TEAM – 18

Loan Approval Prediction using Credit Card Score Analysis

J. Sri Sai Samhitha (BL.EN.U4CSE21072) , Amrita Vishwa Vidyapeetham, Bengaluru, India.
K. Adarsh Sagar (BL.EN.U4CSE21075) , Amrita Vishwa Vidyapeetham, Bengaluru, India.
Kundula Haritha (BL.EN.U4CSE21107) , Amrita Vishwa Vidyapeetham, Bengaluru, India.

Abstract: Among all the industries, the banking sector has a large application of the domains like Machine learning, analytics and Data Science. These domains are a paramount importance in the field of loan prediction which help in fraud detection, risk management system, maintaining efficiency, transparency and helping in continuous improvement. Considering the current situation of banking sector, there is a tremendous increase in the trend of taking loans , by which the banking sectors are facing new challenges such as deciding on which customer to approve loan. Due to the increasing rate of loan defaults, it's a difficult task for the banking authorities to assess the loan requests and deal with the risks of customers defaulting the loan. Hence, this project will be assessing important features (attributes) such as credit card score along with annual income, loan amount of the bank customers (instances) which helps the bank decide how risky the borrower is and if they should lend them loan or not. This project focuses on developing an automatic loan prediction system where the machine learns and predicts whether to approve the loan or not based on the customers eligibility criteria. This project works with supervised learning (a type of machine learning), where the machines are trained using well labelled training data (input data which is tagged with the correct output) based on which it predicts the target attribute. The target attribute is divided into either of Approved / Rejected (2 classes) based on whether the loan is approved or not, hence we are working on Binary Classification which is a supervised learning algorithm that refers to thoseclassification tasks that have two labels. A comparative analysis between the machine learning techniques including Logistic Regression, Random Forest, Decision Tree, K- Nearest Neighbor, Support Vector Machine, and XG-Boost is performed to figure out the best suitable model. The evaluation metrices of all the models are calculated and compared, out of which the XG Boost model had shown the best performance.

Keywords: Machine Learning, Data Science, Credit card score, Supervised Learning, Binary Classification.

1. INTRODUCTION

Machine Learning, Data Science, and Data Analytics areas are all crucial in the study of credit card scores and predicting loan approval procedures. A wide range of Machine Learning approaches and algorithms improve the evaluation of creditworthiness, permitting the banking sector to calculate the risk involved in issuing loans to specific clients.

Using Data Analytical tools, banks are able to find patterns and trends in the customer data, which will be aiding the banks to decide on whom to lend loans, and whom not to. Using the Data Science methods, banks are deriving many new insights from huge amount of financial data, helping them to make their process of lending loans in a better way for their customers to be satisfied. All the above domains together play a crucial role in understanding the customers credit score details, history and the risk associated in lending loans to them, finally helping the banking sector to make more optimized decisions regarding lending loans.

Problem Statement: In the banking sector, as the demand for loans is increasing drastically, there are many new challenges faced by the banking sector. One of the main challenges faced by the banks is determining which customers should be granted loans. Assessing the loan requests, evaluation of customer details such as credit card score, loan history, annual income, loan amount and other crucial features has become a complex and challenging task for all the banks, to decide on which customer to lend the loan.

In this project, dataset with 12 features consisting of 4269 bank customers details is considered, to analyze the customers risk level and take decisions regarding approving the loans to the customers.

The currently existing methods in banking sector for loan approval have limited accuracy in analyzing the customer's financial status, as they are based on simplified models, and may not capture the important features leading to incorrect evaluations of potential risks. These procedures are lengthy and time consuming which makes the customers dissatisfied and cause delays in the process of decision-making,

The current methods used do not consider all the risk factors causing incorrect calculations, finally leading to the selection of the wrong borrowers of loans and financial losses for the banks. Some banks don't use all the analytical tools and Machine Learning techniques due to which important valuable insights are missed out leading to inaccurate decisions. Human interference and manual assessments lead to bias and inconsistency, affecting the fairness and accuracy of the process of loan approval. These are the challenges faced by the banking sector to make accurate decisions and mitigate the financial risks while lending loans.

This project involves the development of an automatic loan prediction system which uses machine learning techniques for making decisions on whom to lend the loan. A comparison of various machine learning models including Logistic Regression, Random Forest, Decision Tree, K- Nearest neighbor, Support Vector Machine, and XG-Boost, has been performed to determine the most suitable model for predicting loan approval. The performance evaluation measures of each model have been compared to make a decision.

Our Contributions:

- Searched through various datasets and found out the dataset which had the most relevant features to analyze the loan approval prediction.

- Studied various papers on the topic loan approval prediction using machine learning models.
- Exploration of various machine learning models for loan approval prediction, analyzed their evaluation metrices and found out some models which can efficiently predict the target variable for our project.

2. LITERATURE REVIEW

2.1 Background

The project focuses on building a model that predicts if the bank approves the loan or not to a given customer based on several features of the customers. In the current generation, it's a very enormous, tuff and tedious task for the bank to decide on whom to trust and approve the loan. The banks need to analyze various customer features to decide on a trustworthy customer. Various features such as their annual income, loan amount requested, their credit cardscore and 12 different features are analyzed of 4269 customers for the bank to make accurate decisions on who is a reliable customer for the bank to trust and approve loan to.

U. E. Orji, in paper [1] they havefocused on the significance of the advanced Machine Learning models, in the loan approval prediction of the banking sector. In order to train the machine learning models, the research has been conducted on the dataset "Loan Eligible Dataset" obtained from Kaggle. The dataset was preprocessed for modelling using the techniques of SMOTE, one - hot encoding and Normalization. The study highlights on 6 Machine Learning Algorithms used to estimate the loan eligibility which includes Logistic Regression Algorithm, K-Nearest, Support Vector Machine (SVM), Decision Tree, Bagging and Boosting. The research revealed that all the ML models had great results in terms of accuracy and other metrices like Precision, Recall and F1 score. All the model's accuracy were in the range of 80% to 96%. The highest accuracy obtained was of the Random Forest Model having 95.55% and the least obtained accuracy was of the model Logistic Regression with 80%.

J. Sinha in [2] , have highlighted the relevanceof machine learning models, in implementing an accurate model for predicting the approval of loan to a customer by analyzing the customers credit risks. They tested the dataset named "Bank loan status dataset" collected from Kaggle, which includes the details of 100514 customers and 19 unique attributes of each customer. The dataset is preprocessed and is split into training and testing data of 80:20 ratio. Several Machine Learning approaches used in this work are Logistic Regression, K closest neighbor and Random Forest Classifier. They have generated test metrices such accuracy score, recall, precision, specificity and F1 score for each of the machine learning algorithms and are displayed in a confusion matrix. The paper research suggests that we cannot use logistic regression as the technique for the model as it has highest precision (0.98) and lowest recall value (0.20). Comparing the K-nearest neighbor and Random Forest Classifier, they picked Random Forest Classifier as it has greater accuracy (0.80) and higher F1(0.43) score than K-nearest neighbor.

R. Priscilla, their paper [3] focuses on the developing a loan approval prediction model to ease the duty of banks to decide on the trustworthy clients to approve loans and lower the danger of financial loss for banks. They explored different Machine Learning methods to construct a loan approval prediction model which include Logistic Regression (LR), Support Vector Classification (SVM), Random Forest Classifier, K-Neighbors Classifier, Gradient Boosting, Perceptron, Decision Tree Classifier. They have evaluated an existing dataset, analyzed the skewness and divided the dataset into training and testing dataset and preprocessed the dataset. Then they have compared the evaluation metrices of each ML method which showed that Random Forest Classifier has the greatest accuracy of 96,82%, maximum average precision and average recall of 95.1%% accordingly. The Perceptron has the lowest accuracy of 90.52%, along with the lowest average precision and Average recall of 88.17% and 89.02% respectively. Through this research they identified Random Forest Classifier as the best method to build the model.

V. Singh, in the paper [4] , have made employed of machine learning methods like XG Boost, Random Forest and Decision Tree to forecast loan acceptance for the new applicants in the bank. They have considered two datasets with various analyzed on the important features like Loan_id, Applicant_income , Credit_history etc. to findthe best bases to predict our target variable. They have also depicted the various steps involvedin this loan approval prediction methodology in a flowchart. They have included the architectural diagram for the methodology and enlightened the advantages of loan approval prediction model in the current banking sector field.

P. S. Saini,in [5] they conducted a comparative analysis with 4 ml(i.e, random forest classifier ,k-nearest neighbors, logistic regression, support vector classifier) algorithms on the loan eligible dataset from Kaggle, each model was assessed usingaccuracy,f1 score, and roc score .Out of all methods random forest model obtained best accuracy of 98.04% yet support vector classifier had lesser accuracy of 68.71%.The data set has been pre-processed by handling missing values, encoding categorical variables, and feature scaling then this data set was separated into two parts: dependent variable and independent variable, split it into training data and test data using a 60/40 ratio .Each model was then trained and predictions were made as random forest classifier: accuracy(98.044%) ,f1 score(0.9857),roc score(0.9739) k-nearest neighbor classifier: accuracy(78.491%), f1 score(0.853), roc score(0.0.709) support vector classifier: accuracy(68.715%),f1-score(0.8133),roc-score(0.0.508),logistic-regression: accuracy(79.608%),f1 score(0.866),roc score(0.0.696). The random forest classifier attained the greatest accuracy score of 98.04%, accompanied by logistic regression at 79.61%, k-Nearest Neighbors classifier at 78.49%, and support vector classifier at 68.71%. random forest classifier defeated the remaining models in terms of roc score with score of 0.974.

S. K. Hegde,[6] according to the article they suggested, machine learning model has been developed utilizing decision tree, XG Boost, Random Forest, and Logistic regression approaches for approval of loan based on different factors, where the dataset has been acquired from dream home financing firm.

Compared to other machine learning algorithms, the Logistic regression technique has provided superior prediction result. The dataset is initially separated using k-fold cross-validation and to evaluate the characteristics and target variables of dataset they applied various visualization approaches. Accuracy acquired on various ml algorithms include decision tree (70.51%), support vector machine (78.63%), Random Forest (80.93%), Logistic regression (80.94%), XG Boost (80.45%). Among these ml algorithms logistic regression has greater accuracy.

M. A. Sheikh in the paper [7] have built Logistic regression model, a machine learning technique, has been derived by training the dataset considered and different the different measures of performances are computed. In data pre-processing, apart from the traditional cleaning techniques, data reduction technique is used to deal with huge data, such has PCA (principal Component Analysis) and along with-it data mining techniques are performed on the data. The different measures of performances like Confusion matrix, Accuracy, Precision, Recall, and F1 score has been calculate with the best-case accuracy obtained is 0.811%.

A. Gupta in their paper [8] presents a machine learning-based approach to anticipate loan acceptance, addressing the expanding issues in the banking business. It also describes the differences between supervised and unsupervised machine learning, and explores the relevance of machine learning in the banking business. The Algorithms utilized for prediction in this work include Logistic regression, Random Forest, Correlation between parameters, and these algorithms were classified on the pre-processed data set.

Yashna Sayjadah in their paper [9] of Credit card default prediction using machine learning approaches and in this they use decision tree, random forest to predict the default credit card score and these factors have 82% accuracy and the actual prediction is done using data pre-processing, data partitioning, data visualization with machine learning algorithm and logistic regression. The dataset which is used is the 30000 instances and 24 attributes generated by credit card users.

Varun S in their paper [10] of Credit Score Analysis Using Machine Learning in this they develop binary classifiers based on machine learning models utilizing actual data and the evaluation of the credit risk of the client dataset, different credit risk analysis approaches are applied. Here they utilize logistic regression, Decision Tree, Support Vector Machine, K-Nearest Neighbors and the keyword employed Credit Risk Analysis, Machine Learning, Performance Ranking, Binary Classifiers, Relevant Features.

Jay Lohokare1 in their paper [11] *of* Automated data collection for credit score calculation based on financial transactions and social media in this, they discuss online purchases made through transactions and messages sent via social media. Credit score, bank transactions, loan approval, CIBIL, SMS, and social media are the keywords used.

Ch. Naveen Kumar in their paper [12] of Customer loan eligibility prediction using machine learning algorithms in banking sector in this they used data cleansing decision tree, random forest, support vector machine, K-nearest neighbor, and decision tree with AdaBoost and the keywords used are Banking sector, feature selection, loan prediction, machine learning. The data set is acquired from numerous banks.

2.2 Related works

"**Table 1.** Related works of Credit Card Score Analysis and predicting Loan approval"

Paper	Tasks	Architecture	Datasets
[1] https://ieeexplore.ieee.org/abstract/document/9803172?casa_token=x63YpsHBRmQAAAAA:eoZ9QgjdhS4kI9fekc2FprXA7nEsSd8Ak5i4u_qbYkuZnNBk8W2gpzcuOzOmiXy11wdsig	Building Machine Learning Models for loan eligibility prediction. Analyzing Evaluation Metrices of multiple Machine Learning Models and find out the best model .	Logistic Regression Algorithm, K-Nearest, Support Vector Machine (SVM), Decision Tree, Bagging and Boosting.	Loan Eligible Dataset – Kaggle.
[2] https://ieeexplore.ieee.org/abstract/document/9725599?casa_token=PUxsrrOvCmoAAAAA:uIJkz9-AxNHie4mr9EtBv3ma_bV628YvEoOYQDbTonfjD7baMxkbsPze8PGpc_YzyNmXAcUjjQ	Building Machine Learning Models for loan eligibility prediction. Analyzing Evaluation Metrices of multiple Machine Learning Models and find out the best model.	Logistic Regression, K nearest neighbor and Random Forest Classifier.	Bank loan status dataset - Kaggle.

[3] https://ieeexplore.ieee.org/abstract/document/10083650?casa_token=GiaWbJO0Sk8AAAAA:kViugfrbCGHi_bulH9QRL2up3klWpKLHSTSM6T3eJUQJt-duGCKmcqDZO-YpRDQGTJFqZBiqFA	Building Machine Learning Models for loan eligibility prediction. Analyzing Evaluation Metrices of multiple Machine Learning Models and find out the best model.	Logistic Regression (LR), Support Vector Classification (SVM), Random Forest Classifier, K-Neighbors Classifier, Gradient Boosting, Perceptron, Decision Tree Classifier, K-Fold.	loan approval data.
[4] https://ieeexplore.ieee.org/abstract/document/9498475?casa_token=c-gT9kFFC5AAAAAA:wS1SZfYjLngCeGGM-ypLy0FN1F72srq9bon03hU7wJlCf_CI5jYWgvIZe-tH09euxLbw1G3QSw.	Loan approval prediction model using machine Learning algorithms. Methodology's flowchart and Architectural diagram for implementation in the banking sector.	Logistic Regression, K nearest neighbor and Random Forest Classifier.	Bank loan status dataset – Kaggle.

[5] https://ieeexplore.ieee.org/abstract/document/10182799?casa_token=i_nDirUcV4UAAAAA:UMXobP7kvdGSVPF2gi3cwwSH4SIOvoP-10bDCiZeGnaN1Z8FUvZpp4YGVkMASRxI-ItHt_SFcw	Building Machine Learning Models for loan eligibility prediction. Analyzing Evaluation Metrices of variousMachine Learning Models and findingout the best suitable model.	Logistic Regression (LR), Support Vector Classification(SVC), Random Forest Classifier, K-Neighbors Classifier.	Loan Eligible Dataset from Kaggle.
[6] https://ieeexplore.ieee.org/abstract/document/10083580?casa_token=lKBxDSJhtP4AAAAA:kmpOeJfp85iFsDz4mP7OqUUxEHke1bYvBI70ZcvGn-9CxSv0Xq0LCymVPyTc0Cpx5wYzYD085Q	Building Machine Learning Models for loan eligibility prediction. Analyzing Evaluation Metrices of various Machine Learning Models and find out best suitable model.	Logistic Regression(LR) XG-Boost, Random Forestand Decision Tree.	Dream House Finance Company from Kaggle.
[7] https://ieeexplore.ieee.org/abstract/document/9155614?casa_token=HX_WrSpdSv4AAAAA:r5W1PBS7XsSRCCuCvT4zutoW8Tx0H1dUb5Nm3erHqMJbW0i41zHNVulI5Bd2hO78-DdT4SEycA	Building Logistic Regression Model for loan eligibility prediction. Analyzing Evaluation Metrices of various Logistic Regression Model and find out best suitable model.	Logistic Regression (LR).	Dataset related to loan approval data from Kaggle.

[8] https://ieeexplore.ieee.org/abstract/document/9336801/?casa_token=-UTo0ZuOA54AAAAA:OnIPutLJBLxAy0A08DDyeyId_B4Eq5Ce9Wo6CwfuI5qt9QQ8EVAW91uOoQjw35txeeZmjvMse_E	Machine learning-based loan approval predictionsystem. Significance of supervised and unsupervised learning. Tested algorithms Logistic Regression and Random Forest.	Logistic Regression (LR), Random Forest, Correlation between parameters.	Dataset related to loan Approval data from Kaggle.
[9] https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8776802	Building Machine Learning Models for loan eligibility prediction. Credit Score, Data mining, Machine Learning, Banking	Data Pre-processing, Feature Selection, Data Partitioning, data visualization	The dataset used is generated from credit card operations by the users
[10] \ https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7977024	Building Machine Learning Models for loan eligibility prediction. Credit Score; bank transactions; loan approval; CIBIL; SMS; social media;	Data Pre-Processing, Data balancing and Feature Selection, Logistic Regression, Decision Tree, Support Vector Machine	Real-time data from a Philippine bank as its dataset.
[11] https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9835725&tag=1	Building Machine Learning Models for loan eligibility prediction. —Credit Risk Analysis, Machine Learning, Performance Ranking, Binary Classifiers, Relevant Features	Credit Score; bank transactions; loan approval; CIBIL; SMS; social media; R	Data set is taken from the various social media users

[12] https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9835725	Building Machine Learning Models for loan eligibility prediction. Banking sector, feature selection, loan prediction, machine learning.	Decision Tree, Random Forest,Support Vector Machine, K-Nearest Neighbor, Ensembled model	Data set is taken from the various bank users

2.3 Motivation

As we can see the tremendous increase in the trend of taking loans, there are many new challenges that are being faced by the banking sector nowadays. The main motive behind this project is to address one of the main issues concerned by banks that is analyzing various features and finding out the customers whom the bank can rely upon and lend loan to. Finally, the objective is to provide the banking systems with a model which can accurately predict the customers whom the bank can trust on for sanctioning loans.

3. PROPOSED METHODOLOGY

3.1 Data Preprocessing

Several data preprocessing steps have been performed on the dataset which includes checking for null values in dataset, removing the spaces in the column names, performing label encoding of the data and scaling the data. Using the isnull (). sum () function it is found that there are no null values in any of the columns of the dataset and any spaces present in the names of the columns have been removed. Later on, label encoding has been performed on the columns of education, self-employed and the target variable loan_status and where all the categorical values in the columns have been mapped to numerical values. Label encoding converts the data into numerical format, which the machine learning algorithms can understand and work easily with. Label encoding also reduces the memory space used. Scaling is performed on all the numerical input features where all the independent variables are standardized in a fixed range of values, so that all the input features contribute equally for the models to learn and determine the loan status (target variable).

3.1 Algorithms

We have trained and tested several machine learning models on the dataset to find the best suitable model for loan price prediction. The machine learning algorithms used are as below:

3.2.1 Logistic Regression Algorithm

Logistic Regression Algorithm is a supervised machine learning model which is often used for classification tasks. This approach works out the link between the independent (input) variables and the binary dependent (output) variable and finds out the chances of a new data instance of belonging to a certain class. For the given dataset, this model has been evaluated on the testing set and had predicted the outcome of loan_approval class of each of the testing dataset as yes (1) or no (0).

3.2.2 Decision Tree

Decision tree is a supervised machine learning technique which is used for both classification and regression applications. It represents a hierarchical structure similar to a tree and consist of the root node representing the entire dataset, internal node which represents the choice made with respect to an input attribute, leaf nodes representing final class labels and branches which represent the outcome based on the conditions on the internal nodes. This algorithm works by splitting the entire dataset into smaller sets recursively by selecting the suitable feature by calculating the entropy and information gain at each step. The information gain is calculated using the below formula:

$$IG(T, a) = H(T) - H(T|a) \tag{5}$$

In (5),

$IG(T, a)$= Information gain of dataset T , with respect to attribute a.

a = value of attribute.

$H(T|a)$= conditional entropy of dataset T given attribute a.

$H(T)$= entropy on the dataset T.

3.2.3 Random Forest Algorithm

Random Forest is a supervised machine learning technique which is based on ensemble learning. The method works by mixing differing number of decision trees which act on subsets of the provided dataset. The average of the findings of each decision tree is taken as the final result. This approach assists to increase the performance as well as the accuracy of the model. As the number of decision trees integrated rises, the total accuracy of random forest likewise increases. This algorithm is trained and tested on the dataset and

determined the accuracy, precision, recall and F1-score to evaluate the performance of the model on the dataset.

3.2.4 Support Vector Machine Algorithm

Support Vector Machine is a supervised machine learning technique which is used for both classification and regression tasks but performs best on binary classification problems. This method attempts at figuring out most optimum hyperplane which will be categorizing the points into their respective classes. Support vectors are the locations which are closest to the hyperplane, and this technique works out the hyperplane such that the distance between the support vectors and the hyperplane will be greatest. This approach works on both linearly separable data and for non - linearly separable data by applying kernel functions. We have tested this model or our dataset and evaluated the performance metrices of this model on our dataset. The decision function of SVM algorithm is given as:

$$f(x) = \sum_{i=1}^{N} \alpha_i \, y_i K(x_i, x) + b \qquad (6)$$

In (6),

$f(x)$ = decision function.

N = number of support vectors.

α_i = Lagrange multipliers.

y_i = class label of ith support vector.

$K(x_i, x)$ = Kernal function.

b= bias term.

3.2.5 K- Nearest Neighbor Algorithm

K-Nearest Neighbor (KNN) classifier is a supervised machine learning technique used mostly for addressing classification issues. This algorithm considers the classes of k points which lie nearest to the test instance. The k nearest points to the test intake are found out by calculating the Euclidean distance or Manhattan distance. The majority or the mean class of all the classes of k nearby points is the predicted class of new instance. We have analyzed the performance of KNN classifier on the dataset considered by calculating the performance evaluation metrices.

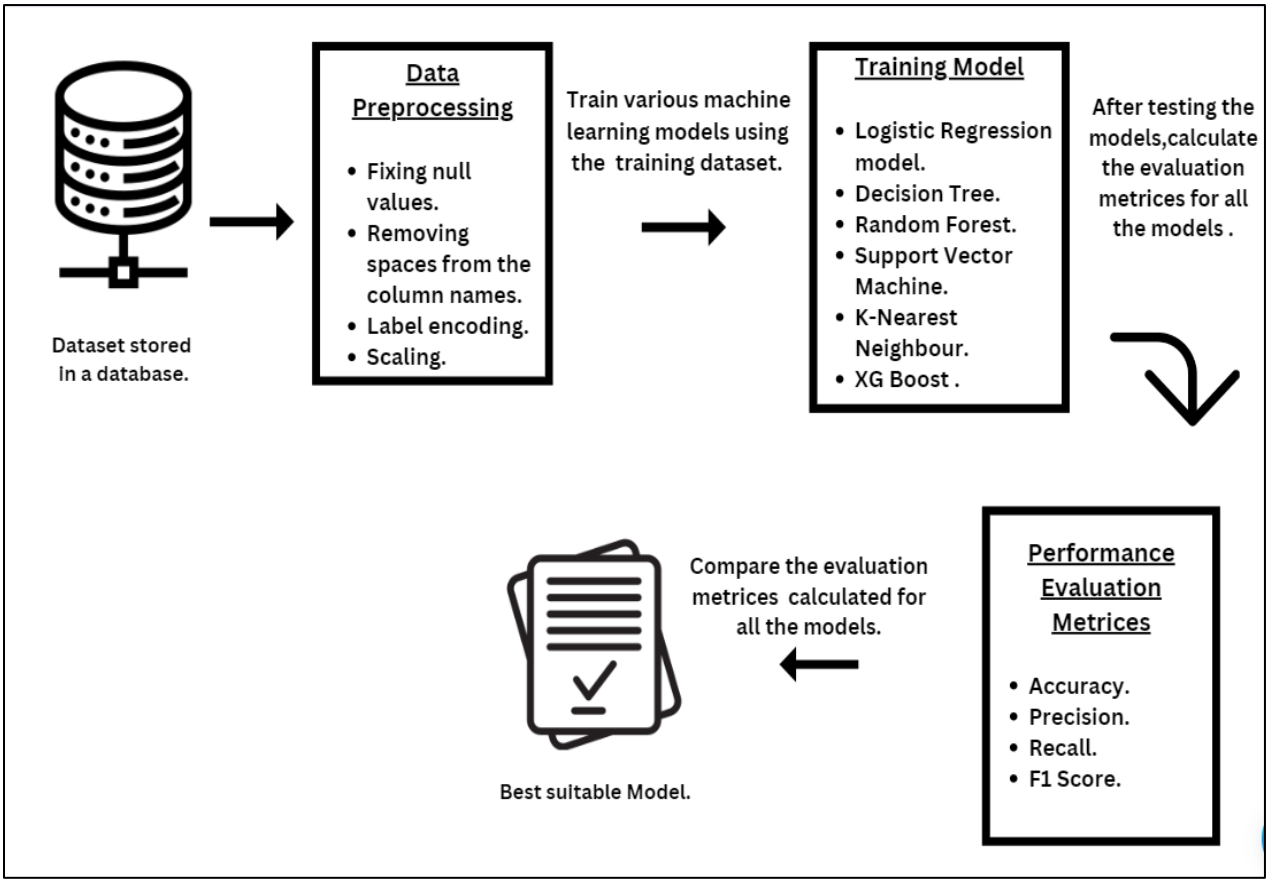
3.2.6 XG-Boost Algorithm

Extreme Gradient Boost method is a machine learning algorithm which employs an ensemble technique of decision trees with gradient boosting. This approach is often utilized on structured data. It takes the predictions provided by numerous decision trees and merges them, where one decision tree corrects the errors created by the preceding decision tree. By

doing this the overall accuracy of the predictions provided by the model is boosted. In this research, the model is trained and tested using the dataset and the performance of this model is assessed using accuracy, precision, recall and F1-score values.

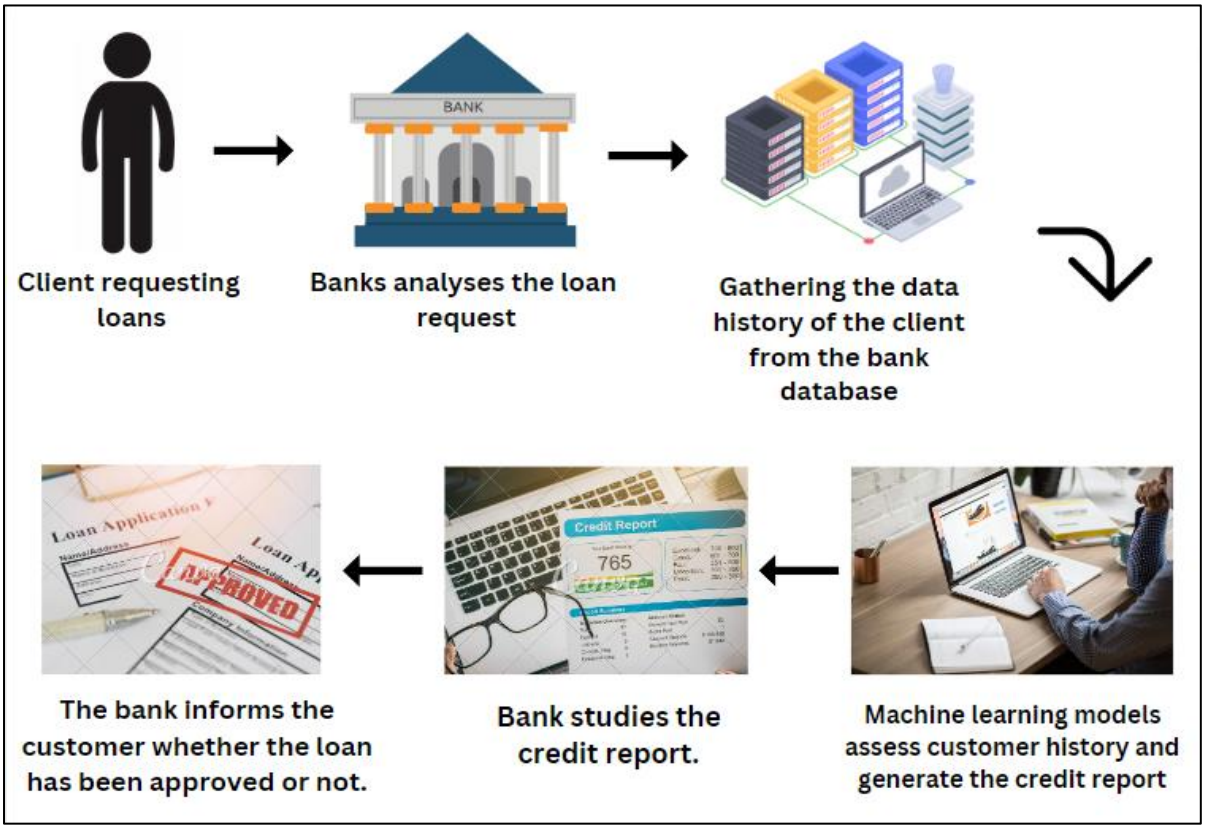
3.2 Proposed Architecture

Work flow of proposed architecture



"**Fig.1** An architecture of the proposed model."

Real time environment



"**Fig.1** Real time scenario of loan approval prediction in banks."

1.EXPERIMENTAL ANALYSIS AND RESULTS DISCUSSION

4.1 Experimental setup and Dataset

We have made used of python language to train and test the various machine learning models on the considered dataset. The experimental environment used to implement the python codes is Jupyter notebook on windows operating system. Jupyter is a very user-friendly and open -source platform where we can run python codes using large number of inbuilt libraries with ease. The dataset which we considered is from the online platform Kaggle and is published by Kai, student at DIT University Dehradun. The dataset consists of 4219 instances and 13 important features which effectively contribute in predicting the loan approval status. The features that are present in the dataset includes loan id, number of dependents, education status, Self employed, annual income, loan amount, loan term, credit score, residential assets value, commercial assets value, luxurious assets value, bank asset value and the target variable Loan status. Based on these features of each of the 4219 customers, the loan approval status of all the customers is predicted using the machine learning models.

4.2 Evaluation parameters and formulations

Evaluation Indices

Evaluation metrices or the performance metrices are used the examine the performance and efficiency of the machine learning models on the provided dataset. The performance metrices covered in this study comprises Accuracy, Precision, Recall and F1-Score. Out of all the instances classified by the model , the total number of instances which are correctly classified as the loan is approved is represented by truly positive (TP) ,the total number of data instances which are correctly classified as the loan is not approved is represented by truly negative (TN), the total number of data instances which are wrongly classified as loan is approved is represented by falsely positive (FP) and the total number of data instances which are wrongly classified as loan is not approved is represented by falsely negative (FN).

Accuracy: It tells out of all the data instances classified by the model, how many data instances were classified correctly by the model.

$$Accuracy = \frac{TP + FP}{TP + FP + TN + FN}$$

Precision: Precision tells out of all the data instances classified as positive by the model, how many data instances are actually positive.

$$Precision = \frac{TP}{TP + FP}$$

Recall: It indicates out of all the data examples that are genuinely positive, how many of the cases did the model accurately categorize as positive.

$$Recall = \frac{TP}{TP + FN}$$

F1- Score: Harmonic mean of precision and recall.

$$F1\ Score = \frac{2 * (Precision * Recall)}{Precision + Recall}$$

4.3 Comparison methods

To find out the best suited model to predict the approval of loan, the evaluation metrices for the machine learning models including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, K- Nearest Neighbor and XG-Boost have been generated. A comparison study of the performance metrices of all the described above algorithms is done and the most appropriate algorithm is picked.

4.4 Results and discussion

Various Machine learning models were trained and evaluated on the dataset examined by dividing it in the ratio of 80:20 and the performance assessment metrices accuracy, precision, recall and F1 Score are calculated for each of the model. The evaluation metrices for each of the model is compared as given below in Table 1. It is noticed that the values of accuracy, precision, recall and F1 score achieved is higher for XG Boost Algorithm compared to Logistic Regression Model, Decision Tree, Random Forest, Support Vector Machine. Hence it is shown that XG Boost Algorithm is the best suited algorithm for forecasting the loan approval status for the dataset.

Table 1 Comparison of performance evaluation metrices for various machine learning models.

Algorithm	Accuracy	Precision	Recall	F1 score
Logistic Regression	0.908	0.908	0.908	0.908
Decision Tree	0.974	0.97	0.974	0.974
Random Forest	0.976	0.976	0.976	0.976
Support Vector Machine	0.627	0.393	0.627	0.484
K- Nearest Neighbor	0.594	0.574	0.594	0.579
XG Boost	0.981	0.981	0.981	0.981

5. CONCLUSION

In this paper, the implementation of an effective and reliable loan approval prediction system is discussed using various machine learning algorithms including Logistic Regression Model, Decision Tress, Random Forest, Support Vector Machine, K Nearest Neighbor and XG Boost algorithm. The performance evaluation metrices which are calculated for all the machine learning algorithms are compared, and it proves that XG Boost Algorithm is the best and efficient algorithm for predicting the approval of loan most accurately.

REFERENCES

[1] Orji, U.E., Ugwuishiwu, C.H., Nguemaleu, J.C. and Ugwuanyi, P.N., 2022, April. Machine Learning Models for Predicting Bank Loan Eligibility. In *2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON)* (pp. 1-5). IEEE.

[2] Sinha, J., Astya, R., Tripathi, K., Verma, A. and Verma, M., 2021, December. Machine Learning based Loan Allocation Prediction System for Banking Sector. In *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)* (pp. 1614-1619). IEEE.

[3] Priscilla, R., Siva, T., Karthi, M., Vijayakumar, K. and Gangadharan, R., 2023, January. Baseline Modeling for Early Prediction of Loan Approval System. In *2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF)* (pp. 1-7). IEEE.

[4] Singh, V., Yadav, A., Awasthi, R. and Partheeban, G.N., 2021, June. Prediction of modernized loan approval system based on machine learning approach. In *2021 International Conference on Intelligent Technologies (CONIT)* (pp. 1-4). IEEE.

[5] Saini, P.S., Bhatnagar, A. and Rani, L., 2023, May. Loan Approval Prediction using Machine Learning: A Comparative Analysis of Classification Algorithms. In *2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)* (pp. 1821-1826). IEEE.

[6] Hegde, S.K., Hegde, R., Marthanda, A.V.G.A. and Logu, K., 2023, February. Performance Analysis of Machine Learning Algorithm for the Credit Risk Analysis in the Banking Sector. In *2023 7th International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 57-63). IEEE.

[7] Sheikh, M.A., Goel, A.K. and Kumar, T., 2020, July. An approach for prediction of loan approval using machine learning algorithm. In *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)* (pp. 490-494). IEEE.

[8] Gupta, A., Pant, V., Kumar, S. and Bansal, P.K., 2020, December. Bank Loan Prediction System using Machine Learning. In *2020 9th International Conference System Modeling and Advancement in Research Trends (SMART)* (pp. 423-426). IEEE.

[9] Sayjadah, Y., Hashem, I.A.T., Alotaibi, F. and Kasmiran, K.A., 2018, October. Credit card default prediction using machine learning techniques. In *2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA)* (pp. 1-4). IEEE.

[10] Varun, S., Theagarajan, A. and Shobana, M., 2023, April. Credit Score Analysis Using Machine Learning. In *2023 International Conference on Networking and Communications (ICNWC)* (pp. 1-5). IEEE.

[11] Lohokare, J., Dani, R. and Sontakke, S., 2017, February. Automated data collection for credit score calculation based on financial transactions and social media. In *2017 International Conference on Emerging Trends & Innovation in ICT (ICEI)* (pp. 134-138). IEEE.

[12] Kumar, C.N., Keerthana, D., Kavitha, M. and Kalyani, M., 2022, June. Customer Loan Eligibility Prediction using Machine Learning Algorithms in Banking Sector. In *2022 7th International Conference on Communication and Electronics Systems (ICCES)* (pp. 1007-1012). IEEE.