International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Dynamic Prediction of Loan Approval using Machine Learning Techniques

Siva Adapa

Student, Rajam, Vizianagaram, 532127, India.

ABSTRACT

There are many improvements in the banking industry nowadays as a result of technological innovation. The yield or loss of a bank mostly depends on loans, or whether the borrowers are making their payments on time or not. The bank can lower its questionable asset by identifying the loan defaulters. Techniques for machine learning can be used to anticipate loan defaulters. Various machine learning techniques exist, including Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF). There are many improvements in the banking industry nowadays as a result of technological innovation. The yield or loss of a bank mostly depends on loans, or whether the borrowers are making their payments on time or not. The bank can lower its questionable asset by identifying the loan defaulters. Techniques for machine learning can be used to anticipate loan defaulters. Various machine learning techniques exist, including Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF).

Keywords: Decision Tree, Logistics Regression, Random Forest

1. Introduction

People today depend on bank loans to meet their requirements. In recent years, the number of loan applications has been rising at an extremely rapid rate. Loan acceptance always involves some level of risk. The loan amount being repaid by the bank's customers is something that the executives take extremely seriously. The judgments on loan approval are not always accurate, even after extensive safeguards and analysis of the loan applicant data. Automation of this procedure is required to make loan approval less hazardous and to prevent losses for banks. The field of artificial intelligence (AI) is one that is currently in development. Numerous issues in the actual world are resolved by the use of AI. An AI method called machine learning is highly helpful in prediction systems. A model is developed through machine learning using training data. The model created by the machine learning training process is employed while making the prediction. Loan prediction uses machine learning methods including Logistic Regression (LR), Random Forest (RF), and Decision Tree (DT). For loan acceptance, the models take into account the customer's personal characteristics, such as age, purpose, credit history, credit amount, and credit term. Using a portion of the available data, the machine learning algorithm tested the system after training it. The machine learning algorithms can be used to make predictions by first applying them to a sample of test data. The machine learning techniques are useful in the prediction of loan approval.

2. Literature Survey

In paper [1] A. Gupta, V. Pant, S. Kumar and P. K. Bansal proposed an efficient prediction system using supervised learning approach. This paper proposed machine learning algorithm on a loan prediction dataset and deploy the result using HTML, CSS at the local server. The proposed system is prepared in such a way that, it can accept new testing data and should also take part in training data and predict accordingly. The advantage of this model is, it can reduce the frauds and bank's time, employs checking every applicant on a priority basis at the time of loan approval.

In paper [2] P. Maheshwari and C. V. Narayana discussed how data science can impact the banking sector to improve their analysis of identifying risk by preprocessing the historical data of customers and building the model using machine learning techniques. This paper uses Exploratory Data Analysis (EDA), is to provide basic insights of any dataset. The main objective of EDA is to extract the essential patterns and visualize them in graphs, plots. In this paper a prototype model is built which can be used by the organization in making the right decision to approve or reject the loan.

In paper [3] V. Singh, A. Yadav, R. Awasthi and G. N. Partheeban discussed who is providing the loan they have to verify or set some criteria that who is taking the loan is able to return or not. The main objective of this paper is to predict whether a new applicant granted the loan or not using machine learning algorithms trained on the historical data set. The advantage of this system is that providing some conditions by setting the algorithms and just by evaluating the details, the paper concludes eligibility criteria that client is eligible or not.

In paper [4] M. A. Sheikh, A. K. Goel and T. Kumar proposed that bank should not only target the rich customers for granting loan but it should assess the other attributes of a customer and predicting the loan defaulters. In this paper, the process of prediction starts from cleaning and processing of data,

imputation of missing values, experimental analysis of data set and then model building to evaluation of model and testing on test data. The advantage of this model is those applicants who have high income and demands for lower amount of loan are more likely to get approved and predict accurately.

In paper [5] H. Ramachandra, G. Balaraju, R. Divyashree and H. Patil proposed a detailed analysis of loan approval methods using machine learning approach is provided. This paper describes how the algorithms have been implemented to predict the loan approval of customers and the output tested in terms of the predicted accuracy. The most widely used algorithms, such as Logistic Regression, Decision Tree, and Random Forest Technique, have been examined and reviewed in greater depth for loan prediction.

3. Data Collection

The dataset for the bank loan prediction system is drawn from the Kaggle competition and represents applicants of various ages and genders. The data set contains thirteen attributes, including information on education, marital status, income, assets, etc.

There are 981 records total of applicants, each with the values of the relevant numerical and categorical data. We manage missing values in the pre-processing and feature engineering of the data, normalise it, and then process it using an ML method. Data from the dataset are separated into training and test sets. The model is trained using machine learning methods and forecasts the performance of the system using test data, which is covered in more detail in the following section.

4. Machine learning

Machine learning is a branch of science that allows computers to learn without explicit programming. Machine learning is one of the most exciting technologies ever created. As the name suggests, the computer's capacity for learning is what gives it a more human-like character. In more places than one may imagine, machine learning is currently being actively deployed.

A thorough understanding of machine learning is essential for any aspiring data scientist or analyst who wishes to convert a sizable volume of unstructured data into trends and forecasts. The Machine Learning Foundation - Self Paced Course, which was designed and assembled by subject-matter experts with years of expertise, can help you learn this skill straight now. Here, cross validation is used.

For evaluating the effectiveness of machine learning models, cross validation is a highly helpful technique. Understanding how the machine learning model would generalise to a different collection of data is useful. Using this strategy, you want to gauge how accurate your model's predictions will be in real-world situations.

You will be given a known data set (training data set) and an unknown data set whenever you are given a machine learning assignment (test data set). In order to check for overfitting and gauge how well your machine learning model would generalise to independent data, such as the test data set provided in the challenge, you would "test" it using cross validation during the "training" phase.

For one cross validation round, you must divide your initial training data set into two parts:

Set for cross-validation training

Validation set or cross validation testing set

Your machine learning model will be trained using the cross-validation training set, and its predictions will be tested using the validation set. You may gauge how accurate your machine learning model's predictions are by contrasting the model's predictions on the validation set with the actual labels of the data points in the validation set.

In order to lessen the variation, numerous cross validation rounds are performed using different cross-validation training sets and cross-validation testing sets. The data from all the rounds are averaged to determine the machine's accuracy.

5. Methodology

In this paper we are using three Machine Learning algorithms which are used to find out the correct prediction of loan approval.

Machine learning algorithms which are used to make a model are as given below:

1. Logistic Regression
2. Decision Tree
3. Random Forest

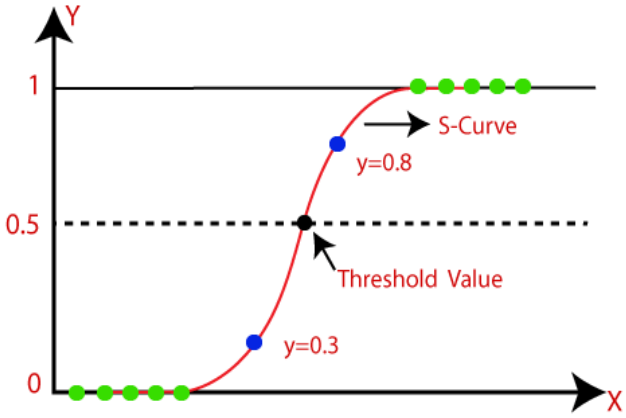
Method 1: Logistic Regression and Decision Tree:

Logistic Regression:

One of the most well-known and effective Machine Learning algorithms, categorised as Supervised Learning, is logistic regression. The statistical classification issues are solved using logistic regression. Fundamentally, the purpose of logistic regression is to establish the link between a dependent binary variable and a nominal or other independent variable. In order to forecast the results of a specific dependent variable, one uses logistic regression. As a result, the result must have a limited value. It can be either True or False, Yes or No, 1 or 0, etc. In essence, logistic regression classifies the data and reduces outliers using a logistic function.

$$Logistic\ function = \frac{1}{1+e^{-x}}$$

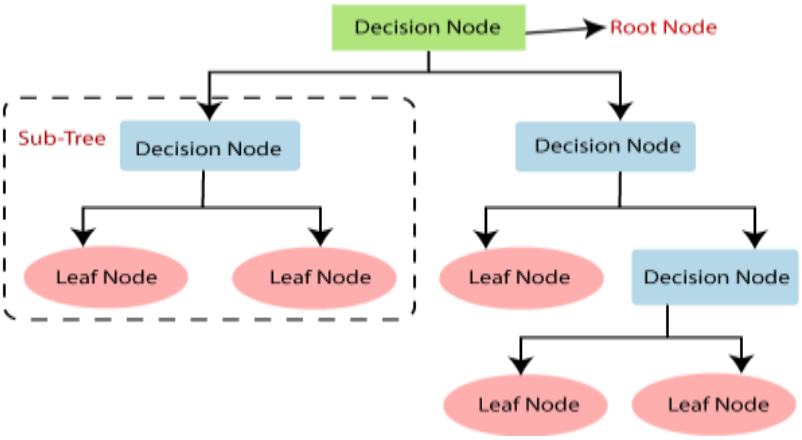
In logistic regression, we fit a "S" shaped logistic curve instead of a regression line, which predicts either 0 or 1. The logistic regression's value must lie within the range of 0 and 1, and because it cannot go beyond these bounds, it takes the shape of a "S." The sigmoid function or logistic function is the name given to the S-shaped curve.



Decision Tree:

Problems involving classification can be solved using the supervised learning technique known as the decision tree. It is a tree-structured classifier, where internal nodes stand in for the dataset's features, branches for the rules of classification, and each leaf node for the result.

- A decision tree has two nodes: the Decision Node and the Leaf Node. A choice is made using a Decision node, which has several branches,
- whereas a Leaf node is the result of that decision and does not have any additional branches.
- Based on the characteristics of the provided dataset, judgments or tests are run.
- It is known as a decision tree because, like a tree, it begins with the root node and grows on subsequent branches to form a structure resembling a tree.



The Working process can be explained in the below steps:

Step-1: Begin the tree with the root node, says S, which contains the complete dataset.

Step-2: Find the best attribute in the dataset using **Attribute Selection Measure (ASM).**

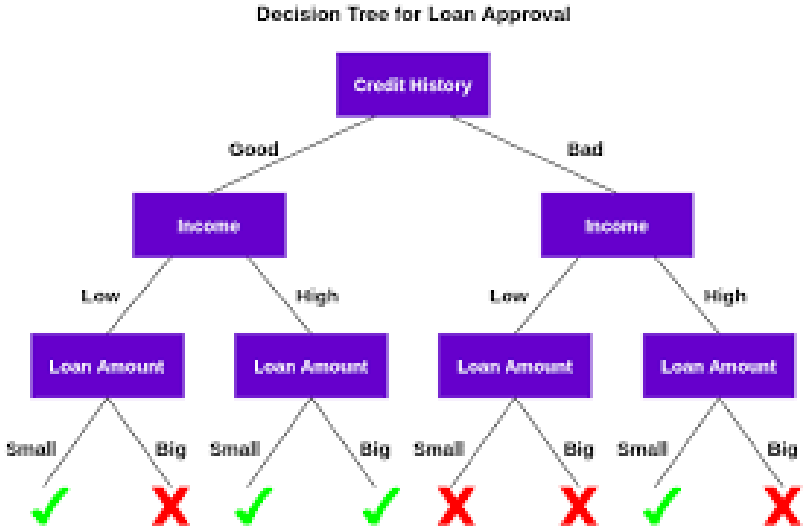
Step-3: Divide the S into subsets that contains possible values for the best attributes.

Step-4: Generate the decision tree node, which contains the best attribute.

Step-5: Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

Method 2: Random Forest:

- Popular machine learning algorithm Random Forest is a part of the supervised learning methodology. It can be applied to ML issues involving both classification and regression.
- It is based on the idea of ensemble learning, which is a method for mixing different classifiers to solve a challenging problem and enhance the performance of the model.
- Random Forest is a classifier that, as the name implies, "contains a number of decision trees on various subsets of the provided dataset and takes the average to enhance the predictive accuracy of that dataset." Instead, then depending on a single decision tree, the random forest uses forecasts from each tree and predicts the result based on the votes of the majority of predictions.
- Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.



The Working process can be explained in the below steps:

Step-1: Select random data points from the training set.

Step-2: Build the decision trees associated with the selected data points (Subsets).

Step-3: Choose the number N for decision trees that you want to build.

Step-4: Repeat Step 1 & 2.

Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

6. Results and Discussion

- In this paper, machine learning algorithms are used to predict loan acceptance. The Confusion Matrix (CM) is used to analyze and determine the performance of the proposed loan prediction model.
- The method includes dataset collection, feature selection, training, testing and comparison of results of prediction accuracy of machine learning algorithms.

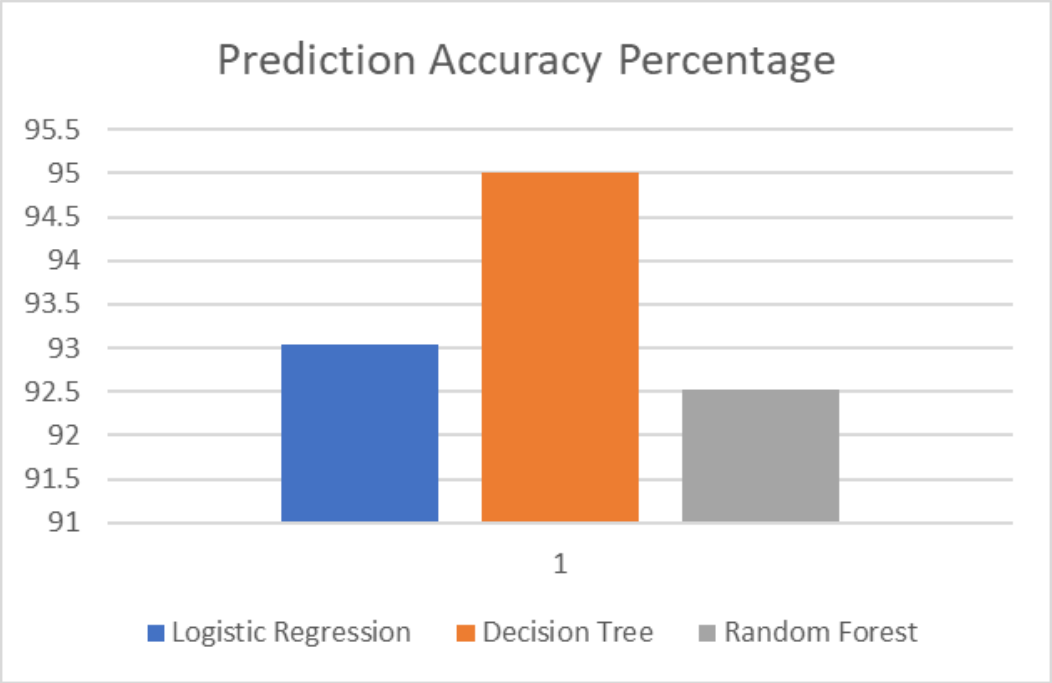
- Dataset is obtained from Kaggle competition. The features like gender, education, married status, income, assets, etc. are selected for training and testing.
- For training 80 percent data is used and 20 percent data is used for testing. The prediction accuracy of the different ML approaches is calculated and compared.
- The table below represents the resultant data, which is subsequently used for training and testing the predictive models. After preparing the final dataset, apply the machine learning techniques to get the accuracy

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History
0	LP001015	Male	Yes	0	Graduate	No	5720	0	110.0	360.0	1.0
1	LP001022	Male	Yes	1	Graduate	No	3076	1500	126.0	360.0	1.0
2	LP001031	Male	Yes	2	Graduate	No	5000	1800	208.0	360.0	1.0
3	LP001035	Male	Yes	2	Graduate	No	2340	2546	100.0	360.0	NaN
4	LP001051	Male	No	0	Not Graduate	No	3276	0	78.0	360.0	1.0
5	LP001054	Male	Yes	0	Not Graduate	Yes	2165	3422	152.0	360.0	1.0
6	LP001055	Female	No	1	Not Graduate	No	2226	0	59.0	360.0	1.0
7	LP001056	Male	Yes	2	Not Graduate	No	3881	0	147.0	360.0	0.0
8	LP001059	Male	Yes	2	Graduate	NaN	13633	0	280.0	240.0	1.0
9	LP001067	Male	No	0	Not Graduate	No	2400	2400	123.0	360.0	1.0
10	LP001078	Male	No	0	Not Graduate	No	3091	0	90.0	360.0	1.0
11	LP001082	Male	Yes	1	Graduate	NaN	2185	1516	162.0	360.0	1.0
12	LP001083	Male	No	3+	Graduate	No	4166	0	40.0	180.0	NaN
13	LP001094	Male	Yes	2	Graduate	NaN	12173	0	166.0	360.0	0.0
14	LP001096	Female	No	0	Graduate	No	4666	0	124.0	360.0	1.0
15	LP001099	Male	No	1	Graduate	No	5667	0	131.0	360.0	1.0
16	LP001105	Male	Yes	2	Graduate	No	4583	2916	200.0	360.0	1.0
17	LP001107	Male	Yes	3+	Graduate	No	3786	333	126.0	360.0	1.0
18	LP001108	Male	Yes	0	Graduate	No	9226	7916	300.0	360.0	1.0
19	LP001115	Male	No	0	Graduate	No	1300	3470	100.0	180.0	1.0
20	LP001121	Male	Yes	1	Not Graduate	No	1888	1620	48.0	360.0	1.0
21	LP001124	Female	No	3+	Not Graduate	No	2083	0	28.0	180.0	1.0
22	LP001128	NaN	No	0	Graduate	No	3909	0	101.0	360.0	1.0
23	LP001135	Female	No	0	Not Graduate	No	3765	0	125.0	360.0	1.0
24	LP001149	Male	Yes	0	Graduate	No	5400	4380	290.0	360.0	1.0

Figure: Test Data Set

- The results of prediction accuracy of machine learning algorithms are shown in below table.

S.No.	Machine learning Algorithm	Prediction Accuracy Percentage
1	Logistic Regression	93.04
2	Decision Tree	95.0
3	Random Forest	92.53



Prediction Percentage Accuracy

- The effective algorithm is detected by comparing the accuracy based on results.
- All the three algorithms results similarly but Decision Tree algorithm results effectively in terms of accuracy score (95.0%).

7. Conclusion

Today's fast-growing IT industry needs to discover new technology and update the old technology that helps us to reduce human intervention and increase the efficiency of the work. In this paper, the Machine Learning algorithms like Logistic Regression (LR), Decision Tree (DT)and Random Forest (RF)are studied. In the analysis from the above algorithms, we can propose that Decision Tree algorithm is the best algorithm for loan approval prediction. Decision Tree (DT) is mostly used because it is easy to develop and provide most accurate predictive analysis. But, by considering the accuracy, all the three algorithms have similar accuracy.so, in future these algorithms can be applied on other data sets available for loan approvals to further investigate the accuracy.

8. REFERENCES

[1]. A. Gupta, V. Pant, S. Kumar and P. K. Bansal, "Bank Loan Prediction System using Machine Learning," 2020 9th International Conference System Modeling and Advancement in Research Trends (SMART), 2020, pp. 423-426, doi: 10.1109/SMART50582.2020.9336801.

[2]. P. Maheshwari and C. V. Narayana, "Predictions of Loan Defaulter - A Data Science Perspective," 2020 5th International Conference on Computing, Communication and Security (ICCCS), 2020, pp. 1-4, Doi: 10.1109/ICCCS49678.2020.9277458.

[3]. V. Singh, A. Yadav, R. Awasthi and G. N. Partheeban, "Prediction of Modernized Loan Approval System Based on Machine Learning Approach," 2021 International Conference on Intelligent Technologies (CONIT), 2021, pp. 1-4, doi: 10.1109/CONIT51480.2021.9498475.

[4]. M. A. Sheikh, A. K. Goel and T. Kumar, "An Approach for Prediction of Loan Approval using Machine Learning Algorithm," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020, pp. 490-494, doi: 10.1109/ICESC48915.2020.9155614.

[5]. H. Ramachandra, G. Balaraju, R. Divyashree and H. Patil, "Design and Simulation of Loan Approval Prediction Model using AWS Platform," 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), 2021, pp. 53-56, doi: 10.1109/ESCI50559.2021.9397049.

[6]. M. J. Hamayel, M. A. Abu Mohsen and M. Moreb, "Improvement of personal loans granting methods in banks using machine learning methods and approaches in Palestine," 2021 International Conference on Information Technology (ICIT), 2021, pp. 33-37, doi: 10.1109/ICIT52682.2021.9491636.

[7]. C N Sujatha, Abhishek Gudipalli, Bh Pushyami, N Karthik, B N Sanjana, "Loan Prediction Using Machine Learning and Its Deployement On Web Application", 2021 Innovations in Power and Advanced Computing Technologies (i-PACT), pp.1-7, 2021.

[8]. J. Lohokare, R. Dani and S. Sontakke, "Automated data collection for credit score calculation based on financial transactions and social media," 2017 International Conference on Emerging Trends & Innovation in ICT (ICEI), Pune, 2017, pp. 134-138.doi: 10.1109/ETIICT.2017.7977024

[9]. M. Bayraktar, M. S. Aktaş, O. Kalıpsız, O. Susuz and S. Bayracı, "Credit risk analysis with classification Restricted Boltzmann Machine," 2018 26th Signal Processing and Communications Applications Conference (SIU), Izmir, 2018, pp. 1-4.doi: 10.1109/SIU.2018.8404397

[10]. S. Yadav and S. Thakur, "Bank loan analysis using customer usage data: A big data approach using Hadoop," 2017 2nd International Conference on Telecommunication and Networks (TEL-NET), Noida, 2017, pp. 1-8.doi: 10.1109/TEL-NET.2017.8343582