International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 09 | Sep 2021 www.irjet.net p-ISSN: 2395-0072

Loan Approval Prediction Using Machine Learning

Kritika Pathak[1], Shazia Shaikh[2]

[1-2]Student, Dept. of Electronics and Telecommunications Engineering, Vivekanand Education Society's Institute of Technology, Mumbai, India

---

Abstract - *As we all know that now-a-days there's a rising in banking sector, ensuing several folks applying for bank loans. looking for the mortal to whom the loan should be approved could be a troublesome. Thus, in this paper, we've proposed a model that predicts the authorization or rejection of associate degree mortal. This will be done by taking into account some machine learning techniques by predicting the model with the information of the previous records of the folks applied for loan.*

***Key Words*: Banking Sector, loan, predict, machine learning, logistic Regression**

1. INTRODUCTION

Loans area unit the core business of banks. The most profit comes directly from the loan's interest. The loan corporations grant a loan with an associate degree of intensive method of verification and validation. However, they still don't have assurance if the mortal is in a position to repay the loan with no difficulties. Loan Prediction is extremely useful for worker of banks similarly as for the mortal additionally. The aim of this Paper is to supply fast, immediate method for the meriting candidates. It will offer special advantage to the bank. The Loan Prediction System will calculate the load of every option participating in loan process and on new check information same options area unit processed with relevance of their associated weight .A threshold time will be set for the mortal to see whether or not his/her loan is sanctioned or not. Loan Prediction System permits to jump to specific application so it is checked on priority basis.

2. LITERATURE SURVEY

Paper 1: Loan Credibility Prediction System using Data Mining Techniques

Abstract: As we all know that now-a-days there's a rising in banking sector, ensuing several folks applying for bank loans. Looking for the mortal to whom the loan is approved could be a troublesome method. Data processing techniques are getting very popular today attributable to the wide handiness of giant amount and therefore the want for remodeling such data into knowledge. Techniques of data mining enforced in numerous domains like retail business, telecommunication business, biological information analysis, etc. During this paper, we had a tendency to plan a model that predicts loan approval/rejection of associate degree mortal by taking help of data processing techniques. This will be done by training the model with the info of the previous records of the folks applied for loan.

Paper 2: An Approach for Prediction of Loan Approval using Machine Learning Algorithm

Abstract: Banks have several commodities to sell however, main supply of financial gain of any banks is on its credit line. So, they'll earn from interest of these loans that they credit. A bank's profit or a loss depends to an oversized extent on loans i.e., whether or not the purchasers area unit return the loan or defaulting. By predicting the loan defaulters, the bank will scale back its Non-Performing Assets. This makes the study of this development important. Previous analysis during this era has shown that there are a large number of strategies to check the matter of dominant loan default. However, because the right prediction is important for the maximization of profits, it's essential to check the character of the various strategies and their comparison. So, it becomes necessary in this predictive analytic to check the matter of predicting loan defaulters: The logistic regression model. The information is collected from Kaggle for learning and prediction. Logistic Regression models are performed and therefore the totally different measures of performances are computed. The models are compared on the idea of the performance measures like sensitivity and specificity. The ultimate results have shown that the model turn out totally different results. Model is marginally higher as a result of it includes variables (personal attributes of client like age, purpose, credit history, credit quantity, credit length, etc.) aside from bank account info (which shows wealth of a customer) that ought to be taken under consideration to calculate the likelihood of getting loan properly. Therefore, by employing a logistic regression approach, the proper customers to be targeted for granting loan is simply detected by evaluating their chances. The model concludes that a bank shouldn't solely target the main clients for granting loan however it ought to assess the opposite attributes of a customer similarly that play a really necessary half in credit granting choices and predicting the loan defaulters.

3. METHODOLOGY

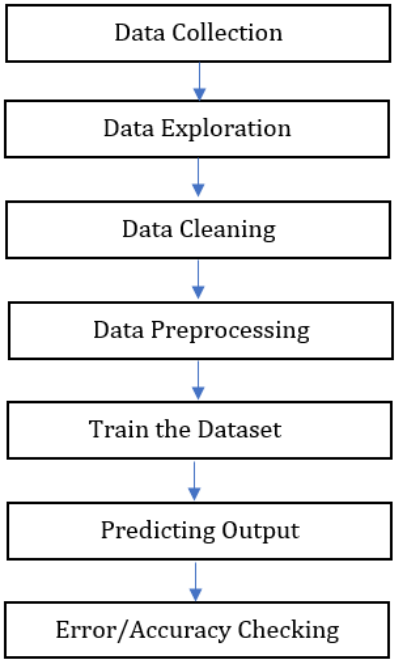


Fig -1: Proposed System

The given problem is a supervised classification problem as we are finding whether the person is reliable for a loan or not that is Yes or No.
This can be solved with any of the algorithms listed below:

i. Decision tree
ii. Logistic regression
iii. Random Forest

These algorithms are some of the few algorithms that can be used to solve the problem.

3.1. Collection of Data

The input dataset is the whole bank dataset of customers who applied for the loan approval. The dataset is a CSV file. The dataset can be read into the python environment by using the read_csv() method in pandas. So, we should import pandas into the present python environment. Some features of the Customers dataset are Loan_ID, Married, Gender, Dependents, Education, Self-employed, Applicant Income, ApplicantIncome, Loan_Amount_Term, Loan Amount, Credit History, Loan_Status and Property Area.

```
0    Loan_ID
1    Gender
2    Married
3    Dependents
4    Education
5    Self_Employed
6    ApplicantIncome
7    CoapplicantIncome
8    LoanAmount
9    Loan_Amount_Term
10   Credit_History
11   Property_Area
12   Loan Status
```

Fig -2: Attributes Of Dataset

3.2. Data Exploration

Various Libraries and packages were imported which was required to explore the data. After that, some top rows were looked at a glance. Also, we checked if the dataset contains nulls values or not.

3.3. Data Cleaning

Some of the values in the data set were null values. All the null values were dropped using the function dropna. This is a very important step in order to obtain a reliable dataset.

```
df.isnull().sum()
```

```
Loan_ID              0
Gender               0
Married              0
Dependents           0
Education            0
Self_Employed        0
ApplicantIncome      0
CoapplicantIncome    0
LoanAmount           0
Loan_Amount_Term     0
Credit_History       0
Property_Area        0
Loan_Status          0
dtype: int64
```

Fig -3: Data Cleaning

3.4. Data Preprocessing

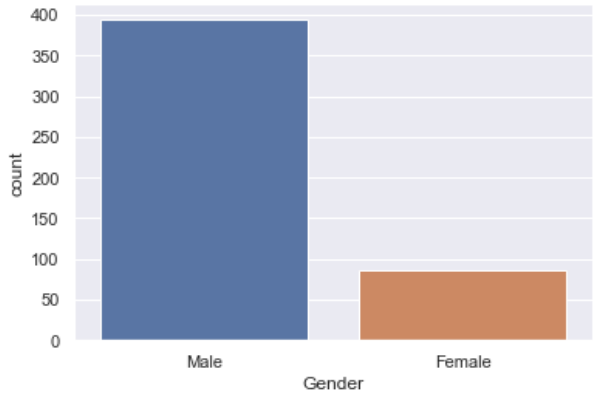
Some attributes of the dataset were of the article sort. we have a tendency to therefore regenerate the article sort variable to number as machine learning algorithms typically deem the mathematical process which needs its input to be of number sort. Also, we've got to Standardize the information as models tend to perform higher once the options square measure on a relative scale.

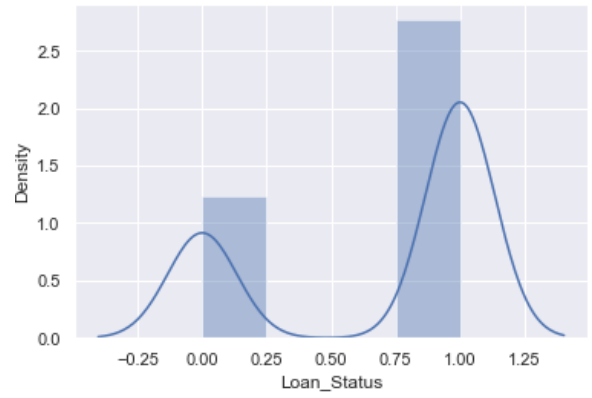
3.5. Data Visualizations

The value count was done and so information was envisioned. Even when the data analysis, there's still no distinctive issue to work out loan standing. Categorical information was regenerated into numerical information.

```
sns.countplot(x='Gender',data=df)
```

```
<AxesSubplot:xlabel='Gender', ylabel='count'>
```





3.6. Data Modelling

After the data is visualized, the data is modeled/trained. For this, the packages of 3 algorithms (Logistic regression, Decision tree and Random forest) were then imported. The model was then outlined and also the accuracy score was evaluated. Logistic Regression was the simplest work with the very best accuracy score of eighty-three.

It is applicable for categorical dependent variables employing a given set of freelance variables. Thus, the end result should be a categorical or distinct price. The output is often either Y or N, 0 or 1, true or false, etc. however rather than giving the precise price as zero or one, it offers some probabilistic values that lie between zero and one. In logistic regression, instead of fitting a curve, we have a tendency to work an"S" formed logistic operator, that predicts 2 greatest values (0 or 1). The curve from the logistic operation demonstrates the likelihood of one thing, as an example, despite whether or not the cells square measure harmful or not, a mouse is rotund or not supported on its weight, and so on. It's a major rule as a result of it will offer possibilities and classify the employment of various kinds of information and simply determines the foremost effective variables that square measure used for classification. The S-structure curve is additionally referred to as the sigmoid operation or the logistic operation.

$$\log(1/1-y) = b0+b1x1+b2x2+b3x3+\ldots\ldots+bnx$$

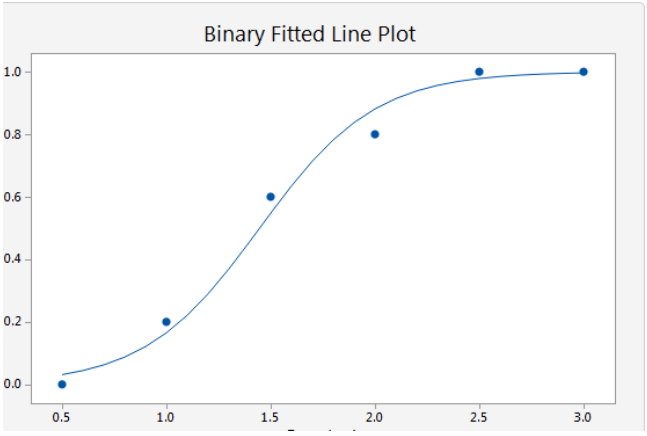


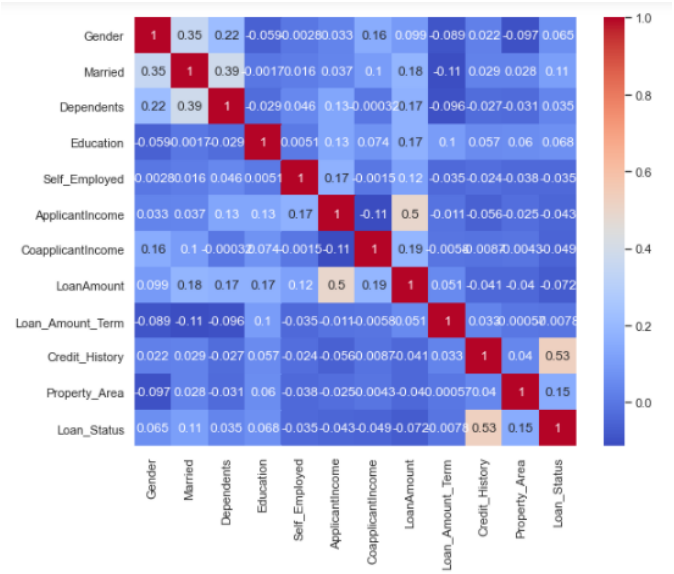
Chart -1: Logistic Function

The Logistic Regression was employed to fit the test data set and prediction result was displayed Successfully.

4. EXPERIMENTAL ANALYSIS

Based on the information given by the loan applicant, we will predict whether or not the loan of the applicant is approved or not. The applicant must provide these values, and supported by these, the model can predict whether or not the loan is going to be approved or not.

After we have a tendency to apply Label encryption on Dependents and Property_Area and Dummy encryption on the remaining options.



Train-Test split applies to any supervised learning rule. Here the total dataset is get divided into 2 datasets as Train and check to create a model and to ascertain the performance of a model and every dataset gets divided into 2 once more as independent options and dependent options (only in supervised).

· Train Dataset: to train the machine learning model for learning functions.

· Test Dataset: to understand the performance of a trained machine learning model.

Logistic function: G(z)=1/(1+e-z)

Accuracy = (True Positives + True Negatives)/Total Sample)
Our accuracy was 0.8317667058123509.

Precision: exactness = (Number of True Positive)/(True Positive + False Positive)
Our exactness score was 0.8117647058823529.

Recall: Recall = (True Positives)/(True Positive + False Negative)
Our Recall score was 0.9726573293456218.

F1 Score: F1 Score = 2/((1/Precision) + (1/Recall))
Our F1 Score was 0.8317654110854729

5. CONCLUSIONS

In our model by employing a logistic regression model we finally predicted whether or not the loan is approved or not. So, to implement this, numerous input variables were required to get the output.

When a program takes the computer file input it offers the output within the type of binary i.e., either 0 or 1. If the output is one then '1' is going to be displayed and it indicates that the loan is approved. If the output is zero then '0' is going to be displayed and it indicates that the loan isn't approved. Here, we have a tendency to had enforced a loan credibility prediction system that helps the organizations in creating the proper call to approve or reject the loan request of the purchasers.

In this model, a Logistic Regression rule or algorithm is employed for the prediction. Incorporation of alternative techniques that vanquish the performance of common data processing models needs to be enforced and tested for the domain.

6. REFERENCES

[1]. Dileep B. Desai, Dr. R.V.Kulkarni "A Review: Application of Data Mining Tools in CRM for Selected Banks", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (2), 2013, 199 –201.M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[2]. J.H. Aboobyda, and M.A. Tarig, "Developing Prediction Model of Loan Risk in BanksUsing Data Mining", MachineLearning and Applications: An International Journal (MLAIJ), vol. 3, no.1, pp. 1–9, 2016. K. Elissa, "Title of paper if known," unpublished.

[3]. A.B. Hussain, and F.K.E. Shorouq, "Credit risk assessment model for Jordanian commercial banks: Neuralscoring approach", Review of Development Finance, Elsevier, vol. 4, pp. 20–28, 2014. JAC: A JOURNAL OF COMPOSITION THEORYVolume XIII, Issue V, MAY 2020ISSN: 0731-6755Page No: 324

[4]. T. Harris, "Quantitative credit risk assessment using support vector machines: Broad versus Narrow default definitions", Expert Systems with Applications, vol. 40, pp. 4404–4413, 2013.