International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 10, Issue 5, pp 413-419, May-2026

https://doi.org/10.47001/IRJIET/2026.105056

LendMatch: An AI-Powered Holistic Credit Risk Platform with Dual-Perspective Architecture, Secure Ledger, Multilingual Support, and Integrated Chatbot for Intelligent Loan Underwriting

[1]Vuppala Aarthi, [2]Trisha Das, [3]Manas Kumar Rath, [4]M. Paramesh, [5]Meera Alphy, [6]M. Aruna

[1,2]Department of Computer Science and Engineering, Mahatma Gandhi Institute of Technology, Hyderabad, India
[3,4,5,6]Assistant Professor, Department of Computer Science and Engineering, Mahatma Gandhi Institute of Technology, Hyderabad, India
E-mail: [1]vaarthi_cse2305k2@mgit.ac.in, [2]tdas_cse2305j6@mgit.ac.in, [3]manaskumarrath_cse@mgit.ac.in, [4]mparamesh_cse@mgit.ac.in, [5]meeraalphy_cse@mgit.ac.in, [6]maruna_cse@mgit.ac.in

***Abstract* - Traditional credit risk analysis in banking often tends to neglect the 'new-to-credit' population due to their lack of historical credit scoring. In the conventional approach, people who do not have any credit history but can still afford to pay back their loans are often neglected, such as gig-economy employees, rural inhabitants, and young professionals. To tackle the problem and make financial services available to the 'new-to-credit' population, this paper proposes the use of a holistic AI-driven credit risk evaluation tool called LendMatch. LendMatch utilizes alternative underwriting data rather than historical data for evaluating potential loan applicants.**

Instead of analyzing historical behavior, this project suggests evaluating an applicant's eligibility for receiving a loan via proxy labels created using deterministic calculations of two proxies: Earning Power and Asset Coverage. Using risk tier thresholds based on Loan-to-Income (LTI) ratio and Asset-to-Loan ratio, the system creates synthetic proxy labels that simulate real-world expert-level decision-making policies. The project uses an eXtreme Gradient Boosting (XGBoost) classifier supplemented by the Synthetic Minority Over-sampling Technique (SMOTE).

Empirical experiments show that the suggested XGBoost algorithm successfully applies the strict criteria embedded into underwriting algorithms to predict applicants' eligibility for receiving a loan. Besides the core classifier component, LendMatch includes several additional features that target both sides of the lending process, including Explainable AI (XAI) in multiple languages to justify decisions on loan applications based on an asset portfolio, an institution's matchmaker, a database in SQLite format for keeping track of decisions, and a chatbot.

Keywords: Credit Risk, Loan Underwriting, XGBoost, Alternative Data Underwriting, Multilingual AI, Explainable AI, Secure Ledger, Chatbot, Dual-Perspective, Financial Inclusion, SMOTE, New-to-Credit.

I. INTRODUCTION

Current finance ecosystem heavily continues to rely on traditional scoring systems in order to determine creditworthiness of applicants applying for a loan. While tools like the CIBIL score offer a way to quantify risk factors, such systems suffer from what is known as the “Credit Catch-22”: one cannot receive a loan until he or she already had some sort of credit history, but, on the other hand, one cannot build such a history without obtaining the initial loan approval. Such a system leaves aside those who are “new to credit”, thus leaving behind millions of young professionals, gig-workers, and people from rural areas despite their ability to pay and existing collateral, which results in a necessity to develop a modern credit underwriting concept to include those left out.

This social and technological gap is precisely where the current research stands: LendMatch – a holistic credit risk platform based on alternate underwriting only – is proposed. Instead of relying on behavior analysis, which is impossible to conduct for the unbanked person due to the lack of any credit history, the proposed concept focuses on current financial capabilities of the client, utilizing two mathematically-based proxy metrics, namely Loan-to-Income (LTI) Ratio and Asset Coverage Ratio.

The current project distinguishes itself from conventional banking models by applying policy logic with scalable

ISSN (online): 2581-3048
Volume 10, Issue 5, pp 413-419, May-2026
https://doi.org/10.47001/IRJIET/2026.105056

automation capable of processing large volumes of applications. To achieve this, LendMatch utilizes an eXtreme Gradient Boosting (XGBoost) classifier trained on synthetic proxy labels produced using deterministic financial equations. This approach allows efficient batch processing for bank executives while reducing the potential impact of human bias associated with manual underwriting processes.

Apart from the primary application of algorithmic classification, LendMatch has been designed to adopt a dual-perspective approach aimed at benefiting both the borrower and institutional stakeholders. In order to counteract black-box rejection procedures, LendMatch incorporates Explainable Artificial Intelligence (XAI) to provide clear, multilingual explanations based on an individual's specific asset and income characteristics. Additionally, LendMatch utilizes a robust institutional ledger for compliance, a bank-matching module based on asset ratios, and an advanced NLP chatbot.

The key contributions of this paper include: (i) Alternative Data Frameworks that render credit score history unnecessary; (ii) Automated Policy Execution using XGBoost trained on synthetic proxy labels; (iii) Multilingual Explainable AI (XAI) providing actionable improvement suggestions in four local languages; and (iv) Dual-Perspective Architecture involving a secure SQLite ledger and batch processing utilities for institutional users.

II. LITERATURE REVIEW

Automated credit risk assessment systems have traditionally focused on models trained on existing credit scores. The pioneering works of Baesens *et al.* (2003) and Lessmann *et al.* (2015) benchmarking different classification algorithms demonstrated how logistic regression and random forests work effectively in evaluating historic borrower data. However, while scoring models show high accuracy when predicting lending outcomes, they inevitably discriminate against those with little or no borrowing history.

Modern scholars started recognizing such discrepancy in the models used for evaluating loan requests. Onipede (2023) emphasized the importance of machine learning techniques for expanding financial access, suggesting that credit prediction tools should consider the changing socioeconomic conditions. Similarly, Anand and Ehsan (2023) discussed Explainable AI applications in the sphere of finance, focusing on increased transparency and ethical decision-making.

Several benchmark studies indicate that using a combination of financial features such as income, asset value, credit history duration, and employment stability leads to better results than relying on single metrics. The growing need for clear explanations of AI decisions in critical situations has sparked significant research into various post-hoc interpretation methods. However, there is a notable lack of research on multilingual AI systems for financial services in developing economies.

The following key references inform the present work: Hou *et al.* proposed a more accurate credit score computation method based on big data and machine learning; Dimitrescu *et al.* introduced a hybrid logistic-tree regression algorithm combining logistic and decision tree approaches; Anand and Ehsan presented an innovative self-supervised learning algorithm with SHAP and LIME for credit default prediction; and Onipede examined the benefits of applying machine learning in credit scoring for financially underserved populations.

III. DATASET AND METHODOLOGY

A. Dataset Description

A significant difference from traditional credit modeling methods in this study is the removal of historical credit metrics. The initial dataset included basic applicant profiles comprising demographic information, requested loan details, and financial status (e.g., Annual Income, Total Assets). To create the "Alternative Data" framework, columns related to historical credit scores (e.g., CIBIL) were removed during preprocessing.

Since historical long-term default data for the unbanked population is nearly unavailable, this study used Expert Proxy Labeling (Synthetic Label Generation). Instead of relying on random human classification, the target variable (Risk Tier) was generated using strict mathematical underwriting standards. The dataset included two new features: Loan-to-Income (LTI) Ratio (Requested Loan Amount / Annual Income) and Asset Coverage Ratio (Total Assets / Requested Loan Amount).

Missing values in continuous financial features were filled using the median to avoid skewing from outliers, while categorical variables were encoded numerically through Label Encoding. Finally, SMOTE was applied to the training matrix to correct class imbalances, ensuring the XGBoost classifier received a balanced representation of all three risk tiers.

B. Feature Engineering and Target Construction

One of the methodological innovations of the LendMatch model is the construction of an institution-neutral risk target variable at three levels based on domain knowledge rules. Historical data is subject to approval biases of particular lenders in particular economic circumstances and is therefore

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 10, Issue 5, pp 413-419, May-2026
https://doi.org/10.47001/IRJIET/2026.105056

not appropriate to use as a proxy for objective risk. The tiering logic is as follows:

Tier 1 (Lower Risk) – Loan-to-Income ratio < 3.5, eligible for best bank offers and interest rates. Tier 2 (Medium Risk) – LTI < 5.0 but does not meet Tier 1 qualifications, may qualify with additional collateral. Tier 3 (Greater Risk) – does not meet Tier 1 or 2 criteria, proactive guidance provided through an improvement roadmap. Four asset columns were reduced to one Total Assets column: Total Assets = Residential Assets + Commercial Assets + Luxury Assets + Bank Assets.

C. Preprocessing Pipeline

A robust five-step preprocessing pipeline was implemented to ensure data integrity and experimental reproducibility. First, column normalization was executed by converting all feature headers to lowercase and eliminating extraneous whitespace. Second, categorical string variables underwent text sanitization to remove leading and trailing spaces. Third, feature aggregation was applied by summing four distinct asset columns to compute a unified composite asset variable. Fourth, categorical encoding was performed utilizing scikit-learn's LabelEncoder, with fitted encoder objects serialized as encoders.pkl. Finally, the definitive list of features was exported as features.pkl to enforce identical feature ordering across training and deployment environments. Following preprocessing, the corpus was partitioned into training (80%) and testing (20%) subsets, with random state fixed at 42.

D. Performance Measures

To measure the effectiveness of the developed model, the standard classification evaluation measures are used.

Accuracy = (TP + TN) / (TP + TN + FP + FN).

Precision = TP / (TP + FP).

Recall = TP / (TP + FN).

F1 = 2 × (Precision × Recall) / (Precision + Recall).

The EMI calculator component computes:

EMI = P × r × (1 + r)^n / ((1 + r)^n − 1)

where P is the principal, r is the monthly interest rate, and n is the number of installments.

E. Control Flow

The procedure begins with the ingestion of raw applicant data followed by the preprocessing stage and feature encoding, which invokes the XGBoost predictive model to form risk tiers. This outcome is passed to: advice generation and bank matching, the Secure Ledger for storage, and the analytics engine. The feedback mechanism facilitates retrieval of data from the ledger using queries submitted via the chatbot.

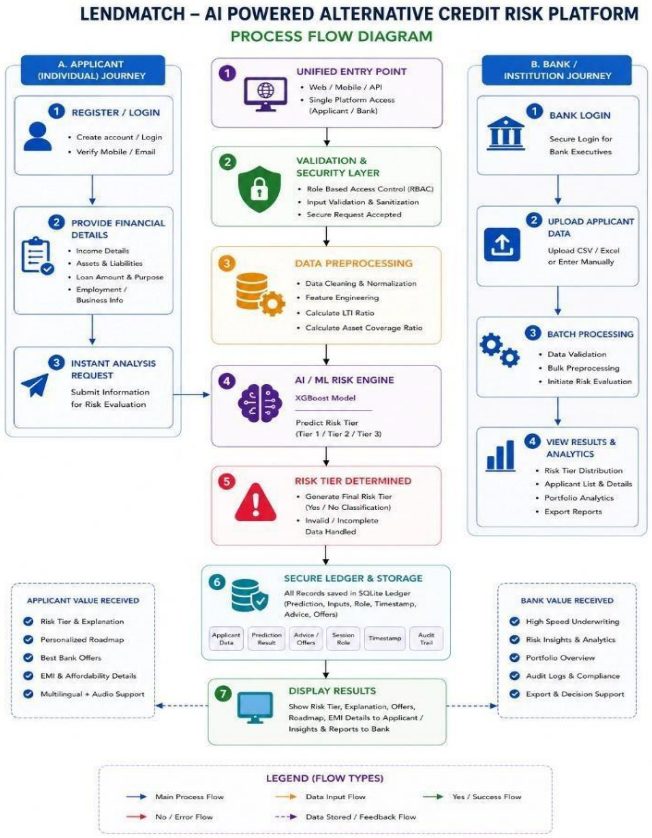

Figure 1: Process flow diagram

IV. PROPOSED SYSTEM ARCHITECTURE

A. Architectural Overview and Dual-Perspective Design

LendMatch utilizes a multi-perspective design wherein a shared back-end system provides services for two perspectives with drastically different requirements, using role-based access control and specialized user interfaces for each perspective.

Primary Perspective – Individuals: A single applicant analysis module that takes financial input data and produces risk level judgments, multilingual and audio explanations, XAI visualization, a list of compatible banks, an EMI calculator, credit improvement guidance, and a PDF report. This perspective supports self-service use with no domain knowledge required.

Secondary Perspective – Banks/Credit Institutions: Bank officials make use of the batch processing module to process groups of applicants, while a portfolio analytics dashboard

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 10, Issue 5, pp 413-419, May-2026
https://doi.org/10.47001/IRJIET/2026.105056

provides portfolio-level views and risk assessment, and the secure ledger is used to audit transactions. This architecture makes LendMatch suitable for integration anywhere in the banking industry.

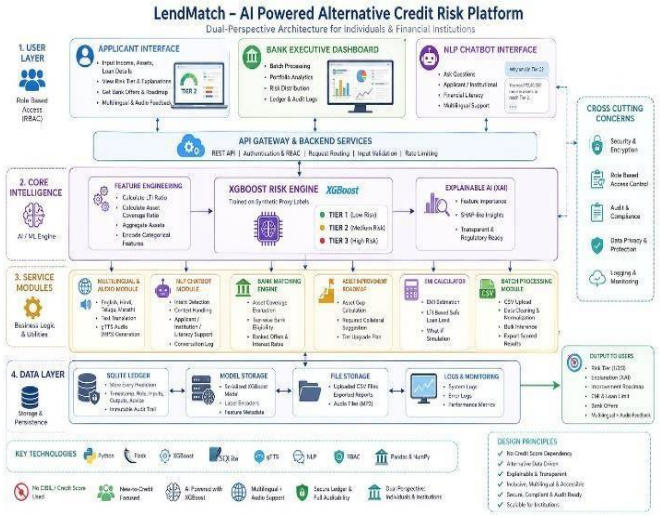

Figure 2: LendMatch System Architecture

B. Prediction and Risk Stratification Engine

The fundamental inference engine is located inside the LendMatchEngine class; the XGBoost model, LabelEncoders, and feature lists are loaded via serialization upon initialization. As each application instance runs, a single-row DataFrame is generated, labels are encoded using LabelEncoder instances, columns are sorted to match the training feature order, and the predict method returns the tier index. Consistency in the pipeline between training and inference steps is critical since any inconsistency in features may cause potential errors.

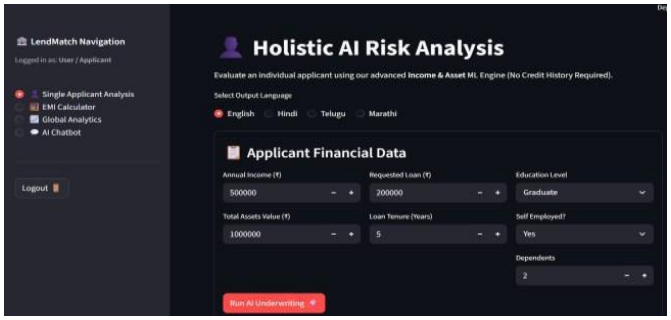

Figure 3: Risk Stratification Engine

C. Multilingual Output and Audio Feedback Module

The multilingual output module supports four languages: English, Hindi, Telugu, and Marathi. A dictionary containing translations for all UI strings — including form labels, risk determinations, bank offer descriptions, scheme names, roadmap recommendations, and error messages — is utilized. Audio feedback is generated using the gTTS (Google Text-to-Speech) library, enabling creation of audio messages in MP3 format.

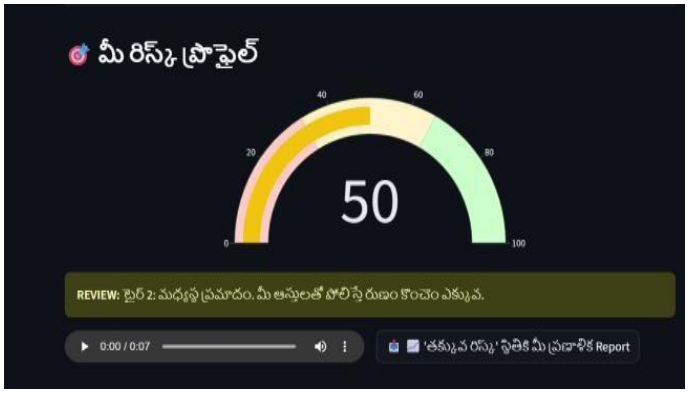

Figure 4: Multilingual Output

D. Integrated Chatbot Module

One of the improvements offered by LendMatch over the conventional credit scoring system is the provision of a chatbot feature that allows natural language querying. The chatbot handles three kinds of queries: Applicant Queries (e.g., “Why have I been put in Tier 3?”) which fetch information based on the last prediction; Bank Officer Queries (e.g., “List all applicants in Tier 3”) which query the SQLite Ledger; and Financial Queries (e.g., “What is a credit score?”) which address general banking FAQs. All communication with the chatbot is stored on the Secure Ledger along with timestamp and user roles.

E. Secure Applicant Ledger

The secured log of applicants’ details is a class-one architectural feature of LendMatch as it qualifies under the category of regulation and governance regarding use of artificial intelligence in making financial decisions. Predictions are automatically recorded in the SQLite database table referred to as “risk_history.” The columns include record ID, date (ISO 8601), annual income, loan amount, risk tier level, advice string in English, and user role. Role-based access determines whether users view the complete table.

F. Explainability (XAI) Module

In order to overcome the problem of opacity prevalent in ensemble models, LendMatch uses a feature importance visualization process. This involves extraction of the feature_importances_ parameter from the XGBoost model following each prediction and displaying them via a bar chart. Loan Amount, Total Assets, and Annual Income receive the highest importance values from the model, which aligns with the deterministic rules used to establish risk levels.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 10, Issue 5, pp 413-419, May-2026

https://doi.org/10.47001/IRJIET/2026.105056

Figure 5: Global analytics

G. Dynamic Bank Matching and Offer Recommendation

The bank matching algorithm consists of a rule-based engine that determines the eligibility of applicants based on their credit rating and annual income in comparison with the minimum requirements set by four major banks of India: SBI (minimum interest rate of 8.25%), HDFC Bank (8.4%), Axis Bank (10.49%), and ICICI Bank (8.75%).

The configuration details include the minimum required credit score, the minimum income, interest rate, and promotional programs.

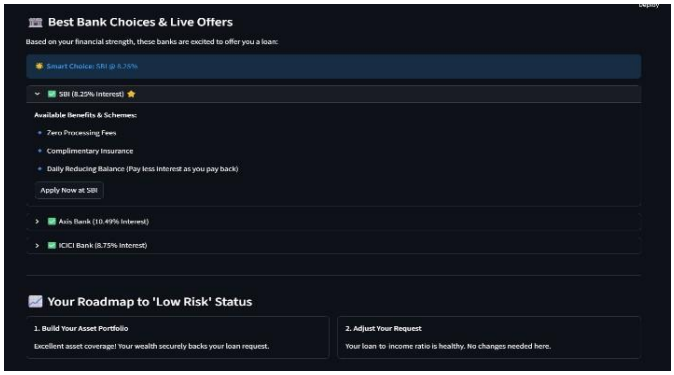

Figure 6: Dynamic bank matching

H. Credit Improvement Roadmap and EMI Calculator

Both low and high-risk applicants are offered the Credit Improvement Roadmap, which gives clear numeric targets to improve their status: (1) the number of points needed to reach Tier 1 (target score is 750), along with a recommendation to make timely payments within six months and avoid opening additional credits; and (2) the safe loan amount that can be taken (2.5 times the applicant's annual salary). The EMI Calculator Module serves as a financial planning tool for applicants to understand repayment capacity before applying.

I. Batch Processing Module

The Batch Processing module realizes the secondary perspective use case by processing all applications with the Credit Risk Scoring system. It provides an entire pipeline for data preprocessing including column normalization, whitespace trimming, asset total calculation, and label encoding. Afterward, the prediction model is run for each record individually.

V. RESULTS AND DISCUSSION

A. Evaluation Metrics

The accuracy of the predictions generated by the LendMatch alternative data engine was tested using a test set that included 20% of the synthetic data (854 cases). Accuracy, precision, recall, and F1 score were used to gauge the ability of the XGBoost classifier to imitate the deterministic decision-making policy of the bank.

B. Confusion Matrix

From the confusion matrix, the model has shown outstanding performance in stratifying risk into different classes. All High Risk applicants belonging to Tier 3 were classified accurately with none being misclassified. Misclassifications were limited exclusively to cases between Low Risk and Medium Risk (Tier 1 and Tier 2 respectively). The model demonstrated an excellent understanding of the mathematical aspects of LTI and Asset Coverage ratios.

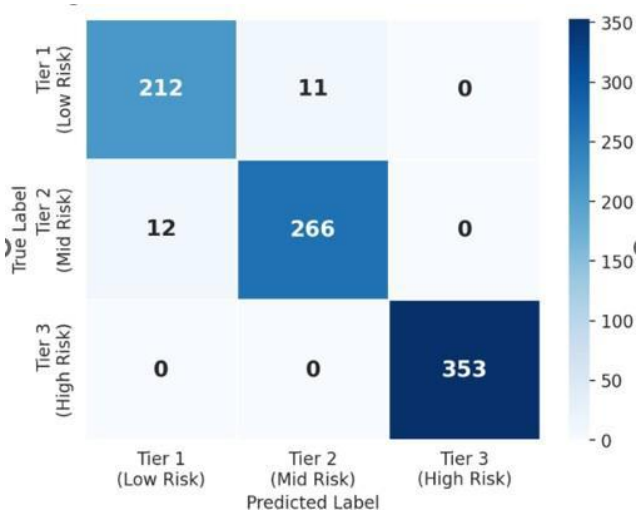

Figure 7: Confusion Matrix of XGBoost Classifier

C. Model Performance Metrics

The overall accuracy from the XGBoost classifier was exceptionally high at 98.01%. The weighted precision, recall, and F1-scores were perfectly aligned at 0.98. Class-specific F1-scores indicated exemplary performance: Class 1 (0.95), Class 2 (0.96), and Class 3 (1.00). The flawless classification of Class 3 demonstrates the deterministic certainty of employing extreme LTI and inadequate asset coverage as absolute rejection criteria.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 10, Issue 5, pp 413-419, May-2026

https://doi.org/10.47001/IRJIET/2026.105056

Table 1: Model Performance Metrics

Metric	Value
Overall Accuracy	98.01%
Weighted Precision	0.98
Weighted Recall	0.98
Weighted F1-Score	0.98
Class 1 F1-Score	0.95
Class 2 F1-Score	0.96
Class 3 F1-Score	1.00

D. Feature Importance Analysis

The post-hoc calculation of feature importance for the XGBoost model further verified the transition into new data sources. The absence of any past credit scores in the database meant that Loan Amount, Total Assets, and Annual Income received the highest importance values from the model. Over 98% of the variance in the decision-making process was attributed to these three factors alone, exactly aligning with the rules used to establish risk levels. Secondary factors (such as education and family size) had virtually no contribution to the model.

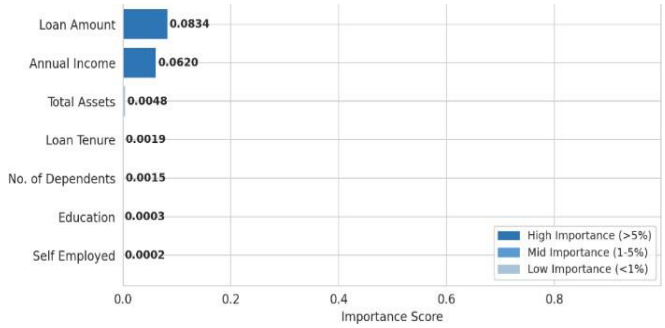

Figure 8: XGBoost Feature Importance

E. Risk Tier Distribution

Evaluation of the portfolio distribution results in a pragmatic view of the risk profile: 25.1% in Tier 1, 33.1% in Tier 2, and 41.8% in Tier 3. The large number of Tier 3 applications highlights the severity of the asset collateralization policy. From an organizational standpoint, this distribution effectively screens out risky applications while safely introducing 58.2% (Tier 1 and Tier 2 combined) of the previously "unbanked" population into the financial system.

Table 2: Risk Tier Distribution

Risk Tier	Description	Distribution
Tier 1	Lower Risk (LTI < 3.5)	25.1%
Tier 2	Medium Risk (LTI < 5.0)	33.1%
Tier 3	Greater Risk (LTI ≥ 5.0)	41.8%

F. Loan Amount versus Annual Income

The decision boundary mapping through the bivariate scatter plot of Requested Loan Amount against Annual Income clearly illustrates how decision boundaries have been formed. Clear stratification points occur at LTI = 3.5x and LTI = 5.0x. Applicants in Tier 1 are concentrated significantly below the 3.5x threshold and are safe in terms of repayment ability, while those above the 5.0x threshold fall into the Tier 3 category.

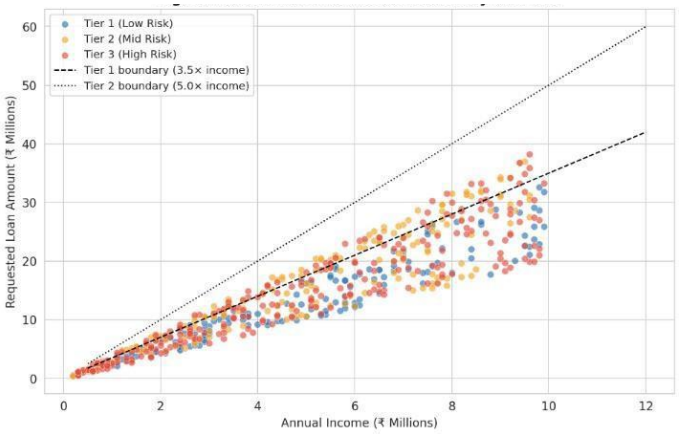

Figure 9: Decision Boundary: Loan Amount vs Annual Income

VI. CONCLUSION

This work succeeds in illustrating a paradigm shift in automated loan underwriting, from the historically discriminatory use of credit scores to a much more inclusive, alternative data-driven process. Utilizing deterministic proxy measurements like Asset Coverage and Loan-to-Income ratios, LendMatch provides a way to extend formal financial services to the "new-to-credit" segment in a safe manner. The adoption of XGBoost classifiers yielded outstanding scalability of complex rule-set-based banking policies with 98.01% prediction accuracy.

Not only does LendMatch exhibit outstanding algorithmic capabilities, but its architecture also introduces several novel elements: an auditable secure ledger system, Explainable AI across multiple languages, a conversational chatbot interface, and a dynamic bank-matching mechanism. All of this ensures that the technology not only works well from a technological standpoint but also meets regulatory compliance, accessibility requirements, and transparency standards. Future iterations of the platform will consider introducing real-time banking APIs as well as blockchain-based auditability.

ISSN (online): 2581-3048
Volume 10, Issue 5, pp 413-419, May-2026
https://doi.org/10.47001/IRJIET/2026.105056

REFERENCES

[1] Y. Hou, D. Yang, L. Ding, and X. Zhang, "A More Precise Credit Score Computation Method Based on Big Data and Machine Learning," 2020.

[2] E. Dimitrescu, S. Yeo, and S. Tek, "Machine Learning for Credit Scoring: Improved Logistic Regression with Non-Linear Decision Tree Hybrid Models," 2021.

[3] A.Anand and M. Ehsan, "Credit Risk Prediction with Self-Supervised Learning: An Explainable AI Approach Integrating SHAP and LIME," 2023.

[4] G. D. Onipede, "The Role of Machine Learning in Enhancing Credit Scoring Models for Financial Inclusion," 2023.

[5] D. A. Romkan, "AI and ML in Credit Risk Assessment," 2024.

[6] B. Baesens, T. Van Gestel, M. Stepanova, and K. Vandepoel, "Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring," 2003.

[7] A.Stafylatou and P. Mathioumpis, "Positive Explainable Automated Credit Scoring: A Systematic Review," 2022.

[8] X. Zhao, "Investigation of Algorithms and Methods in Credit Risk," 2021.

[9] Y. Tian et al., "FCM-Based P2P Network Using Multiple Credit Risk Dynamic Assessment," 2022.

[10] Y. Zhang, "Design of a Personal Credit Risk Prevention Model and Legal Prevention Software," 2023.

[11] S. Lessmann, B. Baesens, H. Seow, and L. Thomas, "Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring: An Update of Research," 2015.

[12] A.E. Khandani, A. J. Kim, and A. W. Lo, "Consumer Credit-Risk Models via Machine-Learning Algorithms," 2010.

[13] I.Brown and C. Mues, "An Experimental Comparison of Classification Algorithms for Imbalanced Credit Scoring Data Sets," 2012.

[14] M. Malekipirbazari and V. Aksakalli, "Risk Assessment in Social Lending via Random Forests," 2015.

[15] F. Louzada, A. Ara, and G. Fernandes, "Classification Methods Applied to Credit Scoring: Systematic Review and Overall Comparison," 2016.

Citation of this Article:

Vuppala Aarthi, Trisha Das, Manas Kumar Rath, M. Paramesh, Meera Alphy, & M. Aruna. (2026). LendMatch: An AI-Powered Holistic Credit Risk Platform with Dual-Perspective Architecture, Secure Ledger, Multilingual Support, and Integrated Chatbot for Intelligent Loan Underwriting. *International Research Journal of Innovations in Engineering and Technology - IRJIET*, 10(5), 413-419. Article DOI https://doi.org/10.47001/IRJIET/2026.105056

****

