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: 7.53
Volume 4, Issue 2, May 2024

Loan Status Prediction with Machine Learning Algorithms using Python

M Ramaraju[1], J Ranjeeth[2], G Sai Rohith[3], M Aditya[4], S Sravan[5], A Sanjay[6]
Assistant Professor, Department of Computer Science & Engineering[1]
UG Student, Department of Computer Science and Engineering[2,3,4,5,6]
Christu Jyothi Institute of Technology & Science, Jangoan, Telangana, India

Abstract: *In the basic environment of a banking system, all the banks have a range of products to sell but the most prominent source of income of most of the banks is mainly dependent on the credit line. So, the earnings come from the interest of that particular loans which are being credited. The profit or loss of a bank primarily depends to a great extent on loans that is whether customers of the bank are paying the loan back or defaulting. By estimating the loan delinquent, the Non-Performing Assets are reduced by the bank. Thus, it makes the study very important for this phenomenon. Various researches done in this era shows that there are so various methods for studying the problem in order to control the loan default. Since the right predictions are very important for the products to be maximized, it is very important to study the nature and structure of the various methods and their comparison. A very important advance in predictive analytics is used to understand the problems of identifying the loan defaulters (i)Data Collection, (ii) Cleaning of Data and (iii)Evaluation of Performance. The test and experiments found that the Naïve Bayes model has a more better performance than other models in terms of loan forecasting*

Keywords: Machine learning, Decision Tree, prediction, Python

I. INTRODUCTION

Loan Prediction is very helpful for bank employees as well as for the applicants. The main motive of the Paper is to provide quick, fast and a very easy way to choose the eligible applicants. Dream housing Finance Company deals in various types of loans. They have presence across all cities, towns and village areas. Applications are first done by customers for loan after that company/bank validates the customer eligibility for getting the loan. They have presence across all cities, towns and village areas. Applications are first done by customers for loan after that company/bank validates the customer eligibility for getting the loan. Then company/bank wants to automate the eligibility loan process (real time) based on details of customers provided while completing the application form. The details include Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and other. This project has taken the previous customers data of various banks to whom on a set of rules loan were approved. So, the machine learning model is done on that method to get approximate results. Our main goal of this project is to predict the loan safety. To predict loan safety, the KNN, logistic regression and decision tree algorithm are used. First the data is cleaned in order to avoid the values missing in the data set. Machine learning (ML) helps to study about the computer algorithms that automatically improves through experience and data use. It falls under the part of artificial intelligence. The algorithms of machine learning help in building a model based on sample data, known as data training, in order to have predictions or decisions without being programmed explicitly to do so. Machine learning algorithms help us in a variety of applications, for example in medicine, filtering of email, recognition of speech, and computer vision, where it is difficult to produce conventional algorithms to perform the task needed.

A machine learning subset is related closely to statistics of computation, which focuses on prediction making with the help of computers; but not every machine learning is statistical understanding of learning. The study of optimization of mathematical delivers methods, theory and domains of applications to the field of machine learning. Data mining[2][8] is the study in related field, focusing on data exploration analysis by unsupervised knowledge gathering. Some applications of machine learning consist of data and neural networks that mimics the biological brain workings. In its application of business problems, machine learning is referred also to as predictive analytics.

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: 7.53
Volume 4, Issue 2, May 2024

Data mining can be used to analyze data from different perspectives and to extract use full knowledge from it. It is a knowledge discovery process. The various steps in extracting knowledge from raw data is depicted in fig.(1) Some of the data mining techniques are classification, clustering, prediction and sequential patterns, neural networks, regression etc. [3]. Classification is most widely used applied data mining technique, which uses a set of pre-classified examples to develop a model which can be used to classify the population of records at large. Fraud detection and credit risk applications are some examples to classification technique. This approach frequently uses Decision tree-based classification Algorithm. In classification, a training set is used to build the model as the classifier which can classify the data items into its appropriate classes. Model validation is done using a test

Fig.1: Steps in knowledge extraction

Certain areas where data mining is used are, banking industry, marketing, risk management and customer relationship management.

It is one of the most common areas of data mining in the banking industry. Marketing department analyses the consumer behavior with reference to product, price and distribution channel. The reaction of the customers to the existing and new products can also be known. Banks uses this information to promote the products, improve quality of products and services, and gain competitive advantages. Bank analysts can analyze the past trends, determine the present demands and forecast the customer behavior of various products and services, in order to grab more business opportunities.

This process is also used form an aging risks in the banking industry. Bank executives must know the credibility of customers they are dealing with. Offering new customers credit cards, extending existing customers‟ lines of credit, and approving loans can be risky decisions for banks ,if they do not know anything about their customers.

Banks provide loans to their customers by verifying the various details relating to the loan, such as amount to loan, lending rate, repayment periodic. Even though, banks are cautious while providing loan, there are chances of customers going to loan default zone. Data mining technique helps to distinguish borrowers who repay loans promptly from those who default.

Risk Management [5] is widely used for managing risks in the banking industry. Bank executives need to know the credibility of customers they are dealing with. Offering new customers credit cards, extending existing customers‟ lines of credit, and approving loans can be risky decisions for banks, if they do not know anything about their customers. Banks provide loans to their customers by verifying the various details relating to the loan, such as amount of loan, lending rate, repayment period etc.

Data mining is used in all the three phases of a customer relationship cycle like customer acquisition, increasing value of the customer and customer retention.

In this project we've got used Streamlit (that's an open- supply app framework in Python language [7]), to construct an internet utility primarily based totally on system learning. Python libraries like Scikit-learn [6], SciPy and plenty of extra are prepared with this framework for the making of this internet utility. Basically, this utility is a totally easy utility which simplest offers the mortgage [4] prediction end result primarily based totally on a few countable enter parameters. One of the number one blessings of the usage of streamlit to increase this utility is that it is able to be shared with all people with the assist of the hyperlink that receives generated every time we run the utility in pycharm terminal. This offers simplicity and area of expertise to the venture.

III. PROBLEM DEFINITION

This project aims a provide a solution to automate a very important process in the field of banking and finance. All over the world banks, housing finance companies deal in a multitude of types of personal and business loans. These companies validate the eligibility of the companies when they apply for the loans. Our project will automate this process by analyzing the online form that customers will fill as a requirement by employing machine learning and

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: 7.53
Volume 4, Issue 2, May 2024

calculate whether the customer is eligible for the loan or not. This form consists of details like sex, marital status, qualifications, details of dependents, annual income, amount of loan, credit history of applicant and others.

IV. MOTIVATION

As it is very difficult to predict the possibility of payment of loan by the customer, using machine learning we'll automate the process. Loan approval is a very important process for banking organizations and recovery of loans is a major contributing parameter in the financial statements of a bank. Therefore, precise calculation and correct approval is very important.

V. PROPOSED MODEL

This system predicts whether the loan is approved or rejected. These are the steps that are employed in the process.

- Data Collection
- Data Pre-processing (Data Cleaning)
- Model Selection
- Model Evaluation
- Classification
- Result (output)

VI. PROPOSED ALGORITHM

The following suggests the pseudo code for the proposed mortgage prediction method

A. Load the statistics:

B. Determine the schooling and checking out statistics

C. Data cleansing and pre-processing.

- Fill the lacking values with imply values concerning numerical values.
- Fill the lacking values with mode values concerning specific variables.
- Outlier treatment.

D. Apply the modeling for prediction

- Removing the burden identifier
- Create the goal variable (primarily based totally at the requirement). In this approach, goal variable is mortgage- status
- Create a dummy variable for specific variable (if required) and break up the schooling and checking out statistics for validation.
- Apply the model: KNN method, SVM method, Logistic Regression, Decision Tree

E. Determine the accuracy observed via way of means of confusion Matrix

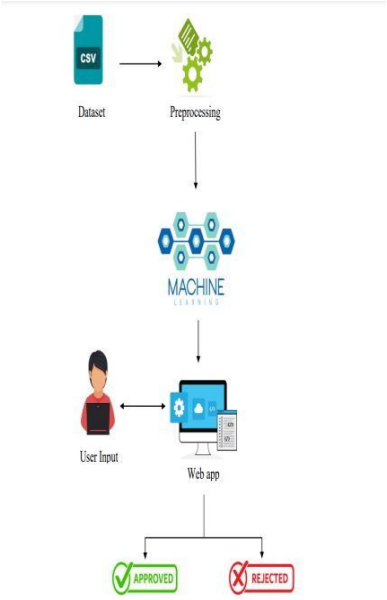
F. According to the confusion Matrix maximum accuracy turned into located for Logistic Regression i.e. , 0.814 ~ 81%.

G. Logistic Regression (Due To the Highest Accuracy) is utilized in an internet App hosted via way of means of streamlit in an effort to take within side the information of clients and display the outcome.

VI. SYSTEM ARCHITECTURE

To predict loan status using machine learning, we first preprocess the data, handling missing values and encoding categorical variables. Next, we fit a Logistic Regression model to the training set using scikit-learn library. This involves importing the Logistic Regression algorithm, initializing a classifier object, and fitting the model to the training data. Once trained, we use the model to predict loan statuses on the test set. To evaluate the model's performance, we calculate accuracy and generate a confusion matrix. The accuracy indicates the proportion of correctly classified instances, while the confusion matrix provides a detailed breakdown of correct and incorrect predictions, allowing us to assess performance across different loan statuses. By following these steps, we can build and assess the effectiveness of a Logistic Regression model for predicting loan statuses.

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: 7.53
Volume 4, Issue 2, May 2024



VII. MODELS USED

Logistic Regression [9]:

In statistics, the logistic version (or logit version) is used to version the opportunity of a sure elegance or occasion present which includes pass/fail, win/lose, alive/lifeless or healthy/sick. This may be prolonged to version numerous lessons of occasions which includes figuring out whether or not an photo consists of a cat, dog, lion, etc. Each item being detected within side the photo could be assigned a opportunity among zero and 1, with a sum of one. Logistic regression is a statistical version that during its simple shape makes use of a logistic feature to version a binary established variable, despite the fact that many extra complicated extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic version (a shape of binary regression). Mathematically, a binary logistic version has a established variable with viable values, which includes pass/fail that is represented with the aid of using a hallmark variable, wherein the 2 values are labeled "zero" and "1".

Decision Tree [10]:-

A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes).

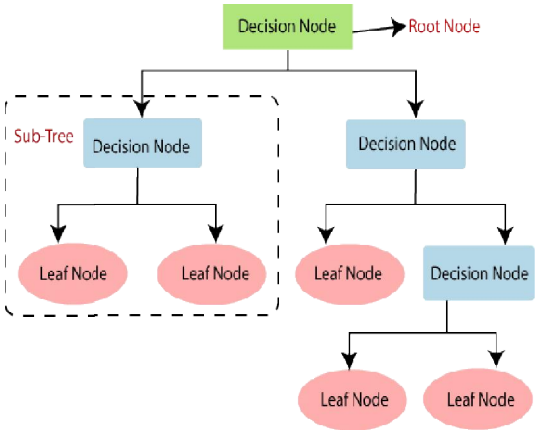


Fig 2: Decision Tree

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: 7.53
Volume 4, Issue 2, May 2024

The paths from root to leaf represent classification rules. In decision analysis, a decision tree and the closely related influence diagram are used as a visual and analytical decision support tool, where the expected values (or expected utility) of competing alternatives are calculated. A decision tree consists of three types of nodes. Decision nodes – typically represented by squares Chance nodes – typically represented by circles End nodes – typically represented by triangles

KNN (K-Nearest Neighbors):

The abbreviation KNN stands for "K-Nearest Neighbor". It is a supervised machine learning algorithm. The algorithm can be used to solve both classification and regression problem statements. The number of nearest neighbors to a new unknown variable that has to be predicted or classified is denoted by the symbol „K". KNN calculates the distance from all points in the proximity of the unknown data and filters out the ones with the shortest distances to it. As a result, it"s often referred to as a distance-based algorithm.

Random Forest Algorithm[11]:

Random Forest is a versatile and powerful ensemble learning algorithm widely used in supervised machine learning tasks such as classification and regression. It operates by constructing a multitude of decision trees during training and outputs the mode (for classification) or average prediction (for regression) of the individual trees.

Each decision tree in a Random Forest is trained on a bootstrap sample of the training data, meaning that each tree is trained on a random subset of the dataset with replacement. Additionally, at each split in the tree, a random subset of features is considered, adding randomness to the model and reducing the risk of over fitting.

Random Forest combines the predictions of multiple decision trees, which helps to improve the model's accuracy and robustness, particularly when dealing with noisy or complex datasets. Furthermore, it provides a built-in mechanism for feature importance estimation, allowing users to identify the most influential features in the dataset.

One of the significant advantages of Random Forest is its ability to handle high- dimensional data with ease, making it suitable for a wide range of applications. It is also less sensitive to outliers and does not require extensive hyper parameter tuning compared to other algorithms.

VIII. IMPLEMENTATION METHODOLOGY

A. Data Collection

The dataset that is collected for predicting the loan failure is put into the training set and the testing set. Generally 8020 proportion is used to break the training set and testing set. The data model which was created using Decision tree is used on the training set and hung on the test take fineness, Test set forecasting is done. Following are the attributes:

Variable	Description
Loan_id	*Unique loan id*
Gender	*Male/Female*
Married	*Applicant*
Dependent	*Number Of Dependents*
Education	*Applicant education (graduate/under graduate*
Self-employed	*Self-employed(Y/N)*
Applicant income	*Applicant income*
CoApplicationince	*Co application income*
Loan Amount	*Loan amount in hundreds*
Loan_Amount_term	*Term of loan in months*
Credit_history	*Credit history meets guidelines*
Property area	*Urban /semi urban /rural*
Loan status	*Loan approved(Y/N)*

Table 3: Variable vs. Description

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: 7.53
Volume 4, Issue 2, May 2024

Preprocessing:

The dataset that is collected may have some missing data that may cause problems and incorrect prediction of the outcome. To get the correct and accurate result the data need to be preprocessed and so it'll have greater effectiveness of the algorithm. We should remove the outliers and we need to convert the variables. We use the chart function to correct these issues.

Train model on training data set:

Next thing we do is to train the model based on the two datasets - training and testing. We divide our train dataset into two tract train and testimony. Then we train the model on this training on this part to make the testimony part easy. In this way, we can validate our sooth sayings as we've the true sooth sayings for the testimony part (which we don't have for the test dataset)

Correlating attributes:

Based on the correlation among attributes it was noticed more likely to pay back their loans. The attributes that are important and individual have to include Property area, education, loan measure, and originally credit History, which is considered as important. Co-plot and Boxplot are used to associate the attributes in Python platform. E. The proposed model is applied: The Proposed model for finding the output, i.e., predicting whether the loan will be ultimately approved, is applied. The process is executed step by step as explained in page no 03. In the next segment we get the results as we finally input the data of a hypothetical customer in our streamlit web app

IX. RESULTS

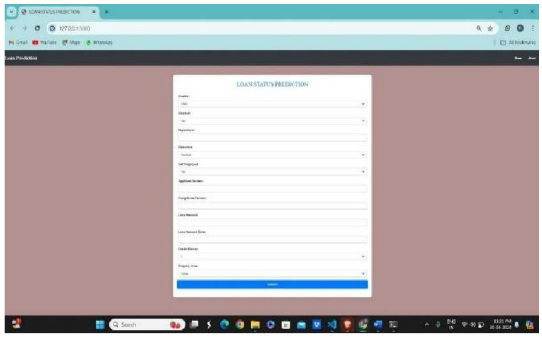


Fig 3: HOME PAGE

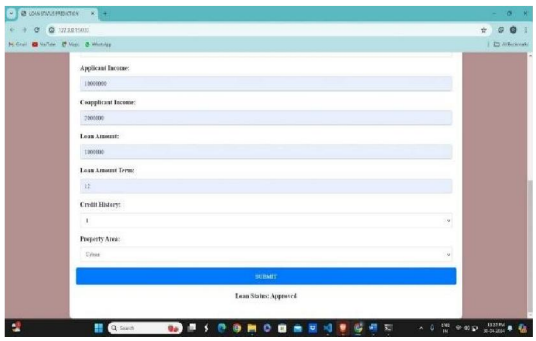


Fig 4: LOAN APPROVED

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: 7.53
Volume 4, Issue 2, May 2024

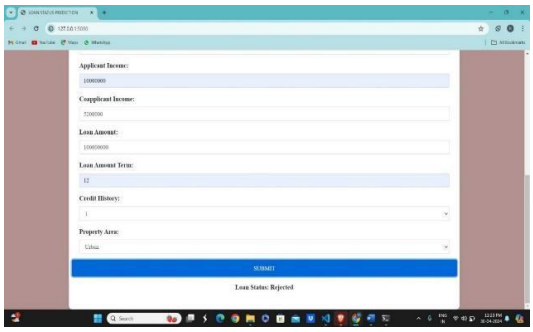


Fig 5: LOAN REJECTED

These results are obtained using Streamlit [12], according to our test and train dataset, which was scaled as per requirements of our data.

X. CONCLUSION

This project helped us to learn about the complicated system of the loan prediction system and the best model that can work with this particular project. It works correctly and fulfills all requirements of bankers. This system properly and accurately calculates the result. It predicts the loan is approve or reject to loan applicant or customer very accurately.

REFERENCES

[1] Kumar Arun, Garg Ishan, Kaur Sanmeet ,"Discuss function of ML in banking system", Loan Approval Prediction based on Machine Learning Approach, Volume 18, Issue 3, Ver. I,e-ISSN: 2278-0661,p-ISSN: 2278-8727,MayJun. 2016.

[2] Shiva Agarwal, "Describe the concepts of data mining", Data Mining: Data Mining Concepts and Techniques, INSPEC Accession Number: 14651878,Electronic ISBN:978-0-7695-5013-8, 2013. [

[3] Aboobyda, J. H., and M. A. Tarig. "Developing Prediction Model of Loan Risk in Banks Using Data Mining." Machine Learning and Applications: An International Journal (MLAIJ)3.1, 2016.

[4] A kindaini, Bolarinwa. "Machine learning applications in mortgage default prediction." University of Tampere, 2017.

[5] Amir E. Khandani, Adlar J. Kim and Andrew Lo, "Consumer credit-risk modelsvia machinelearning algorithms and risk management in banking system",J. Bank Financ., vol. 34, no. 11,pp. 27672787, Nov. 2010.

[6] Aurélien Géron, "Focus on implementing ML programs using the library Scikit-Learn", Publisher – O"Reilly Media, Inc, Edition – Second Edition.

[7] Yuxi (Hayden) Liu, "discuss about the Python Programming", Python Machine Learning By Example, Edition – Third Edition, Publisher – Packt Publishing.

[8] Jiawei Han, Micheline Kamber, Jian Pei , "study the concepts of data mining", Data Mining Concepts and Techniques , Edition- third edition,

[9] Ted Dunning, Ellen Friedman, "discuss logistic regression principles", Machine Learning Logistics, ISBN: 9781491997611Publisher(s): O'Reilly Media, Inc, Released October 2017.

[10] Lior Rokach , "discuss Data mining and decision tree" ,Data Mining with Decision Trees: Theory and Applications: Theory and Applications 81 (Series In Machine Perception And Artificial Intelligence) , Second Edition ,23 October 2014.

[11] Nello Cristianini, John Shawe-Taylor , "discuss SVM principle", An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Hardcover – 23 March 2000.

[12] Tyler Richards, "Implement the streamlit framework",Getting Started with Streamlit for Data Science: Create and deploy Streamlit web applications from scratch in Python, ISBN-13: 978-1800565500ISBN-10: 180056550X ,1st Edition