IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

Impact Factor: 5.731

Volume 9, Issue 1, September 2021

Analysis of Loan Availability using Machine Learning Techniques

Sharayu Dosalwar[1], Ketki Kinkar[2], Rahul Sannat[3], Dr Nitin Pise[4]

UG Student, School of Computer Engineering & Technology[1,2]
UG Student, School of Electronics & Communication Engineering[3]
Professor, School of Computer Engineering & Technology[4]
MIT World Peace University, Pune, Maharashtra, India

Abstract: *In the banking system, banks have a variety of products to provide, but credit lines are their primary source of revenue. As a result, they will profit from the interest earned on the loans they make. Loans, or whether customers repay or default on their loans, affect a bank's profit or loss. The bank's Non-Performing Assets will be reduced by forecasting loan defaulters. As a result, further investigation into this occurrence is essential. Because precise forecasts are essential for benefit maximisation, it's crucial to analyse and compare the various methodologies. The logistic regression model is an important predictive analytics tool for detecting loan defaulters. In order to assess and forecast, data from Kaggle is acquired. Logistic Regression models were used to calculate the various performance indicators. The models are compared using performance metrics like sensitivity and specificity. In addition to checking account details (which indicate a customer's wealth), the model is significantly better because it includes variables (customer personal attributes such as age, objective, credit score, credit amount, credit period, and so on) that should be considered when correctly calculating the probability of loan default. As a result, using a logistic regression approach, the appropriate clients to target for loan issuance can be easily identified by evaluating their plausibility of loan default. The model implies that a bank should assess a creditor's other attributes, which play a critical role in credit decisions and forecasting loan defaulters, in addition to giving loans to wealthy borrowers.*

Keywords: Logistic regression, Loan prediction, Data analysis, Machine learning models.

I. Introduction

Banks have many products to sell in our banking system, but their main source of income is their credit lines. As a result, they are likely to profit from the interest on the loans they make. Loans, or whether customers repay or default on their loans, affect a bank's profit or loss. The bank can minimize its Non-Performing Assets by forecasting loan defaulters. Because precise predictions are crucial for maximising earnings, it's essential to look at the different methodologies and compare them.

A logistic regression model is a critical approach in predictive analytics for analysing the problem of predicting loan defaulters. Kaggle data is taken in order to investigate and predict. Logistic Regression models were used to calculate the various performance measures. Model is significantly better because it includes variables (personal attributes of customers include graduation, dependents, credit score, credit amount, credit period, and so on.) other than checking account information (which indicates a customer's wealth) that should be considered when correctly calculating the probability of loan default. As a result, by evaluating the likelihood of default on a loan, the right customers to target for loan granting can be easily identified using a logistic regression approach. The model predicts that a bank should not solely approve loans to wealthy consumers rather should also consider a customer's other characteristics, which play an important role in credit decisions and predicting loan defaulters.

As the demand for products and services rises, so does the amount of capital credit given, and people are more eager to take credit than ever before. As a result, computer software has replaced the human interface as more people from all over the world (Urban, Rural, and semi-urban) push for a high demand for credit.

A Machine Learning software algorithm has been developed in order to construct a robust and efficient software algorithm that classifies individuals based on 13 characteristics (Gender, Education, Number of Dependents, Marital Status, Employment, Credit Score, Loan Amount, and others) whether they would be eligible for a loan or not.

Although this is the first line of command, it will undoubtedly lower the workload of all other bank employees because the process will be automated to identify client segments and those who are qualified for a loan amount, allowing them to target those clients individually. And this will indicate whether or not the loan applicant meets the eligibility criteria for loan approval based on those 13 elements. To provide a convenient, prompt, and accurate method of selecting deserving applicants for loan eligibility. To determine the model's accuracy, calculate the accuracy score, which is the level of precision displayed by the model when forecasting the applicant's loan eligibility. There are many Machine Learning models that can also be used for the prediction of the loan eligibility of an applicant. Some of the models have been discussed below:

1.1 Support Vector Machine (SVM)

Support vector machines are learning models that employ an association learning technique to examine features and identify pattern information, which is then used to classify applications. SVM can reliably translate their inputs into high-dimensional feature spaces using the kernel method, resulting in a productive-regression. A support vector machine (SVM) is a supervised machine learning algorithm that includes classification techniques to solve two-group classification problems. SVM models can categorise new 'text' after being given sets of labelled training data for each category. They have two key advantages over newer algorithms like neural networks: greater speed and better performance with a limited number of samples (in the thousands). This makes the approach particularly well suited to text classification issues, where it's common to only have access to a few thousand tagged samples.

1.2 Decision Trees

All attributes or features must be discretized in order for the decision tree's basic algorithm to work. The most information gain of features is used to determine feature selection. IF-THEN rules can be used to represent the knowledge shown in a decision tree. Decision Tree Analysis is a general-purpose predictive modelling tool with applications in a variety of fields. In general, decision trees are built using an algorithm that determines multiple ways to segment a data set based on certain conditions. It is one of the most popular and practical supervised learning algorithms. Decision Trees are a supervised non-parametric learning method that may be utilised for both classification and regression applications. The goal is to learn simple decision rules from data attributes to develop a model that predicts the value of a target variable.

1.3 Random Forest (RF)

Random forest is a group learning system for characterization (and relapse) that works by building a large number of Decision trees over time and generating the class that is the mode of the classes generated by individual trees.. Random forest is a supervised learning technique that can be used to classify and predict data. However, it is mostly employed to solve categorization issues. As we all know, a forest is made up of trees, and more trees equal a healthier forest. The random forest technique, similarly, builds decision trees from data samples, extracts predictions from each, and then votes on the best alternative. It's an ensemble technique that's better than using a single decision tree because it averages the outcomes to avoid over fitting.

1.4 Linear Models (LM)

Although the Linear Model is quantitatively indistinguishable from other regression analyses, it has limitations in terms of its applicability for a variety of qualitative and quantitative variables.

1.5 Logistic Regression (LR)

Logistic regression is a technique for describing data and explaining the relationship between one or more independent factors and one or more dependent binary variables. Logistic regression, like all other regression analyses,

DOI: 10.48175/IJARSCT-1895

is a predictive technique that is employed when the dependent variable is categorical. The logistic regression statistical model is a prominent statistical model for binary classification, or predictions of the kind this or that, yes or no, A or B, and so on. Although logistic regression can be used for multi-class classification, we shall concentrate on its most basic application in this paper. It's one of the most common machine learning methods for binary classifications, converting the input to 0 or 1. When the dependent variable has a binary solution, logistic regression is the best regression methodology to use. Logistic Regression, like all other forms of regression systems, is a type of predictive regression system. The link between one dependent binary variable and one or more independent variables is evaluated using logistic regression.

1.6 XGBoost Classifier

XGBoost is a machine learning method that has recently dominated Kaggle tournaments for structured or tabular data. XGBoost is a high-speed and high-performance implementation of gradient boosted decision trees. There are few frills in the library because it is laser-focused on computing speed and model performance. It does, however, include a lot of advanced functions. The algorithm's implementation was designed to maximise computation time and memory resources. To train the model, one of the design goals was to make the most of available resources.

1.7 K-Nearest Neighbors (KNN)

The KNN algorithm is a supervised machine learning method that may be applied to both classification and regression prediction problems. In industry, nonetheless, it is mostly used to solve classification and prediction problems. The KNN method predicts the values of new data points using ‘feature similarity,' this implies that a value will be assigned to the new data point based on how closely it resembles the points in the training set. It's a versatile method because it can be used for both classification and regression. KNN can be used in the banking system to forecast whether or not a person is eligible for a loan, as well as to determine a person's credit rating by comparing them to others who share similar characteristics.

1.8 Naive Bayes

Naive Bayes is a machine learning model that is used for huge amounts of data. It is recommended that you utilise Naive Bayes if you are working with data that has millions of records. When it comes to NLP tasks like sentimental analysis, it performs admirably. It's a simple and quick categorization algorithm. A Naive Bayes classifier is a machine learning model that separates objects based on specific variables' properties. It's a classification algorithm based on the Bayes theorem. For each class, such as the likelihood of data points linked with a given class, membership probabilities are predicted.

II. Literature Survey

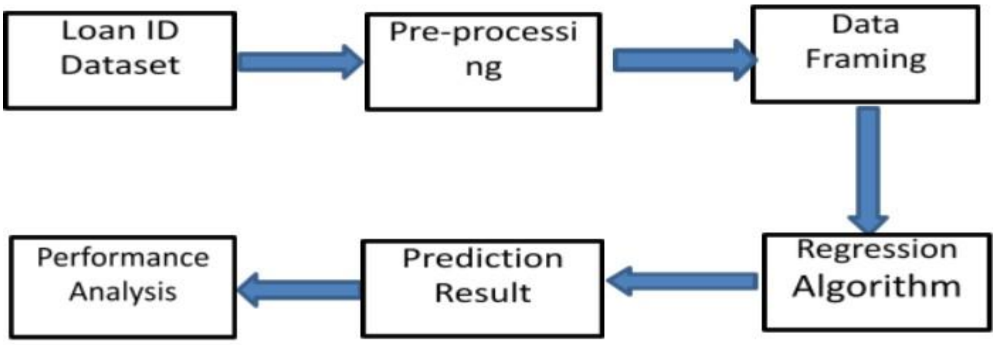
Sheikh, M. A., Goel, A. K., and Kumar, T. [1] used data from previous customers of various banks who had loans approved based on a set of criteria. To get accurate results, the machine learning model is trained on that record. The study's main purpose is to forecast the loan's safety. The logistic regression algorithm is used to predict loan safety. To avoid missing values in the data set, the data is first cleaned. The model was trained using a data set of 1500 cases with 10 numerical and 8 categorical attributes. Various parameters such as CIBIL Score (Credit History), Business Value, Customer Assets, and soon have been taken into account when crediting a loan to a customer. Vaidya [2] talks about logistic regression and how to represent it mathematically. His study employs logistic regression as a machine learning technique to actualize the predictive and probabilistic methods to a particular problem of loan approval prediction. This study employs logistic regression to determine if a loan for a set of records belonging to an applicant will be authorised. It also covers some of the Machine Learning mode's other real-world uses.

Zhang, H., Li, Z., Shahriar, H., Tao, L., Bhattacharya, P., and Qian, Y. [3] presented a logistic regression analysis using Python on imbalanced datasets and determined different classification thresholds based on the data proportion of imbalanced datasets. The research of Zou, X., Hu, Y., Tian, Z., and Shen, K. [5] focused on the logistic mathematical

DOI: 10.48175/IJARSCT-1895

model, the definition of the error function, the gradient descent method for calculating the regression coefficient, and the Sigmoid function improvement. Therefore, the number of repetitions has been reduced, the classification impact has been improved, and the accuracy has remained nearly unchanged. Kumar Arun, Garg Ishan, and Kaur Sanmeet[6] have demonstrated how to reduce the risk factor when picking a safe individual in order to save time and money for the bank. This is performed by mining Big Data of previous records of persons to whom the loan was previously provided, and the machine was taught to get the best accurate result using a machine learning model based on these records/ experiences.

III. Methodology



Block Schematic of Loan Prediction

Figure 1: Block diagram

When an algorithm receives data as input, it produces binary output, which is either 0 or 1. If the result is 1, the number 1 will be displayed, indicating that the loan has been accepted. If the output is 0, the number '0' will be displayed, indicating that the loan has been denied. The prediction process includes phases such as data cleaning and processing, imputation of missing values, experimental analysis of a data set, model creation, and testing on test data. Various input variables were employed to obtain the output in order to implement this.

IV. Experimental Analysis

Model used	**Accuracy**
Logistic Regression	0.785
Decision Tree Classifier	0.662
K Neighbors Classifier	0.619
Naive Bayes	0.779
Random Forest Classifier	0.773
Support Vector Machine	0.650
XGBoost Classifier	0.773

It is observed that Logistic Regression gives better accuracy for loan availability prediction. The reason behind this is that Logistic regression is also known as logit regression or logistic model. It accepts independent features and produces categorical results.

By fitting the features in the logistic curve to the logistic regression model, the probability of occurrence of a categorical output may also be determined. The general logit curve is seen in the diagram below.

The Logistic Regression model can be replaced with the simpler Linear Regression model when the output variable is believed to be continuous. A separate model must be employed to account for the difference when the output variable is not continuous or dichotomous. Following that, numerous models were created to account for the dichotomous nature of the outcome variable. Because of its mathematical clarity and versatility, the Logistic Regression model was chosen above the other models.

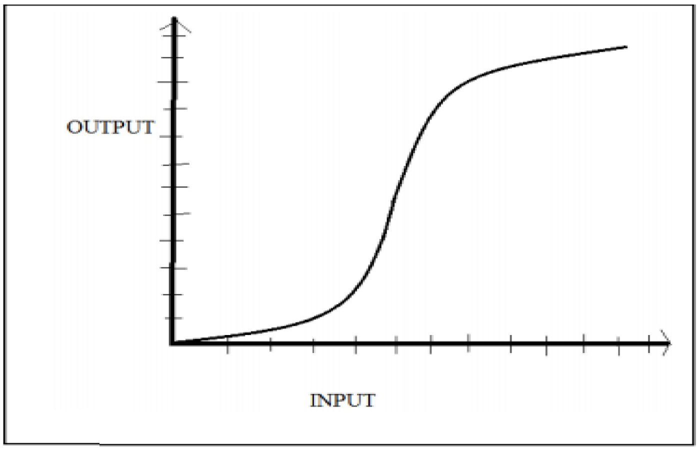


Figure 2: General Logit Curve

It's a prediction analysis. It's used to explain and describe the relationship between a single binary variable and one or more independent variables. Furthermore, the sigmoid function is taken into account in the logistic regression because the outcome is binary[4]. In this model, there can be one or more predictors. With this model, we can describe the target variable's natural log probability in the linear form of the feature variables used as input. This can be expressed mathematically as:

$$\text{logit}(y) = \ln(\frac{p}{1-p}) = \alpha+\beta x$$

The parameters of the logistic regression model are and ß, including p represents the probability of the desired category output and x represents the input feature. We can easily find the probability of the desired result by taking antilog on both sides of (1). The following is a mathematical representation:

$$P = \frac{1}{1+e^{-(\alpha+\beta x)}}$$

If more than one parameter is utilised as a feature and must be used for prediction, The natural log of probability for the desired variable is represented mathematically as follows:

$$\text{logit}(y) = \ln(\frac{p}{1-p}) = \alpha+\beta_1 x_1+\beta_2 x_2+....\beta_l x_l$$

α, $\beta_1\beta_1$ … $\beta_l\ \beta_l$ are the logistic regression parameters, and x1...xl are the characteristics used to fit the model. Using antilog on both sides of the equation yields a result that is similar to but more extended than the second equation, which is given by

$$P = \frac{1}{1+e^{-(\alpha+\beta_1 x_1+\cdots+\beta_l x_l)}}$$

In sectors where a relationship between features must be constructed and a dichotomous output must be achieved, logistic regression is increasingly widely used [3].

IV. Conclusion And Future Scope

In our model prediction of whether the loan would be accepted or not, we achieved the highest accuracy from the logistic regression model. On the dataset, the best case accuracy attained is 0.785. Our model was able to forecast whether the applicants in the dataset would be eligible for the loan when the project was completed. It was also able to anticipate the loan eligibility of a specific applicant by pointing out his row number. Applicants with a high income and smaller loan requests are more likely to be approved, which makes sense because they are more likely to payback their debts. Gender and marital status, for example, do not appear to be considered. The loan credibility prediction system can assist companies in making the best judgement on whether to approve or deny a customer's loan request. This will undoubtedly assist the banking industry in establishing more effective distribution routes. It is necessary to create and test new strategies that outperform the performance of common data mining models for the domain. As a result, in the near future, the so-called algorithm might be made more reliable, efficient, and robust. This prediction module may be integrated with the automated processing system module in the near future. The system is currently trained on an existing training dataset, but algorithms can be implemented in the future to allow additional testing data to be included in the training dataset.

REFERENCES

[1]. Sheikh MA, Goel AK, Kumar T. An Approach for Prediction of Loan Approval using Machine Learning Algorithm. In2020 International Conference on Electronics and Sustainable Communication Systems (ICESC) 2020 Jul 2 (pp.490-494).

[2]. Vaidya A. Predictive and probabilistic approach using logistic regression: application to prediction of loan approval. In2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT) 2017 Jul 3 (pp.1-6).

[3]. TejaswiniJ, Kavya TM, Ramya RD, Triveni PS, Maddumala VR. Accurate Loan Approval Prediction Based On Machine Learning Approach. Journal of Engineering Science. 2020Apr;11(4):523-32.

[4]. Zhang H, Li Z, Shahriar H, Tao L, Bhattacharya P, Qian Y. Improving prediction accuracy for logistic regression on imbalanced datasets. In2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC) 2019 Jul 15 (Vol. 1, pp.918-919).

[5]. Zou X, Hu Y, Tian Z, Shen K. Logistic Regression Model Optimization and Case Analysis. In2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT) 2019 Oct 19 (pp.135-139).

[6]. Arun K, Ishan G, Sanmeet K. Loan approval prediction based on machine learning approach. IOSRJ. Comput. Eng. 2016;18(3):18-21.

[7]. Dutta, P., A Study On Machine Learning Algorithm For Enhancement Of Loan Prediction. International Research Journal of Modernization in Engineering Technology and Science,2021

[8]. Goyal, A. and Kaur, R., 2016. A survey on ensemble model for loan prediction. International Journal of Engineering Trends and Applications (IJETA), 3(1),pp.32-37.

[9]. Ruzgar, B. and Ruzgar, N.S., 2008. Rough sets and logistic regression analysis for loan payment. International journal of mathematical models and methods in applied sciences, 2(1),pp.65-73.

[10]. TVS, J., 2021. Predicting the Loan Status using Logistic Regression and Binary Tree. International Conference on IoT based Control Networks and Intelligent Systems (ICICNIS2020)

[11]. Dutta, P., A Study On Machine Learning Algorithm For Enhancement Of Loan Prediction. International Research Journal of Modernization in Engineering Technology and Science

[12]. Kwofie, C., Owusu-Ansah, C. and Boadi, C., 2015. Predicting the probability of loan-default: An application of binary logistic regression. Research Journal of Mathematics and Statistics, 7(4),pp.46-52.

[13]. Agbemava, E., Nyarko, I.K., Adade, T.C. and Bediako, A.K., 2016. Logistic regression analysis of predictors of loan defaults by customers of non-traditional banks in Ghana. European Scientific Journal,12(1).

[14]. DM, O. and Muraya, M.M., 2018. Comparison of Accuracy of Support Vector Machine Model and Logistic Regression Model in Predicting Individual Loan Defaults. American Journal of Applied Mathematics and Statistics, 6(6),pp.266-271.

[15]. Rath, G.B., Das, D. and Acharya, B., 2021. Modern Approach for Loan Sanctioning in Banks Using Machine Learning. In Advances in Machine Learning and Computational Intelligence (pp. 179-188). Springer, Singapore.

DOI: 10.48175/IJARSCT-1895