Loan Prediction Using Machine Learning Algorithm

Dr. Jim Mathew Philip[1], Harrish Dhaithyaka[2], Gopinath S[3], Ajaikumar N[4]
[1]Assistant Professor, Dept of CSE
[2, 3, 4]Dept of CSE
[1, 2, 3, 4] Sri Ramakrishna Institute of Technology

Abstract- *Loans account for a huge chunk of bank profits. Even though many individuals are looking for loans. Finding a legitimate candidate who will return the loan is difficult. Several misunderstandings may arise while selecting the real candidate when the process is carried out manually. As a result, we are creating a machine learning based loan prediction system that will choose the qualified applicants on its own. Although banks in our financial system can offer a wide range of goods, their primary source of income comes from their credit lines. They may profit from the interest on those loans, therefore. The profitability or loss of a bank is mostly determined by the loans it makes, namely whether its clients are making their loan repayments. The bank's non-performing assets can be decreased by foreseeing loan defaulters. Because of this, it is crucial to examine this phenomenon. The subject of reducing loan default may be studied using a wide variety of techniques, according to earlier research from this era. The nature of the various approaches must be studied in order to compare them, though, as accurate forecasts are crucial for maximizing earnings. Predicting loan defaulters is a challenging subject that is studied using a crucial predictive analytics methodology. Both bank employees and applicants will benefit from this. The loan sanctioning process will be completed in a much shorter amount of time. We are employing a machine learning method to forecast the loan data in this project.*

Keywords- Loan Prediction, Machine Learning, Supervised Learning, Random Forest.

I. INTRODUCTION

Most banks and other financial institutions' main line of business is loan distribution. The banks are having a difficult time issuing loans due to the rise in application volume. A more laborious step for the bank to complete manually is the identification and verification of potential borrowers. In order to identify potential borrowers from the enormous pool of loan applications with the use of machine learning algorithms, this project intends to develop a subsystem that can relate to the current banking system.

1.1 Background History:

A loan is the primary source of income for banks. The money the bank makes from the loan's accounts for the vast bulk of its profits. There is no certainty that the chosen hopeful is the right hopeful, even if the bank approves the loan after a drawn-out verification and testifying process. This process takes more time when carried out manually. We can predict if a certain candidate is secure or not, and machine literacy has automated the entire testimony process. Loan forecasting has several advantages for both bank employees and potential borrowers.

1.2 Problem Statement:

To increase accuracy and reduce fraud, it is necessary to create and deploy a system that uses machine learning and data mining to forecast whether a user will receive a loan from a bank or not. All around the nation, banks, home finance companies, and certain NBFCs deal with different loan kinds including mortgages, personal loans, business loans, etc. These businesses operate in rural, semi-urban, and urban settings. These businesses verify a customer's eligibility for a loan after the consumer applies for one. The use of machine learning techniques in this research offers a way to automate this procedure. In order to apply for a loan, the consumer will complete out an online form. This form includes information on the applicant's sex, marital status, qualifications, dependents' specifics, annual income, loan amount, and credit history, among other things. A machine learning system can automate this procedure by first identifying the customer groups that qualify for loan amounts so that the bank can concentrate on these clients.

1.3 Scope of the Project:

The goal of this strategy is to make the selection of qualified applicants quick, simple, and rapid. It may provide banks special benefits. The credit forecasting system may automatically calculate the weights for each feature that participates in credit processing, and the new test data will process the same characteristics for the supplied weights. To establish if the applicant can approve the loan, the model can set a deadline. We can easily access certain applications and prioritize our checks thanks to credit analysis. Because this technology is just for bank and financial business management

authorities, the whole forecasting process is conducted in confidence, and no stakeholders are permitted to change the process.

1.4 Existing System:

Before making loans available to qualified applicants, bank staff carefully check each applicant's information. The process of reviewing each applicant's data is time-consuming. The artificial neural network model for predicting the credit risk of a bank. A feedforward back propagation neural network is used to forecast the credit default. the practice of integrating two or more classifiers to create an ensemble model for better prediction. The random forest method was used with the bagging and boosting approach. The goal of the classifier process is to improve the data's effectiveness and efficiency. The author of this book discusses several enable strategies for multiclass classification as well as binary classification. The authors' novel approach for ensemble is COB, which provides efficient classification performance but also compromises with noise and outlier classification data. Eventually, they came to the conclusion that the ensemble-based approach enhances the training data set's outcomes. It takes a lot of time and effort to review all of the applicants' information. Human mistake might happen as a result of meticulously examining every detail. Loans could be assigned to applicants who aren't qualified.

II. SYSTEM SPECIFICATION

2.1 Dataset Description:

The bank loan prediction dataset is taken from Kaggle competition dataset which belong to different age group and gender of applicants. There are totally 13 attributes such as Loan_id, Gender, Married, Dependents, Education, Self_Employed, Applicant Income, Co-applicant Income, Loan Amount, Loan Amount Term, Credit History, Property Area, Loan Status.

Variable	Description
Loan_ID	Unique Loan ID
Gender	Male/ Female
Married	Applicant married (Y/N)
Dependents	Number of dependents
Education	Applicant Education (Graduate/ Under Graduate)
Self_Employed	Self employed (Y/N)
ApplicantIncome	Applicant income
CoapplicantIncome	Coapplicant income
LoanAmount	Loan amount in thousands
Loan_Amount_Term	Term of loan in months
Credit_History	credit history meets guidelines
Property_Area	Urban/ Semi Urban/ Rural
Loan_Status	(Target) Loan approved (Y/N)

There are totally 981 records of applicants with the values of their concerning attributes in categorical and numerical data. In pre-processing and feature engineering of the data, we can handle the missing value and normalize the data so we can further process it into ML algorithm. The dataset is further divided into training and testing. The model is trained on machine learning algorithms and predict the system on test data.

2.2. Performance Metrics:

To measure the success rate of the model, the best metric was the precise prediction of a loan default. The profitability of the investor or the financial institution depends on the decision of the model. These are the error types (false positive, false negative, true positive and true negative) which were used for determining a conservative evaluation of the loan default rate.

- **True Positive (TP):** These are cases where the model predicts that the loan will be repaid and the original outcome in the dataset is the same.
- **True Negative (TN):** These are cases where the model predicts that the loan will default and the original outcome in the dataset is the same.
- **False Positive (FP):** These are cases where the model predicts that the loan will be repaid, but the original outcome in the dataset is that it will default.
- **False Negative (FN):** These are cases where the model predicts that the loan will default, but the original outcome in the dataset is that it will not default.

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

Fig 2.1(Performance Metrics)

The above figure is a confusion matrix which displays the error types. Using this, other performance metrics like precision, accuracy, true positive rate, false positive rate are calculated. In the above figure, ‘n’ denotes the total number of cases. There are two possible Predicted and Actual classes: ‘YES’ and ‘NO’. Actual ‘Yes’ means that the loan was originally paid off and Actual ‘NO’ means the loan

wasn't paid off. Predicted 'YES' means that the model classifier predicted that the loan would be paid off and Predicted 'NO' means that the model classifier predicted that the loan would not be paid off. Thus, the classifiermade a total of 165 predictions that were equal to the number of actual outcomes. 'TN', 'FP', 'FN' and 'TP' denote True Negatives, False Positives, False Negatives and True Positives respectively.

The best metric to evaluate the model is the precision of the algorithm to predict whether a customer is going to repay the loan. This is achieved by training the model on the training dataset and then predicting (based on the features) the faithfully paying customers from those that default. The training results in being able to measure the precision, practicality and realism of the model. The precision, accuracy, true positive rate and false positive rate of the model are measured as follows:

- Precision= (True Positive/ (True Positive + False Positive))
- Accuracy= (True Positive/ (True Positive+ False Positive+ True Negative + False Negative)
- True Positive Rate: (True Positive/ (True Positive + False Negative))
- False Positive Rate: ((False Positive/ (False Positive + True Positive)))

So, for the confusion matrix in Fig. 4.1, the precision, accuracy, true positive rate and false positive rate are calculated as follows:

1. Accuracy: (TP + TN) / Total
2. Precision: TP/ Predicted YES
3. True Positive Rate: TP/ Actual YES
4. False Positive Rate: FP/ Actual NO

2.3 Methodology and Flowchart:

The goal of this loan prediction system is to develop an automated system that, using the customer's data, forecasts the loan approval status. This loan prediction technology will be made available as an independent online service that can relate to current bank applications.

The loan prediction system's flowchart is shown in Figure 2.2. Training data and Testing data make up the two categories of the data collection. The testing data will be used to assessthe model's correctness, and the training data will be used to train the model using a machine learning method. To extract the crucial characteristics from the raw dataset, necessary pre-processing and feature extraction activities must be done. The model is trained using a variety of supervised learning algorithms after feature extraction, and the best approach is chosen.

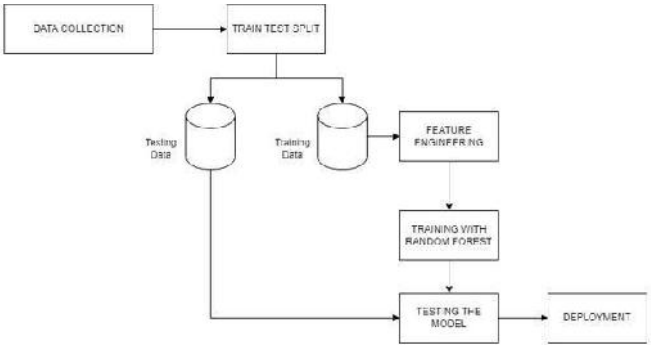

Fig 2.2(Flowchart)

2.4 Data Cleaning and Pre-processing:

Data cleaning and pre-processing are the methods that eliminate noise from the data and put the raw data into a suitable form so that a machine learning algorithm may easily train on it. Several of the cells in the downloaded csv file are filled with null values. The column's mean values must be used to fill up these null values.

```
Loan_ID              0
Gender              13
Married              3
Dependents          15
Education            0
Self_Employed       32
ApplicantIncome      0
CoapplicantIncome    0
LoanAmount          22
Loan_Amount_Term    14
Credit_History      50
Property_Area        0
Loan_Status          0
dtype: int64
```

Fig 2.3(Number of Null Valued Cell in each columns)

As shown in above Figure 2.3, the attributes "Gender " has 13 missing values, "Married" has 3 missing values, "Dependents" has 15 missing values, "Self_Employed" has 32 missing values,"LoanAmount" has 22 missing values, "Loan_Amount_Term" has 14 missing values and "Credit_History" has 50 missing values.

In the dataset the attributes such as Gender, Married, Dependents, Self_Employed, and Credit_History has categorical data such as Male/Female,0/1 etc…, so these values can be replaced with the mode values of respective column.

The attributes such as LoanAmount and Loan_Amount_Term contains continuous values so Iterative

Imputer along with Random Forest Regression is used to fill the null values.

2.5 Exploratory Data Analysis:

Exploratory data analysis (EDA), which frequently makes use of data visualization techniques, is used to examine and summarize large data sets. It makes it simpler to find patterns, identify anomalies, test hypotheses, or verify assumptions by determining how to alter data sources to achieve the answers you need. EDA offers a deeper knowledge of data, variables, and the relationships between them and is generally used to examine what data might disclose beyond the formal modelling or hypothesis testing assignment. It can also assist in determining the suitability of the statistical methods you are contemplating using for data analysis.

2.5.1 Univariate Analysis:

Let us analyse the end result of loan prediction weather the loan application got accepted or not.

Fig 2.4 (Univariate Analysis of Target Variable)

As shown in figure 2.4 we can conclude that almost 69% of the loan applications are getting accepted.

Categorical Features:

- Gender(M/F)
- Married(Y/N)
- Self_Employed(Y/N)
- Credit_History(0/1)

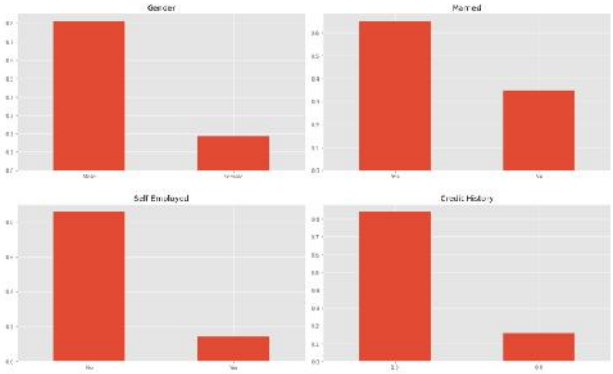

Fig 2.5(Univariate Analysis of Categorical Variable)

Inference:

- Almost 80% of the loan applicants are males as per training set.
- About 70% of the applicants are married.
- Nearly 75% of the loan applicants are graduates.
- Nearly 85 to 90% of the loan applicants are self-employed.

Ordinary Features:

- Dependents
- Education
- Property or Area Background.

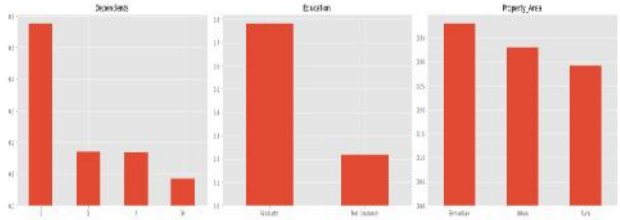

Fig 2.6(Univariate Analysis of Ordinary Features)

Inference:

- Almost 58% of the applicants do not have any dependents.
- Highest number of applicants are from semi-urban area, followed by urban area.
- Around 80% of the applicants are graduates.

Numerical Features:

- Applicant Income
- Co-applicant Income

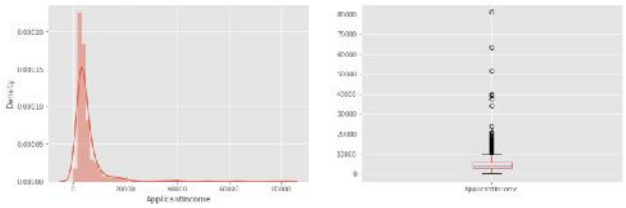

Fig 2.7(Univariate Analysis of Applicant Income)

Inference: It can be inferred that data in the applicants income is towards left and it is normally distributed. The box plot confirms that there is a presence of outliers. The outliers can be inferred as income disparity in the society.

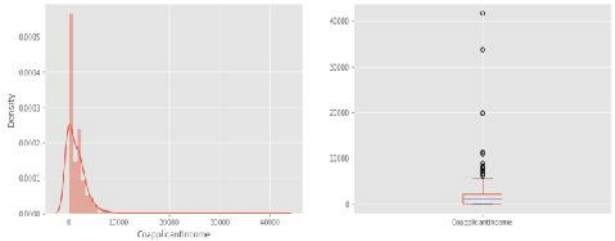

Fig 2.8(Univariate Analysis of Co-applicant Income)

Inference: The above plot infers that the co-applicant income is lesser than the applicants income and lies between the range 5000 to 15000 and has some outliers.

2.5.2 Bivariate Analysis:

Categorical Independent vs Target:

Gender vs Loan Status

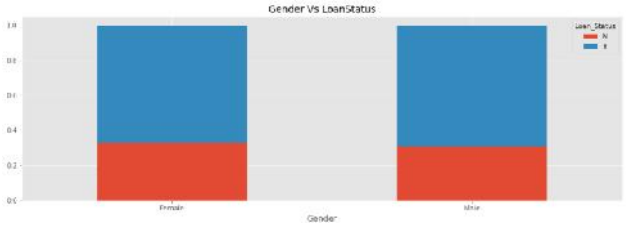

Fig 2.9(Bivariate Analysis of Gender vs Loan Status)

Inference: As per the training data there difference between male and for loan approval rates

.

Marriage vs Loan Status:

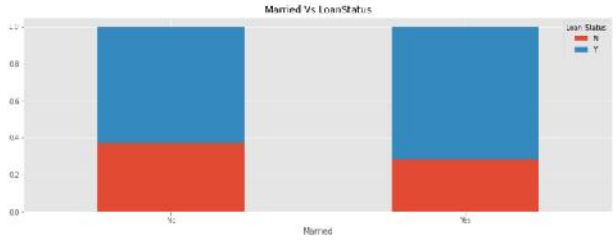

Fig 2.10(Bivariate Analysis of Marriage vs Loan Status)

Inference: Married applicants have slightly higher chance for loan approvals.

Dependency vs Loan Status:

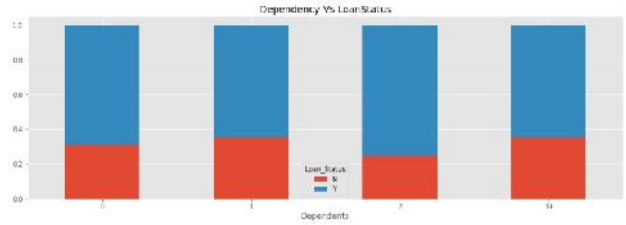

Fig 2.11(Bivariate Analysis of Dependency vs Loan Status)

Inference: Applicants with no dependents or 2 dependents have higher opportunity of loan approval. But this does not correlate well.

Education vs Loan Status:

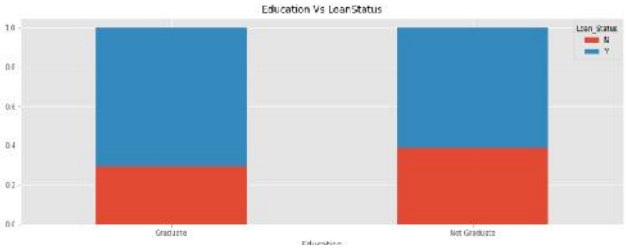

Fig 2.12(Bivariate Analysis of Education vs Loan Status)

Inference: Graduates have higher chance of loan approval compared to non-graduates.

Self-Employed vs Loan Status:

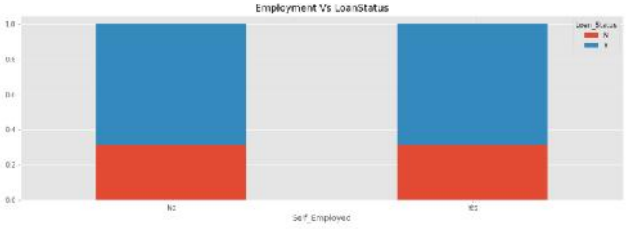

Fig 2.13(Bivariate Analysis of Self-Employed vs Loan Status)

Inference: Self_Employed applicants have slightly less chance for loan approval, but the situation is not that bad.

Credit History vs Loan Status:

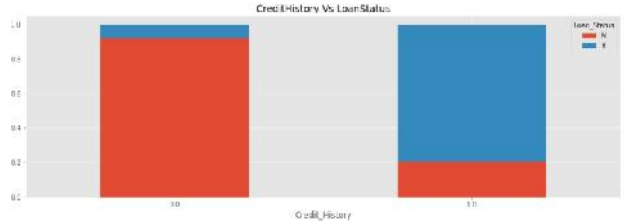

Fig 2.14(Bivariate Analysis of Credit History vs Loan Status)

Inference: Applicants with Credit_History 1 have higher chace for loan approval.

Property Area vs Loan Status:

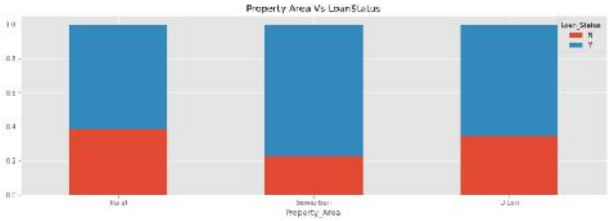

Fig 2.15(Bivariate Analysis of Property Area vs Loan Status)

Inference: The applicants from the semi urban area have slightly higher chance of getting loan approved

Numerical Data vs Target:

Applicant Income vs Loan Status:

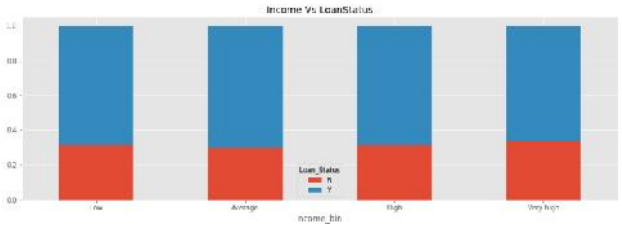

Fig 2.16(Bivariate Analysis of Applicant Income vs Loan Status)

Inference: It can be inferred that applicants income does not affect the chances of loan approval which contradicts out hypothesis in which we assume, if the applicants income is high the chance of loan approval is high.

Co-applicant Income vs Loan Status:

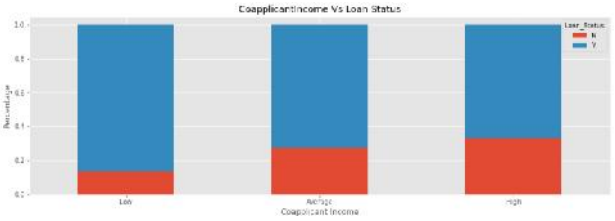

Fig 2.17(Bivariate Analysis of Co-applicant Income vs Loan Status)

Inference: This data shows that lesser the coapplicant income higher the chance of loan approval. But we cannot come to the conclusion with the help of this plots, because most of the applicants do not have coapplicants so the coapplicant income will be 0. So it cannot be inferred as a standalone attribute. So we can derive a new attributes as Total_Income, which will be the sum of ApplicantIncome and CoapplicantIncome.

Total Income vs Loan Status:

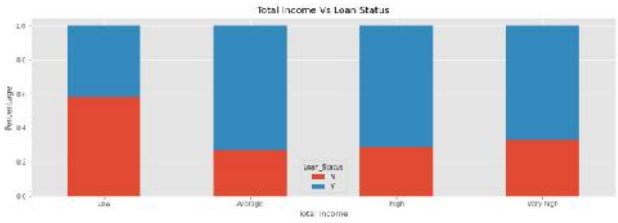

Fig 2.18(Bivariate Analysis of Total Income vs Loan Status)

Inference: We can see that the proportion of loan getting approved is low when the income is low compared to Average, High and Very High.

Loan Amount vs Loan Status:

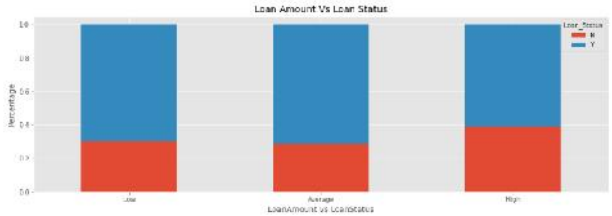

Fig 2.19(Bivariate Analysis of Loan Amount vs Loan Status)

Inference: The proportion of the approved loan is higher for low and average amount, while comparing to High amount. This attribute supports our hypothesis that chance of approval of the loan will be high when the loan amount is less.

2.5.2 Correlation with Heat Map:

A heat map is a data visualization technique that shows magnitude of a phenomenon as color in two dimension. The variation in color may be due to hue or intensity, giving

obvious visual cues to the reader about how the phenomenon is clustered or varies over space. There are two fundamentally different categories of heat map.

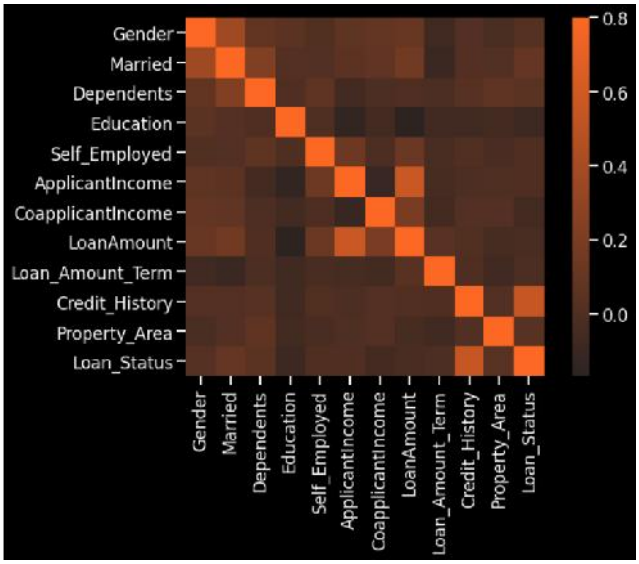

Fig 2.20(Heat Map for the dataset)

The variable with dark color represents the higher correlation. As shown in fig 4.20, we can infer that the most correlated variables are (ApplicantIncome – LoanAmount) and (Credit_History – Loan_Status). LoanAmount is also correlated with CoapplicantIncome.

2.6 Model Training:

Random Forest is a classifier consisting of a collection of tree-structured classifiers {h(x, Θk) k=1,2, ….}, where the {Θk } are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x.

An ensemble of decision trees is produced using Random Forest. Breiman used the randomization strategy, which integrates well with bagging or random subspace approaches, to produce diversity among base decision trees. Breiman used the Random Forest and performed the following procedures to create each individual tree: If there are N records in the training set, then N records from the original data are picked at random but with replacement. This is known as a bootstrap sample. This sample will serve as the tree's training set. There will be an increasing number m forest if there are M input variables. Every tree is developed to its full potential. Pruning is not done. In this way, multiple trees are induced in the forest; the number of trees is pre-decided by the parameter Ntree. The depth of the tree can be controlled by a parameter nodesize which is usually set to one.

Once the forest is trained and built as explained previously, to classify a new instance, it is run across all the trees generated in the forest. Each tree provides a category that applies to all the trees that have developed in the forest. A new instance is classified by each tree, and this classification is recorded as a vote. The classes with the most votes overall (majority voting) are proclaimed as the classifications for new instances after adding the votes from each tree.

About one-third of the original instances are lost throughout the forest-building process when a bootstrap sample set is created by sampling with replacement for each tree. OOB (Out Of Bag) data refers to this collection of instances. The process of estimating each individual tree's error in the forest using its own OOB data collection is known as OOB error estimation. The Random Forest method also has a built-in capability to determine the proximity and relevance of variables. Outliers and missing values are replaced using the proximities.

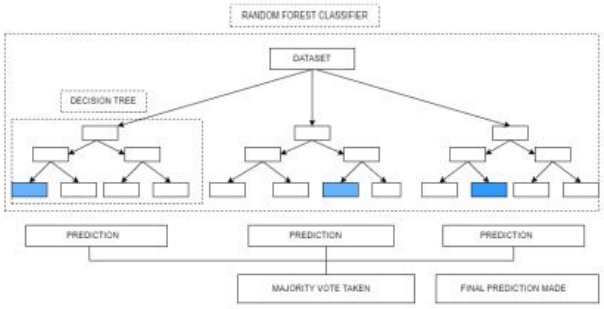

Fig 2.21(Random Forest Classifier)

As shown in above Figure 2.21 the random forest algorithm comprises of n number of decision trees and the final prediction is made based on majority vote taken in the forest.

Accuracy in Random Forest Algorithm:

The Generalization error (PE*) of Random Forest is given as,

$$PE^{*} = P_{x,y}(mg(X,Y)) < 0$$

Where mg(X,Y)is Margin function. The Margin function measures the extent to which the average number of votes at (X, Y) for the right class exceeds the average vote for any other class. Here X is the predictor vector and Y is the classification.

$$mg(X,Y) = av_k I(h_k(X) = Y) - \max_{j \neq Y} av_k I(h_k(X) = j)$$

Here I(.) is Indicator function.

Margin is directly proportional to confidence in the classification.

Strength of Random Forest is given in terms of the expected value of Margin function as,

$$S = E_{X,Y}(mg\ (X,\ Y))$$

The generalization error of ensemble classifier is bounded above by a function of mean correlation between base classifiers and their average strength (s) . If ρ is mean value of correlation, an upper bound for generalization error is given by,

$$PE^* \leq \rho\ (1 - s^2) / s^2$$

III. CONCLUSION

3.1 Summary:

We can now say that, for a loan prediction system, the Random Forest method is more accurate when compared to other supervised learning algorithms like Linear Regression, KNN Algorithm, Decision Tree, etc. The project's goal is to create a system that is simple to connect with the current classifying banking programs. By eliminating the need for manual application verification, it will enable loan-eligible applicants to receive quick loan approval. By minimizing the effort required for manual verification, which may waste a lot of time and effort on the part of the bank and may result in human mistake, it may assist banks in swiftly identifying potential borrowers.

3.2 Future Works:

The accuracy of the model may be increased in the future by collecting more real-time data from different financial institutions and by focusing more on feature engineering to extract and assess new variables for a more precise prediction of loan evaluation. To increase accuracy, the data collected during the feature engineering step may be taught using a variety of supervised learning algorithms.

REFERENCES

[1] Anshika Gupta, Vinay pant, Sudhanshu kumar, Pravesh Kumar Bansan - “Bank Loan Prediction System Using machine learning” 9th International ConferenceSystem Modeling and Advancement in Research Trends(SMART 2020).

[2] Mohammad Ahmad sheikh, Amit Kumar Goel, Tapas Kumar - “An approach for prediction of loan approval using Machine Learning Algorithm” Proceedings of the International Conference on Electronics and Sustainable Communication Systems(ICESE 2020).

[3] Pidikiti Supriya, Myneedi pavini, Nagarapu Saisushma - “ Loan Prediction by using Machine Learning Models” International Journal of Engineering and Techniques.(IJARSCT 2022)

[4] Anuja Kadam, Pragati Namde, Sonal Shirke, Siddhesh Nandgaonkar, Dr.D.R.Ingle - “Loan Credibility Prediction System using data mining techniques” International Research Journal of Engineering and Technology(IRJET 2021).

[5] Jayan Kokru, Abhijeet Shrikant Ghodke, Prathmesh Chavan, Siddharth Chand, prof. Sagar Mane - “Bank Loan Approval Prediction System Using Machine Learning Algorithms" International Journal of Advanced Research in Science, Communication and Technology (IJARSCT 2022)

[6] Vishal Singh, Ayushman Yadav, Rajat Awasthi, N. Partheeban – “Prediction of Modernized Loan Approval System Based on Machine Learning Approach” International Conference on Intelligent Technologies (CONIT 2021).

[7] Ugochukwu. E. Orji, Chikodili. H. Ugwuishiwu, Joseph. C. N. Nguemaleu, Peace. N. Ugwuanyi – “Machine Learning Models for Predicting Bank Loan Eligibility” International Conference on Disruptive Technologies for Sustainable Development (NIGERCON 2022).