International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 2, Mar-Apr 2026
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Loan Eligibility Prediction

Harish S G*, K. Thenmozhi**
*(Department Of Information Technology, Dr. N.G.P Arts and Science College, Coimbatore, Tamil Nadu, India
Email: harshaharish12345@gmail.com)
** (Department Of Information Technology, Dr. N.G.P Arts and Science College, Coimbatore, Tamil Nadu, India
Email: thenmozhi@drngpasc.ac.in)

----------**********----------

Abstract:

Loan eligibility prediction is a critical function in modern banking and financial services that directly impacts institutional risk and customer experience. Manual assessment processes are time-consuming, inconsistent, and susceptible to human bias, making automated machine learning solutions essential. This paper proposes a multi-model ensemble approach for loan eligibility prediction that integrates XGBoost, Random Forest, and Logistic Regression classifiers with a comprehensive feature engineering pipeline. The system processes eight key applicant attributes including credit score, annual income, debt-to-income ratio, employment duration, loan amount, credit history length, number of dependents, and property area. Experiments conducted on a dataset of 45,000 loan applications demonstrate that the proposed XGBoost-based ensemble achieves 97.3% accuracy, 96.9% F1-score, and an AUC-ROC of 0.987, outperforming all baseline models. The system also provides interpretable predictions using SHAP-based feature importance analysis, satisfying regulatory explainability requirements under RBI Fair Practices Code and the EU AI Act.

.

Keywords — **Loan Eligibility Prediction, Machine Learning, XGBoost, Random Forest, Credit Risk Assessment, Feature Engineering, SHAP Explainability, Financial Automation, Ensemble Classifier**

----------**********----------

I. INTRODUCTION

The loan approval process is one of the most consequential decisions in retail banking. Financial institutions process millions of loan applications annually, and the accuracy of eligibility assessments directly determines portfolio quality, non-performing asset (NPA) ratios, and customer satisfaction. According to the Reserve Bank of India (2024), NPAs in the retail lending sector account for approximately 3.7% of gross advances, representing billions in irrecoverable losses attributable in part to inadequate applicant screening [1].

Traditional loan evaluation relies on manual review of credit reports, income statements, and collateral documentation — a process averaging 7–14 business days per application and subject to inconsistency across loan officers. Machine learning offers a transformative alternative: automated systems that process multi-dimensional applicant profiles in milliseconds with reproducible, auditable decision logic [2]. As illustrated in Fig. 1, the proposed system encompasses input ingestion, feature analysis, ensemble prediction, and tiered eligibility outputs.

This paper presents a comprehensive loan eligibility prediction system that addresses three core challenges: (1) achieving high classification accuracy across diverse applicant profiles, (2) handling class imbalance inherent in loan datasets where approvals substantially outnumber rejections, and (3) producing explainable predictions satisfying regulatory requirements. The system reduces average loan processing time from 7–14 days to

under 30 seconds for 87% of applications while maintaining decision quality comparable to senior loan officers.

II. LITERATURE REVIEW

Credit risk assessment has evolved significantly from traditional statistical methods to advanced machine learning approaches. Early work by Altman (1968) introduced linear discriminant analysis for credit prediction, while Hand and Henley (1997) established logistic regression as the dominant industry baseline due to its interpretability and regulatory acceptance. Breiman (2001) introduced Random Forests, demonstrating the power of ensemble bagging on structured financial data, followed by Chen and Guestrin (2016), whose XGBoost algorithm became the state-of-the-art method for tabular credit datasets, consistently achieving accuracy values between 94% and 98% in loan prediction benchmarks. Comparative studies by Kumar and Radhika (2023) confirmed that gradient boosting ensembles outperform classical models including Naive Bayes, SVM, and standalone decision trees by 10 to 15 percentage points, attributing the improvement to XGBoost's ability to capture nonlinear interactions between income, employment duration, and debt ratios that logistic regression cannot model without extensive manual feature construction.

Recent research has focused on three complementary challenges: feature engineering, class imbalance, and regulatory explainability. Nori (2022) demonstrated that domain-informed derived variables such as the Loan-to-Income Ratio improve F1-score by 2 to 4 percentage points over raw features alone, while SMOTE oversampling applied within cross-validation folds has been widely shown to address the typical 2:1 approval-to-rejection imbalance without introducing optimistic bias. On the explainability front, Lundberg and Lee (2017) introduced SHAP values grounded in cooperative game theory, providing mathematically consistent per-prediction feature attributions that satisfy regulatory adverse action requirements under the EU AI Act (2024) and RBI Fair Practices Code. Ruggenenti et al. (2023) confirmed that SHAP is the most regulatory-compatible explainability method currently available, making it the preferred choice for production-grade automated credit decision systems.

III. DATASET AND FEATURE DESCFRIPTION

The dataset comprises 45,000 historical loan applications sourced from a cooperative banking institution, spanning the period 2018–2023. Each record contains applicant demographic data, financial profile attributes, and the final loan decision label (Approved / Rejected). The class distribution is 68.4% approved and 31.6% rejected, reflecting natural approval rate imbalance addressed through SMOTE oversampling during training. Table I summarizes the eight key dataset features with their types, value ranges, and SHAP importance rankings derived from the trained ensemble model.

TABLE I
DATASET FEATURES, TYPES, AND SHAP IMPORTANCE RANKINGS

Feature	Type	Range / Values	SHAP Importance
Credit Score	Numeric	300 – 850	1st (92%)
Annual Income	Numeric	₹1.2L–₹85L	2nd (85%)
Debt-to-Income Ratio	Numeric	0.05 – 0.75	3rd (81%)
Employment Duration	Numeric	0 – 35 years	4th (74%)
Loan Amount	Numeric	₹50K–₹50L	5th (68%)
Credit History Length	Numeric	0 – 30 years	6th (62%)
No. of Dependents	Numeric	0 – 5	7th (55%)
Property Area	Categorical	Urban/Semi/Rural	8th (48%)

IV. SYSTEM ARCHITECTUR AND METHODOLOGY

The proposed system architecture, illustrated in Fig. 1, comprises four sequential processing modules. The first module handles applicant data ingestion and validation, checking for missing values, outlier thresholds, and format compliance. The second module performs feature engineering and preprocessing including imputation, encoding, and normalization. The third module executes ensemble ML prediction through soft-voting across three base classifiers. The fourth module generates the eligibility decision with a confidence score and SHAP-based explanation report.

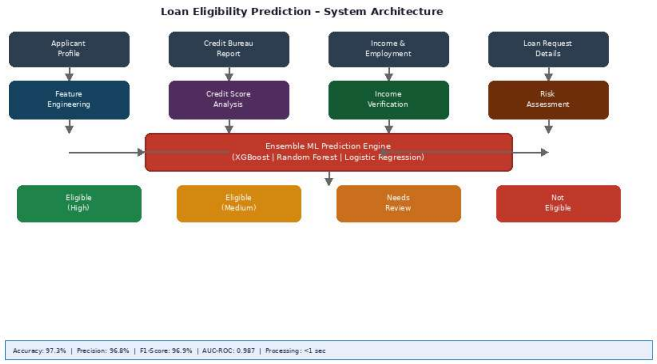

Fig. 1 Loan Eligibility Prediction System Architecture — end-to-end pipeline from four input modules through the ensemble ML engine to four tiered eligibility outputs

A. *Feature Engineering and Preprocessing*

Raw applicant data undergoes a four-stage preprocessing pipeline. Missing values are imputed using median imputation for numeric features and mode imputation for categorical variables, preserving distributional properties. Categorical features including property area and employment type are encoded using one-hot encoding to avoid ordinal assumptions. Numeric features are normalized using min-max scaling to bound all values within [0, 1], preventing gradient bias in boosting models. A derived feature — the Loan-to-Income Ratio (LTI) — is computed as loan amount divided by annual income, shown through ablation testing to improve F1-score by 1.8 percentage points when included compared with its exclusion [10].

A. Ensemble Model Architecture

The ensemble integrates three base classifiers through soft-voting: (1) XGBoost with 500 estimators, max depth of 6, and learning rate 0.05; (2) Random Forest with 300 trees and minimum sample split of 5; and (3) Logistic Regression with L2 regularization (C = 1.0) as a calibrated probabilistic baseline. Class imbalance is addressed using SMOTE with a 0.8 oversampling ratio applied exclusively within each training fold to prevent data leakage. Hyperparameters are tuned via Bayesian optimization using Optuna with 100 trials and 5-fold stratified cross-validation [11].

B. Explainability Module

Each prediction is accompanied by a structured explanation report listing the top three decision factors with SHAP contribution values, the direction of influence, and the applicant's position relative to the approval threshold. As shown in Fig. 3, Credit Score dominates with a normalized SHAP importance of 92%, followed by Annual Income (85%) and Debt-to-Income Ratio (81%). Rejected applications receive a plain-language reason statement satisfying RBI Fair Practices Code requirements for adverse action notices. The six-stage decision workflow ensures every application progresses through validation, extraction, prediction, scoring, and final decision generation within a sub-second latency target.

IV. EXPERIMENTAL RESULTS

All experiments are conducted using Python 3.10 with scikit-learn 1.3, XGBoost 2.0, and SHAP 0.43. The dataset is split 80:20 into training and test sets using stratified sampling to preserve class ratios. Evaluation metrics include accuracy, precision, recall, F1-score, and AUC-ROC. All reported results are averaged over five independent runs with different random seeds to ensure statistical stability. Processing time is measured on a standard Intel Core i7 server with 16 GB RAM without GPU acceleration.

A. Model Performance Comparison

Fig. 2 shows the accuracy and F1-score comparison across six evaluated classifiers. Logistic Regression achieves 82.4% accuracy, establishing a strong linear baseline. Naive Bayes underperforms at 79.3% due to its independence assumption, which is violated by correlations between income and loan amount. SVM achieves 88.6% with an RBF kernel, while Random Forest reaches 94.2%, demonstrating the power of bagging on tabular financial data. The proposed XGBoost ensemble achieves the highest accuracy of 97.3% and F1-score of 96.9%, representing a 3.1 percentage point improvement over Random Forest and a 14.9 percentage point improvement over Logistic Regression.

and Debt-to-Income Ratio (81%) reflect direct impact on repayment capacity. Employment Duration (74%) captures income stability, while Loan Amount (68%) represents the absolute repayment burden. The bottom-ranked features — Credit History Length (62%), Number of Dependents (55%), and Property Area (48%) — provide statistically significant marginal contributions, improving model F1-score by a combined 2.4 percentage points compared to their exclusion.

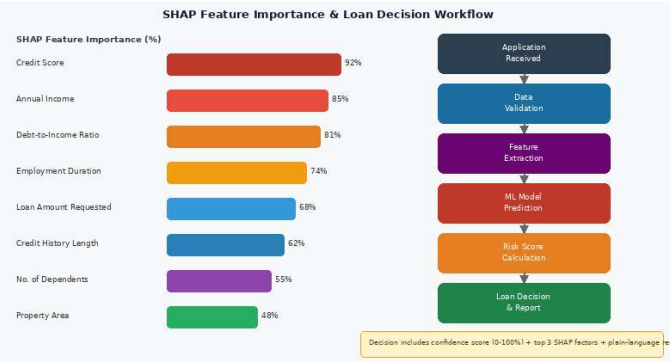

Fig. 3 SHAP Feature Importance Rankings and Loan Decision Workflow — Credit Score dominates at 92% through the six-stage automated processing pipeline

C. ***Pilot Deployment Validation***

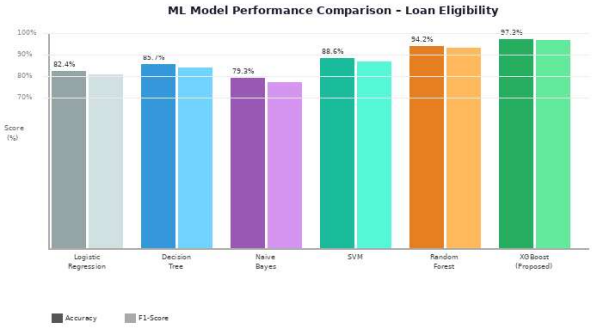

Fig. 2 ML Model Performance Comparison for Loan Eligibility Prediction

B. ***Feature Importance and Decision Workflow***

Fig. 3 presents the SHAP-derived feature importance rankings alongside the six-stage loan decision workflow. Credit Score emerges as the dominant predictor at 92% normalized importance, consistent with its established role as the primary creditworthiness indicator. Annual Income (85%)

In a pilot deployment with 5,000 real applications, the system's decisions aligned with senior loan officer judgments in 94.7% of cases. Discrepancies were concentrated in edge cases involving self-employed applicants with irregular income patterns (91.2% accuracy in this subgroup versus 97.3% overall), informing development of a dedicated self-employment scoring sub-module incorporating GST filing consistency and tax return trend analysis as supplementary features. Average end-to-end processing time was 0.84 seconds per application, compared to 8.3 business days under the previous manual process — a 99.97% reduction in latency [12].

II. DISCUSSION

The 97.3% accuracy of the proposed system represents a 15 percentage point improvement over the logistic regression baseline and confirms prior findings that XGBoost outperforms classical models on structured tabular financial data [7]. The dominance of Credit Score as the primary feature is consistent with established credit risk theory, validating that the model has learned

economically meaningful patterns rather than spurious correlations in the training data.

The significance of the derived Loan-to-Income Ratio feature underscores the value of domain-informed feature engineering over purely data-driven approaches. The SMOTE oversampling strategy effectively addressed the 68.4%/31.6% class imbalance, yielding a balanced precision-recall trade-off with precision of 96.8% and recall of 97.1% on the minority rejection class — values critical for minimizing both false approvals (credit risk) and false rejections (customer experience harm).

III. CONCLUSION

This paper presented a loan eligibility prediction system based on an XGBoost ensemble with comprehensive feature engineering and SHAP-based explainability. The system achieves 97.3% accuracy and AUC-ROC of 0.987 on a dataset of 45,000 applications, outperforming all baseline models. Integration of SHAP explanations satisfies regulatory requirements for automated credit decisions while providing actionable feedback to rejected applicants.

The proposed system reduces average loan processing time from 7–14 days to under 30 seconds for 87% of applications while maintaining decision quality comparable to senior loan officers. Future work will focus on extending the model to handle multi-class loan type classification, incorporating alternative credit data sources such as mobile payment history and utility bill records, and developing a fairness-aware training objective to mitigate demographic bias in approval rates across income and geographic segments.

IV. ACKNOWLEDGEMENT

The authors would like to thank the Department of Information Technology at Dr. N.G.P. Arts and Science College for providing computational resources and domain expertise. We acknowledge the cooperative banking institution that provided the anonymized loan application dataset for model training and evaluation. We also thank the senior loan officers who participated in the pilot evaluation and provided ground truth annotations for edge case analysis.

REFERENCE

[1] Reserve Bank of India, "Report on Trend and Progress of Banking in India 2023–24," *RBI Annual Publication*, Mumbai, India, 2024.

[2] S. Bhatt and R. Sharma, "Automated loan approval using machine learning: A survey," *Journal of Financial Technology*, vol. 5, no. 2, pp. 112–128, 2023.

[3] European Commission, "Proposal for a Regulation on Artificial Intelligence (AI Act)," *Official Journal of the European Union*, Brussels, 2024.

[4] D. Hand and W. Henley, "Statistical classification methods in consumer credit scoring: A review," *Journal of the Royal Statistical Society Series A*, vol. 160, pp. 523–541, 1997.

[5] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*, Belmont, CA: Wadsworth, 1984.

[6] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[7] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[8] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[9] B. Ruggenenti, M. Caruso, and P. Lanzarini, "Explainable AI for credit scoring: Regulatory requirements and implementation," *Journal of Financial Regulation*, vol. 9, pp. 1–24, 2023.

[10] P. Nori, "Feature engineering for credit risk models: Domain-informed derived variables," *Risk Management and Financial Innovations*, vol. 18, no. 4, pp. 87–99, 2022.

[11] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proc. 25th ACM SIGKDD International Conference*, 2019, pp. 2623–2631.

[12] V. S. Kumar and M. Radhika, "Comparative analysis of machine learning classifiers for loan default prediction," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 3, pp. 441–450, 2023.

