Loan Approval Prediction System Using Machine Learning

P. L. Srinivasa Murthy[1], G. Soma Shekar[2], P. Rohith[3], G. Vishnu Vardhan Reddy[4]

[1]Associate Professor and Head, Department of Computer Science and Engineering, Institute of Aeronautical Engineering, Hyderabad, India.
[2, 3, 4] Department of Computer Science and Engineering, Institute of Aeronautical Engineering, Hyderabad, India.
Email: [1]pl.srinivasamurthy@iare.ac.in,[2]somashekargandhe@gmail.com, [3]pippirirohith@gmail.com, [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 martial 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. Mutiple 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**.**

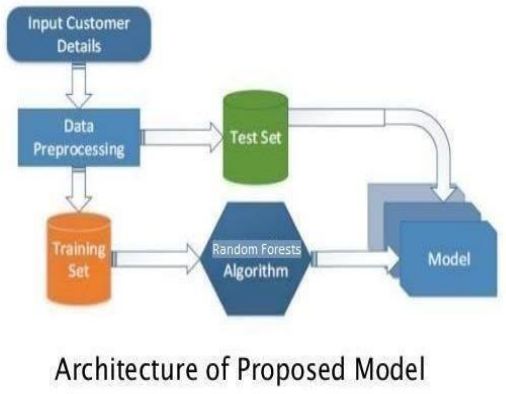


Fig. 1. Block Diagram of Loan Approval Prediction System

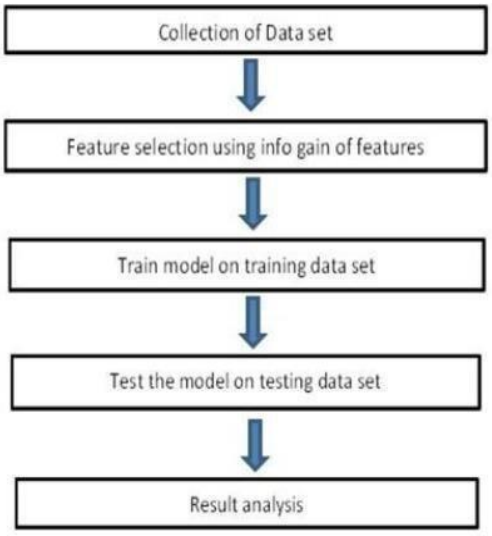


Fig.2. Flowchart Diagram of Loan Approval Prediction System

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.

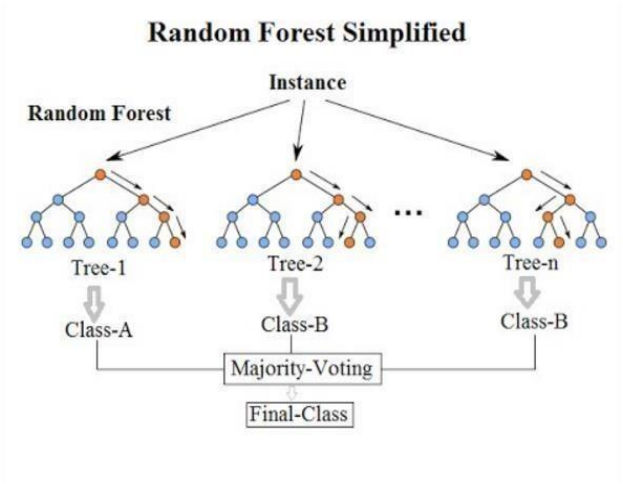


Fig. 3. Random Forest

It builds the number of decision tree models from the trained dataset and all these models are combined and form a new model [5] which satisfies all the tree models. The obtained model is said to be known as strong model since it satisfies all the decision tree models.

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.

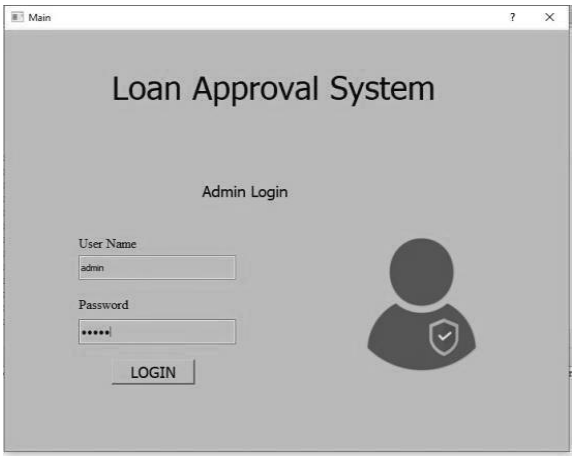


Fig. 4. Admin Login

The admin home consists of two options:

- Loan Approval Prediction
- Classifier performance

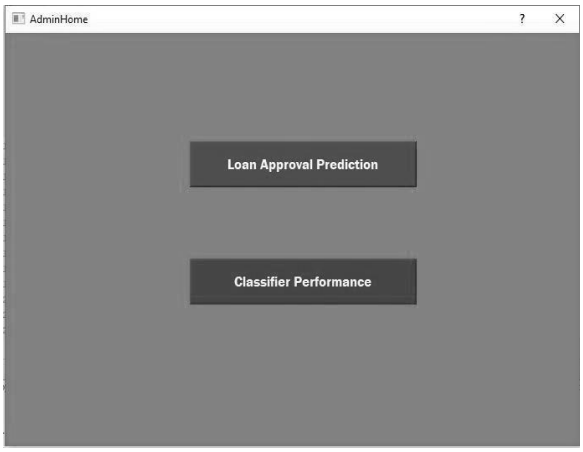
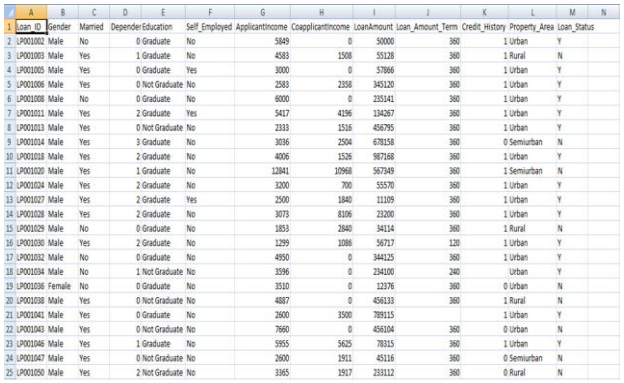


Fig. 5. Admin Home

Loan Approval Prediction is chosen for prediction of loan status whereas classifier performance gives the prediction results of three algorithms:

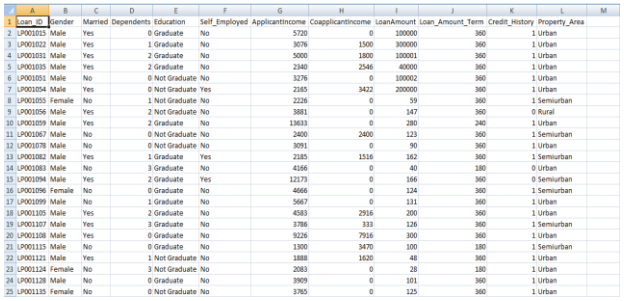
- K Nearest Neighbour
- Decision Tree
- Random Forest



Loan_ID	Gender	Married	Dependents	Education	Self_Employed	Applicantincome	Coapplicantincome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
LP001002	Male	No	0	Graduate	No	5849	0	50000	360	1	Urban	Y
LP001003	Male	Yes	1	Graduate	No	4583	1508	55128	360	1	Rural	N
LP001005	Male	Yes	0	Graduate	Yes	3000	0	57866	360	1	Urban	Y
LP001006	Male	Yes	0	Not Graduate	No	2583	2358	345120	360	1	Urban	Y
LP001008	Male	No	0	Graduate	No	6000	0	235141	360	1	Urban	Y
LP001011	Male	Yes	2	Graduate	Yes	5417	4196	134267	360	1	Urban	Y
LP001013	Male	Yes	0	Not Graduate	No	2333	1516	456795	360	1	Urban	Y
LP001014	Male	Yes	3	Graduate	No	3036	2504	678158	360	0	Semiurban	N
LP001018	Male	Yes	2	Graduate	No	4006	1526	987168	360	1	Urban	Y
LP001020	Male	Yes	1	Graduate	No	12841	10968	567349	360	1	Semiurban	N
LP001024	Male	Yes	2	Graduate	No	3200	700	55570	360	1	Urban	Y
LP001027	Male	Yes	2	Graduate	Yes	2500	1840	11109	360	1	Urban	Y
LP001028	Male	Yes	2	Graduate	No	3073	8106	23200	360	1	Urban	Y
LP001029	Male	No	0	Graduate	No	1853	2840	34114	360	1	Rural	N
LP001030	Male	Yes	2	Graduate	No	1299	1086	56717	120	1	Urban	Y
LP001032	Male	No	0	Graduate	No	4950	0	344125	360	1	Urban	Y
LP001034	Male	No	1	Not Graduate	No	3596	0	234100	240	1	Urban	Y
LP001036	Female	No	0	Graduate	No	3510	0	12376	360	0	Urban	N
LP001038	Male	Yes	0	Not Graduate	No	4887	0	456133	360	1	Rural	N
LP001041	Male	Yes	0	Graduate	No	2600	3500	789115		1	Urban	Y
LP001043	Male	Yes	0	Not Graduate	No	7660	0	456104	360	0	Urban	N
LP001046	Male	Yes	1	Graduate	No	5955	5625	78315	360	1	Urban	Y
LP001047	Male	Yes	0	Not Graduate	No	2600	1911	45116	360	0	Semiurban	N
LP001050	Male	Yes	2	Not Graduate	No	3365	1917	233112	360	0	Rural	N

Fig. 6. Train Data

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.



Loan_ID	Gender	Married	Dependents	Education	Self_Employed	Applicantincome	Coapplicantincome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area
LP001015	Male	Yes	0	Graduate	No	5720	0	100000	360	1	Urban
LP001022	Male	Yes	1	Graduate	No	3076	1500	300000	360	1	Urban
LP001031	Male	Yes	2	Graduate	No	5000	1800	100001	360	1	Urban
LP001035	Male	Yes	2	Graduate	No	2340	2546	40000	360	1	Urban
LP001051	Male	No	0	Not Graduate	No	3276	0	100002	360	1	Urban
LP001054	Male	Yes	0	Not Graduate	Yes	2165	3422	200000	360	1	Urban
LP001055	Female	No	1	Not Graduate	No	2226	0	59	360	1	Semiurban
LP001056	Male	Yes	2	Not Graduate	No	3881	0	147	360	0	Rural
LP001059	Male	Yes	2	Graduate	No	13633	0	280	240	1	Urban
LP001067	Male	No	0	Not Graduate	No	2400	2400	123	360	1	Semiurban
LP001078	Male	No	0	Not Graduate	No	3091	0	90	360	1	Urban
LP001082	Male	Yes	1	Graduate	Yes	2185	1516	162	360	1	Semiurban
LP001083	Male	No	3	Graduate	No	4166	0	40	180	0	Urban
LP001094	Male	Yes	2	Graduate	Yes	12173	0	166	360	0	Semiurban
LP001096	Female	No	0	Graduate	No	4666	0	124	360	1	Semiurban
LP001099	Male	No	1	Graduate	No	5667	0	131	360	1	Urban
LP001105	Male	Yes	2	Graduate	No	4583	2916	200	360	1	Urban
LP001107	Male	Yes	3	Graduate	No	3786	333	126	360	1	Semiurban
LP001108	Male	Yes	0	Graduate	No	9226	7916	300	360	1	Urban
LP001115	Male	No	0	Graduate	No	1300	3470	100	180	1	Semiurban
LP001121	Male	Yes	1	Not Graduate	No	1888	1620	48	360	1	Urban
LP001124	Female	No	3	Not Graduate	No	2083	0	28	180	1	Urban
LP001128	Male	No	0	Graduate	No	3909	0	101	360	1	Urban
LP001135	Female	No	0	Not Graduate	No	3765	0	125	360	1	Urban

Fig. 7. Test Data

The above Fig. 7 represents the test data. The test data consist of various attributes such as salary, martial status, loan accounts except the loan approval status. The loan approval status is obtained. When we deploy the test data to the model which is build from the trained data.

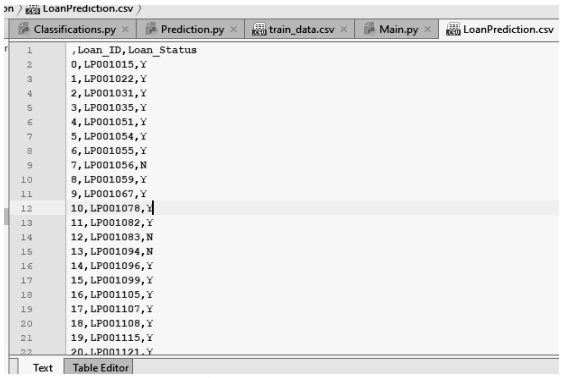


Fig. 8. Loan Status

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

[1] K. Hanumantha Rao., G. Srinivas., A. Damodhar., M.Vikas Krishna., Implementation of Anomaly Detection Technique Using Machine Learning Algorithms: International Journal of Computer Science and Telecommunications, Vol. 2, Issue 3, 2011.

[2] S.S. Keerthi., E.G. Gilbert., Convergence of a generalize SMO algorithm for SVM classifier design, Machine Learning, Springer, Vol. 4, Issue 1, pp. 351-360, 2002.

[3] Andy Liaw., Matthew Wiener., Classification and Regression by random Forest, Vol. 2, Issue 3, pp. 9-22, 2002.

[4] Ekta Gandotra., Divya Bansal., Sanjeev Sofat., Malware Analysis and Classification: A Survey, Journal of Information Security, Vol. 05, Issue 02, pp. 56-64, 2014.

[5] Rattle data mining tool, http://rattle.togaware.com/rattle-download.html.

[6] Aafer Y., Du W., Yin H., Droid APIMiner: Mining API-Level Features for Robust Malware Detection in Android, Security and privacy in Communication Networks, Springer, pp 86-103, 2013.

[7] J. R. Quinlan., Induction of Decision Tree, Machine Learning, Vol. 1, No. 1. pp. 81-106.