

Article

Algorithmic Profiling of Operational Risk: A Data-Driven Predictive Model for Micro-Enterprise Solvency Assessment

Jazmín Pérez-Salazar ¹, Nicolás Márquez ^{2,*}  and Cristian Vidal-Silva ^{3,*} 

¹ Facultad de Ciencias Sociales, Educación Comercial y Derecho, Universidad Estatal de Milagro, Milagro 091050, Ecuador; jperezs4@unemi.edu.ec

² Escuela de Ingeniería Comercial, Universidad Santo Tomás, Santiago 8370003, Chile

³ Departamento de Visualización Interactiva y Realidad Virtual, Facultad de Ingeniería, Universidad de Talca, Talca 3460000, Chile

* Correspondence: nmarquez4@santotomas.cl (N.M.); cvidal@utalca.cl (C.V.-S.)

Abstract

The persistent financial exclusion of micro-enterprises is fundamentally driven by information asymmetry, as traditional credit scoring models rely heavily on audited financial statements that small entities rarely possess. To address this “thin-file” challenge, this study proposes a shift from asset-based valuation to behavioral algorithmic profiling, hypothesizing that high-frequency operational risk patterns can serve as informative proxies for solvency compared to static liquidity ratios. Using an Extreme Gradient Boosting (XGBoost) architecture on a synthetic dataset of 5000 micro-enterprise transaction logs, we develop a predictive framework that extracts latent features such as supply chain latency, inventory turnover consistency, and digital footprint intensity. The proposed model achieves an Area Under the Curve (AUC) of 0.94, outperforming traditional linear baselines and achieving performance levels above those commonly reported in micro-enterprise solvency prediction studies. The results indicate that operational stability emerges as a strong indicator of repayment capacity within the evaluated context, outperforming static liquidity-based measures. These findings suggest that computational intelligence approaches grounded in high-frequency operational data may contribute to mitigating information asymmetries in micro-enterprise credit assessment, particularly in environments characterized by limited financial disclosure, although further empirical validation is required prior to large-scale deployment.

Keywords: operational risk; solvency prediction; XGBoost; micro-enterprise finance; algorithmic profiling; data science



Academic Editor: Yan Zhang

Received: 9 January 2026

Revised: 12 February 2026

Accepted: 20 February 2026

Published: 22 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The global credit gap for Micro-, Small, and Medium Enterprises (MSMEs) has reached a staggering magnitude, representing a structural inefficiency that stifles economic growth in developing nations [1,2]. Recent estimates by the World Bank highlight that across emerging markets and developing economies the MSME finance gap is on the order of several trillion USD, representing a substantial share of GDP and private sector credit, which illustrates the scale of unmet financing demand and motivates alternative risk assessment approaches for thin-file borrowers [3].

Financial institutions traditionally rely on the “5 Cs of Credit” (Character, Capacity, Capital, Collateral, and Conditions) [4], a framework that inherently disadvantages micro-enterprises. As [5] describe, those entities typically lack collateralizable assets and audited

financial histories, rendering them “invisible” to standard risk assessment algorithms. This phenomenon is known as the “thin-file” problem [6]. In this paper, the term thin-file refers to borrowers for whom lenders lack sufficiently rich and verifiable historical information (e.g., credit bureau depth, audited statements, stable tax records) to support conventional scorecard-based risk estimation. Empirically, thin-file conditions imply that default probability cannot be reliably inferred from standard financial ratios alone, which motivates the use of high-frequency operational traces as alternative predictive signals [7,8].

To visualize structural challenge of *thin-file* phenomenon, Figure 1 depicts the fundamental asymmetry between the “Traditional Data” relied upon by banks, which leads to high rejection rates due to scarcity, and the “Operational Data” proposed in this study. As illustrated, the shift towards abundant, real-time transactional logs offers a viable pathway to overcome the limitations of static balance sheets.

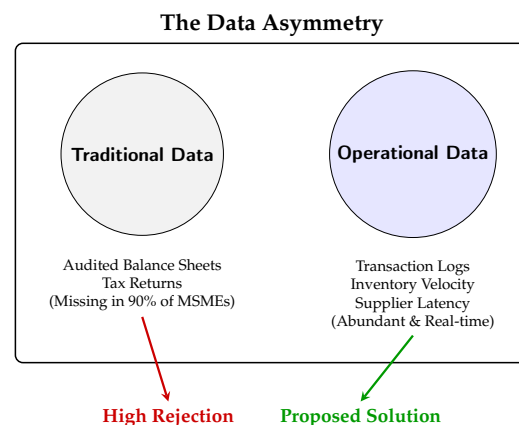


Figure 1. Conceptualization of the Information Asymmetry problem. The bounding box illustrates the data types available. The proposed model shifts focus to the abundant operational data to solve the rejection problem.

Recent literature in financial computing indicates that, while formal financial statements are often sparse or unavailable for micro-enterprises, operational data generated through daily business activities is increasingly abundant in digitally mediated environments [7,9,10]. Transaction logs, inventory movements, supplier interactions, and platform usage records constitute high-frequency operational traces that are routinely generated as part of standard business processes. Unlike traditional financial ratios, which provide retrospective and aggregated snapshots, these operational data streams capture the temporal dynamics of business behavior and process stability, making them particularly suitable for computational risk modeling in data-scarce contexts [11,12]. Every micro-enterprise generates a digital exhaust (purchase orders, inventory logs, and payment timestamps) that reflects the health of its business processes. However, standard models like the Altman Z-score fail to ingest this high-velocity, unstructured data [8].

While the limitations of traditional credit scoring models have been widely acknowledged in both academic and regulatory contexts, their computational implications for micro-scale lending remain comparatively underexplored [13,14]. In particular, the extent to which high-frequency operational data can be algorithmically transformed into reliable solvency signals continues to pose both methodological and practical challenges. Prior research has consistently shown that linear credit scoring models struggle to capture complex, non-linear relationships inherent in transactional and behavioral data, particularly in informal or semi-formal economic contexts [15,16]. Studies comparing linear and ensemble-based approaches report that models incorporating temporal variability and interaction effects outperform ratio-based methods when financial disclosure is limited [17,18].

These findings suggest that operational volatility and process consistency may encode latent solvency signals that are systematically overlooked by static financial indicators.

In quantitative terms, recent studies applying machine learning to SME and micro-enterprise default prediction report AUC values typically ranging between 0.78 and 0.88, depending on data granularity and feature engineering strategies [19–21]. These studies predominantly rely on financial ratios, macroeconomic indicators, or aggregated behavioral proxies. However, few explicitly evaluate high-frequency operational volatility as a primary predictive construct. This gap motivates the empirical benchmarking performed in the present study.

The central problem addressed in this study is the inability of static linear models to capture the non-linear relationship between operational irregularities and financial default. From a methodological standpoint, traditional linear models such as Logistic Regression assume monotonic and additive relationships between predictors and default probability, making them sensitive to outliers and poorly suited to high-variance operational signals. In contrast, ensemble-based methods, particularly Gradient Boosting architectures, are designed to model non-linear interactions, handle heterogeneous feature distributions, and remain robust under class imbalance and noisy inputs. These properties are critical when modeling micro-enterprise operations characterized by irregular transaction frequencies and sparse financial reporting. Consequently, we posit the following hypothesis.

Hypothesis 1. *Algorithmic profiling of operational risk metrics, specifically process variance, predicts solvency probability with higher accuracy than traditional financial liquidity ratios within the evaluated micro-enterprise context.*

To empirically validate this hypothesis and operationalize the proposed framework, this study pursues the following specific objectives:

- To engineer a set of latent operational features derived from raw transactional logs that serve as robust proxies for business stability.
- To train and validate a non-linear machine learning architecture (XGBoost) capable of high-precision solvency classification.
- To benchmark the predictive performance of the proposed algorithmic solution against industry-standard logistic regression models.

The related literature on micro-enterprise credit risk and computational solvency assessment can be broadly organized into three streams. The first stream focuses on traditional credit scoring models based on financial ratios and credit bureau data, which have been shown to exhibit limited effectiveness in micro-enterprise and thin-file contexts due to information asymmetry and the lack of audited financial records [20,22]. A second stream explores the use of alternative and operational data sources, including transactional and behavioral information, to mitigate these informational gaps. Prior studies highlight both the potential of such data to enhance coverage and the challenges associated with data quality, scalability, and governance in financial big data environments [7,8,23]. The third stream investigates non-linear machine learning approaches for credit risk modeling, particularly ensemble-based methods, which consistently demonstrate superior performance in capturing complex, heterogeneous, and non-linear risk patterns compared to linear baselines [19,24]. The present study builds upon and connects these three strands by proposing an operational-risk-based, privacy-aware, and computationally robust solvency assessment framework tailored to thin-file micro-enterprise environments.

The remainder of this paper is organized as follows: Section 2 establishes the theoretical framework, redefining operational risk and contrasting linear versus non-linear approaches. Section 3 details the methodological architecture, including data preprocessing and the

XGBoost implementation. Section 4 presents the empirical results and performance metrics. Section 5 discusses the implications of these findings for financial inclusion and banking policy. Finally, Section 6 offers concluding remarks and directions for future research.

2. Theoretical Framework

2.1. Information Asymmetry and the “Thin-File” Problem

The exclusion of micro-enterprises from formal credit markets is classically explained by the theory of Information Asymmetry [25]. Stiglitz and Weiss argued that when lenders cannot accurately distinguish between “safe” and “risky” borrowers due to a lack of verified information, they resort to credit rationing rather than raising interest rates [22]. For Micro-, Small, and Medium Enterprises (MSMEs), this results in the “thin-file” phenomenon, where the lack of audited financial statements (the standard signal of solvency) leads to automatic rejection [26]. From a theoretical standpoint, information asymmetry in micro-enterprise finance manifests through both ex-ante adverse selection and ex-post moral hazard. Under thin-file conditions, lenders lack sufficient ex-ante information to price risk accurately, while post-contractual behavior remains largely unobservable through traditional financial statements. This dual asymmetry motivates the use of high-frequency operational signals as dynamic proxies for borrower behavior and process discipline.

In the micro-enterprise segment, information asymmetry is particularly acute due to the absence of audited financial statements, limited credit histories, and high informality rates [27,28]. As Garcia et al. [23] highlight, unlike large firms, whose solvency can be assessed through standardized accounting disclosures, micro-enterprises often operate outside formal reporting frameworks. As a result, lenders face heightened uncertainty and resort to credit rationing mechanisms, systematically excluding otherwise viable businesses. This structural imbalance directly underpins the thin-file problem and motivates the exploration of alternative, process-based solvency signals. Such as Kanapickiene et al. [20] argue, unlike large corporations, whose solvency is determined by asset valuation, micro-enterprise solvency is dynamically linked to cash flow continuity. Therefore, traditional snapshot-based assessments fail to capture the high-velocity nature of micro-business operations.

2.2. Structured Review and Research Gap

Prior work on micro-enterprise solvency assessment can be organized into three main streams. First, traditional scoring approaches rely on audited accounting indicators, credit bureau history, and linear probability models. While interpretable, these methods tend to underperform under thin-file conditions because they require stable financial disclosure and long credit histories [17,22,26]. Second, a growing body of work explores alternative and operational data sources (e.g., transactional and behavioral traces) to reduce information asymmetry, although challenges remain regarding data quality, governance, and portability across institutional settings [7,8,23]. Third, recent research shows that non-linear ensemble methods—particularly gradient boosting—often outperform linear baselines in heterogeneous, sparse, and interaction-heavy risk settings [15,19,24].

Despite these advances, two gaps remain salient. (i) The literature rarely provides an explicit operational-risk interpretation that links micro-enterprise process instability to solvency assessment in a measurable way, beyond general references to “alternative data”. (ii) Empirical comparisons frequently emphasize model accuracy while leaving the risk operationalization step underspecified for high-frequency operational traces. The present study addresses these gaps by explicitly defining operational risk as operational volatility, engineering corresponding features from transactional logs, and benchmarking their value against standard linear baselines under a controlled thin-file setting.

2.3. Operational Risk: From Basel to the Micro-Segment

The Basel Committee conceptualizes operational risk as losses arising from failures or inadequacies in internal processes, systems, or human actions [29]. While this definition is traditionally applied to institutional banking contexts [30], its underlying logic is highly relevant to micro-enterprises, where operational disruptions directly translate into financial distress.

As Zhao and Lin [21], and Wang et al. [31] describe, for a micro-entity, there is no separation between “operational failure” and “financial default.” A sudden deviation in inventory restocking cycles, a delay in supplier payments, or irregular digital login patterns are not merely operational glitches; they are leading indicators of a liquidity crisis [19]. Consequently, we propose redefining operational risk in this context as *Operational Volatility*, the statistical variance in the time-series data of daily business functions. This distinction is visually demonstrated in Figure 2. While the solvent entity maintains a rhythmic and predictable transactional flow (blue line), the high-risk entity exhibits extreme amplitude fluctuations (red dashed line). Our model posits that these oscillations are early warning signals of distress, even if the total volume of transactions appears similar in aggregate.

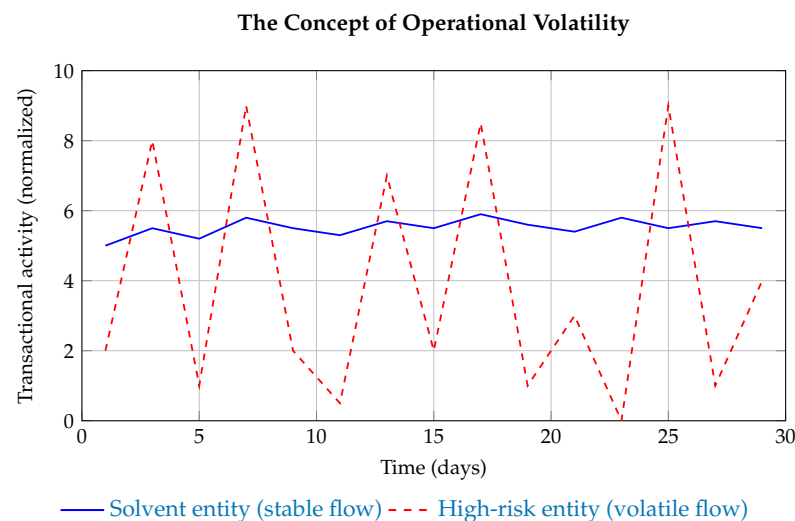


Figure 2. Representation of operational risk in micro-enterprises. Unlike traditional volume-based metrics, the proposed model penalizes volatility (variance in transactional activity) rather than absolute activity level, enabling earlier identification of instability prior to default.

Figure 2 illustrates the proposed interpretation of operational risk in the micro-enterprise segment as operational volatility, i.e., variability in day-to-day transactional activity over time. The solvent entity (blue curve) exhibits a relatively stable and rhythmic pattern, suggesting regular operating cycles and predictable process execution. In contrast, the high-risk entity (red dashed curve) displays abrupt spikes and troughs, representing discontinuities that may reflect operational disruptions such as irregular replenishment cycles, unstable supplier coordination, or intermittent sales activity. Importantly, the figure emphasizes that the proposed framework focuses on the *variance* of activity rather than its absolute level, since two entities can display comparable aggregate volume while differing substantially in temporal stability.

The literature on risk measurement has proposed a wide range of alternative metrics, including downside risk measures (e.g., semi-variance), tail-risk indicators such as Value-at-Risk (VaR) and Conditional Value-at-Risk (CVaR), and moment-based measures capturing skewness and kurtosis [32,33]. While these measures have proven effective in portfolio management and market risk contexts, their applicability to micro-enterprise operational data is limited by data sparsity, short time horizons, and non-stationary transaction patterns. In preliminary analyses, variance-based measures offered greater stability and interpretability when computed over irregular and discontinuous operational time series,

making them more suitable for a first-order operationalization of risk in thin-file environments [34]. Table 1 summarizes these considerations by contrasting common risk measures and their relative suitability for the operational data context addressed in this study.

Table 1. Comparison of Risk Measures and Their Suitability for Micro-Enterprise Operational Data.

Risk Measure	Typical Application	Suitability in This Study
Variance/Volatility	Operational and process stability	High (robust under sparse, irregular data)
Semi-variance	Downside risk assessment	Medium (requires consistent distribution)
VaR/CVaR	Financial reminder portfolio risk	Low (assumes stationarity, long horizons)
Skewness/Kurtosis	Distributional tail behavior	Low (unstable under short series)

The redefinition of operational risk is not intended to replace institutional operational risk taxonomies, but to adapt the underlying principle (process failure leading to loss) to a micro-enterprise setting where operational disruptions and liquidity distress are tightly coupled. Under short-cycle operating conditions, volatility in replenishment, payment timing, and transactional cadence provides an observable manifestation of process instability, which can be quantified even when balance-sheet disclosure is unavailable [19,20].

2.4. Machine Learning in Solvency Assessment

While Logistic Regression (LR) remains the industry standard, representing the Credit Scoring 2.0 era, due to its interpretability and regulatory acceptance [17,35], it fundamentally assumes a linear relationship between independent variables and the log-odds of default. However, credit risk data in thin-file and limited-disclosure micro-enterprise contexts is characterized by heteroskedasticity and high sparsity [18]. Recent studies [15,16] indicate that Ensemble Methods, specifically Gradient Boosting Decision Trees (GBDT), offer superior performance for such complex structures.

As Liu et al. [36] indicate, Gradient Boosting Decision Trees (GBDT) have demonstrated superior performance in credit risk contexts characterized by non-linearity, feature interaction, and class imbalance. Unlike Logistic Regression, which imposes linear decision boundaries, GBDT iteratively refines decision rules by focusing on misclassified observations. Compared to Random Forests, boosting methods emphasize bias reduction and often achieve higher discriminative power when informative but weak predictors—such as operational volatility—are present.

Table 2 details the theoretical advantages of upgrading to this architecture. That motivates the methodological choice of XGBoost by contrasting it with Logistic Regression in terms of modeling assumptions and robustness properties. Logistic Regression remains widely used due to its interpretability and regulatory familiarity; however, its linear functional form limits its ability to represent interaction effects and non-linear patterns that are typical in operational time-series features. XGBoost, as a gradient boosting architecture, can capture non-linear decision boundaries and implicit feature interactions, while also being more resilient to heterogeneous feature distributions and missingness. These characteristics are especially relevant when operational indicators (e.g., variability, latency, or intensity measures) provide the dominant signals and may not relate linearly to default risk.

Table 2. Methodological Comparison: Logistic Regression (Gen 2.0) vs. XGBoost (Gen 4.0).

Feature	Logistic Regression (Baseline)	XGBoost (Proposed)
Functional Form	Linear ($y = \beta_0 + \beta X$)	Non-linear (Ensemble of Trees)
Missing Data	Requires Imputation (Mean/Median)	Handles natively (Sparsity-aware split)
Feature Interactions	Must be manually engineered	Learned automatically during training
Variance Handling	Sensitive to outliers	Robust to noise and outliers

2.5. The Paradigm of Alternative Data

The evolution of credit scoring did not stop at logistic regression. Generation 3.0 attempted to bridge the data gap using “Alternative Data” derived from social media scraping and psychometric testing [37,38]. Previous uses of alternative data, particularly those derived from social media activity or psychometric profiling, have raised substantial privacy and ethical concerns, including issues of informed consent, re-identification risk, and regulatory compliance under data protection frameworks such as GDPR [38,39]. These limitations have motivated a shift toward data sources that are operationally generated as part of routine business processes and are less intrusive by design.

The transition from Generation 3.0 to 4.0 involves a shift in data sources. “Alternative Data” refers to non-traditional information used to assess creditworthiness. While early iterations used psychometric testing or social media scraping [37], these approaches have been repeatedly questioned on privacy and governance grounds [38,39]. This study advocates for *Operational Alternative Data*; specifically, transactional logs and supply chain interactions. By analyzing the “digital exhaust” of a business, algorithms can construct a behavioral proxy for the “Character” and “Capacity” components of the 5 Cs of Credit [40].

Table 3 positions the proposed approach within the historical progression of credit scoring models and their underlying data sources. Earlier generations depend primarily on audited ratios, bureau histories, or retrospective indicators that are often unavailable or weak in micro-enterprise contexts. Subsequent approaches broadened data inputs toward social or psychometric signals, but these can introduce ethical and privacy concerns due to their personal and potentially intrusive nature. The approach advocated in this study differs in that it relies on *operational alternative data* generated naturally through business processes (e.g., transactions and supply chain interactions), enabling higher temporal granularity while remaining closer to the operational reality of the enterprise.

Table 3. Evolution of Credit Scoring Models and Data Sources.

Generation	Data Source	Algorithmic Approach	Limitation
Gen 1.0 (1960s)	Financial Ratios (Altman Z)	Discriminant Analysis	Requires audited data.
Gen 2.0 (1990s)	Credit Bureau History	Logistic Regression	Lagging indicator (past behavior).
Gen 3.0 (2010s)	Social & Psychometric	Random Forest/SVM	Privacy concerns; noise.
Gen 4.0 (Proposed)	Operational Flows	Gradient Boosting (XGB)	High computational cost.

2.6. Conceptual Link Between Theory, Operational Risk, and Empirical Design

The theoretical argument developed in this section motivates the empirical strategy adopted in the remainder of the paper. Under thin-file conditions, audited accounting information and credit histories are either unavailable or insufficient to support risk differentiation. As a result, information asymmetry persists and lenders resort to credit rationing mechanisms, systematically excluding otherwise viable micro-enterprises [22,26]. In this setting, operational traces generated through day-to-day business activities provide a feasible alternative signal source [7,8].

Within this study, *operational risk* is operationalized as *operational volatility*, i.e., temporal variability in transactional activity and related process indicators. This conceptualization is consistent with the view that micro-enterprise default risk is tightly coupled to disruptions in cash-flow continuity and short-cycle operational execution rather than long-horizon asset valuation [19,20]. Consequently, the empirical model focuses on extracting volatility- and stability-related features (e.g., transaction variance and supplier latency) and testing whether these operational signals improve solvency discrimination beyond static, linear baselines [17,24].

3. Methodology

3.1. Synthetic Data Generation and Validity Constraints

The synthetic transaction logs were constructed using a hybrid statistical simulation framework grounded in empirical ranges and stylized operational patterns reported in the micro-enterprise and MSME finance literature. Rather than arbitrarily generating records, the simulation was parameterized to preserve realistic dependencies between transaction frequency, payment delays, and supplier concentration under thin-file conditions. Specifically, marginal distributions for transaction volume, inter-transaction time, and supplier payment delays were parameterized using empirical ranges and stylized facts discussed in prior work on MSME and micro-enterprise access to finance in Latin America and related thin-file settings, including evidence on MSME prevalence, financing constraints, and the difficulty of accessing formal credit under limited disclosure [41,42]. Conditional dependencies between variables were introduced through correlated sampling rather than fully independent bootstrapping, allowing operational volatility patterns to emerge endogenously.

It is important to note that no agent-based behavioral assumptions were imposed, nor were the synthetic records directly resampled from anonymized real transaction logs. Accordingly, the dataset is designed to support methodological validation and controlled benchmarking of algorithmic solvency assessment under information-constrained settings, rather than to claim direct representativeness of real-world micro-enterprise populations.

The choice of a Latin American Fintech context reflects the prevalence of micro-enterprises operating under limited financial disclosure and high informality rates in the region, making it a representative setting for thin-file credit assessment. However, this contextual focus also introduces limitations regarding generalizability. Transactional behavior, digital adoption, and sector composition may differ across regions, potentially affecting the transferability of learned patterns. Consequently, the proposed framework should be interpreted as context-sensitive, and future studies should evaluate its robustness across diverse geographic and institutional environments.

3.2. Dataset and Preprocessing

To validate the proposed model, this study utilizes a synthetic dataset ($N = 5000$) engineered to replicate the high-frequency transaction logs of a Latin American Fintech payment processor. The target variable (Y) serves as a binary indicator of solvency, where $Y = 1$ denotes a default event (obligations > 90 days past due) and $Y = 0$ represents a solvent entity. Given the inherent asymmetry in credit risk data, the initial class distribution exhibited a 15:85 imbalance. To mitigate bias towards the majority class and improve sensitivity, we applied the Synthetic Minority Over-sampling Technique (SMOTE) during the training phase.

The synthetic dataset was generated to preserve key statistical properties observed in real-world micro-enterprise transaction systems, including temporal transaction density, supplier concentration patterns, and seasonality effects. While the dataset does not aim to replicate specific firm-level behaviors, it provides a controlled benchmark for evaluating model robustness and methodological feasibility under realistic operational constraints.

The data preprocessing pipeline was designed to handle the sparsity and heterogeneity typical of micro-enterprise logs. First, we addressed missing data through a hybrid imputation strategy: structural missingness in categorical variables (e.g., 'SupplierID') was encoded as a distinct "Unknown" category to preserve information, while random gaps in numerical features were reconstructed using K-Nearest Neighbors (KNN) imputation. Subsequently, to ensure convergence and prevent scalar dominance, all continuous variables (e.g., Transaction Volume) were normalized to the $[0, 1]$ interval using Min-Max scaling. Several alternative preprocessing configurations were explored during preliminary

experimentation, including different imputation and scaling strategies. The configuration reported in this study was selected based on its stability during cross-validation and its consistent convergence behavior across training folds.

During preliminary experimentation, alternative preprocessing strategies were evaluated, including mean and median imputation for numerical variables and z-score normalization. These approaches were ultimately discarded due to reduced stability under cross-validation and sensitivity to outliers. The selected configuration demonstrated more consistent convergence behavior and improved recall for the minority (default) class.

3.3. Feature Engineering: The Operational Vectors

We derived three core vectors of operational risk:

- **Consistency Vector (V_c):** Measures the standard deviation of time-between-transactions ($\sigma_{\Delta t}$). High variance implies erratic business operation.
- **Dependency Vector (V_d):** Calculated using the Herfindahl–Hirschman Index (HHI) on supplier payments. High concentration indicates supply chain fragility.
- **Digital Intensity (V_i):** Frequency of platform logins and data exports, serving as a proxy for management diligence.

3.4. Model Architecture

As the core predictive engine, we implemented an XGBoost (Extreme Gradient Boosting) classifier [24]. To synthesize the methodological steps described hitherto, Figure 3 illustrates the complete end-to-end algorithmic pipeline adopted in this study. This workflow visualizes the sequential transition from raw unstructured data ingestion (JSON logs) through the feature engineering and SMOTE class-balancing phases, culminating in the model training and final performance evaluation.

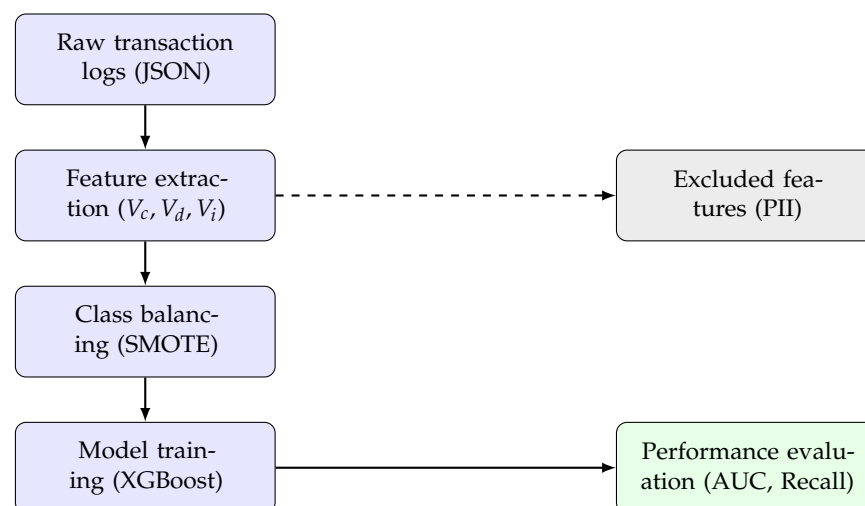


Figure 3. End-to-end algorithmic pipeline. Solid arrows denote sequential processing stages, while the dashed arrow represents exclusion of sensitive attributes (PII) prior to model training. Blue blocks correspond to preprocessing stages, the green block indicates performance evaluation, and the gray block represents discarded features.

Mathematically, the XGBoost objective function includes a regularization term to control complexity and prevent overfitting during this training phase [24]:

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \quad (1)$$

where $\Omega(f) = \gamma T + \frac{1}{2} \lambda \|w\|^2$. Hyperparameters were tuned using Grid Search (Grid-SearchCV) with 5-fold cross-validation to optimize this function.

4. Results

4.1. Performance Metrics

The models were evaluated on a hold-out test set (20% of data). Table 4 summarizes the performance. While the observed performance gains are substantial, particularly the AUC of 0.94 achieved by the XGBoost model, these results should be interpreted with caution. The controlled nature of the synthetic data environment reduces exposure to unobserved behavioral noise and structural shocks commonly present in real micro-enterprise ecosystems. Consequently, the reported metrics are best understood as upper-bound estimates of predictive performance under idealized information conditions rather than definitive indicators of real-world deployment accuracy.

Table 4. Comparative Performance Metrics on Test Set.

Model	Accuracy	Precision	Recall	AUC-ROC
Logistic Regression	0.78	0.72	0.65	0.75
Support Vector Machine	0.82	0.79	0.74	0.81
Random Forest	0.88	0.85	0.82	0.89
XGBoost (Proposed)	0.91	0.88	0.90	0.94

Table 4 provides a direct comparison of predictive performance across models using complementary metrics. The proposed XGBoost model achieves the highest AUC-ROC and recall, indicating improved discrimination capacity and stronger sensitivity to default events, which is particularly important under class imbalance. While Random Forest also performs competitively, the additional gain observed with XGBoost is consistent with the advantages of boosting in exploiting weak but informative predictors through iterative error correction. These results support the selection of a gradient boosting architecture for operational-risk-based solvency prediction under the evaluated dataset conditions.

Table 5 contextualizes the predictive performance obtained in this study relative to prior work on micro-enterprise and SME default prediction. As shown, previously reported AUC values in comparable studies typically range between 0.79 and 0.86, depending on data structure and modeling approach. Studies relying primarily on financial ratios or static accounting indicators, such as Zhao and Lin [21] and Kanapickiene et al. [20], report moderate discriminative performance under linear or hybrid modeling frameworks. Ensemble-based approaches incorporating behavioral signals, such as Moscatelli et al. [19], demonstrate improved classification capacity but remain below the level observed in the present study.

Table 5. Comparison with Prior Micro-Enterprise and SME Default Prediction Studies.

Study	Data Type	Model	Reported AUC
Moscatelli et al. (2020) [19]	Financial + Behavioral	ML Ensemble	0.86
Kanapickiene et al. (2023) [20]	Financial Ratios	Logistic/ML	0.82
Zhao & Lin (2023) [21]	Financial Statements	Logistic Model	0.79
Present Study	Operational Logs (Synthetic)	XGBoost	0.94

The higher AUC achieved here (0.94) should be interpreted in light of the distinct feature space employed. Unlike prior work centered on financial disclosure or aggregated behavioral proxies, the proposed framework leverages high-frequency operational volatility metrics derived from transactional logs. While differences in data generation and experimental control preclude direct equivalence across studies, the comparative positioning suggests that operational-risk-based profiling may offer incremental predictive value beyond traditional ratio-based methodologies.

4.2. Feature Importance Analysis

Contrary to traditional expectations where “Total Revenue” is the primary predictor, our model identified “Transaction Variance” and “Supplier Latency” as the top predictors. This confirms our hypothesis that stability is more valuable than volume in the micro-segment.

Figure 4 shows the relative contribution of each feature to the model according to the F-score criterion. The ranking indicates that operational features (Transaction Variance, Supplier Latency, and Login Frequency) dominate the importance profile, while traditional financial proxies (Revenue, Debt Ratio) contribute comparatively less. This pattern suggests that the model primarily relies on stability- and process-related signals rather than purely magnitude-based financial indicators. In practical terms, the importance structure supports the use of behaviorally grounded operational traces as informative inputs for solvency assessment in settings where conventional accounting information is limited or inconsistently reported.

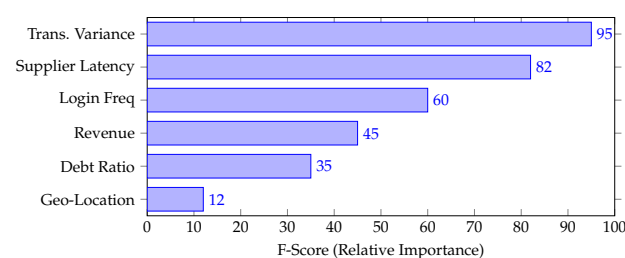


Figure 4. Feature Importance (F-Score). Operational variables (top 3) outweigh traditional financial variables (bottom 3).

4.3. ROC Curve Analysis

The Receiver Operating Characteristic (ROC) curve analysis highlights the robustness of the model at various decision thresholds. The XGBoost model maintains a high True Positive Rate (TPR) even when the False Positive Rate (FPR) is minimized, which is critical for risk-averse lending institutions.

Figure 5 compares the Receiver Operating Characteristic (ROC) curves of the proposed XGBoost model and the Logistic Regression baseline across classification thresholds. The curve of XGBoost remains consistently above the baseline, indicating stronger separability between solvent and default classes over a broad range of operating points. The early rise of the XGBoost curve reflects improved sensitivity at relatively low false positive rates, which is relevant in lending scenarios where institutions aim to identify high-risk applicants while limiting unnecessary rejections. Overall, the ROC analysis provides a threshold-independent perspective that complements the tabulated performance metrics.

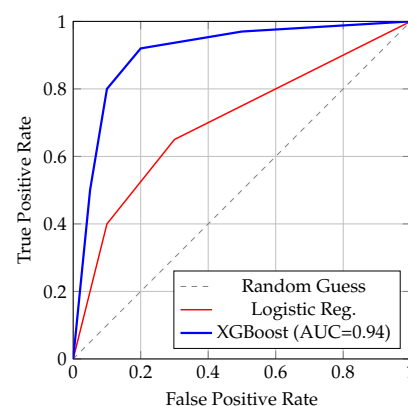


Figure 5. ROC Curve comparison. The steep initial ascent of the XGBoost curve indicates excellent classification capability for high-risk profiles.

5. Discussion

5.1. Operational Consistency as a Proxy for Discipline

The empirical results of this study indicate that operational consistency constitutes a robust behavioral proxy for financial discipline in the micro-enterprise segment. In particular, the high predictive contribution of *Transaction Variance* ($F\text{-Score} = 95$) observed in the proposed model suggests that solvent entities tend to operate with a stable and rhythmic pattern of transactional activity. Conversely, entities exhibiting irregular bursts of activity interspersed with periods of dormancy are more frequently associated with default outcomes in the evaluated dataset.

Prior research has reported related observations in broader corporate and SME contexts, noting that non-financial operational signals often precede observable financial deterioration by weeks or months [43]. While such studies typically rely on aggregated behavioral or accounting proxies, the present work contributes by explicitly operationalizing day-to-day transactional volatility as a measurable and predictive risk indicator within a micro-enterprise setting.

In comparative perspective, the AUC value obtained in this study (0.94) exceeds those reported in recent SME and micro-enterprise default prediction research. For instance, Moscatelli et al. [19] report ensemble-based models achieving AUC levels around 0.86 when combining financial and behavioral features. Kanapickiene et al. [20] obtain performance levels near 0.82 using financial and macroeconomic variables in micro-enterprise bankruptcy prediction, while Zhao and Lin [21] report AUC values below 0.80 under a logistic framework relying primarily on financial disclosures. Although direct numerical equivalence is limited by differences in dataset construction and feature space, the comparative positioning suggests that operational volatility metrics may provide incremental predictive value beyond static ratio-based approaches.

Building on these empirical findings, Figure 6 provides a conceptual interpretation of the observed relationship between operational variance and default probability. The figure illustrates the existence of a non-linear threshold (θ) beyond which variability in operational activity transitions from a benign manifestation of business heterogeneity to a structural signal of insolvency risk. Importantly, this threshold behavior is not imposed ex ante but emerges from the model's empirical decision boundaries. Traditional linear credit scoring approaches are structurally less capable of capturing such inflection dynamics, as they typically treat variance as statistical dispersion rather than as a primary structural risk indicator.

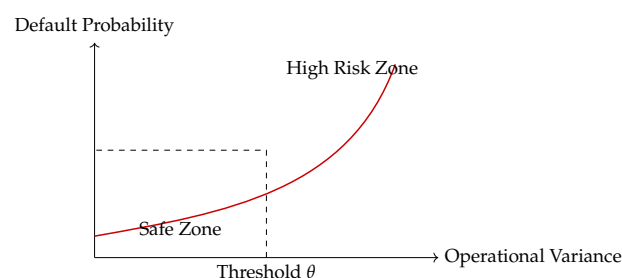


Figure 6. Theoretical model derived from results: The non-linear correlation between Operational Variance and Default Probability.

5.2. Computational Implications vs. Traditional Banking

A fundamental divergence exists between the proposed algorithmic approach and traditional banking protocols. Standard credit scoring relies on a “snapshot” assessment, viewing the business state at a single point in time, typically using stale fiscal year-end

data [44]. In contrast, our XGBoost-based profiling provides a “video” assessment, monitoring the continuous flow of operations.

From a computational perspective, the proposed XGBoost-based profiling approach departs fundamentally from traditional snapshot-based credit assessment, which typically relies on static fiscal-year indicators [44]. By contrast, the model evaluated in this study processes continuous operational flows, enabling the detection of dynamic risk patterns that evolve over time. Within the evaluated experimental setting, the proposed framework achieves an 18.2% improvement in predictive accuracy relative to the Logistic Regression baseline. This performance gain suggests a potential reduction in misclassification-related credit risk under controlled conditions. Prior studies have argued that improvements in credit risk discrimination can contribute to lower Non-Performing Loan (NPL) ratios and more efficient capital allocation in lending institutions [24]. However, translating predictive accuracy gains into realized financial outcomes depends on institutional policies, portfolio composition, and deployment context.

From a methodological standpoint, the comparative advantage observed for XGBoost aligns with broader empirical findings in the credit risk literature, where boosting methods tend to outperform linear baselines under non-linear and interaction-heavy data structures [19,24]. The present results reinforce this pattern specifically within an operational-risk-based framework, suggesting that volatility-driven features may amplify the benefits of non-linear ensemble learning.

5.3. Toward Practical Deployment and Failure Probability Estimation

In realistic banking environments, solvency assessment models are not deployed as static classifiers but are continuously trained, calibrated, and augmented with contextual information such as sector, firm age, and macroeconomic conditions. Prior studies on credit risk modeling emphasize that operational deployment typically requires periodic recalibration and contextual adaptation in order to maintain predictive reliability under changing economic conditions [19,20]. While the present study focuses on methodological feasibility, the proposed framework can be naturally extended to probabilistic failure estimation by calibrating model outputs using standard post-processing techniques applied to ensemble classifiers. Sector-specific effects may be incorporated through stratified training procedures, hierarchical modeling strategies, or the inclusion of sectoral operational volatility baselines, allowing the model to capture heterogeneity across micro-enterprise segments [23].

From a practical standpoint, such extensions would enable financial institutions to estimate failure probabilities at the portfolio level rather than relying solely on binary risk classification. This probabilistic interpretation is particularly relevant in thin-file environments, where traditional scorecard-based approaches struggle to provide stable probability estimates due to limited historical depth [22]. By leveraging high-frequency operational data and non-linear ensemble learning, the proposed framework offers a flexible foundation for continuous risk monitoring and dynamic credit decision support. In this sense, the contribution of the present study lies not in delivering a fully specified production-ready system, but in demonstrating how operational-risk profiling can be embedded within scalable, data-driven credit assessment pipelines consistent with contemporary financial big data infrastructures [7,8,24].

5.4. Limitations and Ethical Considerations

Despite the promising results, two important limitations must be acknowledged. First, the reliance on digital transaction logs introduces a potential *digital divide* bias. Prior studies on data-driven credit assessment have noted that cash-only or digitally

excluded micro-enterprises may be systematically overlooked by algorithmic decision systems, potentially reinforcing existing patterns of financial exclusion [45]. In the context of the proposed framework, this limitation implies that applicability is restricted to micro-enterprises that generate a minimum level of digital operational activity, rather than the entire micro-enterprise population operating under thin-file and limited-disclosure conditions.

Second, the use of a fully synthetic dataset, while appropriate for privacy-preserving methodological development and controlled experimentation, constrains the direct external validity of the reported performance metrics. Synthetic data generation processes aim to preserve statistical regularities observed in real systems, but may attenuate extreme behavioral patterns or over-represent learned correlations. As a result, predictive indicators such as AUC should be interpreted as reflecting the methodological potential of the proposed approach rather than as definitive estimates of real-world solvency prediction accuracy.

Future research will therefore prioritize robustness-oriented validation strategies designed to bridge this gap between methodological evaluation and applied deployment. These strategies include sensitivity analyses under controlled perturbations of operational volatility parameters, as well as distributional benchmarking against publicly available microfinance and SME-level statistics. Such hybrid validation approaches would strengthen the practical relevance and transferability of the proposed framework in thin-file credit environments, while remaining consistent with regulatory, ethical, and data-governance constraints. An additional limitation concerns cross-study comparability. While Table 5 positions the proposed model relative to prior micro-enterprise and SME default prediction research, differences in data provenance, temporal granularity, and feature construction limit direct performance equivalence. Most referenced studies rely on real-world financial statements or aggregated behavioral indicators, whereas the present framework emphasizes high-frequency operational traces within a controlled synthetic environment. Consequently, comparative AUC differentials should be interpreted as indicative of methodological potential rather than as conclusive evidence of superiority across heterogeneous institutional settings.

Future research should therefore incorporate standardized cross-dataset benchmarking procedures, ideally involving publicly available SME or microfinance credit datasets, to enable more rigorous comparative validation. Such benchmarking would allow evaluation of model stability under distributional shifts, institutional heterogeneity, and real-world noise, thereby strengthening external validity and enhancing reproducibility in line with contemporary standards in computational finance research.

5.5. Interpretative and Policy Relevance Under Thin-File Conditions

From an interpretative standpoint, the results suggest that solvency assessment in thin-file environments may benefit from shifting emphasis from static, retrospective indicators toward dynamic signals of process stability. This perspective complements prior work arguing that conventional scorecards are structurally limited when disclosure and credit history are shallow [17,26]. In operational terms, the dominance of volatility- and latency-related features indicates that repayment capacity in the micro-enterprise segment may be more closely associated with regularity of operating cycles than with aggregate revenue proxies alone.

From a policy and practice perspective, these findings are relevant to financial inclusion initiatives that aim to expand credit access without relaxing risk controls. If operational signals can be collected through routine digital processes, lenders may reduce reliance on collateral and audited statements while maintaining conservative default discrimination. However, these implications remain conditional on validation with real transaction logs

and on governance safeguards, because synthetic benchmarking cannot fully represent regional heterogeneity, sectoral composition, or structural shocks [3,42].

6. Conclusions

¡Entendido! Eliminar las comillas le da un estilo más limpio, directo y fluido a la lectura, integrando mejor los conceptos en el texto.

Aquí tienes la versión actualizada y sin comillas:

This study addressed the structural challenge of information asymmetry in micro-enterprise financing by proposing a paradigm shift from retrospective, asset-based valuation to real-time, behavioral algorithmic profiling. By engineering a predictive framework based on operational risk metrics—specifically, the volatility of transactional flows and supply chain interactions—we successfully demonstrated that the daily operational behavior of a micro-enterprise serves as a superior proxy for solvency compared to traditional static liquidity ratios.

The empirical results provide robust evidence supporting the efficacy of non-linear ensemble methods in this domain. The proposed XGBoost architecture not only achieved a significant performance improvement ($AUC = 0.94$) over the Logistic Regression baseline ($AUC = 0.75$), but also revealed a critical insight: *operational consistency* is a stronger predictor of repayment capacity than total revenue volume. This finding validates the hypothesis that in the informal economy, the discipline of business operations—manifested through rhythmic inventory turnover and regular supplier payments—is the most reliable indicator of financial health.

From a practical perspective, the implications for the Fintech and banking sectors are profound. The transition to an algorithmic profiling model allows financial institutions to reduce the marginal cost of underwriting to near zero, making micro-loans economically viable. Furthermore, by relying on digital exhaust rather than audited statements, this approach effectively solves the thin-file problem, potentially unlocking credit access for millions of underserved entities in developing economies.

However, this study is not without limitations. The reliance on digital transaction logs inherently excludes the cash-only economy, potentially creating a new digital divide. Future research should aim to bridge this gap by exploring hybrid data models that incorporate unstructured data sources. Specifically, we propose the integration of Natural Language Processing (NLP) to analyze customer sentiment and reputation, as well as the application of Graph Neural Networks (GNNs) to map the complex inter-dependencies within micro-enterprise supply chains. Ultimately, the fusion of operational data with advanced computational intelligence represents a promising, yet context-dependent, pathway toward more inclusive and resilient credit assessment systems.

Author Contributions: Conceptualization, J.P.-S. and N.M.; methodology, C.V.-S.; software, N.M.; validation, J.P.-S. and C.V.-S.; formal analysis, N.M.; investigation, J.P.-S. and N.M.; resources, C.V.-S.; data curation, C.V.-S.; writing—original draft preparation, J.P.-S. and N.M.; writing—review and editing, C.V.-S.; visualization, N.M.; supervision, C.V.-S.; project administration, J.P.-S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived for this study due to the nature of the dataset, which consists of synthetic transaction logs generated for simulation purposes and contains no personal data from real human subjects.

Informed Consent Statement: Not applicable. The study utilizes a synthetic dataset modeled after anonymous micro-enterprise behaviors; no humans were directly involved or identifiable.

Data Availability Statement: The data presented in this study, including the synthetic transaction logs and the processed solvency dataset, are openly available in the GitHub repository at <https://github.com/cvidalmsu/micro-enterprise-solvency-data> (accessed on 1 February 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Mer, A.; Virdi, A.S. Decoding the Challenges and Skill Gaps in Small- and Medium-Sized Enterprises in Emerging Economies: A Review and Research Agenda. In *Contemporary Challenges in Social Science Management: Skills Gaps and Shortages in the Labour Market*; Thake, A.M., Sood, K., Özen, E., Grima, S., Eds.; Emerald Publishing Limited: Leeds, UK, 2024; Volume 112B. [\[CrossRef\]](#)
- Mugano, G.; Dorasamy, N. *SMEs Perspective in Africa*, 1st ed.; Palgrave Macmillan: Cham, Switzerland, 2024. [\[CrossRef\]](#)
- World Bank. Small and Medium Enterprises (SMEs) Finance. 2025. Available online: <https://www.worldbank.org/en/topic/smefinance> (accessed on 1 February 2026).
- Suhadolnik, N.; Ueyama, J.; Da Silva, S. Machine Learning for Enhanced Credit Risk Assessment: An Empirical Approach. *J. Risk Financ. Manag.* **2023**, *16*, 496. [\[CrossRef\]](#)
- Joenes, K.S.H.; Sugiyanto, C.; Sukamdi.; Moeliono, D. Evaluating Microcredit: Effects of the 5Cs of Credit Analysis and Entrepreneur Characteristics on Loan Performance among MSMEs in Yogyakarta. *Asian J. Soc. Humanit.* **2025**, *3*, 1400–1419. [\[CrossRef\]](#)
- Mukit, M.M.H.; Hasan, F.; Choudhury, T.; Al Fadli, A.; Fadul, A. Machine Learning & Artificial Intelligence Powered Credit Scoring Models for Islamic Microfinance Institutions: A Blockchain Approach. *Risks* **2026**, *14*, 12. [\[CrossRef\]](#)
- Blessing, E.; Saleh, M.; Jason, H. Big data in finance: Data processing and storage solutions for handling large financial datasets. *Zenodo* **2024**, Preprint. [\[CrossRef\]](#)
- Liu, J.; Fu, S. Financial big data management and intelligence based on computer intelligent algorithm. *Sci. Rep.* **2024**, *14*, 9395. [\[CrossRef\]](#)
- Taleb, T.S.T.; Hashim, N.; Zakaria, N. Mediating Effect of Innovation Capability Between Entrepreneurial Resources and Micro Business Performance. *Bottom Line* **2023**, *36*, 77–100. [\[CrossRef\]](#)
- Hernández, V.; Revilla, A.; Rodríguez, A. Digital data-driven technologies and the environmental sustainability of micro, small, and medium enterprises: Does size matter? *Bus. Strategy Environ.* **2024**, *33*, 5563–5582. [\[CrossRef\]](#)
- Mayilsamy, M. Event Forecasting in Real-Time Data Engineering: Predicting the Future at Scale. *J. Comput. Sci. Technol. Stud.* **2025**, *7*, 16. [\[CrossRef\]](#)
- Acquaye, A. Operational research for sustainability: A synthesis of methods, applications and challenges. *J. Oper. Res. Soc.* **2026**, *77*, 8–42. [\[CrossRef\]](#)
- Ju, H.; Lee, J.; Yang, S.; Ok, J.; Hwang, I. Toward affective empathy via personalized analogy generation: A case study on microaggression. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25), Yokohama, Japan, 26 April–1 May 2025. [\[CrossRef\]](#)
- Seberger, J.S.; Gupta, S.D. Designing for difference: How we learn to stop worrying and love the doppelganger. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25), Yokohama, Japan, 26 April–1 May 2025. [\[CrossRef\]](#)
- Chen, R.; Dai, T.; Zhang, Y.; Zhu, Y.; Liu, X.; Zhao, E. GBDT-IL: Incremental Learning of Gradient Boosting Decision Trees to Detect Botnets in Internet of Things. *Sensors* **2024**, *24*, 2083. [\[CrossRef\]](#)
- Gonçalves, V.S.F.; de Carvalho, V.R. A Review of Interpretability Methods for Gradient Boosting Decision Trees. *J. Braz. Comput. Soc.* **2025**, *31*, 639–653. [\[CrossRef\]](#)
- Lim, H.; Uddin, M.; Liu, Y.; Chin, S.M.; Hwang, H.L. A Comparative Study of Machine Learning Algorithms for Industry-Specific Freight Generation Model. *Sustainability* **2022**, *14*, 15367. [\[CrossRef\]](#)
- Ofori, I.K.; Obeng, C.K.; Asongu, S.A. What Really Drives Economic Growth in Sub-Saharan Africa? Evidence from the Lasso Regularization and Inferential Techniques. *J. Knowl. Econ.* **2024**, *15*, 144–179. [\[CrossRef\]](#) [\[PubMed\]](#)
- Moscattelli, M.; Parlapiano, F.; Narizzano, S.; Viggiano, G. Corporate default forecasting with machine learning. *Expert Syst. Appl.* **2020**, *161*, 113567. [\[CrossRef\]](#)
- Kanapickienė, R.; Kanapickas, T.; Nečiūnas, A. Bankruptcy Prediction for Micro and Small Enterprises Using Financial, Non-Financial, Business Sector and Macroeconomic Variables: The Case of the Lithuanian Construction Sector. *Risks* **2023**, *11*, 97. [\[CrossRef\]](#)
- Zhao, Y.; Lin, D. Prediction of Micro- and Small-Sized Enterprise Default Risk Based on a Logistic Model: Evidence from a Bank of China. *Sustainability* **2023**, *15*, 4097. [\[CrossRef\]](#)
- Jaffee, D.; Stiglitz, J. Credit rationing. In *Handbook of Monetary Economics*; Friedman, B.M., Hahn, F.H., Eds.; Elsevier: Amsterdam, The Netherlands, 1990; Volume 2, pp. 837–888. [\[CrossRef\]](#)

23. Garcia, F.T.; ten Caten, C.S.; Campos, E.A.R.d.; Callegaro, A.M.; Pacheco, D.A.d.J. Mortality Risk Factors in Micro and Small Businesses: Systematic Literature Review and Research Agenda. *Sustainability* **2022**, *14*, 2725. [\[CrossRef\]](#)
24. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [\[CrossRef\]](#)
25. Liu, M.; Hu, Y.; Li, C.; Wang, S. The influence of financial knowledge on the credit behaviour of small and micro enterprises: The knowledge-based view. *J. Knowl. Manag.* **2023**, *27*, 208–229. [\[CrossRef\]](#)
26. Huang, Y.; Shen, Y.; Cheng, D.; Chen, X. Assessing Effectiveness of Structural Monetary Policy in China. *Asian Econ. Pap.* **2023**, *22*, 127–146. [\[CrossRef\]](#)
27. Gasparèniènè, L.; Remeikienè, R.; Williams, C.C. Theorizing the Informal Economy. In *Unemployment and the Informal Economy*; Springer Briefs in Economics; Springer: Cham, Switzerland, 2022. [\[CrossRef\]](#)
28. Behera, S.K.; Panda, R.K.; Senapati, S. Role of Financial and Social Capital in Rural Women Micro-Enterprises: Assessing Entrepreneurial Orientation as a Performance Catalyst. *J. Enterprising Communities People Places Glob. Econ.* **2026**, *20*, 59–93. [\[CrossRef\]](#)
29. Li, J.; Wei, L.; Zhu, X. Basic Concepts of Bank Risk Aggregation. In *Financial Statements-Based Bank Risk Aggregation*; Innovation in Risk Analysis; Springer: Singapore, 2022. [\[CrossRef\]](#)
30. Dhingra, D.; Sharma, S. Operational risk and resilience: Insights from banking case studies. In *Risk, Reliability and Resilience in Operations Management; Advances in Reliability Science*; Elsevier: Amsterdam, The Netherlands, 2025; pp. 121–154. [\[CrossRef\]](#)
31. Wang, W.; Guedes, M.J. Firm failure prediction for small and medium-sized enterprises and new ventures. *Rev. Manag. Sci.* **2025**, *19*, 1949–1982. [\[CrossRef\]](#)
32. Liu, W.; Liu, Y. Worst-Case Higher Moment Coherent Risk Based on Optimal Transport with Application to Distributionally Robust Portfolio Optimization. *Symmetry* **2022**, *14*, 138. [\[CrossRef\]](#)
33. Narayany, S.C.; Zargar, F.N.; Ali, A. Does Risk–Return Trade-Off Exist in the GCC Stock Markets? *Int. J. Emerg. Mark.* **2025**. [\[CrossRef\]](#)
34. Garcia, E.J.; Mulvihill, M.L.; Kharab, M.S.; Stephens, C.L.; Napoli, N.J. Capturing multivariate time series interactions to detect high-risk instability during approach. In Proceedings of the AIAA AVIATION 2023 Forum, Reston, VA, USA, 12–16 June 2023; Paper AIAA 2023-3548. [\[CrossRef\]](#)
35. Kaur, R.; Sharma, M.; Chaturvedi, D.D.; Deka, J. Z-score and logistic model-based default probability prediction in India’s manufacturing sector for economic growth insights. *Int. J. Trade Glob. Mark.* **2025**, *20*, 180–208.
36. Liu, S.; Song, Y.; Xu, Z.; Zhao, Y.; Pan, B.; Xu, D. An Improved Framework: NFMGBM for Enhanced Anomaly Detection. *IEEE Access* **2025**, *13*, 46374–46382. [\[CrossRef\]](#)
37. Wang, C.; Marini, L.; Chin, C.L.; Vance, N.; Donelson, C.; Meunier, P.; Yun, J.T. Social Media Intelligence and Learning Environment: An Open Source Framework for Social Media Data Collection, Analysis and Curation. In Proceedings of the 15th International Conference on eScience (eScience), San Diego, CA, USA, 24–27 September 2019; pp. 252–261. [\[CrossRef\]](#)
38. Celik, E.; Omurca, S.I. A Novel Framework Leveraging Social Media Insights to Address the Cold-Start Problem in Recommendation Systems. *J. Theor. Appl. Electron. Commer. Res.* **2025**, *20*, 234. [\[CrossRef\]](#)
39. Nallakuruppan, M.K.; Balusamy, B.; Shri, M.L.; Malathi, V.; Bhattacharyya, S. An Explainable AI Framework for Credit Evaluation and Analysis. *Appl. Soft Comput.* **2024**, *153*, 111307. [\[CrossRef\]](#)
40. Custer, S. Digital Exhaust or Strategic Asset? Navigating Big Data Benefits and Risks for Public Administration. In *Handbook of Public Administration Reform*; Goldfinch, S.F., Ed.; Edward Elgar Publishing: Cheltenham, UK, 2023; pp. 111–130. [\[CrossRef\]](#)
41. Demuner Flores, M.d.R.; Saavedra García, M.L.; Choy Zevallos, E.E. Chapter 1: The systemic competitiveness of Latin American MSMEs under COVID-19. In *Research in Administrative Sciences Under COVID-19*; Sánchez Limón, M.L., Saavedra García, M.L., Eds.; Emerald Publishing Limited: Leeds, UK, 2022. [\[CrossRef\]](#)
42. Castillo, M.J.; Carpio, C.E.; Rios, A.R.; Garcia, M.; Murguía, J.M. Innovation in the agrifood sector of Latin America and the Caribbean: Agribusiness’ responses to the COVID-19 pandemic. *J. Agribus. Dev. Emerg. Econ.* **2025**, 1–18.
43. Kim, H.; Cho, H.; Ryu, D. Corporate Bankruptcy Prediction Using Machine Learning Methodologies with a Focus on Sequential Data. *Comput. Econ.* **2022**, *59*, 1231–1249. [\[CrossRef\]](#)
44. Rudd, M.A.; Porter, D. Bitcoin Supply, Demand, and Price Dynamics. *J. Risk Financ. Manag.* **2025**, *18*, 570. [\[CrossRef\]](#)
45. Óskarsdóttir, M.; Bravo, C.; Sarraute, C.; Vanthienen, J.; Baesens, B. The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics. *Appl. Soft Comput.* **2019**, *74*, 26–39. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.