IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Loan Eligibility Prediction Using Machine Learning

Shaik.Bhanu[1], Dr. Y. A. Siva prasad [2]

[1]Research Scholar, Department of CSE, Sri Venkateswara College of Engineering, Tirupati
[2]Associate Professor, Department of CSE, Sri Venkateswara College of Engineering, Tirupati

Abstract:
With technological advancements and the expansion of businesses, the demand for loans has increased significantly, both for personal and business purposes. Due to limited assets, banks cannot grant loans to every applicant. Identifying the right candidates for loans is a complex and time-consuming process. Banks aim to deliver loans to individuals who can repay them on time and provide maximum profit. Thus, there is a need for a system that can analyse and streamline this process, saving time and resources. This paper aims to develop a more accurate loan prediction model using machine learning to reduce the risk involved in selecting appropriate loan applicants. By mining previous loan records and using bank loan rules, we will train a machine learning model to predict loan eligibility. We will use the sklearn library for our model and the train-test-split method to split the dataset. Various models, including Logistic Regression, Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM), will be employed to achieve accurate results. Our experiments indicate that the Random Forest classifier provides the best accuracy.

Keywords: Logistic Regression, Decision Tree, Random Forest, Support Vector Machine.

I. Introduction

Granting loans is a critical revenue stream for banks, as it is a primary source of profit. Therefore, the precise allocation of loans is paramount. Incorrectly approving loans can lead to financial instability and diminished profits. Banks aim to invest their resources in reliable borrowers who will generate substantial interest returns. Before approving a loan, banks meticulously evaluate various factors such as the applicant's ability to repay, their financial stability, and the purpose of the loan. This manual evaluation process is time-consuming for both the bank and the customer [1].

A machine learning system can streamline this process by accurately predicting the suitability of a loan applicant. Automation ensures the integrity of the results, prevents tampering, and accelerates the approval process, significantly reducing the time required for decision-making and paperwork. This benefits both bank employees and customers by facilitating quicker loan processing [2].

The system's predictions are based on multiple factors, although in some cases, a single strong indicator might suffice for loan approval. Historical loan data will be mined to train the model. Data mining, which involves

extracting valuable information from large datasets, includes techniques like classification, clustering, and association, with classification being the primary focus here [3].

We will use several classification methods, including Decision Trees, Neural Networks, Support Vector Machines (SVM), Logistic Regression, and k-Nearest Neighbours (k-NN). Machine learning models require two datasets: a training dataset to develop the model and a test dataset to evaluate its accuracy. These datasets are derived from a single, larger dataset. We will utilize the train_test_split function from the model selection module to divide the data accordingly [4].

Decision Trees solve problems by representing each class label as a leaf node and attributes as internal nodes. Any boolean function on discrete attributes can be depicted using a decision tree. Random Forest, a supervised learning method, can address both classification and regression issues. It leverages ensemble learning, which combines multiple classifiers to solve complex problems and enhance performance. Random Forest generates a set of decision trees from random subsets of the training data and aggregates their votes for the final prediction.

Logistic Regression, another supervised classification algorithm, predicts categorical dependent variables, resulting in discrete outcomes. It is analogous to Linear Regression but tailored for classification tasks.

The goal of this project is to develop a simplified loan prediction model. Implemented in Python using Google Colab, we employ the pandas library for data manipulation and the seaborn library for data visualization. This model aims to streamline the loan approval process, ensuring quick and reliable predictions.

II Literature Review

We reviewed several research studies focused on loan prediction models. One study by G. Arujothi and Dr. C. Senthamarai highlighted the complexity of predicting credit defaulters and emphasized the necessity of a machine learning approach to enhance efficiency and conserve resources. They suggested that using R software in conjunction with Min-Max normalization and the K-Nearest Neighbour (K-NN) classifier could significantly improve the accuracy of loan approval predictions [6]. Another study by Aboobyda and Tagir from the University of Khartoum, Sudan, explored the effectiveness of j48, Bayes Net, and Naive Bayes algorithms for loan prediction [7],[8],[9]. Their findings indicated that the j48 algorithm was the most effective for precise credit approval predictions, utilizing the Weka application for model implementation and evaluation.

Further research by Kumar Arun, Garg Ishan, and Kaur Sanmeet examined the performance of various classification models, including Random Forest, SVM, LM, Nnet, and ADB, in predicting loan approvals [10]. Glorfeld and Hardgrave developed a highly efficient model using neural networks to assess the creditworthiness of loan applicants, achieving a correct prediction rate for 75% of applicants. In their paper, Andy Liaw and Matthew Wiener discussed the application of Random Forest for classification and regression tasks. Additionally, Stephan Dreiseitl and Lucila Ohno-Machado explored the use of artificial neural networks and logistic regression models, highlighting the effectiveness of logistic regression in building predictive systems [11].

These studies collectively suggest that machine learning models, including those using neural networks, logistic regression, and ensemble methods like Random Forest, are highly effective for predicting loan approvals in the current financial landscape [12].

III Proposed Model

Loan Prediction Model Using Machine Learning Techniques

The Figure 1 illustrates a comprehensive workflow for developing a machine learning model to predict loan eligibility. The process begins with raw loan data and involves several key steps:

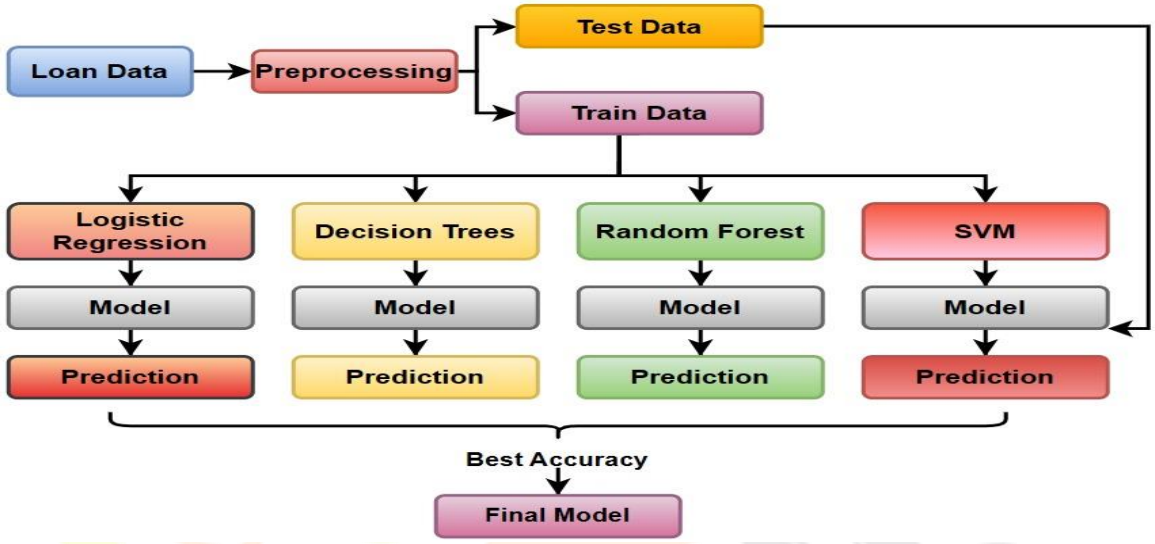

Figure 1 Proposed Model of the work

Data Preprocessing: The raw loan data undergoes preprocessing to clean and transform it into a suitable format for analysis. This step includes handling missing values, normalizing data, and encoding categorical variables. Preprocessing ensures that the data is consistent and ready for model training.

Data Splitting: After preprocessing, the data is divided into two subsets: training data and test data. The training data is used to build the model, while the test data is reserved for evaluating the model's performance.

Model Training: Four different machine learning algorithms are employed to train models on the training data:

Logistic Regression: This algorithm is used to model the probability of a binary outcome, in this case, whether a loan should be approved or not. It is effective for understanding the relationship between the dependent variable and one or more independent variables.

Decision Trees: This model splits the data into branches to make decisions based on feature values, providing clear decision rules for loan approval.

Random Forest: An ensemble method that combines multiple decision trees to improve prediction accuracy and reduce overfitting. It aggregates the votes from different trees to make the final prediction.

Support Vector Machine (SVM): This algorithm finds the optimal hyperplane that separates the data into classes, making it suitable for binary classification tasks like loan approval.

Prediction: Each trained model is then used to make predictions on the test data. The performance of each model is evaluated to determine its accuracy and effectiveness.

Model Selection: The model that achieves the highest accuracy on the test data is selected as the final model. This model represents the best balance between complexity and performance and is used for predicting loan eligibility in real-world scenarios.

Final Model: The selected model is deployed for making future predictions on loan applications. It helps banks make data-driven decisions, improving efficiency and reducing the risk of loan defaults.

IV Results and Discussion

In our study, we implemented and evaluated four machine learning algorithms to predict loan eligibility: Random Forest, Support Vector Machine (SVM), Logistic Regression, and Decision Tree. Each algorithm was trained using the pre-processed loan data, and their performance was assessed based on the accuracy of their predictions on the test data. The results are summarized as follows:

Random Forest: Achieved the highest accuracy of 94% after fine-tuning the hyperparameters. The robustness of the Random Forest algorithm, which leverages multiple decision trees and the concept of ensemble learning, contributed significantly to its superior performance. By aggregating the predictions of various trees, Random Forest minimizes overfitting and improves generalization to unseen data.

Support Vector Machine (SVM): Produced an accuracy of 86.76%. While SVM is effective for high-dimensional spaces and provides a clear margin of separation between classes, it requires careful tuning of parameters such as the kernel type and regularization term. Despite these efforts, SVM did not outperform Random Forest in our experiments.

Logistic Regression: Recorded an accuracy of 84.44%. This algorithm, known for its simplicity and interpretability, performed reasonably well but was less accurate than Random Forest and SVM. Logistic Regression assumes a linear relationship between the independent variables and the log-odds of the dependent variable, which might not fully capture the complexities of the loan approval data.

Decision Tree: Attained an accuracy of 86.67%. Decision Trees are intuitive and easy to visualize but are prone to overfitting, especially with complex datasets. Although the accuracy was comparable to SVM, it fell short of Random Forest, which combines multiple decision trees to enhance performance.

Discussion

The results indicate that the Random Forest algorithm is the most effective for predicting loan eligibility, with an accuracy of 94%. This superior performance can be attributed to several factors:

Ensemble Learning: Random Forest aggregates the results of multiple decision trees, reducing the risk of overfitting and improving the model's ability to generalize from the training data to the test data.

Hyperparameter Tuning: Fine-tuning the hyperparameters, such as the number of trees in the forest and the maximum depth of each tree, played a crucial role in optimizing the Random Forest model's performance.

Robustness to Noise and Overfitting: By averaging multiple trees, Random Forest is less sensitive to noise in the training data and can handle the complexity and variability inherent in the loan dataset more effectively than single models like Decision Trees or Logistic Regression.

While SVM and Decision Trees also demonstrated solid performance, they were less effective than Random Forest in this context. SVM's requirement for parameter tuning and Decision Trees' tendency to overfit likely contributed to their lower accuracy. Logistic Regression, although straightforward and interpretable, may not fully capture the nonlinear relationships in the data, leading to its comparatively lower accuracy. Figure 2 shows accuracy of the decision tree model. Figure 3 depicts the accuracy of the random forest classifier.

Figure 2. Accuracy of the Decision Tree Model

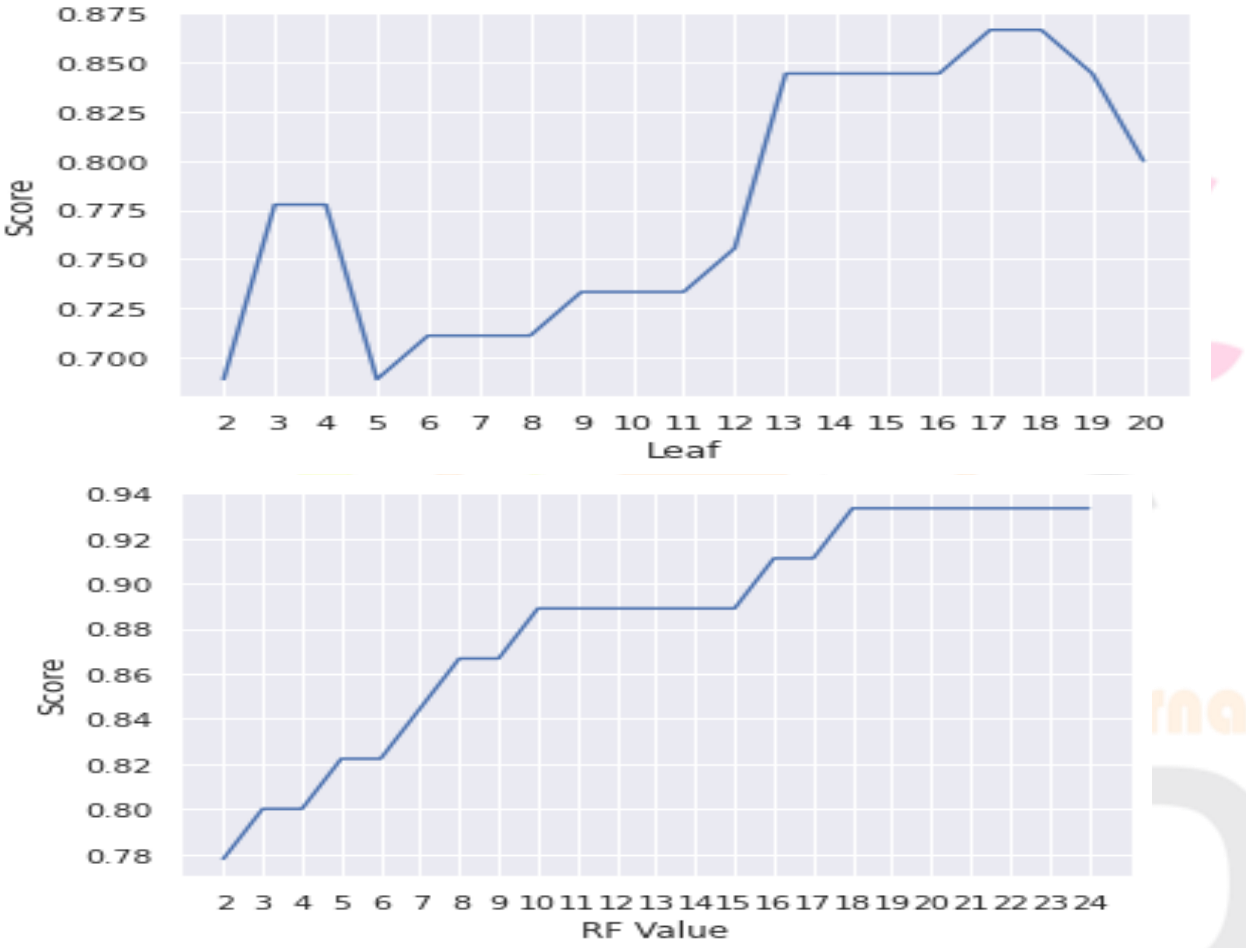

Figure 3. Accuracy of the Random Forest Classifier

In conclusion, the Random Forest model stands out as the best performer for predicting loan eligibility in our study. Its high accuracy and robustness make it a valuable tool for banks to make data-driven decisions, reducing the risk of loan defaults and optimizing resource allocation. Future work could explore further enhancements, such as incorporating additional features or leveraging other ensemble methods, to continue improving predictive accuracy.

.

V. Sample Code on training Models:

Logistic Regression

```
LRclassifier = LogisticRegression(solver='saga', max_iter=500, random_state=1)
LRclassifier.fit(X_train, y_train)

y_pred = LRclassifier.predict(X_test)

print(classification_report(y_test, y_pred))
print(confusion_matrix(y_test, y_pred))

from sklearn.metrics import accuracy_score
LRAcc = accuracy_score(y_pred,y_test)
print('LR accuracy: {:.2f}%'.format(LRAcc*100))
```

```
SVCclassifier = SVC(kernel='rbf', max_iter=500)
SVCclassifier.fit(X_train, y_train)

y_pred = SVCclassifier.predict(X_test)

print(classification_report(y_test, y_pred))
print(confusion_matrix(y_test, y_pred))

from sklearn.metrics import accuracy_score
SVCAcc = accuracy_score(y_pred,y_test)
print('SVC accuracy: {:.2f}%'.format(SVCAcc*100))
```

VI. Conclusion:

In this study, we developed and evaluated a machine learning-based loan prediction model using four different algorithms: Random Forest, Support Vector Machine (SVM), Logistic Regression, and Decision Tree. Our goal was to identify the most accurate model for predicting loan eligibility, thereby streamlining the loan approval process and improving decision-making efficiency for banks.

The results of our experiments demonstrated that the Random Forest algorithm outperformed the other models, achieving an impressive accuracy of 94% after fine-tuning the hyperparameters. This high performance is attributed to the ensemble learning approach of Random Forest, which mitigates overfitting and enhances generalization by combining multiple decision trees. While SVM and Decision Trees also showed solid performance with accuracies of 86.76% and 86.67%, respectively, they were less effective compared to Random Forest. Logistic Regression, with an accuracy of 84.44%, was straightforward and interpretable but fell short in capturing the complex, nonlinear relationships within the loan data.

VII References

[1] Arun, K., Ishan, G., & Sanmeet, K. (2019). Loan approval prediction based on machine learning approach. IOSR J. Comput. Eng, 18(3), 18-21.

[2] Yasaswini, P., Aruna Sri, P., Pratyusha, Y., Reddy, P. S., & Kumari, S. Analysis and Forecasting of bank loan approval data using machine learning algorithms.

[3] Kadam, A. S., Nikam, S. R., Aher, A. A., Shelke, G. V., & Chandgude, A. S. (2021). Prediction for loan approval using machine learning algorithm. International Research Journal of Engineering and Technology (IRJET), 8(04).

[4] Mella, N. V. V. P., & Sai, R. R. LOAN APPROVAL PREDICTION [5] Tejaswini, J., Kavya, T. M., Ramya, R. D. N., Triveni, P. S., & Maddumala, V. R. (2020). Accurate loan approval prediction based on machine learning approach. Journal of Engineering Science, 11(4), 532-532.

[5] Sarkar, A. (2021). Machine learning techniques for recognizing the loan eligibility. International Research Journal of Modernization in Engineering Technology and Science, 3(12), 1135-1142.

[6] G. Arutjothi, Dr C. Senthamarai, "Prediction of Loan Status in Commercial Bank using Machine Learning Classifier," Proceedings of the International Conference on Intelligent Sustainable Systems, (2023).

[7] P. Supriya, M. Pavani, N. Saisushma, N. Kumari and K. Vikas, "Loan Prediction by using Machine Learning Models," International Journal of Engineering and Techniques, (2019).[8] R. Salvi, R. Ghule, T. Sanadi, M. Bhajibhakare, "HOME LOAN DATA ANALYSIS AND VISUALIZATION," International Journal of Creative Research Thoughts (IJCRT), (2021).
[9] B. Srinivasan, N. Gnanasambandam, S. Zhao, R. Minhas, "Domain-specific adaptation of a partial least squares regression model for loan defaults prediction," 11th IEEE International Conference on Data Mining Workshops, (2021).
[10] M. V. Reddy, Dr B. Kavitha, "Neural Networks for Prediction of Loan Default Using Attribute Relevance Analysis," International Conference on Signal Acquisition and Processing, (2010).
[11] G. Chornous, I. Nikolskyi, "Business-Oriented Feature Selection for Hybrid Classification Model of Credit Scoring," IEEE Second International Conference on Data Stream Mining & Processing August (2022)
[12] M. Sheikh, A. Goel, T. Kumar, "An Approach for Prediction of Loan Approval using Machine Learning Algorithm," International Conference on Electronics and Sustainable Communication Systems (ICESC), (2020).