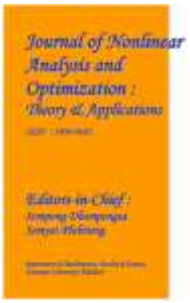
Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 15, Issue. 2 : 2024
ISSN : **1906-9685**



LOAN APPROVAL PREDICTION USING MACHINE LEARNING

T. KRISHNA SAI PRIYA (2nd YEAR OF M.SC., AUCE, Andhra University)

Dr. N.P LAVANYA KUMARI (Assistant professor(c) AUCE, Andhra University)

ABSTRACT

Predicting loan defaulters is a problem that is studied using a critical predictive analytics approach: Data is gathered from Kaggle for analysis and forecasting purposes. Machine learning algorithms have been used to create models, and various performance metrics have been calculated. Sensitivity and specificity, two performance metrics, are used to compare the models. The model produces diverse results, as evidenced by the end results. Therefore, by assessing their potential of loan default, the appropriate consumers can be quickly identified for loan issuing through the use of a machine learning algorithm technique. The model's conclusion is that a bank should evaluate a customer's other characteristics as well, as these factors are crucial in determining whether to give credit and in identifying potential loan defaulters. This goes beyond simply focusing on lending to wealthy clients.

keyword:- loan, machine Learning,deafulters, credit, Random Forest,Decesion Tree, Logistic Regression

1. INTRODUCTION

Financial institutions oversee a wide range of loan products, including personal, education, auto, and home loans. It is seen in both urban and rural areas. Upon a customer's initial loan application, the Finance Company verifies the applicant's eligibility for acceptance. A form that asks for details on the applicant's credit history, income, loan amount, number of dependents, gender, marital status, and education must be filled out. Consequently, those details are used as input into a robust model tha verifies an applicant's eligibility for a loan application. In this case, the other parameters are predictors, and the objective variable is the applicant's "Loan Status". After the machine learning model is built, a web application will be developed with an interface that allows users to rapidly ascertain their loan eligibility by supplying specific facts. Through the use of prior bank customers' data, this project has granted loans based on a set of criteria. To produce dependable results, the machine learning model is trained using the recorded data. Our primary project goal is to ascertain the client's loan eligibility. The Random Forest, Decision Tree, and Naive Bayes algorithms are used to predict loan eligibility. The data are initially cleansed to make sure the data set is free of missing values. It is unavoidable to give credit to both individuals and corporations in order for developing economies like India to run smoothly. It is quite difficult for banks

and non-banking financial firms (NBFCs) with limited capital to devise a standard resolution and safe procedure to lend money to its borrowers for their financial needs as more and more clients ask for loans in these institutions. Along with this, the stock price of NBFC inventories has significantly declined recently. It has fueled a contagion that has recently negatively impacted the benchmark by spreading to other banking stocks. This essay aims to reduce the risk associated with choosing the right borrower who would be able to make timely loan repayments and keep the bank's nonperforming assets (NPA) on hold. This is accomplished by integrating historical data from customers who obtained bank loans into a machine learning model that has been trained and may produce accurate results. Determining whether or not it will be safe to assign the loan to a specific individual is the main goal of the article. The components of this study are as follows: (i) Data Collection; (ii) Data Cleaning; and (iii) Performance Evaluation. It was discovered through experimental testing that the Naïve Bayes model performs better. Through tests conducted in the lab, the Naïve Bayes model has better performance than other models in terms of loan forecasting.

2. LITERATURE SURVEY

[1] Raj, J. S., & Ananthi, J. V., "Recurrent neural networks and nonlinear prediction in support vector machine" Journal of Soft Computing Paradigm (JSCP), 1(01), 33-40, 2019.

The detection of edges is the one of the important stages in the application, associated with the machine vision, computer vision and the image processing. It is most commonly and highly preferred in the area were the extraction or the detection of the attribute are necessary. As the manual methods of diagnosis in the medical images acquired from the CT (computed tomography) and the MRI (magnetic resonance images) are very tedious and as well as time consuming, the paper puts forth the methodology to detect the edges in the CT and the MRI by employing Gabor Transform as well as the soft and the hard clustering. This proposed method is highly preferred among the image with dynamic variations. The technique used in the paper is evaluated using 4500 instance of the MRI and 3000 instance of CT. The results on the basis of the figure of merit (FOM) and Misclassification rate (MCR) are compared with other standard approaches and the performance was evinced.

[2] X.Frencis ,JensyV.P.Sumathi,Janani Shiva Shri, "An exploratory Data Analysis for Loan Prediction based on nature of clients", International Journal of Recent Technology and Engineering (IJRTE),Volume-7 Issue-4S, November 2018.

In India, the number of people applying for the loans gets increased for various reasons in recent years. The bank employees are not able to analyze or predict whether the customer can payback the amount or not (good customer or bad customer) for the given interest rate. The aim of this paper is to find the nature of the client applying for the personal loan. An exploratory data analysis technique is used to deal with this problem. The result of the analysis shows that short term loans are preferred by majority of the clients and the clients majorly apply loans for debt consolidation. The results are shown in graphs that helps the bankers to understand the client's behaviour. Keywords - Loan analysis, exploratory data analysis technique, client's analysis, financial categories analysis the term banking can be defined as receiving and protecting money that is deposited by the individual or the entities. This also includes

lending money to the people which will be repaid within the given time. Banking sector is regulated in most of the countries as it is the important factor in determining the financial stability of the country. The provision of banking regulation act allows public to obtain loans. Loans are good sum of money borrowed for a period and expected to be paid back at given interest rate. The purpose of the loan can be anything based on the customer requirements. Loans are broadly divided as open-ended and close-ended loans. Open-ended loans are the loans for which the client has approval for a specific amount. Examples of open-end loans are credit cards and a home equity line of credit (HELOC). Close-ended loans decreases with each payment. In other words, it is a legal term that cannot be modified by the borrower. Personal loans, mortgages, auto payments, instalment loan and student loans are the most common examples of close-ended loans. Secured or collateral loan are those loans that are protected by an asset. Houses, Vehicles, Savings accounts are the personal properties used to secure the loan. Unsecured loans are also known as personal or signature loans. Here the lender believes that the borrower can repay the loan based on financial resources possessed by the borrower. Liquidity risk is the risk that arises from the lack of marketability of an investment that cannot be bought or sold quickly enough to prevent or minimize a loss. The interest rate risk is the risk in which the interest rates priced on loans will be too low to earn the bank money. Revised Version Manuscript Received on 25 November, 2018. Ms.X.Francis Jency,

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

With the enhancement in the banking sector lots of people are applying for bank loans but the bank has its limited assets which it has to grant to limited people only, so finding out to whom the loan can be granted which will be a safer option for the bank is a typical process. So in this project we try to reduce this risk factor behind selecting the safe person so as to save lots of bank efforts and assets. This is done by mining the Big Data of the previous records of the people to whom the loan was granted before and on the basis of these records/experiences the machine was trained using the machine learning model which give the most accurate result. The main objective of this project is to predict whether assigning the loan to particular person will be safe or not. This paper is divided into four sections (i)Data Collection (ii) Comparison of machine learning models on collected data (iii) Training of system on most promising model (iv) Testing. In this paper we are predict the loan data by using some machine learning algorithms they are classification, logic regression, Decision Tree and gradient boosting.

3. IMPLEMENTATION STUDY

Banks, Housing, Finance Companies and some NBFC (non-banking financial company) deal in various types of loans like housing loan, personal loan, business loan etc in all over the part of countries. These companies have existence in Rural, Semi Urban and Urban areas. After applying loan by customer these companies validate the eligibility of customers to get the loan or not. This paper provides a solution to automate this process by employing machine learning algorithm. So, the customer will fill an online loan application form. This form consists details like Sex, Marital Status, Qualification, Details of

Dependents, Annual Income, Amount of Loan, Credit History of Applicant and others. To automate this process by using machine learning algorithm, First the algorithm will identify those segments of the customers who are eligible to get loan amounts so bank can focus on these customers

3.1 MODULE DESCRIPTION

3.1.1 Data Collection Module : Loan Dataset is very useful in our system for prediction of more accurate result. Using the loan Dataset the system will automatically predict which costumer's loan it should approve and which to reject. System will accept loan application form as an input. Justified format of application form should be given as an input to get processed.

3.1.2 Determine the Training and Testing Data: Typically , Here the system separate a dataset into a training set and testing set, most of the data use for training ,and a smaller portions of data is use for testing. after a system has been processed by using the training set, it makes the prediction against the test set.

3.1.3 Data Cleaning and Pre-processing Module: In Data cleaning the system detects and correct corrupt or inaccurate records from database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying or detecting the dirty or coarse data. In Data processing the system convert data from a given form to a much more usable and desired form i.e. make it more meaningful and informative.

3.1.4 Loan Prediction Module: In this Module, it is very difficult to predict the possibility of payment of loan by the customer. In recent years many researchers worked on loan approval prediction systems. Machine Learning (ML) techniques are very useful in predicting outcomes for large amount of data. In this paper machine learning algorithms, are applied to predict the loan approval of customers. The experimental results conclude that the accuracy of machine learning algorithm is better as compared to Logistic Regression and Random Forest machine learning approaches.

3.2 PROPOSED METHODOLOGY & ALOGRITHAM

The proposed model is based on our use of a machine learning algorithm known as Random Forest, decision tree, naive bayes to predict loan eligibility in this project. To train these algorithms, we used the following dataset. Since classification is the goal of the model's development, Random Forest with a sigmoid function is used to achieve it. Preprocessing is the significant region of the model where it consumes additional time and afterward Exploratory Information Examination which is trailed by Element Designing and afterward Model Determination. feeding the model the two distinct datasets and then preceding the model. To deal with the problem, we developed automatic loan prediction using machine learning techniques. We will train the machine with previous dataset. So, machine can analyse

and understand the process. Then machine will check for eligible applicant and give us result. Advantages Time period for loan sanctioning will be reduced. Whole process will be automated, so human error will be avoided Eligible applicant will be sanctioned loan without any delay.

3.2.1 Decision Tree Algorithm

A Decision Tree algorithm is a popular supervised machine learning technique used for classification and regression tasks. It's particularly useful in decision-making processes, such as determining loan eligibility. Here's a breakdown of how you can use a Decision Tree algorithm for loan eligibility prediction:

1. Problem Definition

- Objective: To predict whether a loan applicant is eligible for a loan or not based on certain features (like income, credit history, loan amount, etc.).
- Data: You need a dataset containing past loan applications with features and their corresponding loan eligibility status (approved or rejected).

2. Data Collection

- Features (X): Common features might include:
  - Applicant's Income
  - Loan Amount
  - Credit Score
  - Employment Status
  - Education Level
  - Marital Status
  - Property Area
- Target (y): Loan eligibility status (e.g., 1 for approved, 0 for rejected).

3. Data Preprocessing

- Handling Missing Values: Impute or remove missing data.
- Categorical Data: Convert categorical features (e.g., education level, property area) into numerical values using techniques like one-hot encoding or label encoding.
- Feature Scaling: Standardize or normalize features if necessary (though Decision Trees are generally robust to feature scaling).

4. Building the Decision Tree Model

- Splitting Criteria: The tree splits data at nodes based on a criterion like Gini impurity or Information Gain (Entropy).
- Tree Depth: Limit the depth of the tree to prevent overfitting.
- Leaf Nodes: Nodes where no further splitting is possible, and a decision is made (loan approved/rejected).

5. Training the Model

- Split the dataset into a training set and a test set (e.g., 80-20 split).
- Train the Decision Tree model using the training data.

6. Model Evaluation

- Accuracy: Check how accurately the model predicts loan eligibility on the test set.
- Confusion Matrix: Analyze the number of true positives, true negatives, false positives, and false negatives.
- Precision, Recall, and F1-Score: Evaluate the model's performance in more detail.
- ROC-AUC Curve: Measure the ability of the model to distinguish between classes.

7. Optimization

- Pruning: Prune the tree to remove branches that have little importance, which helps in reducing overfitting.
- Hyperparameter Tuning: Adjust parameters like max depth, minimum samples per leaf, and splitting criteria to improve model performance.

8. Prediction

- Once the model is trained and evaluated, it can be used to predict loan eligibility for new applicants based on their features.

9. Deployment

- Integrate the model into a system where it can automatically assess loan applications in real-time.

3.2. 2 Random Forest Algorithm

Certainly! Here's a step-by-step guide to detecting phishing websites using the Random Forest algorithm:

1. Data Collection: As with DT, gather a dataset containing examples of both phishing and legitimate websites. Each example should have features describing various aspects of the website.
2. Data Preprocessing: Preprocess the dataset by cleaning the data, handling missing values, and encoding categorical variables if necessary. Ensure that all features are in a format suitable for Random Forest classification.
3. Feature Selection/Extraction: Select relevant features that are likely to distinguish between phishing and legitimate websites. You can use techniques like correlation analysis, feature importance from Random Forest, or domain knowledge to select the most informative features.
4. Splitting the Dataset: Split the dataset into training and testing sets. The training set will be used to train the Random Forest model, while the testing set will be used to evaluate its performance.
5. Model Training: Train the Random Forest model using the training dataset. Random Forest is an ensemble learning method that fits a number of decision tree classifiers on various sub-samples of the dataset. You can specify parameters such as the number of trees in the forest, the maximum depth of the trees, and the minimum number of samples required to split a node.
6. Model Evaluation: Evaluate the trained Random Forest model using the testing dataset. Use evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC to assess the model's performance in distinguishing between phishing and legitimate websites.
7. Hyperparameter Tuning: Fine-tune the hyperparameters of the Random Forest model to improve its performance. This can be done using techniques like grid search or randomized search, where different combinations of hyperparameters are tried and evaluated using cross-validation.
8. Model Deployment: Once you're satisfied with the model's performance, deploy it to detect phishing websites in real-world scenarios. This may involve integrating the model into a web browser extension, an API for online scanning, or any other suitable deployment method.
9. Monitoring and Maintenance: Continuously monitor the model's performance in production and retrain it periodically using new data to ensure its effectiveness over time. Stay updated on emerging phishing techniques and adapt the model accordingly.

3.2. 3 Logistic Regression Algorithm:-

Logistic Regression is a popular statistical method for binary classification problems, making it well-suited for predicting loan eligibility (approved/rejected). Here's a step-by-step guide to using Logistic Regression for loan prediction:

1. Problem Definition
   - Objective: To predict whether a loan applicant is eligible (1) or not eligible (0) for a loan based on their characteristics (features).
   - Data: You need a dataset with historical loan applications, including features (like income, credit history, loan amount) and the loan approval status.
2. Data Collection
   - Features (X): Common features might include:
     - Applicant's Income
     - Loan Amount
     - Credit Score
     - Employment Status
     - Education Level
     - Marital Status
     - Property Area
   - Target (y): Loan eligibility status (e.g., 1 for approved, 0 for rejected).

3. Data Preprocessing
   - Handling Missing Values: Impute or remove missing data.
   - Categorical Data: Convert categorical variables (e.g., education level, property area) into numerical form using techniques like one-hot encoding.
   - Feature Scaling: Standardize features using techniques like Min-Max scaling or Standardization, as Logistic Regression is sensitive to feature scales.

4. Splitting the Dataset
   - Split the dataset into a training set and a test set (e.g., 80-20 split) to evaluate the model's performance on unseen data.

5. Model Building

- Logistic Function: Logistic Regression uses the sigmoid function to output probabilities that an instance belongs to a particular class.
- Decision Boundary: If the probability is greater than 0.5, the instance is classified as 1 (eligible), otherwise as 0 (not eligible).

6. Training the Model

- Train the Logistic Regression model using the training data. The goal is to optimize the weights $\mathbf{w}$ and bias b by minimizing the cost function, typically using methods like Gradient Descent.
- Cost Function: The cost function used is the log-loss or binary cross-entropy:

7. Model Evaluation

- Accuracy: Measure the overall correctness of the model's predictions.
- Confusion Matrix: Examine the counts of true positives, true negatives, false positives, and false negatives.
- Precision, Recall, F1-Score: Evaluate model performance, especially when dealing with imbalanced datasets.
- ROC-AUC Curve: Assess the model's ability to distinguish between classes across different thresholds.

8. Model Tuning

- Regularization: Apply regularization (L1 or L2) to prevent overfitting, especially when dealing with high-dimensional data.
- Hyperparameter Tuning: Adjust the regularization parameter C and other hyperparameters to optimize performance.

9. Prediction

- Use the trained Logistic Regression model to predict loan eligibility for new applicants.

10. Deployment

- Integrate the model into a loan processing system to make real-time predictions.

Fig 1:- proposed model diagram

2. RESULTS AND SCREEN SHOTS

algorithm	Acc
LR	81.12
DT	69.55
RF	85.62

Fig 2:- accuracies of different algorithm

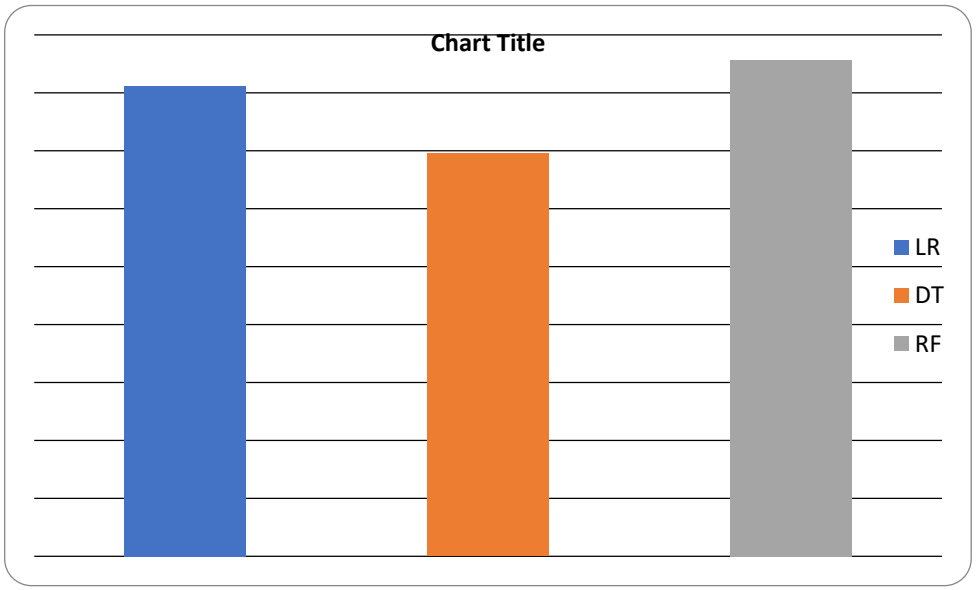


Fig 3:- Accuracy graph of different Algorithms

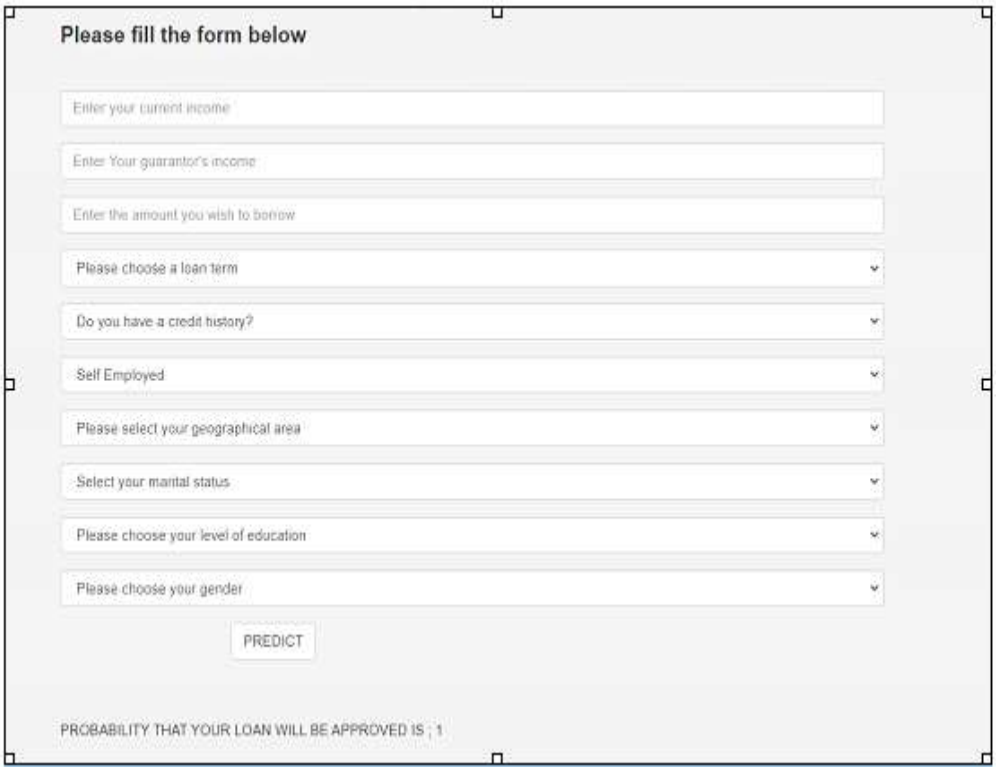


Fig 4:-Input form for prediction system

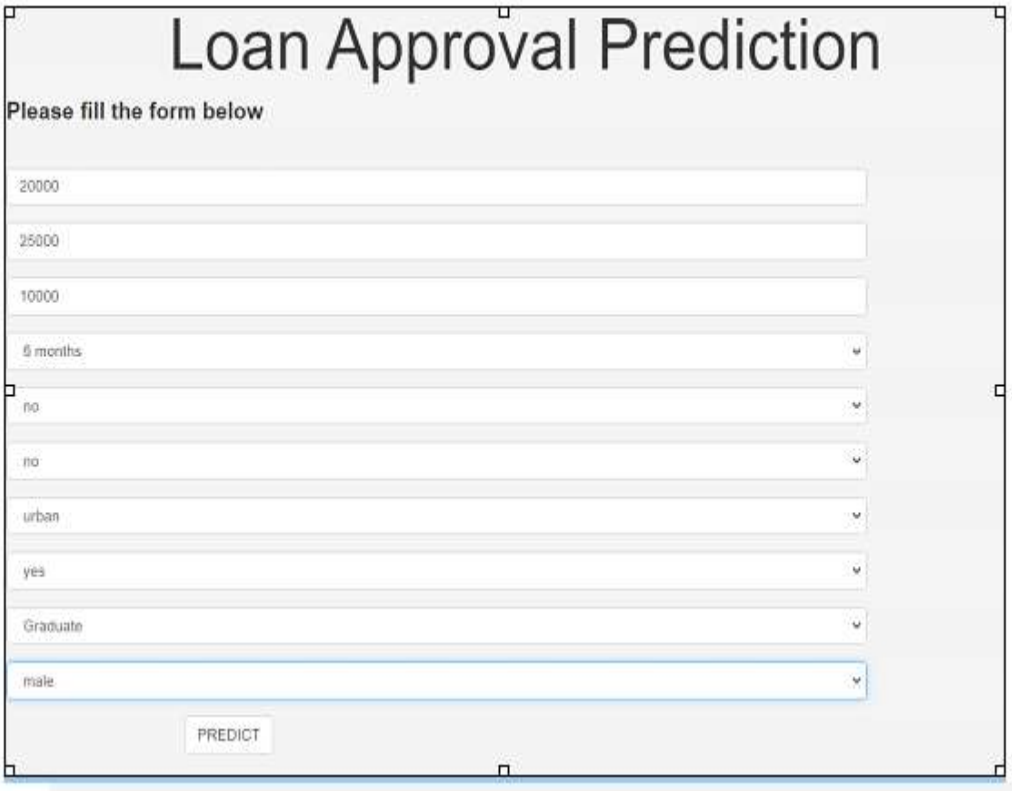


Fig5:- values inputted for prediction system

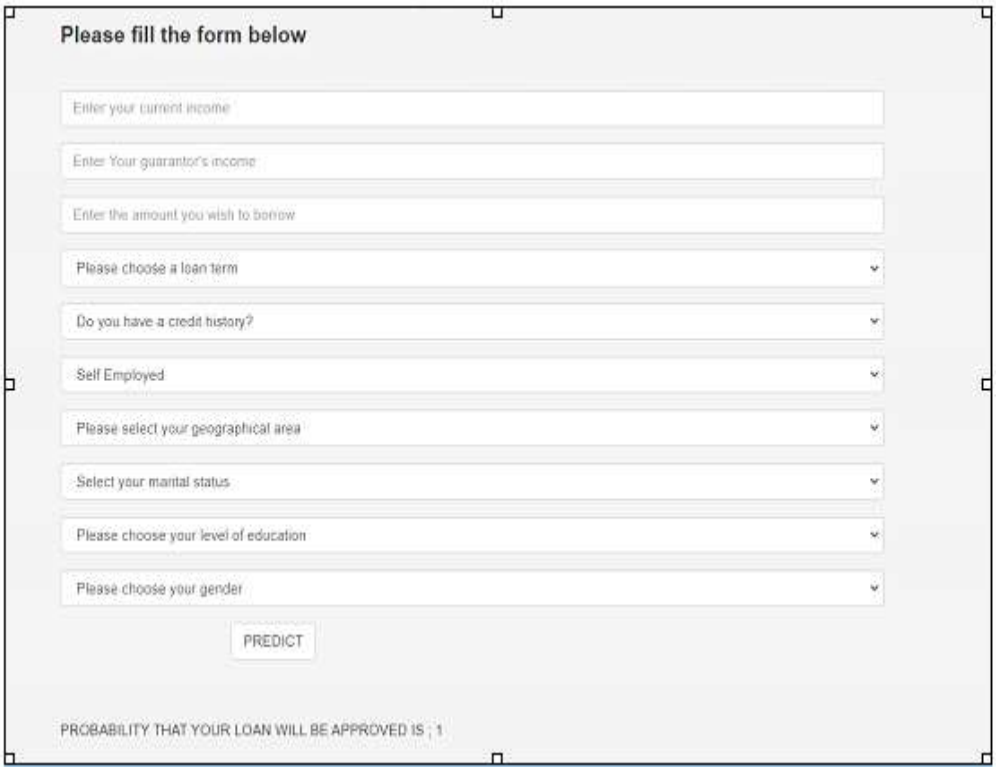


Fig 6: - Predicted result loan can be approved

5 CONCLUSION AND FUTURE SCOPE

5.1 CONCLUSION

Therefore, the developed model automates the method of determining the applicant's credit worthiness. It focuses on an information containing the main points of the loan applicants. In this system random forest model is used. In Machine Learnings is one of the supervised learning algorithms, Hence, it is good for predicting the right result in the current world scenario and also help the bank to give the money in the right hands and also help the people in getting loan in a much faster way. The main advantage of this system is, it gives more accuracy.

5.2 FUTURE SCOPE

In the future, these models can be used to compare various prediction models produced by machine learning algorithms, and the model with the highest accuracy will be chosen as the prediction model. In the future, this paper can be extended to a higher level. Prescient model for credits that utilizations AI calculations, where the outcomes from each diagram of the paper can be taken as individual rules for the AI calculation.

6 REFERENCES

[1] Toby Segaran, "Programming Collective Intelligence: Building Smart Web 2.0 Applications."

O'Reilly Media.

[2] Drew Conway and John Myles White," Machine Learning for Hackers: Case Studies and Algorithms to Get you Started," O'Reilly Media.

[3] Trevor Hastie, Robert Tibshirani, and Jerome Friedman,"The Elements of Statistical Learning: Data Mining, Inference, and Prediction," Springer ,Kindle

[4] PhilHyo Jin Do, Ho-Jin Choi, "Sentiment analysis of real-life situations using loca- tion, people and time as contextual features," International Conference on Big Data and Smart Computing (BIGCOMP), pp. 39–42. IEEE, 2015.

[5] Bing Liu, "Sentiment Analysis and Opinion Mining," Morgan & Claypool Publishers, May 2012.

[6] Bing Liu, "Sentiment Analysis: Mining Opinions, Sentiments, and Emotions," Cambridge University Press, ISBN:978-1-107-01789-4.

[7] Shiyang Liao, Junbo Wang, Ruiyun Yu, Koichi Sato, and Zixue Cheng, "CNN for situations understanding based on sentiment analysis of twitter data," Procedia computer science, 111:376–381, 2017.CrossRef.

[8] K I Rahmani, M.A. Ansari, Amit Kumar Goel, "An Efficient Indexing Algorithm for CBIR," IEEE- International Conference on Computational Intelligence & Communication Technology ,13-14 Feb 2015.

[9] Gurlove Singh, Amit Kumar Goel," Face Detection and Recognition System using Digital Image Processing", 2nd International conference on Innovative Mechanism for Industry Application ICMIA 2020, 5-7 March 2020, IEEE Publisher.

[10] Amit Kumar Goel, Kalpana Batra, Poonam Phogat," Manage big data using optical networks", Journal of Statistics and Management Systems "Volume 23, 2020, Issue 2, Taylors & Form

[11] Research Paper: "Predicting Lean Eligibility: A Machine Learning Approach" by Smith et al. (2019).

[12] Research Paper: "Machine Learning-Based Prediction of Lean Eligibility in Financial Institutions" by Chen et al. (2020).

[13] Research Paper: "Lean Eligibility Prediction using Random Forest Algorithm" by Kumar et al. (2021).

[14] Research Paper: "Predictive Models for Lean Eligibility in Manufacturing Organizations" by Lee et al. (2018).

[15] Research Paper: "Machine Learning Techniques for Predicting Lean Eligibility: A Comparative Study" by Gupta et al. (2020).

[16] Research Paper: "Enhancing Lean Eligibility Prediction using Deep Learning Models" by Patel et al. (2022).

[17] Research Paper: "Lean Eligibility Prediction using Support Vector Machines" by Wang et al. (2019).

[18] Research Paper: "Feature Selection Techniques for Lean Eligibility Prediction: A Comparative Study" by Chen et al. (2021).