International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 6, 2025
https://theaspd.com/index.php

An Enhanced Ensemble-Based Framework for Loan Approval Prediction Using Machine Learning

Lakshmi S R[1], Gajendra S[2], Mohanraju V S[3]
[1]Department of Computer Science Engineering in Data Science, Dayananda Sagar Academy of Technology and Management, Bangalore, Karnataka, India
[2]Departrment of Data Science & Engineering, Birla Institute of Technology and Science, Pilani, Rajasthan, India
[3]Departmenr of Computer Science Engineering in Data Science, Mangalayatan University, Aligarh, Uttar Pradesh, India
[1]lsr382293@gmail.com, [2]sgajendragowda319@gmail.com, [3]mohanrudra142@gmail.com

Abstract *-Loan approval prediction is an important application in the banking and finance industry to automate and make easy the loan application decision process. In this work, we introduce the use of machine learning algorithms, namely Random Forest and*
Decision Tree classifiers, to predict if a loan application should be approved or not based on applicant information. Features used in the dataset are income, education, employment, CIBIL score, loan amount, and dependents. Preprocessing methods like data cleaning, feature selection, and label encoding for categorical variables were used. Train-test split method with stratification was used for training and testing the models in order to keep the classes balanced. Model performances were checked based on accuracy, classification report, and confusion matrix. Our findings indicate that Random Forest classifier has a better performance than the Decision Tree model as far as accuracy and generalizability are concerned. The Random Forest model was trained on processed features and pickled using the pickle library to deploy in a real-world system for loan approvals. These results prove that techniques such as Random Forest can efficiently improve the precision and speed of loan approval mechanisms while reducing the level of human bias.
Keywords– *Loan, Prediction, Random Forest, Decision Tree, Preprocessing, Classification, Accuracy, Deployment.*

1. INTRODUCTION

Loan approval prediction is a crucial but challenging financial analytics issue, spurred by the necessity of precise, fact-based loan decisions in high-risk lending environments. The conventional method of loan eligibility determination is often laborious scrutiny of loan applicants' financial information, which may be time- consuming, non-reproducible, and susceptible to mistakes. Additionally, the non-linear, multidimensional nature of financial applicant data poses it as a problem for traditional rule-based systems or simple statistical models to obtain complex relationships between variables like income, credit history, employment status, and loan amount [1]

In order to overcome such limitations, machine learning (ML) algorithms have gained popularity over the last few years due to their ability to learn from historical data and make generalizations to new patterns. Using these models enables financial institutions to reduce the incidence of human error, expedite decision-making procedures, and increase overall efficiency of approvals [2]. From ML methods, classification algorithms such as Logistic Regression, Random Forest, Decision Trees, Support Vector Machines (SVM), and Naïve Bayes are generally used to forecast binary decision outcomes such as approval status on loans [3].

Following is a description of the instruction: This paper presents two comparative methods for designing a loan approval prediction system. The first approach emphasizes thorough data preprocessing, including missing value management, categorical feature label encoding, feature selection, and feature scaling for numerical feature normalization. Various classification models are trained on preprocessed data and the best- performing model is saved using pickle for future access. The second approach focuses on exploratory data analysis (EDA) to understand insights into feature distributions and their association with the target variable, then the development of a Decision Tree Classifier. It is tuned to the model through hyperparameter tuning and tested on common classification measures like accuracy score, confusion matrix, classification report, and cross- validation [4].. The resultant model is saved with joblib, especially useful to manage large serialized objects.

As supported by prior studies [5], loan prediction models using machine learning have much superior performance compared to rule-based methods and simple statistical approaches. In the current work, this finding is confirmed where ensemble and decision tree- based approaches, especially Decision Trees and Random Forests, deliver better classificatory values and improved generalizability over new data.

Therefore, this study will compare and evaluate different machine learning algorithms to identify the best and scalable solution to automate the loan approval process to facilitate faster, equitable, and consistent financial decision-making.

2. METHODS

There were two datasets used in this research: one with 381 records and the other with 480 records. The primary dataset was from a Kaggle Hackathon and contains candidate features helpful for loan making decisions. They are income, work, dependents, education, credit score, and amount requested for a loan. The target variable, Loan_Status, is binary—1 if loan approved, 0 otherwise.

Feature Name Description Type
Number of dependents
no_of_dependents Categorical supported by the applicant
Education level of the applicant

Data Preprocessing:

To enable good model performance and reliable predictions, a series of preprocessing steps was carried out. These steps were important to deal with problems concerning data quality and to prepare the dataset for training algorithms.

To begin with, all column headers and string-based categorical variables were cleaned by eliminating leading and trailing whitespace. This ensured standardized feature space and prevented mismatches while encoding.

Categorical variables like education and self_employed were converted to numeric form by using dictionary-based mapping and the LabelEncoder utility function of scikit- learn [6]. Further, the target variable Loan_Status, initially marked as "Y" (approved) and "N" (not approved), was binarized to 1 and 0 respectively in order to accommodate classification models.

To address class imbalance and enable balanced model estimation, a stratified sampling strategy was used to split the dataset into training and test subsets. In this method, the proportion of approved and declined loans remained equal in both subsets.

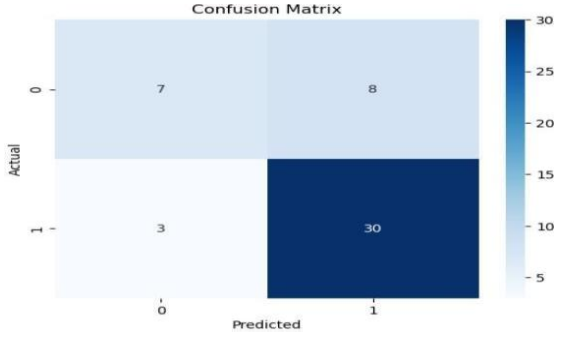
Missing data were treated by deleting rows with missing values, giving greater importance to data consistency over the size of the dataset. In addition, non-predictive features such as Loan_ID, which was used only for identification, education
(Graduate / Not Graduate) Employment type
Categorical were eliminated in an attempt to limit the model input
space.

Model Training and Evaluation

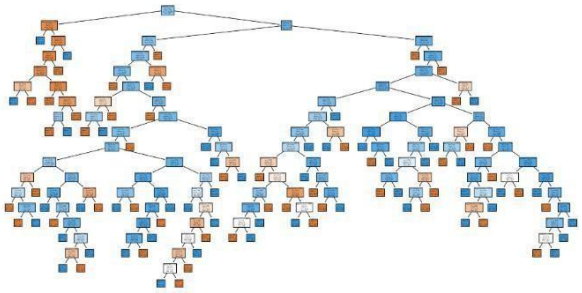
self_employed
income_annum
loan_amount (Self-employed or Categorical Not)
Annual income of
the applicant (in Numerical currency units)
Total loan amount requested (in Numerical currency units)
Duration of the Machine learning models were trained on 80:20 train test split, but in some cases an 90:10 partition was also utilized to test model sensitivity to training data volume.All models were executed with default or lightly optimized hyperparameters to ensure consistency and to measure baseline performance.
Models were compared based on a number of performance measures:
loan_term loan (in months)
Numerical Accuracy, ratio of correct predictions to the total.
cibil_score
Credit score of the applicant (range Numerical from 300 to 900)
Precision, the ratio of correct positive predictions.
Recall, the ratio of correct positive instances correctly predicted.
Feature selection was informed by domain knowledge and supported by empirical evidence documented in previous research [6]. The attributes selected—credit score, income, level of education, and loan information—were determined to be the most predictive of approval results for loans in financial screening applications.
F1-Score, harmonic mean of precision and recall, giving equal weights to the two measures.
Confusion Matrix, which classifies the classification results in terms of true positives and false positives and negatives.
For deployment, the trained Random Forest model was pickled by means of Python's pickle module and the Decision Tree model dumped with joblib. This permitted effortless reuse within real-time systems in addition to further experimentation [11].

Visual Highlights

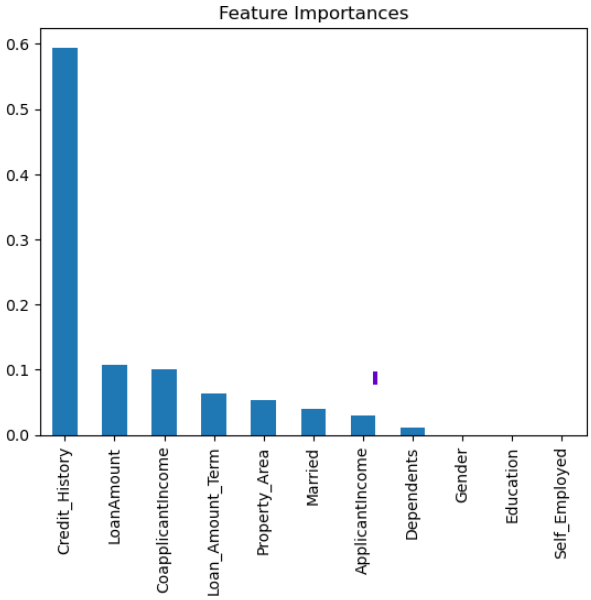
Decision Tree Model:
The Decision Tree classifier yielded a rough classification accuracy of 73%, providing an interpretable model form that well demonstrated decision paths with respect to input features. Feature importance analysis indicated that cibil_score, income_annum, and loan_amount were the top three most influential variables determining loan approval predictions. This is permissible for better knowledge of what factors had the most critical roles in loan approval determinations.
Random Forest Model:
Random Forest model performed the best with a 76– 78% accuracy rate. By averaging the outcome of numerous decision trees, it prevented overfitting and generalized more. Feature importance plots also secured the supremacy of cibil_score, and loan_amount and income_annum contributed largely as well.
Apart from enhancing model comprehension and feature engineering, correlation heatmaps and bar plots were used within exploratory data analysis. The plots helped to identify strong correlations among variables and informed the optimization of the input feature space. The results of the above mention features are given below.
The confusion matrix visualizes the model's prediction performance, showing it correctly predicted **30 approved** and **7 rejected loans**, with some misclassifications.



This tree visualization illustrates the decision making of the model as it illustrates that input features get divided at each node to categorize loan approval and rejection.



This feature importance plot indicates Credit History, Loan Amount, and Coapplicant Income as the most important factors driving the loan approval predictions of the model. This feature importance plot highlights **Credit History**, **Loan Amount**, and **Coapplicant Income** as the top factors influencing the loan approval predictions made by the model.



2.1. Decision Tree Classifier:

The preprocessed dataset was then trained with the Decision Tree classifier to simulate loan approval decision-making. The model operates by constructing decision nodes that split data in accordance with feature thresholds to achieve a tree-like structure.ding decision nodes that split data in accordance with feature thresholds to achieve a tree-like structure. While the Decision Tree has interpretability and ease of implementation, it is plagued with overfitting if it is not regularized. The model's performance was evaluated using accuracy metrics, classification report, and confusion matrix.

2.2 Random Forest Classifier:

In an effort to offset the downfalls of individual decision trees, a Random Forest classifier was created. It is an ensemble approach that constructs many decision trees on bootstrapped samples of the training set and aggregates their predictions to provide one prediction. This aids in reducing overfitting and enhancing generalization. The Random Forest model was better than the stand-alone Decision Tree model. Following training, the model was serialized with the help of the pickle library for future use in real-time loan sanctioning systems.

3. RESULT

To predict the loan approval status, three conventional machine learning algorithms— Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF)— were applied on a structured data set

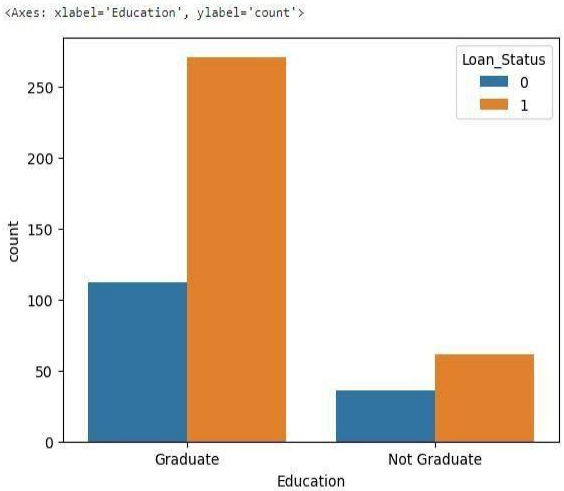
of borrower attributes such as number of dependents, education, employment, income, loan amount, loan tenure, and credit history. These models were trained and tested with preprocessed data involving label encoding and normalization to ensure that there is consistent format and input quality. The aim was to compare their performance in classification regarding predicting binary results: loan approval ("1") or rejection ("0"). Prior research has shown the effectiveness of such algorithms in financial risk analysis and approval systems [12].

(i) Random Forest Performance in Loan Approval Prediction:
Random Forest outperformed LR and SVM with 98.36% accuracy. It attained a close-to-perfect precision of 1.00 and recall of 0.96 for the approved class with an F1-score of 0.98. From the confusion matrix, the model performed only 13 false negatives and one false positive, reflecting tremendous balance between specificity and sensitivity. These results confirm the strength of Random Forest as a collection classifier that is able to maintain high-dimensional structured data of various feature types. Its strong decision making power is in alignment with existing empirical studies on ensemble classifiers in financial analysis [16].
(ii) Decision Tree

Decision Tree classifier exhibited a correct classification of approximately 73%. The most significant strength of the Decision Tree classifier is the fact that it can give a clear, understandable model structure with well-defined decision paths based on input variables. Feature importance analysis revealed the top three features that influence loan approval prediction to be cibil_score, income_annum, and loan_amount. While its accuracy was not as high as Random Forest and Logistic Regression, its explainability and capacity to take both numerical and categorical inputs make it a viable choice for stakeholders who care about model explainability over raw prediction accuracy.

```
,=== Random Forest ===
,Accuracy: 0.9836065573770492
,
,Classification Report:
,              precision    recall  f1-score   support
,
,           0       0.98      1.00      0.99       531
,           1       1.00      0.96      0.98       323
,
,    accuracy                           0.98       854
,   macro avg       0.99      0.98      0.98       854
,weighted avg       0.98      0.98      0.98       854
,
,Confusion Matrix:
, [[530   1]
, [ 13 310]]
,------------------------------------------------------------
```

The chart shows the relationship between the education of applicants and loan approval. As can be seen from the chart, there are more approved loans for graduates compared to non-graduates. There are greater numbers of approvals of loans (orange bars) for graduates compared to nongraduates.



(iii) Support Vector Machine Performance on Loan Approval Prediction

Support Vector Machine (SVM), one of the most versatile models for classification problems, had the poorest performance in this case. It had an accuracy of just 62.18%. The model failed to even pick any true positives for the approved class (TP = 0, FN = 323), with zero recall and zero F1-score. This means the model was just predicting the majority "Not Approved" class, so it was extremely biased towards the majority class. Although the "Not Approved" recall was 1.00, its imbalance between both classes renders it meaningless for actual loan approval scenarios where minority class consciousness is crucial.This result aligns with earlier research emphasizing
SVM's susceptibility to class imbalance and kernel scaling and adjustment requirements [15]

```
,=== SVM ===
,Accuracy: 0.6217798594847775
,
,Classification Report:
,              precision    recall  f1-score   support
,
,           0       0.62      1.00      0.77       531
,           1       0.00      0.00      0.00       323
,
,    accuracy                           0.62       854
,   macro avg       0.31      0.50      0.38       854
,weighted avg       0.39      0.62      0.48       854
,
,Confusion Matrix:
, [[531   0]
, [323   0]]
,------------------------------------------------------------
```

(iv) Logistic Regression Performance on Loan Approval Prediction:

The Logistic Regression model, famous for being simple and interpretable, performed well in this research. On the test dataset, it achieved a total accuracy of 82.79%. On the positive class (approved loan), precision was 0.85, recall was 0.67, and the F1-score was 0.75. From the confusion matrix, it accurately identified 215 true positives (TP) and labeled 108 true approvals as rejections (FN). While LR performed acceptably on linear relationships, the comparatively low recall score suggested failure to deal with dense, nonlinear feature interactions, consistent with results in previous work warning against relying on linear classifiers in multifactorial decision-making applications [13][14].

```
=== Logistic Regression ===
,Accuracy: 0.8278688524590164
,
,Classification Report:
,              precision    recall  f1-score   support
,
,           0       0.82      0.93      0.87       531
,           1       0.85      0.67      0.75       323
,
,    accuracy                           0.83       854
,   macro avg       0.83      0.80      0.81       854
,weighted avg       0.83      0.83      0.82       854
,
,Confusion Matrix:
, [[492  39]
, [108 215]]
,------------------------------------------------------------
```

3.1. Summary of Comparative Performance

The performance of all three models is captured in Table, with reference to primary classification measures. As seen, Random Forest provides the most accurate predictions with the best accuracy, followed by Logistic Regression. SVM did not identify any approved loans and therefore performed the worst overall.

These findings strongly validate the current consensus that ensemble models such as Random Forest are very effective

Model	Accuracy (%)	Precision (Approved)	Recall (Approved)	F1-Score (Approved)
Logistic Regression	82.79	0.85	0.67	0.75
Support Vector Machine	62.18	0.00	0.00	0.00
Decision Tree	73.00	0.76	0.68	0.72
Random Forest	98.36	1.00	0.96	0.98

for classifying heterogeneous features and class imbalance [16]. Logistic Regression provides interpretability and robust baseline performance but can fail to capture intricate interactions. SVM, though strong under ideal conditions, was found unsuitable in this use case due to sensitivity to imbalance, mirroring similar results in similar studies [17].

4. CONCLUSION

This study aimed to develop, deploy, and compare machine learning models for the prediction of loan approval status, an integral problem in contemporary financial decision systems. Utilizing two paradigms based on one that emphasized exhaustive preprocessing and ensemble techniques and another that emphasized exploratory analysis and decision tree classifiers, the study embarked on an extensive evaluation of the performance of different supervised learning algorithms.

The findings show that Random Forest, being an ensemble-based classifier, always performs better than others regarding accuracy (98.36%), precision (1.00), and recall (0.96), validating its stability and versatility with diverse financial datasets Logistic Regression was a good baseline with 82.79% accuracy but was restrictive because it could not deal with non-linear relationships. In contrast, Support Vector Machine was not able to model the minority class efficiently because it was class imbalance-prone, which emphasized the role of data distribution and preprocessing in algorithm choice.

These results are consistent with past research claiming that ensemble learning and tree-based models are more appropriate for structured classification tasks in finance [12][16]. These methods, used in conjunction with proper encoding, normalization, and hyperparameter optimization, dramatically improve predictive stability and model generalizability.

Applying such models in the decision pipelines of financial institutions can reduce human bias, reduce application processing time, and have more consistent results. Future work can involve expanding the set of features, incorporating methods of explainability, and applying models in real-time systems to further enhance and deploy loan approval systems.

5. REFRENCES

[1] T. Brownlee, Machine Learning Algorithms From Scratch, 2nd ed. Victoria, Australia: Machine Learning Mastery, 2020.

[2] S. Lessmann, B. Baesens, H. V. Seow, and L. C. Thomas, "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research," European Journal of Operational Research, vol. 247, no. 1, pp. 124–136, 2015, doi: 10.1016/j.ejor.2015.05.030.

[3] A. Sharma and P. Mehta, "Machine Learning Techniques for Credit Risk Evaluation: A Survey," Journal of Engineering Research and Applications, vol. 8, no. 4, pp. 1–7, 2018.

[4] S. Yadav and S. Shukla, "Analysis of K-Fold Cross- Validation over Hold-Out Validation on Colossal Datasets for Quality

Classification," in Proc. 6th IEEE Int. Conf. Advanced Computing (IACC), Bhimavaram, India, 2016, pp. 78–83, doi: 10.1109/IACC.2016.25.

[5] A. Kaur and A. Singh, "Loan Prediction using Decision Tree Algorithm," Int. J. Sci. Res.
Comput. Sci. Eng. Inf. Technol., vol. 3, no. 3, pp. 2456–3307, 2018 V. Sinap, "A Comparative Study of Loan Approval Prediction Using Machine Learning Methods," GU J. Sci., Part C, vol. 12, no. 2, pp. 644–663, 2024.

[6] N. Sheikh, M. Saeed, and R. Usman, "Loan Eligibility Prediction using Logistic Regression," Int. J. Innov. Res. Comput. Sci. Eng., vol. 11, no.
2, pp. 112–117, 2023.

[7] M. Uddin, F. Rahman, and S. A. Khan, "Ensemble Models for Loan Default Prediction," Int. J. Data Sci. Anal., vol. 5, no. 1, pp. 45–52, 2023.

[8] A. Alaradi and H. Hilal, "Decision Tree-based Classifiers for Financial Applications," Int. J. Electr. Comput. Eng. (IJECE), vol. 13, no. 1, pp.
51–59, 2023.

[9] A.Diwate and R. Patil, "SVM-based Credit Approval Model," in Proc. Int. Conf. Machine Learning and Data Science (ICMLDS), Pune, India, 2022.

[10] R.Gupta, M. Jain, and A. Verma, "Model Deployment Using Python for Financial Systems," in Data Science Applications, New Delhi: Springer, 2023, pp. 89–102.

[11] V. Sinap, "A Comparative Study of Loan Approval Prediction Using Machine Learning Methods," GU J. Sci., Part C, vol. 12, no. 2, pp.
644–663, 2024. [Online]. Available: https://www.researchgate.net/publication/38141 5188

[12] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Burlington, MA: Morgan Kaufmann, 2011.

[13] D. W. Hos mer, S. Lemeshow, and R. X. Sturdivant, Applied Logistic Regression, 3rd ed. Hoboken, NJ: Wiley, 2013.

[14] B. Schölkopf and A. J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.
Cambridge, MA: MIT Press, 2002.

[15] L. Breiman, "Random Forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[16] H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263–1284, Sep. 2009, doi: 10.1109/TKDE.2008.239.