*Loan Eligibility Prediction Using*

*Ensemble Learning.*

Kotu Lahari  
Department of Computer Science and Engineering

Bapatla Engineering College

(Autonomous)

(Affiliated to Acharya Nagarjuna University)

Bapatla, India

[kotulahari18@gmail.com](mailto:kotulahari18@gmail.com)

Muthangi Sriteja  
Department of Computer Science and Engineering

Bapatla Engineering College

(Autonomous)

(Affiliated to Acharya Nagarjuna University)Bapatla, India  
[sritejamuthangi336@gmail.com](mailto:sritejamuthangi336@gmail.com)

Nayani Madhu Sai Hemanth  
Department of Computer Science and Engineering

Bapatla Engineering College

(Autonomous)

(Affiliated to Acharya Nagarjuna University)Bapatla, India  
[madhunayani12@gmail.com](mailto:madhunayani12@gmail.com)

Kondapalli kalyani  
Department of Computer Science and Engineering

Bapatla Engineering College

(Autonomous)

(Affiliated to Acharya Nagarjuna University)Bapatla, India  
[kalyanikondapalli2@gmail.com](mailto:kalyanikondapalli2@gmail.com)

*Abstract*— In the dynamic landscape of financial services, accurate assessment of loan eligibility remains a critical challenge for lenders. Leveraging ensemble learning techniques, this study aims to develop a robust predictive model for evaluating loan applicants' eligibility based on a comprehensive set of features. Key predictors, including Revolving Utilization Of Unsecured Lines, age, Number Of Time 30-59 Days PastDue Not Worse, DebtRatio, Monthly Income, Number Of Open CreditLines And Loans, Number Of Times 90 Days Late, Number Real Estate Loans Or Lines, Number Of Time 60-89 Days Past Due Not Worse, and Number Of Dependents, are analyzed to enhance prediction accuracy. By harnessing the power of ensemble learning algorithms, such as random forests and gradient boosting machines, this research endeavors to improve the loan approval process and mitigate credit risk. Through rigorous experimentation and evaluation on real-world datasets, the efficacy of the proposed model is assessed using performance metrics such as accuracy. The findings of this study hold the potential to revolutionize lending practices, facilitate informed decision-making, and enhance risk management strategies in the financial industry.

Keywords— Django, Logistic Regression, Decision Trees, Random Forests, Support Vector Machines.

# Introduction

In the contemporary era of banking and finance, the accurate assessment of loan eligibility stands as a pivotal challenge for financial institutions. The ability to precisely determine whether an individual qualifies for a loan not only influences the institution's risk management strategies but also significantly impacts customer satisfaction and retention. In response to this critical need, the "Loan Eligibility Prediction Using Ensemble Learning" project endeavors to employ advanced ensemble learning techniques to enhance the loan approval process. By harnessing the power of ensemble learning algorithms, this research seeks to develop a predictive model capable of effectively evaluating the eligibility of loan applicants based on diverse sets of features such as income, credit score, employment status, and requested loan amount.

This paper presents a comprehensive exploration of the application of various ensemble learning algorithms for loan eligibility prediction. Through rigorous experimentation and analysis, we aim to ascertain the efficacy of different models, including logistic regression, decision trees, random forests, support vector machines, and gradient boosting machines, in accurately predicting loan eligibility. Additionally, we delve into the realm of feature selection and engineering to optimize the predictive power of the models. Furthermore, the study investigates ensemble learning techniques, such as bagging and boosting, to further enhance the performance of the predictive model.

The primary objectives of this research are twofold: firstly, to develop a robust and accurate predictive model for loan eligibility assessment, and secondly, to contribute to the advancement of ensemble learning applications in the domain of finance. Through extensive experimentation on real-world datasets and evaluation using pertinent performance metrics such as accuracy, precision, recall, and F1-score, this study endeavors to provide valuable insights into the development of practical solutions for loan eligibility prediction. Ultimately, the findings of this research hold the potential to revolutionize the loan approval process, facilitate informed decision-making in the financial industry, and pave the way for enhanced risk management strategies.

This paper presents a comprehensive exploration of the application of various ensemble learning algorithms for loan eligibility prediction. Through rigorous experimentation and analysis, we aim to ascertain the efficacy of different models, including logistic regression, decision trees, random forests, support vector machines, and gradient boosting machines, in accurately predicting loan eligibility. Additionally, we delve into the realm of feature selection and engineering to optimize the predictive power of the models. Furthermore, the study investigates ensemble learning techniques, such as bagging and boosting, to further enhance the performance of the predictive model.

# Literature Survey

Literature Survey is the summary of previously done work on topics. Which explains the what work has been done by other authors. The main purpose of this is to understand the existing work and it helps us for our present work.

In the paper [1] proposed by Ugochukwu .E. Orji evaluated the loan default prediction of the Chinese peer-to-peer (P2P) market using R.F, XGBoost, GBM, and Neural Network machine learning models. Their four models exceeded 90% accuracy, with RF being the superior model. This research is closely related to our study in terms of methods used and algorithms deployed.

In the paper [2] proposed by Mohammad Ahmad Sheikh, Tapas Kumar , “An Approach for Prediction of Loan Approval Using Machine Learning Algorithm”. This paper has taken the data of previous customers of various banks to whom on a set of parameters loan were approved. So the machine learning model is trained on that record to get accurate results. Our main objective of this research is to predict the safety of loan to predict loan safety, the logistic regression algorithm is used.

In the paper [3] proposed by Ashlesha Gupta, Anishka Gupta, Vinay Pant Loan, “Approval Prediction System based on Machine Learning Approach”. This model is used for the banking system or anyone who wants to apply for a loan. It will be very helpful in bank management. From the analysis of the data, it is very clear that it reduces all the frauds done at the time of loan approval. Time is also very precious for everyone through this not only the bank but also the waiting time of the applicant will also reduce. As it seems, it will not deal with some special cases when only one parameter is enough for the decision, but it is quite efficient and reliable in some instant. In the future, this prediction module can be more improved and integrated. The system is prepared on the previous training data but in the future, it is possible to make changes to software, which can accept new testing data and should also take part in training data and predict accordingly.

In the paper [4] proposed by Vishal Singh ,Ayushman Yadav, “Prediction of Modernized Loan Approval System Based on Machine Learning Approach” .This Prediction of a modernized loan approval system is incredibly helpful for banks and also the clients. This system checks the candidate on his priority basis. Customer can submit his application directly to the bank so the bank will do the whole process, no third party or stockholder will interfere in it. And finally, the bank will decide that the candidate is deserving or not on its priority basis. The only object of this research paper is that the deserving candidate gets straight forward and quick results.In this system, we take some data from the user like his monthly income, marriage status, loan amount, loan duration, etc. Then the bank will decide according to its parameters whether the client will get the loan or not.

# EXISTING SYSTEMS

## Till now loans are processed by various banks through pen and paperwork.When the large no of customers’ apply for bank loan these bank take lot of time to approve their loan. After approval of loan by the banks,there is no surety that the chosen applicant is capable of paying loan or not. Many banks use their own software’s for the loan approval. In existing system datamining algorithms for the loan approval; this is the old technique for the approval of loan.

### A. Architecture

A diagram of a computer

Description automatically generated

Figure 1: Architecture

Data Collection: The System collect the dataset, initial input

Data Preprocessing: This step involves cleaning and transforming the raw data into a more suitable format for analysis. This may include handling missing values, outliers, and categorical variables.

Data Preparation: After preprocessing, the data is prepared for modeling. This could involve feature scaling, and splitting the data into training and testing sets.

Graphically Analyzing data: This step refers to the visualization of data and results. Graphs and plots can help understand the data and the performance of the model better.

Train and Test: Split the Data into Training and testing data.

Machine Learning: Selecting the Machine Learning Model that is Random Forest Regression.

Result: This is the output of the machine learning model after it has been trained and tested.

User Interface: This is the part of the system that users interact with. A web interface using Django Web Framework, HTML, CSS, and JS.

## Interface

In order to create a user interactive web page we used a framework called Django.

Django is a popular web framework for building web applications using Python. It provides a set of tools and libraries that streamline the development process, allowing developers to focus on creating their application's functionality rather than dealing with repetitive tasks.

## *Algorithms*

Logistic regression is a fundamental and widely used statistical technique for binary classification tasks such as loan eligibility prediction. Despite its name, logistic regression is a classification algorithm rather than a regression algorithm. In logistic regression, the goal is to model the probability that a loan application belongs to a certain class (e.g., approved or rejected) based on input features. The algorithm estimates the probability using a logistic (sigmoid) function, which outputs values between 0 and 1. To make predictions, logistic regression uses a decision threshold (often 0.5). If the predicted probability is above the threshold, the loan application is classified as approved (positive class); otherwise, it's classified as rejected (negative class).

The Naive Bayes algorithm is a popular and effective method for classification tasks, especially in natural language processing and document categorization. Its strength lies in its simplicity and efficiency. The algorithm is based on Bayes' theorem, which describes the probability of a hypothesis given the evidence. In the context of classification, Naive Bayes calculates the probability of each class given a set of features, assuming that these features are conditionally independent. This assumption simplifies the computation and makes the algorithm scalable even with large datasets. To use Naive Bayes for classification, we first calculate the prior probability of each class based on the training data. Then, we compute the likelihood of the features given each class.

The Support Vector Machine (SVM) algorithm is a powerful supervised learning technique used for loan eligibility prediction and various other classification tasks. SVM is particularly effective in scenarios where clear boundaries between classes exist within the dataset.

The Random Forest algorithm is a powerful machine learning technique used for loan eligibility prediction and other classification tasks. It belongs to the ensemble learning methods, which combine multiple individual models (in this case, decision trees) to improve overall performance and generalization.

## Dataset

The data comprises of different factors . Dataset is classified into 10 different parameters which is considered important during the Loan approval process.

It typically includes features such as Revolving Utilization of unsecured lines, age, Number of time 30-59 days past due not worse, Debt ratio, Monthly income, Number of open credit lines and loans, Number of times 90

day late, Number real estate loans or lines, Number of time 60-89 days past due not worse, Number Of Dependents.

|  |  |
| --- | --- |
| **Variable Name** | **Type** |
| **SeriousDlqin2yrs** | **Y/N** |
| RevolvingUtilizationOfUnsecuredLines | percentage |
| age | integer |
| NumberOfTime30-59DaysPastDueNotWorse | integer |
| DebtRatio | percentage |
| MonthlyIncome | real |
| NumberOfOpenCreditLinesAndLoans | integer |
| NumberOfTimes90DaysLate | integer |
| NumberRealEstateLoansOrLines | integer |
| NumberOfTime60-89DaysPastDueNotWorse | integer |
| NumberOfDependents | integer |

Table 1. Data Set Attributes and values

## WorkFlow

* Data Collection: Gather data from various sources including university admission records, student profiles, and university rankings.
* Preprocessing Data : Handle missing values-Impute missing data or remove incomplete records. Also splitting the data.
* Model Selection: Choosing the most suitable machine learning algorithm. In this project we have selected Random Forest .
* Deployment: Deploy the machine learning model as a web application using frameworks such as Django.

# evaluation

To evaluate the performance of the model, you would typically use a separate dataset (e.g., a validation set or a test set) that the model hasn't seen during training. This ensures an unbiased assessment of its generalization capabilities. The most common metrics for evaluating classification models include accuracy, confusion matrix.

**Confusion Matrix:** Confusion Matrix is a performance measurement for the machine learning classification problems where the output can be two or more classes. It is a table with combinations of predicted and actual values. The confusion matrix provides a detailed breakdown of the model's predictions for each class, showing the number of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (TN).

* TP: Instances correctly classified as positive by the model.
* FP: Instances incorrectly classified as positive by the model.
* TN: Instances correctly classified as negative by the model.
* FN: Instances incorrectly classified as negative by the model.

**Accuracy:** Accuracy measures the proportion of correctly classified samples out of the total number of samples. While it's a straightforward metric, it might not be sufficient if the classes are imbalanced. The following Equation (6.1) represents the formula for calculation accuracy.

|  |  |  |
| --- | --- | --- |
|  | Results The below figure shows the loan approval for a person who will repay the loan amount.  A screenshot of a computer  Description automatically generated |  |

##### Acknowledgment

The authors would like to express our sincere gratitude to Mr. V. Naveen Kumar for guidance, support throughout the duration of this project. Thank you to all the research paper authors whose work contributed to the success of our project.

# References

[1] Ugochukwu .E. Orji ; "Machine Learning Models For Predicting Bank Loan Eligibility" 2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON).

[2] Anshika Gupta;Vinay Pant;Sudhanshu Kumar;Pravesh Kumar Bansal2020 9th International Conference System Modeling and Advancement in Research Trends (SMART)]

[3] Mohammad Ahmad Sheikh;Amit Kumar Goel;Tapas Kumar2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)

[4] Raj, J. S., & Ananthi, J. V., "Recurrent neural networks and nonlinear prediction in support vector machine" Journal of Soft Computing Paradigm (JSCP), 1(01), 33-40, 2019.

[5] Aakanksha Saha, Tamara Denning, VivekSrikumar, Sneha Kumar Kasera. "Secrets inSource Code: Reducing False Positives usingMachine Learning", 2020 InternationalConference on Communication Systems &Networks (COMSNETS), 2020.

[6] X.Frencis Jensy,V.P.Sumathi,Janani Shiva Shri, "An exploratory Data Analysis for Loan Prediction based on nature of clients", International Journal of Recent Technology and Engineering (IJRTE),Volume-7 Issue-4S, November 2018.

[7] Pidikiti Supriya, Myneedi Pavani, Nagarapu Saisushma,Namburi Vimala Kumari, k Vikash,"Loan Prediction by using Machine Learning Models", International Journal of Engineering and Techniques.Volume 5 Issue 2, Mar Apr 2019 .

[8] Nikhil Madane, Siddharth Nanda,"Loan Prediction using Decision tree", Journal of the Gujrat Research History,Volume 21 Issue 14s, December 2019.

[9] Vaidya and Ashlesha, Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval, 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT). IEEE, 2017.

[10] Amin, Rafik Khairul and Yuliant Sibaroni, Implementation of decision tree using C4. 5 algorithm in decision making of loan application by debtor (Case study: Bank pasar of Yogyakarta Special Region), 2015 3rd International Conference on Information and Communication Technology (ICoICT). IEEE, 2015.