

## Loan Prediction Using Random Forest Algorithm

Sukkala Tharun Kumar Goud<sup>1</sup>, Mohammed Shoaib<sup>2,\*</sup>, Satya Bharadwaj<sup>3</sup>, Y. Sreenivasulu<sup>4</sup>

### Abstract

*With the growth of the banking sector, more and more people are applying for bank loans. All of these loans are not allowed. The principal income of bank assets arises from the interest earned on the loan. Bank profit or loss depends largely on the amount of the loan, i.e., whether customers pay off the loan or fail. The main purpose of banks is to invest their assets in secure customers. Today, many banks approve loans after a number of verifications and verification processes but yet there is no guarantee that the selected customer is safe or not. By predicting defaulters, the bank can reduce its non-performing Assets. This makes a study of this situation very important. Previous research in this period has shown that there are many ways to learn the problem of controlling loan default. But as appropriate predictions are very important for increase in profit, it is important to study the type of different methods and their comparisons. A very important method in predictive analysis is used to study the problem of predicting those who fail to borrow money. It is therefore important to use a variety of strategies in the banking sector to select a customer who pays the loan on time. In this report, we use a random forest algorithm to separate the data. The random forest algorithm creates a model from a trained database and this model is used in test data and we obtain the required output.*

**Keywords:** Database, Random Forest algorithm, Loan Prediction System, Banking Sector, Secure Customers

### INTRODUCTION

The two most pressing issues in the banking sector are: 1) How dangerous is a borrower? 2) Should we lend to a borrower in the event of an accident? The answer to the first question determines the interest rate of the borrower. The interest rate, among other things (such as the amount of time period), assesses the borrower's risk, i.e., the higher the interest rate, the higher the borrower's risk. We will then decide whether the applicant is eligible for a loan based on the interest rate. Lenders (investors) lend to borrowers to obtain interest-bearing payments [1]; that is, the lender makes repayment (interest) only if the borrower repays the loan. However, even if he does not repay the loan, the lender

loses the loan. Banks lend to customers to obtain payment confirmation. Some were not paying their debts. They could not pay them for many reasons. The bank maintains insurance to reduce the likelihood of failure in the event of an error. The insured amount can cover all the loan or only part of it [2]. Banking procedures use manual processes to determine whether a borrower is eligible for a loan based on the results. Manual procedures worked very well but were not sufficient for a large number of loan applications. At that point, the decision would be long overdue. As a result, the loan model learning model can be used to assess a client's credit status and develop strategies [3]. This model identifies and introduces key

**\*Author for Correspondence**  
Mohammed Shoaib

<sup>1-3</sup>Student, Department of Electronics and Communication Engineering, Sreenidhi Institute of Science and Technology, Hyderabad, Telangana, India

<sup>4</sup>Associate Professor, Department of Electronics and Communication Engineering, Sreenidhi Institute of Science and Technology, Hyderabad, Telangana, India

Received Date: March 14, 2022

Accepted Date: April 20, 2022

Published Date: April 30, 2022

**Citation:** Sukkala Tharun Kumar Goud, Mohammed Shoaib, Satya Bharadwaj, Sreenivasulu Y. Loan Prediction Using Random Forest Algorithm. Journal of Advanced Database Management & Systems. 2022; 9(1): 33–39p.

features of the borrower that influence the credit status of the customer. Finally, it produces scheduled performance (loan status). These reports make the job of a bank manager easier and faster. Loan lending is a major part of the business of many banks. The bulk of the bank's revenue comes from loans that are distributed to customers. These banks use interest rates on loans that are distributed to customers [4]. The main purpose of banks is to invest their assets in secure customers. To date, many banks are still processing debt after the verification and retrenchment process. But so far, no bank can guarantee that a customer who has been selected to apply for a loan is safe or not. To avoid this situation, we have introduced the Loan Prediction System Using Python. Loan Prediction System is software that assesses the suitability of a particular customer who is able to repay a loan or not. This program tests various parameters such as customer credit status, revenue, costs and various features. This process applies to most clients of a set of qualified data [5]. By considering these factors, the required model was developed. This model is used in test data to obtain the required output. Produced output will be by authorized or unauthorized means [6].

## **SYSTEM ANALYSIS**

### **Introduction**

Computer finance, World Wide Web communications and automated software support have completely changed the basic business concept and business model. With the advent of computers and network technology, the power in the fields of information production and data collection has improved dramatically. Data mining plays an important role in computer science to solve many data-related problems. The data mining strategy involves the use of sophisticated data analysis tools to detect previously unknown patterns, valid patterns, and relationships in a large data set [7].

### **Purpose of the Project**

Loan prediction is very useful for bank employees and the applicant as well. The purpose of this study is to provide a quick, quick and easy way to select eligible applicants. Loan Prediction System allows you to jump to a specific application for critical testing [8].

### **Existing System**

To date, loans have been processed by various banks using pen and paper. When a large number of clients apply for a bank loan, these banks take longer to approve the loan. After bank approval has been approved, there is no guarantee that the selected applicant will be able to repay the loan or not. Many banks use their software to obtain a loan. In the existing system we use data mining algorithms to approve loans; this is an old form of loan approval. Many data sets are combined to form standard data sets, and different machine learning algorithms are used to produce results. But these processes are not up to standard [9]. As a result, the central banks are struggling financially. To resolve this issue, we are introducing a new approach to loan approval [10–12].

### **Proposed System**

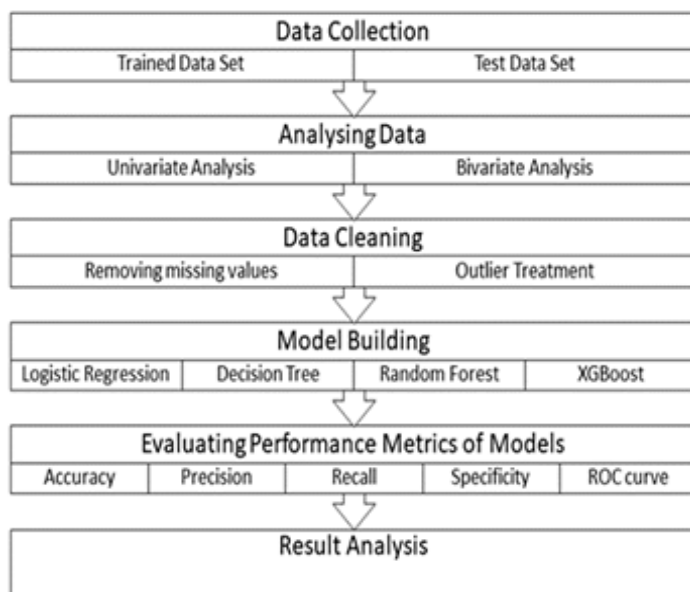
Loan approval system is a software used to authorize loans in the banking sector. In the proposed program, we used a machine learning algorithm. Machine Learning is the process by which a symmetric model is constructed from an existing database; this model is used for new database tests. The program contains a qualified database and a test database. A trained database is used to model. This model is used for database testing to obtain the required result. We used the ensemble method to model. The random forest algorithm uses this integration method and creates a model from existing training data.

## **SYSTEM DESIGN**

### **Introduction**

The loan loyalty program analyses the applicant's information and distinguishes it between the applicant with secure credit and the applicant with risky credit. An applicant with a secure credit has a higher chance of repaying the loan immediately while an applicant with risky credit has a higher

chance of failing to repay the loan amount. Profitability of financial institutions definitely depends on the accuracy of the model. Using a random forest division algorithm, it is very effective to build an effective forecasting model that helps banks make informed decisions about safe and risky loan applications. The random forest section produces high accuracy. The machine learning model is trained using a training data set. All new applicants' information completed during the application form serves as a set of test data. Based on the training data sets, the model will predict whether the loan will be approved or not. We have 13 total features with 12 independent variables and 1 dependent variant namely loan status on the trained database and 12 independent variables on the test database. Debt ID, Gender, Spouse, Dependents, Education, Self-Employment, Location Local, and Loan borrow are all categories.



**Figure 1.** Data flow chart.

### Data Flow Diagrams

A data flow diagram is a graphical tool used to describe and evaluate system data flow (Figure 1). These are the middle tool and the base on which some parts are made. The conversion of data from input to output, by processing, logically and independently, describes system-related body parts. These are known as logical data flow diagrams (Figure 2).

### UML Diagrams

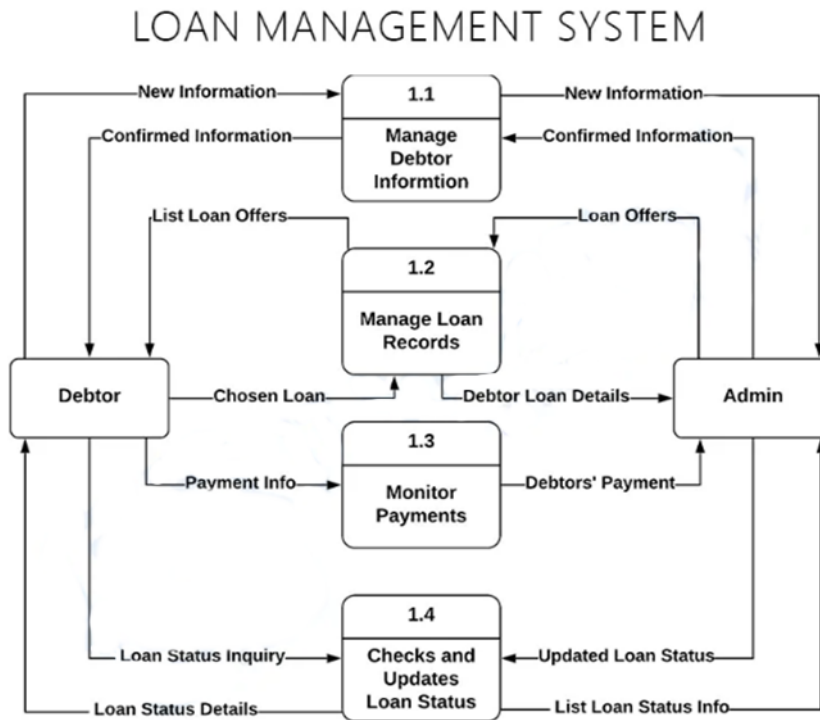
A UML diagram is a diagram based on the UML (Unified Modelling Language) that is used to visually represent a system along with its main actors, roles, actions, artefacts or classes in order to better understand, alter, maintain, or document the system information (Figures 3 and 4).

## IMPLEMENTATION

### Ensemble Learning

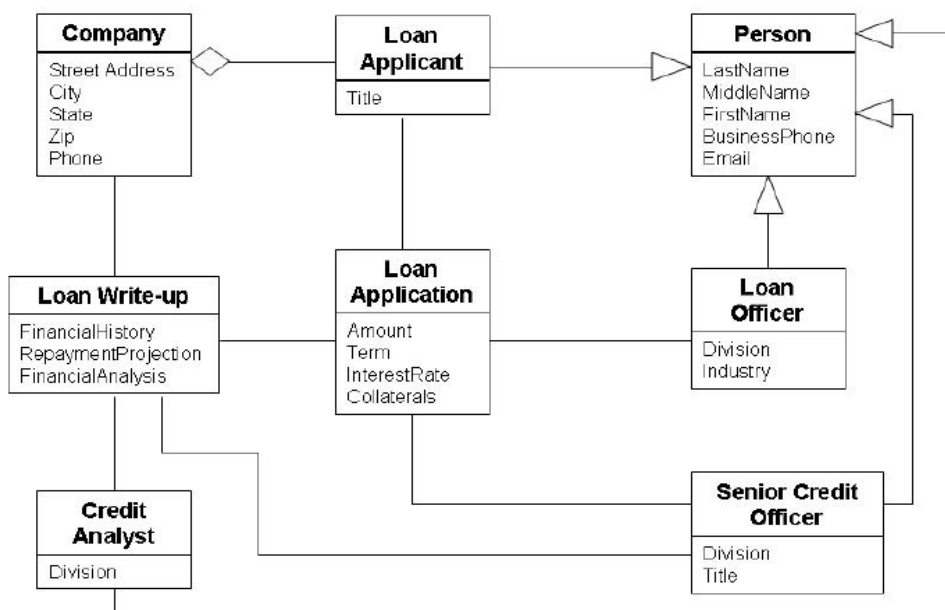
We used integrated learning to create a program. Shared reading is a method that contains many numbers of weak dividers. We upload our training database to these dividers to get the result. These dividers may use different algorithms such as one reader categorizing using an additional tree algorithm, another separator may use a decision algorithm, another editor may use a random forest algorithm etc. When we submit our training database to these dividers, they produce output according to their algorithm function. This method is known as Heterogeneous classifying or Ensemble learning. We can also take all student filters as the same algorithm. But when we upload the same training database, these dividers produce the same output. To avoid this problem, we have provided a separate

training database for these dividers, so that each category provides a different output. While providing the output, these class dividers by considering the various factors make up the model. Each separator can build a model according to the training database provided. In these different dividers, we combine all these dividers to create a classification model that satisfies all these dividers. This category is considered to be the Solid Divorce. This is called the strong classifier because it produces more accurate output and smaller errors. Random Forest also uses the Ensemble Learning Technique [10].

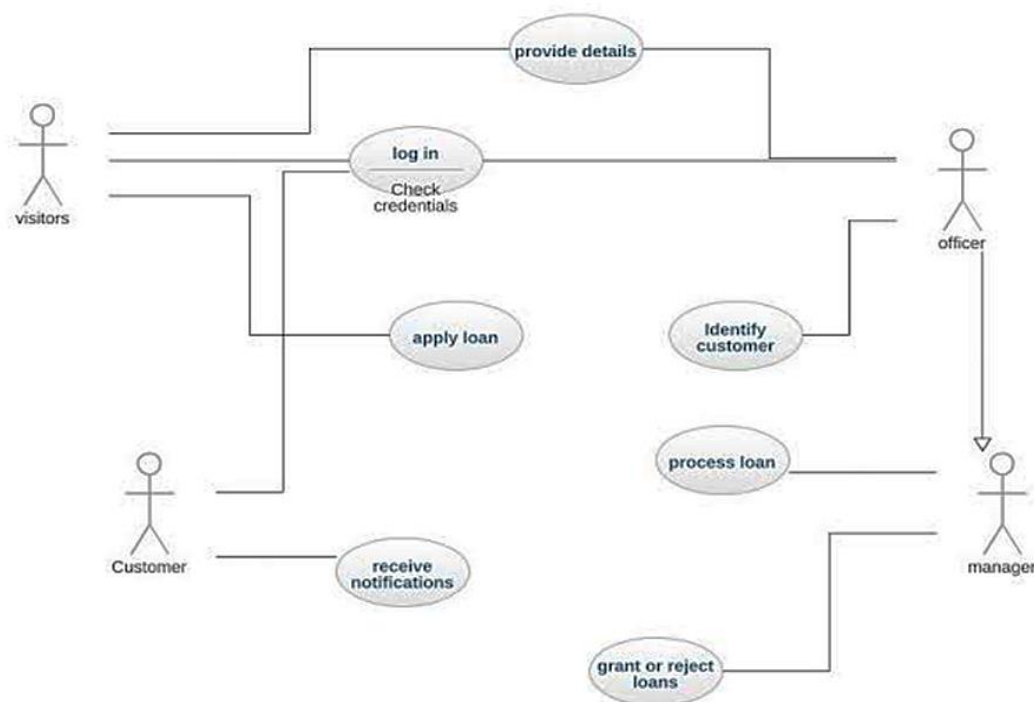


## DATA FLOW DIAGRAM LEVEL 1

**Figure 2.** Data flow diagram.



**Figure 3.** Class diagram.



**Figure 4.** Use case diagram.

### Random Forest Algorithm

As we discussed earlier, the Random Forest algorithm uses the Ensemble Learning Technique.

#### Functional Random Forest Algorithm

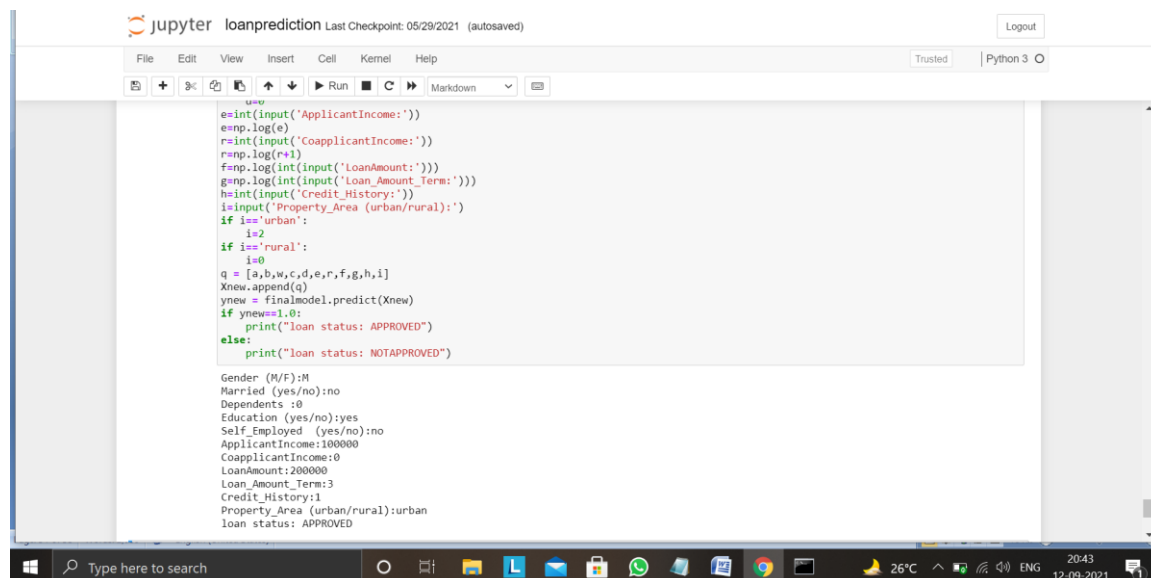
The Random Forest algorithm follows the rules of the decision tree. The difference in them is that the decision tree algorithm provides output by considering only one aspect whereas the Random Forest Algorithm compares multiple decision trees and provides a satisfactory result for most decision trees. Random Forest Algorithm builds a robust model that satisfies many of the decision tree models, this model is used in experimental databases to obtain the required output. First, in order to build a model, we need a training database; in that training database, we are considering a training database set known as bootstrap dataset-1. This database-1 of bootstrap contains a set of variables. In these variables, we look at only two variables; from these two variables, making the root node of one variable producing more accurate output than the other variable. In this format, we create a resolution tree from the bootstrap-1 database. We are processing another database collection from the training database, say bootstrap dataset-2. Just as we build the bootstrap-1 database decision tree, in the same way we create the bootstrap-2 database. We have to follow these steps until we find a lot of decision trees. After finding many cut trees, we compare the whole decision tree and build a model that satisfies all these cut trees. The model found is known as the Strong Model. In this way, the model is built from the training database [11].

### OUTPUT SCREEN

Below here we have shown in Figure 5 the output screen of loan prediction on Jupyter.

### CONCLUSION

From a practical point of view, the system can be put to good use in determining who is eligible for a loan. The software works well and can be used for all banking needs. This program can be easily downloaded from any operating system. As technology goes online, this system has a wide range of future days. This program is very secure and reliable.



**Figure 5.** Output Screen of the loan prediction.

With the use of the Random Forest Algorithm, the system returns very accurate results. There is no problem if there are not many clients applying for a loan. This system receives N and no data of customers. Next time we can add more algorithms to this system to get the most accurate results. Now tests have proven that a model that uses a pre-segment selection feature can help bank officials make the right decision more accurately. This proposed approach will protect the bank from further misuse, fraudulent applications etc. by identifying customers with their critical payment capabilities, especially in the cooperative banking sector. Tests proved that the accuracy of the categories increased significantly after the selection of the feature. The proposed algorithm has produced better accuracy than the available methods.

## Future Scope

In the future we wish to improve the data mining application using the Random Forest selection algorithm and it certainly helps bank officials to make the right decision when a new customer goes to the bank to take out a loan. The proposed algorithm for selecting a mixed feature to differentiate customer loyalty behaviour in the banking industry may be used for a few other applications in the future, especially binary separation problems such as various disease forecasting, various test results prediction etc.

## REFERENCES

1. Liaw A, Wiener M. Classification and regression by Random Forest. R news. 2002 Dec 3; 2(3): 18–22.
2. Gandotra E, Bansal D, Sofat S. Malware analysis and classification: A survey. J Inf Secur. 2014; 5(2): 56–64.
3. Keerthi SS, Gilbert EG. Convergence of a generalized SMO algorithm for SVM classifier design. Mach Learn. 2002 Jan; 46(1): 351–60.
4. Rao KH, Srinivas G, Damodhar A, Krishna MV. Implementation of anomaly detection technique using machine learning algorithms. International journal of computer science and telecommunications (IJCST). 2011 Jun; 2(3): 25–31.
5. Jency XF, Sumathi VP, Sri JS. An exploratory data analysis for loan prediction based on nature of the clients. Int J Recent Technol Eng (IJRTE). 2018 Nov; 7(4): 176–179.
6. Saha A, Denning T, Srikumar V, Kasera SK. Secrets in Source Code: Reducing False Positives using Machine Learning. In IEEE 2020 International Conference on COMMunication Systems & NETworkS (COMSNETS). 2020 Jan 7; 168–175.

- 
7. Chambers JM. Computational methods for data analysis. New York: A Wiley Publication in Applied Statistics; 1977.
  8. Quinlan JR. Induction of decision trees. *Machine learning*. 1986 Mar; 1(1): 81–106.
  9. Cowell RG, Dawid AP, Lauritzen SL, Spiegelhalter DJ. *Graphical Models and Expert Systems*. Berlin: Springer; 1999.
  10. Sreedevi E, PremaLatha V, Prasanth Y, Sivakumar S. A Novel Ensemble Learning for Defect Detection Method with Uncertain Data. In *Applications of Artificial Intelligence for Smart Technology*, IGI Global. 2021; 67–79.
  11. Togaware. Rattle data mining tool. [Online]. Available from <http://rattle.togaware.com/rattle-download.html>
  12. Random Forest. Mean Decrease Accuracy. [Online]. Available from <https://dinsdalelab.sdsu.edu/metag.stats/code/randomforest.html>