Loan Prediction Model Using Logistic Regression, Decision Tree, Random Forest

[1]Prof. Satish Jaywant Manje, [2]Mr. Dhiraj Raghunath Manje, [3]Mr. Rahul Pandurang Bhere, [4] Mr. Rahul Kailas Pawade

[1]Asst. Professor,[2,3,4]UG Student,[1,2,3,4]Computer Engg. Dept. Shivajirao S. Jondhle College of Engineering & Technology, Asangaon, Maharashtra, India. [1]satishmanje93@gmail.com, [2]996dhirajmanje @gmail.com, [3]rahulbhere60@gmail.com, [4]rahulpawade003@gmail.com

Abstract- Currently in banking sector, the main profitable parameter compared to others is on its credit line. Giving loans to customers and collecting it with given interest rate is important process in order to gain profit or a loss depends large extend to customer will payback or not. So in order to reduce the risk, there need to predict whether the customer is loan defaulter to reduce banks non-profitable assets. The profit can be maximized by achieving accurate predictions hence the comparison of different approaches are needed. These very important approaches in predictive analytics are prepared to manage the prediction of loan defaulters: The Logistic regression, Decision Tree, Random forest.

The selected three models are better as compared to others because the account some includes variables such as personal attributes not like others considering financial status of applicant should be taken into account to calculate probability of loan approval. Therefore, by using appropriate approaches, the right customers to be prioritized for granting loan can be easily detected by evaluating their chances so called eligibility of default on loan.

Keywords:- Loan, Prediction, Logistic regression, Decision tree, Loan defaulters, Random Forest

I. INTRODUCTION

In India, peoples are highly applying for loans due to certain reasons. It is being hard for bank employees to check and predict whether the customer trustworthy to approve the loan with interest rate. The Data analysis helps to reduce the complexity of data and give appropriate results helpful for user. The data analysis technique used to analyze the behavior of data is quite exploratory.[1], [2]

The main motive behind the paper is to identify the nature of loan applicant; Loan prediction system is very useful tool for the bank-employee and also to the customer. It reduces the risk factor. Loan prediction system provides results with the help of trained models for loan approval. The datasets used in purpose to train the models are the past records collected.[2],[3]

Following very important approaches in loan predictive analytics are used to achieve the prediction of loan eligibility:

1. Logistic Regression Model
2. Decision Tree Model
3. Random Forest Model

II. AIM AND OBJECTIVE

a) Aim

As the improvement in banking sector, count of loan applicants are increasing but due to limited fund available in banks; it is being necessary to identify who is safer to approve the loan. It will be really helpful to both bank and applicant. Hence the primary aim of our paper is to minimize the risk present in approving loan to applicant. Therefore, by using different machine learning approaches, the right customers to be targeted for granting loan can be easily detected by evaluating their likelihood of default on loan.[2],[3]

b) Objective

Predictive Analysis:

The main objective of the paper is to create a prediction model for loan approval and to classify that an applicant applying for loan is eligible or not or person is loan defaulter using data collected by bank employee.

Risk Minimization:

In present, the counts of people applying for the loans are incrementing due to financial conditions. It is being hard for employee to select the rightful applicant for loan. Hence risk behind the loan approval can be minimized with the usage of these prediction models.

III. LITERATURE SURVEY

Paper 1: Loan Approval Prediction with the help of Machine Learning Approach:

In this paper they tried to minimize this risk factor behind approving loan to the safe person so as to save lots of bank efforts and money. With the help of these past records/experiences by identifying the behavioral pattern of data; the machine was trained using ML models to give the most accurate result.[2],[4]

Paper 2: Loan Prediction using Machine Learning Models:

In this given paper, they introduced about every fundamental process occurred in predictive analysis. Data cleaning and processing were the initial stages of analysis. The best accuracy for the given test set was 81.1%.The approval of loan was mostly depends on the credit history of candidate. Less the credit history greater the chances of not being approved. Those applicants who were applying for low amount but having high income get easily approved. [2].

Paper 3: Probabilistic and predictive approach using logistic regression: prediction of loan approval:

In this given paper, they have implemented Logistic regression models as Predictive analysis tools. It is used for given problem of prediction of loan approval. simply, It gives the prediction about whether the customer is trustworthy to approve the loan or not as per given information using logistic regression approach. Certain limitations were also introduced i.e. it requires large data sample for parameter estimation, unable to provide continuous output. [3].[8]

IV. EXISTING SYSTEM

Loan prediction model of existing system is a powerful tool for a range of possible circumstances. Convenient but still had some limitations to it. The history of prediction model is long and there had been many obstacles in its evolution. Although constantly increasing variations and evolution in technology managed to overcome many obstacles. The different prediction models use different approaches on dealing with these limitations. There are two important observations to make about existing system of loan prediction. For the prediction of loan approval, only logistic regression approach had been used. [1],[4]

V. COMPARATIVE STUDY

Sr. No.	Paper Name	Author/Publication	Technology	Advantages	Disadvantages
1.	Loan Approval Prediction with the help of Machine Learning Approach	Kumar Arun, IOSR	Decision Trees,RandomForest,Support Vector Machine,LinearModel,Neural network	Six ML based classification models have been used for prediction	The system is trained on old training dataset.
2.	Loan Prediction using Machine Learning Models	Pidiketi Supriya	Decision tree algorithm	The best accuracy on dataset test is 0.811.	In this model only one method used for prediction.
3.	Probabilistic and predictive approach using logistic regression : prediction of loan approval.	Ashlesha Vaidya/ Computer Science Engineering SRM University,	logistic regression, decision trees,Artificialneural networks(ANN) and BayesianNetworks.	Logistic regression is widely used in data analytics where analyzing of the pre existing data within all kinds of organization is required.	Logistic regression requires a large sample for parameter estimation.

Table no. 01 – ***Comparative analysis***

VI. PROBLEM STATEMENT

Problem of existing is the less accuracy for predictive analysis. Only one prediction model is available to predict the loan approval. Also prediction credibility is a challenge not only in the field of prediction. A prediction model can be useful for banks if the probability is gives and result it delivers is applicable to real world facts. The prediction model developer can never be one-hundred percent sure if this is the case.

VII. PROPOSED SYSTEM

Basically, the proposed system gives the predicted value a customer with details then the person is eligible or not for loan, by using the right information provided by applicant as input in real time or collected by bank employee. Proposed system's prediction model includes Decision tree model & Random forest model along with Logistic regression model present in existing system.

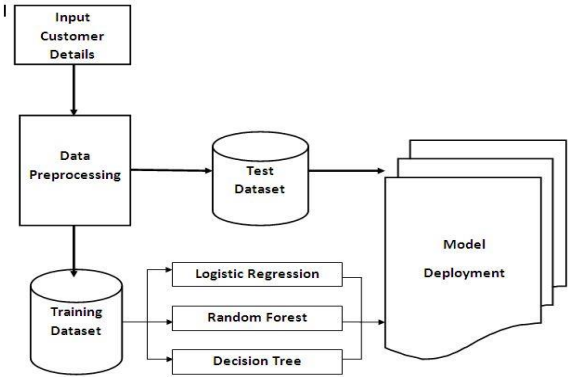

Fig 01. ***System overview***

The Simple GUI is very helpful for naïve users like customer or applicant and bank employee hence also we have created an UI using the Node-red software and Flask for the loan status prediction, this UI will allow the users to predict the loan status very easily and the User interface is user friendly not at least one complication in using the interface, and it can be used just by entering some necessary details into the UI in real time it'll give the predicted value like if the customer is useful to the customer is eligible or not and to give some scenarios about eligibility.

VIII. ALGORITHM

Step 1: Start

Importing libraries;

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

Step 2: loading dataset

data←pd.read_csv("loan_data.csv")

Step 3: Taking care of null values or All null values removed

data.apply(lambda x: sum(x.isnull()),axis=0)

Step 4: Data visualization

sns.pairplot(data)sns.heatmap(data.corr(), annot = True)

a←data["Gender"].value_counts().to_numpy()

b←data["Married"].value_counts().to_numpy()

c←data["Dependents"].value_counts().to_numpy()

Step 5: Analyzing the data

X ← data.iloc[:, 1: 11].values

y ← data.iloc[:, 11].values

Step 6: Label Encoding

From sklearn.preprocessing

import LabelEncoderle = LabelEncoder()

fori in range(0, 5):

X[:,i] ← le.fit_transform(X[:,i])

X[:,9] ← le.fit_transform(X[:,9])

y ← le.fit_transform(y)

OneHotEncoding *fromsklearn.preprocessing import OneHotEncoder*

one ← OneHotEncoder()

z ← one.fit_transform(X[:,9:11]).toarray()

X ← np.delete(X, 9, axis ← 1)

X ← np.concatenate((z,X), axis = 1)

Step 7: Splitting into train and test

From sklearn.model_selection

import train_test_split

X_train, X_test, y_train, y_test ← train_test_split(X, y, test_size = 1/3, random_state =0)

Step 8: Feature Scaling

from sklearn.preprocessing import StandardScaler

sc ← StandardScaler()

X_train ← sc.fit_transform(X_train)

X_test ← sc.fit_transform(X_test)

Step 8: Use all of three machine learning models to train models:

#Fitting all ML models to the Training set

From sklearn.tree import DecisionTreeClassifier

From sklearn.linear_model import LogisticRegression

from sklearn.ensemble import RandomForestClassifier

dt ← DecisionTreeClassifier(criterion ← 'entropy',random_state ← 0)

lr←LogisticRegression(random_state=0)

rf ← RandomForestClassifier(n_estimators = 10, criterion = 'entropy', random_state = 0)

dt.fit(X_train, y_train)

rf.fit(X_train, y_train)

lr.fit(X_train, y_train)

Predicting the Test set results

y_pred ← classifier.predict(X_test)y_pred

Measuring Accuracy

fromsklearn import metrics

print('The accuracy of model is: ', metrics.accuracy_score(y_test, y_pred))

Making confusion matrix

fromsklearn.metrics import confusion_matrixprint(confusion_matrix(y_test, y_pred))

Step 9: Evaluation of models

fromsklearn.metrics import r2_score, mean_absolute_error

foralgo,model in fit_models.items():

yhat ← model.predict(X_test)

Step 10: launch the web application.

Step 11: Enter the values for prediction.

Step 12: Result display you are eligible or not for loan approval prediction & as well as you get some tips regarding bank loan.

Step 13: Stop

IX. MATHEMATICAL MODEL

1. Logistic Regression

The Logistic regression is a mathematical approach that used in the describing the relationship between some

independent variable to a numerous dependent in variable or a dichotomous dependent. The regression function is a employed because of the proposed covariates are combination of continuous and categorical random variable , whereas the dependent variable default is a dichotomous. [1],[7]

Mathematical equation for sigmoid function used in logistic regression is

$S(x) = 1/(1+e^{-x})$

Algorithm:

Step 1: Data preprocessing

Step 2: Fitting Logistic Regression to our training set

model import LogisticRegression

lr←LogisticRegression(random_state=0)

lr.fit(X_train, y_train)

Step 3: Testing the model

pred ← lr.predict(X_test) y_pred

Step 4:Measuring Accuracy

fromsklearn.metrics import accuracy_score

accuracy_score(y_test,y_pred)

Step 5:import sklearn.metrics as metrics

fpr,tpr,threshold ← metrics.roc_curve(y_test, pred)

roc_auc ← metrics.auc(fpr, tpr)

plt.title("Logistic Regression")

plt.plot(fpr,tpr,'b',label = 'auc = %0.2f'%roc_auc)

plt.legend(loc = 'lower right')

plt.plot([0,1],[0,1],'r--')

plt.xlim([0,1])

plt.ylim([0,1])

plt.ylabel('tpr')

plt.xlabel('fpr')

2. Decision Tree

In decision tree approach, it builds the classification models to acquire rules like classification gradient model. The rules are like data featuring, creating a tree structure and decision nodes which are related to attributes. To produce the purest node, After splitting criterion of the model, the attribute with best score will be chosen as the purest node in given model. hence the derivation of the root node for the subsequent is done.[5],[6]

Algorithm:

Step 1: Data preprocessing

Step 2: Training the Decision Tree Classification with the help of Training set

from sklearn.tree import DecisionTreeClassifier

dt ← DecisionTreeClassifier(criterion = 'entropy', random_state = 0)

dtr.fit(X_train, y_train)

Step 3: testing the prediction results

pred ← dtr.predict(X_test) y_pred

Step 4: Measuring Accuracy

from sklearn import metrics

print('The accuracy of Decision Tree Classifier is: *'metrics.accuracy_score(y_test, y_pred))*

Step 5: import sklearn.metrics as metrics

fpr,tpr,threshold ← metrics.roc_curve(y_test, y_pred)

roc_auc ← metrics.auc(fpr, tpr)

plt.title("Decision Tree Curve")

plt.plot(fpr,tpr,'b',label = 'auc = %0.2f'%roc_auc)

plt.legend(loc = 'lower right')

plt.plot([0,1],[0,1],'r--')

plt.xlim([0,1])

plt.ylim([0,1])

plt.ylabel('tpr')

plt.xlabel('fpr')

3. Random Forest

In this model Random forest they have their own features of mathematical model. For each Random forest, the libraries i.e. Scikit-learn, scikit-py will calculate a node importance. The Gini importance method is used, hence by assuming only two child nodes:

Algorithm

Step 1: Data preprocessing

Step 2: Training the Random forest model with the help of Training set *from sklearn.ensemble import RandomForestClassifier*

rf ← RandomForestClassifier(n_estimators = 10, criterion = 'entropy', random_state = 0)

rf.fit(X_train, y_train)

Step 3:Predicting the Test set results

pred ← rf.predict(X_test) y_pred

Step 4: Measuring Accuracy

metrics.accuracy_score(pred, y_test))

Step 5:*import sklearn.metrics as metrics*

fpr,tpr,threshold ← metrics.roc_curve(y_test, y_pred)

roc_auc ← metrics.auc(fpr, tpr)

plt.plot([0,1],[0,1],'r--')

plt.xlim([0,1])

plt.ylim([0,1]).

X. SYSTEM ARCHITECTURE

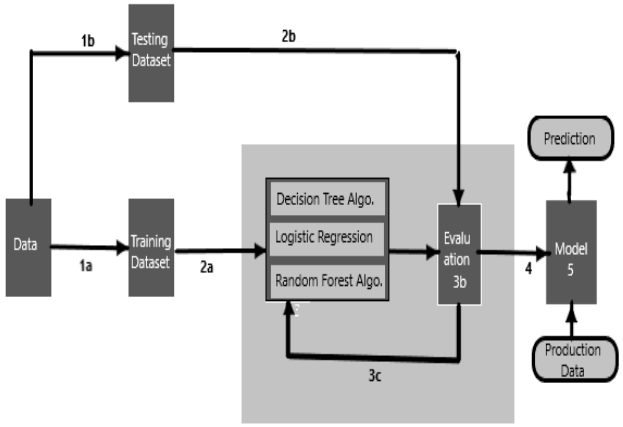

Fig 02. ***System Architecture***

1. Data Collection:

The accuracy of our model is mostly depending on the data this step generally refers to representation of data. The data is collected from well-known website Kaggle, which will be used as training and testing dataset.

2. Data Preprocessing:

Preprocessing include removing null values, empty cells, missing values, conversion of data type and randomization of data to remove preformed in order that helps to stop misleading.

3. Choose a Model:

Multiple machine learning approaches are used to get proper result. In this paper, we have used Logistic regression, Decision tree and also Random forest models. One of them having highest accuracy will be used to give prediction.

4. Training the Model:

Training of the models is done to learn the data patterns and relationship between variables. Larger the dataset for training more evaluation needed to be done and more accurate results it gives.

5. Evaluate the Model:

In evaluation step, the trained models are tested with already separated testing dataset to evaluate the accuracy.

7. Make Predictions:

The information provided by user is used as data to make prediction. Based on the information given the prediction is provided. The prediction is given in real time as soon as the input data is provided.

8. Display Prediction result (Web Application):

An end to end web application is developed to predict the Loan output. The web application must be built with Node-red in JSON or Flask (python) in HTML format with the machine learning trained models.

XI. ADVANTAGES

- Simple and user friendly interface, which is comfortable for naïve user like bank employee to evaluate the loan status of applied customer.
- Logistic Regression give the accurate result of the prediction up to 83% which is the algorithm we used or prediction.
- It is composed using the JSON and Python for the web usage in real time.
- As per the data available, the models give prediction in real time.

XII. DESIGN DETAILS

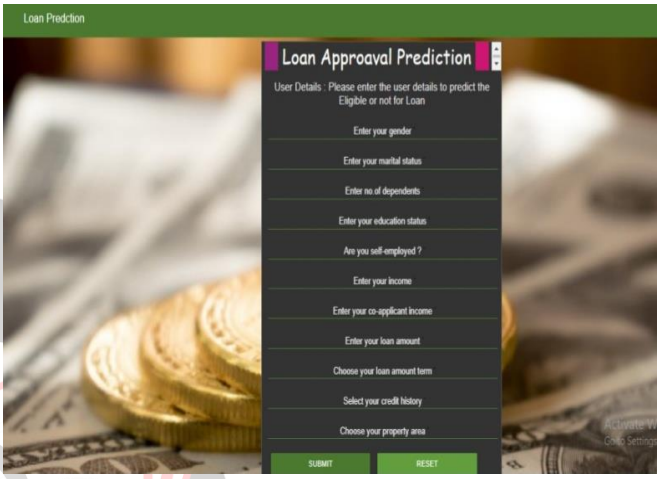

Fig 3. ***Main page***

The web application is created using Node-red in json format and deployed on IBM Cloud using Watson studio of IBM. The user can input all needed details to check eligibility here and check prediction.

XIII. CONCLUSION

Thus, we have tried to implement the paper present by Mohammad Ahmad Sheikh, "An Approach for Prediction about Loan Approval using Machine Learning Algorithm," and according to implementation; the conclusion is that the logistic regression algorithm is most accurate in three of them hence adopted to build a user interface for predicting loan eligibility. The accuracy is compared with other algorithms i.e. random forest and decision tree. It can be seen that the logistic regression algorithm gives high accuracy than the other algorithms in the prediction of loan default and has strong ability of generalization; but still there is no definitive standards that which algorithm should be used.

REFERENCE

[1]. "An Approach for Prediction of Loan Approval about Machine Learning Algorithm," Mohammad Ahmad Sheikh, Proceedings of the International Conference on Electronics and Sustainable Communication Systems (2020).

[2]. PidikitiSupriya, "Loan Prediction using Machine Learning Models," International Journal of Engg. and Techniques, Mar-Apr 2019.

[3]. Ashlesha Vaidya, "Predictive and probabilistic approach by logistic regression : prediction of loan approval," Computer Science Engineering SRM University, Chennai July 3-5, 2017, IIT Delhi

[4]. Kumar Arun, "Loan Approval Prediction based on Machine Learning Approach" IOSR-JCE (NCRTCSIT-2016).

[5]. Nikhil Madane, Siddharth Nanda, "Loan Prediction Analysis using Decision Tree", December 2019.

[6]. Sivasree M.S., Rekha Sunny T., "Loan Credibility Prediction System using Decision Tree Algorithm", (IJERT), 09 September 2015.

[7]. "Prediction of Loan Approval using Machine Learning" Rajiv Kumar, Vinod Jain, Premsagar Sharma, International Journal of Advanced Science and Technology, (2019)

[8]. "Bank loan analysis using customer usage data: A big data approach using Hadoop," 2017 2nd International conference on Telecommunication & Networks.

