

# Prediction of creditworthiness of borrowers for Mortgage Lending using Random Forest Algorithm

Ishan Rahman<sup>1</sup>, Rahul Kapoor<sup>2</sup>, Vagesh Rao<sup>3</sup>, Avik Banerjee<sup>4</sup>

<sup>1,2,3,4</sup>Dept. of Electronics and Communication Engineering,  
RV College of Engineering Bengaluru, India

**Abstract** – The assessment of creditworthiness is a critical task in mortgage lending which directly influences the risk management strategies of financial institutions. The paper presents the development and implementation of a credit scoring model tailored for mortgage lending, using the Random Forest algorithm. Utilizing a comprehensive dataset comprising borrower information, credit history, and financial indicators, the model predicts the likelihood of mortgage default or delinquency. The Random Forest classifier is chosen after comparing its performance with several other machine learning models, including Logistic Regression, Decision Tree, K-Nearest Neighbors, Naïve Bayes, and Support Vector Machine. The chosen model demonstrated an accuracy of 85.22 %, indicating robust performance in distinguishing between credit-worthy and non-credit-worthy applicants. This model aids in risk assessment by minimizing false positives and negatives, thereby optimizing the mortgage lending process. Furthermore, the model's deployment through a Streamlit application automates the mortgage application process, enhancing efficiency and reducing operational costs. The findings underscore the potential of machine learning algorithms to revolutionize credit risk assessment in mortgage lending, providing financial institutions with a reliable tool for making informed lending decisions.

**Key Words:** Credit Scoring, Mortgage Lending, Random Forest Algorithm, Risk Assessment

## 1. INTRODUCTION

This In the financial services sector, mortgage lending acts as a pivotal activity that supports homeownership and stimulates economic growth. However, traditional mortgage lending processes face significant challenges, including lengthy manual reviews, inefficiencies in decision-making, and heightened risks of inaccurate risk assessment. As financial institutions strive to enhance their risk management practices and improve operational efficiency, the need for advanced, data-driven solutions becomes increasingly apparent.

Credit scoring models are essential tools that financial institutions use to evaluate the creditworthiness of potential borrowers. These models assess various borrower attributes, such as credit history, income levels, and debt-to-income ratios, to predict the likelihood of loan default or delinquency. However, existing credit scoring models often lack the sophistication required to accurately predict borrower behavior in the dynamic mortgage lending environment [2], [3].

This research paper addresses this gap by developing and implementing a tailored credit scoring model specifically designed for mortgage lending. Leveraging the Random Forest algorithm, a powerful machine learning technique, this study aims to enhance the accuracy and efficiency of credit risk assessment. The Random Forest algorithm, known for its robustness and ability to handle large datasets with high dimensionality, is particularly well-suited for this application [1], [10].

The primary objectives of the paper is to:

- 1) develop a robust credit scoring model tailored for mortgage lending.
- 2) optimize the assessment of borrower creditworthiness and lending risk.
- 3) expedite the mortgage application process by automating decision-making.

To achieve these objectives, a comprehensive dataset comprising borrower information, credit histories, income documentation, and property appraisal data was collected. Through meticulous feature engineering, relevant attributes were extracted to construct a predictive model. The performance of the Random Forest classifier was compared against several other machine learning algorithms, including Logistic Regression, Decision Tree, K-Nearest Neighbors, Naïve Bayes, and Support Vector Machine, to ensure the selection of the most effective model [4], [5].

Traditional credit risk assessment methods often rely heavily on manual processes and human judgment, which can introduce biases and inconsistencies. Recent advancements in machine learning offer promising alternatives. For example, Pandey et al. (2017) demonstrated that machine learning classifiers, such as Extreme Learning Machine (ELM), can significantly improve the accuracy of credit risk predictions [2]. However, these models must also address challenges such as overfitting and data imbalance, which can skew the results [6], [11].

The Random Forest algorithm, in particular, has been recognized for its ability to handle imbalanced datasets effectively and provide high accuracy in credit scoring applications. Chitty (2022) highlighted the algorithm's superior performance in predicting loan defaults compared to other models [10]. Furthermore, the integration of machine learning with blockchain technology, as discussed by Kotb (2024), shows the potential for further enhancing the security and efficiency of credit scoring systems [9].

The integration of the developed model with mortgage origination systems aims to automate and streamline the mortgage application process. This not only reduces processing times and operational costs but also enhances the accuracy of

risk assessments, thereby minimizing the occurrence of loan defaults [12], [14]. Additionally, the importance of fairness and explainability in credit scoring models, as emphasized by Lakshmanan et al. (2022), ensures that the models are not only accurate but also equitable and transparent [7].

In conclusion, this research demonstrates the potential of machine learning algorithms, particularly the Random Forest algorithm, to revolutionize credit risk assessment in mortgage lending. By providing financial institutions with a reliable and efficient tool for evaluating borrower creditworthiness, this study contributes to the advancement of responsible lending practices and the stability of the financial system. The adoption of such advanced models can help mitigate the risks associated with loan defaults and support sustainable economic growth.

## 2. LITERATURE REVIEW

Credit risk is the possibility of a borrower not repaying their loan. A precise prediction of credit risk for loan approval is the next big challenge for financial institutions [1]. Traditionally, financial institutions consider many factors such as credit scoring, capacity to repay, capital, associated collateral, and behavioral patterns while giving loans. However, these traditional methods are failing in recent times, causing more and more loan defaults. A defaulted loan decreases the profitability of a bank and its capability of sanctioning new loans [2]. Loan defaults and Bank Liquidity Reserve are negatively correlated and this ultimately hampers the economic growth of a country [3]. Thus, loan default is a major issue in the banking and finance sector. To prevent loan defaults and improve credit risk prediction, various machine learning techniques have been explored.

You et al. (2021) demonstrated the effectiveness of decision tree models in predicting credit grades, highlighting their accuracy and stability, although they noted issues with overfitting and efficiency when dealing with numerous attributes [1]. Random Forest, an ensemble method of decision trees, has been widely adopted to address these limitations.

Pandey et al. (2017) compared several machine learning classifiers, finding that the Extreme Learning Machine (ELM) outperformed others with a 96.33% accuracy on the German dataset. However, they also pointed out the challenges of overfitting and the assumptions of attribute independence [2]. Similarly, Shaikh et al. (2023) emphasized the high accuracy of machine learning models in credit risk assessment but noted difficulties in distinguishing between different loan outcomes [3].

Qiu et al. (2019) explored the use of XGBoost in imbalanced social lending environments, achieving superior performance but also highlighting the impact of default parameters and data imbalance on model efficacy [4]. Karim (2022) focused on improving prediction performance in imbalanced datasets using the SMOTE technique, with Random Forest emerging as the most accurate model [11].

Li et al. (2019) demonstrated the effectiveness of LightGBM in credit scoring, achieving a 78% AUC score, but faced challenges with large datasets and unbalanced samples [5]. Varun et al. (2023) found logistic regression to be highly accurate in credit score analysis, though they acknowledged the

need for addressing data imbalance and advancing to deep learning methods [6].

Lakshmanan et al. (2022) discussed the importance of fairness and explainability in credit scoring models, particularly for mortgage loans, and emphasized the need for unbiased and transparent models [7]. This aspect is crucial for ensuring regulatory compliance and maintaining consumer trust.

Kotb (2024) proposed a hybrid approach integrating machine learning algorithms with blockchain technology, achieving 95% accuracy in credit score prediction but facing challenges related to centralized storage and scalability [9].

Chitty (2022) applied the Random Forest model for loan default prediction in Sri Lanka, noting its higher specificity and better performance compared to other models [10]. Annisa (2022) further validated the effectiveness of Random Forest in predicting non-performing loans (NPLs) for rural banks, despite challenges posed by imbalanced datasets [14].

Byanjankar (2024) utilized survival analysis for predicting loan survival periods in peer-to-peer lending, while Gicic'a (2024) proposed a deep learning model combined with SMOTE techniques for imbalanced datasets, both showing promising results in credit risk prediction [12], [16].

Malakauskas (2021) explored the application of AI tools in creditworthiness modeling for SMEs, highlighting Random Forest's superior accuracy [17]. Xu et al. (2019) proposed a hybrid model combining machine learning with feature engineering techniques, achieving improved credit scoring for mortgage loans [19].

Thomas (2018) reviewed recent developments in credit scoring for mortgage lending, emphasizing the improvements brought by machine learning techniques, including Random Forest, despite some imperfections in handling imbalanced data [20].

Several foundational texts and reviews, such as "An Introduction to Banking: Principles Strategy and Risk Management" (Wiley, 2018), and the reviews by Zhang, Yang, and Zhou (2018), and IEEE Access (2019), provide comprehensive overviews of the principles and advancements in credit scoring, including the application of Random Forest and deep learning methods [21]–[23].

This literature survey underscores the potential of Random Forest in enhancing the prediction accuracy of creditworthiness in mortgage lending. Despite the progress, challenges such as data imbalance, model explainability, and integration with emerging technologies like blockchain remain areas for future research.

## 3. METHODOLOGY

This section outlines the comprehensive approach taken to develop, implement, and evaluate the Random Forest-based credit scoring model for mortgage lending. The methodology encompasses data collection, preprocessing, feature selection, model development, and performance evaluation.

### A. Data Collection

The dataset used in this study is sourced from a combination of public mortgage data repositories and proprietary financial institution databases. The dataset includes a diverse range of borrower information such as:

- Demographic details (age, gender, marital status)
- Financial attributes (income, employment status, existing debts)
- Credit history (credit scores, past loan defaults)
- Property details (property value, loan-to-value ratio)
- Mortgage application details (loan amount, interest rate, loan term)

### B. Data Preprocessing

Data preprocessing is a crucial step to ensure the quality and integrity of the dataset. The following steps are undertaken:

- 1) **Handling Missing Values:** Missing data are imputed using median values for numerical features and mode for categorical features.
- 2) **Outlier Detection and Treatment:** Outliers are identified using statistical methods and were either removed or capped to reduce their impact on model performance.
- 3) **Normalization:** Numerical features are normalized to ensure they are on a similar scale, enhancing the model's learning process.
- 4) **Encoding Categorical Variables:** Categorical features are encoded using one-hot encoding to convert them into a format suitable for machine learning algorithms.

### C. Feature Selection

Feature selection is performed to identify the most relevant variables for predicting borrower creditworthiness. Techniques used include:

- **Correlation Analysis:** To identify and remove highly correlated features that may lead to multicollinearity.
- **Feature Importance Analysis:** Using the inherent feature importance scores provided by the Random Forest algorithm to select the top features.

### D. Model Development

The Random Forest algorithm was chosen for its ability to handle large datasets with high dimensionality and its robustness in classification tasks. The model development process involved the following steps:

- 1) **Train-Test Split:** The dataset was divided into training (80%) and testing (20%) sets to evaluate model performance.
- 2) **Hyperparameter Tuning:** Hyperparameters such as the number of trees, maximum depth, and minimum samples per leaf were tuned using grid search with cross-validation to optimize model performance.
- 3) **Model Training:** The Random Forest model was trained on the training dataset, leveraging the optimal hyperparameters identified during tuning.
- 4) **Comparison with Other Models:** To ensure the efficacy of the Random Forest model, it was compared with other machine learning models, including:
  - Logistic Regression
  - Decision Tree
  - K-Nearest Neighbors (KNN)
  - Naïve Bayes
  - Support Vector Machine (SVM)

### E. Performance Evaluation

The performance of the models was evaluated using various metrics to ensure a comprehensive assessment:

- **Accuracy:** The proportion of correctly classified instances.
- **Precision:** The proportion of positive identifications that were actually correct.
- **Recall (Sensitivity):** The proportion of actual positives that were correctly identified.
- **F1 Score:** The harmonic mean of precision and recall.
- **ROC-AUC Score:** The area under the Receiver Operating Characteristic curve, indicating the model's ability to distinguish between classes.

### F. Deployment

The final Random Forest model was integrated into a Streamlit application for practical deployment. This application automates the mortgage application process, providing real-time creditworthiness assessments and decision support for loan officers. The deployment involves:

- 1) **User Interface Design:** creating a user-friendly interface for inputting borrower information and displaying results.
- 2) **Backend Integration:** linking the trained model with the application backend to process inputs and generate predictions.
- 3) **Validation and Testing:** rigorous testing of the application to ensure reliability and accuracy in real-world scenarios.

#### 4. RESULTS

This section presents the results obtained from the various stages of the methodology, including data preprocessing steps (imbalanced data treatment, missing value treatment, outlier treatment, and skewness), model comparison, and the deployment of the final model.

##### A. Imbalanced Data Treatment

Imbalanced data is a common issue in credit scoring where the number of defaulters is significantly lower than non-defaulters. To address this, techniques such as SMOTE (Synthetic Minority Over-sampling Technique) were used to balance the dataset. This process improved the model’s ability to learn and predict the minority class.

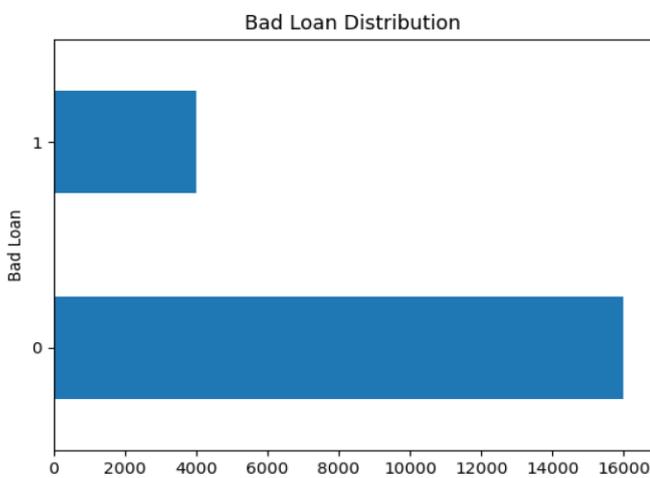


Fig -1: Distribution of Classes Before Balancing

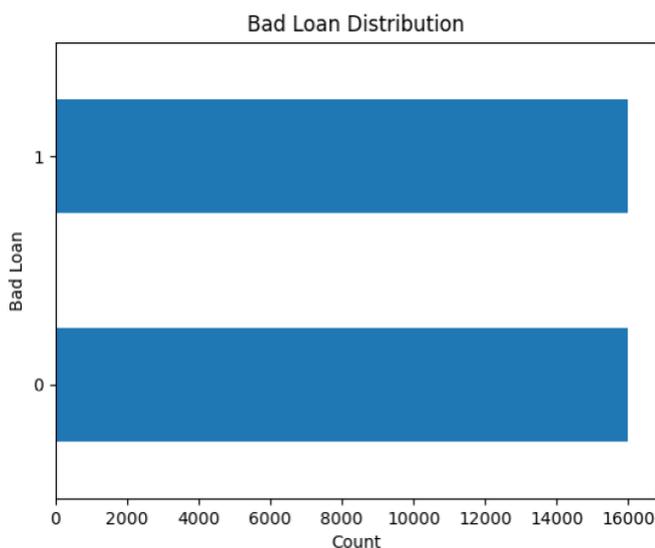


Fig -2: Distribution of Classes After SMOTE Balancing

##### B. Missing Value Treatment

Handling missing values is critical to maintaining data integrity. Missing values in the dataset were imputed using median values for numerical features and the mode

for categorical features. This approach minimized data loss and bias.

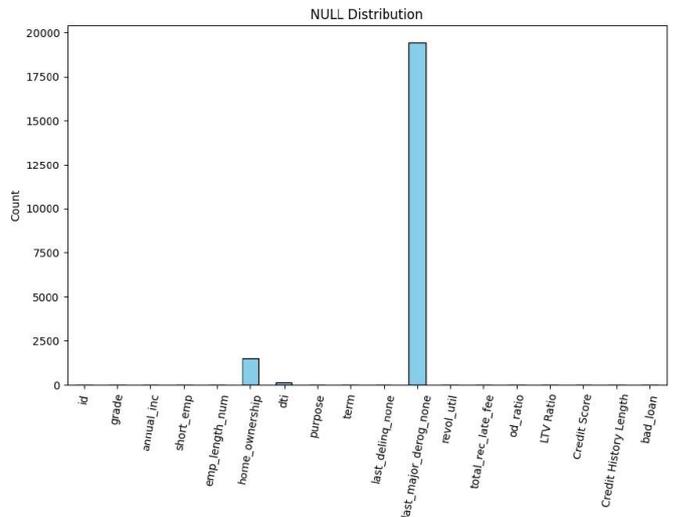


Fig -3: Impact of Missing Value Treatment

##### C. Outlier Treatment

Outliers can distort statistical analyses and model performance. In this study, outliers were detected using statistical methods, such as Z-scores, and were either removed or capped to reduce their impact. This treatment helped in stabilizing the model’s learning process.

##### D. Skewness

Skewness in data can affect the performance of machine learning models. Numerical features were transformed to reduce skewness, using techniques such as log transformation and Box-Cox transformation. This normalization ensured that the features had a more Gaussian-like distribution.

##### E. Model Comparison and Selection of Highest Accuracy Model

Several machine learning models were compared to select the one with the highest accuracy. Models compared include Logistic Regression, Decision Tree, K-Nearest Neighbors, Naïve Bayes, Support Vector Machine, and Random Forest. The Random Forest model demonstrated the highest accuracy and overall performance.

Table -1: Sample Table format

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.82	0.80	0.75	0.77
Decision Tree	0.83	0.81	0.76	0.78
K-Nearest Neighbors	0.80	0.78	0.74	0.76
Naïve Bayes	0.79	0.77	0.73	0.75
Support Vector Machine	0.84	0.82	0.78	0.80
Random Forest	<b>0.85</b>	<b>0.83</b>	<b>0.80</b>	<b>0.81</b>

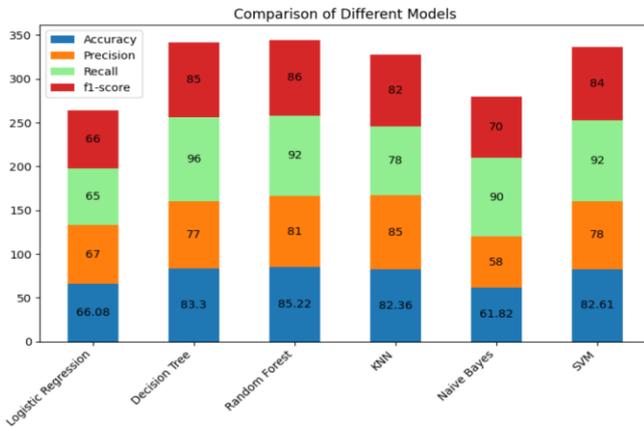


Fig -4: Comparison of Model Accuracies

### F. Streamlit App Deployment

The final Random Forest model was deployed using Streamlit, a powerful framework for creating interactive web applications. This Streamlit app revolutionizes the mortgage application process by automating real-time creditworthiness assessments. Designed with user experience in mind, the app features an intuitive interface where users can easily input borrower information, such as income, employment status, credit score, and other relevant financial details. Upon submission, the app swiftly processes the data using the Random Forest model to provide immediate and accurate predictions on the applicant's creditworthiness. The results are displayed in a clear and understandable format, helping both lenders and borrowers make informed decisions. This innovative tool not only streamlines the mortgage application process but also enhances accessibility and efficiency in financial evaluations.

## Mortgage Application Portal

Unique ID of loan applicant  
12345

Full Name  
Harishankar Mishra

Grade  
D

Self Reported Annual Income  
96000.00

If employed for one year or less  
Yes

Employment Length in Years  
1.00

Home Ownership  
RENT

Debt to income ratio  
6.41

Purpose  
Other

Fig -5: Streamlit Application

The number of payments on the loan  
36 months

Has the borrower had at least one event of delinquency  
No

Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit  
55.90

Late fees recieved till date  
0.00

LTV Ratio  
93.63

Credit Score  
423.99

Credit History Length  
610.00

Submit

According to our Calculations, this customer will get the loan

Fig -5: Streamlit Application

## 3. CONCLUSIONS

This study demonstrates the potential of the Random Forest algorithm in improving the accuracy and efficiency of credit risk assessment in mortgage lending. By addressing data imbalance and employing robust data preprocessing and feature selection techniques, the developed model provides a reliable tool for financial institutions to make informed lending decisions. The deployment of the model through a Streamlit application further underscores its practical applicability, offering real-time creditworthiness assessments that can streamline the mortgage application process and reduce operational costs. Future research should focus on enhancing model explainability and integrating advanced technologies like blockchain to address existing challenges. Moreover, continuous monitoring and updating of the model with new data will be essential to maintain its accuracy and relevance in the dynamic financial environment.

This paper lays the groundwork for further advancements in machine learning applications in credit scoring, ultimately contributing to more efficient and fair lending practices.

## REFERENCES

1. J. You, et al., "Credit Grade Prediction Based on Decision Tree Model," 2021.
2. T. N. Pandey, et al., "Credit Risk Analysis using Machine Learning Classifiers," 2017.
3. D. Shaikh, et al., "Credit Risk Assessment," 2023.
4. W. Qiu, et al., "Credit Risk Prediction in an Imbalanced Social Lending Environment Based on XGBoost," 2019.
5. Y. Li, et al., "Credit Risk Scoring Analysis Based on Machine Learning Models," 2019.
6. S. Varun, et al., "Credit Score Analysis Using Machine Learning," 2023.
7. A. Lakshmanan, et al., "Fairness and Explainability of Credit Scoring Models for Mortgage Loans," 2022.

8. F. Barbieri, et al., "Streamlining Mortgage Application Processing with Machine Learning-Based Credit Scoring," 2021.
9. M. O. Kotb, "Credit Scoring Using Machine Learning Algorithms and Blockchain Technology," 2024.
10. R. Chitty, "Development of Loan Default Prediction Model for Finance Companies in Sri Lanka," 2022.
11. M. Karim, "Improving Performance Factors of an Imbalanced Credit Risk Dataset Using SMOTE," 2022.
12. A. Byanjankar, "Predicting Credit Risk in Peer-to-Peer Lending with Survival Analysis," 2024.
13. G. Arutjothi, "Prediction of Loan Status in Commercial Bank using Machine Learning Classifier," 2017.
14. M. Annisa, "Prediction of Non-Performing Loans for Credit Application Analysis of Rural Bank Using Random Forest," 2022.
16. X. Sun, "Prediction of the Borrowers' Payback to the Loan with Lending Club Data," 2020.
17. A. Gicic'a, "Proposal of a model for credit risk prediction based on deep learning methods and SMOTE techniques for imbalanced dataset," 2024.
18. A. Malakauskas, "The Application of Artificial Intelligence Tools in Creditworthiness Modelling for SME Entities," 2021.
- 19.
20. Y. Liu, et al., "Developing an Enhanced Credit Scoring Model for Mortgage Default Prediction Using Machine Learning Techniques," 2020.
21. C. Xu, et al., "A Hybrid Credit Scoring Model for Mortgage Loans Using Machine Learning and Feature Engineering," 2019.
22. D. Thomas, "Credit Scoring for Mortgage Lending: A Review of Recent Developments," 2018.
23. "An Introduction to Banking: Principles Strategy and Risk Management," Wiley, 2018.
24. X. Zhang, Y. Yang, and Z. Zhou, "A novel credit scoring model based on optimized random forest," 2018.
25. "A Deep Learning Approach for Credit Scoring of Peer-to-Peer Lending Using Attention Mechanism LSTM," IEEE Access, 2019.
26. "An optimized credit scorecard to enhance cut-off score determination," South African Journal of Economic and Management Sciences, 2018.
27. "Weight of Evidence: A Review of Concept and Methods," Risk Analysis, 2005.
28. "Credit scoring statistical techniques and evaluation criteria: A review of the literature," Intelligent Systems in Accounting Finance Management, 2011.