Customer Loan Prediction Using Supervised Learning Technique

L. Udaya Bhanu[1], Dr. S. Narayana[2]

[1]M.Tech Student, Dept. of Computer Science & Engineering, Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh, India
[2]Professor&Mentor, Dept. of Computer Science & Engineering, Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh, India

DOI: 10.29322/IJSRP.11.06.2021.p11453
http://dx.doi.org/10.29322/IJSRP.11.06.2021.p11453

Abstract- Customer loan prediction is usually life time issue so; each and every retail bank faces the issue at the minimum lifetime. If done exactly, it can spare a lot's of man hours at the conclusion of a retail bank. If Company wants to semi automate the loan acceptability process (real time) based on customer detail provided while filling online application form. These subtle elements are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. To automate this method, they have given an issue to recognize the customers segments; those are allowed for loan amount total so they can clearly target these customers. We need to predict whether or not a loan would be approved. In a classification problem, we need to predict separate values based on a given set of self-sufficient variable(s). What's our objective is to implement machine learning model so as to classify, to the best doable degree of accuracy, and dataset gathered from Kaggle. Random forest classification method shows best accuracy in classifying given on loan candidates using python help on Jupyter notebook.

Index Terms- 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 wherever it's. Now a day's several banks/financial agencies approves loan after a relapse method of verification and validation however still there's no surety whether or not the chosen candidate is the worthy right candidate out of all candidates. Through this method we are able to predict whether that particular candidate is safe or not and the whole method of validation of attribute is automated by machine learning technique [8][6]. The disadvantage of this model is that it emphasizes completely different weights to every issue however in reality sometime loan can be approved on the premise of 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 the candidate. 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

Random forest is ensemble learning method for both classification and replaces issues. The advantage of random decision forest is reduce over fitting and helps to improve the accuracy and runs efficiently on a large datasets and work on both continuous and categorical values and predict analysis of data with help of test data.

Bhoomi Patel, Harshal Patil, Jovita Hembram, Shree Jaswal are 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].

Xin Li, Xianzhong Long, Guozi Sun, Geng Yang, and Huakang Li This paper mainly introduces the main application of LSTM-SVM model in user loan risk prediction, and elaborates the current economic background, 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 in this paper actually has few limits and needs to be improved in future research [7].

Aakanksha, Tamara Denning, Vivek Srikumar, Sneha Kumar Kesera[8] this paper is mainly used for voting classifier (combination of logistic regression, naïve bayes, SVM). They able

to reduce the number of FP considerably. This work represents the group of generic passwords to reduce misclassification. Arutjothi [9] present a new credit scoring model, which depends on the hybrid feature selection model and C4.5 classifier. This is depend on hybrid system not only has a strong mathematical basis, but also has higher accuracy and more benefits.

Mrunal Surve, Priya Shinde, Sandip Pandit, Pooja Thitme and Swati Sonawane in this paper, they mainly focus to identify and analyze the risk in giving a loan of commercial banks. To analyze risk in giving loan they have used data mining techniques. It includes analyzing and processing information from various agency/assets and summarize into valuable information [12]. They have used C4.5 classification algorithm for predicting the risk percentage for an individual to give loans.

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 for that candidate eligible for loan or not. The main purpose of the system is applicant loan approved or not based on train models [6]

IV. PROPOSED MODEL

In Machine Learning, we are using semi-automated extraction of knowledge of data for identifying whether a loan would be approved or not [6][8]. Classification could be a supervised learning within which the response is categorical that's its 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 we try to develop user interface flexibly graphics concepts in mind, associated through a browser interface. Our goal is to implement machine learning model so as to classify, to the best potential degree of accuracy, master card fraud from a dataset gathered from Kaggle. once initial knowledge exploration, we have a tendency to knew we might implement a random forest model for best accuracy reports.

Random forest, as it was a good candidate for binary classification. Python sklearn library was used to implement the project, We used Kaggle datasets for Credit card fraud detection, using pandas to data frame for class ==0 for no fraud and class==1 for fraud, matplotlib for plotting the fraud and non fraud data, train_test_split for data extraction (Split arrays or matrices into random train and test subsets) and used Logistic Regression machine learning algorithm for fraud detection and print predicting score according to the algorithm. Finally Confusion matrix was plotted on true and predicted.

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. They are:

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

For this process we have mainly two methods:

a. Probabilistic model selection
b. Resampling methods

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

b. Preprocessing

Data mining methods are used in preprocessing for normalize the data which is collected from kaggle. There is 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 method for reduce the null values then recover the data with help 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. The range is the distinction between the first most and original minimum. It preserves the shapes of the first original distribution.

```
(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)
```

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.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

d. Machine Learning Methods

Machine learning is a subset of AI that trains machines with vast 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 kind of IF-THEN rules. This model is associate degree extension of C4.5 classification algorithms represented by Quinlan.

Random Forest

Random forests are a classifying learning framework for characterization (and backslide) that work by building a very large number of Decision trees at planning time and yielding the class that's the mode of the classes surrender by individual trees.

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 dimensional attribute 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 able to 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 machine learning 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 grows [5].

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

5. EXPERIMENT AND RESULTS

A. 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 then evaluate the confusion matrices(accuracy, precision, recall, f1-score) and finally model building i.e., used methods Procedures are designed to detect errors in data at a lower level of detail. *Data validations* have been included 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 user should 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 designed to speak effectively with the user. The system has been designed with popup menus.

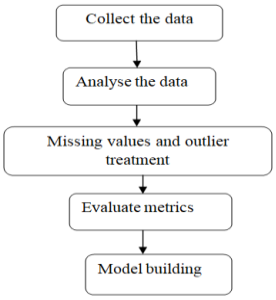


Fig (A): overview of experiment

B. 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.

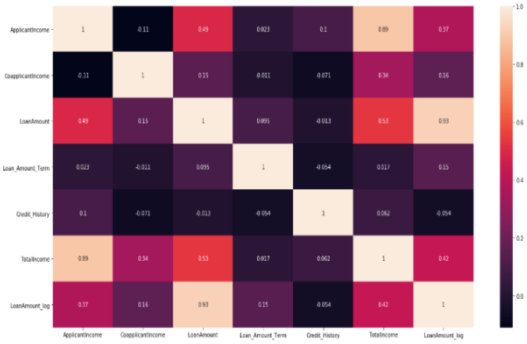
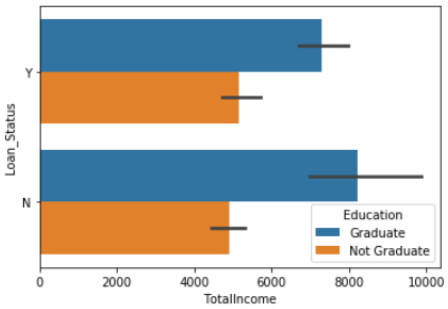


Fig (B): Heat map

C. 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 uncertainty around that estimate using error bars.

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



D. 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

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

V. EVALUTION MODELS

Need for confusion matrix:

Classification (predict category) models have multiple output categories. Most error measures will tell us the total error in our model but we cannot use it to find out individual instances of errors in our model. Confusion matrix helps us identify the correct predictions of a model for different individual classes as well as the errors. The main matrix:

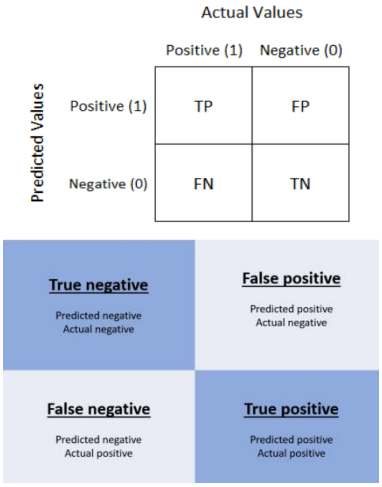


Fig (a): Confusion matrix

- **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 to classify positive values correctly.

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

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

Fig (c): Precision

- **Recall:** To calculate the models ability to predict positive 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 define problem of group of predicted vales.

VI. 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 and meeting to all or any Banker necessities. This part is often simply 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 can be integrated with the module of machine-driven processing system.

VII. 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] Yu Jin and Yudan Zhu, "A data-driven approach to predict default risk of loan for online Peer-to-Peer (P2P) lending," School of Information, Zhejiang University of Finance and Economics, 310018 Hangzhou, China.

[2] https://www.kaggle.com/telco-churn

[3] Bhoomi Patel, Harshal Patil, Jovita Hembram, Shree Jaswal "Loan default forecasting using data mining" Department of Information Technology, St. Francis Institute of Technology, Mumbai, India (2020)

[4] 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).

[5] 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.)

[6] 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).

[7] 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 and Telecommunications, Nanjing, 210023, China.

[8] 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)

[9] Arutjothi .G, Dr. C. Senthamarai. "Credit Risk Evaluation using Hybrid Feature Selection Method. Software engineering and technology (2017)

[10] Ch. Balayesu and S Narayana, "An Improved Algorithm for Efficient Mining of Frequent Item Sets on Large Uncertain Databases" in International Journal of Computer Applications, Volume 73, No. 12 July 2013, Page No. 8-15.

[11] Bala brahmeswara kadaru et al."A novel ensemble decision tree classifier using hybrid feature selection measures for parkinson's disease prediction", Int. J. Data science (IJDS), ISSN: 2053-082X, Vol.3, No.4,2018.

[12] Mrunal Surve, Pooja Thitme, Priya Shinde, Swati Sonawane, and Sandip Pandit. "Data mining techniques to analyze risk giving loan (bank)" Internation Journal of Advance Research and Innovative Ideas in Education Volume 2 Issue 1 2016 Page 485-490

AUTHORS

First Author – L. Udaya Bhanu, M.Tech Student, Dept. of Computer Science & Engineering, Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh, India
Second Author – Dr. S. Narayana, Professor&Mentor, Dept. of Computer Science & Engineering, Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh, India