International Journal of Security and Privacy in Pervasive Computing
Volume 14 • Issue 1

A Systematic Survey of Automatic Loan Approval System Based on Machine Learning

Vandana Sharma, J. C. Bose University of Science and Technology, India*
https://orcid.org/0000-0002-9612-7721

Rewa Sharma, J. C. Bose University of Science and Technology, India

ABSTRACT

The banking sector is an integral part of an economy as it helps in capital formation. One of the most critical issues of banks is the risk involved in loan applications. Employing machine learning to automate the loan approval process is a significant advancement. For this topic, all classification algorithms have been tested and assessed in previous researches; however, it is still unclear which methodology is best for a particular type of dataset. It is still difficult to identify which model is the most effective. Since each model is dependent on a certain dataset or classification approach, it is critical to create a versatile model appropriate for any dataset or attribute collection. The aim of the study is to provide detailed analysis of previous studies and to propose a predictive model for automatic loan prediction using four classification algorithms. Exploratory data analysis is performed to obtain correlation between various features and to get insights of banking datasets.

KEYWORDS

Bank Loans, Classification Algorithm, Confusion Matrix, Exploratory Data Analysis, Indian Banking System, Loan Prediction, Loan Risk, Predictive Model

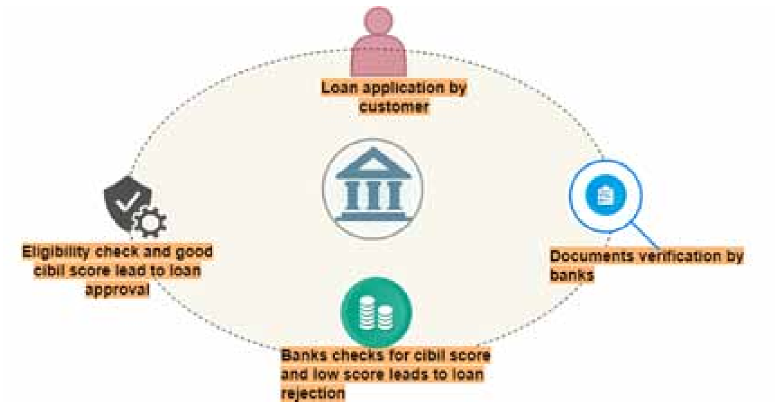
INTRODUCTION

The Indian banking sector is a strong, well- capitalized, and well-regulated industry. Despite the fact that the government has been injecting capital into public sector banks via recapitalization bonds for the past two years, liquidity has become a major concern following the Covid outbreak. The banking sector is quietly struggling to meet the problems it faces, which include preserving capital sufficiency, asset quality, and growth. Bad loans are one of the major issues. Loan recovery is one of the issue that has harmed the banking sector, which is struggling to preserve the quality of its assets. In order to avoid wrongdoings, proper examination and severe application methods are required in loan approval process.

Banks receive numerous amount of loan applications on daily basis. Banks would lose money if loan repayments are delayed, thus they must carefully examine the loan approval procedure. When they first authorize a loan, they do a lot of paperwork, which results in a lot of data. Banking has improved its study of determining the potential of risk through client profile, prior expenditures, and consumer transaction history, among other things, in recent years. Yet we can see assessing the loan related risk is still one of the primary concern for every banks. It usually involves steps portrayed in Figure 1.

DOI: 10.4018/IJSPPC.304893
*Corresponding Author

Figure 1. Existing method in banks



During the process of loan approval there are difficulties on both ends of the transaction whether it is the lender or the borrower. The bank staff checks the customer profile thoroughly on a variety of parameters. They mainly checks for the default risks that should be as low as possible. The applicant must have made timely payments on past loans and should possess good credit score. As a result, loan processing takes time. One requires a good Cibil score to get their loan approved. Cibil score depicts the credit profile of an applicant. It is a three digit numeric summary of credit history having predetermined range usually in between 300 and 900. The necessity of maintaining credit history is visible in the whole process. There comes a situation when applicant is fresh without having any credit history, in such instances, there is high possibility of rejection of loan.

Technology inclusion in banking sector is one of the thing which can work out in longer term. Thus motivated by the recent advancement in automation process and problems being faced by banking industry a model is proposed for automating the loan approval system. Many researches provided solution to this particular problem using different algorithms. The aim of this study is to perform detailed analyses of previous research works to find their limitations, to extract patterns among the types of classifiers applied, dataset used and maximum efficiencies achieved by each algorithms. Findings of literature survey are used to select algorithms and dataset for proposed model. In this paper four classification algorithms Logistic Regression, Random Forest, Support Vector Machine and Gradient Boosting are applied simultaneously.

The summarizing details of the findings of this work are as follows:

- Theoretical Background provide details of Indian banking system and classification of bank loans. Applications of Machine Learning, Data Analytics and Predictive Analysis are also discussed in reference to banking sector issues.
- Research Approach section presents analysis and results of methodologies used in providing solution to the loan prediction problem. Results of findings are illustrated using different charts.
- In Related Work section, summary of the datasets and algorithms used, with accuracies and limitations is provided.
- Proposed Methodology section discusses about various classification algorithms and evaluation techniques which can be applied.
- Results and Discussion section provides outcomes of exploratory data analysis and accuracies achieved by the proposed model using classification reports.
- Conclusion and Future work section discusses about the results and limitations of this work and provides suggestions that can be applied in future studies.

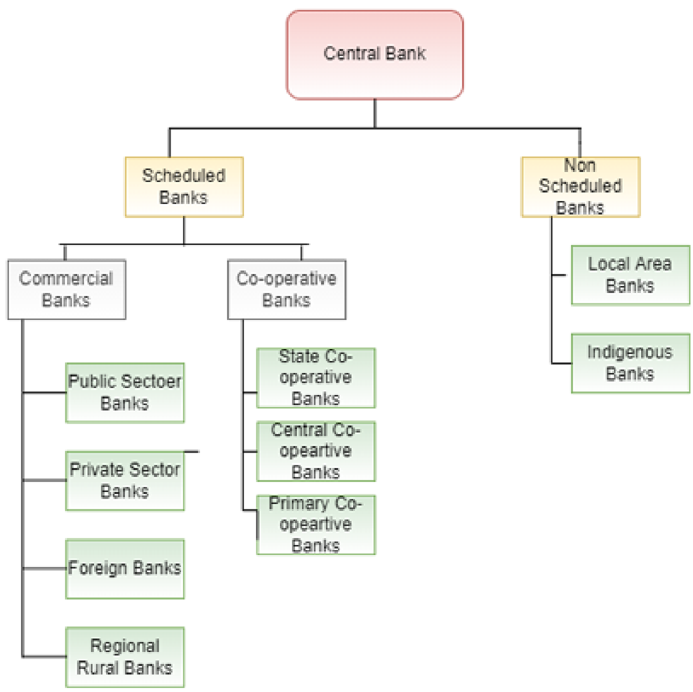
THEORETICAL BACKGROUND

In this section the Indian banking sector and its entities are discussed. Further classification of different types of loan is provided. Consecutively an overview of data science, machine learning predictive analysis and classification is also provided.

Indian Banking Sector

Banks provide various financial services like safe deposition of customer's currency and its exchange, loan and mortgage services, wealth management, lending facility, overdraft services and many others. It provide indispensable services for both consumers and businesses. Therefore, banks are regulated by the national government or central authority. Several different kinds of banks are there in India including corporate banks, retail banks, commercial banks and investment banks. The classification of different types of Indian banks is given in Figure 2.

Figure 2. Types of banks



There are two broad categories under which banks are classified in India- Scheduled Banks and Non Scheduled Banks.

Scheduled Banks can be defined as "Banks which have been included in the second scheduled of the RBI Act, 1934." The Rules and Regulations of Scheduled Banks are made by Reserve Bank of India (RBI), a central and highest monetary authority of India. These banks are allowed to borrow money from Reserve Bank of India and need to deposit amount in it to maintain Cash Reserve Ratio.

Non Scheduled Banks are those who does not comply with RBI guidelines. These banks also need to follow Cash Reserve Ratio conditions but they can have Cash reserve Ratio fund with themselves as there is no compulsion for its deposition from the RBI. Banks under this category usually do not borrow from RBI for its daily banking activities except in some emergency situations.

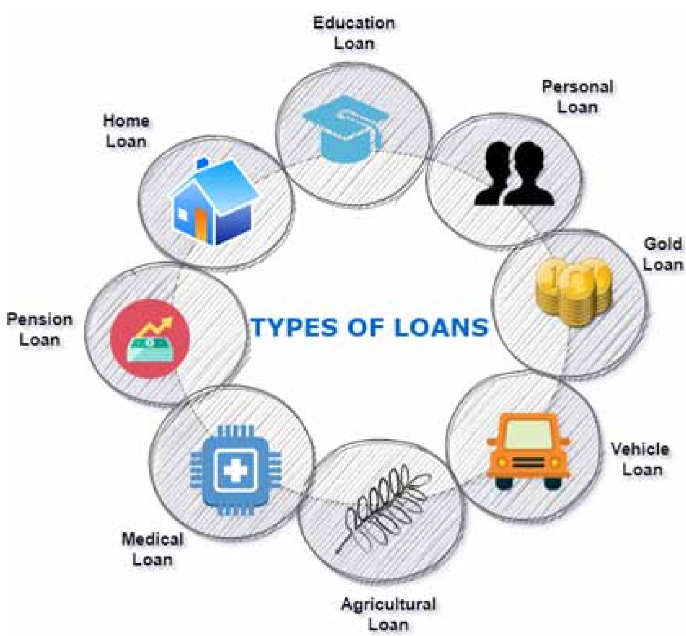
Bank Loans

The general people can get loans from any bank of their choice under the stipulations of banking regulation act as per their requirements. The loans are money in bulk or in good amount borrowed for a specific length of time and need to be returned in installments at some predetermined interest rates.

Loans can be classified into two types- Open-ended and Close-ended loans in terms of credit management. Open-ended loans are the ones in which customer provides agreement for certain amount. These loans does not have any certain end date. Examples are such as credit cards, debit cards and a home equity line of credit (HELOC). Close-ended loans are typically an installment loan that are provided for a predetermined sum and repaid over time in installment payments. Examples are personal loans, home mortgage loan, auto payments, installment loan, and car loan.

Based on the security provided other two types: of loans are Secured loans and unsecured loans. Secured loans are the ones that are protected by a collateral or tangible assets. This collateral can be kept in case one is not able to repay the loan in specific pre-determined time. It tends to have lower interest rates and longer tenure as compare to unsecured loans. Examples are Home Loans, Car Loan. Unsecured loans are offered without collateral. The lender provides loans taking into consideration of the personal property or resources owned by the borrower. It tends to have higher interest rates because of higher risk but faster processing as compared to secured loans. Examples are Personal Loans, Small Business Loans. Further based on the purpose and pledged assets different kinds of loans exists as provided in Figure 3.

Figure 3. Types of loans



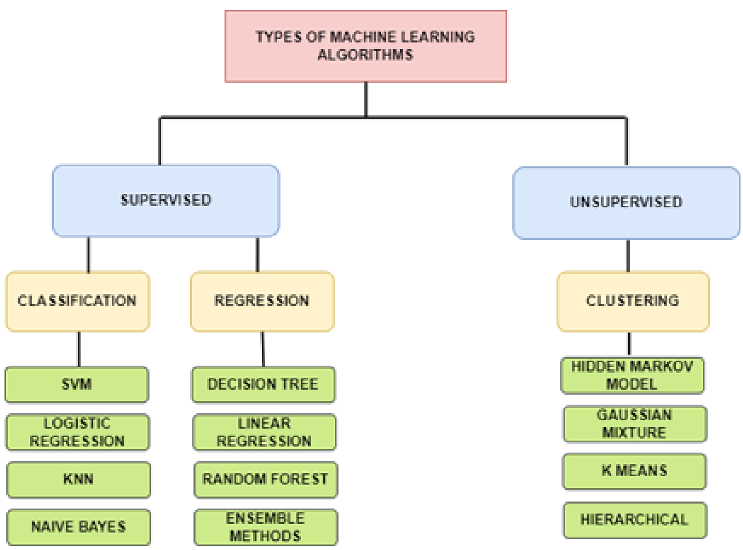
Technology Inclusion in Banking Sector

In present scenario banking sector is generating trillions of data on daily basis. It is beyond human ability to manage or analyze this much data manually and to transform it to have some fruitful knowledge. Data Science is a blanket term which include data analytics, data mining, machine learning and many other related disciples. With the help of Data Mining we can search in large stores of data by using automatic tools to discover hidden patterns of the data or to get future trends by analyzing such data with the help of mathematical algorithms. By using data mining we can easily distinguish the loan borrowers whether he can repay loan on time or not. Various phases of Data Mining are described in Figure 4.

Figure 4. Phases of data mining

Artificial intelligence (AI) is associated with transforming human intellect into machines. Machine Learning (ML) is a subfield of Artificial Intelligence that makes predictions using statistical models. ML algorithms are employed in the banking industry to identify fraud, automate trading processes, and give financial advice to investors. For example, we can train ML algorithms with the help of some financial data containing details of customer income, age, occupation, credit history etc. to identify whether he qualifies for getting a loan or not. There are mainly two types of ML algorithms- Supervised and Unsupervised. *Supervised learning* develop models with the help of both input and output data. *Unsupervised learning* group and interpret data with the help of only input data. Further there are various algorithms available for each type of learning. Figure 5 describes types of algorithms for both learning processes.

Figure 5. Types of machine learning algorithms



Predictive analytics uses machine learning techniques to develop predictive models. Various types of data values can be used to train these models to help various fields in predicting the new values over time. There are two types of models: classification models predicting class membership, and regression models predicting numerical value. Built-in algorithms in predictive analytics software packages may be utilized to create predictive models. These algorithms are referred to as 'classifiers,' and they identify the category data belongs to. Decision Trees, Regressions, and Neural Networks are the most often used Predictive Models.

Classification is a most common technique that helps in analyzing and categorizing the available data with the help of classifiers to provide accurate predictions. The training set data is used to build model and test set data is used to validate the model. Examples of Classification technique includes Fraud detection, Loan defaulter detection, Credit risk detection .The frequently used algorithm for Classification is Decision Tree. As depicted in Figure 6, Data Analytics, Artificial Intelligence, Machine Learning and Predictive Analysis are interrelated to each other.

Figure 6. AI ML and Data Analytics relationship

RESEARCH APPROACH

To finish the study process, this survey takes a methodical approach as described in Figure 7.

Figure 7. Research methodology

RESEARCH QUESTIONS

In the banking industry, the two most important questions are:

1) How risky is the customer?
2) Should the bank approve or deny the customer's loan based on the risk?

Considering these two questions two research questions are formulated for the study:

Hypothesis One: How to predict the validity of an applicant for loan approval?
Hypothesis Two: How to help and mitigate the loan default rate?

Method of Searching Research Paper

It is imperative to explore relevant research articles for getting precise knowledge for any research study. Relevant keywords plays a key role in identifying the research findings. Thus significant keywords are used for finding out research publications with the aid of research questions that are formulated in the study. Various substrings are applied with the help of conjunctions and disjunctions. Table 1 contains a list of different strings (S1, S2, and S3) that are utilized.

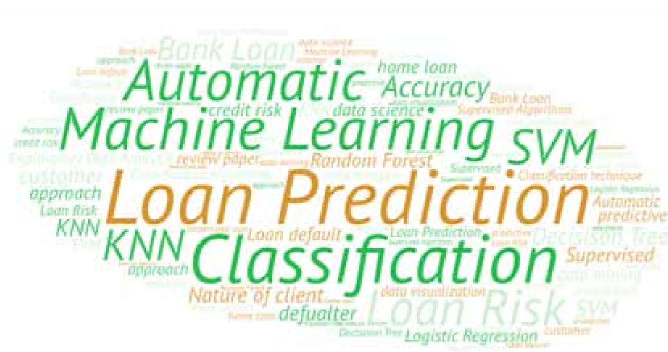
S = S1.S2.S3 or S = S1+S2+S3

Table 1. Keywords used

String	Keywords
S1	Loan, Automatic Loan, Loan Prediction
S2	ML, Machine Learning
S3	Algorithms, Classification Algorithms

The titles of all the research papers are explored to make a world cloud portrayed in Figure 8. Loan Prediction, Machine Learning, and Classification Algorithms are the primary keywords considered in the word cloud. All relevant papers are discovered using correct keyword selection to the best of knowledge.

Figure 8. Word cloud of research papers



Categorization of Papers

Following the discovery of relevant articles, the next step is to categorize them using loan prediction system approaches. Following that, a thorough literature assessment of these study publications is conducted. Approximately 90 research publications were finalized and then categorized on the basis of types of modelling algorithm used, first is using classification models and another is regression and mathematical based models.

Information Extraction

The information is obtained on the basis of methodology and strategies employed from the literature review. These methods are the foundations of this comparison based study. Since the nature of the study is classification based the focus is upon work implemented by classification based algorithms mainly.

Comparing and Preparing Final Report

In this stage a detailed statistical analysis is performed for all research findings.

The outcomes in Figure 9, Figure 10 and Figure 11 show the distribution of selected research papers based on publication year, dataset and algorithms respectively. Nearly 55% of research papers are considered from 2020 to 2021.

Figure 9. Publication year of research papers

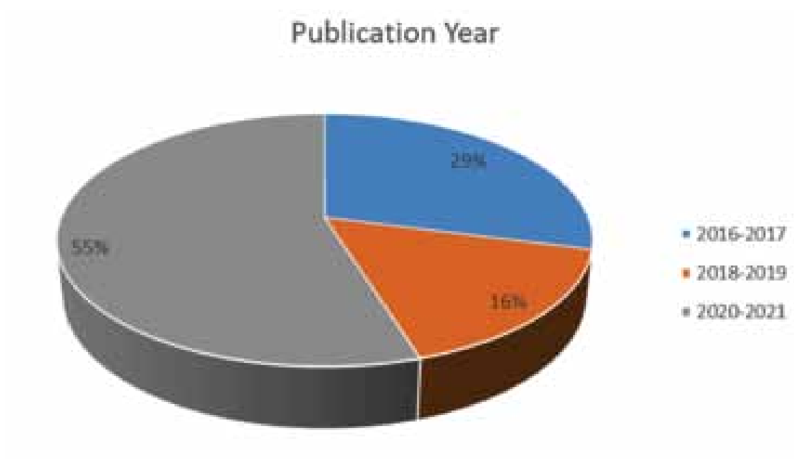


Figure 10. Algorithms used by research papers

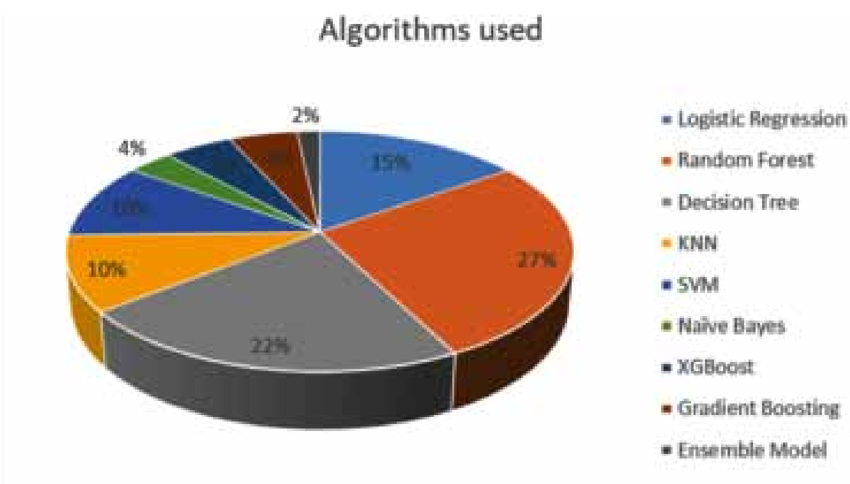
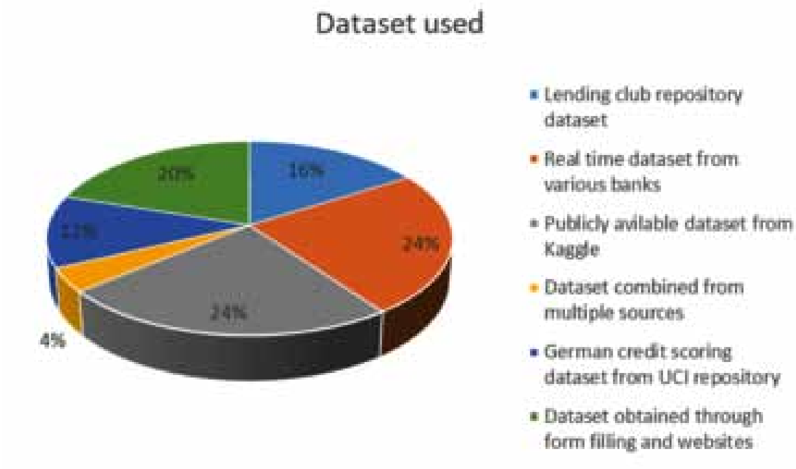


Figure 11. Dataset used by research papers



RELATED WORK

Different research papers have proposed different techniques for solving loan prediction problem but since the nature of our study is classification based the focus is drawn on the work done through supervised learning based classification algorithms mainly. Chilagani Raviteja et al (Alsaleem & Hasoon, 2020) suggested that solely ML classifiers are not sufficient enough to build a model for identifying loan defaulters. The data science techniques must be applied to improve the predictions. Important features were selected as predictors after doing EDA. The Random Forest algorithm is used and performance was measured on the basis of false positive rate. Suliman Mohamed Fati (Diwate et al., 2021) used a historical dataset to build a model to predict the validity of loan. Three algorithms Logistic Regression, Decision Tree and Random Forest were applied. Pre-processing was done to find missing values, outlier detection and removal. EDA was performed for finding out characteristics. Model was evaluated with different performance metrics including AUC and ROC curve. Anchal Goyal et al (Goyal & Kaur, 2016a) proposed an ensemble model for loan prediction. Eleven machine learning models with nine properties are built with the help of R language. Comparison of performance was made using different parameters such as Accuracy, Gini, Auc, Roc. The feature importance is calculated using Real Coded Genetic Algorithms. X. Francis Jency et al (Hsieh & Hung, 2010) proposed Exploratory Data Analysis (EDA) as a method for predicting loan amounts depending on the nature of the client and their requirements. Different visualization graphs are made with the help of pandas and matplotlib libraries of python. Many outcomes were provided by the visualization graphs such as preference of short term loans rather than long term by the customers. Mehul Madaan et al (Khadse, 2020) focused on finding the probability of loan default of an applicant using Random Forest and Decision Tree. Data was cleaned before doing Exploratory Data Analysis. Dataset was divided into 70:30. Performance was evaluated using Precision, Recall, F1 Score and Support with the help of confusion matrix. P. Maheswari et al (Sudhakar & Reddy, 2016) used Logistic Regression, Random Forest and KNN for building a model for predicting loan defaulters using statistical measures. Pre-processing was done on the dataset followed by EDA and feature selection using PCA and LDA. The model was validated using performance evaluation metrics. Aboobyda Jafar Hamid et al (Murthy et al.,) proposed three algorithms- j48, Bayes Net, and naïve Bayes to create predictive model for classifying customers on the basis of customer behaviour and previous payback credits. Weka application was used to build a model. K. Gana Sai Prasad et al (Patil & Dharwadkar, 2017) focused on enhancing the model accuracy of Random Forest in loan approval prediction model. The machine was trained to assume a linear boundary between the loan defaults and the non-defaults using SVM, fitted by R-Statistical software. Model analysis was done using a confusion matrix. As per result, if the probability is greater than 75% it is high chance of paying back of loan by the applicant. Kavita Khadse (Prasad et al., 2019) work is based on published resources. Exploratory Analysis is the nature of the study, and the study process includes creating functional and Ml workflows for the required systems, performing attribute selection using Univariate and Multivariate selection, segregation of training and testing datasets, feature engineering, Random Forest Classifier application, and finally, generation of Confusion Matrix. Random Forest feature engineering is used, with the 10-fold cross validation AUC values guiding the process. Mohammad Ahmad Sheikh et al (Ramesh, 2017) used logistic regression to target customers for granting loan by evaluating their likelihood of default on loan. Logistic Regression with sigmoid function was used. The study outcome showed that customers having bad credit score would fail to get approval for the loan and applicants having high income and demanding lower loan amount are more likely to get loan approval. Pidikiti Supriya et al (Supriya et al., 2019) proposed a loan prediction model using SVM, Decision Tree, KNN and Gradient Boosting. Outlier detection and removal, as well as imputation removal processing, were done during the pre-processing stage for the dataset followed by model building and evaluation of the model. Ashlesha Vaidya (Vaidya, 2017) discussed about logistic regression and its mathematical representation. The work introduced predictive and probabilistic approach for loan approval problem. It was concluded that if the probability was more than 0.5, the loan should be approved; otherwise, the application should be rejected. Maan Y. Alsaleem et al (Somayyeh & Abdolkarim, 2015) worked on a dataset of 1000 loans and their repayment status using ML algorithms. The findings were assessed by comparing the performance of each algorithm using different metrics, with neural networks seeming to have the best accuracy when compared to the other methods. The research was conducted using the Weka Version 3.8.4 environment for model development and testing.

The summary is provided in table containing details of the applied methods and algorithms, dataset used and the efficiencies achieved.

Table 2. Summary of Research findings

Title of the paper	Year	Dataset used	Method used	Best Efficiency	Limitations
Intelligent Defaulter Prediction Using Data Science Process (Alsaleem & Hasoon, 2020)	2021	Lending Club Repository Dataset	Random Forest	93.40%	To increase classifier accuracy, deep learning neural network architectures-based classifiers with model diagnostics might be researched.
Prediction Of Loan Status In Commercial Bank Using Machine Learning Classifier (Arun et al., 2016)	2017	Lending Club Repository Dataset	Combination of Min-Max normalization and KNN	75.08%	A model accuracy can be increased by modifying present iteration level 30 based on KNN model.
A Comparative Study Of Machine Learning Algorithms For Predicting Loan Default And Eligibility (Arutjothi, 2017)	2021	NA	Logistic Regression, Decision Tree, Random Forest	NA	Since Random Forest provides best accuracy it should be applied on various others datasets.
Customer Loan Prediction Using Supervised Learning Technique (Bhanu & Narayana, 2021)	2021	Publicly available dataset from Kaggle	Random Forest, Logistic Regression, Decision Tree, KNN, SVM	Random Forest-82% Logistic Regression-73% Decision Tree-72% KNN-59% SVM-78%	Model faced various forms of computer glitches and error in content. Model efficiency can be increased using techniques like dynamic weight adjustment thus making it more reliable.
Loan Approval Prediction Using Machine Learning (Raviteja & Santosh, 2021)	2021	Real time Dataset	SVM	81.11%	Other classification algorithms can be applied to check for better efficiency.

Table 2 continued on next page

Table 2 continued

Title of the paper	Year	Dataset used	Method used	Best Efficiency	Limitations
Machine Learning-Based Prediction Model For Loan Status Approval (Diwate et al., 2021)	2021	Publicly available dataset from Kaggle	Logistic Regression, Decision Tree, Random Forest	Logistic Regression-91% Random Forest-86% Decision Tree-82%	Realistic dataset with more features can be used. Accuracy can be improved using feature extraction and a mixed machine learning approach.
Loan Prediction Using Decision Tree and Random Forest (Gautam et al., 2020)	2020	Real time Dataset	Random Forest, J48 Decision Tree classifier	Random Forest-85.75% Decision Tree- 63.39%	This prediction module can be used with the automated processing system module.
A Survey On Ensemble Model For Loan Prediction (Goyal & Kaur, 2016a)	2016	NA	Bagging, Boosting and Stacking	NA	Article provides a good theoretical background for ensemble modelling but lacks in providing deeper insights to the process of applying ensemble model to the dataset.
Accuracy Prediction for Loan Risk Using Machine Learning Models (Goyal & Kaur, 2016b)	2016	Publicly available dataset from Kaggle	Eleven machine learning models	81.25% for Tree model of genetic algorithm	More number of seed values should be tested for acquiring best efficiency.
Bank Loan Prediction System Using Machine Learning (Gupta et al., 2020)	2020	Publicly available dataset from Kaggle	Logistic Regression, Random Forest	NA	It is possible to use a larger real-time dataset for this model.
Machine Learning Based Loan Prediction System Using SVM and KNN (Jency et al., 2018)	2020	Data provided by the customers by through webpage	SVM, KNN	NA	The validity of collected data through web application is indefinite. Higher number of instances of one particular type may provide biasedness to the model. More real time dataset can be applied.
Prediction For Loan Approval Using Machine Learning Algorithm (Rawate & Tijare, 2017)	2021	Real time Dataset	SVM, Naïve Bayes	NA	The efficiency achieved for proposed model is not provided for Naïve Bayes algorithm
Loan Default Prediction Using Decision Trees And Random Forest: A Comparative Study (Khadse, 2020)	2021	Lending Club Repository Dataset	Random Forest, Decision Tree	Decision Tree-73% Random Forest-80%	Updated dataset containing present scenarios of loan status must be applied.
Predictions of Loan Defaulter - A Data Science Perspective (Sudhakar & Reddy, 2016)	2020	Lending Club Repository Dataset	Logistic Regression, Random Forest, KNN	Logistic Regression-80% Random Forest-79 % KNN-78%	More cross validation approaches like Stratified K fold technique may be applied other than Grid Search CV.
Loan Approval Prediction System Using Machine Learning (Madaan et al., 2021)	2020	Dataset combined from multiple sources	Random Forest	NA	Ensemble approach should be applied with different algorithms to get more accurate results other than Random Forest.
Developing Prediction Model Of Loan Risk In Banks Using Data Mining (Murthy et al.,)	2016	Real time Dataset	DT J48, Bayes Net and Naive Bayes	J48-78.37% bayesNet-77.47% naiveBayes-73.87%	Accuracies achieved by the model are quite low. Collected data available in ARFF format can be preprocessed further and various feature selection method may be used to increase efficiency.
Analysis of Banking Data Using Machine Learning (Madane & Nanda, 2019)	2017	Germen Credit Dataset from UCI ML Repository	Supervised Artificial Neural Network	72% (Dataset 1) 98% (Dataset 2)	Since dataset 1 provided less accuracy this data should be preprocessed and labelled more efficiently.

Table 2 continued on next page

Table 2 continued

Title of the paper	Year	Dataset used	Method used	Best Efficiency	Limitations
Customer Loan Approval Classification By Supervised Learning Model (Patil & Dharwadkar, 2017)	2019	Publicly available dataset from Kaggle	Random Forest, SVM	NA	More tuning parameters can be applied to increase efficiency. Model may be applied on various other real time datasets.
Applications Of Machine Learning In Loan Prediction System (Prasad et al., 2019)	2021	NA	Random Forest	94%	Graph results may be used to build a predictive model. Also graphs can be used to make a risk score model.
Predicting Bank Loan Risks Using Machine Learning Algorithms (Ramesh, 2017)	2020	German Credit Scoring Dataset taken from UCI Repository	DT J48, Random Forest, Bayes' Theorem, Multilayer perceptron	DTJ48-73.5% Bayes Net-75% Naïve Bayes-77.5% Random Forest-78.5% Multilayer Perceptron-80%	Data can be preprocessed more efficiently for making number of Yes and No instances equal for loan default in the dataset.
Predictive Analytics For Banking User Data Using AWS Machine Learning Cloud Service (Athreyas et al., 2022)	2017	University of California Irvine Dataset	Logistic Regression	91.15%	Algorithms other than Random Forest must be checked for better efficiency.
Home Loan Data Analysis And Visualization (Kotsiantis, 2007)	2021	Dataset of customer details through form filling	Logistic Regression, Decision Tree, Random Forest	NA	Only theoretical concepts are discussed. Any output or results is not provided.
An Approach For Prediction Of Loan Approval Using Machine Learning Algorithm (Sheikh et al., 2020)	2020	Publicly available dataset from Kaggle	Logistic Regression with sigmoid function	81.11%	Dataset contains small amount of values. Large dataset may be used further to better train model.
Prediction Of Modernized Loan Approval System Based on Machine Learning Approach (Singh et al., 2021)	2021	Dataset containing user inputs	XG Boost, Random Forest, Decision Tree	XGBoost-77% Random Forset-76% Decision Tree-64%	The efficiency of model should be evaluated using some matrices like Accuracy, precision, Recall or F1 Score. Evaluation is missing in proposed model.
Credit Risk Analysis and Prediction Modelling Of Bank Loans Using R (Sudhamathy, 2016)	2016	German Credit Scoring Dataset Taken from UCI Repository	Decision Tree	83.33%	Although model provides very good accuracies it is unable to find credit score for individual customer. Thus credit score model can be built extending the proposed work.
Loan Prediction By Using Machine Learning Models (Supriya et al., 2019)	2019	Real time Dataset	SVM, Decision Tree, KNN, Gradient Boosting	81.11% for Decision Tree	Model accuracy can be increased by using better feature selection or hyper parameter tuning.
Machine Learning Applications In Loan Default Prediction (Tiwari, 2018)	2018	NA	Logistic Regression, KNN, Classification and Regression Tree (CART), Random Forest	86% for Random Forest	CART algorithms provided lowest accuracy which can be increased or solved by using trees pruning.
Predictive And Probabilistic Approach Using Logistic Regression:Application To Prediction Of Loan Approval (Vaidya, 2017)	2017	Real Time Dataset	Logistic Regression	NA	Proposed model does not undergo evaluation process using any curves or matrices.

Note: NA is used for Not Available status

Despite many studies, models, and approaches, determining which model is best is a difficult task. Major limitations traced out during literature review are related to unavailability of dataset than can truly define current needs or trends. Majority of the publicly available historical datasets are outdated.

Figure 12 shows maximum efficiencies achieved by different types of datasets in various researches. Lending club dataset and German credit scoring dataset from UCI repository provided best accuracies in comparison to others.

Figure 12. Efficiencies achieved by datasets

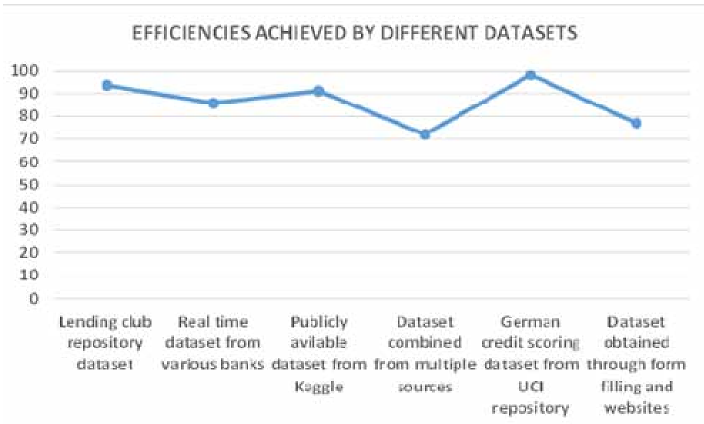
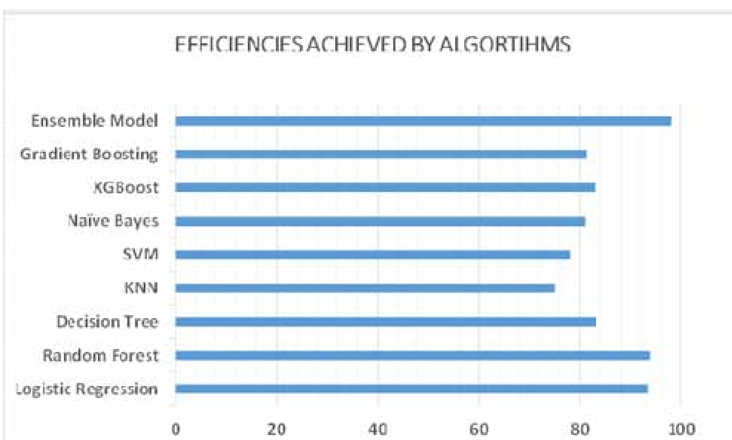


Figure 13 illustrates performance of different algorithms that are been used in different studies. Ensemble models seems out to be best in terms of accuracy. Still very few research papers are there using ensemble models. Other than ensemble approach, logistic regression and random forest provided suitable efficiencies and thus used in large number by the researchers. These two algorithms provided exceptional results

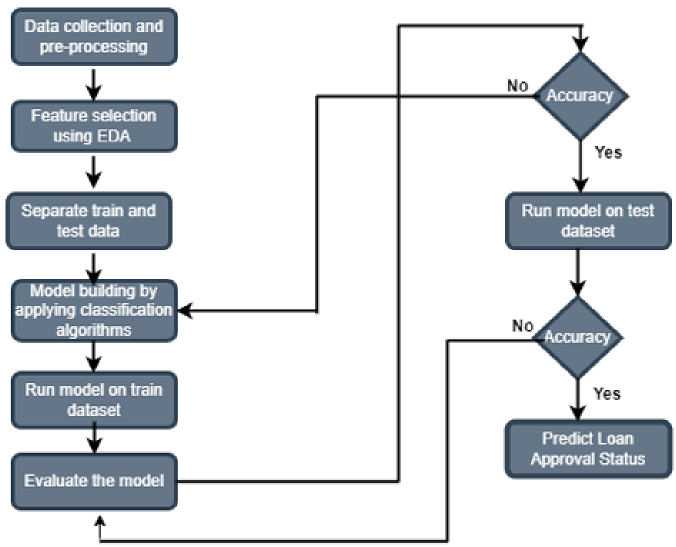
Figure 13. Efficiencies achieved by algorithms



PROPOSED METHODOLOGY

The majority of papers used Random Forest, Logistic Regression, Decision Tree and Ensemble model and provided adequate efficiencies. For this reason, four algorithms are chosen to be applied on the proposed model, which are Logistic Regression, Random Forest, Decision Tree and SVM. SVM despite not providing good accuracies is selected in order to check if its efficiency can be optimized through various hyper parameter tuning and feature selection techniques. Out of all, maximum efficiencies are achieved by Lending Club dataset and UCI repository dataset. Thus a Lending club dataset is selected for proposed model. The workflow of predictive model is portrayed in Figure 14.

Figure 14. Workflow of proposed model



Step 1: Data Collection and Preprocessing

The dataset having details of banking customers is collected from Kaggle. Each variable's description of the dataset is given in Figure 15. There are 614 items in total in the dataset, which comprises 19 columns. The collected dataset may contain some impurities. It may contain inconsequential or missing variables.

Data preprocessing helps in resolving such issues of dataset by cleaning it and performing transformations. The outliers and imputations are handled by detecting and resolving them with the help of proper visualization graphs in preprocessing.

Figure 16 provides visualization of missing values present in dataset variables. Each variable like Sex, Dependents, and Marital Status is handled one by one for filling in the missing values.

Step 2: Feature Selection using Exploratory Data Analysis (EDA)

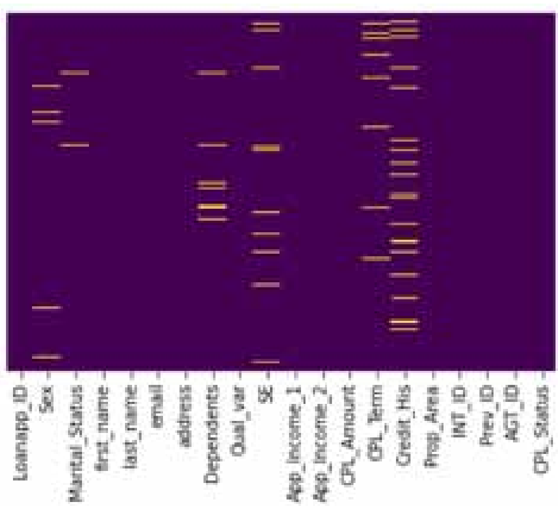
After doing preprocessing task an exploratory data analysis is performed to know about the nature of customer to whom loan is granted. Better understanding of the dataset is obtained by getting insights of each characteristics. EDA is the process of obtaining explanatory solutions to different ostensibly unanswered issues. EDA assists in determining the link between qualities in order to identify abnormalities, as well as providing a statistical summary of connected aspects. All of these issues may be answered with graphs that are properly correlated. After doing EDA a feature selection is done for finding out the most relevant variables in the dataset.

Figure 15. Dataset description

```
#general idea about dataset
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 614 entries, 0 to 613
Data columns (total 20 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Loanapp_ID      614 non-null    object
 1   Sex             601 non-null    object
 2   Marital_Status  611 non-null    object
 3   first_name      614 non-null    object
 4   last_name       614 non-null    object
 5   email           614 non-null    object
 6   address         614 non-null    object
 7   Dependents      599 non-null    object
 8   Qual_var        614 non-null    object
 9   SE              582 non-null    object
 10  App_Income_1    614 non-null    float64
 11  App_Income_2    614 non-null    float64
 12  CPL_Amount      612 non-null    float64
 13  CPL_Term        600 non-null    float64
 14  Credit_His      564 non-null    float64
 15  Prop_Area       614 non-null    object
 16  INT_ID          614 non-null    int64
 17  Prev_ID         614 non-null    object
 18  AGT_ID          614 non-null    object
 19  CPL_Status      614 non-null    object
dtypes: float64(5), int64(1), object(14)
memory usage: 96.1+ KB
```

Figure 16. Missing value description



1) Univariate

Univariate analysis is one of the basic and simplest type of data analysis in which single variable is explored and analyzed separately. It describes patterns and responses of the individual variable.

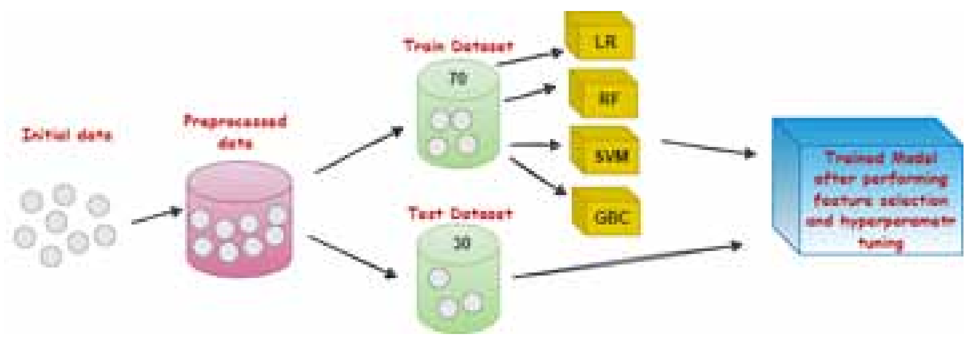
2) Bivariate

After independently analyzing each variable in Univariate analysis, hypotheses may be evaluated in bivariate analysis by again analyzing each variable with the target variable. New features can be created based on domain knowledge that may impact the target variable. The bivariate analysis makes use of two types of data: categorical and continuous data.

3) Multivariate

Multivariate analysis uses more than two variables altogether to gain deeper insight of variable correlation. It tries to find patterns and relationships among multiple dependent variables with regard to target variable.
Step 3: Building a model
In model building process after doing preprocessing task, categorical variables are converted into dummy variables like sex, their self-employment and marital status and qualification before training a model. Then dataset is divided into two parts: training dataset and testing dataset into the ratio 70:30. After that four classification algorithms are applied- Logistic Regression, Random Forest, Support Vector Machine and Gradient Boosted Classifier. Further hyper parameter optimizations, feature selection and various cross validations techniques are applied to validate the model. A k-fold cross validation sampling technique is used for this model. The whole process is depicted in Figure 17.

Figure 17. Model building process



The four classification algorithms chosen for the study process are described below:

1) Logistic Regression

It is a most commonly used classification based supervised ML algorithm. It uses statistical technique to predict probability of an outcome on the basis of one or more independent variable. To calculate the probability it employs a link function known as the sigmoid function to bring the target variable to 0 to 1. The likelihood of a target variable is predicted using log of odds as dependent variable.

2) Random Forest

It is a supervised approach that uses various decision trees which provides better accurate predictions than any single decision tree. It is an ensemble learning approach in which a set of weak models is combined to create a powerful model. The Random Forest Algorithm is based on Decision Tree principles. The distinction is that the decision tree method only considers one aspect, but the Random Forest Algorithm evaluates many decision trees and returns a solution satisfied by majority of decision trees.

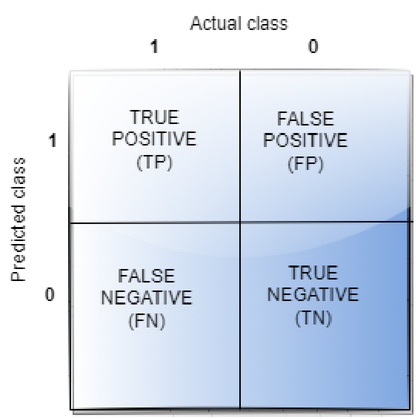
3) Support Vector Machine

It is also known as SVM and is most popular supervised algorithm. The purpose of SVM is to design a hyper plane in N-dimensional space that divides dataset in two categories in best possible optimized way and assign classes to the data points. N is the number of features. The hyper plane is the optimum decision boundary. The goal is to select a hyper plane that has maximum gap in it comparing to training data set points increasing probability of classification for new data points.

4) Gradient Boosted Classifier

This algorithm applies for both regression and classification based problems. This algorithm uses various decision trees and combine the predictions of all these trees to provide the final outcome or prediction. Each decision trees coming next is made by using error function of past trees. It uses weak learning functions collectively to create a strong model for predictions.
Step 4: Evaluating a model
Model evaluation is a technique for quantifying a performance of the developed model. Confusion matrix, Accuracy, Precision, Recall, F1 score, and other approaches are used to evaluate the model performance. The confusion matrix is laid down in Figure 18 indicating types of classes used in performance evaluation.

Figure 18. Confusion matrix



1) Accuracy

Accuracy is the percentage of correct values predicted by a model out of total number of predictions. Its value ranges from 0 to 1.

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

2) Recall

Recall value provides the true positive rate for all the observations. It tells how many fractions of real positive values that are anticipated are positive.

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

3) Precision

Precision tells about the positive prediction value which means out of all positive predicted observations how much percentage of is actually positive.

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

4) F1 Score

F1 Score is basically the harmonic mean of precision and recall values. The F1 score is used to assess both recall and precision.

$$F1Score = \frac{2\mathrm{TP}}{2TP + FP + FN}$$

For the banking sector losing the right customer who deserves the loan in reality leads to a big financial loss and it affects the reputation of institute and its performance. Hence it is crucial to check performance of the model mainly on false positive rate.

RESULTS AND DISCUSSIONS

The following is how exploratory data analysis (EDA) is carried out. To achieve the important insights, Univariate, bivariate, and multivariate analyses is performed. For Univariate analysis various features like gender, marital status, qualification, loan term and many others are used. Analysis of sex variable provided that around 20 percent of total applicants are female only. Figure 19 to Figure 21 are the results obtained by analysis of credit history, applicant area and dependents features respectively.

Credit History is very important feature for banks to lend a loan. If Credit history is 0 it has very high default risk and if it is 1, there is very low default risk.

As visible in applicant area visualization plot most applicants are from Semi Urban area. In bivariate analysis, the focus is to relate every feature to the loan status as it is the target variable. All features are analyzed with loan status only.

In Figure 21 visualization of loan status versus total income is presented. Income is basic requirement to lend a loan. It assures banks that a person would not default the loan EMIs. But, here we can see high income is rejected as well. This may be because of credit history risk or other factors like having more dependents, term of loan would be for very long duration etc.

Figure 19. Credit history visualization

Figure 20. Applicant area visualization

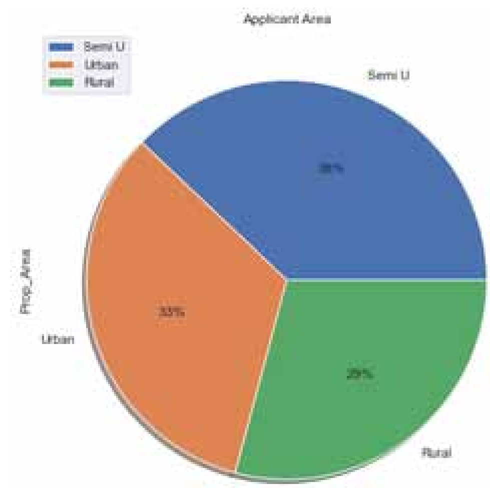
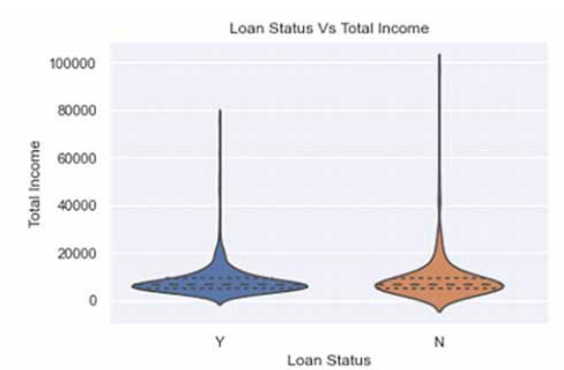


Figure 21. Loan status and Total income visualization



In Figure 22 visualization of credit history with loan status is displayed. Credit history is highly correlated to status of loan whether he or she is lent or not. Here out of 512 applicants having 1 credit history 420 are accepted and out of 102 applicants having 0 credit history 95 are rejected.

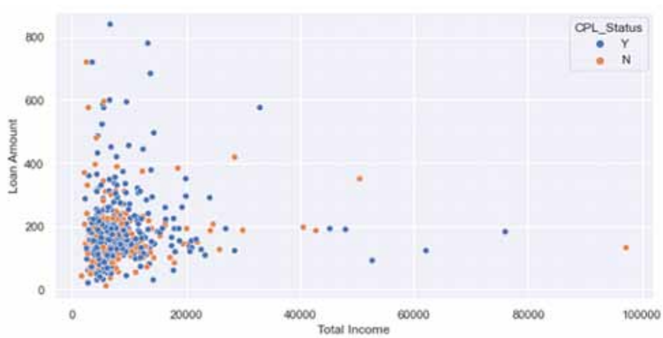
Figure 22. Credit history and Loan status visualization

For multivariate analysis different variables are considered altogether. Since credit history plays a key role in loan approval, credit history is analyzed with other two features, self-employment (SE) and loan status in Figure 23.

Figure 23. Credit history Self-employment and Loan status visualization

The impact of income and loan amount on loan status is analyzed in Univariate and bivariate analysis. Here in Figure 24 one can see the exact points of loan amount which has been accepted or rejected by bank with respect to income.

Figure 24. Total income Loan amount and Loan status visualization



These visualization plots revealed that Credit history is the key to grant a loan. The aftermath by analyzing the data are: those applicants with low credit scores are not provided the loan as they carry higher risk of loan default. Applicant requesting for smaller loan amount and having large income are having high chances of loan approval. Various other parameters like sex and marital status do not appear to be valued by any financial institution. The accuracies achieved by the test dataset for the chosen algorithms are provided from Figure 25 to Figure 28.

Only Gradient Boosting Classifier performed better in comparison to previous research findings. SVM provided worst efficiency overall despite performing hyper parameter tuning.

Figure 25. Logistic Regression

Classification Report of Logistic Regression

	precision	recall	f1-score	support
N	0.92	0.47	0.62	51
Y	0.83	0.99	0.90	134
accuracy			0.84	185
macro avg	0.88	0.73	0.76	185
weighted avg	0.86	0.84	0.82	185

Figure 26. Random Forest

Classification Report of Random Forest

	precision	recall	f1-score	support
N	0.71	0.49	0.58	51
Y	0.83	0.93	0.87	134
accuracy			0.81	185
macro avg	0.77	0.71	0.73	185
weighted avg	0.80	0.81	0.79	185

Figure 27. Support Vector Machine

Classification Report of SVM

	precision	recall	f1-score	support
N	0.00	0.00	0.00	51
Y	0.72	1.00	0.84	134
accuracy			0.72	185
macro avg	0.36	0.50	0.42	185
weighted avg	0.52	0.72	0.61	185

Figure 28. Gradient Boosting

Classification Report of GBC

	precision	recall	f1-score	support
N	0.78	0.49	0.60	51
Y	0.83	0.95	0.89	134
accuracy			0.82	185
macro avg	0.81	0.72	0.74	185
weighted avg	0.82	0.82	0.81	185

CONCLUSION AND FUTURE WORK

In this paper, an automatic loan approval model is developed using four classification algorithms. Previous research findings are thoroughly analyzed for selecting algorithms and dataset for proposed model. F1 score and other metrics are used to evaluate performance of model. Out of all the four algorithms Logistic Regression provides best accuracy overall. Logistic Regression, Gradient Boosting, Random Forest and SVM provided accuracy 0.84, 0.82, 0.80 and 0.72 respectively. Based on F1 score also, Logistic Regression comes out to be best classifier with accuracy of 0.90 for yes instances and 0.62 for no instances. The model predicted very well for both the training and testing dataset. However there were more instances of class 'yes' in the dataset. This can create class imbalance and ultimately biased output. According to results of EDA, Credit history, loan duration, no of dependents and applicant's area are the most critical factor in determining eligibility for loan approval.

This work can be extended further by predicting interest rates for individual customers based on the applicant's information. Also a credit risk score model can be generated that would help in predicting default probability for particular customer. Banks usually operates on diverse and varying data, such model requires more enhancements to adapt with various types of data. Research papers based on ensemble approach provided much better accuracies for model. Thus algorithms like supervised ANN, KNN with Min-Max normalization and ensemble approach are need to be explored more.

CONFLICT OF INTEREST

The authors of this publication declare there is no conflict of interest.

FUNDING AGENCY

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

Alsaleem, M. Y., & Hasoon, S. O. (2020). Predicting Bank Loan Risks Using Machine Learning Algorithms. *AL-Rafidain Journal of Computer Sciences and Mathematics*, *14*(1), 149–158.

Arun, K., Ishan, G., & Sanmeet, K. (2016). Loan approval prediction based on machine learning approach. *IOSR J. Comput. Eng*, *18*(3), 18–21.

Arutjothi, G. (2017). Prediction of Loan Status in Commercial Bank using Machine Learning Classifier. *2017 International Conference on Intelligent Sustainable Systems (ICISS), Iciss*, 416–419.

Athreyas, Kashyap, Bairy, & K. (2022). A Comparative Study of Machine Learning Algorithms for Predicting Loan Default and Eligibility. *Perspectives in Communication, Embedded-Systems and Signal-Processing, 5*(12), 116-118. https://doi.org/10.5281/zenodo.6543983

Bhanu, L., & Narayana, , D. (2021). Customer Loan Prediction Using Supervised Learning Technique. *International Journal of Scientific and Research Publications*, *11*, 403–407. doi:10.29322/IJSRP.11.06.2021.p11453

Bhattad, S., Bawane, S., Agrawal, S., Ramteke, U., & Ambhore, P. B. (2021, May-June). Loan Prediction using Machine Learning Algorithms. *International Journal of Computer Science Trends and Technology*, *9*(3), 143–146.

Diwate, Y., Rana, P., & Chavan, P. (2021). Loan Approval Prediction Using Machine Learning. *International Research Journal of Engineering and Technology*, *8*(5), 1741–1745.

Fati, S. M. (2021). Machine Learning-Based Prediction Model for Loan Status Approval. *Journal of Hunan University Natural Sciences*, *48*(10).

Gautam, K., Singh, A. P., Tyagi, K., & Kumar, M. S. (2020). Loan Prediction using Decision Tree and Random Forest. *International Research Journal of Engineering and Technology*, *7*(8), 853–856.

Gerritsen, R. (1999). Assessing loan risks: A data mining case study. *IT Professional*, *1*(6), 16–21.

Goyal, A., & Kaur, R. (2016a). A Survey on Ensemble Model for Loan Prediction. *International Journal of Advance Research and Innovative Ideas in Education*, *2*(1), 623–628.

Goyal, A., & Kaur, R. (2016b). Accuracy prediction for loan risk using machine learning models. *Int. J. Comput. Sci. Trends Technol*, *4*(1), 52–57.

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

Hamid, A. J., & Ahmed, T. M. (2016). Developing prediction model of loan risk in banks using data mining. *Machine Learning and Applications: An International Journal*, *3*(1), 1–9. doi:10.5121/mlaij.2016.3101

Hsieh, N. C., & Hung, L. P. (2010). A data driven ensemble classifier for credit scoring analysis. *Expert Systems with Applications*, *37*(1), 534–545.

Jency, X. F., Sumathi, V. P., & Sri, J. S. (2018). An exploratory data analysis for loan prediction based on nature of the clients. *International Journal of Recent Technology and Engineering*, *7*(4), 176–179.

Kaarthik, K., Dharanidharan, G., Navalarasu, R. B., & Sabarinathan, G. (2021). Machine Learning Based Loan Prediction System Using Svm and Knn Algorithms. *Turkish J. Physiother. Rehabil*, *32*(2), 3214–3219.

Kadam, A., Nikam, S., Aher, A., Shelke, G., & Chandgude, A. (2021). Prediction for Loan Approval using Machine Learning Algorithm. *International Research Journal of Engineering and Technology*, *8*(04), 4089–4092.

Karthiban, R. (2019). A Review on Machine Learning Classification Technique for Bank Loan Approval. *2019 International Conference on Computer Communication and Informatics (ICCCI)*, 1–6.

Khadse, K. (2020). Applications of machine learning in loan prediction systems. *Linguistica Antverpiensia,* (3), 3658 – 3674.

Kotsiantis, B. (2007). Supervised Machine Learning: A Review of Classification Techniques. *Informatica (Vilnius)*, *31*, 249–268.

Madaan, M., Kumar, A., Keshri, C., Jain, R., & Nagrath, P. (2021). Loan default prediction using decision trees and random forest: A comparative study. *IOP Conference Series. Materials Science and Engineering*, *1022*(1). Advance online publication. doi:10.1088/1757-899X/1022/1/012042

Madane & Nanda. (2019). Loan Prediction using Decision tree. *Journal of the Gujrat Research History, 21*(14s).

Maheswari, P., & Narayana, C. V. (2020, October). Predictions of Loan Defaulter-A Data Science Perspective. In *2020 5th International Conference on Computing, Communication and Security (ICCCS)* (pp. 1-4). IEEE.

Murthy, P. L. S., Shekar, G. S., Rohith, P., & Reddy, G. V. V. (n.d.). Loan Approval Prediction System Using Machine Learning. *Journal of Innovation in Information Technology*, 21–24.

Patil, P. S., & Dharwadkar, N. V. (2017, February). Analysis of banking data using machine learning. In *2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)* (pp. 876-881). IEEE.

Prasad, K. G. S., Chidvilas, P. V. S., & Kumar, V. V. (2019). Customer Loan Approval Classification by Supervised Learning Model. *International Journal of Recent Technology and Engineering*, *8*(4), 9898–9901. doi:10.35940/ijrte.d9275.118419

Ramesh, R. (2017). Predictive analytics for banking user data using AWS Machine Learning cloud service [Análisis predictivo de los datos de los usuarios bancarios mediante los servicios en la nube de aprendizaje automático de AWS]. *Proceedings of the 2017 2nd International Conference on Computing and Communications Technologies, ICCCT 2017*, 210–215.

Raviteja, , & Santosh, . (2021). Intelligent defaulter Prediction using Data Science Process. *Journal of Physics: Conference Series*. Advance online publication. doi:10.1088/1742-6596/1916/1/012001

Rawate, R., & Tijare, P. A. (2017). Review on prediction system for bank loan credibility. *Int. J. Adv. Eng. Res. Dev.*, *4*(12), 860–867.

Salvi, R., Ghule, R., Sanadi, T., & Bhajibhakare, M. (2021). Home loan data analysis and visualization. *International Journal of Creative Research Thoughts*, *9*(1), 3131–3135.

Sheikh, M. A., Goel, A. K., & Kumar, T. (2020). An Approach for Prediction of Loan Approval using Machine Learning Algorithm. *Proceedings of the International Conference on Electronics and Sustainable Communication Systems, ICESC 2020, Icesc*, 490–494. https://doi.org/ doi:10.1109/ICESC48915.2020.9155614

Singh, V., Yadav, A., Awasthi, R., & 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.

Sivasree, M. S., & Rekha Sunny, T. (2015, September). Loan Credibility Prediction System Based on Decision Tree Algorithm. *International Journal of Engineering Research & Technology*, *4*(9). Advance online publication. doi:10.17577/IJERTV4IS090708

Somayyeh, , & Abdolkarim, . (2015). Natural Customer Ranking of Banks in Terms of Credit Risk by Using Data Mining A Case Study: Branches of Mellat Bank of Iran. *Journal UMP Social Sciences and Technology Management*, *3*(2), 307–316.

Sudhakar, , & Reddy, . (2016). Two Step Credit Risk Assessment Model for Retail Bank Loan Applications Using Decision Tree Data Mining Technique. *International Journal of Advanced Research in Computer Engineering & Technology*, *5*(3), 705–718.

Sudhamathy, G. (2016). Credit risk analysis and prediction modelling of bank loans using R. *Int. J. Eng. Technol*, *8*(5), 1954–1966. doi:10.21817/ijet/2016/v8i5/160805414

Supriya, P., Pavani, M., & Saisushma, N. (2019). Loan Prediction by using Machine Learning Models. *International Journal of Engineering and Techniques*, *5*(2), 144–148.

Tiwari, A. K. (2018). Machine learning application in loan default prediction. *Machine Learning, 4*(5), 1–5.

Vaidya, A. (2017). Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval. *8th International Conference on Computing, Communications and Networking Technologies, ICCCNT 2017*. https://doi.org/ doi:10.1109/ICCCNT.2017.8203946

Vandana Sharma is an M.Tech Computer Science and Engineering student in Department of Computer Engineering at J. C. Bose University of Science and Technology, YMCA, Faridabad, India. She has done M.Sc in Computer Science from Department of Computer Science and Applications, Kurukshetra University, Haryana, India. She has completed her B.Sc Hons Computer Science from Shahhed Rajguru College of Applied Sciences for Women, University of Delhi, India.

Rewa Sharma is working as an Assistant Professor in Department of Computer Engineering at J.C. Bose University of Science and Technology, YMCA Faridabad. She has completed her PhD in Computer Engineering from Banasthali University, Rajasthan, India. She has teaching experience of 10 years. She has presented and published many papers in various National/ International conferences and reputed journals. Her research interests include Wireless Sensor Networks, Internet of Things and Machine Learning.