IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

# Machine Learning-Based Loan Approval Prediction and Risk Analysis

**Prachi Panchbhai[1] and Asha Mane[2]**
[1] Student, Department of Masters of Computer Applications
[2]Assistant Professor, Department of Masters of Computer Applications
[1,2] Sadhu Vaswani Institute of Management Studies for Girls, Pune
Savitribai Phule Pune University, Pune, Maharashtra, India.

**Abstract**: *Loan approval depends on the information collected from an applicant. When some of that information is missing, a prediction model may give a different result. This study develops a simple machine learning application for loan approval prediction and examines how reduced applicant information affects the model. A public dataset containing 614 loan records is used. The data is cleaned, missing values are filled, categorical fields are encoded, and the input features are scaled for Logistic Regression. Logistic Regression is used as the main research model, while Random Forest is kept as a second model in the application. With complete information, Logistic Regression gives an accuracy of 86.18%. The accuracy becomes 85.37% at 5% missing information, 83.74% at 10% and 20%, and 79.67% at 30%. The overall fall from the baseline to the 30% setting is 6.51 percentage points. The work also includes feature-wise testing and decision stability as part of the planned analysis. The results show that the amount of applicant information supplied to the model can affect prediction performance.*

**Keywords**: Machine Learning, Loan Approval, Logistic Regression, Random Forest, Credit Risk, Decision Stability

## I. INTRODUCTION

Loan approval is a routine activity in banks and other financial institutions. Before a loan is approved, the applicant is usually assessed using details such as income, employment, credit history, family information and the requested loan amount. In a manual process, these details are checked by people and the decision can take time. The quality of the information available for review also matters.

Machine learning offers a way to support this process by learning patterns from earlier loan applications. Once a model is trained, it can use the information of a new applicant to estimate whether the application is likely to be approved or rejected. Logistic Regression is widely used for binary classification because the output is easy to interpret. Random Forest is another useful method because it combines the results of several decision trees.

This work uses a public loan approval dataset with 614 records. The data contains applicant details such as gender, marital status, dependents, education, employment type, applicant income, co-applicant income, loan amount, loan term, credit history and property area. Loan_Status is the target field with two outcomes: approved and rejected.

The main purpose of this study is not to show that one algorithm is always better than another. Logistic Regression is kept as the main research model and is tested under different levels of reduced applicant information. Random Forest is included in the application as an additional model so that the user can see a second prediction for the same applicant.

### Objectives of the Research

- To develop a machine learning-based loan approval prediction application.
- To use Logistic Regression as the main model for predicting loan approval outcomes.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

- To include Random Forest as an additional model and compare its performance with Logistic Regression.
- To measure the effect of different levels of missing applicant information (5%, 10%, 20%, and 30%) on prediction accuracy.
- To examine how individual applicant features affect loan approval predictions.
- To study the stability of model predictions by calculating the decision flip rate when applicant information is reduced.

## II. METHODOLOGY

The study follows a simple experimental process. The dataset is first cleaned and prepared. The models are then trained using the prepared data. A baseline result is obtained with complete information, after which the available applicant information is reduced in steps. The same Logistic Regression model is used for these tests so that the change in the result can be linked to information availability rather than a change in the algorithm.

### A. Dataset

The study uses a public loan approval dataset containing 614 records. The original fields are listed in Table I. Loan_ID is an identifier, while Loan_Status is the target variable used for prediction.

TABLE I DATASET ATTRIBUTES USED IN THE STUDY

| Attribute | Type | Purpose |
|---|---|---|
| Gender | Categorical | Applicant information |
| Married | Categorical | Marital status |
| Dependents | Categorical | Number of dependents |
| Education | Categorical | Education level |
| Self_Employed | Categorical | Employment type |
| ApplicantIncome | Numerical | Applicant income |
| CoapplicantIncome | Numerical | Co-applicant income |
| LoanAmount | Numerical | Requested loan amount |
| Loan_Amount_Term | Numerical | Loan repayment term |
| Credit_History | Numerical | Credit history status |
| Property_Area | Categorical | Property location |
| Loan_Status | Target | Approved / Rejected |

### B. Data Preprocessing

The raw data contains missing values in some fields. For categorical columns, missing entries are replaced with the mode of the column. For numerical columns, the median is used. The Dependents value written as 3+ is converted to 3. Loan_ID is removed because it is only an identifier and is not a useful applicant feature.

Categorical variables are changed into numerical columns using one-hot encoding. The numerical inputs used by Logistic Regression are scaled with StandardScaler. The target Loan_Status is converted into a binary value, with rejected represented by 0 and approved represented by 1.

### C. Train-Test Split

The data is divided into 80% training records and 20% testing records. Stratified splitting is used so that the proportion of approved and rejected cases remains similar in the two sets. The random state is fixed at 42 so that the same split can be reproduced.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

**D. Model Development**

Logistic Regression is the main model in the study. It estimates the probability of the approved class and then assigns the applicant to an approved or rejected class. Random Forest is trained on the same prepared data and is used as the second model in the application. It combines predictions from multiple decision trees.

**E. Information Reduction Experiment**

The first test is the baseline, where the complete prepared applicant information is used. The next tests reduce the available information by 5%, 10%, 20% and 30%. The Logistic Regression model remains fixed for all five conditions. The purpose is to observe how much the test accuracy changes as information is removed.

The research design also includes feature-wise testing. One applicant field can be removed at a time and the resulting prediction can be compared with the prediction obtained from the complete input. This helps identify which applicant fields deserve closer attention in the analysis. Decision stability is studied by checking whether the model changes an approval into a rejection, or a rejection into an approval, after information is changed.

**F. Evaluation Measures**

Accuracy is the main measure reported for the information-reduction experiment. The application also calculates precision, recall, F1 score and ROC-AUC for the two models. Decision stability is expressed using the decision flip rate: Decision Flip Rate = (Number of Changed Decisions / Total Decisions) × 100.

**G. Research Workflow**

The overall workflow is shown in Fig. 1. The process moves from data collection and cleaning to missing-value handling, feature preparation, model training, baseline testing, information reduction, feature-wise analysis and decision stability.

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/IJARSCT-40438

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

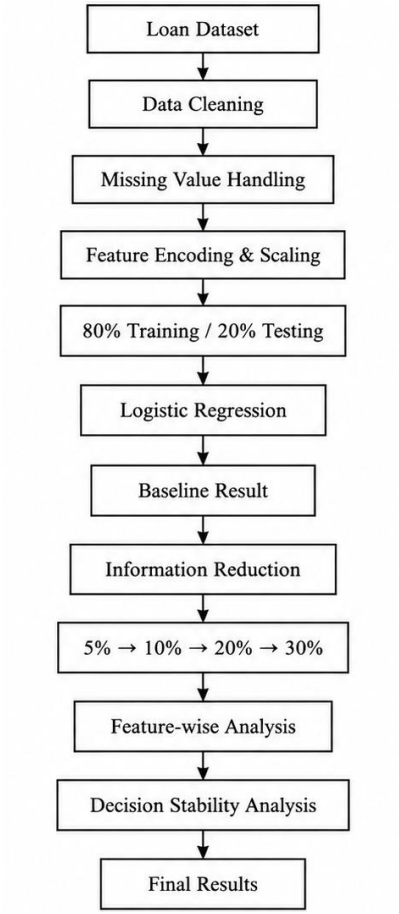



Figure 1: Research workflow used for the loan approval study

## III. LITERATURE REVIEW

Many researchers have studied how machine learning can help banks predict loan approval and assess financial risk. Logistic Regression is a popular choice because it is easy to understand, while Random Forest can find more complex patterns in applicant data.

Previous research also shows that data quality plays an important role in making predictions. When information is missing or incomplete, the model's performance may change. Therefore, preparing the data properly and checking model performance using measures such as accuracy, precision, recall, F1-score, and ROC-AUC are important steps.

In this project, I compare Logistic Regression and Random Forest to understand how they perform when different amounts of applicant information are reduced. I also study individual features and changes in prediction results to understand how missing information affects the model's decisions.

## IV.RESEARCH GAP

Many studies have explored the use of machine learning for loan approval prediction. However, less attention may be given to understanding how prediction accuracy changes when some applicant information is missing. Comparing different models under different levels of information reduction can help provide a clearer picture of this issue.

This project addresses this area by comparing Logistic Regression and Random Forest using different levels of information reduction. It also examines individual applicant features and the decision flip rate to understand whether

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

missing information changes the model's predictions. The aim is to better understand the importance of complete information in loan approval prediction.

## V. MODELING AND ANALYSIS

### A. Logistic Regression

Logistic Regression is used for the main analysis because the target has two classes. The model receives the prepared applicant fields and returns a probability for the approved class. A class label is then assigned from that output. Another advantage of this model for this study is that its behaviour can be examined while the input information is changed.

### B. Random Forest

Random Forest is included as an additional model in the application. It uses multiple decision trees and combines their outputs. In the application, the model gives a second prediction and approval probability for the same applicant. It is not used as the main model for the missing-information experiment.

### C. Application Output

The application provides a form in which the user enters applicant information. After the prediction button is selected, Logistic Regression and Random Forest each return a loan decision and an approval probability. When the two models disagree, the application can display a warning so that the user knows that the two predictions are not the same.

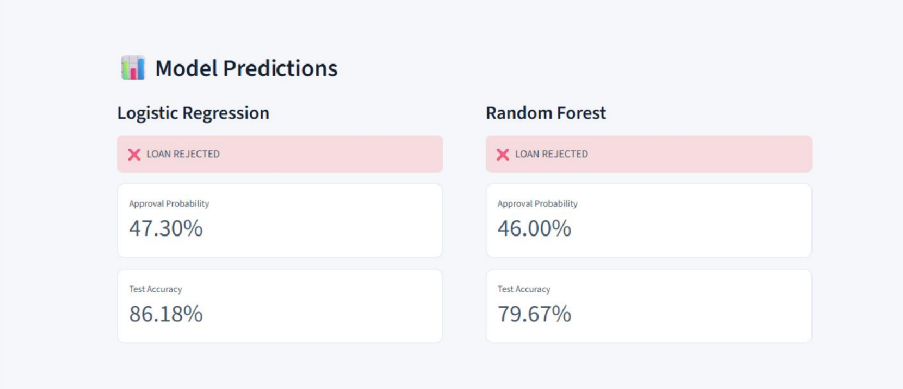



Figure 2: Example output from the loan approval prediction application

### D. Model Performance

Table II gives the performance values currently reported by the application. These numbers are included to describe the implemented models. The main research result remains the change in Logistic Regression accuracy when information is reduced.

TABLE II PERFORMANCE OF THE IMPLEMENTED MODELS

| Model | Accuracy | Precision | Recall | F1 Score | ROC-AUC |
|---|---|---|---|---|---|
| Logistic Regression | 86.18% | 84.00% | 98.82% | 90.81% | 84.86% |
| Random Forest | 79.67% | 85.71% | 84.71% | 85.21% | 79.61% |

## VI. RESULTS AND DISCUSSION

### A. Effect of Missing Applicant Information

The information-reduction experiment shows an overall fall in Logistic Regression accuracy as more applicant information is made unavailable. With all information present, the model reaches 86.18%. At 5% missing information the result is 85.37%. The accuracy becomes 83.74% at both 10% and 20%, and it decreases to 79.67% at 30%.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

TABLE III LOGISTIC REGRESSION ACCURACY AT DIFFERENT INFORMATION-REDUCTION LEVELS

| Missing Applicant Data | Accuracy | Change from Baseline |
|---|---|---|
| 0% | 86.18% | 0.00 percentage points |
| 5% | 85.37% | -0.81 percentage points |
| 10% | 83.74% | -2.44 percentage points |
| 20% | 83.74% | -2.44 percentage points |
| 30% | 79.67% | -6.51 percentage points |

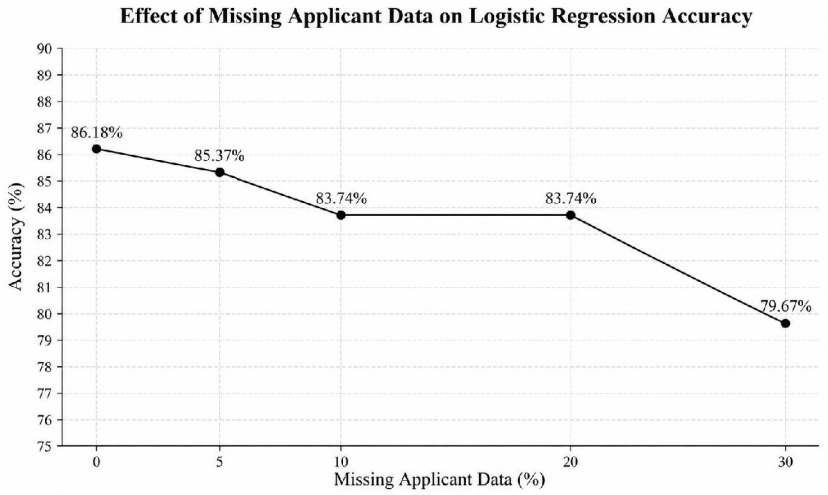



Figure 3: Logistic Regression accuracy at different levels of missing applicant information

The baseline value of 86.18% decreases to 79.67% at 30% missing information. The difference is 6.51 percentage points. The fall is not the same at every step: the accuracy changes by 0.81 percentage points at 5%, by 2.44 percentage points at 10%, remains at the same value at 20%, and then drops by a further 6.51 percentage points from the baseline at 30%. This pattern shows that the relationship between information loss and accuracy is not perfectly linear.

**B. Random Forest Performance**

Random Forest is used in the application as a second prediction model. Its current accuracy is 79.67%, precision is 85.71%, recall is 84.71%, F1 score is 85.21% and ROC-AUC is 79.61%. The purpose of showing these values is to give the application user another view of the same applicant, rather than to make model comparison the central research question.

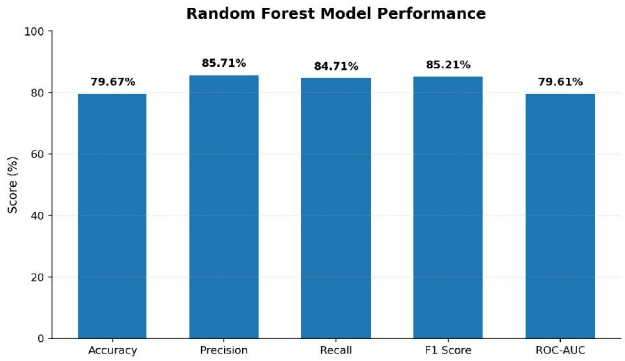



Figure 4: Performance measures of the Random Forest model

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/IJARSCT-40438

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

### C. Decision Stability

Accuracy gives an overall result for a test set, but it does not show whether a particular applicant can receive a different decision after the available information changes. For this reason, decision stability is included in the research design. A decision flip occurs when an applicant moves from approved to rejected or from rejected to approved after information is reduced or changed. The flip rate can be used as an additional way to describe how stable the model decision is.

### D. Interpretation

The present results support a clear observation: reducing the information supplied to the Logistic Regression model can lower its prediction accuracy. This is important for loan-related systems because the model can only use the values that are provided to it. The result does not mean that every missing field will change a decision. It means that, across the tested conditions, removing information changes the model performance.

Keeping the same Logistic Regression model throughout the main experiment makes the result easier to understand. The algorithm, train-test split and preprocessing approach are kept consistent, while the information available to the model is changed. This makes information availability the main variable being examined.

## VII. CONCLUSION

This study developed a machine learning application for loan approval prediction using Logistic Regression and Random Forest. Logistic Regression was used as the main research model, while Random Forest was included as an additional model in the application. The central experiment examined what happens when applicant information is progressively reduced.

The Logistic Regression model achieved 86.18% accuracy with complete information and 79.67% accuracy when the tested missing-information level reached 30%. The difference is 6.51 percentage points. The result indicates that the amount of information supplied to the model can affect its prediction performance.

The study also provides a basis for feature-wise analysis and decision-stability analysis. These measures can give more detail about the behaviour of a prediction model than a single accuracy value. The application is suitable for academic demonstration and research use. A real loan decision would still require proper financial checks, institutional policies and human review.

## REFERENCES

1. D. P. Singh, A. Khare, P. N. Purusothaman, R. Lal, S. S. Chauhan and K. C. Kalaiarasan, “Predictive Modeling for Bank Loan Approval: From Data to Decisions,” Procedia Computer Science, vol. 259, pp. 1426-1431, 2025. doi: 10.1016/j.procs.2025.04.097.
2. L. Hota, P. K. Jain and A. Kumar, “A Comparative Performance Assessment for Prediction of Loan Approval in Financial Sector,” Procedia Computer Science, vol. 258, pp. 298-307, 2025. doi: 10.1016/j.procs.2025.04.267.
3. R. Kurniawan, “Application of Random Forest Algorithm on Credit Risk Analysis,” Procedia Computer Science, vol. 245, pp. 740-749, 2024. doi: 10.1016/j.procs.2024.10.300.
4. N. Uddin, M. K. U. Ahamed, M. A. Uddin, M. M. Islam, M. A. Talukder and S. Aryal, “An ensemble machine learning based bank loan approval predictions system with a smart application,” International Journal of Cognitive Computing in Engineering, vol. 4, pp. 327-339, 2023.
5. D. Maloney, S.-C. Hong and B. N. Nag, “Two class Bayes point machines in repayment prediction of low credit borrowers,” Heliyon, vol. 8, e11479, 2022.
6. N. Bussmann, P. Giudici, D. Marinelli and J. Papenbrock, “Explainable Machine Learning in Credit Risk Management,” Computational Economics, vol. 57, pp. 203-216, 2021. doi: 10.1007/s10614-020-10042-0.
7. S. Shi, R. Tse, W. Luo, S. D’Addona and G. Pau, “Machine learning-driven credit risk: a systemic review,” Neural Computing and Applications, vol. 34, pp. 14327-14339, 2022. doi: 10.1007/s00521-022-07472-2.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

8. A. Bitetto, P. Cerchiello, S. Filomeni, A. Tanda and B. Tarantino, "Can we trust machine learning to predict the credit risk of small businesses?," Review of Quantitative Finance and Accounting, vol. 63, pp. 925-954, 2024. doi: 10.1007/s11156-024-01278-0.
9. C. Guan, H. Suryanto, A. Mahidadia, M. Bain and P. Compton, "Responsible Credit Risk Assessment with Machine Learning and Knowledge Acquisition," Human-Centric Intelligent Systems, vol. 3, pp. 232-243, 2023. doi: 10.1007/s44230-023-00035-1.