e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

PREDICTING LOAN APPROVAL USING ML

Nikhil Bansode[*1], Adarsh Verma[*2], Abhishek Sharma[*3],
Varsha Bhole[*4]

[*1,2,3,4]University Of Mumbai Department Of Information Technology, A.C Patil College Of Engineering, Navi Mumbai, India.

ABSTRACT

Taking out loans from banks has become very common in today's world. Banks' main business is lending money. The primary source of benefit is the interest on the loan. However, because the bank has limited funds to distribute to a limited number of people, determining who the loan can be given to and who would be a better choice for the bank is standard procedure. Credit firms issue a loan after a lengthy period of authentication and confirmation. They are also concerned about whether the borrower will be able to repay the loan without difficulty. Many researchers have been exploring systems for determining loan acceptance in recent years. Machine learning can bring an additional reliable predictive modeling method to the banking business, which is still needed. The primary aim of this paper is to determine if a loan granted to some organization or a specific person would be accepted.

Keywords: Classification, Exploratory Data Analysis, Loan, Loan Approval, Machine Learning, Prediction, Python.

I. INTRODUCTION

For banking institutions, loan approval is a crucial move. The loan applications were either accepted or rejected by the system. Loan recovery is a significant contributor to a bank's financial statements. Because data in banks is growing at such a rapid pace today, bankers must analyze a person's data before approving a loan. It is extremely difficult to predict whether the customer will be able to pay back the loan. Methodologies of Machine Learning (ML) are highly beneficial for determining what will happen when dealing with massive amounts of information. Machine learning assists specifically data learning from its own experiences, as well as data prediction and decision-making. The Python programming language is chosen in our research since it comes with all of the essential tools and libraries. Python is one of the most commonly used and preferred languages in Machine learning and artificial intelligence. As data is the most valuable resource on the globe, it has sparked a revolution in computer science. Machine Learning algorithms have provided numerous data analysis solutions. We would use five machine learning algorithms to predict customer loan approval: the Decision Tree algorithm, the Random Forest algorithm, the Logistic Regression algorithm, SVM, and K Neighbors algorithm. Our primary goal is to use machine learning concepts to calculate a customer's loan status and predict an immediate and precise outcome that assists the lender in analyzing the situation, providing better services, and reducing risk by selecting the appropriate person, saving the lender time and money. We would also test various machine learning algorithms and select the best among them.

The following is a breakdown of the paper's structure: An overview of relevant literature surveys on Research articles on Loan Prediction is included in Section II. Data Description can be found in Section III. The study's methodology for generating results is covered in Section IV. The research results are in Section V. Section VI brings the paper to a conclusion and the future work on this project.

II. LITERATURE SURVEY

Rajiv Kumar and Vinod Jain, in their research paper [1] used the Python programming language to implement the logistic tree, decision tree, and random forest algorithms. They chose the Decision tree method as the most efficient after comparing the evaluation of three kinds of machine learning approaches in terms of prediction accuracy. They didn't fill in the blanks or properly categorize the data however, this can be fixed by filling in the blanks and properly categorizing the data.

Pidikiti Supriya and Myneedi Pavani, in their research paper [2] claim to have pre-processed the data to remove inconsistencies in the dataset. They've also compiled a list of Correlating Characteristics that were found to make people more likely to repay their debts. To divide the dataset into training and testing operations, the

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

80:20 rule was used. The Python platform's corplot and boxplot are used to find the correlation between attributes. However, they haven't utilized any other method to compare accuracy results other than a decision tree. This may be avoided by training datasets with multiple algorithms and comparing their efficiency.

Kumar Arun and Garg Ishan, in their research paper [3] tested a total of six different machine learning approaches, including neural networks, support vector machines, random forests, decision trees, linear models, and Adaboost. There are four sections to this study. (i) Gathering of data (ii) Model evaluation using ML on the collected information (iii) System training using the most feasible model (iv) After the system has been trained on the most promising model, it is put to the test. R programming language was used to create this system. They didn't represent the data results for easier comprehension and comparison, but this problem can be solved by offering data visualization in the form of graphs or other matrix forms.

Authors in [4]. Initially, the information was cleansed. The next step was exploratory data analysis and feature engineering. They had done visualization through graphs. For loan prediction, four models are used. Decision Tree, Naive Bayes, Support Vector Machines, and Logistic Regression methods are the four methods. They determined confidently showing the Naive Bayes model is very capable of delivering superior results to other models after thoroughly studying positive attributes and constraints.

Authors in [5] said a set of data was obtained from the banking sector. The data set is in the ARFF (Attribute-Relation File Format) format, which Weka understands. They used exploratory data analysis to solve the challenge of granting or rejecting loan requests, as well as short-term loan projection. In their research, they did an exploratory data analysis. For prediction, two machine learning classification models are used Decision Tree and Random Forest. In their analysis, they chose the random forest method.

III. DATA DESCRIPTION

We have got the loan data set through Kaggle [14]. The redundant and identical entries were deleted once the dataset was normalized. There is a chance that the data received possibly involve some null values, which could cause inconsistencies. Data must have been pre-processed to boost the algorithm's efficiency. Outliers must be eliminated, and variable conversion must be performed. The dataset gathered for forecasting loan default customers is divided into two groups: training and testing. Our data set includes a total of 13 columns. The response variable is Loan Status, and the remaining variables/factors determine whether the loan would be approved or not.

The characteristics are as follows:

Table 1. Loan Prediction Parameters

Variable	Description	Category	Type
Loan_ID	Loan ID is unique	Qualitative	Integer
Gender	Man/ Woman	Categorical	Character
Married	Applicant married status (Y/N)	Categorical	Character
Dependents	Dependents count	Qualitative	Integer
Education	Education of the applicant	Categorical	String
Self_Employed	Self-employed person (Y/N)	Categorical	Character
ApplicantIncome	Applicants' earnings	Qualitative	Integer
CoapplicantIncome	Co-applicant's earnings	Qualitative	Integer
LoanAmount	Amount of the loan in thousands	Qualitative	Integer
Loan_Amount_Term	The loan's duration in months	Qualitative	Integer
Credit_History	Credit history complies with rules	Qualitative	Integer
Property_Area	Rural/Urban/Semi-Urban	Categorical	String
Loan_Status	Approval of the loan (Y/N)	Categorical	Character

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

IV. METHODOLOGY

WORKING OF THE MODEL

Based on the data provided by the borrower, an organization must automate the loan qualifying method (in real-time). Data such as Loan Amount, Gender, Marital Status, Income, Credit History, Education, Number of Dependents, and a few other details while completing a request form. As shown in Table 1. To make things simple, they created a system that allows them to identify types of applicants, who are qualified for a loan amount and approach them specifically. Since we need to classify everything before determining if the loan status is Yes or No, therefore this is considered as a classification issue. The system can quickly determine if a loan application is likely to be granted or rejected. Figure. 1, shows the working of the proposed model step by step.

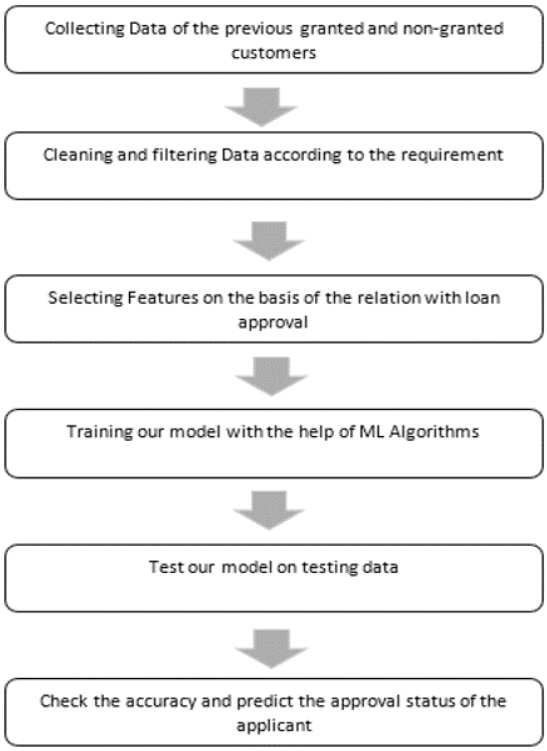

Figure 1: Proposed methodology

ARCHITECTURE OF PROPOSED MODEL

Our project makes use of a variety of algorithms to help us achieve a precise result. The Python programming language, which is among the most often used and popular languages in AI and ML because it comes with all of the necessary tools and libraries has been used in our project. It has several libraries, like pandas for the filtering process, matplotlib for plotting the data, data visualization, and exploratory data analysis. We have also used sklearn which is Scikit-learn which includes several clustering, regression, and classification algorithms that are commonly used in AI and machine learning. Numpy is used to deal with the multidimensional array and data structures.

Seaborn library has been used for data visualization. The model then applies this technique to pre-defined data set including all the information about our customers. In a linear pattern, the algorithms are executed one after the other. The data is then analyzed, segregated, and provided into the model to train it. As shown in Figure. 2. After each algorithm, the precision rate is displayed. We have trained our model with many algorithms to get a precise result. The Random Forest Algorithm, Decision Tree Algorithm, Logistic Regression Algorithm, SVM, and K Neighbors algorithm will all be used, with a 70% training set and a 30% testing set. We have discovered that logic regression, decision trees, and random forests have superior precision. Following the testing procedure, the model predicts if the current candidate based on the conclusion is a good candidate for getting a loan acceptance, it draws from the training data sets. As a result, the better we are in determining the capable borrower, the more beneficial it is to the organization.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

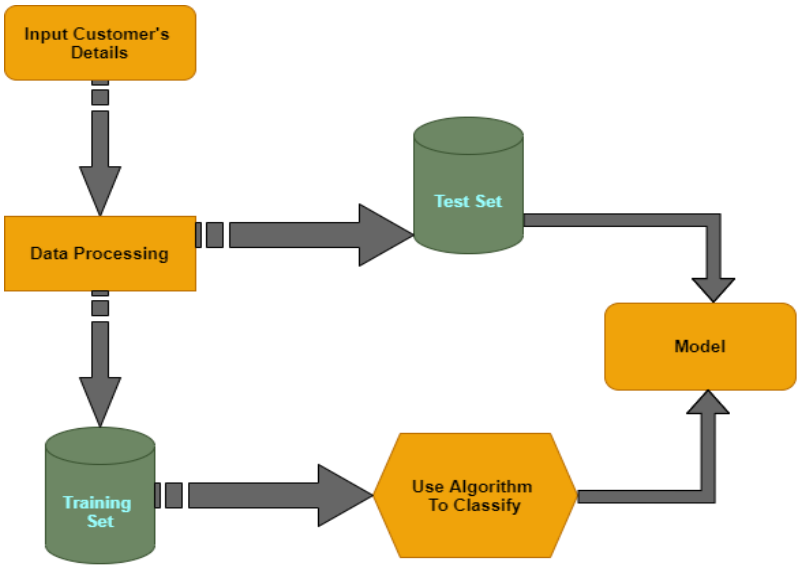

Figure 2: The proposed model's architecture.

This design is based on the category of Supervised learning, with the classification problem. A confusion matrix is used to visualize the data. It displays the results of the prediction model that was created. The model includes a variety of graphs and diagrams that aid in visualizing the model's data flow. Each graph illustrates a collection of procedures carried out on the data set.

ALGORITHMS

- Decision Tree- It's a supervised non-parametric machine learning technique. It can be used for regression as well as classification. On both input and output variables, it works in categorical and continuous modes. All attributes or features must be discretized by the basic algorithm of the decision tree. The most detailed features are chosen for inclusion in the feature set. IF-THEN rules can be used to interpret the knowledge represented in a decision tree.
- Logistic Regression- The LR approach is a popular way to solve binary classification problems. Binary LR makes use of binary dependent variables. The variables used should be relevant. Many of the model's independent variables should be self-contained. The sample size for LR should be large.
- Random Forest- It is most commonly used in classification and regression analysis disciplines. During training, the RF algorithm builds a large number of decision trees. An approach to characterization involves constructing an enormous number of Decision trees over the duration and obtaining the class that would be the mode of the classes produced by the independent tree.
- SVM- Support vector machines (SVM) are a class of supervised learning methods used for classification, regression, and outlier detection. A discriminative classifier is another name for an SVM classifier. In high-dimensional spaces, it is beneficial. When the number of dimensions exceeds the number of samples, the method serves its purpose. It is well-known for its kernel trick for dealing with nonlinear input spaces. SVM discovers an optimal hyperplane, which assists in the classification of new data points.
- K Neighbors- K-Nearest Neighbour is a method for Supervised Learning. It is a non-parametric algorithm, which appears to mean it makes no assumptions about the underlying data. It assumes a correlation between the new case/data and existing cases and places the new case in the category that is most similar to the existing categories.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

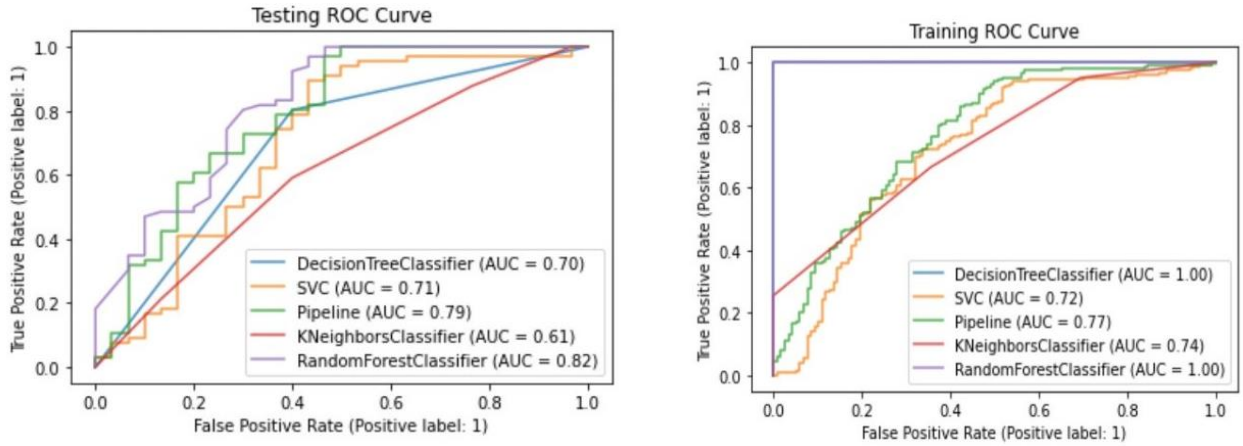

Figure 3: ROC Curves.

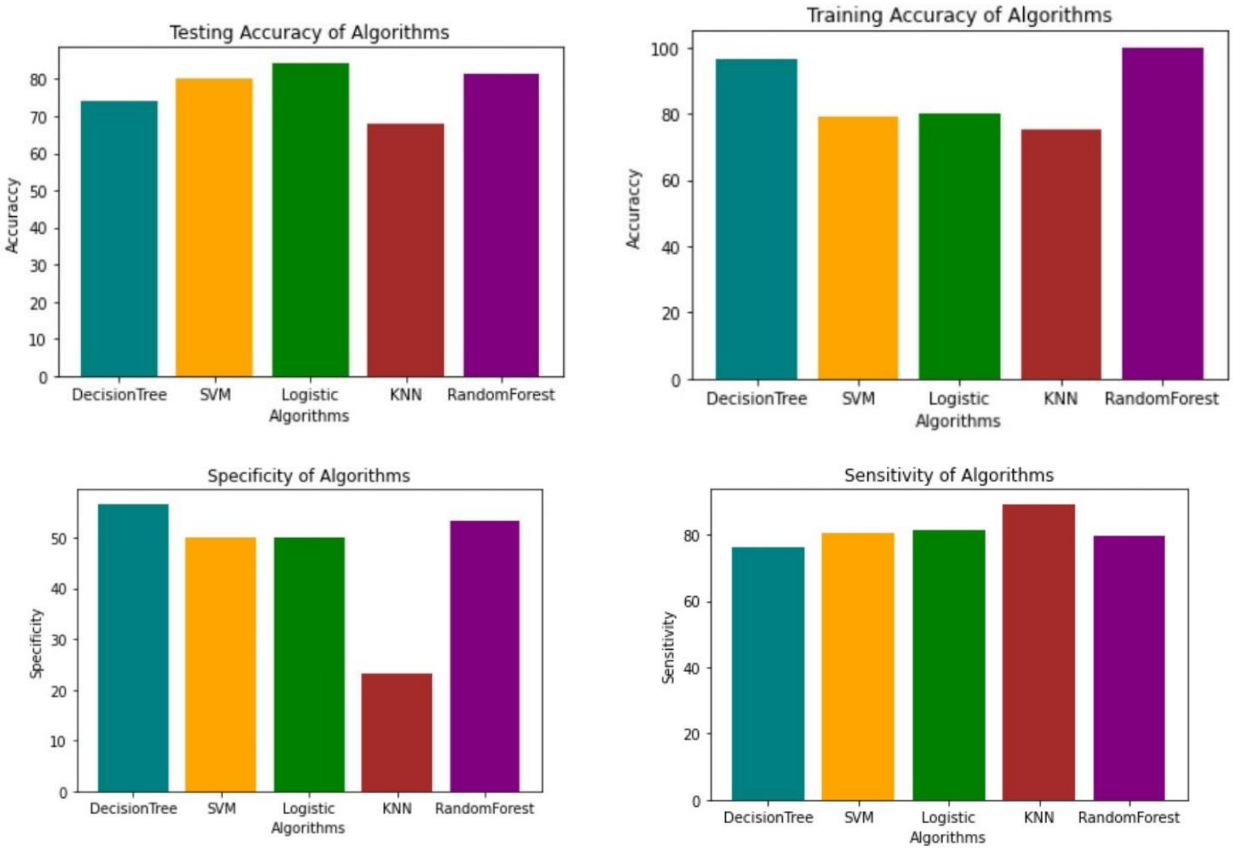

Figure 4: Algorithms Bar Graph Comparison.

V. SCOPE

The scope of the project includes:

- Assists the lender in analyzing the situation.
- Gives better services for use.
- Reduce the risk factor by choosing the right person.
- Save time and money for the lender.

VI. RESULTS AND DISCUSSION

To acquire an accurate result, we used a variety of strategies to train our model. With a 30 percent testing set and 70 percent training set, the Logistic Regression Algorithm, the Decision Tree Algorithm, Random Forest Algorithm, SVM, and K Neighbors were implemented. Logistic regression, on the other hand, has the highest accuracy of all the algorithms.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

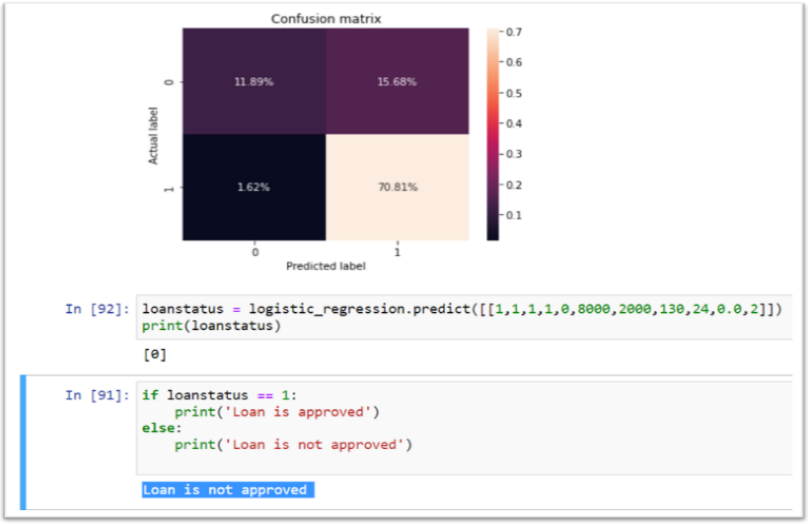

Figure 5: Rejected state.

In Figure. 5, after giving the input we can see the loan status as rejected. As the customer was not eligible.

Figure 6: Accepted state.

In Figure. 6, after giving the input we can see the loan status as accepted. Because the customer was an eligible candidate. Our model has a prediction accuracy of 84.376 %, indicating that it can predict defaulters. A heatmap was used to analyze everything. The correlation matrix is represented visually as a heatmap. It greatly helps in the speedy identification and verification of relationships between columns.

Table 2: Confusion Matrix

	Predicted No:	Predicted Yes:	
Actual No:	22	29	51
Actual Yes:	3	131	134
	25	160	

Table 3: Accuracy Table

Algorithms	**Accuracy**
Logistic regression	84.376 %
Random forest	82.293 %
SVC	80.200 %

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

K Neighbors	76.700 %
Decision tree	71.873 %

Accuracy is just a part of the whole model. Other things are taken into consideration when training the model and analyzing the results. For example, the sensitivity and specificity of a particular algorithm are also evaluated from these two terms. These two play a crucial role in evaluating and plotting the ROC Curve.

$$Specificity = \frac{TN}{TN + FP} \quad Sensitivity = \frac{TP}{TP + FN}$$

To calculate the sensitivity and specificity, one should first evaluate True Positive, True Negative, False Positive, and False Negative. These values can be calculated from the confusion matrix, as shown in Table 2. And accuracy of every algorithm is mentioned in Table 3.

VII. CONCLUSION

This system would be able to determine the status of the loan whether it would get approved or denied swiftly in real-time. Displays accuracy with various algorithms. We have compared the Logistic regression algorithm to two other algorithms, random forest, and decision tree Table III. However, of all the algorithms, Logistic regression has the highest accuracy. Also, it can fill the missing values of the datasets, treat categorical values, scalability problems, overfitting problems, and provide a good visualization of the data using a confusion matrix. Applicants who have a poor credit history are likely to be rejected, especially to the risk of not repaying the loan. Applicants with high income who request low-interest loans have a stronger chance of being accepted, which is logical because they have a strong chance to repay their debts. Few essential characteristics, such as marital status and gender, appear to be overlooked by the organization, but the number of dependents is taken into consideration. The libraries are used professionally and are sufficient for now because we chose the Python programming language, but many aspects require additional exploration. Many areas of our project are left unexplored and might be studied and explored further.

For further research, applicants' Age, past health records, as well as the type of occupation they have will be utilized to evaluate the ambiguity factor of paying debts, and possible defaults of corporate loans for businesses and startups can be forecasted. Another method could be developed to forecast defaulters on different types of loans as well. We used a medium-sized data set to train our model, which may have influenced the outcome; therefore, a big and well-defined data set is required for more accurate results. This paperwork could be expanded to a higher level in the future, allowing the software to be improved to make it more dependable, secure, and accurate. The system has been trained using current data sets that may become older in the future, allowing it to participate in fresh testing to pass new test cases.

VIII. REFERENCES

[1] Rajiv Kumar, Vinod Jain, Prem Sagar Sharma- Prediction of Loan Approval using Machine Learning- International Journal of Advanced Science and Technology Vol. 28, No. 7, (2019).

[2] Pidikiti Supriya, Myneedi Pavani, Nagarapu Saisushma- Loan Prediction by using Machine Learning. International Journal of Engineering and Techniques - Volume 5 Issue 2, Mar-Apr 2019

[3] Kumar Arun, Garg Ishan, Kaur Sanmeet- Loan Approval Prediction based on Machine Learning Approach- IOSR Journal of Computer Engineering, p-ISSN: 2278-8727 PP 18-21.

[4] E. Chandra Blessie, R. Rekha- Exploring the Machine Learning Algorithm for Prediction the Loan Sanctioning Process- (IJITEE) ISSN: 2278-3075, Volume-9 Issue-1, November 2019.

[5] Kshitiz Gautam, Arun Pratap Singh, Keshav Tyagi - Loan Prediction using Decision Tree and Random Forest- (IRJET) Volume: 07 Issue: 08 | Aug 2020.

[6] Sujoy Barua, Divya Gavandi- Swindle: Predicting the Probability of Loan Defaults using CatBoost Algorithm- ICCMC 2021.

[7] Bhoomi Patel, Harshal Patil, Jovita Hembram- Loan Default Forecasting using Data Mining- (INCET) Belgaum, India. Jun 2020.

[8] Sourav Kumar, Amit Kumar Goel- Prediction of Loan Approval using Machine Larning Technique- International Journal of Advanced Science and Technology Vol. 29, pp. 4152 – 4161 (2020).

[9] Soni PM, Varghese Paul- Algorithm For the Loan Credibility Prediction System-(IJRTE) Volume-8, June 2019.

[10] X. Francis Jency, V.P.Sumathi, Janani Shiva Sri-An Exploratory Data Analysis for Loan Prediction Based on Nature of the Clients, -(IJRTE) Vol. 7, pp. 176-179, No. 48, 2018.

[11] S. Vimala, K. C. Sharmili - Prediction of Loan Risk using NB and Support Vector Machinel, vol. 4, no. 2, pp. 110-113 (ICACT 2018).

[12] K. Ulaga Priya, S. Pushpa- Exploratory analysis on prediction of loan privilege for customers using random forest- International Journal of Engineering & Technology, 7 (2018) 339-341.

[13] Deepak Ishwar Gouda, Ashok Kumar, Anil Manjunatha Madivala- Loan Approval Prediction Based On Machine Learning- e-ISSN: 2395-0056 Volume: 08 Issue: 01 | Jan 2021 (IRJET).

[14] https://www.kaggle.com/burak3ergun/loan-data-set

[15] A. Ibrahim, R. L., M. M., R. O. and G. A., "Comparison of the CatBoost Classifier with other Machine Learning Methods", International Journal of Advanced Computer Science and Applications, vol. 11, no. 11, 2020. Available: 10.14569/ijacsa.2020.0111190.

[16] J. Tejaswini, M. Kavya, D. Ramya, S. Triveni and V. Rao Maddumala, "ACCURATE LOAN APPROVAL PREDICTION BASED ON MACHINE LEARNING APPROACH", Journal of Engineering Sciences, vol. 11, no. 0377-9254, 2020.

[17] A. Gupta, V. Pant, S. Kumar and P. Kumar Bansal, "Bank Loan Prediction System using Machine Learning", Ieeexplore.ieee.org, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9336801. [Accessed: 08- Apr- 2022].

[18] A. Vaidya, "Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval", Ieeexplore.ieee.org, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/8203946. [Accessed: 08- Apr- 2022].

[19] M. Ahmad Sheikh, A. Kumar Goel and T. Kumar, "An Approach for Prediction of Loan Approval using Machine Learning Algorithm", Ieeexplore.ieee.org, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9155614/authors#authors. [Accessed: 08- Apr- 2022].

[20] V. Singh, A. Yadav and R. Awasthi, "Prediction of Modernized Loan Approval System Based on Machine Learning Approach", Ieeexplore.ieee.org, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9498475/authors#authors. [Accessed: 08- Apr- 2022].