IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 10, October 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.524

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 10, October 2024

|DOI: 10.15680/IJIRSET.2024.1310064|

Forecasting Loan Repayment Prediction System: Machine Learning Models and Implications

M Poornanivetha[1], Salma S[2]

Assistant Professor, Department of CSA, The Oxford College of Science, Bangalore, Karnataka, India[1]

PG Student, Department of CSA, The Oxford College of Science, Bangalore, Karnataka, India[2]

ABSTRACT: A Loan Repayment Prediction System leverages machine learning to assess the likelihood of a borrower repaying a loan, helping financial institutions mitigate risks and make informed lending decisions. The core of this system involves building predictive models using historical loan data, borrower profiles, and various socio-economic factors. These models analyze features such as credit score, income level, loan amount, employment history, and past repayment behaviour to estimate the probability of default or successful loan repayment.

The system typically employs algorithms like Logistic Regression, Decision Trees, Random Forest, or Gradient Boosting Machines to perform binary classification (repayment vs. default). Advanced models may incorporate deep learning or ensemble techniques to improve accuracy. Feature engineering plays a crucial role, as relevant borrower information and loan terms must be appropriately processed to improve model performance. Additionally, techniques like cross-validation and hyper parameter tuning are used to ensure the model generalizes well to unseen data.

Once trained, the model predicts repayment probabilities for new loan applicants. Lenders can use these probabilities to adjust interest rates, set loan limits, or reject high-risk applications, ultimately reducing the chance of loan defaults.

Key advantages of this system include automated decision-making, improved risk assessment, and the ability to process large datasets quickly. However, it requires careful handling of bias, interpretability, and regulatory considerations to ensure fair lending practices. Overall, a loan repayment prediction system can greatly enhance the efficiency and accuracy of lending decisions in the financial industry.

KEYWORDS: Decision Tree, Random Forest Classifier, Gradient Boosting

I. INTRODUCTION

The Loan Repayment Prediction System utilizing machine learning is designed to assist financial institutions in evaluating the likelihood of borrowers repaying loans, ultimately reducing the risk of defaults. With the increasing volume of loan applications, traditional risk assessment methods based on credit scores and manual reviews can be time-consuming and less effective. Machine learning offers a more sophisticated, data-driven approach, enabling automated, accurate, and efficient predictions of repayment behaviour.

The system relies on historical loan data, borrower demographics, credit history, income, and employment details to train predictive models. These models can detect patterns and relationships between various factors, allowing lenders to identify high-risk applicants more accurately. Common algorithms used include Decision Trees, Random Forest Classifier, and more advanced techniques like Gradient Boosting.

By providing insights into the probability of loan default or successful repayment, this system empowers financial institutions to make informed decisions, adjust loan terms, and optimize interest rates based on individual risk profiles. Moreover, it improves the overall efficiency of the loan approval process by automating decision-making and reducing human biases.

II. LITERATURE REVIEW

The Loan Repayment Prediction System using machine learning has been the subject of extensive research, with various studies focusing on different predictive models and features. Early work primarily used traditional statistical

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 10, October 2024

|DOI: 10.15680/IJIRSET.2024.1310064|

methods like Logistic Regression to model default risk based on credit scores, income, and employment history. However, these approaches often struggled to capture non-linear relationships and complex interactions between variables.

In recent years, machine learning techniques such as Decision Trees, Random Forests, Gradient Boosting Machines (GBMs), and Support Vector Machines (SVMs) have gained popularity due to their superior predictive power and ability to handle large datasets. These models can incorporate additional features such as behavioural data, transaction history, and macroeconomic indicators, improving prediction accuracy. Ensemble methods and hybrid models, combining multiple algorithms, have shown further performance enhancements.

III. METHODOLOGY

The methodology for a Loan Repayment Prediction System using machine learning involves several key steps. First, historical loan data is collected and preprocessed, including handling missing values, outlier detection, and feature scaling. Next, relevant features such as borrower demographics, credit scores, income, loan terms, and repayment history are selected or engineered. Various machine learning algorithms—such as Decision Trees, Random Forest, and Gradient Boosting—are then trained on the dataset to classify borrowers as likely to repay or default. The model is optimized through techniques like cross-validation and hyperparameter tuning, and its performance is evaluated using metrics like accuracy, precision, recall, and AUC-ROC.

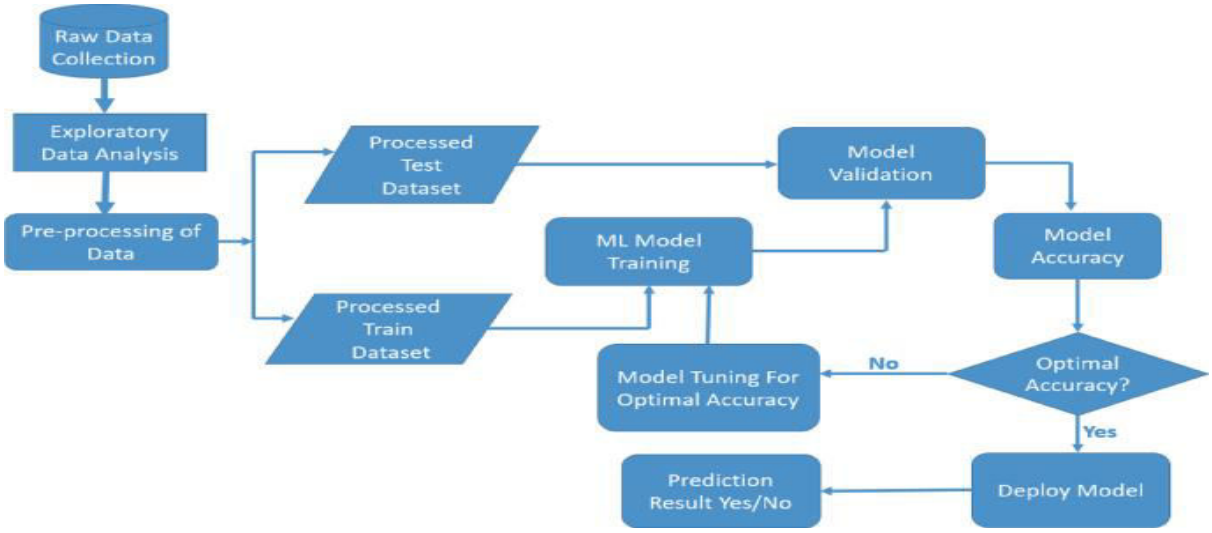

Fig. 1 Methodology Used

A. Data Collection

Data collection is a critical step in building a Loan Repayment Prediction System using machine learning. The quality and variety of data directly impact the model's ability to predict loan repayment accurately. Typically, data is sourced from financial institutions, including historical loan records, borrower profiles, and repayment histories.

Key data points include:

- **Borrower information**: demographics (age, gender, education), income, employment status, and credit scores.
- **Loan details**: loan amount, interest rate, loan type (secured/unsecured), repayment period, and prior loan applications.
- **Repayment history**: payment dates, missed payments, partial payments, or defaults.

In addition to borrower and loan-specific data, external data such as macroeconomic indicators (e.g., inflation rates, unemployment rates) may be integrated to capture external factors influencing repayment behaviour.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 10, October 2024

|DOI: 10.15680/IJIRSET.2024.1310064|

Data must be cleaned and preprocessed, ensuring completeness and consistency. Features with missing values can be imputed, and outliers must be handled appropriately to maintain model performance and prevent biases. This well-prepared dataset becomes the foundation for training predictive models.

B. Data Preprocessing

The collected data is often raw and may contain noise or missing values, so preprocessing is essential to prepare the data for model training. This stage involves:

- **Data cleaning**: Handling missing values, removing outliers, and normalizing the data.
- **Feature extraction and selection**: Identifying key features such as credit policy, installment, log.annual.inc
- **Labelling**: We need to convert the label values into numeric, it can be done using LabelEncoder.

C. Data Visualization

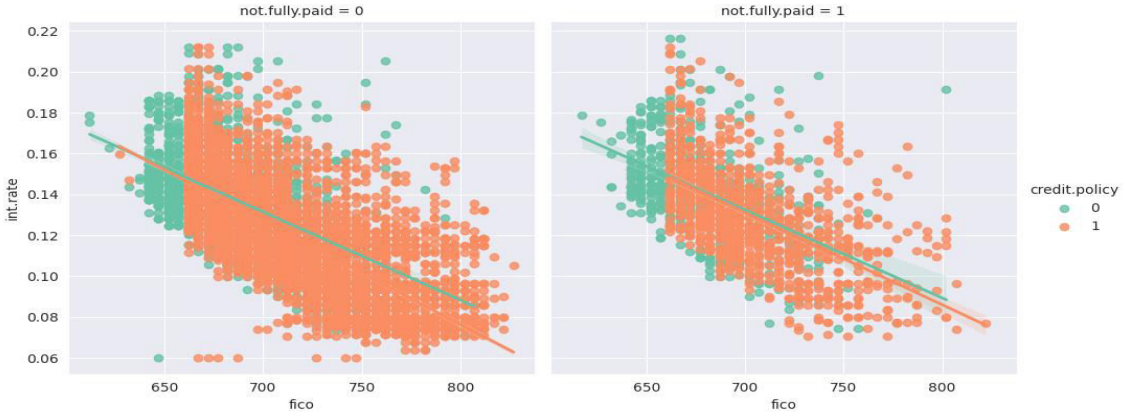

Fig. 2 Credit policy

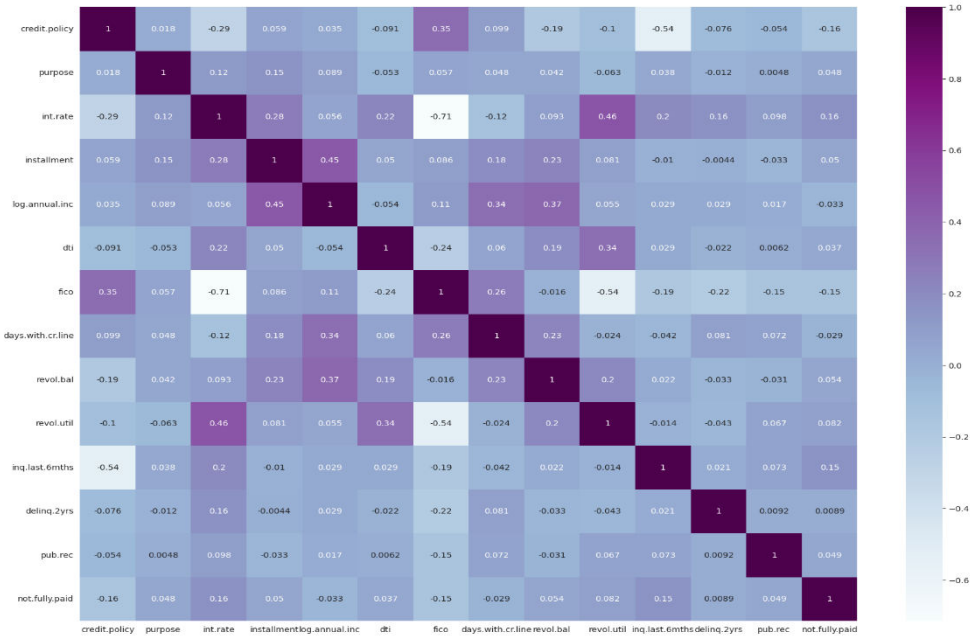

Fig. 3 Visualization of the loan repayment prediction dataset

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 10, October 2024

|DOI: 10.15680/IJIRSET.2024.1310064|

D. MODEL SELECTION

In a Loan Repayment Prediction System, selecting the right machine learning model is critical to ensuring high predictive accuracy. Various models can be employed, each with unique strengths. Here's an overview of key models:

1. **Decision Trees**: A simple, interpretable model that partitions the dataset based on feature values to predict whether a loan will be repaid or default. However, decision trees are prone to overfitting, particularly with complex datasets.

We got **Accuracy of 84.58%** using Decision Tree Classifier.

2. **Bagging with Decision Trees**: Bagging (Bootstrap Aggregating) reduces overfitting by training multiple decision trees on different subsets of data and aggregating their predictions. This helps in improving model stability and accuracy.

Bagging is not improving the score of model and giving only **73.10% of mean Score.**

3. **AdaBoost with Decision Trees**: AdaBoost (Adaptive Boosting) enhances decision trees by focusing on misclassified instances. It iteratively adjusts weights for incorrectly classified data points, leading to improved model accuracy.

It giving the same result of **84%** and not improving our Model.

4. **Random Forest Classifier**: An ensemble of decision trees, Random Forest reduces variance by averaging the predictions of many trees. It offers improved accuracy and robustness compared to individual decision trees, making it ideal for complex datasets.

We got the **Accuracy of 85%** with random Forest Classifier

5. **AdaBoost with Random Forest**: This combines AdaBoost's boosting strategy with Random Forest's ensemble approach, further improving accuracy, particularly in cases with imbalanced data or complex patterns.

We got **Accuracy of 84.7%** using this model.

6. **Gradient Boosting**: Gradient Boosting builds trees sequentially, each new tree correcting the errors of the previous one. It is highly effective but can be computationally expensive and more prone to overfitting if not properly tuned.

We got **Accuracy of 84.4%** using this model

E.Model training

- The training process involves splitting the data into training and test sets to avoid overfitting and ensure generalization.
- **Split Dataset**: Divide the dataset into training, validation, and test sets. Common splits are 70% training, 15% validation, and 15% testing.
- **Cross-Validation**: Use K-Fold Cross-Validation to train the model and avoid overfitting.

F. Experimental Setup

The experiments are conducted using Python and popular Deep Learning Frameworks. The data set is split into training, validation and test sets, with a typical split of 70% for training, 15% for validation and 15% for testing. Hyper parameter tuning is performed using grid search or random search methods to find the optimal configuration for the model.

G. Confusion Matrix

The Confusion Matrix is an essential tool to evaluate the model's performance. It provides insight into how well the model distinguishes between borrowers who will repay the loan and those who will default.

The Confusion Matrix has four key components:

1. True Positives (TP): The number of borrowers who were predicted to repay the loan and actually did repay.
2. True Negatives (TN): The number of borrowers who were predicted to default and actually defaulted.
3. False Positives (FP): The number of borrowers predicted to repay but ended up defaulting (Type I error).
4. False Negatives (FN): The number of borrowers predicted to default but actually repaid the loan (Type II error).

The confusion matrix is typically displayed as follows:

	Predicted: Repay	Predicted: Default
Actual: Repay	TP	FN
Actual: Default	FP	TN

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 10, October 2024

|DOI: 10.15680/IJIRSET.2024.1310064|

Metrics derived from the confusion matrix:

- Accuracy: (TP+TN)/(TP+TN+FP+FN)(TP + TN) / (TP + TN + FP + FN)(TP+TN)/(TP+TN+FP+FN)
- Precision: TP/(TP+FP)TP / (TP + FP)TP/(TP+FP)
- Recall (Sensitivity): TP/(TP+FN)TP / (TP + FN)TP/(TP+FN)
- F1 Score: The harmonic mean of Precision and Recall, providing a balance between them.

Confusion Matrix:

	Precision	Recall	F1 score	Support
0	0.85	1.00	0.92	2431
1	0.39	0.02	0.03	443
Accuracy		0.84	2874	
Macro avg	0.62	0.51	0.47	2874
Weighted avg	0.78	0.84	0.78	2874

IV. RESULTS

While Computing different Ensemble Learning Technologies, We Found that Most of the Bagging and Boosting algo are giving similar result with minimum difference in accuracy. Even though in all these Ensembles-
We Found that the Best Model for this Data Set is Random Forest with Accuracy of 85%.

V. INTERPRETATION

- **Accuracy** of 85% indicates that the model is correct 85% of the time across all predictions.
- **Precision** and **Recall** of approximately 84.7% suggest that when the model predicts a positive outcome, it is relatively accurate, and it successfully identifies about 84.7% of actual positive cases.
- The **F1 Score** of 84.7% reflects a good balance between precision and recall, making it suitable for applications like mental stress detection where both false positives and false negatives can have significant implications.

VI. CONCLUSION

A Loan Repayment Prediction System powered by machine learning offers financial institutions a data-driven solution to accurately assess the likelihood of borrowers repaying loans. By utilizing advanced algorithms like Random Forest, this system can effectively analyze borrower profiles, loan characteristics, and historical repayment patterns to predict default risk. Machine learning models, particularly ensemble methods like Random Forest, offer improved predictive accuracy, robustness, and the ability to handle large and complex datasets.

Overall, machine learning-based loan repayment prediction systems hold great potential for transforming the lending process, enabling automated, faster, and more informed credit decisions while mitigating financial risks and promoting sustainable growth in the financial sector.

VII. FUTURE SCOPE

The future scope of a Loan Repayment Prediction System using machine learning is promising, driven by advancements in technology and data analytics. As financial institutions increasingly adopt big data analytics, the incorporation of diverse data sources—such as social media activity, transaction histories, and alternative credit scoring models—can enhance predictive accuracy.

Moreover, the integration of real-time data processing capabilities will enable lenders to assess borrower risk dynamically, adapting to changes in economic conditions or borrower behaviour. The rise of explainable AI (XAI) will also improve model interpretability, fostering trust among stakeholders and ensuring compliance with regulatory requirements.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 10, October 2024

|DOI: 10.15680/IJIRSET.2024.1310064|

Additionally, incorporating advanced techniques such as deep learning and natural language processing can lead to more nuanced insights and better risk assessment. Ultimately, these advancements can lead to more efficient, transparent, and responsible lending practices in the financial industry.

REFERENCES

1. Antipov, E., & Bakaev, A. (2020) “A new approach to predicting credit default risk using machine learning.” Journal of Risk Management in Financial Institutions, 13(4), 325-335.
2. Do, H. G., & Vinh, T. H. (2020) “Applying machine learning techniques to predict loan default.” International Journal of Data Science and Analytics, 9(1), 55-64.
3. Tavakoli, M. (2018) “Credit Scoring Using Machine Learning Techniques: An Empirical Study on the Australian Consumer Loan Market.”
4. Kleinbaum, D. G., & Klein, M. (2010) “Logistic Regression: A Self-Learning Text.”
5. Churty, A., & Vangala, S. (2018) "Predicting loan default risk using machine learning techniques." Proceedings of the International Conference on Data Science and Advanced Analytics (DSAA).
6. Khandani, A. E., Kim, A. J., & Lo, A. W. (2010) "Consumer credit-risk models via machine-learning algorithms." Journal of Banking & Finance, 34(11), 2767-2787.
7. Malhotra, P., & Malhotra, S. (2017)."Loan Default Prediction: A Comparative Study of Machine Learning Techniques."International Conference on Advanced Computing Technologies (ICACT).
8. Yun, H. K., & Kim, H. (2019) "The effect of deep learning on loan default prediction." Expert Systems with Applications, 116, 10-20.
9. Data Sources- Kaggle Datasets Looked for datasets related to loan default or credit scoring which can be useful for building and testing models.
10. Paper - "Machine Learning for Credit Risk Modelling: A Review" A comprehensive review of machine learning applications in credit risk modellin

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details