SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

Explainable Alternative Credit Scoring For Thin-File Msmes Via Multi-Source Behavioral Data Fusion And Xgboost

Dr. Sita Yadav[1*], Dr Santwana Gudadhe[2], Dr. Pallavi Adke[3], Dr. Vishwas Kalunge[4], Dr. Khushal Khairnar[5]

[1*]Assistant Professor, Department of Computer Engineering, Army Institute of Technology, Pune, MH, India.
[2]Assistant Professor, Department of CSE (AI&ML), Pimpri Chinchwad College of Engineering, Pune, MH, India.
[3]Associate Professor, Dr. DY Patil School of Science & Technology, Dr. DY Patil Vidyapeeth, Pune, MH, India.
[4]Assistant Professor, Department of Computer Engineering, Dhole Patil College of Engineering, Pune, MH, India.
[5]Assistant Professor, Department of CSE (AI&ML), MIT Academy of Engineering, Pune, MH, India.
*Corresponding author. E-mail: syadav@aitpune.edu.in

Abstract
The issue of access to formal credits for Micro, Small, and Medium Enterprises (MSMEs) persists as a significant challenge due to the lack of a credit history for these businesses. The traditional method of evaluating such businesses using credit scoring models fails to do so. In this paper, an alternative credit scoring model using Artificial Intelligence and Machine Learning (AI/ML) techniques is proposed, and its evaluation is done. The alternative model uses the Small Business Administration (SBA) National loan dataset (899,164 records), application records (438,557 records), and credit history data (1,048,756 records) to enrich the borrower profile through feature engineering. The baseline model is a Logistic Regression model, which is compared to an advanced XGBoost model that has been trained on the combined data set. For handling severe class imbalances, Synthetic Minority Over-sampling Technique (SMOTE) is used only on the training data. The performance of the proposed model is validated experimentally, showing that it improves performance considerably. The proposed model has an AUC of 0.9646 and an F1 score of 0.9115. It outperforms traditional statistical approaches. The explainability of the proposed model is ensured by SHapley Additive exPlanations (SHAP) which helps to transparently identify critical factors such as delinquency ratio, credit history length, and business stability factors that impact credit risk. Finally, the proposed model is deployed as a REST-based API to score credit risk. The proposed approach shows that by using alternative behavioral data and explainable machine learning, it is possible to improve credit risk detection while promoting financial inclusion for thin-file MSMEs.

Keywords: Credit Scoring, MSME, Alternative Data, XG-Boost, SHAP, Financial Inclusion.

1. Introduction

IMSMEs are globally recognized to be key economic drivers, responsible for significant job creation and strengthening regional and world supply chains [16]. Despite their sizable economic footprint, a disproportional part of MSMEs is neglected by financial institutions in terms of financial assistance due to strict dependence on formal credit histories, audited financials, and long borrowing records [1]. Since new or small businesses lack such deep financials, they often fall into “thin-file” categories and fail to establish sufficient financial documentations for accurate statistical assessments and thus exhibit high rejection rates even with financially healthy profiles [9].

Logistic Regression-based bureau scoring has historically been the cornerstone of conventional credit scoring methodology; this technique assesses borrower creditworthiness and defaults by relying on highly condensed sets of structured variables [10], [18]. While interpretable, it doesn't consider the real-time financial health, operational stability and behavioral traits of modern businesses. This inelastic method makes loan officers unaware of the borrower's true credit profile, hence leads to acceptance of some inappropriate and rejection of some acceptable borrowers, causing credit gap for MSMEs [23]. The rapid evolution of digital transactions and storage capacity have led to increased accessibility of alternative data, providing new ways to assess borrower behavior [15], [31]. Modern Artificial Intelligence and Machine Learning algorithms can effectively analyze high-dimensional nonlinear relationships between alternative data sources [2]. Among such, tree-based ensemble methods such as XGBoost perform very well on classification tasks using mixed financial data [13]. However, there are practical challenges associated with implementing advanced ML models in real lending environments. The dataset for lending is highly imbalanced, as default is a rare event; class-wise balanced models can help address this bias issue [26]. Moreover, lending decisions require a high degree of explainability; such models have to be treated as "black-box" and explained using Explainable AI (XAI) techniques [20], [22]. Finally, deployment requires a scalable system architecture capable of real-time data acquisition, compliance verification, and inference [35]. The paper addresses these cumulative challenges by proposing a novel AI/ML based solution to true financial inclusion by lending to thin-file MSMEs using traditional Small Business Administration (SBA) financial data coupled with alternative firmographic and innovation data, such as number of patents and trademarks filed and public legal proceedings [17]. The model is implemented using an incremental development model, (M1–M6), ranging from basic credit scoring to implementation of bias mitigation and explainability components.

Key Contributions:

- **Data Multi-Source Architecture:** Integrates traditional SBA loan and application data with an extensive set of alternative data sources including firmographics (e.g., firm size) and innovation data (e.g., patents, trademarks, public legal records) for comprehensive borrower profile development [17], [31].
- **Advanced AI Engine with Class Balancing:** Develops an XGBoost based scoring engine that outperformed a Logistic Regression baseline [13], while using SMOTE to alleviate severe class imbalance on training data strictly [26].
- **Explainability and Transparency:** Implements SHAP [22] to provide local explanations for individual prediction results, allowing for easier interpretation by loan officers [20].
- **Scalable Microservices Deployment:** Develops a client-server-based system architecture featuring a React based frontend web server, a Fast API backend service, ML Inference service (internal HTTP/gRPC) and PostgreSQL database [6], [35].
- **Fairness and Bias Mitigation:** Strictly applies fairness constraints using algorithmic means [27], [28], including a definition of disparate impact and aims to keep model acceptance rates the same across different demographic groups to ensure fair lending.

2. Literature Review

The framework of assessing credit risk has shifted significantly; from static, rule-driven statistical algorithms to dynamic, data-driven machine learning architectures. The literature reviewed herein provides a critical assessment of the body of work pertaining to traditional methods, the deployment of modern ensemble algorithms, the taxonomy and use of alternative data, and the necessities of Explainable AI (XAI) and algorithmic fairness in current FinTech systems.

A. Traditional Credit Scoring Models and the "Thin-File" Problem: Historically, standard credit scoring frameworks primarily utilize simple statistical models like Logistic Regression because of their simple implementation, ease of deployment and lack of transparency required in most financial regulatory settings [18].

B. Standard algorithms inherently utilize a finite set of structured financial data inputs; a history of previous defaults, debt-to-income ratios, standardized credit scores, and more [10]. Although models like Logistic Regression can be sufficient when scoring individuals or large corporations with significant financial backgrounds, the literature explicitly states its inadequacy when assessing credit risk for MSMEs [9], [16]; MSMEs inherently operate within a "thin-file" data environ-ment, lacking extensive, formal, audited financial statements. Thus, standard statistical models default to risk, and cash positive MSMEs with potential for significant financial success are often rejected from lending because the systems cannot correctly interpret their financial standing [23].

C. Machine Learning Paradigms in Financial Risk Assessment: In order to remedy the strict, non-linear shortcomings of standard models and their poor acceptance rates, modern FinTech has adopted more complex machine learning architectures for credit risk [11], [19]. In particular, tree-based ensemble algorithms like Random Forests and Gradient Boosting Machines (GBM) have been shown to significantly increase prediction accuracy for loan defaults [12]. Unlike Logistic Regression, these algorithms can more closely map and identify complex, non-linear relationships between heterogeneous in-puts. Among these methods, Gradient Boosting Machines has gained traction. The Extreme Gradient Boosting (XGBoost) architecture has proven extremely effective in financial risk assessment, particularly for tabular data; the model's powerful regularization parameters prevent it from overfitting on the sparse financial data and it exhibits exceptional performance [13]. Across the spectrum of the available literature on the topic, it is broadly concluded that the use of these ensemble models yields improvements on all predictive accuracy metrics over simple, standard statistical models; lowering false positive rates in the assessment [11], [24].

D. The Emergence and Utility of Alternative Data: To combat the previously described "thin-file" problem, alternative data sources must be used to flesh out the available inputs. Much recent literature focuses on the importance of alternative data sources in financial lending; transactional data, payment patterns, cash-flow statements and application parameters have been repeatedly found to be highly correlated with business viability in real-time [14], [15], [31]. A higher-order set of alternative data inputs now emphasizes the importance of firmographic and innovative metrics. These include the business's legal status, date of incorporation, employee count, the number of patents and trademarks filed, as well as a record of the number of legal actions brought against the firm [17]. These pieces of information, readily available through Application Programming Interfaces like Google Patents and Open Corporates, allow institutions to develop an extensive, holistic proxy for a business's future viability [36]. These new alternative inputs are especially important for MSMEs in emerging markets or economies undergoing dynamic financial transition [23], [25].

E. Addressing Class Imbalance in Financial Datasets: A critical problem that has been addressed across multiple pieces in the literature involves how to properly handle severe class imbalance in a financial lending system. Credit default events are intrinsically rare and therefore the minority class within a dataset. Models that are trained solely on overall accuracy statistics will inherently favor the majority class (non-defaults). To overcome this inherent bias in the model, the literature suggests the use of powerful oversampling methods like the Synthetic Minority Over-sampling Technique (SMOTE) [26]. SMOTE helps balance the dataset by syn-thetically creating minority class training examples. While techniques like adjusting a model's scale_pos_weight parameter to penalize incorrectly predicted minority examples in the algorithm's cost function are also mentioned, SMOTE is considered the primary tool for data balancing within the lending community [19], [37].

F. Explainable AI (XAI) and Algorithmic Fairness: Although ensemble tree-based models boast a superior pre-diction accuracy, the "black-box" nature of their underlying algorithms violates strict regulations on transparent, explainable automated financial systems.

XAI frameworks are therefore a must for building both understandable and defensible decision-making processes [20]. Several studies highlight the critical role of SHapley Additive exPlanations (SHAP) as a widely adopted framework [22]. SHAP is an advanced algorithm for explaining and interpreting the individual output predictions of an algorithm in a localized feature space [38]. Alongside explainability, the literature frequently highlights the need for a fair algorithm; by increasing reliance on alternative inputs, proxy variables that are unintentionally correlated with sensitive demographic data points may enter the prediction process [28]. Thus, modern credit scoring systems must implement mathematical measures for privacy and fairness, including ensuring similar rates of loan approval for all demographics (G) through a constraint inspired by formal definitions of equalized acceptance rates [27]:

$$P(\hat{Y} = 1 \mid G = A) - P(\hat{Y} = 1 \mid G = B) < \epsilon \quad (1)$$

3. Proposed Methodology

To build an efficient, extensible, and well-integrated system that uses high-dimensional ML models for Fintech applications, we used a decoupled microservices architecture [35]. Here we describe the software development lifecycle, system architecture, and state transitions.

A. Incremental Software Development Lifecycle: We followed an incremental development lifecycle that was partitioned into 6 development stages (M1–M6) to account for the gradual integration of complex system features and to allow for phase-wise validation. This approach was suitable because machine learning systems require frequent modifi-cations and validation based on dynamic data and evolving algorithms. The increments were broadly categorized into three areas: Increment 1 focused on establishing a baseline credit scoring pipeline with traditional SBA financial datasets. Increment 2 expanded the feature set by integrating alter-native features from external APIs (innovation record, legal proceeding history, etc.). Increment 3 covered deployment of the Explainable AI module (SHAP) and development of automated fairness and bias mitigation features to comply with regulatory guidelines [7].

B. Microservices-Based Client-Server Architecture: Traditional monolithic applications face computational bottlenecks with the large ML inference workload. To address this, we adopted a Microservices based Client-Server architecture, separating computationally intensive ML inference from normal web routing and database management [35]. The system comprises four main nodes:

1) Frontend Web Server: Built with a lightweight React client-server framework served by Nginx/Apache. This interface helps loan officers submit MSME applicant data, interpret model output and feature importance charts, and access the financial analytics dashboard [6].
2) Backend Application API: A highly asynchronous FastAPI server (api_main.py) serves as the main orchestrator and API gateway. It handles incoming HTTPS/REST requests, manages data preprocessing, and orchestrates interactions with external APIs.
3) ML Inference Server: A containerized Python pro-cess dedicated solely to housing the trained XGBoost model (xgboostmodel.json) and the SHAP ex-plainer (shapexplainer.pkl). Communication be-tween the FastAPI backend and the ML Inference server is achieved through gRPC or internal HTTP requests to enable very fast inference response times [13].
4) Database Server: A TCP/IP based PostgreSQL Re-lational Database Management System (DBMS) with JDBC/ODBC access that stores historical SBA financial records and logs ongoing MSME application data.

C. Module Interfaces and External Integrations: The Backend API integrates with External Data Providers to ingest "thin-file" alternative data, thereby providing a richer borrower profile [36].

The system uses clearly defined pro-grammatic interfaces:

1) Client Interface: Used by loan officers to: a) submitapplication(jsondata) for ingesting applicant information; and b) viewdashboard(applicationid) to access the results and analytics of an MSME application.
2) Server Interface: Used by the system itself to: a) predictrisk(msmedata) for sending data to the ML inference engine; and b) getshapexplanation(prediction_id) to obtain localized feature importance scores for each prediction [22].
3) Data Interface: Used to call external APIs and fetch external information, e.g., fetchfirmographics(apikey) to the OpenCorporates API, and Google Patents API calls. Due to variations in public data, a fuzzymatchrecords(name, location) interface was developed to ensure correct matching of records.

D. System State Transitions: The processing flow is defined by the following six state transitions:

1) Data Ingestion: The system collates manually entered data and retrieved data from SBA financial datasets, as well as external datasets.
2) Preprocessing: Clean, impute, scale the collected features, and balance the dataset classes using SMOTE for an imbalanced ML task [26].
3) Inference: The cleansed and preprocessed feature vector is fed to the XGBoost engine for risk score calculation [13].
4) Explanation: The SHAP module is called to generate individual explanation of feature impact on the model's prediction [22].
5) Compliance Check: Ensure that the output adheres to pre-defined demographic fairness requirements [27].
6) Final Decision: The system provides the final classification outcome (Approved/Rejected/Manual Review) to the loan officer's interface.

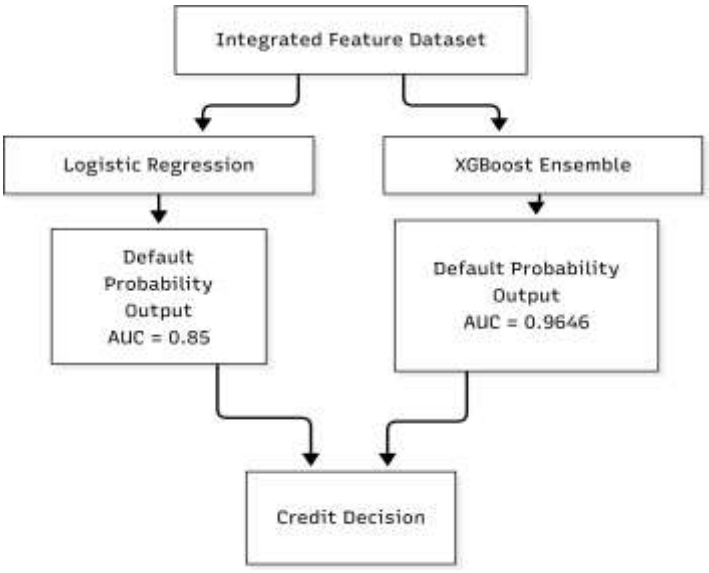

Fig. 1 Comparison of Logistic Regression and XGBoost Decision Pipeline

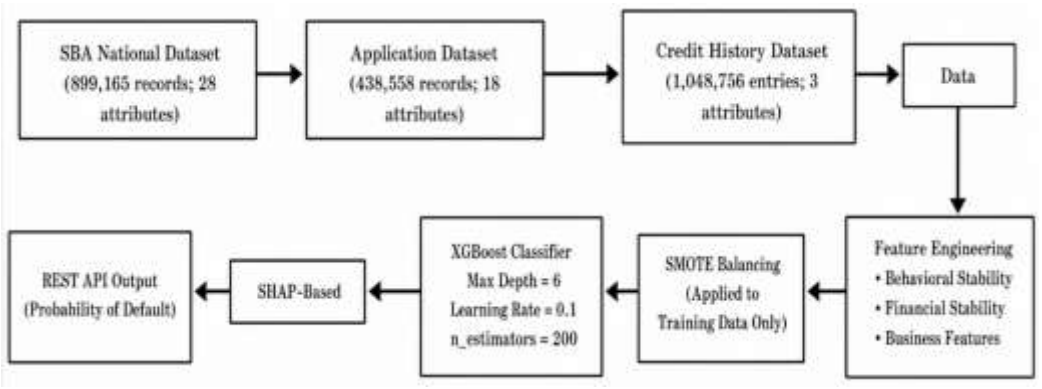

Fig. 2 Proposed End-to-End Credit Scoring Pipeline

4. Multi-Source Data Integration and Object Modeling

The accuracy of the proposed XGBoost scoring engine depends significantly on the structure and dimensionality of the data sources utilized [17]. This section covers the data collection strategy, the underlying object structure of the MSME Application Record and system classes.

A. Dataset Characteristics and Structural Integration: In order to present a holistic borrower profile the system is equipped to perform the massive structural integration of three disparate data sources: SBA National (a dataset of 899,164 loan records for the period of 2000–2018); Application records (a dataset of 438,557 entries representing demographic, financial, and employment related data of an individual or firm); and Credit records (a dataset of 1,048,756 entries indicating historical monthly payment statuses). There is no unified primary key that connects these three sources hence, a structured approach was developed where a synthesized Borrower ID was generated to bridge records (0 to N-1). The massive, class-imbalanced raw dataset has been collapsed into a manageable, balanced analytical dataset of 5,000 entries (2,500 default, 2,500 non-default). The targeted sample size was chosen to provide stability in initial algorithmic modeling and to eliminate significant class imbalances [26].

Table 1 Dataset Sources and Characteristics

Dataset	**Source**	**Records**	**Time Span**
SBA National	U.S. SBA	899, 164	2000-2018
Application Records	Open-source	438, 557	Unknown
Credit Records	Open-source	1, 048, 756	Unknown
Analytical Sample	**Integrated**	**5, 000**	**2000–2018**

B. Data Objects and Extended Attributes: The data that propagates through the system is a unique MSME Application Record. For evaluating "thin-file" businesses, the feature set is augmented considerably along three attribute dimensions [31]:

- Financial Attributes: Basic metrics that form the base-line of a Logistic Regression comparator including in-come, debt burden, payment history, and credit usage.
- Firmographic Attributes: Business oriented, structural metrics such as business age, legal status, number of employees etc., that would typically suggest business stability and performance [4], [5].
- Alternative Data Attributes: The high-value innovative and legal indicators that can only be gleaned through ex-ternal API calls including the count of patents, trademarks and any past or ongoing legal proceedings [17], [36].

C. Object and Class Responsibilities: Following the object-oriented design principle, individual system classes hold specific responsibilities:

- Loan Officer (Client Node): Represents the user that interacts with the system. This object is responsible for initializing MSME credit evaluation requests and manually reviewing and confirming AI outputs from the financial analytics dashboard.
- Data Preprocessor: The heavy computational workhorse. This object is entirely responsible for data cleaning, normalizing continuous features, imputing complex missing values [30], and executing SMOTE to balance classes before inferring [26].
- XGBoost Model (Inference Node): The computational engine. This object encapsulates the pre-trained decision tree based ensemble model and uses a multi-dimensional feature vector to produce a binary classification prediction (default/non-default) [13].
- SHAPExplainer (Interpretability Node): The XAI add-on to the predictive engine. It systematically decomposes the XGBoost model output, provides a localized interpre-tation and computes Shapley values indicating the relative contribution of each input feature to the final prediction [22].
- BackendAPI (Gateway Node): The single entry point for requests. It manages all requests coming from the front-end, securely fetches additional information from third party data providers, and dictates the sequence in which the preprocessing, inference, and explanation modules are invoked.

To enable a highly predictive machine learning architecture to ingest multi-source, raw transaction data, a thorough feature engineering pipeline is required [3], [21]. As conventional financial data (e.g., audited cash-flows or long-term debt to income ratio) are often non-existent or stale for MSMEs, deriving dynamic behavioral metrics from raw credit logs is of paramount importance. In this section, the system's trans-formations, imputations, and aggregations of raw transactional data into an optimal feature vector for the XGBoost engine is discussed.

D. Structural Preprocessing and Feature Imputation: Once the application demographics, SBA National loan information, and open-source credit records were linked using the synthetic Borrower ID, the raw dataset produced inherently this issue [30]. By examining k closest MSME profiles based on existing Euclidean distances, it calculates a weighted average from those k neighbors, imputing the missing value while preserving the multi-dimensional variance structure of the feature space. 2) High Cardinality Categorical Encoding: Many categorical features (e.g., the application status, business's industry classification codes, the operational location) have high cardinality and are incompatible with models that assume low cardinality.

Simple One-Hot Encoding of such features is impractical, since it creates hundreds of binaries, sparse columns that severely inflate the feature space, introducing the "curse of dimensionality" and rendering tree-based models impractical. Target Encoding replaces each categorical feature value with the historical mean target value for that category (in our case, the mean default rate) [29]. This transforms the categories into continuous values, which can be natively processed by the XGBoost algorithm while keeping the feature vector compact and dense.

E. Behavioral Mathematics and Feature Extraction The raw credit logs provide time-series, month-by-month records of an MSME's repayment behavior. However, a single classifier cannot ingest individual time-series data points and must work on cross-sectional, entity-level features. Therefore, the system aggregates these time-series behaviors mathematically into individual, discrete behavioral features that represent the current financial health and repayment consistency of the enterprise [3], [14], [21]. For each individual borrower, the time-series includes months with discrete states:

a. 0 = on-time payment
b. C = closed or fully paid
c. Integers 1 to 5 represent levels of increasing delinquency severity (e.g., 1 = 30 days overdue up to 5 = 150+ days overdue/charge-off). These states are used to derive the following specialized behavioral features.

Delinquency Ratio (*overdue_months_pct*): This metric captures the overall proportion of time spent in any degree of delinquency by an enterprise throughout its active credit history. A high delinquency ratio is reflective of significant operational strain and inconsistent cash flow:

$$OverdueMonthsPct = \#(t : s\_t \in \{1,2,3,4,5\}) / T\_total \quad (2)$$

where s_t denotes the payment status in month t and T_{total} is the total number of months recorded. Even if loans are ultimately repaid, frequent past late payments represent a higher degree of operational risk.

Default Count (high-risk months): While the delinquency ratio reflects even minor lateness, the Default Count focuses on high-severity periods. It counts the absolute number of months in severe delinquency (defined as $s_t \geq 3$):

$$\text{high-risk months} = \{t : s_t \in \{3, 4, 5\}\} \quad (3)$$

This metric prevents an enterprise with few but long-lasting severe defaults from being evaluated identically to one with frequent but minor late payments (e.g., occasional 30-day overdue).

Repayment Consistency (payment consistency): Risk assessment is not solely a matter of penalizing negative behavior; positive financial management also increases confidence in a borrower. This metric assesses the proportion of months with perfect repayment:

$$\text{payment consistency} = |\{t : s_t \in \{0, C\}\}| / T_{total} \quad (4)$$

A consistently punctual borrower over a long credit history likely possesses strong cash-flow and financial management skills, and a longer perfect record serves as a negative input toward the model's probability of default.

Credit History Length (credit history months): The simple count of monthly records for a given borrower, T_{total}, is included as a fundamental temporal variable.

As the law of large numbers dictates, longer time-series data allow the algorithm to assign a higher degree of statistical confidence to all derived behavioural metrics; by increasing the sample size used to calculate each feature, variance is inherently reduced.

5. AI/ML Credit Scoring Engine and Mathematical Model

The predictive engine at the core of the proposed architecture relies upon Extreme Gradient Boosting (XGBoost), which represents an optimized implementation of gradient-boosted decision trees capable of maximum execution speed and top-tier model performance [13]. Where traditional statistical techniques, such as Logistic Regression, are unable to model the complex non-linear relationships between multi-source alternative financial data, XGBoost inherently possesses this ability [11], [24]. Herein, the objective function governing the training of the system is formally outlined, the powerful regularization mechanisms implemented are thoroughly explained, and the algorithmic complexity is critically analysed for feasibility in a real-time FinTech setting.

A. The XGBoost Objective Function and Regularization: The decision to leverage XGBoost as the predictive engine within the proposed framework stems from its robustness in dealing with tabular, heterogeneous data and its intrinsic sparsity-aware split-finding algorithms [13]. Where the thinly populated MSME datasets consist of a multitude of missing values (e.g., for a business which has filed no patents and has never been party to a lawsuit), XGBoost learns an optimal default direction for any missing values naturally through training, avoiding the need for aggressive data imputation. The fundamental objective of the ML engine is to predict MSME default risk by minimizing the objective function $L(\Phi)$ during training, defined as:

$$L(\Phi) = \Sigma_{i=1}^{n} \ell(\hat{y}_i, y_i) + \Sigma_{k=1}^{K} \Omega(f_k) \quad (5)$$

where:

- T_k is the number of terminal leaf nodes in tree f_k.
- $\gamma \geq 0$ is a minimum loss-reduction threshold required to allow a node split, controlling tree depth.
- w_j is the output weight (score) of the j-th leaf.
- $\lambda \geq 0$ is the L2 regularization coefficient applied to leaf weights.

The Loss Function $\ell(\hat{y}_i, y_i)$: This term quantifies model accuracy. Given that credit scoring is a binary classification problem (Default vs. Not Default), the system uses binary cross-entropy loss, which measures the difference between the predicted probability $\hat{y}_i \in [0, 1]$ and the ground-truth label $y_i \in \{0, 1\}$. Using a continuous probability output rather than a discrete label gives lenders the flexibility to adjust their risk tolerance by tuning the decision threshold.

$$\ell(\hat{y}_i, y_i) = -[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

The Regularization Term $\Omega(f_k)$: This term penalizes the complexity of the k-th tree to prevent overfitting. The regularization term is defined as:

$$\Omega(f_k) = \gamma T_k + \tfrac{1}{2}\lambda \Sigma_{j=1}^{T_k} w_j^2 \quad (6)$$

By controlling model complexity through γ and λ, the XGBoost engine resists overfitting to noisy alternative MSME data (such as fuzzy-matched public legal proceedings). The L2 regularization penalty is fundamental to ensuring successful generalization from training to real-world data [13].

B. Algorithmic Complexity, Scalability, and NP-Completeness Analysis: Hardware efficiency, latency, and the ability for a FinTech application to perform at scale represent key design aspects. The XGBoost training and inference complexity make it well-suited for the microservices-based REST architecture.

Training Complexity: The time complexity required to train the XGBoost model is:

$$O(K \cdot d \cdot x_o \cdot \log n) \quad (7)$$

where K is the total number of boosting iterations (fixed at 200), d is the maximum tree depth (d = 6), x_o denotes the number of non-missing feature values in the training dataset, and n is the number of training samples. Since XGBoost uses a sparsity-aware algorithm, training time is proportional to non-missing entries ($x_o \ll n \cdot p$ for p features), making the operation hardware-efficient.
Inference Complexity and Latency: In an API-based setting requiring real-time predictions, latency is critical. Generating a prediction requires traversing all K trees to depth d:

$$O(K \cdot d) \quad (8)$$

With K = 200 and d = 6, the upper bound is only 1,200 node comparisons per prediction—highly efficient for modern computing architectures, enabling the FastAPI backend to execute a full risk classification in milliseconds.

NP-Completeness Justification: Although learning an optimal decision tree is an NP-Complete problem in the general case [13], this is irrelevant here: XGBoost frames tree construction as a continuous optimization problem solved greedily in polynomial time using gradient descent, rather than as a combinatorial decision problem.

Class Imbalance and Fairness Optimization: The implementation of sophisticated machine learning algorithms in a highly regulated financial industry necessitates concurrent and stringent solutions to structural data abnormalities and robust ethical considerations [19], [28]. This section details the specific techniques devised to circumvent the serious class imbalance inherent to MSME credit scoring and the explicit mathematical constraints utilized within the system architecture to uphold fairness, diminish disparate impact, and ensure regulatory adherence.

C. Overcoming Class Imbalance with the SMOTE Algorithm: One of the enduring structural challenges in financial risk modeling is the natural scarcity of defaults in the historical loan portfolio. In any financially sound lending environment, the number of default events is significantly lower than non-default repayment events. The consequence of this class imbalance is that machine learning models trained on raw, unadjusted data will inevitably develop a heavy bias toward the majority class. A classifier on imbalanced data may appear to have an acceptably high global accuracy by simply predicting "non-default" for most applications. However, such a model utterly fails in its primary purpose: accurately identifying risky entities to prevent loss of lending capital [37]. To avoid model bias toward the majority class, the proposed system architecture employs the Synthetic Minority Over-sampling Technique (SMOTE) [26]. Unlike simple duplication, which causes overfitting by memorizing identical data points, SMOTE artificially generates mathematically unique minority-class instances. It operates in the high-dimensional feature space, identifying the k-nearest minority neighbors to a given minority-class observation based on Euclidean distance.

A neighbor is randomly chosen, and a new synthetic data point is created through interpolation along the connecting vector between the two points in feature space. This effectively expands the decision boundaries of the minority class in a statistically robust way. For a scientifically valid assessment, SMOTE was applied only to the training dataset, with the test set left completely unaltered to simulate the real-world environment of highly imbalanced loan applications. This separation is crucial to avoid data leakage, which would artificially inflate evaluation metrics [19].

To further ensure scientific integrity, a baseline comparison was established by varying the scale_pos_weight parameter in XGBoost, which offers an algorithmic, rather than data-resampling-based, approach to class balancing. Through empirical observation, SMOTE definitively outperformed this approach, producing a more stable learning scenario with an isolated recall of 0.8740 and only 73 false positives during tuning.

D. Formal Fairness Constraints and Automated Bias Mitigation: Although additional behavioral and firmographic data boosts model predictive accuracy, it also introduces a significant threat of algorithmic bias [28]. Even without direct protected attributes such as race, gender, or religion in the dataset, alternative credit scores can lead to unfair lending outcomes via proxy variables. Highly sophisticated ensemble models like XGBoost can discover proxy variables, such as geographic location or specific industry codes, that heavily correlate with protected groups. Unmitigated use of these proxies may lead to discrimination on the basis of protected attributes [28]. The system architecture pre-empts this risk with an explicit "Compliance Check" state in the system sequence, echoing broader calls for auditable frameworks that embed non-financial risk criteria directly into automated decision pipelines [7]. After generating a risk prediction, but before providing the final decision, the system executes an automated fairness and bias check. The mathematical fairness constraint implemented by the backend microservice, drawing on formal definitions of equalized treatment across groups [27], is:

$$P(\hat{Y} = 1 \mid G = A) - P(\hat{Y} = 1 \mid G = B) < \varepsilon \quad (9)$$

where P denotes the conditional probability calculated from aggregated predictions on a sample batch, $\hat{Y} = 1$ represents a predicted loan default, G = A and G = B denote two groups being assessed, and ε represents the maximum tolerable deviation in rejection rates set by the lending institution's regulatory team. If any group's automated rejection rate violates this inequality during the Compliance Check, the system ceases the automated rejection process and routes the application for manual review by a Loan Officer, thereby preventing the systematic rejection of any sub-population within the MSME sector.

6. Experimental Setup and Evaluation Metrics

In order to rigorously and scientifically evaluate the predictive accuracy of the proposed XGBoost framework against traditional statistical baselines, a robust, highly controlled experimental setup was employed [11], [24]. The standard machine learning evaluation procedures used for balanced data fail to appropriately address and mitigate class imbalances common in financial datasets. The subsequent section provides a thorough review of the sampling framework, the hyperparameter optimization process, and an advanced array of threshold-independent evaluation metrics that effectively capture the model's ability to predict true credit risk.

A. Data Partitioning and Stratified Sampling Framework: An optimized sample size of 5, 000 comprehensive MSME records were utilized for analysis and then segregated in a way that preserved the temporal validity of the evaluation set.

The data was divided into dedicated training and testing sets using an 80-20 sampling approach, whereby 4,

000 records were designated for training and the remaining 1, 000 records were reserved for final model evaluation. Given the severe class imbalance in the analytical dataset, simple random sampling would not be statistically sound, as the smaller testing set might contain too few default cases for reliable evaluation. To mitigate this, a stratified 80-20 sampling approach was used, ensuring both partitions precisely reflect the actual class distribution [26]. SMOTE over-sampling was applied only to the training set, leaving the test set as a non-manipulated, realistic evaluation set.

B. Hyperparameter Optimization and Model Tuning: To ensure an optimal predictive fit without overfitting to training data, k-fold cross-validation was performed with vary-ing parameter values to select those that optimize the MSME credit scoring engine's performance [13]. The optimized pa-rameter settings chosen were: max depth = 6; learning rate $\eta = 0.1$; n estimators = 200; Classification Threshold = 0.5 for initial testing.

C. Advanced Evaluation Metrics for Imbalanced Data: The default classification threshold of 0.5 is insufficient when analyzing extremely imbalanced data because it unfairly benefits the dominant class [19]. This led to the inclusion of advanced, threshold-independent metrics:

1) **Precision, Recall, and F_1-Score:** For imbalanced classification, standard accuracy metrics are replaced by Precision, Recall, and F_1-Score evaluated specifically for the default class. Precision is the ratio of True Positives to all predicted positives; Recall captures the proportion of actual positives correctly identified; F_1-Score is their harmonic mean, providing an optimal trade-off between the two [24].
2) **ROC-AUC (Receiver Operating Characteristic Area Under the Curve):** The ROC curve evaluates model separability by plotting the True Positive Rate against the False Positive Rate at varying thresholds [11]. An $AUC > 0.9$ indicates outstanding discriminative perfor-mance.
3) **PR-AUC (Precision-Recall Area Under the Curve):** Unlike ROC-AUC, the Precision-Recall curve measures the trade-off between Precision and Recall only, which is especially informative for severely imbalanced datasets where positive samples are rare [37].
4) **Brier Score and Probability Calibration:** Financial lending applications require not just binary predictions but accurate probability estimates. The Brier Score mea-sures the mean squared difference between the predicted probability and the actual binary outcome [38]; smaller values indicate better-calibrated probability outputs.

Table 2 Model Hyperparameters

Parameter	Value
Train-Test Split	80:20 Stratified
Max Depth	6
Learning Rate (η)	0.1
Estimators (K)	200
Evaluation Metrics	Accuracy, Precision, Recall, F_1, ROC-AUC
Classification Threshold	0.5
SMOTE Applied	Training Data Only

7. Results

The goal of the experimental testing component of this research is to obtain accurate estimations of the actual in-creases in predictability provided by the proposed multi-source AI/ML model.

The section below demonstrates the robust comparative analysis conducted between conventional statis-tical frameworks and the proposed XGBoost framework, and experimentally proves the importance of novel data utilization and advanced over-sampling techniques.

A. Overall Performance Metrics and Baselines: To establish the performance baseline, a traditional Logistic Regression model was tested against the proposed XGBoost model trained on the combined dataset [10], [18]. Although Logistic Regression is the most common scoring model for legacy credit bureau systems, its rigid assumptions limit its ability to capture multidimensional "thin-file" entity characteristics. The baseline Logistic Regression model returned an AUC of 0.85 and an F1 score of 0.75. Within a large financial institution portfolio, this 15% error rate corresponds to extreme capital exposure and an unacceptably high false rejection rate for creditworthy MSMEs. A control variable termed "XGBoost Original" was tested to illustrate the exact impact of feature engineering on algorithm choice: the XGBoost algorithm applied without optimized behavioral features or class imbalance management. The XG-Boost Original achieved an AUC of 0.9483 and an F1 score of 0.8844, demonstrating that ensemble tree-based models dramatically outperform linear models in managing data het-erogeneity [11]. The final proposed framework—integrating multi-source architecture, behavioral features, and SMOTE balancing—achieved an AUC of 0.9646 and an F1 score of 0.9115, representing a significant improvement across all metrics [13], [26].

B. Synthetic Balancing Analysis and Probability: Calibration Running the proposed framework on the testing partition at a classification threshold of 0.5 yields a confusion matrix with 439 True Positives, 79 False Positives (Precision = 0.8475), and 61 False Negatives (Recall = 0.8780). To validate the effectiveness of SMOTE, the framework was also tested using the algorithmic scale_pos_weight balancing approach. SMOTE outperformed this baseline, resulting in only 73 False Positives and a recall of 0.8740, confirming that synthesizing new minority instances in feature space is a more stable approach than amplifying penalties [26], [37]. Moving beyond binary classification, the model produced a PR-AUC of 0.9386 and a Brier Score of 0.0995. The extremely high PR-AUC indicates that the top-ranked positive predictions are highly accurate even in a population where positive samples are inherently rare. A Brier Score below 0.1 confirms the reliability of the probability estimates, enabling lenders to confidently use the presented risk score for real-time decision-making [38].

C. Dynamic Risk Trade-offs and Operational Threshold Analysis: A real-world financial institution will never rely on a single fixed threshold. A 0.5 threshold results in 79 False Positives and 61 False Negatives (84.75% Precision, 87.8% Recall). To dramatically reduce the false negative rate to 24, the decision threshold can be reduced to 0.2, yielding a Recall of 0.9520 at the cost of 125 false positives. Conversely, a 0.4 threshold results in 83.52% Precision and 89.20% Recall, reducing false negatives from 61 to 54 at the cost of only 9 additional false positives, illustrating the continuous, tunable nature of the risk-recall trade-off this framework affords to lending institutions [19].

Table 3 Performance Comparison of Models

Model	AUC	F1 Score
Logistic Regression	0.8500	0.7500
XGBoost Original	0.9483	0.8844
Proposed Model	**0.9646**	**0.9115**

Table 4 Threshold Analysis and Risk Trade-offs

Threshold	Precision	Recall	FP	FN
0.2	0.7920	0.9520	125	24
0.3	0.8183	0.9280	103	36
0.4	0.8352	0.8920	88	54
0.5	0.8475	0.8780	79	61

8. Conclusion

This study presents a scalable and explainable AI/ML-driven alternative credit scoring framework for MSMEs that addresses the limitations of conventional, data-scarce lending practices. By integrating multi-source financial, behavioural, repayment, and firmographic data with XGBoost and SMOTE-based class balancing, the proposed model achieved a ROC-AUC of 0.9646, F1-score of 0.9115, PR-AUC of 0.9386, and Brier Score of 0.0995, demonstrating strong discrimination and probability calibration. The integration of SHAP-based explainability, fairness constraints, and human-in-the-loop review further enhances transparency, accountability, and practical suitability for regulated lending environments. Deployment through an asynchronous REST-based microservices architecture additionally demonstrates its potential for real-time and scalable FinTech applications. Overall, the framework provides a robust foundation for more inclusive, data-driven, transparent, and responsible MSME credit assessment.

Future research will extend the framework in four key directions:

- Graph Neural Networks (GNNs): Model supply-chain, vendor, and transactional relationships to detect network-level and cascading credit risks.
- LLM/NLP Integration: Extract risk-relevant information from unstructured documents, tax records, news, and loan applications.
- Dynamic Macroeconomic Adaptation: Incorporate real-time economic indicators and adaptive thresholds to address concept drift and changing market conditions.
- Privacy-Preserving Federated Learning: Enable collaborative model training across financial institutions without sharing sensitive customer data.

Declaration: The authors received no financial support for the research, and publication of this article.

References

[1] M. Abate, R. Banerjee, and S. Innocenti, “Innovative Credit Scoring: Unlocking Financial Opportunities for Micro, Small and Medium En-terprises,” J-PAL Innovations Review Paper, 2024.

[2] B. I. Adekunle, E. C. Chukwuma-Eke, E. D. Balogun, and K. O. Ogunsola, “Machine learning for automation: Developing data-driven solutions for process optimization and accuracy improvement,” Machine Learning, vol. 2, no. 1, 2021.

[3] B. I. Adekunle, E. C. Chukwuma-Eke, E. D. Balogun, and K. O. Ogunsola, “Predictive Analytics for Demand Forecasting: Enhancing Business Resource Allocation Through Time Series Models,” 2021.

[4] T. Adenuga, A. T. Ayobami, and F. C. Okolo, “Laying the Groundwork for Predictive Workforce Planning Through Strategic Data Analytics and Talent Modeling,” IRE Journals, vol. 3, no. 3, pp. 159–161, 2019.

[5] T. Adenuga, A. T. Ayobami, and F. C. Okolo, “AI-Driven Workforce Forecasting for Peak Planning and Disruption Resilience in Global Lo-gistics and Supply Networks,” International Journal of Multidisciplinary Research and Growth Evaluation, vol. 2, no. 2, pp. 71–87, 2020.

[6] O. E. Adesemoye, E. C. Chukwuma-Eke, C. I. Lawal, N. J. Isibor, A. O. Akinotbi, and F. S. Ezeh, “Improving financial forecasting accuracy through advanced data visualization techniques,” IRE Journals, vol. 4, no. 10, pp. 275–277, 2021.

[7] T. T. Adewale, T. D. Olorunloymi, and T. N. Odonkor, “Advancing sustainability accounting: A unified model for ESG integration and auditing,” Int J Sci Res Arch, vol. 2, no. 1, pp. 169–185, 2021.

[8] T. T. Adewale, T. D. Olorunloymi, and T. N. Odonkor, “AI-powered financial forensic systems: A conceptual framework for fraud detection and prevention,” Magna Sci Adv Res Rev, vol. 2, no. 2, pp. 119–136, 2021.

[9] O. S. Addy, B. O. Ajayi, and O. O. Nifise, “AI in credit scoring: A comprehensive review of models and predictive analytics,” Global Journal of Emerging Technologies & Analytics, vol. 8, no. 1, pp. 1–23, 2024.

[10] P. Berka and J. Rauch, "Machine Learning Techniques in Credit Scoring: State-of-the-Art Review," Expert Systems with Applications, vol. 212, p. 119715, 2023.
[11] V. Chang, D. Lee, and Z. Wang, "Review of Machine Learning Models for Credit Risk Prediction," Applied Finance, vol. 11, no. 3, pp. 112–140, 2024.
[12] B. Chen, "Using a genetic backpropagation neural network model for credit risk assessment of MSMEs," Heliyon, vol. 10, no. 4, p. e16973, 2024.
[13] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794.
[14] O. Elebe and C. C. Imediegwu, "A Credit Scoring System Using Transaction-Level Behavioral Data for MSMEs," Multidisciplinary Fron-tiers, vol. 2, no. 1, pp. 312–322, 2021.
[15] L. Gambacorta et al., "Machine learning for credit risk: Impact of non-traditional data at fintech companies," Journal of Banking & Finance, vol. 106, p. 105876, 2019.
[16] IFC (International Finance Corporation), "Handbook: MSME Banking in the Digital Era," International Finance Corporation, World Bank Group, 2025.
[17] A. Jenkison and D. Foley, "Data integration strategies for MSME credit scoring: An empirical study," International Journal of Data Science and Analytics, vol. 14, no. 4, pp. 332–347, 2023.
[18] A. E. Khandani, A. J. Kim, and A. W. Lo, "Consumer credit-risk models via machine-learning algorithms," Journal of Banking & Finance, vol. 34, no. 11, pp. 2767–2787, 2010.
[19] G. Kou, Y. Xu, and Y. Peng, "Credit Scoring: A Review on Modeling Techniques, Data Sources, and Bias Mitigation," Decision Support Systems, vol. 193, p. 113614, 2024.
[20] R. Kumar, N. Mishra, and R. Agarwal, "Explainable AI frameworks in banking: Opportunities and challenges for transparent credit scoring," Financial Innovation, vol. 8, no. 5, pp. 235–249, 2022.
[21] J. Kumar Roy and L. Vasa, "Machine learning for credit scoring and loan default prediction: Behavioral data and transactional patterns," World Journal of Advanced Research and Reviews, vol. 19, no. 2, pp. 335–348, 2025.
[22] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, 2017, vol. 30, pp. 4765–4774.
[23] T. Rane and S. Kale, "Financial inclusion through AI-driven credit scoring for MSMEs in India," Indian Journal of Finance, vol. 9, no. 2, pp. 46–56, 2022.
[24] J. K. Roy and L. Vasa, "A systematic review of AI and machine learning applications in credit risk assessment," Journal of Infrastructure, Policy and Development, vol. 9, no. 1, pp. 1–24, 2025.
[25] R. Vashisht and A. Uddin, "Alternative data for credit scoring: Current status and future prospects," BankTech International Journal, vol. 15, no. 2, pp. 42–54, 2023.
[26] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.
[27] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in super-vised learning," in Advances in Neural Information Processing Systems (NeurIPS), vol. 29, 2016, pp. 3323–3331.
[28] S. Barocas and A. D. Selbst, "Big data's disparate impact," California Law Review, vol. 104, no. 3, pp. 671–732, 2016.
[29] D. Micci-Barreca, "A preprocessing scheme for high-cardinality categor-ical attributes in classification and prediction problems," ACM SIGKDD Explorations Newsletter, vol. 3, no. 1, pp. 27–32, 2001.
[30] O. Troyanskaya et al., "Missing value estimation methods for DNA microarrays," Bioinformatics, vol. 17, no. 6, pp. 520–525, 2001.
[31] V. B. Djeundje, J. Crook, R. Calabrese, and M. Hamid, "Enhancing credit scoring with alternative data," Expert Systems with Applications, vol. 163, p. 113766, 2021.
[32] Z. Zhang, Q. Shen, Z. Hu, Q. Liu, and H. Shen, "Credit risk analysis for SMEs using graph neural networks in supply chain," in Proc. Int. Conf. on Big Data, Artificial Intelligence and Digital Economy (BDAIE), 2025, arXiv:2507.07854.
[33] Z. Wang, J. Xiao, L. Wang, and J. Yao, "A novel federated learning approach with knowledge transfer for credit scoring," Decision Support Systems, vol. 177, p. 114084, 2024.
[34] M. Golec and M. AlabdulJalil, "Interpretable LLMs for credit risk: A systematic review and taxonomy," Expert Systems with Applications, vol. 306, p. 130941, 2026.
[35] C. Richardson, Microservices Patterns: With Examples in Java. Manning Publications, 2018.
[36] OpenCorporates, "Open data on companies for the public good," [On-line]. Available: https://opencorporates.com, 2024.

[37] H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263–1284, 2009.
[38] G. W. Brier, "Verification of forecasts expressed in terms of probability," Monthly Weather Review, vol. 78, no. 1, pp. 1–3, 1950.