors of our country. Thus, loan approval is considered as one of the most important and crucial part of the banking system. Above all, recovery of loan is one the major contributing parameter in the bank statements. With the advancements in the field of artificial intelligence and machine learning techniques many researchers worked on building one such model called loan application status prediction i.e either the loan will be approved or it will be rejected based on factors like applicant income, dependents on applicant, co-applicant income, credit history and many more. This article incudes 5 classification machine learning algorithms used for predicting loan approval status with the Logistic Regression classification model giving the accuracy score of 95%. The other metrices like Precision, F1 score, Recall, ROC_AUC score are also evaluated.

DATASET

Raw data is something which has been collected from source but in its very initial phase. It has not been processed or cleansed to provide any useful information. Dataset considered here is the Loan application status prediction dataset and is of open source.

The dataset has total of 614 rows and 13 columns/features. The dataset is in a raw form i.e we will be performing EDA (Exploratory Data Analysis) i.e some pre-processing and data cleaning techniques to covert the data into useful information for our machine learning model.

Briefing about the features of dataset along with there data type:



```
Loan_ID                 object
Gender                  object
Married                 object
Dependents              object
Education               object
Self_Employed           object
ApplicantIncome          int64
CoapplicantIncome      float64
LoanAmount             float64
Loan_Amount_Term       float64
Credit_History         float64
Property_Area           object
```

Loan ID: contains alpha numeric characters.
Gender: Male or Female
Married: Yes or No
Depdendents: 0,1,2,3+
Education: Graduate, Not graduate
Self_employed: Yes or No
Applicant Income: Numeric
Coapplicant Income: Numeric
Loan Amount: Numeric
Loan_Amount_Term: Numeric
Credit History: 0 or 1
Property_Area: Urban, Rural, Semi_urban

The label of the dataset:

```
Loan_Status             object
```

Loan Status: Yes or No

Here, the loan id is the unique id of each applicant. Gender tells whether the person is male or female. Married feature tells us whether the person is married or not. Dependents shows how many members in the family is dependent on the loan applicant. Education tells us how much the person is educated. We also look out if the person is self-employed or salaried. We may also look out for the applicant income and co applicant income also the loan amount he has applied for and the most important is the credit history of the person whether he has any previous loans or not or he is not the defaulter. After taking all these factors into account we may check whether the person is illegible for loan or not.

Observations

- There are some missing values in the dataset.
- As we have to predict the loan approval status which makes it a classification problem.
- Some of the data is in a categorical form we need to convert it to numeric type.
- As we have some unrealistic values in some of the columns.
- The values in the dataset are not scaled it is important to perform scaling techniques
- As the dataset is not balanced, sampling techniques are to be considered.

DATA ANALYSIS

In this Phase we have created some graphical representations of various features and label:

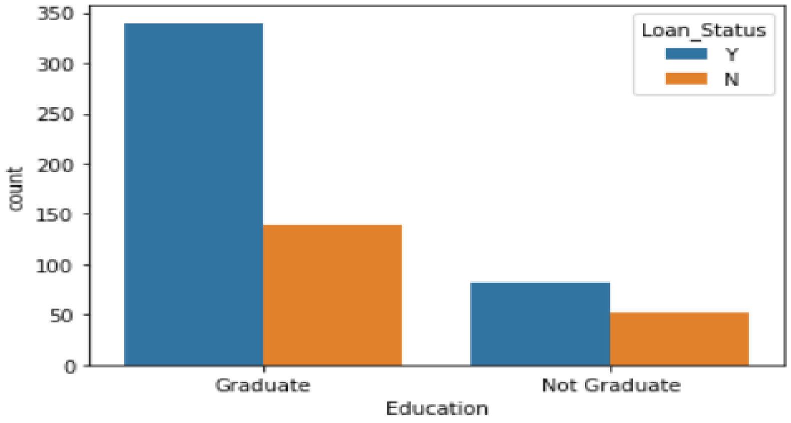


Figure 1: Says loan status is approved for those applicants who are graduate.

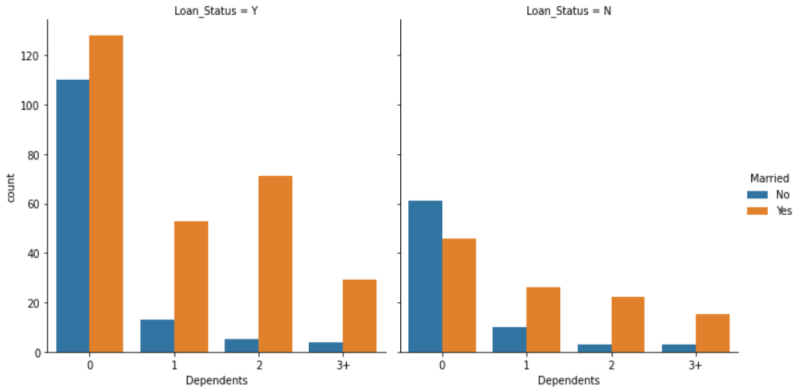
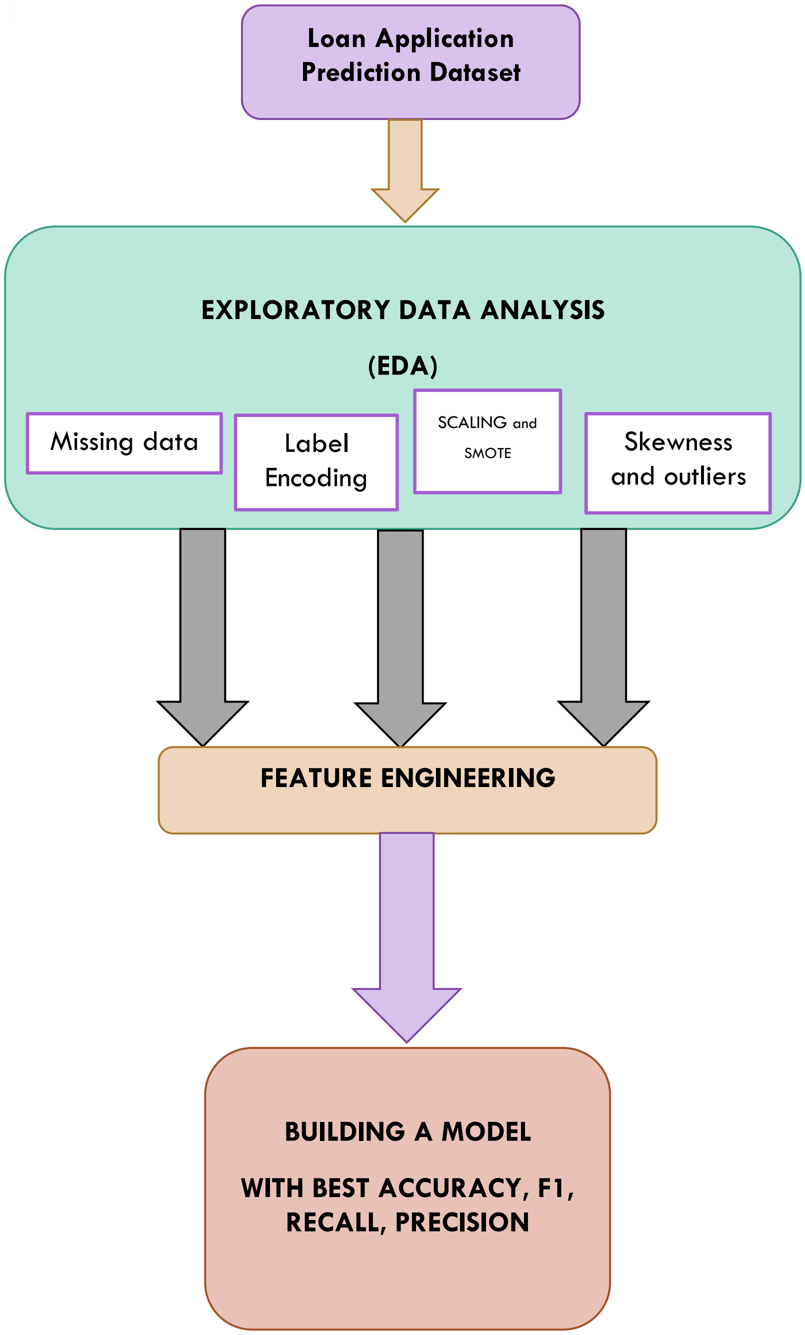


Figure 2: Says loan application status is approved for thoes who have 0 dependents on them. i.e if they are married and there partners are earning in major of the cases loan status is approved.

Observations:

- Loan status will be approved for the one having credit history score as 1.
- Loan status is approved majorly for the males as compared to females.
- Maximum Loan cases are approved for the property area in semi urban location.

Methodology for Loan Application Status Prediction:-



Pre-Processing Analysis

Preprocessing is one of the important steps in building a model. In this phase we usually deal with missing values if any. In case of any irrelevant value we will remove that data. Improving the quality of the data is the major concern in this phase so as to get the best accuracy for the model. Transforming all the categorical data into numeric so as to feed the data to algorithm is carried out. Unrealistic values/ Outliers are the important part to be dealt with. Scaling techniques to be followed if required. Sampling is crucial to balance the dataset.

EDA Concluding Remarks

- ✓ Replacing missing values with mode value for categorical data and with mean value for numeric data
- ✓ Label encoding technique is followed for transforming categorical into numeric
- ✓ Skewness and outliers are removed using power transformation and z-score respectively.
- ✓ Standard scaling technique is implemented.
- ✓ Combined sampling technique is followed to balance the dataset.
- ✓ VIF factor was calculated to check the multicollinearities.

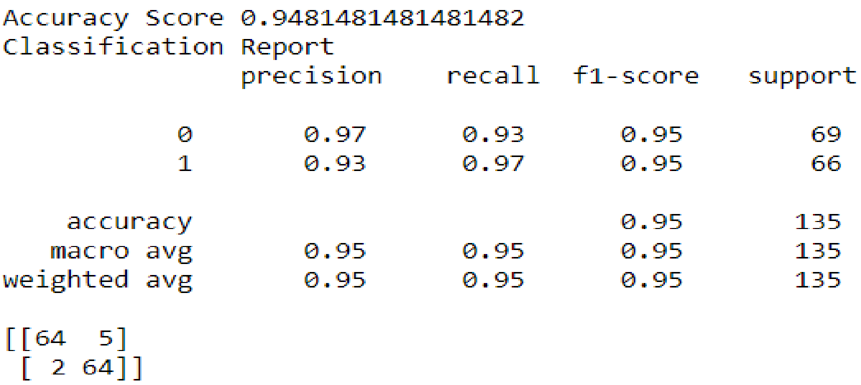
Classification Experiments

We have performed various classification algorithms like Logistic Algorithm, Decision Tree Algorithm, Support Vector Mechanism, Random Tree Classifier, Ada Boost Classifier to check the accuracy. Logistic Regression Classification model predicting the best results along with the accuracy of 95%. Logistic Regression is a predictive analysis algorithm based on the concept of probability. Cost function of which is also known as sigmoid function.

Evaluation Matrix

Evaluation matrix plays a very important role in evaluating the performance of any classification model. The metrices that are evaluated here are the accuracy score, ROC_AUC score, classification matrix which includes F1, recall, Precision and the confusion matrix. Time taken to test the model on dataset plays a very crucial role.

Here, for Logistic Regression the accuracy score we are getting is 95% and the classification matrix i.e precision, F1, recall also scores to 97%, 95%, 93% respectively.



```
Accuracy Score 0.9481481481481482
Classification Report
              precision    recall  f1-score   support

           0       0.97      0.93      0.95        69
           1       0.93      0.97      0.95        66

    accuracy                           0.95       135
   macro avg       0.95      0.95      0.95       135
weighted avg       0.95      0.95      0.95       135

[[64  5]
 [ 2 64]]
```

Figure 3: Summary

As the loan approval status is to be predicted i.e either the loan is approved or rejected which makes it a binary classification we have here plotted the ROC_AUC curve and ROC_AUC score is evaluated. Higher the ROC_AUC score better the performance of model is. In this case AUC_ROC score is 95% and the ROC_AUC curve which plots the TPR and FPR at different thresholds is shows distortion at some points which confirms the presence of noise.

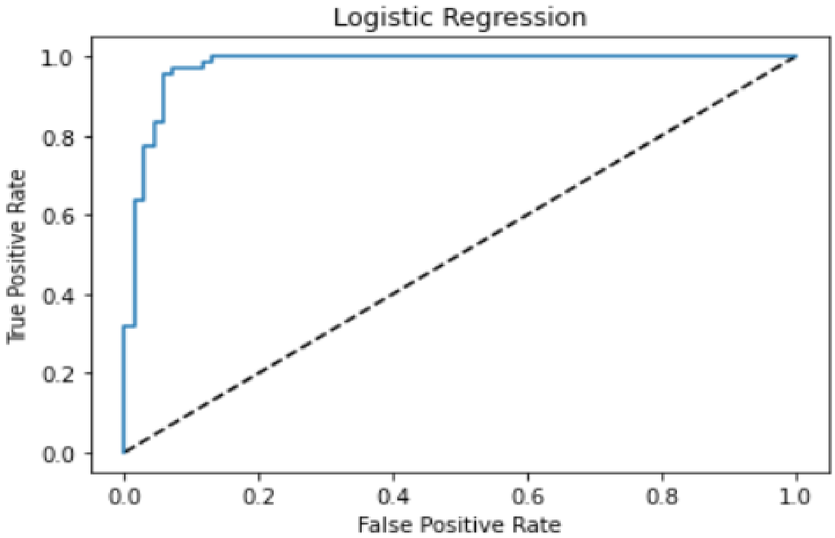


Figure 4: AUC_ROC curve

Conclusion

The above article states the methodology and the classification algorithms used for predicting the status of loan application based on the various features provided. We have analyzed the data through various graphical representations and cleaned the data before passing it to a classification model. We have developed 4-5 classification model and the best results were predicted by the Logistic Regression with 95% accuracy.