Machine Learning Algorithms for Predicting the Loan Status

Dr. M Srinivasa Sesha Sai
Department of Information Technology, KKR & KSR Institute of Technology and Sciences, Andhra Pradesh, INDIA, msssai@gmail.com

Ardhala Bala Krishna
Department of Electrical and Electronics Engineering, Malla Reddy Engineering College for Women, Telangana, Hyderabad, ardhalabalakrishna@gmail.com

Parimala Ganthi
Department of Information Technology, KKR & KSR Institute of Technology and Sciences, Andhra Pradesh, INDIA, ganthiparimala001@gmail.com

Sai Keerthi Kankanala
Department of Information Technology, KKR & KSR Institute of Technology and Sciences, Andhra Pradesh, INDIA, keerthikankanala277@gmail.com

Siripriya Tinnaluri
Department of Information Technology, KKR & KSR Institute of Technology and Sciences, Andhra Pradesh, INDIA, siripriyatinnaluri@gmail.com

Vineela Simhadri
Department of Information Technology, KKR & KSR Institute of Technology and Sciences, Andhra Pradesh, INDIA, vineelachinni839@gmail.com

Abstract

Nowadays, more people in India are asking for loans. Whether a consumer or client pays back a loan determines how many loans a bank makes or loses. In the banking industry, improvement is essential. As the process is manual, many mistakes could be made in identifying the real candidate who is applying for a loan for the first time. Each problem has a fix with us. To make our lives easier and to give us a sense of completion, we have machines. By combining different categorization algorithms with historical candidate data, a machine learning model was created. If the application is allowed or not depends on how the system evaluates the candidate's prior performance data. The loan length, loan size, age, and income (if the client would have a job) are the most crucial variables for deciding there. Bankers can better comprehend the client's behavior by looking at the graphs that show the results. On the basis of the data supplied, it is utilised to establish loan eligibility. When the project's results were uploaded to score the predictions, they came back at 0.7986, which is a decent result.

Keywords: *loan recovery, machine learning, categorization algorithms, banking industry, loans.*

I. INTRODUCTION

Loans are currently the most important need in the entire world. Banks receive most of the overall profit through this alone. Every business retails the products, and the bank's product is finances. Banks build money through taking in funds from contributors and other sources and then they will lend money out to consumers. The main part of the bank's profit comes from the loans where it earns more profit. Although the bank has a small amount of money, it obtains loans from consumers in the banking industry. Loans are applied by the people for their needs and to fulfill their necessaries. Some people will apply for the loan for a variety of reasons, such as a business person who wants to grow their company and also for profits. They require a loan in order to deal with the loss of a firm. People in middle-class households seek to meet their necessities, therefore they ask for loans. Customers will apply for loans at the bank every day, and the bank receives a lot of loan applications daily. However, not every application gets chosen. The majority of banks review loan applications using their risk assessment techniques and credit score systems before deciding whether to approve loans for existing customers. Many people are facing difficulty in repaying the loans because of some incidents. So that's why the banks will take all the details of a person whether he/she is payable or not. The Bank reduces the Non-Performing Assets by predicting the loan identifiers. Loan approval of individual consumers' credibility is somewhat difficult to check. It is also a time consuming and risky process. Thus, the bank's aim was to minimize the credit risks of defaulting. To maximize the profits the right predictions are very important. Compared to traditional statistical and Artificial Intelligence techniques, the Machine learning techniques will perform better. It will combine various base models to produce one optimal predictive model in addition to good performance and reliability. Here, we check the new candidate's profile for eligibility of a loan by using Machine Learning with Python to ease the bankers' work.

II. LITERATURE SURVEY

We have gone through numerous research papers to get a clear image of how the machine learning algorithms can predict the genuine applicant. To reduce the significant risk of loss to commercial financial institutions, k-Nearest neighbour scoring model developed using the R language [4]. Min-max normalization along with the 30 an iteration level K-NN model is applied on the publicly available club repository loan data in order to retrieve the pool of both defaulted and non-defaulted loan estimation. Lili lai [3] compared 5 ml models xg boost, adaboost, and multilayer perceptron on loan dataset of Xiamen international bank in which the adaboost and xgboost performed well on the loan default problem whereas, Random Forest (classifier) algorithm, K-nearest neighbor and multilayer perceptron shown weaker output than the others, as a result of overfitting of training dataset.

Singh et al., in 2021 [5], augmented the bank in minimizing the losses and improving the number of credits, by identifying whether the specific type of customer can repay the allocated loan or not, based on their transaction history. This prediction is made by using a system for modernizing loan approval that uses machine learning. [6] successfully validated applicants into four broad categories as 1) An Excellent Credit.2) Good Credit.3) Low Credit.4) and a Bad Credit. Based on these categories the catboost algorithm labels the applicants as approved or rejected. The repayable capabilities of an individual were

retrieved after applying an optical character recognition tool 'pytesseract' to go through the embedded text in the images collected from a typical Kaggle loan dataset.

Patel et al., [7], used a home loan dataset and observed the behavior of each attribute and their impact on target attribute.in further training their model to predict how likely it is that the non-defaulter will turn out to be a fraudster, by training the data using gradient boosting, catboost classifier and logistic regression. Explained the solution through each section of standard default prediction [8] pipeline-data pre-processing, training the model, and the basic architecture of the model, by proposing the use of a federal learning approach.

Initially, to address the class imbalance problem, the Synthetic Minority Oversampling Technique (SMOTE) technique is suggested. Second, concentrate on a default prediction mechanism guarding sensitive data. Security, sensitivity, and the impact of missing data are all mitigated by a decentralized system. The authors are interested in increasing the scalability of the model in the coming days.

A stacking ensemble method is used by Ke et al. [9] for predicting the possible rate of loan repayment of a real provident fund user from user details of the Provident Fund center located in the City of China. A two-layer stacking model was implemented for prediction of loan repayment and to analyze the user behavior. This model had a better result compared to the other 6 approaches in terms of accuracy, the recall score, f1 values and scoring in kappa. A variant of the Grey-Markov approach is implemented [10] to measure and obtain the predictions from the national total loan amount as a reference.

Farmers' microfinance decision-making assessed using support vector machines. [11] Examined the associated conditions, frequent occurrences, and associated likelihood. The base data is obtained through the lending club platform of a huge number of customers. Maheswari, Narayana [12] proposed to improve the analysis of risk identification done by early pre-processing of the past records of user data by building a machine learning model. The major outcomes of the research paper observed that sanctioned loans had higher probability in case of loan default.

The scope of the work describes the relative background of each method [13], its limitations and the criterion for model selection. the base data is collected from analytics vidya competition. classification tools like decision tree, bagging approach along with boosting are implemented to establish a meaningful relation among the attributes. A system to recognise the applicant’s category, who are eligible for a loan amount and specific model, was created by authors in[14]. This model helps the lender in assisting regarding the analysis of situations and usage of better services. The logistic regression algorithm chooses the right person which helps in reducing risk factor.

The kaggle dataset is used for studying and predictions. Support vector machines along with some AI models are used by Yash et al., [15] detects the potential customers to be pointed out for sanctioning the loan by proper evaluation of likelihood of default loan. A high-performing ML learning classifier built using Apache Spark is built [16], which allows the financial organizations to predict the default in already issued loans. This will reduce the economic loss and loan recovery costs which indeed increase the profits.

An application through which the banks can minimize the rate of awful advances by cutting off misfortunes is developed in [17]. By continuous construction and implementation of tasks on a ml application implemented on a cloud-based platform is done in this project along with AI bundle libraries, some python libraries helped in obtaining fruitful examination and highlighting the determined approach.

III. PROBLEM STATEMENT

In many banks, the bankers may consider the credit score and then they can approve the loan to the particular applicant or customer. If the customer is new to applying for a loan, they can face some problems. The financial organization doesn't have the customer's banking-related history for making a decision. So, they must be more careful when choosing a new applicant.

So many financial organizations approve the loan after checking the disk along with the verification and authorization of user data. However, they are not sure they are choosing the right person from the candidates. As are many applicants, it is more difficult for the banks to identify the customer segments that are eligible for the loan amounts to target these customers. Going through the details of the candidates with good scores, they can be easily identified. The essential information of the new applicants is insufficient to choose a good customer among them. So more appropriate data is required.

Machine learning techniques may also result in false predictions like manual approvals. So, this problem requires a more appropriate solution to get good results.

IV. PROPOSED MODEL

In this proposed solution, by using some efficient machine learning algorithms, we will predict whether the client is eligible with a reasonable accuracy rate and decide whether we can approve the loan. The loan prediction model uses different machine learning algorithms to find a trustworthy loan applicant who pays back the loan with interest. It contains some data sets like various attributes provided to the algorithms.

A. Dataset

A dataset is made up of many different pieces of data, however it may be used to train an algorithm to look for predicted patterns throughout the entire dataset. This project dataset contains Loan ID, Gender, Marital status, no. of Dependents, level of Education, Employment, Applicant Income, Co-applicant Income, Applied Loan Amount, Loan Amount Term, Credit History, Property Area, and loan Status.

The non-null count and Data type of each attribute used in the Kaggle dataset are described in the Table-1. All these attributes are used in verification of eligibility process regarding the loan sanctioning.

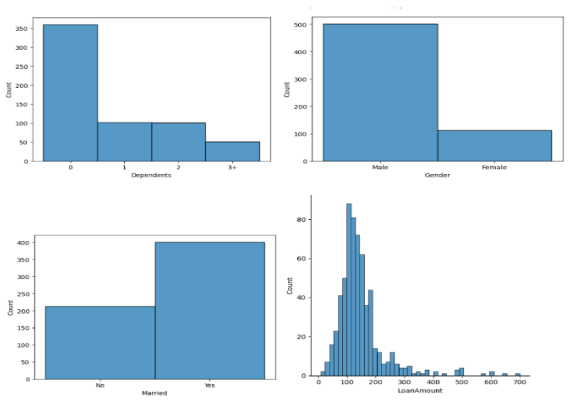
TABLE-1 Attribute Description Table

Column	Non-Null Count	Dtype
Gender	614 non-null	Object
Married	614 non-null	Object
Dependents	614 non-null	Object
Education	614 non-null	Object
Self_Employed	614 non-null	Object
Credit_History	614 non-null	float64
Property_Area	614 non-null	Object
Loan_Status	614 non-null	Object
ApplicantIncomeLog	614 non-null	float64
LoanAmountLog	614 non-null	float64
Loan_Amount_Term_Log	614 non-null	float64
Total_Income_Log	614 non-null	float64

B. Data Pre-processing

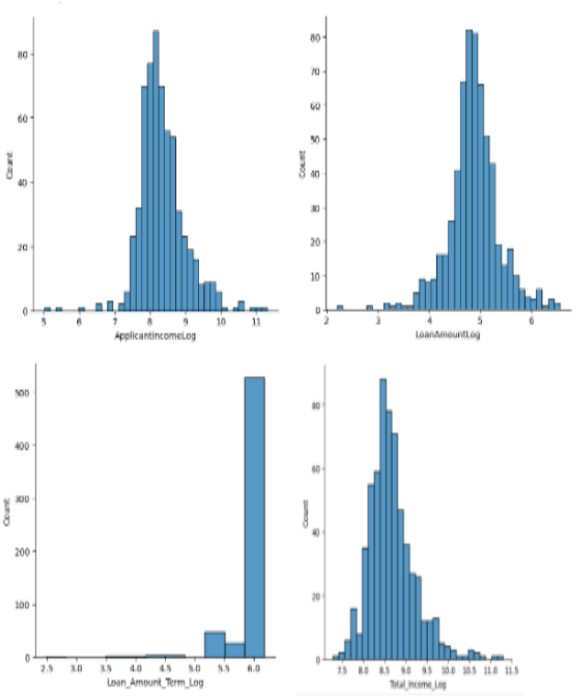
Pre-processing of the dataset is done to handle the numerical missing data and categorical missing data in order to obtain better results. It comprises replacing each group in a categorical feature with the target variable's average value. Handling of missing data is represented in Fig. 1 in the form of bar charts. The missing values of numerical data such as number of dependents, gender, marital status and loan amount is pre-processed and shown in the Pictorial format.

Fig. 1. Handling missing data



Additional columns are added to the existing dataset for handling the numerical and categorical attributes. This is represented in Fig. 2 where, applicant income log, loan income log, loan amount log, total income log are the additional columns created and handled.

Fig. 2. Handling Additional columns



C. Split-Datasets

The input dataset is divided into 2 datasets for training and evaluating the result.

a. Training Dataset

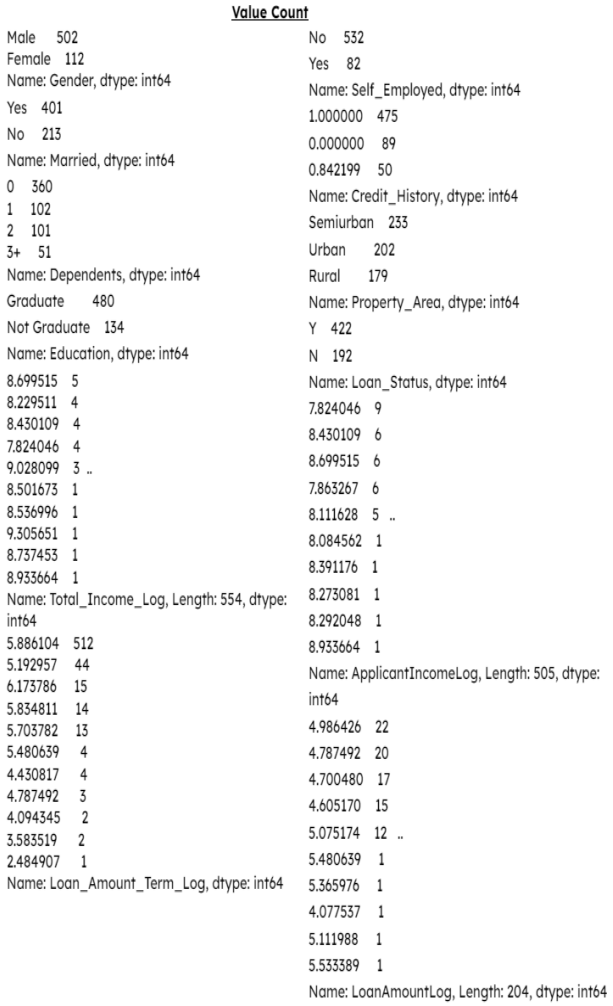
This training set is the biggest one in size and is involved in training or building a machine learning model. This Dataset is pre-processed along with handling of numerical and categorical data.

The dataset is trained on various attributes such as gender of applicant, marital status and the count of the dependents. Further, new columns are created by combining the available columns in the Data processing. Fig. 3 shows the value count of each attribute of training data. Value count of existing data attributes as well as additional columns are calculated and analyzed during the training process.

b. Testing Dataset

Compared to the training dataset, the test dataset contains a different subset of the original data. It utilizes it as a threshold once the model training is complete because it has some similar properties and a similar class probability distribution. Along with the evaluation of performance of the machine learning model based on the trained result. The Testing dataset contains all the attributes similar to training data except the loan status value. This value is predicted after completion of training the machine learning model.

Fig. 3. Value count of Training Data



Value Count

Male 502
Female 112
Name: Gender, dtype: int64

Yes 401
No 213
Name: Married, dtype: int64

0 360
1 102
2 101
3+ 51
Name: Dependents, dtype: int64

Graduate 480
Not Graduate 134
Name: Education, dtype: int64

8.699515 5
8.229511 4
8.430109 4
7.824046 4
9.028099 3 ..
8.501673 1
8.536996 1
9.305651 1
8.737453 1
8.933664 1
Name: Total_Income_Log, Length: 554, dtype: int64

5.886104 512
5.192957 44
6.173786 15
5.834811 14
5.703782 13
5.480639 4
4.430817 4
4.787492 3
4.094345 2
3.583519 2
2.484907 1
Name: Loan_Amount_Term_Log, dtype: int64

No 532
Yes 82
Name: Self_Employed, dtype: int64

1.000000 475
0.000000 89
0.842199 50
Name: Credit_History, dtype: int64

Semiurban 233
Urban 202
Rural 179
Name: Property_Area, dtype: int64

Y 422
N 192
Name: Loan_Status, dtype: int64

7.824046 9
8.430109 6
8.699515 6
7.863267 6
8.111628 5 ..
8.084562 1
8.391176 1
8.273081 1
8.292048 1
8.933664 1
Name: ApplicantIncomeLog, Length: 505, dtype: int64

4.986426 22
4.787492 20
4.700480 17
4.605170 15
5.075174 12 ..
5.480639 1
5.365976 1
4.077537 1
5.111988 1
5.533389 1
Name: LoanAmountLog, Length: 204, dtype: int64

D. Classification Techniques

a. Random Forest Classifier

It is one of the most efficient and less time taking Supervised classification Techniques, that predicts based on the majority of results. Random forest classifier predicts the output with better accuracy and can also maintain even without a large part of data.

This classifier works with very large datasets and many dimensions. Additionally, it increases model accuracy and prevents overfitting.

b. Decision Tree

It is an effective approach for understanding and visualizing the best possible decision options and expected output from the provided data. Decision tree is useful in analyzing risk and rewards regarding each action. It is a graphical representation for finding every possible solution to a problem or option based on predefined parameters.

c. Logistic Regression

One of the Statistical Learning approaches that returns probabilistic value instead of 0 or 1. This is the most effective form of ensemble learning. Logistic regression can be used to quickly pinpoint the variables that will be effective when classifying observations using multiple sources of data.

Machine learning models are used as inputs based on ML techniques and the final method is finalized. The proposed system is tested for further ML models and notes the accuracy.

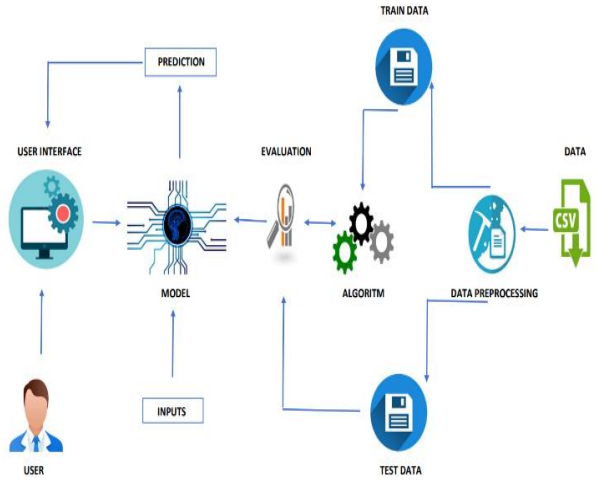
E. Prediction

The Threshold value is finalized. Here the ML model checks the accuracy of each value returned by the algorithms with the threshold values. If the threshold value is greater than the

value of the ML algorithm, that candidate's profile is selected for the loan process. This ML model plays a crucial role in choosing applicants without their credit score or civil rating. This uses a data set with different attributes that decide whether the loan recovery is possible. Fig. 4 depicts the structural approach of proposed ML tactic.

After many checks relating to a loan applicant's features, the 3 ML algorithms return some values that give an accurate payback rate to an applicant. The comparison of Threshold values that our ML algorithm predicts the genuine loan applicant among the vast (huge) no of applications.

Fig. 4. Machine Learning Model Architecture



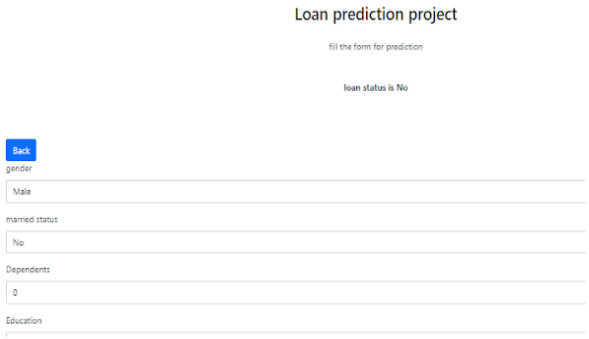
V. RESULTS

The applicant needs to fill all the required fields in the form displayed on the user interface of the loan prediction project. The 'Gender' can be filled as either a male or female. Yes or No should be selected for the 'married status' field. The applicant must specify the number of 'dependents' ranging from 0 to 3+. 'Education' level can be a graduate or non-graduate. The applicant can specify whether he/she is 'self-employed' or not by choosing Yes or no option. 'Property Area' provided along with the application must be among the Urban, Rural and semi-Urban areas.

The remaining fields such as applicant income, co-applicant income loan amount must be numerical values. Credit history of the applicant can be chosen among 1.00, 0.00 and 0.821 values. The loan amount term has to be a numerical value entered by the loan applicant. As a final outcome the 'Loan status' is displayed on the Prediction screen either as 'YES' or 'NO'.

This predicted result is based on the three trained machine learning models' accuracy scores. Fig 6 Resembles the final predicted result obtained. If accuracy of each model exceeds or equals to the fixed threshold value, then the loan eligibility status is predicted as 'YES', otherwise 'NO' will be displayed on the prediction page.

Fig. 6. Loan prediction analysis outcome



VI. CONCLUSIONS

In this project, to predict the loan applicant, we used a loan status model to check whether the customer is valid or not for a valid customer model. Financing institutions or banks can use this model's methodology to decide whether to lend money in response to loan applications. To avoid or reduce the loss of commercial banks, this model will be very helpful to decrease

those losses. To maintain the loan factor process it is designed to show the time period of loan transaction for period. It will be drastically reduced by using this model. Predicting the loan data by using some machine learning algorithms which takes comparatively lower than the time actually needed for a procedure of manners. It could be made more secure, trustworthy and characterized by constant change weight conformation, as we are using trending datasets for reference. In future this model will be trained on old training dataset along with the new ones. Some other distinctive indicators like gender and marital status seem not to be taken into consideration by the model. This project model is an approach that can result in best performance if we participate in all the cost required problems and datasets as input. In our upcoming work, we'll concentrate on adding more dataset attributes so that the upcoming machine learning model can be more efficient.

REFERENCES

[1] A. Gupta, V. Pant, S. Kumar and P. K. Bansal, "Bank Loan Prediction System using Machine Learning," 2020 9th International Conference System Modeling and Advancement in Research Trends (SMART), 2020, pp. 423-426, doi: 10.1109/SMART50582.2020.9336801.

[2] M. A. Sheikh, A. K. Goel and T. Kumar, "An Approach for Prediction of Loan Approval using Machine Learning Algorithm," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020, pp. 490-494, doi: 10.1109/ICESC48915.2020.9155614.

[3] L. Lai, "Loan Default Prediction with Machine Learning Techniques," 2020 International Conference on Computer Communication and Network Security (CCNS), 2020, pp. 5-9, doi: 10.1109/CCNS50731.2020.00009.

[4] G. Arutjothi and C. Senthamarai, "Prediction of loan status in commercial banks using machine learning classifier," 2017 International Conference on Intelligent Sustainable Systems (ICISS), 2017, pp. 416-419, doi: 10.1109/ISS1.2017.8389442.

[5] V. Singh, A. Yadav, R. Awasthi and G. N. Partheeban, "Prediction of Modernized Loan Approval System Based on Machine Learning Approach," 2021 International Conference on Intelligent Technologies (CONIT), 2021, pp. 1-4, doi: 10.1109/CONIT51480.2021.9498475.

[6] S. Barua, D. Gavandi, P. Sangle, L. Shinde and J. Ramteke, "Swindle: Predicting the Probability of Loan Defaults using CatBoost Algorithm," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 2021, pp. 1710-1715, doi: 10.1109/ICCMC51019.2021.9418277.

[7] B. Patel, H. Patil, J. Hembram and S. Jaswal, "Loan Default Forecasting using Data Mining," 2020 International Conference for Emerging Technology (INCET), 2020, pp. 1-4, doi: 10.1109/INCET49848.2020.9154100.

[8] Shingi G, "A federated learning based approach for loan defaults prediction," 2020 International Conference on Data Mining Workshops (ICDMW), 2020, pp. 362-368, doi: 10.1109/ICDMW51313.2020.00057.

[9] L. Ke et al., "Loan Repayment Behavior Prediction of Provident Fund Users Using a Stacking-Based Model," 2021 IEEE 6th International Conference on Cloud

Computing and Big Data Analytics (ICCCBDA), 2021, pp. 37-43, doi: 10.1109/ICCCBDA51879.2021.9442613.

[10] Y. Yang, J. Zeng, J. Wang and Z. Tang, "Prediction of Chinese Financial Total Loan Amount via Unbiased Grey-Fuzzy-Markov Chain Method," 2015 8th International Symposium on Computational Intelligence and Design (ISCID), 2015, pp. 388-391, doi: 10.1109/ISCID.2015.190.

[11] T. Deng, "Study of the Prediction of Micro-Loan Default Based on Logit Model," 2019 International Conference on Economic Management and Model Engineering (ICEMME), 2019, pp. 260-264, doi: 10.1109/ICEMME49371.2019.00058.

[12] P. Maheswari and C. V. Narayana, "Predictions of Loan Defaulter - A Data Science Perspective," 2020 5th International Conference on Computing, Communication and Security (ICCCS), 2020, pp. 1-4, doi: 10.1109/ICCCS49678.2020.9277458.

[13] M. Alaradi and S. Hilal, "Tree-Based Methods for Loan Approval," 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI), 2020, pp. 1-6, doi: 10.1109/ICDABI51230.2020.9325614.

[14] Nikhil Bansode, Adarsh Verma, Abhishek Sharma, Varsha bhole, "PREDICTING LOAN APPROVAL USING ML ", International Research Journal of Modernization in Engineering Technology and Science.

[15] Yash Divate, Prashant Rana, Pratik Chavan,"Loan Approval Prediction Using Machine Learning",International Research Journal of Engineering and Technology (IRJET),Volume: 08 Issue: 05 | May 2021.

[16] Aiman Muhammad Uwais and Hamidreza Khaleghzadeh,"Loan Default Prediction Using Spark Machine Learning Algorithms"

[17] H. Ramachandra, G. Balaraju, R. Divyashree and H. Patil, "Design and Simulation of Loan Approval Prediction Model using AWS Platform," 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), 2021, pp. 53-56, doi: 10.1109/ESCI50559.2021.9397049.