Inspira- Journal of Modern Management & Entrepreneurship (JMME)
ISSN : 2231–167X (Print), Impact Factor: 7.866, Volume 15, No. 03, July-September, 2025, pp. 115-121

LOAN FRAUD DETECTION USING DECISION TREE AND RANDOM FOREST MODELS

Shama Rani[1*] & Prof. Anil Kumar Mittal[2]

[1]Research Scholar, University School of Management, Kurukshetra University Kurukshetra and Assistant Professor of Commerce, Government College for Girls, Palwal, Kurukshetra.
[2]Director, Institute of Management Studies, Kurukshetra University, Kurukshetra.

[*]Corresponding Author: shamagovtcollege@gmail.com

***Citation**: Rani, S., & Mittal, A. (2025). LOAN FRAUD DETECTION USING DECISION TREE AND RANDOM FOREST MODELS. Journal of Modern Management & Entrepreneurship, 15(03), 115–121. https://doi.org/10.62823/jmme/15.03.7940*

ABSTRACT

Banking system vulnerabilities generated opportunities for fraudulent activities, which result in both financial losses and reputational harm for banks and their customers. Financial fraud in the banking sector each year results in significant monetary losses. The continuous problem led financial institutions to close multiple banks, which denied potential borrowers access to loans while producing numerous job losses among banking staff. This study leverages past loan fraud records and employs machine learning to detect fraudulent activities in bank loan applications. The integration of data mining technology improves loan administration through deficiency detection in loan applications, before potential future risks that manual credit officer evaluation might miss. In this work, we utilize two machine learning approaches, i.e., Decision Tree and Random Forest.

***Keywords**: Fraud, Decision Tree, Random Forest, Machine Learning, Banking System.*

Introduction

There are serious financial and reputational concerns for banks and other financial institutions as a result of the surge in fraudulent activity brought on by the growing reliance on digital banking and automated loan processing. The study by Siddiqui et al. (2021) establishes loan fraud as obtaining unauthorized credit through knowingly misrepresented information or fabricated financial paperwork. The extensive volume of applications with advanced fraudulent techniques makes human credit officer reviews ineffective at detecting complex fraud operations (Zhang and Li, 2020).Early detection stands as an essential step for reducing and controlling losses according to Swain and Pani (2016). The identification of emerging threats and the tracking of fraudsters' activities becomes possible through the implementation of suitable algorithms alongside software systems and programs. The banking sector uses machine learning techniques as powerful tools to detect fraud because of existing security problems. The examination of previous loan fraud records through these models brings improved accuracy to fraud detection systems while minimizing financial risks by identifying concealed patterns (Gupta and Sharma, 2019).This research utilizes previous loan fraud data through Decision Tree and Random Forest algorithms for identifying fraudulent loan applications. This study brings data mining methods to develop an evidence-based prevention system for financial institution fraud protection.

Related Studies

The rising demand for credithas led to many loan applications, resulting in an increase in workload that manual financial institution evaluations cannot handle efficiently. Current rule-based systems combined with statistical models experience difficulties in learning complex and fast-changing fraudulent patterns, thus leading to major financial costs. The financial sector chose machine learning (ML) as its robust alternative after facing this challenge. Mamun et al. (2022) researched bank loan eligibility prediction using machine learning (ML) models to analyze multiple models by their performance metrics. The research showed how machine learning techniques succeed at analyzing candidate profiles to boost the decision-making capabilities during loan approval assessments. Similarly, Kavitha and Suriakala (2015) conducted a thorough examination of fraud detection strategies because they observed advanced fraud techniques while stressing how crucial data-based intelligent systems have become. Through their study they found that ML systems utilize significant amounts of historical data including age, income, credit history, employment type and loan amount to guarantee more precise prevention of fraudulent loan activities.

Milojević and Redzepagic (2021)conducted research about artificial intelligence (AI) and machine learning (ML) applications for banking risk management specifically related to credit risk and market risk and liquidity risk and operational risk domains. The authors underscored the capacity of AI and ML technology to tackle worldwide financial problems including those which surfaced during the COVID-19 pandemic. Studies demonstrated that combined AI technologies with deep learning algorithms along with machine learning models for big data analysis would bring better operational outcomes and reduced costs and higher client support. They pointed out three main obstacles including model uncertainty together with a shortage of trained staff and difficulties accessing data. A phased AI and ML application model together with extensive risk management practices would allow organizations to obtain maximum benefits while reducing potential risks. Research by Eweoya*et al.* (2019) demonstrated that the decision tree method delivered 75.9% accuracy in fraud prediction tasks better than established statistical and traditional methods that demonstrated poor predictive results in this domain. Kumar et al. (2022) demonstrates that the Decision Tree and Random Forest algorithms deliver efficient classification results for fraudulent and non-fraudulent loan applications. Guerra and Castelli [10] performed a thorough examination of academic research on machine learning (ML) within banking supervision domains focusing on credit risk evaluation and stress assessment processes. Their evaluation included 41 research papers alongside 2 book chapters which revealed five popular machine learning methods: K-nearest neighbors (KNN) and support vector machines (SVM) and decision trees and ensemble models and neural networks. The review demonstrated that machine learning improves both supervision warning systems and predictive analytics technologies and supports better decision-making. They emphasized that a standardized approach such as the ECB's Risk Assessment System (RAS) should replace current fragmented methods because it offers structured data collection methods that boost the practical use and comparison potential in banking oversight operations.

Methodology of Proposed Study

Finance institutions and banks employ automated assessment systems to review loan applicants through a combination of applicant information and financial stability status and requested loan amount and credit history. Loan approval processes gain more accuracy while becoming more efficient and less risky to fraud through automated assessment systems. The initial step before analyzing data requires a definitive definition of the problem statement. The predictive model depends heavily on identifying both dependent and independent variables after which comes the following step.

This study follows a planned methodology, and several essential issues must be evaluated at this phase.

- Which distinct data patterns within the dataset would be most helpful for accurate loan approval predictions?
- What should the approach be for dealing with absent or non-relevant data points?
- What is the most suitable prediction method while selecting the appropriate models for the current task?
- The model shows what precision level it achieves while various algorithm methods display their performance metrics against each other.

The main goal of this study involves developing a system which predicts whether banks should grant or decline loan applications by defining the problem as a binary classification task. The classification model determines which applications belong to the two possible status groups of approval and denial. An organized procedure was implemented to create a predictive model through Decision Tree and Random Forest algorithm development. A reliable loan approval prediction system relies on conducting all phases including data preprocessing and model evaluation in order to achieve accuracy.

Data Collection

The research dataset was acquired from Kaggle which serves as a common platform for academic and research data needs. The dataset contains two parts with distinct functions: the designed section serves for model training and the assessment section is reserved for performance testing.The training dataset serves as the basis for creating predictive models through separate divisions where one component develops the model and another portion validates its performance. Training the model with a larger portion of data (80% or 70%) enables it to discover patterns and data relations.The tested data enables researchers to determine how well their trained model performs. The model uses between 20% and 30% of the total data for evaluation to determine its capability to predict unknown data points. The model's accuracy and effectiveness can be assessed through testing done on this distinct dataset. The testing dataset serves to measure the exactness of the developed model. In classification models accuracy represents how many predicted results match their genuine correspondences. The database contains 32,581 exclusive customer records which feature 13 variables that serve both loan approval and fraud detection functions:

- **Customer Information**
  - customer_id – Unique identifier for each customer
  - customer_age – Age of the customer
  - customer_income – Annual income of the customer
  - home_ownership – Type of home ownership (e.g., owned, rented, mortgage)
  - employment_duration – Length of employment in years
- **Loan Details**
  - loan_intent – Purpose of the loan application
  - loan_grade – Assigned risk grade of the loan
  - loan_amnt – Amount of the loan requested
  - loan_int_rate – Interest rate applied to the loan
  - term_years – Loan repayment term in years
- **Credit & Default History:**
  - historical_default – Record of past loan defaults
  - cred_hist_length – Length of the customer's credit history
- **Target Variable:**
  - Current_loan_status – Status of the loan (approved, denied, or defaulted)

The structured format of this data allows training a predictive model capable of estimating loan approval risk along with fraud potential.

Data Preprocessing

Data preprocessing led to missing value identification so researchers could resolve these issues to strengthen dataset integrity and accuracy. First, we extracted the dataset and converted it to CSV/Excel format for analyzing purposes. A systematic process using Python together with Pandas library detected the degree of missing values across columns through its examination of all columns. Data scientists need to address missing or inconsistent entries because handling these data points forms the basis of developing trustworthy machine learning models. The dataset contains missing values, which are illustrated in Table 1.

Table 1

Column Name	Missing Values
customer_id	3
customer_age	0
customer_income	0
home_ownership	0
employment_duration	895
loan_intent	0
loan_grade	0
loan_amnt	1
loan_int_rate	3116
term_years	0
historical_default	20737
cred_hist_length	0
Current_loan_status	4

As shown in Figure 2, each variable in the data set had its missing value percentage calculated. Historical_default registered the highest percentage of missing data amounting to 63.77% while loan_int_rate recorded 9.56% and employment_duration had 2.75% missing values. The data columns loan_amnt and Current_loan_status contained minimal null values while all features including customer_age, customer_income, and loan_intent contained no missing data at all.

```
>>> missing_percentage = (df.isnull().sum() / len(df)) * 100
>>> print("\nPercentage of Missing Values per Column:\n", missing_percentage)

Percentage of Missing Values per Column:
 customer_id            0.009206
customer_age            0.000000
customer_income         0.000000
home_ownership          0.000000
employment_duration     2.746578
loan_intent             0.000000
loan_grade              0.000000
loan_amnt               0.003069
loan_int_rate           9.562389
term_years              0.000000
historical_default     63.637759
cred_hist_length        0.000000
Current_loan_status     0.012275
```

Figure 2

A combination of appropriate filling techniques was used to handle data gaps by employing median distributions for numerical fields and mode values for categorical fields to preserve data validity while supporting model stability. Processing of the loan_amnt column involved removing currency symbols, followed by converting its values into a numeric format to create consistent data for modelling applications.

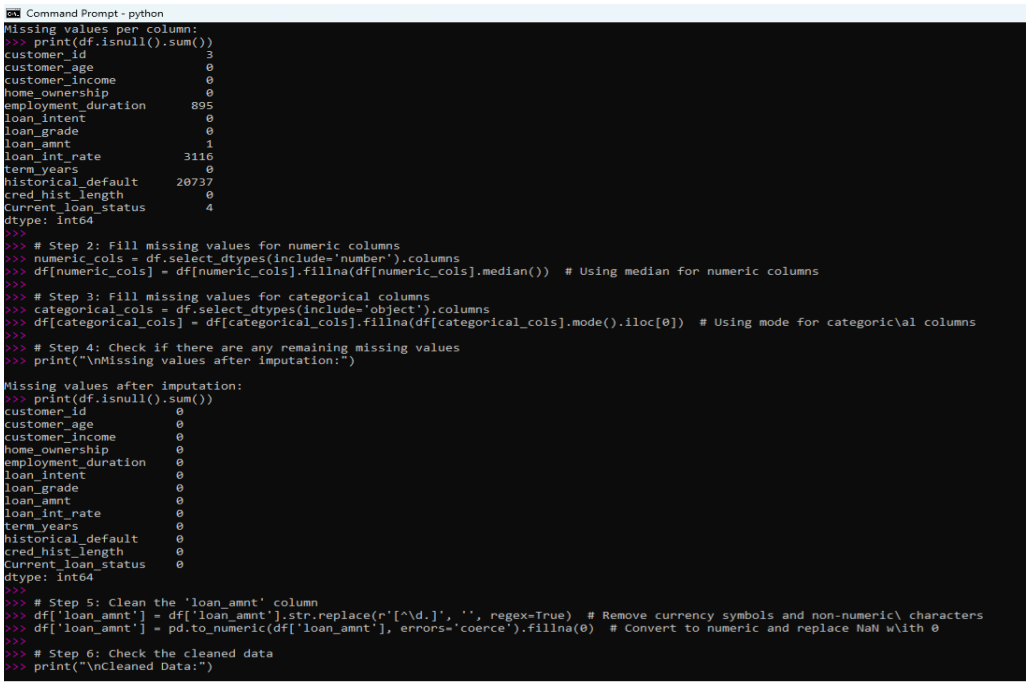

Figure 3

The Python commands for imputing missing data and their corresponding outputs are presented in Figure 3.The programmed data separation process split encoded information into X for features along with y for the target variable. The 'Current_loan_status_NO DEFAULT' column functioned as the target variable while the rest of the dataset elements remained part of the feature set. The dataset split its components into training and testing groups based on an 80:20 proportion and used a fixed random state value at 42 to guarantee result reproducibility. Out of the original set of 32,586 instances the model received 26,068 for training purposes and 6,518 for testing purposes. The stratified partition generates solid model assessment and reduces sampling bias before Machine Learning algorithms such as Decision Tree and Random Forest begin their operation.

Decision Tree Model Performance

A partition of 80% training (26,068 samples) and 20% testing (6,518 samples) subsets was applied to the pre-processed data for removing performance bias in model evaluation. The Decision Tree classifier received training using 80% of the available data while random_state was set to 42 for repeatable results. The model created predictions for the test set hold-outs which allowed measuring performance through overall accuracy and precision and recall and F_1-score metrics for each class.

Table 4

Class	Precision	Recall	F1-Score	Support
0 (Non-default)	0.21	0.24	0.22	1148
1 (Default)	0.83	0.80	0.82	5370
Overall Accuracy	0.7041			6,518

As shown in Table 4, the Decision Tree model reached 70.41% accuracy but displayed inadequate performance for the minority Non-default applications due to poor metrics (precision=0.21 and recall=0.24) that reflected high misclassification rates. The model performed higher effectiveness in detecting defaults (F_1-score = 0.82) when compared to identifying non-defaults. The detection rate of the minority class requires improved techniques such as class-weight adjustment with resampling strategies or ensemble methods according to these results.

Random Forest Model Performance

The data collection underwent preprocessing before the researchers divided it into training groups that comprised 80% of the 26,068 samples and testing groups that included 20% of 6,518 samples. A Random Forest classifier received training from the prepared data in the training set before performing its evaluation against the reserve test data. Through its implementation the model managed to deliver an accuracy of 97% which eclipsed traditional rule-based strategies. Table 5 contains point-by-point information about class metrics.

Class	Precision	Recall	F1-Score	Support
0 (Non-default)	0.96	0.99	0.98	5,142
1 (Default)	0.98	0.86	0.91	1,376
Overall Accuracy	0.97			6,518

Table 5

The Random Forest model (Table 5) achieved impressive results among the majority "Non-default" class by maintaining precision at 0.96 and recall at 0.99, thus preventing minimal false predictions. Evaluation results showed the “Default” class had a precision of 0.98 with a recall of 0.86 because some default cases were not identified. These results confirm the efficacy of ensemble methods in capturing complex patterns and justify their use over simpler rule-based systems. Future work may explore techniques to further boost recall on the minority class, such as class weighting or targeted resampling.

Performance Evaluation Using ROC and AUC

The ROC curve is a visual aid that plots the True Positive Rate (TPR) versus the False Positive Rate (FPR) to assess classification algorithms' performance. The model's overall capacity to discriminate between classes is measured by the AUC (Area Under the Curve) score.

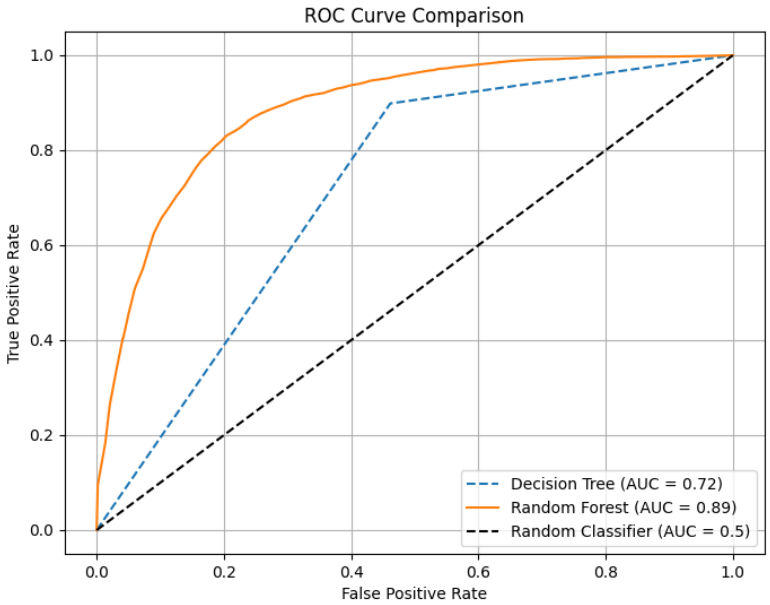

Figure 6

Evaluation of model performance can be conducted through AUC scores that we generated in our analysis.

Table 7

Model	AUC Score	Performance Interpretation
Decision Tree	0.72	Average performance, but is less robust than Random Forest
Random Forest	0.89	Excellent performance, highly robust and generalizable

ROC and AUC analysis serve as a measurement to determine the loan fraud detection capability of Decision Tree and Random Forest models. As demonstrated in Table 7, the Random Forest model displays better classification performance than a Decision Tree model with AUC scores of 0.89 and 0.72, respectively. The ROC curve (Figure 6) shows that the Random Forest model reaches better detection rates, thus proving more effective for identifying fraud.

Limitations of the Study

Some research restrictions limit the accuracy and stability of the obtained findings. Model predictions can be influenced by possible feature selection bias in the dataset obtained from Kaggle. Restrictions in computational ability prevent researchers from conducting tests with sophisticated modelling approaches. The results need to be validated through external testing because they currently only work with this dataset.

Future Perspectives

Future studies can investigate other supervised machine learning techniques, including Support Vector Machines (SVM), Gradient Boosting, and Neural Networks, to improve the precision of fraud detection. Improved model reliability will result from applying advanced resampling methods and cost-sensitive learning approaches to data imbalance problems. The deployment of real-time fraud detection systems throughout multiple financial institutions as well as testing models between institutions, will promote robust adaptive detection systems.

Declarations

No conflicts of interest are present in this study.

References

1. Eweoya, I.O., Adebiyi, A.A., Azeta, A.A. and Azeta, A.E. (2019) 'Fraud prediction in bank loan administration using decision tree', *Journal of Physics: Conference Series*, 1299(1), p. 012037.
2. Guerra, P. and Castelli, M., 2021. Machine learning applied to banking supervision: A literature review. *Risks*, 9(7), p. 136.
3. Gupta, R. and Sharma, A. (2019) 'Machine learning applications in banking fraud detection: A review', *Journal of Financial Technology*, 12(4), pp. 45–58.
4. Kavitha, M. and Suriakala, M. (2015) 'Fraud detection in current scenario, sophistications and directions: A comprehensive survey', *International Journal of Computer Applications*, 111(5), pp. 35–40.
5. Kumar, S., Patel, R. and Mehta, P. (2022) 'Comparative analysis of decision tree and random forest for loan fraud detection', *International Journal of Data Science*, 18(2), pp. 112–130.
6. Mamun, M.A., Farjana, A. and Mamun, M. (2022) 'Predicting bank loan eligibility using machine learning models and comparison analysis', in *Proceedings of the 7th North American International Conference on Industrial Engineering and Operations Management*, June, pp. 12–14.
7. Milojević, N. and Redzepagic, S. (2021) 'Prospects of artificial intelligence and machine learning application in banking risk management', *Journal of Central Banking Theory and Practice*, 10(3), pp. 41–57.
8. Siddiqui, H., Khan, M. and Ali, T. (2021) 'Financial fraud in banking: Challenges and machine learning-based solutions', *Journal of Risk Management*, 25(1), pp. 78–92.
9. Swain, S. and Pani, L.K. (2016) 'Frauds in Indian banking: Aspects, reasons, trend-analysis and suggestive measures', *International Journal of Business and Management Invention*, 5(7), pp. 1–9.
10. Zhang, Y. and Li, J. (2020) 'The role of artificial intelligence in fraud detection: A banking perspective', *Finance AI Journal*, 10(3), pp. 33–50.

❑❑❑