Loan Approval Prediction System Using Machine Learning

P. L. Srinivasa Murthy\textsuperscript{1}, G. Soma Shekar\textsuperscript{2}, P. Rohith\textsuperscript{3}, G. Vishnu Vardhan Reddy\textsuperscript{4}

\textsuperscript{1}Associate Professor and Head, Department of Computer Science and Engineering, Institute of Aeronautical Engineering, Hyderabad, India.
\textsuperscript{2, 3, 4}Department of Computer Science and Engineering, Institute of Aeronautical Engineering, Hyderabad, India.
Email: \textsuperscript{1}pl.srinivasamurthy@iare.ac.in,\textsuperscript{2}somashekgandhe@gmail.com,\textsuperscript{3}pippirirohith@gmail.com,\textsuperscript{4}gudipallyvishnuvardhanreddy@gmail.com

Abstract - With the increase in banking sector many people are applying for loans in bank. All these loans are not approvable. The main income of bank assets comes from gain earned from loans. The main objective of banks is to invest their assets in safe customers. Today many banks approve loan after many process of verification and validation but still there is no surety that selected customer is safe or not. Therefore it is important to apply various techniques in banking sector for selecting a customer who pays loan on time. In this report we use random forest algorithm for the classification of data. Random forests algorithm builds a model from trained dataset and this model is applied on test data and we get the required output.

Keywords - Trained Dataset, Random Forests, Bank Loans, Safe Customers.

1. INTRODUCTION

Loan Distribution is the main business part of many banks. The main portion of banks income comes from the loan distributed to customers. These banks apply interest on loan which are distributed to customers.

The main objective of banks is to invest their assets in safe customers. Up to now many banks are processing loans after regress process of verification and validation. But till now no bank can give surety that the customer who is chosen for loan application is safe or not. So to avoid this situation we introduced a system for the approval of bank loans known as Loan Prediction System Using Python.

Loan Prediction System is a software which checks the eligibility of a particular customer who is capable of paying loan or not. This system checks various parameters such as customer’s marital status, income, expenditure and various factors. This process is applied for many customers of trained data set. By considering these factors a required model is built. This model is applied on the test data set for getting required output. The output generated will be in the form of yes or no. Yes indicates that a particular customer is capable of paying loan and no indicates that the particular customer is not capable of paying loan. Based on these factors we can approve loans for customers.

2. LITERATURE REVIEW

Data Analysis for prediction of loan based nature of clients

The report main intention is to classify the nature of clients for loans. Depending upon the certain factors the report classifies the customers. Classification is done through exploratory data analyses [1].

Exploratory data analysis is a technique that analyzes and summaries the main features from training dataset.

Prediction of Loan Approval using Machine Learning Approach

Machine learning [2] is a phenomenon in which analytical model is build from the trained model.

This model is applied on test data for providing of the accurate results.

Here the author used three algorithms for prediction of loan. They are

1. K Nearest Neighbor
2. Decision Tree
3. Random Forests

The main purpose of this report is to provide immediate and accurate results for the approval of loan to the eligible customers. In banking sector there will be n number of people who apply loans. It is difficult to check customer’s eligibility through paper work. The system can provide accurate results for the n number of people.

Building the model using Random Forest Approach

In this report we have discussed about credit risk and credit analysis. Banking sectors success mainly depends of credit risk analysis. In this report we have used Random Forest [3] approach to build the model. The use of Random Forest is because Random Forest Approach provides accurate results than the K Nearest Neighbor and Decision Tree.

Ensemble model survey for loan prediction

In this report author has used Random Forest approach for building a model. In this report two or more classifiers are combined together and identify a perfect model for loan prediction.

Ensemble method compares two or more models and identifies a perfect model from two or more models for better loan prediction which makes banking sector to make a right choice for approval of loan application.
3. RELATED WORK

3.1. Existing System

Till now loans are processed by various banks through pen and paperwork. When the large no of customers’ apply for bank loan these bank take lot of time to approve their loan. After approval of loan by the banks, there is no surety that the chosen applicant is capable of paying loan or not. Many banks use their own software’s for the loan approval. In existing system we use data mining algorithms for the loan approval; this is the old technique for the approval of loan. Multiple data sets are combined and form a Generalised datasets, and different machine learning algorithms are applied to generate results. But these techniques are not up to the mark. Due to this huge banks are suffering from financial crises. To resolve this issue we introduce a new way for approval of loans.

3.2. Proposed System

Loan Approval System is software used for approval of loan in banking sector. In this proposed system we have used machine learning algorithm. Machine Learning is process in which a symmetric model is build from the existing dataset; this model is applied for the testing of the new dataset. The system consists of trained dataset and test dataset. The trained dataset is used for construction of model. This model is applied on testing dataset for the required result. We have used Ensemble approach for building of the model.

Random forest algorithm uses this ensemble approach and builds a model from the existing training dataset.

4. IMPLEMENTATION

We have used ensemble learning for building of the system.

4.1. Ensemble learning

Ensemble learning [4] is a method which consists of many no. of weak classifiers. We upload our training dataset in these classifiers to obtain result. These classifiers may use different algorithms such as one learner classifier uses svm algorithm, another classifier may use decision algorithm another classifier may use k nearest neighbour algorithm etc.

When we submit our training dataset to these classifiers produce output according to their algorithm function. This method is known as Heterogenous classifying or Ensemble learning. We can also take all learner classifier as same algorithm. But when we upload same training dataset these all classifiers produce same output. To avoid this problem we have give different training dataset to these classifiers, So that each classifier gives different outputs. While providing the output these classifiers considering various factors build a model. Each classifier may build a model according to given training dataset. From these different classifiers we combine all these classifiers a build a new classifier model which satisfies all these classifiers. This classifier is considered as Strong Classifier. This is called Strong classifier because it produces accurate output and less error. Random Forest also uses this Ensemble Learning Technique.

4.2. Random Forest Algorithm

As we discussed earlier Random Forest Algorithm uses Ensemble Learning Technique.

Working of Random Forest Algorithm

Random Forest Algorithm is follows rules of Decision Tree. The difference in them is decision tree algorithm [7] gives the output by considering only one factor where as Random Forest Algorithm compares many no of decision trees and gives the result satisfying majority no of decision trees.
Random Forest Algorithm builds a strong model which satisfies the models many no of decision tree, this model is applied on the testing dataset for getting the required output. Firstly for building a model we require a training dataset, from that training dataset we consider the subset of training dataset known as bootstrap dataset1. This bootstrap [6] dataset1 consists of set of variables. From these variables we consider only two variables from these two variables we make a root node for one variable which produces accurate output than the other variable. In this format we build a decision tree from the bootstrap dataset1. We consider another subset of dataset from the training dataset, let us say bootstrap dataset2. As we build a decision tree for the bootstrap dataset 1one in the same way we build a dataset for the bootstrap dataset2. We have to follow these steps until we get many no of decision tree.

After getting many no of decision trees we compare the entire decision tree and build a model which satisfies these entire decision trees. The obtained model is known as Strong Model. In this way a model is built from the training dataset.

This model is applied on the testing data set and it produces a required output. The obtained output is accurate because a strong model is build from many no. of decision trees

In this format, when we upload a training dataset then the system builds a model using Random forests Algorithm then when upload the testing dataset by using the model system provides the required output. The output consists of two class labels yes/no. Yes indicates that the client is eligible for approval of loan and no indicates that the client is not eligible for the approval of loan. In this fashion the system provides the required output.

5. OUTPUT

To protect system from unauthorized access, we have created admin login module for security purpose. It consists of username and password. By providing the valid username and password we can access the system.

The admin home consists of two options:
- Loan Approval Prediction
- Classifier performance

Loan Approval Prediction is chosen for prediction of loan status whereas classifier performance gives the prediction results of three algorithms:
- K Nearest Neighbour
- Decision Tree
- Random Forest
The above Fig. 6 represents the trained data. The train data consists of various attributes such as salary, martial status, loan accounts, and loan repayments on time etc. According to these factors we build a required model for loan prediction.

The above Fig. 8 represents the loan status. The loan status is obtained after the deployment of test data to the model which is build from the trained data using Random Forest Algorithm. The loan status consists of Customer id and loan status. It indicates for a particular customer loan is approved or not. If loan status is Y (Yes) then the customer is eligible for approval of loan and if it is N (No) then the customer is not eligible for approval of loan.

6. CONCLUSION

From the proper view of analysis this system can be used perfect for detection of clients who are eligible for approval of loan. The software is working perfect and can be used for all banking requirements. This system can be easily uploaded in any operating system. Since the technology is moving towards online, this system has more scope for the upcoming days. This system is more secure and reliable. Since we have used Random Forest Algorithm the system returns very accurate results. There is no issue if there are many no of customers applying for loan. This system accepts data for N no. of customers. In future we can add more algorithms to this system for getting more accurate results.

REFERENCES


