IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 3, May 2023

Loan Status Prediction

Nalla Sai Ram Reddy[1], D. Ram Babu[2], Shaik Sajid[3],Ch Nihal Reddy[4]
[1,3,4]B.TECH Scholars, Dept. of Computer Science and Engineering
[2]Assistant Professor , Dept. of Computer Science and Engineering
Sreenidhi Institute of Science & Technology, Hyderabad, India

Abstract: *Banking Industry always needs a more accurate predictive modelling system for many issues. Predicting credit defaulters is a difficult task for the banking industry. The loan status is one of the quality indicators of the loan. It doesn't show everything immediately, but it is the first step of the loan lending process. The loan status is used for creating a credit scoring model. The credit scoring model is used for accurate analysis of credit data to find defaulters and valid customers. The objective of this project is to create a credit scoring model for credit data. Various machine learning techniques are used to develop the financial credit scoring model. In this project, we propose a supervised machine-learning model for credit data. We use a supervised machine implemented using python. This proposed model provides important information with the highest accuracy. It is used to predict the loan status in commercial banks using a machine learning model.*

Keywords: Customer loan, Prediction, preprocessing, classification models

I. INTRODUCTION

Circulation of the loans is that the core business a part of as good as each and every bank. The principle parcel the bank's resources are straightforwardly came from the benefit acquire from the advances distributed by the banks. The main goal in banking system is to invest their resources in safe hands whereverit's. Now a day's several banks/financial agencies approve loan after a relapse method of verification and validation however stillthere'sno surety whether or not the chosen candidate is the worthyright candidate out of all candidates. Through this method we areable to predict whether that particular candidate is safe or not andthe whole method of validation of attribute is automated bymachine learning technique [8][6]. The disadvantage of this modelis that it emphasizes completely different weights to every issue however in reality sometime loan can be approved on the premiseof single strong part only, that isn't possible through this method. Loan Prediction is useful for member of staff of banks as well as for thecandidate. The aim of this Paper is to apply quick, immediate and easy way to choose the worthy person [6]. It will give special gain to the bank. The Loan Prediction method can automatically compute the heaviness of each attribute taking part in loan processing and on new test data information same issues are prepared with regard to their comparable heaviness. A period breaking point can be set for the candidate to check regardless of whether his/her loan can be affirmed or not. Loan Prediction technique licenses bouncing to explicit candidates with the goal that it very well may be keep an eye on need premise. This Paper is completely overseeing the power of Bank/finance Company, entire procedure of prediction is done secretly no colleagues would have the option to caution the process. Result against specific Loan Id can be ship off different divisions of companies so that they can make a proper move on application. This aides all others divisions to done different conventions. *Data Source* we obtained customer loan dataset from kaggle [4][2]. The dataset consists of various values/variables such as sex, marital status, education, self employed, loan status, applicant income, co-applicant income etc…*Data Description* the dataset has 614 rows and 13 columns. 1 out of 13 columns is the target attribute i.e., default one attribute is target value. The dataset split into train and test data having shape (614, 13) and (367, 12) respectively.

II. LITERATURE SURVEY

SVM works relatively well when there is a clear margin of separation between classes. SVM is more effective in high dimensional spaces and is relatively memory efficient. SVM iseffective in cases where the dimensions are greater than the number of samples.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 3, May 2023

Bhoomi Patel, Harshal Patil, Jovita Hembram, Shree Jaswalare used data mining methodology to predict the likely default from a dataset that contains information about home loan applications, thereby helping the banks for making better decisions in the future [3].

This paper mainly introduces the main application of LSTM-SVM model in user loan risk prediction and elaborates the current economic background, and traditional risk forecasting method. On this basis, the prediction methodology based on LSTM method and SVM method is proposed, and the prediction results are compared with the traditional algorithm, and the feasibility of the model is confirm. However, the LSTM- SVM method proposed inthis paper actually has few limits andneedsto be improved in future research [7]

III. PROBLEM STATEMENT

Finance companies, banks are deals with different kinds of loans such as education loan, shop loans, home loans, personal loans etc all are part of our country loan types. All the companies and banks are present in villages, towns,cities. After customer apply for loan these banks/companies want to validate the customer details forthat candidate eligible for loan or not. The main purpose of the system is applicant loan approved or not basedon train models [6]

IV. PROPOSED MODEL

In Machine Learning, we are using semi-automatedextraction of knowledge of data for identifying whether aloan would be approved or not [6][8]. Classification couldbe a supervised learning within which the response is categorical that'sits values area unit in finite unordered set. To easily the matter of classification, scikit learn are used. The praim primacy of this system is company need not has to maintain a ground team to validate and verify the customer records. They can easily check whether the loan has to be approved or not by this prediction model.

In this paper preprocessing is major part used sklearn method is MinMax scalar i.e., helps normalize the data. Model selection with help of cross validation, train/test split, kfold, GridSearchCV.

A. Model Selection

Model selection is that method of selecting one in every of the models because the final model that addresses the issue. In there we have different steps. Theyare:

- Data filtering
- Data transformation
- Feature selection
- Feature engineering

For this process we have mainly two methods:

- Probabilistic model selection
- Resampling methods

In this paper we are using resampling methods such ascrossvalidation, train/test split, Kflod, GridSearchCV

B. Preprocessing

Data mining methods are used in preprocessing for normalize the data which is collected from kaggle. Thereis a need to convert because dataset may have missing values, noisy data. So, we are using data mining method for cleaning method [10][12]. Before using model selection process we are used preprocessing methodfor reduce the null values then recover the data withhelp of train/test split with help of MinMaxScaler [5].

MinMaxScalar, for each value in every feature MinMaxScalar cipher the minimum value within the feature then divided by the vary. Therange is the distinction between the first most and original minimum.It preserves the shapes of the firstoriginal distribution

C. Feature Engineering

It is the method of using domain data to extract options from data via data processing techniques. These features are wanted to improve the performance of machine learning algorithms. Feature engineering is thought-about as applied Machine learning itself. It is helping for import the models.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 3, May 2023

import numpy as npimport pandas as pd
import matplotlib pyplot as pltimport seaborn as sns

D. Machine Learning Methods

```
(Loan_ID                0
 Gender                13
 Married                3
 Dependents            15
 Education              0
 Self_Employed         32
 ApplicantIncome        0
 CoapplicantIncome      0
 LoanAmount            22
 Loan_Amount_Term      14
 Credit_History        50
 Property_Area          0
 Loan_Status            0
 dtype: int64,
 Male      489
 Female    112
 Name: Gender, dtype: int64,
 Yes     398
 No      213
 Name: Married, dtype: int64,
 No      500
 Yes      82
 Name: Self_Employed, dtype: int64,
 1.0     475
 0.0      89
 Name: Credit_History, dtype: int64)
```

Machine learning is a subset of AI that trains machines withvast volumes of data to think and act like humans without being explicitly programmed. In this paper we are using supervised (Classification methods) methods
Five machine learning classification models have been used for prediction of android applications. The models are available in python open source software. The brief details of each model are described below.

Decision Trees
The basic algorithmic rule of call tree needs all attributes or options ought to be discredited. Feature choice relies on greatest info gain of options.
The data pictured in call tree will delineate within the kindof IF-THEN rules. This model is associate degree extension of C4.5 classification algorithms represented byQuinlan.

Random Forest
Random forests are a classifying learning framework forcharacterization (and backslide) that workby building a very large number of Decision trees at planning time and yielding the class that's the mode of theclasses surrender by individualtrees.

Support Vector Machine
Used SVM to build and train a model prepare a demonstrate utilizing human cell records, and classify cells to whether the tests are benign (mild state) or dangerous (evil state).
Support vector machines are managed learning models that utilize affiliation R-learning calculation which analyze attributes and distinguished design information, utilized for application classification. SVM can beneficially perform a replace utilizing the kernel trick, verifiably mapping their inputs into high dimensionalattribute spaces [8].

Logistic Regression
Logistic regression is supervised learning classification algorithm (try to method connections and conditions between the target prediction output and input attributes) such that we are ab leto anticipate the yield values for new information based on those connections which it learned from the previous information sets [8][6].

K-nearest neighbor (KNN)
The KNN algorithm is a simple supervised machinelearning algorithm that can be utilized to unravel both classification and replace issues. It is easy to implement and understand but significantly slows as the size of that data on use

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 3, May 2023

grows [5]

$$d' = \frac{d - \min(p)}{\max(p) - \min(p)}$$

Experiment overview:

In this experiment firstly collect the data and understand the data with help of (.describe()) and then analyses of data then search for any missing/null/nosy data present in the dataset and are designed to detect errors in data at a lower level of detail. *Data validations* have beenincluded in the system in almost every area where there is a possibility for the user to commit errors. The system won't accept invalid information. Whenever invalid information is keyed in, the system like a shot prompts the user and also the usershould once more key within the information and also the system will accept for the info provided that the info is correct. Validations are enclosed wherever necessary.

The system is designed to be a user friendly one. In alternative words the system has been designedto speak effectively with the user. The system has been designed with popup menus

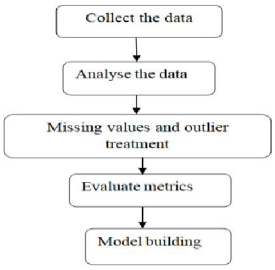


Fig (A): overview of experiment

Major Attributes:

In the below map shows the positive and negative values of attributes and heat map helps us to analyze the data dependent attributes. Loan Amount shows in after log form used

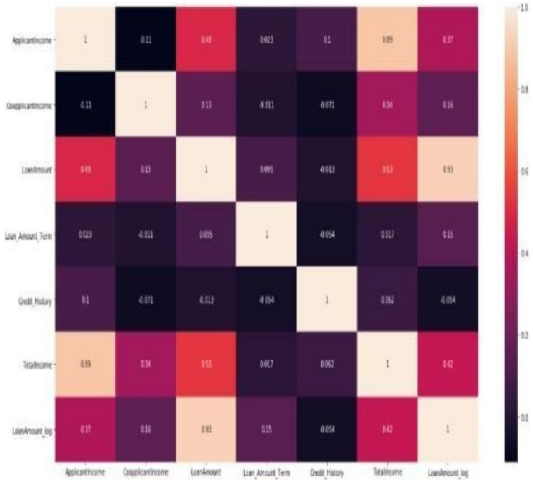


Fig (B): Heat map

Barplot():

A bar plot represents an estimate of central tendency for a numeric variable with the height of each rectangle and provides some indication of the uncertaintyaround that estimate using error bars.

sns.barplot(x=data_train['TotalIncome'],y=data_train['L oan_St atus'],hue=data_train[' Education'])

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 3, May 2023

Pd.crosstab():

Compute a basic cross organization of two (or more) components. By default computes a recurrence table of the components unless an cluster of values and an accumulation work is passed.

Loan_Status	N	Y	All
Credit_History			
0.0	82	7	89
1.0	97	378	475
All	179	385	564

V. RESULTS

Here shows all the methods we build and these methods are evaluate the accuracy, precision, recall, F1-score. And the below table represents the value obtained for the various metrics from the different methods. Here we choose the accuracy so, all methods comparatively SVM is the less accuracy. Therefore we can summarize that random forest is doing prediction well for our data.

Classification Results

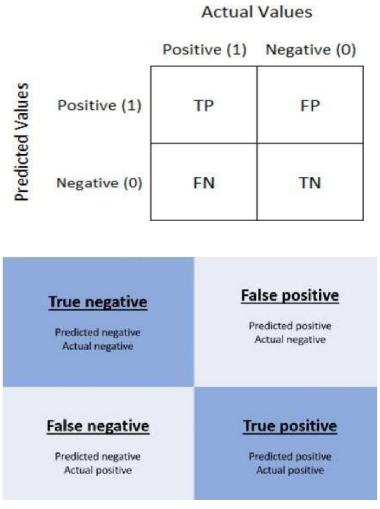
Used Algorithms	Accuracy	Precision	Recall	F1-score
Random forest	82%	0.84	0.82	0.81
Logistic regression	73%	0.73	0.74	0.73
Decision tree	72%	0.72	0.72	0.72
KNN	59%	0.52	0.59	0.53
SVM	78%	0.82	0.78	0.75

Fig (i): Results

VI. EVALUTION MODELS

Need for confusion matrix:

Multiple output categories are available in classification (predict category) models. While most error measures can tell us how much error there is overall in our model, they cannot be used to identify specific instances of error. Confusion matrix aids in identifying both the errors and the right predictions of a model for various distinct classes. The central matrix



IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 3, May 2023

Accuracy: It's worn to find the portion of correctly classified values. It is tell us how often our classifier is right. Sum of all true values divided by total values.

Number of classified samples =TP+TN Total number of samples= TP+FP+TN+FN

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

Fig (b): Accuracy

Precision: It is used to calculate the models ability toclassify positive values correctly.

Number of classified values = TP Number ofactual values = TP+FP

$$\textit{Precision} = \frac{TP}{TP + FP}$$

Fig (c): Precision

Recall: To calculate the models ability to predictpositive values

$$\textit{Recall} = \frac{TP}{TP + FN}$$

Fig (d): Recall

F1- Score: It is also called the F score or the F Measure. Put another way, the f1 score conveys the balance between the precision and the recall.

$$F_1 = \frac{2}{\frac{1}{\text{recall}} \times \frac{1}{\text{precision}}} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

Fig (e): F1-Score

Note: precision and recall are exactly helps to defineproblem ofgroup of predicted vales.

VII. CONCLUSION

In this paper, we have proposed customer loan prediction using supervised learning techniques for loan candidate as a valid or fail to pay customer. In this paper, various algorithms were implemented to predict customer loan. Optimum results were obtained using Logistic Regression, Random Forest, KNN, and SVM, decision Tree Classifier. Compare these five algorithms random forest is the high accuracy. From a correct analysis of positive points and constraints on the part, it can be safely ended that the merchandise could be an extremely efficient part. This application is functioning properly andmeeting to all or any Banker necessities. This part is oftensimply obstructed in several different systems. There are numbers cases of computer glitches, errors in content and most significant weight of option is mounted in machine- driven prediction system, therefore within the close of future the therefore called software system might be created more secure, reliable and dynamic weight adjustment. In close to future this module of prediction canbe integrated with the module of machine-driven processing system.

VIII. FUTURE SCOPE

The system is trained on old training dataset in future software can be made such that new testing data should also take part in training data after some fix time.

REFERENCES

[1]. Octave Iradukunda, Haiying Che, Josiane Uwineza, Jean Yves Bayingana, Muhammad S Bin-Imam, Ibrahim Niyonzima "Malaria Disease Prediction Based on Machine Learning" School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China (2019).

[2]. G. Arutjothi, Dr. C. Senthamarai "Prediction of Loan Status in Commercial Bank using Machine Learning Classifier" department of computer applications government arts college (Autonomous) Salem, India (2017.)

[3]. Mohammad Ahmad Sheikh, Amit Kumar Goel, Tapas Kumar "An Approach for Prediction of Loan Approval using Machine Learning Algorithm" School Of Computer Science And Engineering Galgotias University Greater Noida,India (2019).

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 3, May 2023

[4]. Xin Li, Xianzhong Long, Guozi Sun, Geng Yang, and Huakang Li "Overdue Prediction of Bank Loans Based on LSTM- SVM"Jiangsu Key Lab of Big Data and Security and Intelligent Processing Nanjing University of Posts andTelecommunications, Nanjing, 210023, China.

[5]. Aakanksha, Tamara Denning, Vivek Srikumar, Sneha Kumar Kesera "secrets in source code: reducing false positives using ML" software engineering (Microsoft) school of computing, USA (2020)