Comparing Machine Learning Techniques for Loan Approval Prediction

Krishnaraj P[1], Rita S[2], Jitendra Jaiswal[3]

palaniselvaraj3krishnaraj@gmail.com[1], ritasamikannu@gmail.com[2], jitendra.jaiswal@pibm.in[3]

Digital Subscription,The Hindu,Salem,Tamilnadu[1], Department of Statistics, Periyar University, Salem, Tamilnadu[2], Department of Business Analytics, Pune Institute of Business Management Pune, Maharashtra[3]

Abstract. Automated loan eligibility prediction plays a pivotal role in transforming the lending landscape by harnessing the power of data analytics and machine learning. Its importance lies in revolutionizing the loan approval process, making it more efficient, accurate, and accessible. By swiftly analyzing vast amounts of applicant data, automated systems expedite loan processing, reducing waiting times for borrowers and enhancing overall customer experience. The use of advanced algorithms ensures greater accuracy and fairness in evaluating creditworthiness, mitigating the risk of human biases. Moreover, it empowers lenders to assess a wider range of applicants, including those with limited credit history, fostering financial inclusion. Cost savings, improved fraud detection, and personalized loan offerings are additional benefits, making automated loan eligibility prediction an indispensable tool for modern lending institutions seeking to optimize operations, manage risk prudently, and cater to diverse customer needs effectively. In the last decade, machine learning techniques have obtained significant attention in automating and enhancing loan approval processes. This research paper focuses to provide a comparative analysis of different machine learning techniques applied in multiple areas such as loan approval, exploring their strengths, limitations, and performance metrics. In this paper, we applied three popular machine learning models, namely Logistic Regression, Decision Tree and its extension as Random Forest and to predict and classify the target variable. The primary objective was to perform a comparative analysis of these models and identify the most suitable one for the task at hand. After thorough analysis and comparison of the models' outcomes, we found that the Logistic Regression model demonstrated a slightly superior performance in comparison to the other models.

Keywords: *Loan approval, automated loan, Machine Learning, Logistic Regression*

1 Introduction

Loan approval is the process by which a financial institution or lender assesses an individual's creditworthiness and determines whether to grant them a loan. The approval process involves several factors and considerations, including the applicant's credit history, income, employment

IACIDS 2023, November 23-25, Lavasa, India

DOI 10.4108/eai.23-11-2023.2343174

stability, debt-to-income ratio, and the type of loan being applied for. Here are some key aspects that lenders typically consider during the loan approval process:

In the loan approval process, several key factors are carefully evaluated by lenders to assess the creditworthiness and risk associated with the applicant. Firstly, the applicant's Credit History and Credit Score are thoroughly reviewed, providing insights into their past repayment behavior. A higher credit score signifies a strong credit history, thus increasing the likelihood of loan approval. Secondly, the applicant's Income and Employment Stability are crucial considerations. Lenders analyze the income level and stability of employment to determine whether the applicant possesses the financial capacity to repay the loan consistently. A steady and sufficient income stream significantly improves the chances of loan approval. The Debt-to-Income Ratio is another vital aspect considered by lenders. This ratio reveals the percentage of the applicant's monthly income allocated towards debt payments. A lower debt-to-income ratio indicates a performing financial position and enhances the probability of loan approval. Furthermore, the Loan Amount and Purpose are factored into the approval process. Lenders have specific criteria for various types of loans, such as business loans, home loans, and personal loans, etc. The purpose of the loan and the requested amount influence the lender's decision-making. Applicants are also required to submit various Documentation and Verification materials, including identification proof, income statements, bank records, employment proof, and other relevant paperwork. Lenders meticulously verify these documents to ensure their authenticity and accuracy. In certain cases, Collateral and Security may be necessary, particularly for secured loans. Lenders may demand collateral to secure the loan amount, and the value and quality of the provided collateral play a significant role in determining loan approval. Finally, specific Loan Terms and Conditions are established by lenders, covering aspects such as interest rates, repayment tenure, and any additional fees or charges. These terms are communicated to the borrower upon approval, providing a comprehensive understanding of the loan agreement. Considering these multifaceted factors, lenders can make informed decisions about loan approvals, mitigating risks and ensuring responsible lending practices.

2 Literature Survey

In paper [1], the authors have focused on using machine learning algorithms to predict loan approval. They proposed an approach or methodology that utilizes machine learning techniques to analyze and predict loan approval outcomes. It involves the selection and implementation of various machine learning models, data preprocessing techniques, feature engineering, and performance evaluation measures. On the original data, the accuracy at the best-case is 0.81. It has been observed that applicants with high incomes and modest loan requests are quite feasible to be approved, which makes sense, and are also more likely to rectify their debts.

In this research paper, we will use a dataset which is acquired from the website ‘Kaggle.com’ for research and analysis purposes. The lending institution operates in urban, semi-urban, and rural areas, and receives loan applications from customers seeking financial assistance. Currently, the loan eligibility process requires manual validation, which is time-consuming and prone to errors. To enhance efficiency and accuracy, the firms aim to automate the loan eligibility assessment in real-time using customer details provided while online submission of the application form. The key variables considered during this process include Education, Gender, Marital Status, Income, # of Dependents, Credit History, the Loan Amount, and others.

The objective of this research paper is to develop an automated loan eligibility prediction model that efficiently identifies customer segments eligible for a loan, enabling the finance company to target them more effectively. The study also investigates the impact of various factors on loan approval decisions and highlights the challenges and potential solutions in implementing machine learning models in real-world scenarios. To achieve this, the research will follow these major steps: Data Pre-processing: The initial step involves data cleaning and preparation to ensure the quality and reliability of the dataset for subsequent analysis.

Data preprocessing takes up the majority of the time and effort, followed by performing EDA (Exploratory Data Analysis), feature engineering, and eventually model development [1].

The EDA on the dataset is subjected to the following processes: data normalization, missing-value treatment, selecting the key columns using the filtering technique, generating additional columns, identifying the output variables, and graphical representation of the data. Python is used for quick and effective data processing. [2].

Classifier Selection: Three different machine learning classifiers, namely logistic regression, decision tree, and random forest classifiers, will be employed to forecast the loan eligibility based on the eligibility of customers.

Comparative Analysis: The research will focus on comparing the accuracy and performance of the three classifiers to understand their strengths and weaknesses. Model Selection: After comparing the classifiers, the research will identify and select the most suitable machine learning model for predicting loan approvals, considering both predictive accuracy and computational efficiency.

3 Data Set

The loan applicant's data are loaded into Python using Panda's library. The panda's data frame contains 614 entries and 13 columns. The columns in the Data Frame include 'Application ID' (unique identifier for each loan applicant), 'Gender', 'Dependents', 'Married', 'Self Employed', 'Education', 'Applicant Income' (income of the applicant), 'Coapplicant Income' (income of the co-applicant, if any), 'Loan Amount' (amount of the loan applied for), 'Loan Amount Term' (term or duration of the loan), 'Credit History' (credit history of the applicant), 'Property Area' (area where the property is located), and 'Loan Status' (status of the loan application).

Variables	Description
Application ID	Unique Loan ID
Gender	Female / Male
Married	Applicant's Marital status(Y/N)
Dependents	# of dependents
Education	Highest education of the applicant (Undergraduate/ Graduate)
Self-Employed	Self-employed or not (Y/N)
Applicant Income	Applicant's annual income
Co-applicant Income	Annual income of the Co-applicant
Loan-Amount	Total amount of the loan (in thousands)
Loan Amount Tenure	Tenure of the loan (in months)

Credit History	Following credit-history guidelines (Y/ N)
Property Area	Urban / Rural / Semi-Urban
Loan Status	Whether the loan has been approved (Y/N)

The summary provides us information on the non-null count and data types for each column. Some columns have missing data represented by 'NaN'. The dataset encompasses comprehensive information regarding loan applications, with each row representing a distinct loan applicant. It comprises a set of pertinent variables, including application ID, serving as a unique identifier for each loan application. The variable Gender denotes the gender of the applicant, categorized as "Female" or "Male". Additionally, the variable Married indicates whether the applicant is married or not, with values "Y" indicating "Yes" and "N" for "No." The variable Dependents represents the number of dependents the applicant has, with values ranging from 0 to 3 or more, and may also be missing if no dependents are associated with the application.

Education indicates the applicant's educational qualification, categorized as "Graduate" or "Undergraduate". Moreover, Self-employed is a binary variable, taking "Y" for "Yes" and "N" for "No," signifying if the applicant is self-employed or not. The dataset also includes essential financial details, such as the income of the Applicant and Co-applicant, and the Loan-Amount, measured in thousands. Loan-Amount Term specifies the loan duration in months, reflecting the time given for repayment.

Credit History, observed as a binary variable, showcases the credit history status of the applicant, where "1" indicating a satisfactory history and "0" representing an unsatisfactory one. Property Area delineates the location of the property for which the loan is sought, with three categories: "Urban," "Semi Urban," and "Rural." Finally, Loan-Status serves as the target-variable, denoting whether the loan application was approved ("Y" for "Yes") or not ("N" for "No"). The dataset holds great potential for conducting diverse analyses and building machine learning models to gain insights into the factors influencing loan approval and predict the likelihood of approval based on applicant characteristics.

4 Proposed Machine Learning Model

*A. **Decision Tree***

The decision tree technique is a visual representation of a classification algorithm or decision-making process that employs a tree-like structure. Its structure resembles a flowchart, with each internal node standing in for a decision, each branch for a consequence of that option, and each leaf node for the ultimate choice or classification result.

Although it is more extensively utilised as a classification tool, decision trees are frequently employed in machine learning and data mining for both classification and regression applications. It can deal with both the categorical and continuous variables but mostly applicable for categorical data [5].

*B. **Random Forest***

Several decision trees are combined to generate a more reliable and accurate forecasting model as part of an ensemble learning technique called Random Forest. In machine learning, it is a

well-liked approach for both regression and classification applications. A collection of decision trees are referred together as a "forest". It generates random forests and then probes through them for answers. To minimise the case of overfitting difficulties, random forest assesses each tree separately instead of just one. The more trees there are, the more accurate problem-solving becomes [5].

*C. **Logistic regression***

Logistic regression in machine learning is a statistical model applied for binary classification perspectives, where the outcome variable is categorical with two classes (e.g., yes/no, true/false, 0/1) [6].

*D. **Overfitting***

In supervised machine learning, overfitting is a general problem that cannot be totally avoided. This is due to the limitations of the algorithms, which are too complex and need too many parameters, or the limitations of the training data, which may be little or include a lot of noise [7]. It could also be identified in the case when while training, we are achieving very high accuracy but while testing it shows very poor accuracy.

In other words, rather of identifying underlying patterns and correlations, the model deals with the noise and quirks of the training data. Poor performance and incorrect forecasts might result from overfitting. However, there are a number of methods for reducing the overfitting issue.

*E. **Hyperparameter tuning***

The process of determining the ideal settings for a machine learning model where values are determined by hit and trial aspects rather than mathematical approaches, is known as hyperparameter tuning. Hyperparameters are configuration options that are chosen by the user before to training the model rather than ones that are learnt from the data. The effectiveness and generalizability of the model can be dramatically impacted by tuning these hyperparameters. Learning rate, regularisation strength, the number of hidden layers, and batch size are some frequent hyperparameters [8 -10].

5 Exploratory Data Analysis

Based on the exploratory analysis of the dataset using python libraries, it is evident that loan approval status shows a positive trend, with approximately two-thirds of applicants being granted loans. In terms of gender, there are significantly more male applicants, outnumbering females by around three times. The marital status of the applicants indicates that nearly two-thirds of the population in the dataset is married, and married individuals are more likely to be granted loans.

Regarding dependents, most applicants have zero dependents, and this category is also more likely to be accepted for a loan. Moreover, education plays a significant role in loan approval, with about five-sixths of the population being graduates, and they have a higher proportion of loan approvals compared to non-graduates. Regarding employment status, approximately five-sixths of the population is not self-employed. Moving on to property area, a larger number of applicants are from semi-urban areas, and they are also more likely to be granted loans compared to applicants from other regions [9].

Applicants with a credit history are found to be far more likely to have their loan applications accepted, indicating the importance of a good credit history. Lastly, it is observed that most of

the loans taken are for a longer duration of 360 months (30 years). This information from EDA provides valuable insights into the loan approval trends and various factors influencing the decision-making process.

6 Model Comparison

The decision tree-based model has been developed using 'DecisionTreeClassifier' from the library 'Scikit-learn' for binary classification. After training the model on the provided training data (X_train) and labels (y_train), predictions are generated and accuracy as well as F1 score are computed for the training set (). To gauge its performance on unseen data, 5-fold cross-validation is applied, offering insights into the model's generalization capabilities beyond the training data and the final analysis can be observed in following Table – 1.

Table – 1: Accuracy and F1 Score

Metric	Training Dataset	Test Dataset
Accuracy	1.0	0.7047
F1 Score	1.0	0.6604

Overfitting Concerns: Based on the above metrics, it becomes evident that the Training Accuracy surpasses the Test Accuracy using the default settings of the Decision Tree classifier. This discrepancy suggests that the model is likely overfitting the training data. To address this issue, we will undertake Hyper-parameter tuning to fine-tune the model and assess if it alleviates the overfitting problem. First, we tried tuning 'Max_Depth' of tree.

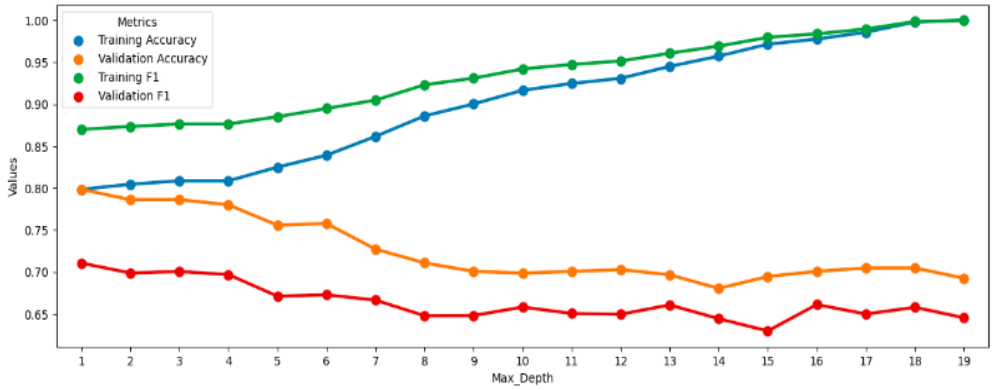

Figure – 1: Optimum Test Accuracy

From above graph in Figure - 1, we could observe that considering 'Max_Depth' = 3 will yield optimum Test accuracy and F1 score Optimum Test Accuracy ~ 0.805; Optimum F1 Score: ~0.7

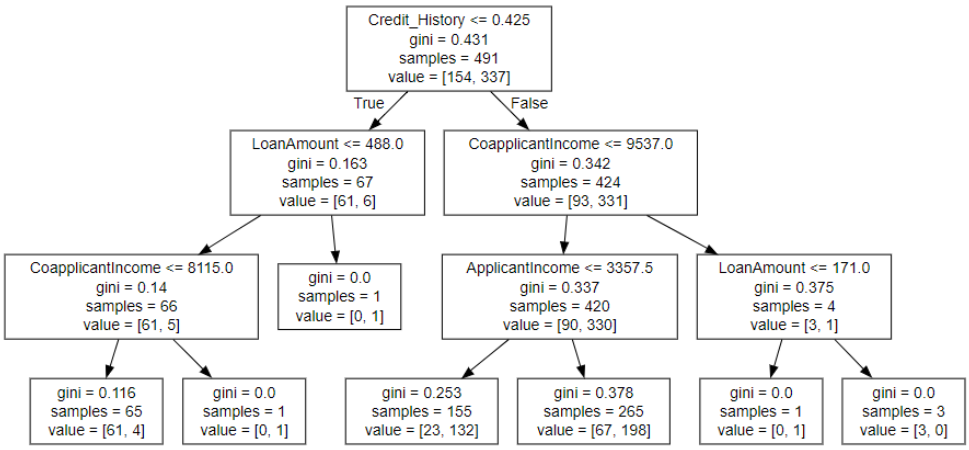

Figure -2: Decision Tree

From above decision tree in Figure - 2, we could see that some of the leafs have less than 5 samples hence the classifier might overfit. We can sweep hyper-parameter 'min_samples_leaf' to further improve test accuracy by keeping max_depth to 3.

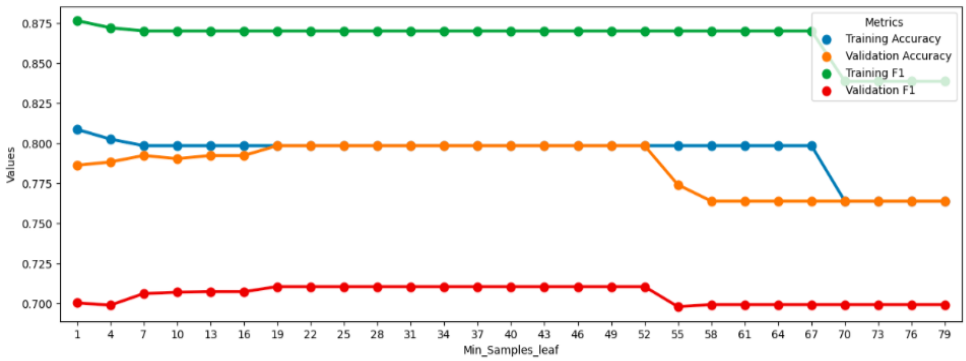

Figure-3: Training and Validation Plot

From above plot in Figure - 3, we will choose Min_Samples_leaf to 35 to improve test accuracy.

Let's use this Decision Tree classifier on unseen test data and evaluate Test Accuracy, F1 Score and Confusion Matrix

Test Accuracy: 0.8536 & Test F1 Score: 0.9032

Mis-classifications: It can be seen that majority of the misclassifications are happening because of Loan Reject applicants being classified as Accept.

Let's look into Random Forest Classifier if it can reduce misclassifications in the Table -2.

Table- 2: Random Forest Classifier

Metric	Training Data Set	Validation Mean	Test Mean
Accuracy	0.7983	0.7983	0.8536
F1 Score	0.8699	0.7028	0.9032

Random Forest gives same results as Decision Tree Classifier. Finally, we will try Logistic Regression Model by sweeping threshold values.

Logistic Regression does slightly better than Decision Tree approach and Random Forest. Depending on the above Test/Train curves, we can keep threshold to 0.4.

Now Finally let's look at Logistic Regression Confusion Matrix Test Accuracy: 0.8617 & Test F1 Score: 0.9081 in the following Figure – 4.

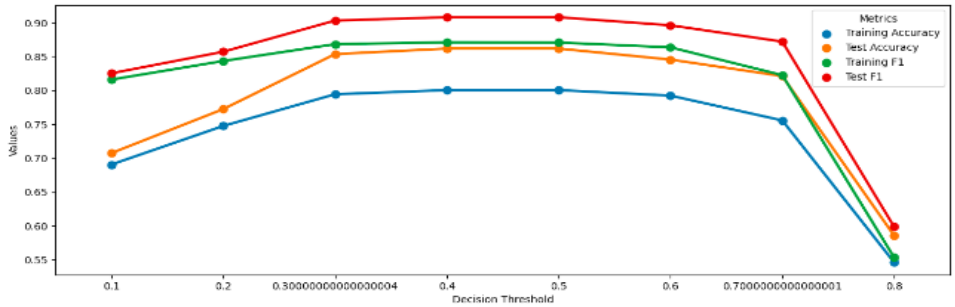

Figure – 4: Logistic Regression

7 Conclusion

With the creation of confusion matrix for the three techniques as Decision Tree, Random Forest Classifier and Logistic Regression, we found that they are very similar to each other as per their performances. In this analysis, we did extensive analysis of input data and were able to achieve Test Accuracy of 86 %.

The data in this analysis suggests that logistic regression, decision tree, and random forest classifier exhibit similar patterns in terms of their confusion matrices. A confusion matrix enables to achieve a detailed breakdown of the predictions through different considered models, including true positives, false positives, true negatives, and false negatives. It allows for an empirical evaluation of the model's accomplishments.

An extensive analysis of the data was conducted, likely involving preprocessing steps, feature engineering, and data exploration. This analysis aims to ensure the data is appropriately prepared for the models. Lastly, in our comparison the logistic regression technique achieved a test

accuracy of 86%. Test accuracy measures the proportion of correctly predicted instances in the test dataset. An accuracy of 86% indicates that the logistic regression model performed reasonably well in making accurate predictions on unseen data.

References

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

[2] J. X. and V. P. Sumathi, J. Sri, "An exploratory data analysis for loan prediction based on the nature of the clients," International Journal of Recent Technology and Engineering, vol. 7, pp. 176-179, 2019.

[3] A. Hamid and T. Ahmed, "Developing Prediction Model of Loan Risk in Banks Using Data Mining," Machine Learning and Applications: An International Journal, vol. 3, pp. 1-9, 2016. doi: 10.5121/mlaij.2016.3101.

[4] P. Supriya, M. Pavani, N. Saisushma, N. Vimala Kumari, and K. Vikash, "Loan Prediction by using Machine Learning Models," International Journal of Engineering and Techniques, vol. 5, issue 2, pp. 144-148, Mar-Apr 2019.

[5] N. Pandey, R. Gupta, S. Uniyal, and V. Kumar, "Loan Approval Prediction using Machine Learning Algorithms Approach," International Journal of Innovative Research in Technology (IJIRT), vol. 8, issue 1, pp. 898-902, June 2021, ISSN: 2349-6002.

[6] A. Jadhav, P. Wankhande, P. Balure, A. Pawar, and P. P. Halkarnikar, "Loan Approval Prediction using Machine Learning," International Research Journal of Modernization in Engineering Technology and Science, vol. 05, issue 05, pp. 4956-4957, May 2023, e-ISSN: 2582-5208.

[7] X. Ying, "An Overview of Overfitting and its Solutions," Journal of Physics: Conference Series, vol. 1168, pp 1-6, 022022, 2019. doi: 10.1088/1742-6596/1168/2/022022.

[8] L. Yang and A. Shami, "On Hyperparameter Optimization of Machine Learning Algorithms: Theory and Practice," Preprint, July 2020.

[9] P. Probst, A.-L. Boulesteix, and B. Bischl, "Tunability: Importance of Hyperparameters of Machine Learning Algorithms," Journal of Machine Learning Research, vol. 20, pp. 1-32, 2019.

[10] K. E. Hoque and H. Aljamaan, "Impact of Hyperparameter Tuning on Machine Learning Models in Stock Price Forecasting," in IEEE Access, vol. 9, pp. 163815-163830, 2021, doi: 10.1109/ACCESS.2021.3134138.