Loan Default Prediction using Machine Learning Models

***Vijay Kumar*[1]*, Rachna Narula*[2]*, Akanksha Kochhar*[3]**
[1, 2, 3]*Assistant Professor, Department of Computer Science and Engineering, Bharati Vidyapeeth's College Of Engineering, Affiliated to GGSIPU, Delhi, India*

****Corresponding Author***
E-Mail Id: *- vijay.kumar@bharatividyapeeth.edu*

ABSTRACT

Borrowing from financial institutions has become commonplace in today's society. Many people submit loan applications each day for a variety of reasons. But not every one of these candidates is reliable, and not everyone is accepted. A significant part of bank loans are frequently not returned each year, leaving the bank with enormous losses. Making a choice to approve a loan involves significant risks.Therefore, the goal of this project is to gather credit data from a variety of sources and then use various machine learning techniques to extract key information. With the use of this model, businesses can decide whether to approve or reject consumer loan requests. This article examines actual bank credit data, performs numerous

Keywords*:-Machine learning, bank credit, classification, confusion matrix, predictive analysis.*

INTRODUCTION

One of the most important factors affecting us- The economy and financial state of the country Bank-regulated credit system Banks play an important role in the market economy. An organization's success or failure largely depends on the industry's ability to assess credit risk. The banking system uses a manual process for checking whether a borrower is a defaulter or not. Most likely the manual interaction will be more exact and viable, yet this cycle can't work when there are countless advance applications simultaneously. On the off chance that there happens a period like this, the dynamic cycle will consume most of the day and furthermore heaps of labor supply will be required. Assuming that we can do the advance expectation it will be exceptionally useful for candidates and furthermore for the workers of banks. In this way, the errand is to group the borrower as fortunate or unfortunate i.e., whether the borrower will actually want to repay the obligations or not.

Any organization or bank finds it challenging to forecast a borrower's situation, i.e., whether they will be delinquent or not in arrears in the future. “As you are aware, there are numerous methods for assessing credit risk. Used when calculating risk magnitude. One of the primary responsibilities of the banking sector is credit risk. Predicting credit defaults is fundamentally an issue of binary categorization. Loan amount; creditworthiness for obtaining a loan is determined by customer history. However, creating such a model is a particularly challenging endeavor due to the rising demand for loans. An example of a model that a company could use to decide whether to approve or reject a customer's loan application. In this project, machine learning is trained using data.

LITERATURE REVIEW

In this field, plenty of time has been put in.

Research articles exhibit a variety of methods with respectable outcomes and analysis. Some of the most significant papers and their accuracy are highlighted in the following section:

Table 1:*-Literature Review*

SOURCE	**DESCRIPTION**	**RESULTS**	**YEAR**
[1]	Chosen five models for machine learning classification, the following algorithmshave neutral networks, discriminant analysis, naïve byes, k-nearest neighbors, Linear Regression, ensemble Learning method and Decision Trees.	Each of these algorithms achieved an accuracy rate between 76% to over 80%	2020
[2]	Chosen 4 models for machine learning classification:Support Vector Machine (SVM), Decision Tree, Logistic Regression, Random Forest.	Accuracy: Random Forest-76.42% Support Vector Machine-79.67%Decision Tree-70% Logistic Regression-75.60%	2021
[3]	Xgboost, Adaboost, LightGBM, Random Forest, Decision tree, and KNN Algorithms have been used	XgBoost-0.9180 AdaBoost-0.9187 LightGBM-0.9189 Random Forest-0.9188Decision tree-0.8497 KNN- 0.9167	2022
[4]	Decision Tree, Naive Bayes, Support Vector Machines, and Logistic Regression methods are the four methods	Logistic regression has the highestaccuracy of all the algorithms.	2022
[5]	Chosen two models for machine learning models which includes Logistic Regression and Random Forest.	Logistic Regression - 0.7214Random Forest:- 0.7947	2022
[6]	These algorithms has been used RandomForest , Naïve Bayes, Decision Tree, LogisticRegression, K Nearest Neighbor	Random Forest-79.03% Naïve Bayes-85.48% ,DecisionTree- 79.03%, Logistic Regression-88.70%, K Nearest Neighbor -80.64%	2022

DATA PRE-PROCESSING

A. Dataset

To assess the risk of loan defaults, we obtained historical loan approval data from Keggle. This dataset contains essential information such as DisbursementGross, MIS_Status, and RevLineCr, which are crucial in determining whether a loan should be approved or denied. We used the "National SBA" dataset to evaluate the default risk of each loan application based on these features. The dataset we collected is quite extensive, comprising 8,99,164 rows and 27 columns, providing us with a robust set of data to work with. By leveraging this dataset, we made informed decisions about loan approvals, helping to mitigate the risk of default and ultimately ensuring the long-term sustainability of the lending institution.

B. Data Pre-Processing

The 'MIS_Status' table has an entry with the target variable. We changed them to either 0 or 1 depending on the values they had before. The entire set of data was then normalized based on their mean deviation in order to remove any disparities in the data. Next, the entire set of data was split into three equal halves for training, validation, and testing, with the ratios being 70:15:15.The observations in the training set act as the algorithm's classroom and are used to calibrate a classifier's parameters. A validation dataset is a set of instances used to fine-tune the hyperparameters (i.e., the architecture) of a classifier. The test set is a set of data used to determine whether using a performance metric to judge the model is effective. It is essential that there are no observations from the training set in the test set. If a model correctly fits both the training and test datasets, it has only slightly overfitted.

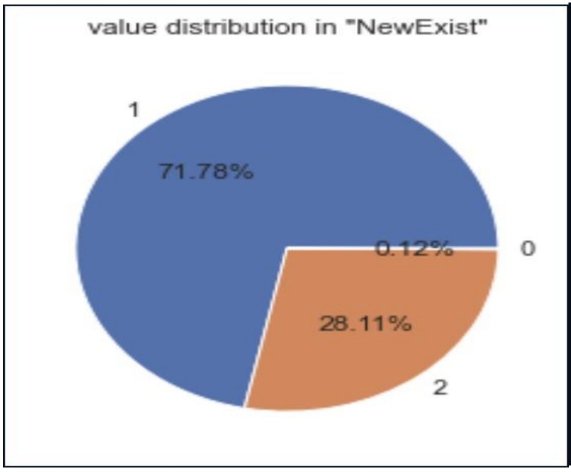

***Fig.1:**-Shows if SBA is new or old*

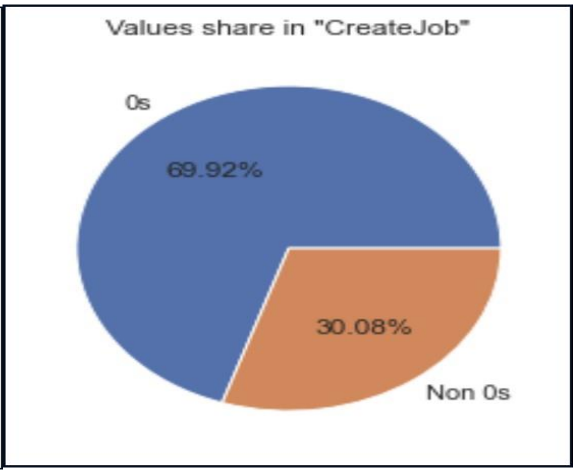

***Fig.2:**-Shows no. of new jobs created*

WORKING METHODOLOGY

The following sections describe all the steps of model training using selected algorithms -

C. Splitting Dataset

In machine learning, it's common practice to split our data into two parts: the training set and the test set. The training set contains labeled examples that the model uses to learn and improve its performance on unseen data.

On the other hand, the test set is used to assess the model's predictions on portions of the data it has not seen before. To employ this splitting approach, we need to import the pandas library. Training Dataset: Dataset of 70% has been used for training. Testing Dataset: Dataset of 30% has been used for testing.

D. Algorithms

- K-Nearest Neighbors (KNN): The KNN algorithm is versatile and can be applied to both classification and regression problems, though it is primarily used for classification. In the kNN algorithm, 'k' refers to the number of nearest neighbors from which we want to obtain a vote [16].
- Logistic Regression: Logistic Regression is a classification algorithm used in supervised learning that predicts the probability of a target variable. The target variable is binary, meaning it can only have two possible outcomes or classes [17].

E. Training and Testing Model

In order to train the model, a split ratio of 7 : 3 was employed, and the K-NN and

Logistic Regression algorithms were imported from the sci-kit library in Python. These algorithms are commonly used for classification tasks in supervised machine learning.

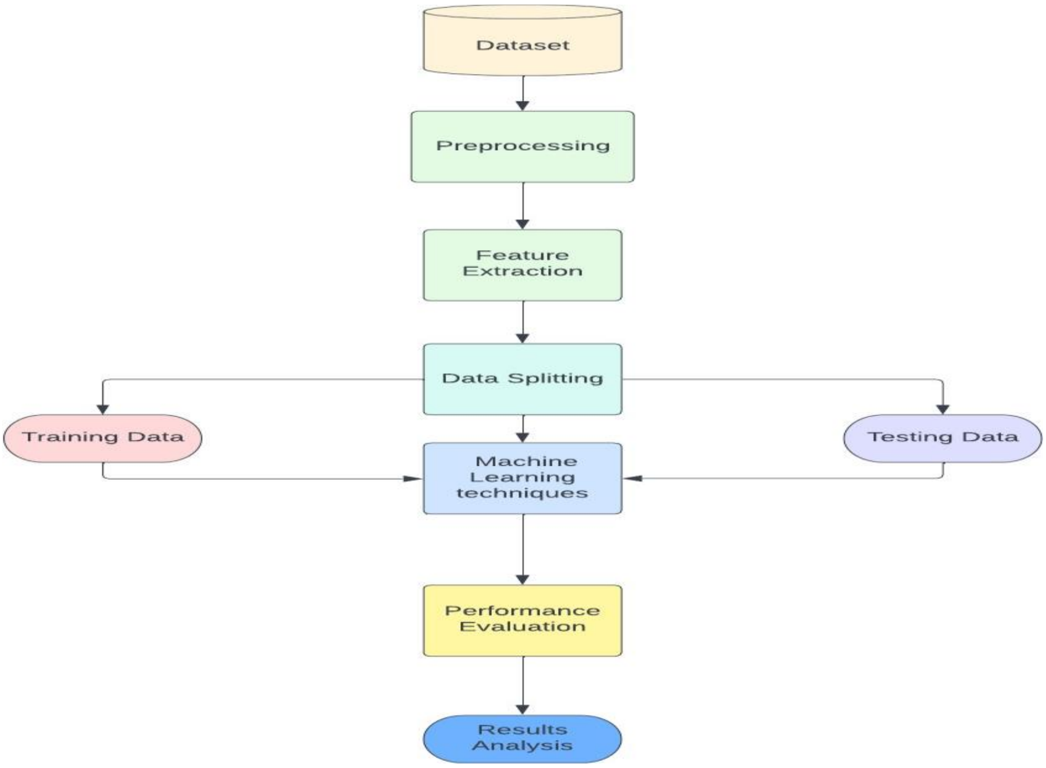

Fig.3:-*Step-wise Process to Train Our Model*

EXPERIMENTAL RESULTS

The accuracy of the model is measured after it is compared to algorithms like logistic regression and K-NN and trained with a split ratio of 7:3.

Table 2:-*Algorithm Accuracy*

Split Ratio of Dataset	**Logistic Regression (%)**	**K-Nearest Neighbors (%)**
7:3	98.30	98.3

Table II shows the obtained accuracy of approx 75% using logistic regression and an accuracy of 90.61% using the KNN algorithm.

CONCLUSION AND FUTURE SCOPE

In conclusion, the banking sector has enormous potential for using machine learning to automate procedures and enhance decision-making. Its article's experiment, which gathers credit information from numerous sources and uses machine learning algorithms to determine creditworthiness, is an excellent illustration of its potential. The precision and effectiveness of loan approval procedures will only increase as machine learning algorithms continue to evolve and new data sources become accessible. Banks may see significant cost savings as a result, while customers may enjoy better service due to quicker loan approvals and more individualized credit determinations. This paper has addressed the loan approval hassle using different machine learning algorithms The study's findings demonstrated the precision of KNN and Logistic Regression models, having true positive 81.84% and 81.96% accuracy respectively. The heatmaps show the performance of the models.

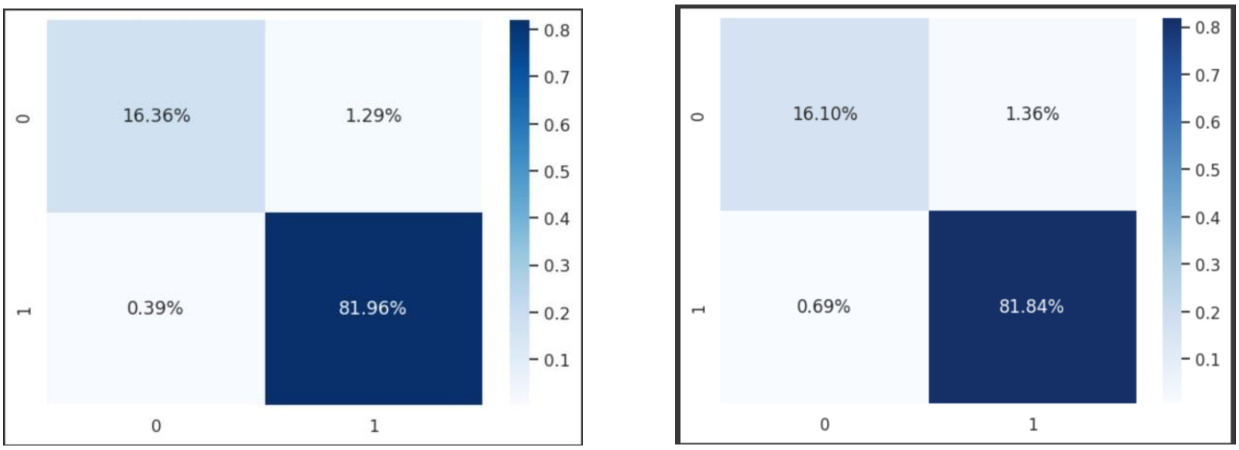

***Fig 4 & 5:**-shows heat map of confusion matrix for Logistic Regression[Fig.4] and KNN[Fig.5]*

The future scope for this project is quite promising as the use of machine learning in the banking industry is becoming increasingly popular. Here are some potential directions for future development:

Application of Deep Learning: Machine Learning approaches have been seen with these types of projects, but use of Deep Learning on this domain can help us forecast unforeseen data and prediction level quality.

Integration of more data sources: The current project collects credit data from multiple sources, but there may be other sources of data that could be used to better assess creditworthiness. For example, social media data could be used to analyze an applicant's online behavior and predict their likelihood of defaulting on a loan.

Development of more sophisticated algorithms: While the current project uses multiple machine learning algorithms to assess creditworthiness, there are always new and more sophisticated algorithms being developed. In the future, more advanced techniques such as deep learning and reinforcement learning could be used to further improve the accuracy of credit assessments.

Increased automation: The current project is an automated banking risk system, but there may be further opportunities to automate the loan approval process. For example, machine learning could be used to automatically approve loans up to a certain amount based on the applicant's credit score and other factors.

REFERENCES

1. Aphale, A. S., & Shinde, S. R. (2020). Predict loan approval in banking system machine learning approach for cooperative banks loan approval. *International Journal of Engineering Trends and Applications (IJETA)*, *9*(8).
2. Nitesh Pandey1 , Ramanand Gupta2 , Sagar Uniyal3 , Vishal Kumar4,AKTU © June 2021| IJIRT | Volume 8 Issue 1 | ISSN: 2349-6002
3. Shubham Nalawade1, Suraj Andhe1, Siddhesh Parab1, Prof. Amruta

Sankhe2 Vol:09 Issue:04 April'22

4. Nikhil Bansode*1, Adarsh Verma*2, Abhishek Sharma*3, Varsha Bhole*4 Vol:05/Issue:04/May 2022
5. Afia Farjana and Muntasir Mamun Vol:05/Issue:04/June/2022
6. Akinsola, J E T. (2017). Supervised Machine Learning Algorithms: Classification and Comparison. International Journal of Computer Trends and Technology (IJCTT). 48. 128 - 138. 10.14445/22312803/IJCTT-V48P126.
7. IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 3, Ver. I (May-Jun. 2016), PP 79-81
8. 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)*, Coimbatore, India, 2020, pp. 490-494, doi: 10.1109/ICESC48915.2020.9155614.
9. Ekta Gandotra, Divya Bansal, Sanjeev Sofat 2014, 'Malware Analysis and Classification: A Survey'
10. K. Hanumantha Rao, G. Srinivas, A. Damodhar, M. Vikas Krishna: Implementation of Anomaly Detection Technique Using Machine Learning Algorithms: Internatinal Journal of Computer Science and Telecommunications (Volume2, Issue3, June 2011).
11. J.M. Chambers. Computational methods for data analysis. Applied Statistics, Wiley, 1(2):1–10, 1077.
12. Puneeth B. R, Ashwitha K, Arhath Kumar, Balachandra Rao, Preethi Salian K, Supravi A P, "An Approach to Predict Loan Eligibility using Machine Learning", *2022 International Conference on Artificial Intelligence and Data Engineering (AIDE)*, pp.23-28, 2022.
13. Narayana Darapaneni, Pramod Srinivas, Keerthi Reddy, Anwesh Reddy Paduri, Lakshmikanth Kanugovi, Pavithra J, Sudha B G, Bharath S, "Tree Based Models: A Comparative And Explainable Study For Credit Default Classification", *2022 IEEE 9th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, pp.1-8, 2022.
14. Rareş Constantin, Moritz Dück, Anton Alexandrov, Patrik Matošević, Daphna Keidar, Mennatallah El-Assady, "How Do Algorithmic Fairness Metrics Align with Human Judgement? A Mixed-Initiative System for Contextualized Fairness Assessment", *2022 IEEE Workshop on TRust and EXpertise in Visual Analytics (TREX)*, pp.1-7, 2022.
15. Changwu Huang, Zeqi Zhang, Bifei Mao, Xin Yao, "Preventing Undesirable Behaviors of Neural Networks via Evolutionary Constrained Learning", *2022 International Joint Conference on Neural Networks (IJCNN)*, pp.1-7, 2022.
16. Narayana Darapaneni, Akshay Kumar, Archanna Dixet, Manikandan Suriyanarayanan, Shabd Srivastava, Anwesh Reddy Paduri, "Loan Prediction Software for Financial Institutions", *2022 Interdisciplinary Research in Technology and Management (IRTM)*, pp.1-8, 2022.
17. Praveen Tumuluru, Lakshmi Ramani Burra, M. Loukya, S. Bhavana, H.M.H. CSaiBaba, N Sunanda, "Comparative Analysis of Customer Loan Approval Prediction using Machine Learning Algorithms", *2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, pp.349-353, 2022.
18. Krishan Kumar Pandey, Abhishek

Giri, Saket Sharma, Anurag Singh, "Predictive Analysis of Classification Algorithms on Banking Data", *2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON)*, pp.1-5, 2021.

19. Richa Manglani, Anuja Bokhare, "Logistic Regression Model for Loan Prediction: A Machine Learning Approach", *2021 Emerging Trends in Industry 4.0 (ETI 4.0)*, pp.1-6, 2021.
20. Ankit Sharma, Vinod Kumar, "An Exploratory Study-Based Analysis on Loan Prediction", *Inventive Communication and Computational Technologies*, vol.383, pp.423, 2023.