Loan Eligibility Prediction Using Machine Learning

Kaivalya Gogula, Nagaraju Chattu

***Abstract:** Technology has made many improvements, and the banking industry is no exception. Submission of loan applications by people are so many everyday, making it more difficult for bank to approve loan. To choose an applicant for loan approval, Banks must consider other bank policies also. Based on a few factors, the bank must choose the proposal that has the best probability of getting granted. It would be time-consuming and unsafe to individually check each applicant before recommending them for loan approval. Based on the prior performance of the person to whom the loan amount was previously accredited, we utilize a machine learning technique in this study to forecast the person who is trustworthy for a loan. This will check the whether the applicant is eligible for the loan or not based upon the any previous loan or running loans whether the applicant is paying back the loan within the deadline or not and it will check many other factors to shortlist the applicant is genuinely eligible for loan or not*

***Keywords:** Machine Learning, Loan Approval, Random Forest, Dataset.*

I. INTRODUCTION

In this Modern world loans are one of the important aspects that required for everyone. Banks will get the maximum amount of Profit through loans in the form of interest. There exist many loans like business loan for business purpose and personal loan for their own purpose. The loans [1,2][9][10][11][12][13][14] are classified into two factors based upon the purpose they used for one is secured loan and other is non secured loan. The bank employeesmanage large number of loan applications. To check each and every application manually is difficult and raises the possibilities of mistakes. The majority of the banks makes money through the loan. Here banks have to know the people who can pay the bank in return in time. For this they have their some perspectives [1] which have to be satisfied. It is difficult to choose the deserving customer from the number of applications. Suppose if they sanction the loan to the undeserving customers due to error made by bank employee due to workload. Then bank should suffer the severe loss due to one error because the undeserving customer will not repay the loan. The main aim to this project is to reduce that error instead of checking each and every application manually here we predicting [8] with the model which will developed using. Machine Learning with Python to ease their work and everything is automated this will check the all the loan applications and shortlist applicants whose are eligible for the loan based upon the eligibility criteria The rest of the applications will be rejected. It is an impartial system that saves the bank time by prioritizing the consideration of each application. The timely completion of all other client formalities by the bank authorities benefits the customers. This will save the lot of time for both bank employees and the applicants. With the help of this mechanism, the applications will consider in the form of Priority Basis [8].

Manuscript received on 24 July 2024 | Revised Manuscript received on 20 August 2024 | Manuscript Accepted on 15 September 2024 | Manuscript published on 30 September 2024.

*****Correspondence Author(s)
Kaivalya Gogula, Masters of Science in Computer/Information Technology Administration and Management, St. Francis College, Brooklyn. E-mail ID: kaivalya.gogula@gmail.com
Nagaraju Chattu*, Masters of science in Business Analytics and Information Systems, University of South Florida, Tampa. E-mail ID: nagarajuchattu6@gmail.com

II. LITERATURE SURVEY

An assertion regarding what one anticipates willhappen in the future is known as a prediction. Every day, many predictions [8] are made. While someare very serious and based on mathematics, others are only guesses. Predicting what will happen in the future, whether it be in a few months, a year, orten years, can help us in a number of ways. A subset of advanced analytics known as predictive analytics [5] uses a number of techniques from data mining, statistics, modelling, machine learning, and artificial intelligence to assess current data and produce forecasts [1,2]. In their 2016 study, Kumar Arun [1] explored several ways to predict how a bank will grant a loan. They provided a model that makes use of machine learning tools like neural networks and SVM. Their evaluation of the literature helped us conduct our research and createa trust worthy bank loan prediction model.

III. PROPOSED MODEL

The Process of predicting the loan application [8] is shown in Fig:1. Each and every phase has differentstep shown in fig.

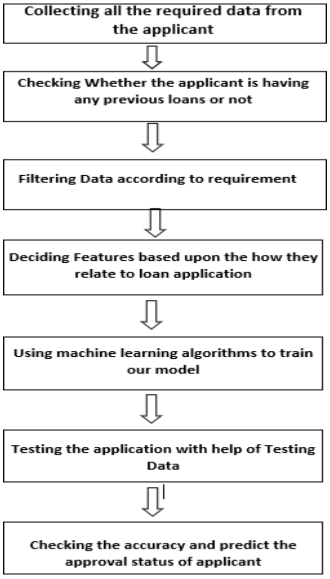

Collecting all the required data from the applicant
Checking Whether the applicant is having any previous loans or not
Filtering Data according to requirement
Deciding Features based upon the how they relate to loan application
Using machine learning algorithms to train our model
Testing the application with help of Testing Data
Checking the accuracy and predict the approval status of applicant

Fig. 1: An Overview of Proposed Methodology

Based upon the Fig:1 whenever the any loan application is received it will gather all the data from the applicant then it will check whether the applicant is holding any pending loan or not after that it will filter the data according to requirement [2]. Then it will check the for what purpose the applicant is applied for the loan then it will validate all the data of the applicant if he meets all the requirements for the loan. Then the loan application will be shortlisted otherwise the loan application will be rejected.

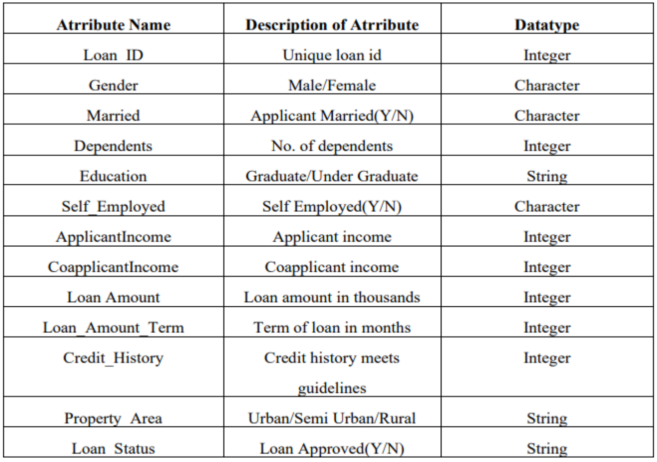
IV. DATASET AND DATA PROCESSING DESCRIPTION

A. Dataset

From this loan prediction system, we are using two different data sets from Kaggle [1,2].

i. First Dataset:

This data set consists of 13 attributes such as gender, loan id, depends etc. First data set attributesare used to validate the business loans [1,2].

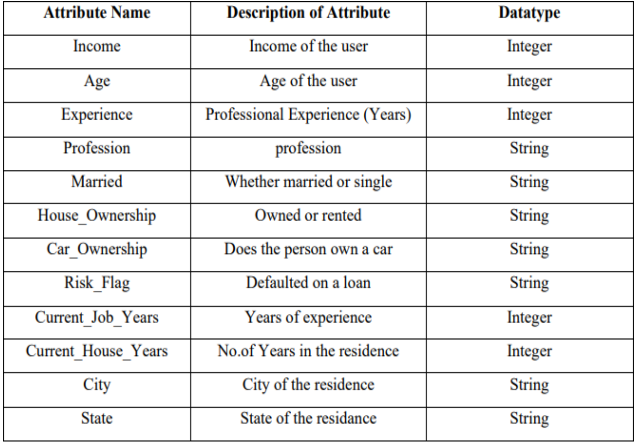


Atrribute Name	Description of Atrribute	Datatype
Loan_ID	Unique loan id	Integer
Gender	Male/Female	Character
Married	Applicant Married(Y/N)	Character
Dependents	No. of dependents	Integer
Education	Graduate/Under Graduate	String
Self_Employed	Self Employed(Y/N)	Character
ApplicantIncome	Applicant income	Integer
CoapplicantIncome	Coapplicant income	Integer
Loan Amount	Loan amount in thousands	Integer
Loan_Amount_Term	Term of loan in months	Integer
Credit_History	Credit history meets guidelines	Integer
Property_Area	Urban/Semi Urban/Rural	String
Loan_Status	Loan Approved(Y/N)	String

Fig. 2: Business Loan Dataset Attribute Names and Information

ii. Second Dataset:

The second dataset consists of the 12 attributes such as Income, Age, Experience, Profession etc. The second data set attributes are used to validate the personal loans [1,2].



Attribute Name	Description of Attribute	Datatype
Income	Income of the user	Integer
Age	Age of the user	Integer
Experience	Professional Experience (Years)	Integer
Profession	profession	String
Married	Whether married or single	String
House_Ownership	Owned or rented	String
Car_Ownership	Does the person own a car	String
Risk_Flag	Defaulted on a loan	String
Current_Job_Years	Years of experience	Integer
Current_House_Years	No.of Years in the residence	Integer
City	City of the residence	String
State	State of the residance	String

Fig. 3: Personal Loan Dataset Attribute Names and Information

B. Data Preprocessing

Data preprocessing include concatenating train and test sets of data and eliminating any unnecessary columns. using isnull () to find the missing values [1,2]. We will infer the missing values after recognizing them. We will now apply mean to null values. Then the iterative imputer will be used to fill in the missing numbers for Loan_Amount and Loan Amount Term and other missing parameter. Now we proceed to map the categorical variables with the integers after imputed all missing values. As the model does not accept any string values, we map the values in order to enter the traininto the model[3].

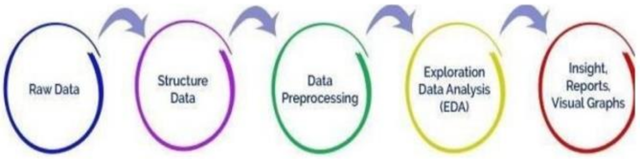

Raw Data
Structure Data
Data Preprocessing
Exploration Data Analysis (EDA)
Insight, Reports, Visual Graphs

Fig. 4: Data Pre-Processing

V. TECHNOLOGY USED

The technology used to develop this project is Machine Learning [3].

A. Machine Learning

Machine Learning [3] is part of the artificial intelligence and computer sciences which focusesabout using data and algorithms to model the waythat humans learn and increase the accuracy of thesystem. Machine learning has been divided into different categories based upon the problems.

i. Supervised Learning:

In the face of uncertainty, supervised machine learning [3] creates a model by employing predictions based on evidence. A model can be trained to generate precise predictions about how it will respond to new data using a known set of input data and known outputs (responses to the data). Use supervised learning when attempting to estimate a result for which there are already known data. By employing methods for classification and regression, supervised learning creates machine learning models.

ii. Unsupervised Learning

It locates latent patterns or underlying structures in unsupervised learning [4][15][16] data. It's employed to remove datasets with input data but no labelled output. Clustering is a well-liked unsupervised learning technique. It is used for exploratory data analysis to find hidden patterns and clusters in the data. A few applications for cluster analysis include product identification, market analysis, and sequence analysis.

iii. Algorithms Used for Prediction:

i. Random Forest:

The Random Forest [6] method improves the adaptability and decision-making capacity of individual trees. Another machine learning technique that incorporates the learning results from various decision trees relies on the ensemble learning theorem. In some use cases of loan and credit risk prediction, some components, or more specifically, those features where removal would improve overall performance, are more important than others.

Given that we are familiar with the principles underlying decision trees and how they choose features based on information acquisition, random forests would take advantage of these advantages to offer better performance.

ii. Logistic Regression Algorithm:

This strategy is used to forecast if the bankwill issue the clients a loan. The model is constructed using classification, followed by Logistic Regression [7]. The use of the sigmoid characteristic is for model growth. The main area where the model spends the most time is pre-processing, which is followed by the analysis of exploratory data, or Function Engineering, and model selection. Two different datasets should be fed into the model before it is applied, followed by Logistic Regression.

iii. Correlation Coefficient Method:

You can determine the link between two quantities with the aid of the correlation coefficient [1]. It provides you with a measurement of how strongly two variables are associated. The Pearson's Correlation Coefficient's value can range from -1 to +1. 1 denotes a strong link between them, while 0 denotes no association.

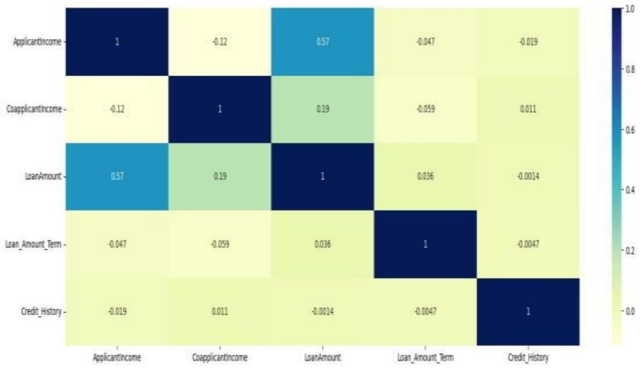
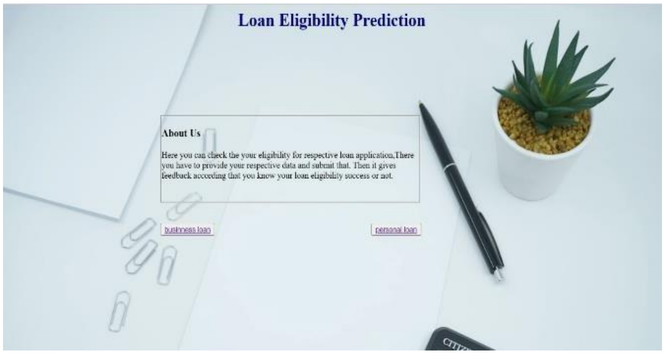


Fig. 5: Correlation

VI. RESULTS

In this Module we validated the application by using the parameters. Below Fig:6 is the home screen of application it will have two type of loan options like business loan and personal loan based upon application requirement bank employee willselect the loan type.



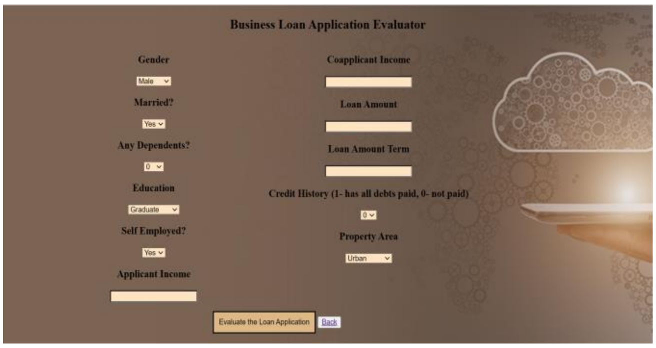
Loan Eligibility Prediction

About Us

Here you can check the your eligibility for respective loan application,There you have to provide your respective data and submit that. Then it gives feedback according that you know your loan eligibility success or not.

Fig. 6: Home Screen

If the bank employee selects the business loan means the window will open like Fig:7 where the employee will enter the applicant parameters andcheck whether the applicant is eligible of the loan or not.

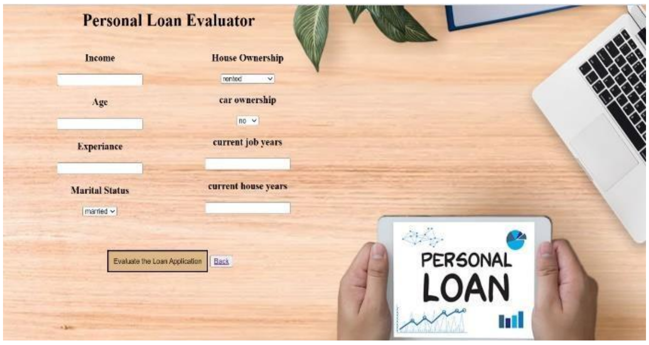


Business Loan Application Evaluator

Gender: Male
Married?: Yes
Any Dependents?: 0
Education: Graduate
Self Employed?: Yes
Applicant Income
Coapplicant Income
Loan Amount
Loan Amount Term
Credit History (1- has all debts paid, 0- not paid): 0
Property Area: Urban

Fig. 7: Business Loan Screen

If the bank employee selects the personal loan in the home screen means the window open like Fig:8 then the employee enters the parameters of the applicant and check whether the applicant eligible for the personal or not.



Personal Loan Evaluator

Income
Age
Experiance
Marital Status: married
House Ownership: rented
car ownership: no
current job years
current house years

Fig. 8: Personal Loan Screen

VII. CONCLUSION AND FUTURE SCOPE

From this project we will conclude that this loan prediction system has been developed using the machine learning algorithms like Random Forest which has accuracy 84 and logistic regression which has accuracy 74 but the random forest will have more accuracy compared to other. Now a days the technology has been changing rapidly by adopting this type technology to the banking system will save the lot of time to both bank employees and customers it will also improve thequality of the work and reduce the human errors and improves the accuracy of the loan prediction.This Project will be very useful to the Banking systems. To develop more accuracy using machine learning algorithms, the work can extend and improved for the automation of loan eligibility prediction by using advanced techniques.

DECLARATION STATEMENT

After aggregating input from all authors, I must verify the accuracy of the following information as the article's author.

- **Conflicts of Interest/ Competing Interests:** Based on my understanding, this article has no conflicts of interest.
- **Funding Support:** This article has not been sponsored or funded by any organization or agency. The independence of this research is a crucial factor in affirming its impartiality, as it has been conducted without any external sway.

- **Ethical Approval and Consent to Participate:** The data provided in this article is exempt from the requirement for ethical approval or participant consent.
- **Data Access Statement and Material Availability:** The adequate resources of this article are publicly accessible.
- **Authors Contributions:** The authorship of this article is contributed equally to all participating individuals.

REFERENCES

1. Kumar Arun, Garg Ishan, Kaur Sanmeet, May- Jun. 2016. Loan Approval Prediction based on Machine Learning Approach, IOSR Journal of Computer Engineering (IOSR-JCE).
2. Dr. K. Kavitha, International Journal of Advanced Research in Computer Science and Software Engineering.
3. K. Hanumantha Rao, G. Srinivas, A. Damodhar, M. Vikas Krishna: Implementation of Anomaly Detection Technique Using Machine Learning Algorithms: International Journal of Computer Science and Telecommunications (Volume2, Issue3, June2011).
4. Clustering Loan Applicants based on Risk Percentage using K-Means Clustering Techniques,
5. Short-term prediction of Mortgage default using ensembled machine learning models, Jesse C.Sealand on july 20, 2018.
6. https://www.ibm.com/in-en/topics/random-forest#:~:text=Random%20forest%20is%20a%20commonly,both%20classification%20and%20regression%20problems.
7. https://www.researchgate.net/publication/357449126_THE_LOAN_PREDICTION_USING_MACHINE_LEARNING
8. https://ieeexplore.ieee.org/document/9336801
9. Nixon, J. S., & Desta, A. W. (2020). Data Mining Application in Predicting Bank Loan Defaulters. In International Journal of Innovative Technology and Exploring Engineering (Vol. 4, Issue 9, pp. 2733–2744). https://doi.org/10.35940/ijitee.d2037.029420
10. Prasad, K. G. S., Chidvilas, P. V. S., & Vasanthamisan, V. K. (2019). Customer Loan Approval Classification by Supervised Learning Model. In International Journal of Recent Technology and Engineering (IJRTE) (Vol. 8, Issue 4, pp. 9898–9901). https://doi.org/10.35940/ijrte.d9275.118419
11. Gupta, R., Gowalker, N., Patil, D. S., & Joshi, Dr. S. D. (2019). Predicting Risk in Sentiment Analysis using Machine Learning. In International Journal of Engineering and Advanced Technology (Vol. 9, Issue 1, pp. 455–460). https://doi.org/10.35940/ijeat.a9540.109119
12. Mukherjee, P., Palan, P., & Bonde, M. V. (2021). Using Machine Learning and Artificial Intelligence Principles to Implement a Wealth Management System. In International Journal of Soft Computing and Engineering (Vol. 10, Issue 5, pp. 26–31). https://doi.org/10.35940/ijsce.f3500.0510521
13. Dubey, S. K., Sinha, Dr. S., & Jain, Dr. A. (2023). Heart Disease Prediction Classification using Machine Learning. In International Journal of Inventive Engineering and Sciences (Vol. 10, Issue 11, pp. 1–6). https://doi.org/10.35940/ijies.b4321.11101123
14. Sharma, P., & Site, S. (2022). A Comprehensive Study on Different Machine Learning Techniques to Predict Heart Disease. In Indian Journal of Artificial Intelligence and Neural Networking (Vol. 2, Issue 3, pp. 1–7). https://doi.org/10.54105/ijainn.c1046.042322
15. Baig, M. A. (2021). An Efficient Cluster Based Routing Protocol (ECCRP) Technique Based on Weighted Clustering Algorithm for Different Topologies in Manets using Network Coding. In Indian Journal of Data Communication and Networking (Vol. 1, Issue 2, pp. 31–34). https://doi.org/10.54105/ijdcn.b5011.041221
16. Patravali, S. D., & Algur, Dr. S. P. (2023). COVID-19 Sentiment Analysis using K-Means and DBSCAN. In International Journal of Emerging Science and Engineering (Vol. 11, Issue 12, pp. 12–17). https://doi.org/10.35940/ijese.l2558.11111223

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)/ journal and/or the editor(s). The Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP) and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.