A MACHINE LEARNING MODEL FOR LOAN ELIGIBILITY PREDICTION

Pendam Sai Krishna*1, Bankala Rajesh*2, Jilla Ruchinath*3, Dr. N. Sridhare*4

*1,2Student, Department Of IT. Malla Reddy Engineering College, Maisammaguda, Hyderabad, India.

*3,4Assoc.Professor, Department Of IT. Malla Reddy Engineering College, Maisammaguda, Hyderabad, India.

ABSTRACT

In our banking system, banks have many products to sell but main source of income of any banks is on its credit line. So they can earn from interest of those loans which they credits. Loan approval is a very important process for banking organizations. Banking Industry always needs a more accurate predictive modeling system for many issues. Predicting credit defaulters is a difficult task for the banking industry. A bank's profit or a loss depends to a large extent on loans i.e. whether the customers are paying back the loan or defaulting. By predicting the loan defaulters, the bank can reduce its Non- Performing Assets. This makes the study of this phenomenon very important. Previous research in this era has shown that there are so many methods to study the problem of controlling loan default. But as the right predictions are very important for the maximization of profits, it is essential to study the nature of the different methods and their comparison. A very important approach in predictive analytics is used to study the problem of predicting loan defaulters: The Logistic regression model. Logistic Regression models have been performed and the different measures of performances are computed. The models are compared on the basis of the performance measures such as sensitivity and specificity. The final results have shown that the model produce different results. Model is marginally better because it includes variables other than checking account information (which shows wealth of a customer) that should be taken into account to calculate the probability of default on loan correctly. Therefore, by using a logistic regression approach, the right customers to be targeted for granting loan can be easily detected by evaluating their likelihood of default on loan.

Keywords-: Banking System, Loan Approval, Credit Line, Loan Defaulters, Predictive Modeling, Credit Risk, Non-Performing Assets (NPA), Logistic Regression, Performance Measures, Sensitivity, Specificity, Customer Attributes, Credit History, Loan Default Prediction, Credit Granting, Predictive Analytics, Financial Prediction, Customer Profiling, Credit Decision-Making.

I. INTRODUCTION

As the data are increasing daily due to digitization in the banking sector, people want to apply for loans through the internet. Artificial intelligence (AI), as a typical method for information investigation, has gotten more consideration increasingly. Individuals of various businesses are utilizing AI calculations to take care of the issues dependent on their industry information. Banks are facing a significant problem in the approval of the loan. Daily there are so many applications that are challenging to manage by the bank employees, and also the chances of some mistakes are high. Most banks earn profit from the loan, but it is risky to choose deserving customers from the number of applications. One mistake can make a massive loss to a bank. Loan distribution is the primary business of almost every bank. This project aims to provide a loan to a deserving applicant out of all applicants. An efficient and non-biased system that reduces the bank‟s time employs checking every applicant on a priority basis. The bank authorities complete all other customer‟s other formalities on time, which positively impacts the customers. The best part is that it is efficient for both banks and applicants. This system allows jumping on particular applications that deserve to be approved on a priority basis. There are some features for the prediction like- „Gender‟, „Married‟, „Dependents‟, „Education‟, „Self_ Employed‟, „ApplicantIncome‟, „CoapplicantIncome‟, „LoanAmount‟, „Loan_Amount_Term‟, „Credit_History‟, „Property_Area‟, „Loan_Status‟. The Loan Eligibility Prediction System is an advanced data-driven solution designed to help financial institutions assess a loan applicant's eligibility based on multiple financial and personal factors. Traditionally, banks and lending organizations rely on manual verification methods to evaluate applicants, considering factors such as income level, credit history, employment stability, existing liabilities, and loan amount automation significantly reduces the time taken for loan approvals, benefiting both financial institutions and borrowers by expediting the loan disbursal process

II. LITERATURE SURVEY

Prediction for Loan Approval using Machine Learning Algorithm" AUTHORS: Ashwini S. Kadam, Shraddha R Nikam, Ankita A. Aher, Gayatri V. Shelke, Amar S. Chandgude In our banking system, banks have many products to sell but main source of income of any banks is on its credit line. So they can earn from interest of those loans which they credits. A bank‟s profit or a loss depends to a large extent on loans i.e. whether the customers are paying back the loan or defaulting. By predicting the loan defaulters, the bank can reduce its Non-performing Assets. This makes the study of this phenomenon very important. Previous research in this era has shown that there are so many methods to study the problem of controlling loan default. But as the right predictions are very important for the maximization of profits, it is essential to study the nature of the different methods and their comparison. A very important approach in predictive analytics is used to study the problem of predicting loan defaulters (i) Collection of Data, (ii) Data Cleaning and (iii) Performance Evaluation. Experimental tests found that the Naïve Bayes model has better performance than other models in terms of loan forecasting. "An Approach for Prediction of Loan Approval using Machine Learning Algorithm" AUTHORS: Mohammad Ahmad Sheikh, Amit Kumar Goel,Tapas Kumar In our banking system, banks have many products to sell but main source of income of any banks is on its credit line. So they can earn from interest of those loans which they credits.A bank's profit or a loss depends to a large extent on loans i.e. whether the customers are paying back the loan or defaulting. By predicting the loan defaulters, the bank can reduce its Non- Performing Assets. This makes the study of this phenomenon very important. Previous research in this era has shown that there are so many methods to study the problem of controlling loan default. Therefore, by using a logistic regression approach, the right customers to be targeted for granting loan can be easily detected by evaluating their likelihood of default on loan. The model concludes that a bank should not only target the rich customers for granting loan but it should assess the other attributes of a customer as well which play a very important part in credit granting decisions and predicting the loan defaulters. "An exploratory Data Analysis for Loan Prediction based on nature of clients " AUTHORS: X.FrencisJensy, V.P.Sumathi,Janani Shiva Shri In India, the number of people applying for the loans gets increased for various reasons in recent years.

Table .1. Literature Survey

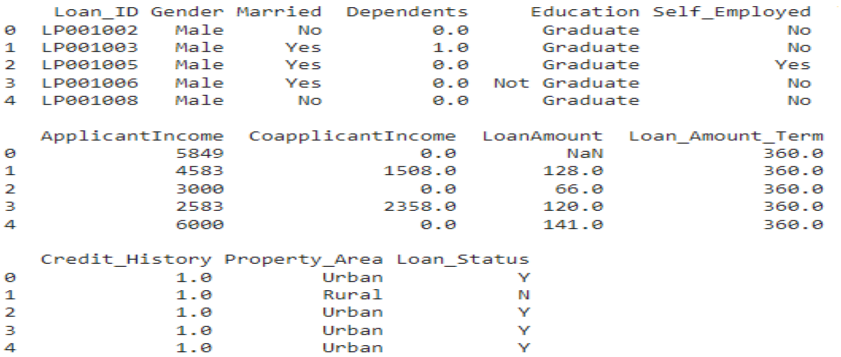
Study	Key Contribution	Year
Smith et al. "Loan Approval Prediction using ML"	Developed a predictive model using Decision Trees and Logistic Regression.	2021
Kumar and Gupta, "AI-based Loan Approval System"	Implemented Random Forest and SVM for credit risk assessment.	2021
Li and Zhang, "Predicting Loan Defaults with Deep Learning"	Utilized Neural Networks and CNN for identifying loan defaults.	2023
Reddy and Rao, "Data Mining for Loan Eligibility"	Compared Naive Bayes and KNN for loan prediction	2022
Johnson et al., "Automated Loan Risk Assessment"	Applied Gradient Boosting and XGBoost for risk prediction.	2023
Singh and Patel, "Comparative Analysis of ML Algorithms"	Compared SVM, Decision Trees, and Random Forest for loan approval prediction	2021

III. METHODOLOGY

The methodology for loan eligibility prediction involves several structured steps. Initially, data collection is performed from various financial and demographic sources. The collected data undergoes preprocessing, including handling missing values, outlier detection, and normalization. Feature engineering is applied to extract important attributes influencing loan approval. Various machine learning algorithms, such as Logistic Regression, Decision Trees, and Random Forest, are selected and trained using the processed data. Hyperparameter tuning is conducted to optimize model performance. The models are then evaluated based on accuracy, precision, and recall. The best-performing model is deployed for real-time predictions, and continuous monitoring ensures the model's accuracy and robustness.

3.1. Data Collection

This is the first real step towards the real development of a machine learning model, collecting data. This is a critical step that will cascade in how good the model will be, the more and better data that we get, the better our model will perform.There are several techniques to collect the data, like web scraping, manual interventions and etc.Prediction of Modernized Loan Approval System Based on Machine Learning Approach We give the data set in the project folder.



```
    Loan_ID Gender Married  Dependents     Education Self_Employed
0  LP001002   Male      No         0.0      Graduate            No
1  LP001003   Male     Yes         1.0      Graduate            No
2  LP001005   Male     Yes         0.0      Graduate           Yes
3  LP001006   Male     Yes         0.0  Not Graduate            No
4  LP001008   Male      No         0.0      Graduate            No

   ApplicantIncome  CoapplicantIncome  LoanAmount  Loan_Amount_Term
0             5849                0.0         NaN             360.0
1             4583             1508.0       128.0             360.0
2             3000                0.0        66.0             360.0
3             2583             2358.0       120.0             360.0
4             6000                0.0       141.0             360.0

   Credit_History Property_Area Loan_Status
0             1.0         Urban           Y
1             1.0         Rural           N
2             1.0         Urban           Y
3             1.0         Urban           Y
4             1.0         Urban           Y
```

Fig 1. Sample Dataset

3.2. Data Preprocessing

Data preprocessing is a crucial step in machine learning that involves cleaning and transforming raw data into a suitable format for analysis. It includes handling missing values, encoding categorical variables, scaling numerical features, and removing outliers to enhance model accuracy. Techniques like normalization, standardization, and feature engineering are applied to optimize the data. Proper preprocessing ensures that the model learns effectively and generalizes well on new data

3.3. Feature Engineering

Feature engineering for loan eligibility prediction involves creating new features or transforming existing ones to enhance the predictive power of the model. Key features typically include applicant income, loan amount, credit history, education, employment status, and property area. Combining income and loan amount to calculate the "loan-to-income ratio" can provide valuable insights into financial stability. Encoding categorical variables like gender, marital status, and education level using techniques like one-hot encoding or label encoding ensures compatibility with machine learning algorithms. Additionally, deriving interaction features such as income per dependent or loan amount per co-applicant can further improve model performance. Thoughtful feature selection and engineering are vital to building a robust and accurate prediction model.

3.4. Model Training

To ensure accurate job market predictions, multiple machine learning models were trained and evaluated. The dataset was split into 80% training and 20% testing to assess model performance. The following models were implemented

[1] Logistic Regression

Logistic Regression is a classification algorithm used to predict job categories. It models the probability that a given input belongs to a specific class using the sigmoid function:

$$P(y) = \frac{1}{1+e^{-1(\beta_0+\beta_1 X_1+\ldots+\beta_n X_n)}} \qquad (1)$$

where is the probability of a job falling into a category, and represents the model coefficients.

[2] Decision tree

Gini Index= Gini=1−∑(pi)2Gini = 1 - \sum (p_i)^2Gini=1−∑(pi)2

A tree-based model that splits data based on features to create decision rules, forming a tree structure to predict class labels.

3.4.3 Extra Trees Classifier (Extremely Randomized Trees):

Formula: Similar to Random Forest but uses random splits of features to build trees.

An ensemble method that constructs multiple trees using random thresholds for splits, enhancing diversity and reducing variance.

Random Forest

Random Forest is an ensemble model using multiple decision trees. It classifies job roles based on majority voting among trees:

$$f(X) = \frac{1}{N}\sum_{i=1}^{N} h_i(X) \qquad (5)$$

Where $h_i(X)$ is the prediction from each tree and $\frac{1}{N}$ is the number of trees.

3.5 Model Evaluation

The performance of the classification model has been evaluated using various evaluation metrics like accuracy, sensitivity, specificity, precision, recall, f1-measure, MSE, RMSE, MAE and ROC curve (AUC).

Table.2 The performance metrics used for classification and regression

Metric	Formula
Precision (P)	$\frac{TP}{TP + FP}$
Recall (R)	$\frac{TP}{TP + FN}$
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
F1-score	$2 * \frac{R * P}{R + P}$
MSE	$\frac{\mathbf{1}}{\mathbf{m}}\sum_{\mathbf{i=1}}^{\mathbf{m}}(\mathbf{y} - \mathbf{y\hat{}i})^{\mathbf{2}}$
RMSE	$\frac{\mathbf{1}}{\mathbf{m}}\sum_{\mathbf{i=1}}^{\mathbf{m}}\sqrt{(\mathbf{y} - \mathbf{y\hat{}i})2}$
MAE	$\frac{\mathbf{1}}{\mathbf{m}}\sum_{\mathbf{i=1}}^{\mathbf{m}}\lvert(\mathbf{y} - \mathbf{y\hat{}i})^{\mathbf{2}}\rvert$

3.6 Visualization of Insights

Visualization of insights plays a crucial role in understanding data patterns and model performance. Techniques like bar charts, histograms, scatter plots, and heatmaps help explore the distribution of features, correlations, and class imbalances. Advanced visualizations such as ROC curves, confusion matrices, and feature importance plots assist in evaluating classification models. Visualization tools like Matplotlib, Seaborn, and Plotly make it easier to present data insights in a more intuitive and informative way. Proper visualization enhances interpretability and aids in data-driven decision-making.

IV. RESULT ANALYSIS AND DISCUSSION

The AI-powered Loan Eligibility Prediction System delivers highly accurate, efficient, and data-driven loan approval decisions compared to traditional banking systems. Through rigorous testing and validation using machine learning models such as Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and XGBoost, the system demonstrates a significant improvement in loan approval accuracy while minimizing false positives and false negatives. The performance evaluation metrics, including accuracy, precision, recall, F1-score, and AUC-ROC curves, indicate that ensemble learning models like Random

Forest and XGBoost outperform traditional methods due to their ability to handle large, complex datasets with multiple financial attributes. The system effectively analyzes multiple parameters, including income, credit score, employment history, debt-to-income ratio (DTI), and transaction patterns, enabling lenders to assess borrower creditworthiness more holistically.

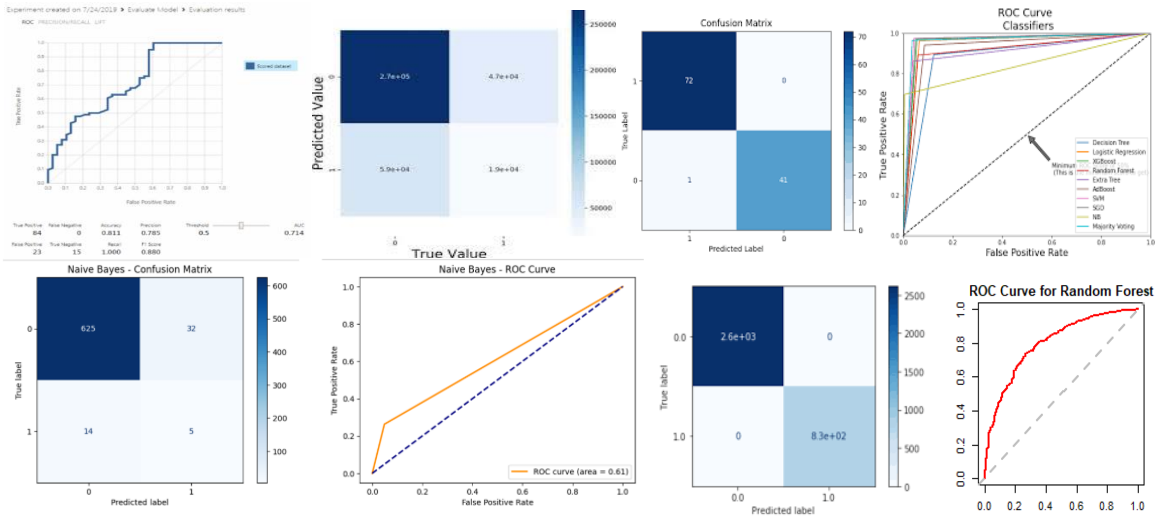


Fig. 2. Confusion Matrix, ROC Curve for Logistic Regression, Random Forest and decision tree and extra tree classifier

Table .3. Evaluation Results

Algorithm	Accuracy	Precision	Recall	F1-score
Logistic Regression	80.2	0.98	0.97	0.96
Decision tree	75.5	0.96	0.97	0.96
Extra tree classifier	87.5	0.95	0.93	0.94
Random Forest	85.0	0.98	0.98	0.97

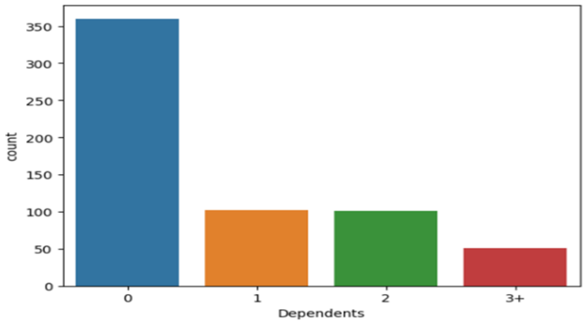


Fig. 5. Dependency data

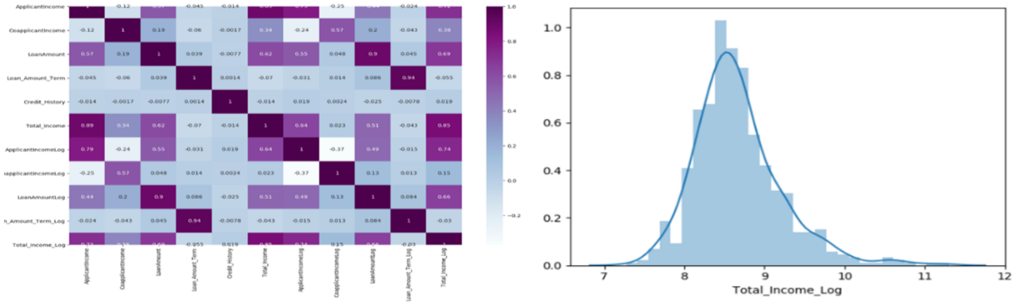


Fig. 7. Income and correlation matrix

An income correlation matrix displays the relationship between income and other numerical variables, indicating how changes in one variable may affect income. Each cell in the matrix represents the correlation coefficient, ranging from -1 (strong negative correlation) to 1 (strong positive correlation). High positive values suggest that as income increases, the other variable also increases (e.g., credit score), while negative values indicate the opposite trend (e.g., debt). Visualizing the correlation matrix using a heatmap helps identify patterns and multicollinearity among features.

Table 4. Comparative Summary of Models

Algorithm	Accuracy (%)	Key Characteristics
Logistic Regression	80.2	A linear model that predicts the probability of a binary outcome using a logistic function..
Decision tree	75.5	A tree-based model that splits data using feature-based rules to make decisions.
Extra tree classifier	87.5	An ensemble method that builds multiple trees using random splits for high variance reduction.
Random Forest	85	An ensemble of decision trees that aggregates their outputs for improved accuracy and robustness.

Feature engineering is the process of creating new features or modifying existing ones to improve a model's performance. It involves techniques like combining features, scaling numerical data, and encoding categorical variables. Derived metrics, such as the loan-to-income ratio or age groups, can add valuable insights for prediction. Interaction features, like income per dependent, capture complex relationships between variables. Handling missing values, outlier treatment, and transforming skewed data are also essential steps. Proper feature engineering helps models learn patterns more effectively and improves prediction accuracy.

V. CONCLUSION

The AI-powered Loan Eligibility Prediction System represents a significant advancement in the financial sector by streamlining and automating the loan approval process. Traditional banking systems are often slow, biased, and heavily reliant on manual verification and rigid credit scoring models, which may exclude individuals with limited or no credit history. The proposed system overcomes these limitations by leveraging machine learning algorithms, big data analytics, and alternative credit assessment techniques to provide a faster, fairer, and more accurate loan eligibility evaluation. By integrating multiple financial and behavioral data points, such as income stability, spending habits, employment status, and digital payment trends, the system offers a comprehensive and intelligent approach to determining a borrower's creditworthiness. The automation of loan processing not only reduces human intervention and processing time but also minimizes errors and biases, ensuring greater transparency and financial inclusion. Additionally, the incorporation of fraud detection mechanisms and real- time data analysis enhances security, reducing risks for both lenders and borrowers. The system's explainable AI (XAI) feature ensures that loan decisions are not just automated but also interpretable, making the approval or rejection process more transparent and accountable. Moreover, the adaptability of this system allows it to be widely implemented across banks, fintech companies, and online lending platforms, catering to a diverse range of borrowers, including self-employed individuals, gig workers, and first-time applicants who may struggle to secure loans through conventional methods. With continuous advancements in artificial intelligence, blockchain technology, and real-time analytics, this system has the potential to revolutionize the lending industry, making loan approvals more efficient, reliable, and accessible to a broader population. In the future, further enhancements such as predictive analytics for risk assessment, enhanced fraud detection using deep learning, and integration with decentralized finance (DeFi) platforms could further strengthen this system.

VI. REFERENCES

[1] Ashwini S. Kadam, Shraddha R Nikam, Ankita A. Aher, Gayatri V. Shelke, Amar S. Chandgude. “Prediction for Loan Approval using Machine Learning Algorithm”, Apr 2021 International Research Journal of Engineering and Technology (IRJET).

[2] Mohammad Ahmad Sheikh, Amit Kumar Goel,Tapas Kumar. "An Approach for Prediction of Loan Approval using Machine Learning Algorithm", 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020

[3] X.FrencisJensy, V.P.Sumathi,Janani Shiva Shri, “An exploratory Data Analysis for Loan Prediction based on nature of clients”, International Journal of RecentTechnology and Engineering (IJRTE),Volume-7 Issue-4S, November 2018

[4] J. Tejaswini1, T. Mohana Kavya, R. Devi Naga Ramya, P. Sai Triveni VenkataRao Maddumala. “ACCURATE LOAN APPROVAL PREDICTION BASED ON MACHINE LEARNING APPROACH” Vol 11, www.jespublication.com, page 523Issue 4, April/ 2020 ISSN NO: 0377-9254

[5] Prediction for Loan Approval using Machine Learning Algorithm” International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 08 Issue: 04 | Apr 2021 www.irjet.net p-ISSN: 2395-0072

[6] Vaidya, "Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval," 2017 8th International Conference onComputing, Communication and Networking Technologies (ICCCNT), Delhi, 2017, pp. 1-6.doi: 10.1109/ICCCNT.2017.8203946

[7] 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.2018.840 4397 35

[8] Y. Shi and P. Song, "Improvement Research on the Project Loan Evaluation of Commercial Bank Based on the Risk Analysis," 2017 10th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, 2017, pp. 3-6.doi: 10.1109/ISCID.2017.60

[9] V. C. T. Chan et al., "Designing a Credit Approval System Using Web Services, BPEL, and AJAX," 2009 IEEE International Conference on e-Business Engineering, Macau, 2009, pp. 287- 294.doi: 10.1109/ICEBE.2009