Loan Approval Prediction using Machine Learning: A Review

Ritika Purswani[1], Sakshi Verma[2], Yash Jaiswal[3], Prof. Surekha M[4]

[1-3]Student, Dept. of Computer Science and Engineering, JSS Academy of Technical Education, Noida, India
[4]Assistant Professor, Dept. of Computer Science and Engineering, JSS Academy of Technical Education, Noida, India

---***---

Abstract- *With the improvement in the banking sector, many people are applying for bank loans, but the bank has limited assets that it can only grant to a limited number of people, so determining who the loan can be granted to and who will be a safer option for the bank is a typical process. This process of predicting if a loan should be approved or not can be automated using Machine Learning. This is accomplished by mining Big Data for previous records of the people to whom the loan was previously granted, following which the machine is trained using machine learning algorithms based on these records/experiences. Previous research in this era has revealed that there are numerous methods for studying the problem of loan default control. However, because accurate predictions are critical for profit maximization, it is critical to investigate the nature of the various methods and compare them. In this paper, various machine learning algorithms that have been used in past are discussed and their accuracy is evaluated.*

***Key Words*: Loans, Machine Learning, Big Data, Accuracy**

1.INTRODUCTION

Almost every bank's fundamental business is loan distribution. The majority of the bank's assets are directly derived from the profit made from the loans distributed by the bank. The primary goal in the banking industry is to place their funds in safe hands. Today, many banks/financial organizations grant loans following a lengthy verification and validation process, but there is no guarantee that the chosen applicant is the most deserving of all applicants. We can forecast whether a given application is safe or not using this approach, and the entire feature validation procedure is automated using machine learning techniques such as Exploratory Data Analysis, Feature Engineering, Logistic Regression.

1.1 Risks involved in loans

There are numerous hazards associated with bank loans, both for the bank and for those who obtain them. Risk analysis in bank loans necessitates an awareness of what risk entails. Risk refers to the likelihood of specific outcomes—or the uncertainty of those outcomes—particularly a current negative danger to achieving a current monetary activity. Credit risk is the possibility that the loan will not be repaid on time or at all; liquidity risk is the possibility that too many deposits will be withdrawn too quickly, leaving the bank short on cash; and interest rate risk is the possibility that interest rates on bank loans will be too low to earn enough money for the bank.

2. LITERATURE REVIEW

Numerous pieces of literature about loan prediction have been published already and are available for public usage.

1. Paper Name: Loan Prediction by using Machine Learning Models

Authors: Pidikiti Supriya, Myneedi Pavani, Nagarapu Saisushma

Description: Data collection and pre-processing, applying machine learning models, training, and testing the data were the modules covered in this paper. Outlier detection and removal, as well as imputation removal processing, were done during the pre-processing stage. To predict the chances of current status regarding the loan approval process, SVM, DT, KNN, and gradient boosting models were used in this method. To divide the dataset into training and testing processes, the 80:20 rule was used. Experimentation concluded that the Decision Tree has significantly higher loan prediction accuracy than the other models.

Results: Accuracy achieved: 0.811

Model used: Decision Tree

2. Paper Name: Credit Risk Analysis and Prediction Modelling of Bank Loans Using R

Authors: Sudhamathy G.

Description: Using the R package, this paper proposed a risk analysis method for sanctioning a loan for customers. Data selection, pre-processing, feature extraction and selection, building the model, prediction, and evaluation were among the steps involved in developing the model. The USI repository provided the dataset for evaluation and prediction. Because the most important and time-consuming step is pre-processing, classification and clustering techniques in R were used to prepare the data for further use. The decision tree classifier was then built using the pre-processed dataset.

Results: Precision achieved: 0.833

3. Paper Name: Developing Prediction Model of Loan Risk in Banks using Data Mining

Authors: Aboobyda Jafar Hamid, Tarig Mohammed Ahmed

Description In this paper, three algorithms - j48, bayesNet, and naive Bayes - were used to create predictive models that can be used to predict and classify loan applications introduced by customers as good or bad by analyzing customer behaviors and previous payback credit. The model was built with the Weka application. It was discovered that the best algorithm for loan classification is the J48 algorithm after applying classification's data mining techniques algorithms such as j48, bayesNet, and Naive Bayes. The J48 algorithm is the best because it is highly accurate and has a low mean absolute error.

Results: Accuracy achieved: 78.37

Model used: j48

4. Paper Name: Loan Prediction Using Ensemble Technique

Authors: Anchal Goyal, Ranpreet Kaur

Description: Eleven machine learning models with nine properties are built in the proposed work to predict the credit risk of customers who have applied for a loan. This paper presented an ensemble model for loan predictions using several parameters such as Accuracy, Gini, AUC, Roc, and others to compare different training algorithms. The main goal of this paper is to evaluate the accuracy of models and to create an ensemble model that combines the outputs of three different models to predict customer loan amounts. The feature importance is calculated using Real Coded Genetic Algorithms. These features aid in predicting a customer's credit risk. The K-fold validation method is used to determine the predictive model's robustness.

Results: Maximum accuracy achieved: 81.25%

Model: Tree model for genetic algorithm

5. Paper Name: An Exploratory Data Analysis for Loan Prediction Based on Nature of the Clients

Authors: X. Francis Jency, V.P.Sumathi, Janani Shiva Sri

Description: This paper proposed Exploratory Data Analysis (EDA) as a method for predicting loan amounts based on the nature of the client and their needs. Annual income versus loan purpose, customer trust, loan tenure versus delinquent months, loan tenure versus credit category, loan tenure versus credit category, loan tenure versus the number of years in current job, and chances for loan repayment versus homeownership were the major factors concentrated during the data analysis. Finally, the purpose of this study was to infer the constraints that the customer faces when applying for a loan, as well as to make a prediction about repayment. Furthermore, the results revealed that customers were more interested in short-term loans than long-term loans.

3. PROPOSED METHODOLOGY

Since the problem of predicting the approval of a loan application is a classification problem, the model is trained using classification algorithms like Logistic Regression, Decision Tree, Random Forest Classifier, Support Vector Machine. The steps involved in building the model are specified in Fig-1.

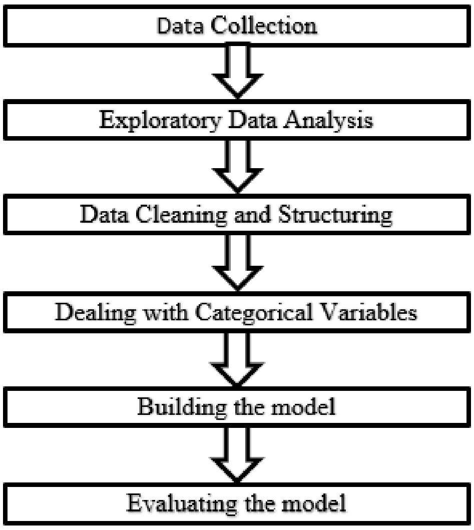
Data Collection
Exploratory Data Analysis
Data Cleaning and Structuring
Dealing with Categorical Variables
Building the model
Evaluating the model

Fig-1: Diagram of Proposed Methodology

4. MACHINE LEARNING ALGORITHMS

4.1 Logistic Regression

Logistic regression (Fig-2) is a popular Machine Learning algorithm that falls under the Supervised Learning technique. It is used to predict the categorical dependent variable from a given set of independent variables. We fit an "S" shaped logistic function that predicts two maximum values instead of a regression line (0 or 1). The output of a categorical dependent variable is predicted using logistic regression. It can be Yes or No, 0 or 1, true or false, and so on, but instead of giving the exact values as 0 and 1, it gives the probabilistic values that fall between 0 and 1.

In previous studies, model accuracy achieved after using Logistic Regression was around 0.81.

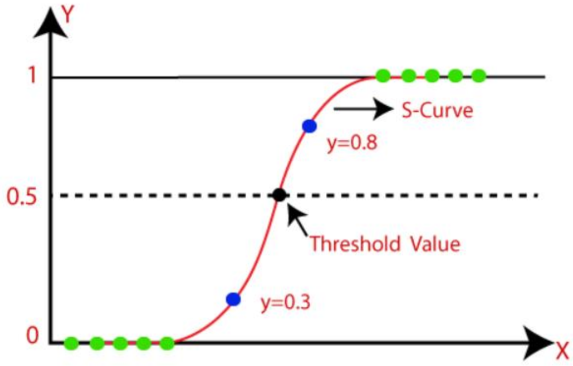

Fig-2: Logistic Regression

4.2 Decision Tree

The Decision Tree (Fig-3) algorithm is a member of the family of supervised learning algorithms. The goal of using a Decision Tree is to generate a training model that can be used to predict the class or value of the target variable (training data) by following the rules in the training data set. When using Decision Trees to predict a class label for a record, we begin at the root of the tree. The values of the root attribute are compared to the values of the record's attribute. We proceed to the next node by following the branch corresponding to that value based on the comparison. The accuracy achieved using Decision Tree by previous authors was varied around 0.82

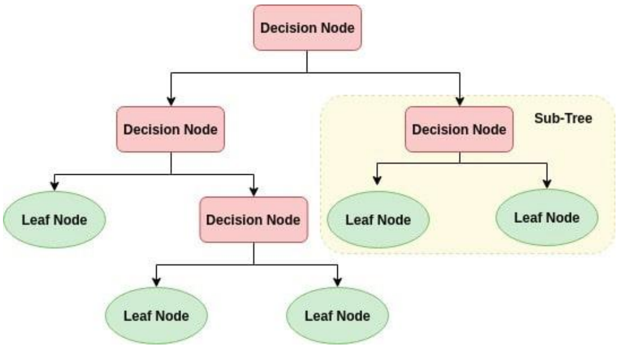

Fig-3: Decision Tree Classifier

4.3 Random Forest Classifier

Random Forest (shown in Figure 4) is a supervised learning algorithm. It creates a forest out of a collection of decision trees that are typically trained using the bagging method. The bagging method's basic premise is that combining different learning models improves the overall result. The random forest takes the predictions from each tree and predicts the final model based on the majority votes of predictions. The greater the proportion of trees in the forest,

the higher the precision and the less chance of overfitting.

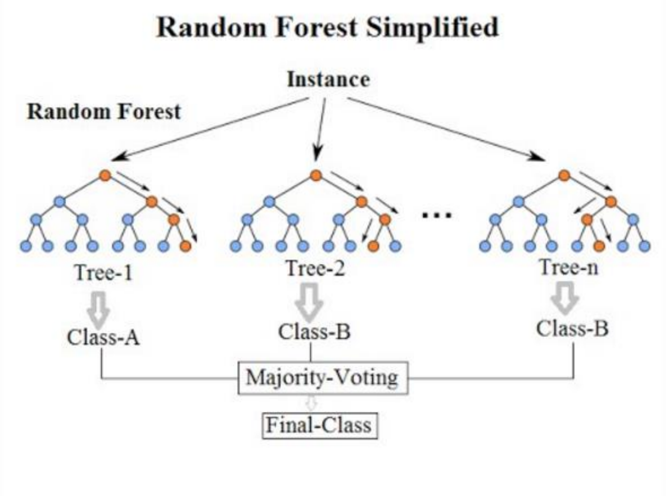

Fig-4: Random Forest Classifier

5. CONCLUSIONS

Loan companies grant loans after a thorough verification and validation process. However, they do not know with absolute certainty whether the applicant will be able to repay the loan without difficulty. The loan Prediction System will allow them to choose the most deserving applicants quickly, easily, and efficiently. It may provide the bank with unique benefits. In this paper, we have reviewed the process of building a Loan Approval Prediction System. Data collection, Exploratory Data Analysis, Data Preprocessing, Model Building, and Model Testing are the analytical processes involved in building this system. We conducted a thorough review of the previous research papers in this field in this paper. The most widely used algorithms, such as Logistic Regression, Decision Tree, and Random Forest Technique, have been examined and reviewed in greater depth.

REFERENCES

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

[2] Sudhamathy G.-"Credit Risk Analysis and Prediction Modelling of Bank Loans Using R", International Journal of Engineering and Technology (IJET), Vol. 8, No. 5, pp. 1954-1966, Oct-Nov 2016

[3] Aboobyda Jafar Hamid and Tarig Mohammed Ahmed - "Developing Prediction Model of Loan Risk in Banks using Data Mining"

[4] Anchal Goyal , Ranpreet Kaur-“Loan Prediction Using Ensemble Technique”, International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 3, March 2016

[5] X. Francis Jency, V.P.Sumathi, Janani Shiva Sri-“An Exploratory Data Analysis for Loan Prediction Based on Nature of the Clients”

[6] X. Francis Jency, V.P.Sumathi, Janani Shiva Sri-“An Exploratory Data Analysis for Loan Prediction Based on Nature of the Clients”