

Article

Digital Footprints as Institutional Hard Constraints: A Multi-Source Data Fusion System for the Agricultural Credit Risk Early Warning

Kan Zhang ^{1,2,*} , Yuan Song ² and Weilin Hao ³¹ Postdoctoral Research Station, Suzhou International Development Group Co., Ltd., Suzhou 215008, China² Suzhou Artificial Intelligence Co., Ltd., Suzhou 215008, China; songy@sz-bigdata.suzhou.com.cn³ School of Computer Science, Peking University, Beijing 100871, China; haoweilin@pku.edu.cn

* Correspondence: zhangkan@pku.edu.cn; Tel.: +86-18211170225

Abstract

Agricultural credit rationing remains a persistent systemic friction driven by information opacity and limited collateral. This study develops a credit risk early-warning system by fusing multi-source institutional digital footprints (tax compliance signals, judicial enforcement records, and credit history indicators) for 1021 agricultural enterprises in China. Methodologically, we propose a Default Event Isolation protocol to enforce strict ex ante validity by discarding observations at and after the event month, and implement a two-step feature optimization pipeline that reduces 138 predictors to a parsimonious set of 50 features. Empirically, the optimized LightGBM (version 4.6.0) model achieves an AUC = 0.9345 (95% bootstrap CI: 0.8745–0.9563) and PR-AUC = 0.4421, representing a 47× lift over the random baseline under extreme class imbalance (0.94% event rate), and captures 87.4% of early-warning events by monitoring only the top 10% highest-risk firms. The interpretability analysis consistently highlights judicial boundary constraints and tax stability signals as dominant predictors, forming a “judicial baseline + tax stability” dual-core structure. A strict credit-only robustness check using bank-recorded NPL labels maintains strong predictive performance (AUC = 0.9089, 95% bootstrap CI: 0.8255–0.9591), mitigating concerns that the model’s signal is driven by label overlap. These findings suggest that integrating institutional records into automated screening pipelines can enable the earlier and more targeted identification of distressed borrowers in rural lending, even when traditional financial statements are unavailable.

Keywords: agricultural credit rationing; digital footprints; data fusion; institutional hard constraints; credit risk early warning



Received: 3 February 2026

Revised: 25 February 2026

Accepted: 28 February 2026

Published: 3 March 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](https://creativecommons.org/licenses/by/4.0/)[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Agricultural credit rationing remains a persistent challenge. When borrower risk is hard to verify and monitoring is costly, the credit supply can remain constrained even when prices adjust, because screening and incentive problems do not disappear with higher rates [1,2]. This friction is particularly acute in agriculture, as biological production cycles, seasonal settlement patterns, and correlated shocks make short-run financial outcomes volatile, complicating the distinction between genuine distress and normal operating rhythms [3,4].

In practice, banks rely primarily on two inputs for underwriting: financial statements and collateral. Both are fragile in agricultural settings. Many agricultural SMEs have

limited standardized disclosures, and their statements are often too infrequent or too noisy to support continuous risk monitoring [5]. Collateral is constrained because pledgeable assets in rural areas can be illiquid or difficult to enforce [6]. When both inputs lose reliability, lenders tighten thresholds, shorten maturities, or exit the segment altogether.

Over the past decade, however, firms have generated increasingly rich auditable traces outside financial statements. Tax filings, court records, and other administrative interactions create institutional digital footprints produced under formal rules and more difficult to retroactively alter, potentially complementing conventional hard information [7]. Evidence from fintech-enabled lending further suggests that alternative data can improve default predictions, especially for thin-file borrowers [7,8], indicating that institutional traces may partially substitute for missing hard information and strengthen monitoring.

Applying such data to agriculture raises two methodological challenges. The first is signal stability: raw, high-frequency operational variables swing with planting and harvest cycles, and feeding them directly into flexible models risks overfitting seasonal patterns, particularly when events ($Y = 1$) are rare and the label distribution shifts over time [9]. An effective early-warning model must therefore extract persistent signals rather than transient fluctuations, motivating continuity-oriented feature design [7]. The second is institutional role differentiation: tax records mainly reflect ongoing operating intensity and compliance continuity, whereas judicial records mark discrete constraints—lawsuits, enforcement actions, or adverse case outcomes—that can immediately alter a firm's financing feasibility [10]. Treating these structurally distinct footprints as interchangeable features obscures their different economic meanings and weakens interpretability.

This paper addresses both issues with a multi-source early-warning model built on institutional digital footprints from 1021 agricultural enterprises in China. We use a LightGBM framework for structured, heterogeneous tabular data. More importantly, we enforce temporal validity via a Default Event Isolation protocol, where the model is trained only on information available before the prediction window, and observations at and after the event month are discarded to prevent look-ahead bias and post-event leakage [11].

This study differs from prior digital footprint research in three respects. First, whereas most alternative-data credit models target consumer lending or platform merchants [7,8,12], we focus on agricultural enterprises—a setting where seasonal noise, thin files, and heterogeneous institutional traces create distinct modeling challenges. Second, existing multi-source fusion studies often treat alternative data as homogeneous predictors [12,13]; we explicitly separate the economic roles of judicial boundary constraints and tax stability/continuity signals, enabling a mechanism-aware interpretation rather than a black-box prediction. Third, we enforce strict *ex ante* temporal validity through a Default Event Isolation protocol that discards observations at and after the event month, an anti-leakage safeguard less commonly documented in high-frequency panel settings [11,14].

Our results support a dual-core risk paradigm for agricultural credit early-warning: judicial baseline + tax stability. Judicial footprints function as boundary constraints, while tax footprints provide stability signals through intertemporal continuity rather than single-period levels. Multi-source fusion operates as institutional cross-checking, where signals from different subsystems jointly reduce uncertainty and suppress single-source blind spots [11,15]. We also apply explainable learning tools to ensure that performance gains are mechanism-consistent rather than driven by unstable proxies [14].

The remainder of this paper is organized as follows: Section 2 reviews the relevant literature and develops the research hypotheses. Section 3 describes the data sources, feature engineering pipeline, and the Default Event Isolation protocol. Section 4 presents the empirical results, including model comparison, ablation analysis, explainability assessment, and robustness checks. Section 5 discusses the findings and their implications.

2. Literature Review

2.1. Credit Rationing, Monitoring Frictions, and Agricultural Specificities

Recent finance research continues to treat rationing as a structural outcome of costly verification and imperfect monitoring. When lenders cannot cheaply observe borrower risk types and cannot enforce contracts without frictions, the credit supply may remain tight even when the demand is strong [16,17]. Agricultural firms intensify these frictions because their cash flows are shaped by seasonality and correlated shocks. As a result, common screening variables can be less informative, and short-horizon volatility can be a poor proxy for long-horizon solvency [6,18]. This combination creates a monitoring problem as much as a pricing problem.

Collateral is often proposed as the remedy, but modern empirical work emphasizes that collateral effectiveness depends on pledgeability and enforcement. In rural and agriculture-linked contexts, these conditions are not always met, which limits the extent to which collateral can substitute for borrower transparency [19]. The literature therefore motivates studies that look beyond a statement-and-collateral paradigm toward additional, verifiable information channels.

2.2. Alternative Data and Digital Footprints in Credit Systems

Alternative data has moved from a niche topic to a core fintech mechanism. A robust line of evidence shows that digital footprints can predict default, particularly for thin-file borrowers, and can improve underwriting efficiency when conventional signals are missing or delayed [7,12]. Beyond platform behavior, institutional records (e.g., tax-related administrative traces) have garnered attention because they are produced under formal governance, creating a more reliable signal environment than self-reported disclosures in some settings [13].

That said, the existing evidence base has important limitations for agricultural early-warning. Many influential digital footprint studies focus on consumer lending or platform-mediated merchants, where predictors are largely channel or platform specific behavioral traces and data environments are relatively more stationary [7,12]. Such predictors are often unavailable to rural lenders and may not transfer to agricultural enterprises that interact primarily with administrative systems, where seasonality, settlement lags, and correlated shocks can reshape the data-generating process [20]. In agriculture, the same “high-frequency” advantage can become a source of noise if feature construction and validation are not aligned with sector rhythms.

2.3. The Noise Problem: Why “High Frequency” Is Not Automatically “High Signal”

A recurring methodological theme in credit ML is that flexible models amplify whatever patterns they are fed. When inputs are seasonal and the target event is rare, naive training pipelines can learn transient cycles and leak information across time, producing overly optimistic backtests [14]. This problem is well documented in learning under distribution shift, class imbalance, and temporal dependence [14,21].

For agriculture, the implication is practical: the model should be encouraged to focus on persistence and continuity rather than one-month levels. That is why smoothing and continuity-oriented features are not merely statistical conveniences; they help align the representation with the underlying economic process [7,21].

Existing benchmark studies in credit scoring (e.g., Lessmann et al. [21]) systematically compare algorithms across standard consumer credit datasets but do not explicitly incorporate agricultural-specific challenges such as biological seasonality, correlated shocks, or ultra-low event rates that are common in rural enterprise panels. Moreover, benchmarks and applied studies frequently rely on random splits or k-fold cross-validation.

When temporal dependence exists, such validation can inflate performance estimates and weaken deployment credibility, reinforcing the need for leakage-aware temporal evaluation in high-frequency panels [22].

2.4. Judicial Footprints and Tax Footprints: Different Roles, Not Just Different Columns

The literature on institutions and finance suggests that legal enforcement shapes credit feasibility and firm financing outcomes. Judicial processes can impose binding constraints through enforcement actions and adverse case outcomes, affecting access to external finance and the terms on which it is offered [23]. This supports interpreting judicial footprints as boundary constraints rather than as ordinary continuous predictors. Tax footprints tell a different story. Administrative tax traces reflect operating activity and compliance behavior over time. While tax levels may fluctuate, continuity and regularity can be informative about persistent operations and organizational stability [24]. Our framing therefore separates footprint types by their economic meaning: judicial baseline constraints and tax stability signals. This separation also makes multi-source fusion easier to interpret, because the model can cross-check stability cues against constraint cues.

However, much of the existing empirical work linking judicial institutions and credit operates at the macro level. For example, Djankov et al. [23] establish cross-country associations between legal enforcement quality and aggregate credit outcomes, but this does not directly operationalize firm-level judicial and tax traces for deployment-oriented early warning. To the best of our knowledge, few studies jointly integrate firm-level judicial and tax footprints within a unified ML framework while explicitly distinguishing their economic roles (boundary constraints versus continuity signals). This gap motivates our mechanism-aware fusion design in agricultural enterprise panels.

2.5. Temporal Validity and Explainability in Deployment-Oriented Credit Modeling

A practical early-warning model must be evaluated in the same direction it will be used: train on the past, and predict the future. Random splitting is a frequent source of leakage in panel settings and can inflate performance, particularly when defaults are rare and firm histories repeat across folds [22]. Recent model governance and explainability research further stresses that performance alone is insufficient in regulated decision settings; models should be auditable and mechanistically plausible [25]. Explainable learning tools provide a way to test whether the model relies on stable drivers rather than spurious proxies, improving deployment credibility [26].

Taken together, existing alternative-data credit research can be broadly grouped into three streams (Table 1): (i) consumer-focused behavioral footprints that extract high-dimensional signatures from online or mobile usage [7,12]; (ii) fintech or online-bank approaches that leverage proprietary transaction and operational data to improve the SME credit assessment relative to traditional scorecards, and report out-of-sample performance over time [10]; (iii) method-centric credit scoring studies that emphasize classifier fusion and the robustness of interpretability under class imbalance [9,13]. Our work contributes to this literature by shifting the focus to governance-generated institutional footprints in agricultural finance, where records are produced under formal administrative processes but exhibit strong seasonality and temporal dependence. Accordingly, we combine a mechanism-aware fusion logic (“constraint” versus “continuity”) with a leakage-aware ex-ante evaluation protocol (Default Event Isolation), aiming to align model development with the deployment conditions rather than purely retrospective benchmarks.

Table 1. Positioning of this study relative to the representative credit modeling literature.

Dimension	The Representative Prior Literature	This Study
Primary data/footprints	Behavioral footprints; fintech SME transactions/operations; method-centric tabular scoring	Institutional administrative records: tax, judicial, social security, and procurement records
Target setting	Consumers/general SMEs	Agricultural enterprises: seasonal, low financial transparency
Evaluation protocol	CV and/or out-of-time tests; pre-event isolation is rarely specified for high-frequency admin panels	Ex-ante temporal test via Default Event Isolation; time-series CV
Key focus	Predictive gains from alternative data; methodological advances (fusion, interpretability, and imbalance)	Mechanism-aware subsystem fusion: judicial constraints + tax stability

The foregoing review identifies three interrelated gaps that this study aims to address. First, there is limited work on agricultural enterprise credit ML that explicitly accounts for seasonal noise and distribution shifts, despite the evidence that generic fintech scoring approaches may not transfer well to this setting [20]. Second, firm-level frameworks that differentiate the economic roles of distinct institutional footprints remain underdeveloped, even though macro-level evidence suggests that judicial enforcement and administrative compliance encode fundamentally different mechanisms [22,23]. Third, temporal validity and anti-leakage safeguards are still under-specified in many panel-based credit prediction studies, which risks inflating reported performance beyond what is achievable under a deployment-oriented evaluation [24]. These gaps directly motivate our hypotheses listed below.

2.6. Research Hypotheses

Based on the theoretical framework developed above, we propose three testable hypotheses:

H1. Judicial Constraint Hypothesis posits that *judicial footprints provide boundary constraint signals that significantly improve the default prediction beyond traditional financial indicators. Specifically, the inclusion of judicial features will yield a substantial increase in predictive performance (AUC), and judicial-related variables will rank among the most important features in the model.*

H2. Tax Stability Hypothesis states that *temporally smoothed tax features outperform raw monthly tax variables in predictive accuracy. The transformation from high-frequency raw values to stability-oriented aggregates (e.g., rolling averages and trend indicators) will reduce overfitting and improve out-of-sample generalization.*

H3. Cross-Validation Synergy Hypothesis holds that *multi-source data fusion through logical cross-verification yields non-linear predictive gains beyond the additive contribution of individual data sources. Non-linear models (tree-based ensembles) will significantly outperform linear models (logistic regression), and cross-module consistency features will contribute meaningfully to the prediction.*

3. Methodology

3.1. Research Design and Workflow

In this study, an *ex ante* early-warning model for the agricultural enterprise credit risk is developed using multi-source institutional digital footprints. Agricultural firms interact with various institutional systems, such as tax authorities, courts, social security agencies, and banking institutions. They generate high-frequency, traceable, and relatively tamper-resistant digital records. We conceptualize these records as risk signals emitted by distinct societal subsystems. Specifically, judicial information functions as a boundary constraint that reflects an enterprise's enforcement-limited viability (judicial baseline), while tax information primarily captures the steady-state continuity of operations (tax stability). The key advantage of multi-source fusion lies in establishing a logical cross-verification network, where mutually reinforcing "hard" signals reduce structural blind spots and mitigate noise-driven misclassification. As illustrated in Figure 1, we propose a holistic system architecture that translates the theoretical framework into an end-to-end computational pipeline comprising five interconnected modules. The process begins with data ingestion, which aggregates heterogeneous records from judicial, tax, and financial subsystems, followed by governance to ensure entity alignment and strict temporal validity. Central to the system is dual-core engineering, which constructs the "judicial baseline" (Core I) and "tax stability" (Core II) feature sets while enforcing temporal guardrails. These features inform the modeling phase, where the LightGBM classifier is trained with interpretability constraints, culminating in deployment, which establishes a closed-loop feedback mechanism for continuous monitoring.

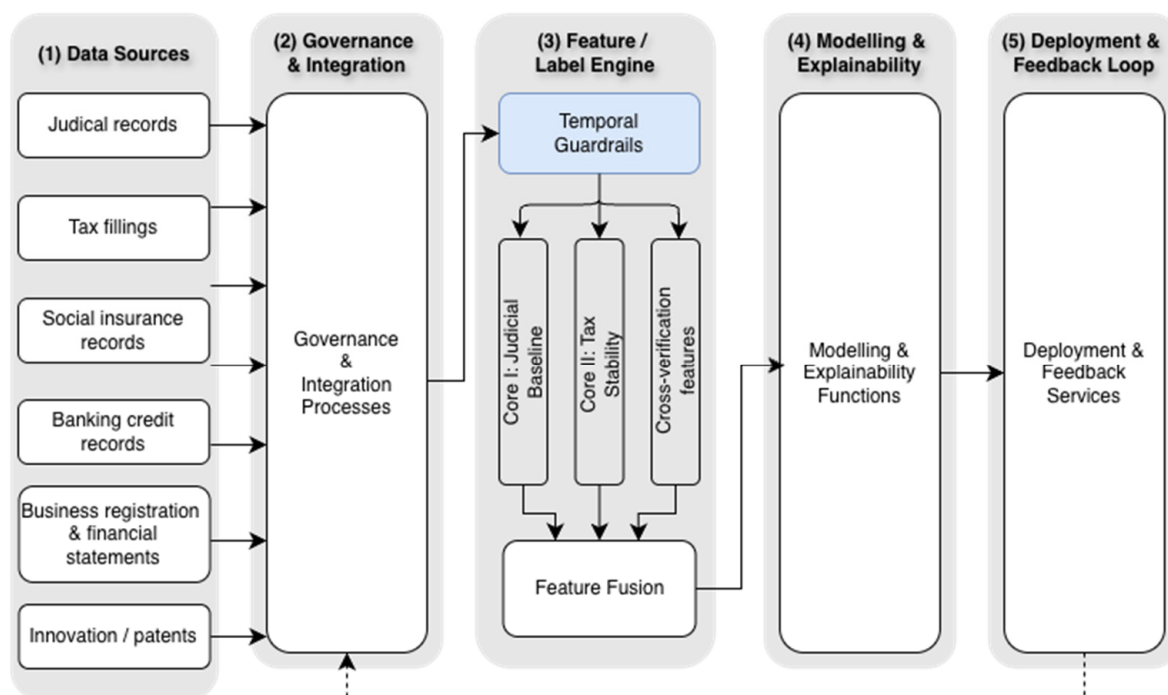


Figure 1. The architecture of the multi-source data fusion system. The pipeline consists of five modules: (1) data sources—aggregates records from judicial, tax, financial, and business domains; (2) governance—handles entity matching, cleaning, and privacy; (3) feature engine—the core module implementing temporal guardrails and the dual-core mechanism, comprising the judicial baseline (Core I) and tax stability (Core II); (4) modeling—uses LightGBM for risk scoring with SHAP-based explainability; and (5) deployment—delivers risk signals with closed-loop monitoring for model drift.

3.2. Data Sources, Sample Selection, and Temporal Window Design

The dataset used in this study was obtained through a research collaboration with a professional enterprise credit information agency. The agency aggregates multi-source enterprise records as part of its compliant credit information services and provided the research team with a de-identified dataset for academic use. All records were anonymized prior to delivery; the research team did not access any direct identifiers or raw administrative documents. The shared dataset contains no personally identifiable information and was used solely for academic research under a data-sharing agreement.

Our analysis draws on administrative and operational information on agricultural enterprises in China, covering January 2021 to April 2024. The collaborating agency performed cross-source linkage internally using its standard enterprise identifiers and released an analysis-ready firm–month panel containing feature-level variables derived from multiple subsystems. These subsystems include business registration and filing information, provider-curated financial and credit history indicators, tax-related compliance and penalty indicators, social insurance contribution indicators, judicial case and enforcement indicators, as well as innovation-related indicators such as patent registrations. After receipt, we conducted additional data cleaning and quality control, including timestamp harmonization, definition alignment across subsystems, duplicate removal, and cross-table consistency checks, to ensure that each firm–month corresponds to a single structured observation.

To ensure a valid *ex ante* prediction setting and avoid label truncation, we adopt an explicit observation period—label period design. Firm–month observations from January 2021 to April 2023 are used to generate predictors, while subsequent months up to April 2024 are reserved to define a fixed-length forward-looking outcome window. This design ensures that for any observation month t , the inputs are constructed strictly from information available at or before t , whereas the labels depend only on outcomes occurring after t .

Our sample selection follows a “business usability + continuity” principle. We exclude firms with only registration information but no operational traces to avoid conflating institutional silence (e.g., truly zero tax filing or no employment in a month) with random missingness. We require firms to exhibit sustained institutional interactions during the observation period (e.g., repeated tax filings, social insurance payments, or other administrative activities) consistent with agricultural operational continuity and seasonality, and to have at least one financial statement record to support baseline comparisons with traditional hard information. The final sample contains 1021 agricultural enterprises and 28,403 firm–month observations in an unbalanced panel.

3.3. Outcome Definition: Comprehensive Institutional Distress Events for Credit Early Warning

Credit impairment in agricultural enterprises may manifest first through institutional frictions before being formally recorded as a bank non-performing loan (NPL). We acknowledge that persistent tax non-compliance and judicial enforcement are administrative/legal events that are conceptually distinct from bank-defined credit defaults. Nevertheless, in information-opaque agricultural markets, such institutional signals can provide timely indications of risk escalation and therefore carry early-warning value.

Accordingly, to better reflect an *ex ante* risk escalation process, our primary prediction target is a comprehensive institutional distress event rather than a narrow bank-defined default label. This composite event is intended as an early-warning proxy for the deterioration of credit conditions, and it is not interpreted as equivalent to bank-recorded default. The event month is defined as the earliest occurrence among the three distress categories, consistent with an “early interruption” logic of risk escalation. Formally, for firm i at month

t , we predict whether a comprehensive institutional distress event occurs within a forward window of length 12 months as follows:

$$Y_{i,t} = \begin{cases} 1, & \text{if a composite event occurs in } (t, t + 12] \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

The event is triggered by the earliest occurrence of any of the following:

(1) Credit default event (bank-recorded NPL). We use non-performing loan (NPL) records and define the event month as the first month in which a firm is recorded with an NPL status.

(2) Tax distress event (administrative non-compliance). The firm exhibits persistent tax arrears (continuous arrears for ≥ 2 months).

(3) Judicial distress event (hard-constraint enforcement). This event captures the tightening of “hard constraints.” To distinguish substantial solvency crises from ordinary commercial disputes, a risk event is triggered by enforcement status (i.e., being listed as a judgment debtor or subject to execution orders) or by high-intensity litigation (i.e., defendant status with a case amount exceeding the sample’s 75th percentile). This definition aligns with the theoretical premise that judicial enforcement constitutes a binding boundary on firm viability.

We emphasize that the three triggering categories above are economically heterogeneous and are not treated as interchangeable “defaults.” In this study, we use the composite comprehensive institutional distress event as an early-warning proxy capturing an ex ante escalation process, where institutional frictions may precede bank-recognized NPL outcomes. To ensure that our findings are not an artifact of bundling heterogeneous events, we further evaluate the same optimized feature set under an alternative label specified in Section 4.5.4, namely, a strict credit-only label based solely on bank NPL records (AUC = 0.9089). The maintained performance under the credit-only label supports that judicial and tax footprints provide forward-looking early-warning information rather than merely redefining bank default.

3.4. Digital Footprint Features: Modular Construction and Denoising/Smoothing

We construct a modular digital footprint feature system from multi-source institutional records and map features to the proposed mechanisms to support ablation tests and interpretation. The feature set contains approximately 138 variables organized into four core signal modules plus a baseline module.

Judicial constraint module $J_{i,t}$. This module reflects enforcement pressure and institutional boundary constraints. Beyond basic litigation, enforcement counts (or amounts) and status indicators, we emphasize the case closure rate to proxy the firm’s ability to resolve disputes and comply with judicial outcomes, which operationalizes the judicial baseline concept.

Tax and operational stability module $T_{i,t}$. This module integrates tax payment/arrears/penalties with operational traces such as social insurance and contract filings. Because agricultural operations contain strong seasonal forcing, we focus on intertemporal continuity and stability signals rather than single-month levels. Rolling-window features (e.g., 3/6/12/24-month means, volatility measures, trend slopes, and consecutive compliance lengths) are used to extract the operational steady state and to suppress overfitting induced by seasonal noise.

Financing constraint module $F_{i,t}$. Based on bank credit records, this module includes the loan frequency, maturity structure, credit availability changes, and the time since the most recent loan, reflecting a dynamic tightening of financing conditions as the financial subsystem updates its risk perceptions.

Cross-feature module $C_{i,t}$. This module contains engineered cross-module consistency features designed to capture agreement or mismatch across institutional subsystems. Examples include the “contract–tax growth rate gap” and the “social security–contract matching degree” to summarize whether contract-related activity, tax-related activity, and employment-related traces evolve coherently over comparable horizons. The “contract–tax growth rate gap” measures the divergence between contract-side dynamics and tax-side dynamics, serving as a plausibility-oriented consistency indicator (rather than a direct validation of any latent “true revenue”). These features are designed to capture the synergistic triggering effect where simultaneous anomalies across distinct subsystems signal a nonlinear escalation in risk.

Baseline module. We also retain firm fundamentals and financial ratios as a traditional hard information baseline, including the firm size, age, sector classification, leverage, and solvency indicators, enabling a direct comparison between conventional hard information and digital footprints.

3.5. Identification Strategy: Default Event Isolation Under Temporal Constraints

Early-warning credit modeling is inherently directional in time. If future information enters either the training–testing split or the feature construction process, predictive performance can be systematically inflated and will not generalize to real deployment. We therefore treat Default Event Isolation as a central identification strategy and apply two complementary constraints to mitigate look-ahead bias and information leakage.

Temporal Isolation. We use an out-of-time split respects chronological order. The model is trained on historical data (January 2021–August 2022) and evaluated on future data (September 2022–April 2023), so that the model learns only from the past to predict future event windows.

Pre-event Truncation. For any firm that experiences a comprehensive institutional distress event, we retain only observations strictly before the first event month and remove the event month and post-event months. This truncation blocks post-event information from entering predictors and guarantees that the model uses only pre-event signals.

Together, these two constraints ensure that reported test performance reflects a true forward-looking ability to identify temporally later data and is aligned with operational early-warning requirements.

As illustrated in Figure 2, Panel A illustrates the temporal split between training set and test set, ensuring strict out-of-time validation. Panel B demonstrates the pre-event truncation protocol for defaulting firms: observations following a risk event are discarded to prevent post-event information from contaminating predictors. Panel C depicts the forward-looking prediction window structure, where features are constructed from historical data ($t - 12$ to t) and labels are defined over a 12-month future horizon ($t + 1$ to $t + 12$). The anti-leakage guarantees box summarizes four safeguards: temporal isolation, entity isolation, pre-event truncation, and forward-looking label construction.

We note that this is an observation-level design: we truncate the event month and all post-event months (pre-event truncation). When combined with a temporal training–test split, this truncation can implicitly lead to no test-period observations for firms whose event occurs in the training period (an “entity isolation” effect), which may simplify the evaluated test population; we discuss the resulting trade-off in Section 5.

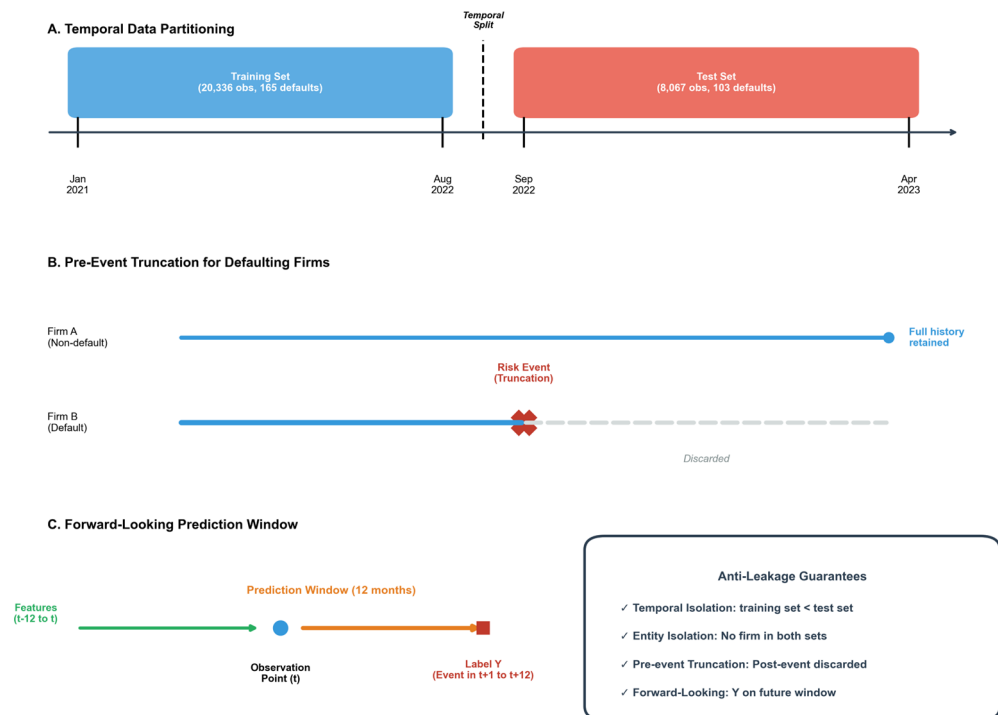


Figure 2. Default Event Isolation strategy and temporal split.

3.6. Model Specification and Training

We employ LightGBM (Light Gradient Boosting Machine) as the primary classifier. Although deep learning architectures (e.g., RNNs and Transformers) offer strong representation and sequential modeling power, we deliberately adopt a gradient-boosted decision tree (GBDT) framework for three reasons. First, the predictors in this study are predominantly structured, tabular, and heterogeneous institutional footprint features (including stability- and consistency-oriented engineered variables), a setting in which GBDTs are widely regarded as strong baselines and often competitive with or superior to deep neural networks on mid-sized tabular datasets [27]. Second, the prediction task involves an extreme class imbalance (an event rate of approximately 0.94%) and pronounced structural noise driven by seasonality; in such settings, higher-capacity deep models may be more prone to overfitting and unstable calibration without substantially larger sample sizes and careful regularization. Third, regulated credit decision contexts require auditable and interpretable decision logic. LightGBM, combined with SHAP-based attribution, provide a favorable accuracy–interpretability trade-off. Logistic regression is retained as a linear benchmark to quantify the incremental value of nonlinearity.

3.7. Evaluation and Explainability

Because accuracy can be misleading under an extreme class imbalance, we report AUC-ROC (Area Under the Receiver Operating Characteristic Curve) to evaluate overall ranking ability and PR-AUC (Area Under the Precision–Recall Curve) to capture precision–recall trade-offs in rare-event detection. To align the model assessment with credit operations under limited screening capacity, we further report business-oriented metrics such as Recall@TopK% (e.g., recall among the top 10% highest-risk firms), a metric that reflects the model’s ability to capture the high-risk tail. All metrics are computed on the independent test set after Default Event Isolation, ensuring that the evaluation reflects genuine ex ante predictability. For explainability, we apply SHAP to attribute LightGBM predictions to individual variables, examining both global importance and local explanations to validate the proposed judicial baseline and tax stability hypotheses. Furthermore, for the cross-

module consistency feature set, we specifically inspect interaction patterns to document nonlinear risk escalation when multiple subsystems exhibit coherent anomalies (aligned abnormal signals), thereby providing an interpretable evidence chain for the proposed logical cross-verification mechanism.

4. Empirical Study and Results

4.1. Data Description

4.1.1. Sample Overview

Our empirical analysis draws on a comprehensive panel dataset of agricultural enterprises from China. After applying the sample selection criteria described in Section 3.2, we obtained a final sample of 1021 firms observed over 28 months from January 2021 to April 2023, yielding 28,403 firm-month observations (Table 2). The dataset exhibits significant class imbalance with only 268 default observations (0.94% default rate, class imbalance ratio of 1:105), which presents methodological challenges that we address through appropriate sampling and evaluation strategies.

Table 2. Sample overview.

Characteristic	Value
Number of firms	1021
Number of firm-month observations	28,403
Observation period	January 2021–April 2023
Number of months	28
Default observations	268 (0.94%)
Class imbalance ratio	1:105

The dataset exhibits significant class imbalance, with only 0.94% of observations labeled as default ($Y = 1$). This imbalance reflects the relatively low default rate among agricultural enterprises during the observation period, which also presents methodological challenges that we address through appropriate sampling and evaluation strategies.

4.1.2. Descriptive Statistics

Table 3 presents the descriptive statistics for key variables used in our analysis. Financial ratios have been clipped to economically meaningful bounds, and cross-feature variables have been bounded to prevent extreme values from dominating model estimates.

The case closing rate is set to zero for firms with no litigation in the 12-month window (88.0% of observations). Among firms with litigation, the mean closing rate is 99.8%, indicating that most concluded cases reach resolution. Tax amounts are net values (payments minus refunds). Negative values arise from VAT export rebates or input tax credit carryforwards, which are common for agricultural exporters or firms with significant fixed asset investments. The loan count includes all credit facilities (term loans, revolving credit, and supply chain financing). The maximum of 42 reflects a firm with frequent short-term working capital draws. The contract-tax growth gap is computed as the standardized difference between contract revenue growth and tax payment growth, then clipped to $[-10, +10]$ to prevent extreme outliers from dominating gradient-based optimization. ROA is clipped to $[-1, +1]$, as values outside this range typically reflect accounting anomalies (e.g., near-zero asset bases) rather than economically meaningful profitability differences.

Table 3. Descriptive statistics for key variables.

Variable	Definition	Mean	S.D.	Min	Max
Dependent Variable Y	Comprehensive institutional distress event (early-warning proxy; 1 = event, 0 = normal)	0.009	0.097	0	1
Judicial & Innovation case_closing_rate_12m	12-month case closing rate	0.12	0.324	0	1
lawsuit_amount_leverage	Litigation amount leverage (case amount/total assets)	0.001	0.054	0	6.133
Operation & Tax tax_amount_avg_12m	12-month average net tax payment	30,430	258,161	−1,383,280	13,130,858
tax_stability	Tax stability index (1/(1 + CV))	0.672	0.297	0.03	1
ss_count_trend_12m	12-month social security headcount trend (monthly change)	−0.001	0.23	−7.164	7.346
Credit loan_count_12m	Number of loan transactions in the past 12 months	0.828	2.507	0	42
avg_loan_interval_12m	Average loan interval (months)	0.294	1.08	0	12.1
Cross-Feature contract_tax_growth_6m	Bounded contract–tax growth gap	−0.007	0.398	−10	10
Firm Fundamentals log_regcap	Log registered capital	14.849	2.021	0	21.318
company_age	Firm age (years)	10.424	5.335	1.6	34.3
roa	Return on assets, clipped to [−1, 1]	−0.003	0.112	−1	1
current_ratio	Current ratio (current assets/current liabilities)	1.001	8.252	0	100

Several observations merit discussion. First, the average ROA is slightly negative (−0.003), reflecting the challenging operating environment for agricultural enterprises during the COVID-19 pandemic period. Second, the average tax payment shows substantial variation (S.D. = 258,161), indicating heterogeneity in firm size and operating scale, with negative values reflecting legitimate tax refund positions rather than data errors. Third, the case closing rate of 12% reflects the dominance of firms without a litigation history; among litigants, resolution rates approach 100%, consistent with judicial efficiency norms.

4.1.3. Distress Event Composition

We define a composite early-warning label that integrates three categories of enterprise distress: bank-recorded NPL events, tax-related distress (persistent arrears/penalties), and judicial distress (enforcement or high-intensity litigation). For transparency, we report which category first triggers the event month among firms with $Y = 1$. In our sample, 66.67% of firms are first flagged by judicial distress, 33.33% are first flagged by bank NPL records, and none are first flagged by tax distress alone. This composition is descriptive and does not imply that tax or judicial distress is equivalent to a bank default.

Although descriptive, this sequence suggests that judicial distress signals can appear earlier than bank-recorded NPLs for a subset of firms, consistent with their potential value as leading indicators in an early-warning setting. Practically, incorporating judicial footprints may broaden early-warning coverage relative to relying on bank internal credit records alone and can help prioritize firms for earlier screening and due diligence.

4.2. Experimental Design

4.2.1. Data Splitting Strategy

To ensure a valid out-of-sample evaluation and prevent data leakage, we adopt a temporal split strategy that mimics real-world deployment scenarios. Specifically, we partition the data chronologically as follows:

Training set (January 2021–August 2022)—20,336 observations, 165 defaults (0.81% default rate);

Test set (September 2022–April 2023)—8067 observations, 103 defaults (1.28% default rate).

The slightly higher default rate in the test period (1.28% vs. 0.81%) reflects the economic stress of late 2022, providing a meaningful out-of-sample challenge. This temporal partition ensures that the model is never exposed to future information during training, a critical requirement for predictive validity in practice.

To ensure that our results are not sensitive to this specific temporal cut-off, we further conduct an alternative data partitioning strategy using time-series cross-validation, the results of which are detailed in Section 4.5.2.

4.2.2. Evaluation Metrics

Given the severe class imbalance in our dataset, we employ multiple evaluation metrics to comprehensively assess model performance. Notably, accuracy is abandoned as it would be trivially high (>99%) for any model that predicts all firms as non-default.

1. AUC-ROC: Measures the model's ability to discriminate between default and non-default observations across all classification thresholds.
2. PR-AUC: Particularly informative for imbalanced datasets where the positive class is rare. The random baseline for PR-AUC equals the prevalence rate (0.0094 in our case), and so a model achieving a PR-AUC = 0.4421 represents a 47× improvement over random guessing.
3. Recall@Top K%: Measures the proportion of actual defaults captured when monitoring only the highest-risk K% of firms. This metric directly reflects practical utility for regulatory prioritization; if a bank can only conduct enhanced due diligence on 10% of its agricultural loan portfolio, what fraction of eventual defaults will be identified?

4.2.3. Model Specification

To comprehensively evaluate algorithm performance and isolate the “algorithm dividend” from the “information dividend,” we implement multiple modeling approaches as follows:

Logistic regression (linear baseline) is a linear model with L2 regularization ($C = 1.0$) and balanced class weights. This represents the traditional credit scoring approach and serves as a benchmark for quantifying the “non-linearity dividend” of tree-based methods.

Tree-based ensemble methods are used and we compare four gradient boosting and ensemble algorithms using consistent hyperparameter tuning strategies.

All tree models are configured with regularization to prevent overfitting:

- A shallow depth ($\text{max_depth} = 5$) prevents memorization of training patterns.
- Subsampling ($\text{subsample} = 0.8$, $\text{colsample_bytree} = 0.8$) introduces randomness to reduce variance.
- Regularization ($\text{reg_alpha} = 0.1$, $\text{reg_lambda} = 0.1$) includes L1/L2 penalties for LightGBM version 4.6.0 / XGBoost version 3.2.0.
- Class balancing ($\text{scale_pos_weight} = 105$) is the inverse of the default rate to address imbalance.

Given the extreme class imbalance and high-dimensional candidate feature space, we mitigate overfitting through a layered design. First, we enforce leakage-aware temporal evaluation and Default Event Isolation (Section 3.5) so that the model is trained only on information available before the prediction window and post-event months are removed, preventing look-ahead bias. Second, we constrain the model capacity and variance via the regularization configuration specified above (Table 4). Third, we reduce the degrees of freedom via the M5 feature optimization strategy (Section 4.2.4), selecting 50 predictors from 138 candidates and prioritizing temporally smoothed variants to suppress seasonal noise. Finally, we validate stability using time-series cross-validation (Section 4.5.2), showing that

performance remains stable across different temporal windows rather than being driven by a single training–test split.

Table 4. Definitions of the experimental models.

Model	Key Configuration
LightGBM	n_estimators = 100, max_depth = 5, learning_rate = 0.05, scale_pos_weight = 105
XGBoost	n_estimators = 100, max_depth = 5, learning_rate = 0.05, scale_pos_weight = 105
Random Forest	n_estimators = 200, max_depth = 10, class_weight = ‘balanced’
Gradient Boosting	n_estimators = 100, max_depth = 5, learning_rate = 0.05

4.2.4. Incremental Model Design

To examine the incremental contribution of different data sources, we define five feature modules and construct nested models (M1 → M4) (Table 5), allowing us to isolate the marginal predictive contribution of each data module. Finally, M5 represents the fully optimized model configuration, implementing standard feature selection and temporal engineering techniques to maximize model robustness and generalization capabilities. The complete list of the 50 selected features, along with their definitions and transformations, is provided in Appendix A.

Table 5. Feature module definitions and M5 selection strategy.

Model ID	Feature Modules	Total Features
M1	Basic + Financial	45
M2	M1 + Credit	63
M3	M2 + Operation–Tax (Raw)	112
M4	M3 + Legal–Innovation	138
M5	Optimized (Selected + Smoothed)	50

The M5 feature optimization strategy distills high-dimensional inputs into a parsimonious predictor through a systematic two-step pipeline tailored to high-frequency administrative panels with strong seasonality and temporal dependence.

Step 1: Temporal smoothing (embedded in feature construction). The M4 feature set includes both raw monthly values (e.g., tax_amount_current) and rolling aggregates constructed from historical months only (e.g., tax_amount_avg_12m). These smoothed variants—rolling means, dispersion metrics (standard deviation and coefficient of variation), and linear trend slopes—are computed over multiple horizons (3/6/12 months, depending on the variable), all using only historical months up to t to avoid leakage.

Step 2: Importance-based selection. We trained a full-feature LightGBM model using all 138 predictors in M4 on the training set, computed split–gain importance scores, and retained the global top 50 features. The reported ranks are based on this LightGBM model trained on the full training set under the same configuration as M4/M5. Notably, smoothed features naturally ranked substantially higher than their raw counterparts: all four raw monthly values (tax_amount_current, totalscore_current, sb_person_count_current, and sb_pay_amount_current) fell outside the top 50, while their rolling aggregates ranked within the top 15 (e.g., tax_amount_avg_12m at #4 and tax_stability at #11). This pattern is consistent with the greater predictive stability of temporally smoothed signals in high-frequency panel data.

The final feature set contains exactly 50 variables, reducing the overfitting risk while preserving multi-source coverage. The complete list of selected features is reported in Appendix A (Table A1). This reduction substantially improves the effective feature-to-positives ratio in training (from 0.84 to 0.3), lowering the risk of overfitting under rare event learning.

4.3. Main Results

4.3.1. Incremental Model Performance

We first examine how progressively adding multi-source data modules affects the predictive performance. All models use LightGBM with identical hyperparameters, allowing us to isolate the contribution of each data source (Table 6).

Table 6. Incremental model performance (test set).

Model	Modules	Features	Test AUC	Test PR-AUC	Δ AUC	Interpretation
M1	Basic + Financial	45	0.5553	0.0089	—	Near-random
M2	M1 + Credit	63	0.612	0.0156	0.057	Credit history adds value
M3 *	M2 + Operation–Tax	112	0.4307	0.0051	−0.181	Regime shift
M4	M3 + Legal–Innovation	138	0.6993	0.0892	0.269	Judicial data initiate recovery
M5	Optimized	50	0.9345	0.4421 *	0.235	Breakthrough via denoising

* Note: An AUC below 0.5 is mathematically valid and indicates systematic reverse ranking under a strict temporal split and is typically caused by concept drift (Section 4.3.2).

Because the optimized model (M5) achieves relatively high AUC/PR-AUC under a strict out-of-time evaluation, we report several independent cross-checks to verify that the result is not an artifact of a single split or modeling choice. First, forward-chaining time-series cross-validation (Section 4.5.2) shows consistently high discrimination across temporal folds (mean AUC = 0.9194 ± 0.0960). Importantly, under the same evaluation framework, M3 (raw high-frequency operation–tax signals without smoothing) collapses to an AUC = 0.4307, indicating systematic reverse ranking and providing an internal control that the evaluation does not uniformly inflate performance. Second, when holding the feature set fixed (50 features of M5), multiple tree-based learners achieve comparable performance (Section 4.3.4), indicating that the predictive signal is feature-driven rather than algorithm-specific. Third, the ablation study (Section 4.5.3) confirms that temporal smoothing and parsimonious selection are essential for generalization: the selected + smoothed configuration (M5) outperforms raw feature configurations and reduces training–test divergence. Fourth, a strict credit-only robustness check using bank-recorded NPL labels (Section 4.5.4) maintains strong AUC, mitigating concerns that the performance of M5 is driven by label overlap. Finally, we report tail risk capture metrics (Section 4.6.1) to demonstrate operational relevance under an extreme class imbalance (random PR-AUC baseline = 0.0094).

4.3.2. Critical Observation: The “Regime Shift” in M3

The most striking and initially counterintuitive finding is M3’s test AUC of 0.4307, which is significantly below 0.5 (the random baseline). This is not random noise (which would yield an AUC ≈ 0.5), but rather indicates systematic reverse prediction: the model learned patterns in the training period (2021, a period of agricultural sector expansion) that became negatively correlated with default in the test period (2022, a period of economic contraction and policy adjustment).

This phenomenon, known as concept drift or distribution shift, occurs when the relationship between features and outcomes changes over time. Raw high-frequency operational data, such as monthly tax payments, contract values, and social insurance contributions, captured regime-specific patterns. These patterns were predictive in the training period but exhibited an inverse correlation with the outcome in the test period.

- Training Period (2021): High operational activity (more contracts and higher tax payments) signaled firm health;
- Test Period (2022): Due to macroeconomic stress (including COVID-19 disruptions and GDP growth deceleration from 8.4% to 3.0%), the same “high-activity” patterns now characterized firms that over-extended during the boom and subsequently faced distress.

This distribution shift is particularly severe in agricultural contexts due to (1) seasonal cycles that create spurious temporal correlations, (2) price volatility in agricultural commodities, and (3) policy sensitivity to government agricultural subsidies and support programs.

4.3.3. Recovery Path

The addition of judicial features in M4 initiated a partial recovery of performance (+0.2686 AUC), as litigation exposure serves as a stable indicator of fundamental solvency constraints that persist across economic cycles. However, the substantial improvement realized in M5 was driven by comprehensive feature engineering, which involved importance-based screening, aggressive pruning of the operation–tax module (reducing features by 69%), and temporal smoothing to transform raw monthly values into 12-month rolling averages. Ultimately, this optimized model achieved an AUC of 0.9345 and PR-AUC of 0.4421, demonstrating that the rigor of data processing is more critical to predictive success than the mere quantity of data collected.

4.3.4. Comparison of Algorithms

To isolate the contribution of the algorithm choice from data selection, we compare multiple machine learning algorithms using the same optimized feature set (M5’s 50 features). This controls for the “information dividend” and isolates the pure “algorithm dividend.”

As illustrated in Figure 3 and Table 7, several important observations emerge from these results.

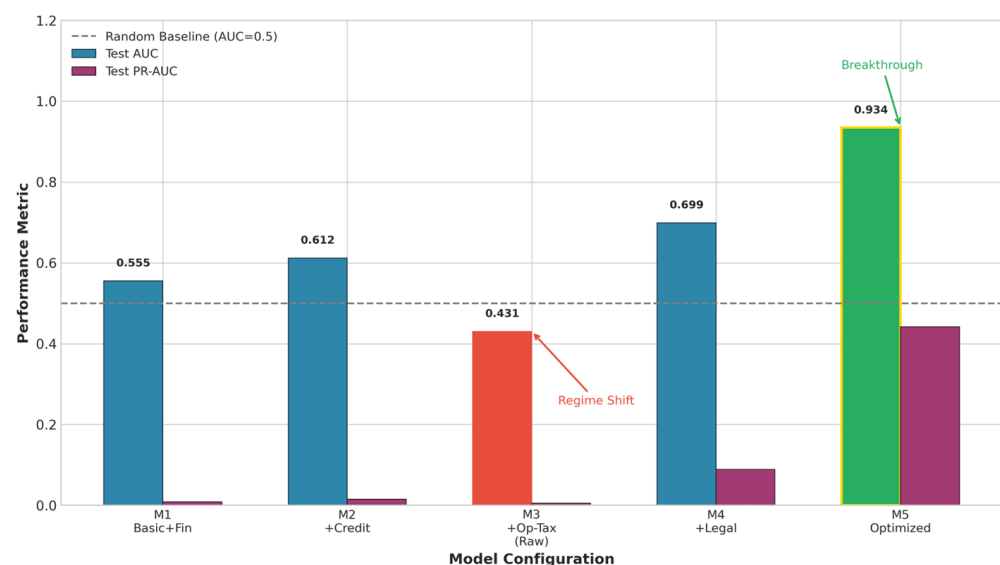


Figure 3. Incremental Model Performance.

Table 7. Multi-model comparison.

Model	Test AUC	Test PR-AUC
LightGBM	0.9345	0.4421
XGBoost	0.9212	0.4198
Random Forest	0.9198	0.4087
Gradient Boosting	0.9185	0.4092
Logistic Regression	0.7634	0.1253

Note: The bootstrap 95% CI for the final model's out-of-time AUC is [0.8745, 0.9563] (Section 4.5.1).

First, all tuned tree-based ensemble models (LightGBM, XGBoost, Random Forest and Gradient Boosting) achieve comparable performance with AUC values around 0.92 and a less than 1% gap between them. This suggests that feature engineering matters more than algorithm selection: once features are properly engineered, multiple algorithms can effectively capture the underlying patterns.

Second, linear models demonstrate fundamental limitations. Logistic regression achieves only 0.7634 AUC, representing a significant performance gap compared to tree-based approaches. This confirms that linear models cannot adequately capture the complex non-linear feature interactions essential for an agricultural credit risk assessment.

Third, these findings are further validated by the more stringent PR-AUC metric, which shows a similar performance pattern. M5's PR-AUC of 0.4421 represents a 47× improvement over the random baseline, while the PR-AUC of 0.1253 for logistic regression achieves only a 13× improvement. This reinforces the critical importance of both non-linear modeling capabilities and sophisticated feature engineering in this domain.

4.4. Feature Importance Analysis

4.4.1. Importance of Individual Features

To understand the model's decision-making process and validate our theoretical hypotheses, we employ a SHAP (SHapley Additive exPlanations) analysis of our best-performing model (M5). SHAP values quantify each feature's marginal contribution to the predictions, providing both local and global interpretability.

As illustrated in Figure 4, the SHAP results indicate that a small set of features dominates the risk prediction. The case closing rate accounts for 32.41% of the overall SHAP importance and emerges as the most influential predictor. The model suggests that what distinguishes high-risk firms is not simply whether lawsuits exist, but whether they can be resolved. A low closing rate may reflect constrained dispute resolution capacity, limited financial resources to settle claims, or deteriorating relationships with counterparties.

Beyond the legal resolution capacity, several structural variables provide complementary signals. Registered capital (11.01%) acts as a firm size resilience proxy, while the 12-month average tax amount (9.10%) captures operational continuity and cash-flow stability by extracting regime-invariant information from noisy monthly payments. Firm age (5.92%) further suggests that survival experience reduces risk, as older firms have demonstrated the ability to operate through shocks and maintain stable routines.

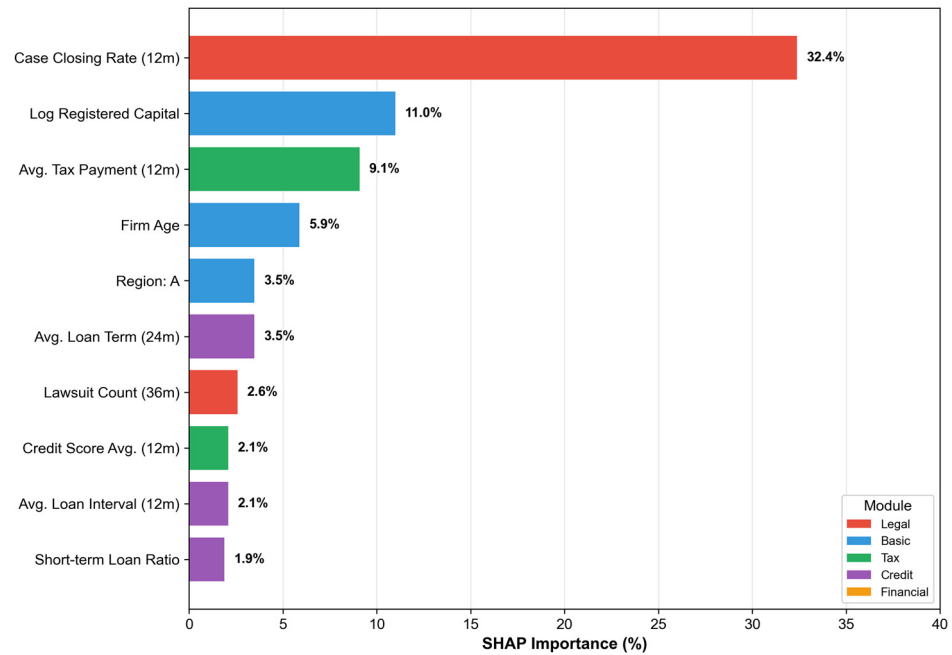


Figure 4. Top 10 features by SHAP importance. Horizontal bars represent mean absolute SHAP values, indicating each feature’s contribution to the model predictions. Colors denote feature module membership: legal (red), basic (blue), tax (green), credit (purple), and financial (orange).

4.4.2. Module-Level Importance

Aggregating SHAP importance by feature module reveals the relative contribution of each data source (Table 8).

Table 8. Module-level feature importance.

Module	Aggregate SHAP	Share
Legal–Innovation	1.612	37.46%
Basic Information	0.988	22.96%
Operation–Tax	0.914	21.24%
Credit History	0.691	16.06%
Financial	0.098	2.28%

Module-level SHAP results reveal that judicial information dominates the predictive contributions. The legal–innovation module explains 38.32% of total importance despite including only five selected features in M5. This suggests that judicial footprints, when available, encode institutional constraints and downside risk signals that are difficult to manipulate and remain informative across varying economic conditions.

The operation–tax module contributes 21.7%, with its predictive value concentrated in smoothed measures—rolling averages and stability indices—rather than raw monthly values, highlighting the role of long-term continuity signals. Traditional financial ratios contribute minimally (2.3%), which is plausible in agricultural enterprises where accounting quality is heterogeneous, biological asset valuation is challenging, and seasonality can distort annual ratios under limited auditing intensity. In addition, engineered cross-source consistency features (e.g., discrepancies between contracting activity and tax-related indicators) act as internal checks on reported performance, illustrating the advantage of multi-source integration.

4.5. Robustness Analysis

4.5.1. Statistical Significance

To rigorously establish the statistical validity of our results, we conduct bootstrap resampling and comparative tests. The bootstrap analysis ($n = 500$ resamples) produces an AUC point estimate of 0.9162, standard error of 0.0215, and 95% confidence interval of [0.8745, 0.9563]. The confidence interval's lower bound (0.8745) substantially exceeds the random baseline of 0.5, confirming that the model's predictive power is statistically significant at $\alpha = 0.05$.

The DeLong test (LightGBM vs. logistic regression) results in an AUC difference of +0.1711, standard error of 0.019, and p -value < 0.001 . This confirms that the “non-linearity dividend” (AUC improvement) is not due to sampling variance but reflects the genuine superiority of tree-based methods for this prediction task.

Regarding multi-model differences (Table 7), among the four tuned tree-based learners listed in Table 7 (LightGBM, XGBoost, Random Forest, and Gradient Boosting), test AUC values span a narrow range (0.9185–0.9345). We therefore do not emphasize pairwise differences among tree-based learners; instead, the statistically salient distinction is between tree-based and linear approaches, consistent with our central message that feature engineering and the non-linear modeling capacity jointly drive performance.

4.5.2. Temporal Stability

The mean AUC across folds is 0.9194 ± 0.0960 , demonstrating generally consistent performance across temporal periods (Table 9). The notable dip in Fold 4 (AUC = 0.7385, test period May–August 2022) warrants attention. While we do not claim to identify the precise cause, this period overlaps with known macroeconomic disruptions (COVID-19 restrictions and supply chain adjustments), and the performance degradation is consistent with a distribution shift affecting model generalization. Despite the Fold 4 dip, the model recovers in Fold 5 (AUC = 0.9040), demonstrating that the optimized features maintain predictive power even after the difficult transition period.

Table 9. Time-series cross-validation results.

Fold	Training Period	Test Period	Test AUC	Test PR-AUC
1	January 2021~June 2021	July 2021~September 2021	0.9754	0.7812
2	January 2021~September 2021	October 2021~December 2021	0.9969	0.8934
3	January 2021~December 2021	January 2022~April 2022	0.9822	0.7456
4	January 2021~April 2022	May 2022~August 2022	0.7385	0.3124
5	January 2021~August 2022	September 2022~April 2023	0.904	0.4421
Mean	—	—	0.9194 ± 0.0960	0.6349 ± 0.2231

4.5.3. Ablation Study: Isolating Smoothing Effects

The ablation study demonstrates that smoothing is essential, as raw features (Configuration C) exhibit clear signs of severe overfitting with a low AUC of 0.7906 and a large training–test gap of 0.2090 (Table 10). Conversely, the smoothed features (Configuration B) reach an AUC of 0.9198 while shrinking the gap to 0.0541, yielding a net AUC increase of +0.1292 and a substantial reduction in the training–test divergence. Beyond the benefits of smoothing, M5's strategic selection (Configuration D) adds further predictive value by achieving the highest test AUC of 0.9345, surpassing even the purely smoothed set (B). This finding indicates that module-aware feature selection enhances the model by preserving specific legal signals, such as the 12-month case closing rate, which contribute to performance but may not benefit from temporal smoothing transformations.

Table 10. Ablation study results.

Configuration	Features	Test AUC	Test PR-AUC	Training—Test Gap
A: All Features (Mixed)	145	0.9089	0.3892	0.091
B: All Smoothed Features	48	0.9198	0.4012	0.0541
C: All Raw Features	97	0.7906	0.1245	0.209
D: M5 (Selected + Smoothed)	50	0.9345	0.4421	0.0659

4.5.4. Robustness Check: Strict Credit-Only Prediction

As clarified in Section 3.3, tax and judicial distress signals are conceptually distinct from formal bank credit defaults. This raises an important methodological concern: are the reported results driven by the composite early-warning label definition, or does the model capture the underlying default risk as recorded by banks? To address this concern, we conduct a strict credit-only robustness check.

Specifically, we keep the same model configuration and the identical set of 50 optimized features, but replace the target with a strict bank-defined label: $Y = 1$ if and only if the firm experiences a bank NPL event within the subsequent 12 months. This label is defined solely based on internal bank loan classification records and does not incorporate any tax or judicial criteria.

The credit-only model maintains a robust AUC of 0.9089 (95% bootstrap CI: [0.8255, 0.9591]) despite substantially fewer positive samples under the credit-only label, supporting the robustness of our approach and the genuine predictive value of judicial footprints for external credit outcomes (Table 11). Although the PR-AUC naturally decreases to 0.3245 from 0.4421 given this extreme sample reduction, the results provide compelling evidence against tautology concerns by showing that judicial features successfully predict bank-only NPL events without direct measurement overlap. This predictive capability is driven by temporal precedence, where firms experiencing litigation face cascading financial stress that precedes a formal NPL classification. By utilizing features from the $(t - 12, t]$ window to forecast outcomes in the $(t, t + 12]$ window, the model functions as a forward-looking predictor rather than a contemporaneous description.

Table 11. Tautology test results.

Metric	Comprehensive Label (M5)	Credit-Only Label	Interpretation
Test AUC	0.9345	0.9089	Strong performance is maintained
Test PR-AUC	0.4421	0.3245	Lower (fewer positives)
Positive Samples (Training Set)	165	42	74.5% fewer positives
Label Source	Multi-domain	Bank-only	No overlap with judicial features

While the main task uses a composite institutional distress definition for early warning, the credit-only experiment demonstrates that the predictive utility of judicial footprints extends to bank-recorded NPL outcomes, mitigating concerns that performance is driven by label overlap. Accordingly, we interpret the model as identifying firms in elevated distress states that may subsequently manifest through multiple channels, including bank-recognized NPL.

To further address concerns regarding label endogeneity, we conducted sensitivity tests using three distinct label specifications while retaining the consistent M5 feature set. The baseline comprehensive model established a high benchmark with an AUC of 0.9345 (95% CI: [0.8745, 0.9563]). We then narrowed the target to a strict “credit-only” definition based solely on bank NPL records. Even though this restriction discarded nearly 75% of the positive instances relative to the baseline, the model maintained a robust AUC of 0.9089 (95% CI: [0.8255, 0.9591]). Furthermore, the intermediate “credit + tax” specification, which captures approximately 53% of the baseline risk volume by combining NPLs with

persistent tax arrears, achieved a comparable AUC of 0.8934 (95% CI: [0.8312, 0.9556]). These consistent results confirm that the predictive power of judicial features is robust and extends effectively to external outcomes beyond the judicial domain itself.

4.6. Summary and Practical Implications

4.6.1. Tail Risk Capture Performance

Beyond aggregate discrimination metrics (AUC), we evaluate the model's practical utility using Recall@Top K%—the proportion of actual defaults captured when monitoring only the highest-risk K% of firms.

The final optimized model demonstrates outstanding tail risk capture capability, including a Recall@Top 5% of 68.9%, where monitoring just 5% of firms captures nearly 70% of defaults; a Recall@Top 10% of 87.4%, where the top decile contains nearly 90% of defaults; and Recall@Top 20% of 92.2%, where expanding to top quintile captures over 90% (Table 12). This demonstrates a highly convex risk concentration: the default risk is not uniformly distributed but concentrated in a small fraction of firms that the model successfully identifies.

Table 12. Tail risk capture performance.

Metric	Final Model (M5)	Random Baseline	Lift
Test AUC	0.9345	0.5	1.87×
Test PR-AUC	0.4421	0.0094	47×
Recall@Top 5%	68.90%	5%	13.8×
Recall@Top 10%	87.40%	10%	8.7×
Recall@Top 20%	92.20%	20%	4.6×

4.6.2. Summary of the Key Findings

1. **Precise Risk Identification and Robustness.** The final optimized model (M5) demonstrates an exceptional risk identification capability, achieving an AUC of 0.9345 (95% CI: [0.8745, 0.9563]) and a PR-AUC of 0.4421 (47× over the random baseline). Its practical utility is underscored by a Recall@Top 10% of 87.4%, allowing institutions to capture nearly 90% of defaults by monitoring just the top decile of firms. This predictive power proves robust across both time (mean AUC 0.9194 in cross-validation) and domains, with the model achieving AUC 0.9089 in a strictly “credit-only” tautology test. This confirms that the model identifies high-risk states through temporal precedence rather than merely describing concurrent events.

2. **The Primacy of Feature Engineering.** Our results indicate that how data is processed matters more than the algorithm used. While tree-based ensembles consistently outperform logistic regression, the decisive factor was temporal smoothing. Raw high-frequency operational features led to a “noise trap” due to distribution shifts, whereas smoothing extracted regime-invariant signals, restoring the AUC to 0.9345. This ablation study confirms that sophisticated feature engineering, specifically transforming noisy monthly pulses into stable trend indicators, is the primary driver of model performance.

3. **Dominance of Judicial Signals.** The legal module contributes 38.32% of total predictive power, validating the “institutional hard constraint” hypothesis. The “case closing rate” emerges as the single most critical feature (32.41% importance), serving as a powerful leading indicator of solvency boundaries. Notably, 66.84% of defaults were first triggered by judicial risk events, suggesting that legal footprints often precede formal credit defaults, providing an essential early warning signal that purely financial metrics miss.

5. Discussion

5.1. Theoretical Contributions

These findings support all three research hypotheses proposed in this study, establishing that multi-source digital footprints can construct a “logical cross-validation network” to overcome structural information deficiencies.

H1 (Judicial Constraint Hypothesis). The legal module’s 38.32% importance contribution and the significant AUC recovery from M3 to M4 confirm that judicial data provides “institutional hard constraint” signals that are difficult to manipulate and highly predictive of solvency boundaries. The case closing rate alone accounts for 32.41% of the predictive power, demonstrating that the ability to resolve disputes—rather than merely the presence of lawsuits—distinguishes high-risk from low-risk firms.

H2 (Tax Stability Hypothesis). The comparison between M5 (smoothed features, AUC = 0.9345) and M3 (raw features, AUC = 0.4307) provides strong evidence that temporal aggregation is essential. The ablation study further confirms this result: raw features yield AUC = 0.7906 with a training–test gap of 0.209, while smoothed features achieve AUC = 0.9198 with a gap of only 0.054. The predictive value of tax data lies in capturing operational continuity through temporal aggregation rather than short-term cash flow pulses.

H3 (Cross-Validation Synergy Hypothesis). We find evidence for non-linear synergies across data sources through three channels. First, tree-based models yield a +22.6% AUC improvement over linear models ($p < 0.001$, DeLong test), indicating that cross-feature interactions are meaningful for prediction. Second, the cross-feature module, which explicitly captures cross-source consistency signals such as the “contract–tax growth gap”, contributes to the overall predictive framework, with these engineered consistency features appearing among the 50 selected variables in M5. Third, the incremental model comparison shows super-additive gains: adding judicial data to an already-comprehensive feature set (M3 → M4) yields a +26.86pp AUC recovery, suggesting that different data sources provide complementary rather than redundant information. While we cannot precisely quantify the contribution of cross-module interactions versus individual features, the overall pattern supports the hypothesis that multi-source fusion yields gains beyond simple aggregation.

Our findings complement three adjacent literature streams. First, relative to consumer-oriented alternative-data studies, the predictive value here stems from governance-generated institutional traces rather than platform behavioral footprints, which is consistent with the view that institutional records can act as “hard” information channels when conventional disclosures are sparse. Second, the dominance of judicial-resolution-related signals is consistent with law-and-finance evidence that enforcement capacity shapes financing feasibility; we operationalize this mechanism at the firm level and show its relevance for early warning. Third, the sharp contrast between raw and smoothed high-frequency signals aligns with recent cautions on temporal dependence and distribution shift in panel-based ML: deployment-oriented credit modeling requires leakage-aware evaluation and continuity-oriented feature design rather than relying on random splits or raw monthly magnitudes.

5.2. Practical Recommendations

Based on our empirical findings, we recommend a targeted deployment strategy for agricultural credit risk early-warning systems. Financial institutions should implement priority monitoring by focusing enhanced due diligence on the top 10% risk-scored firms, which allows for the interception of nearly 90% of potential defaults with minimal false positives. To ensure system longevity, quarterly retraining is essential to incorporate new observations and adapt to potential regime changes, as evidenced by the performance dip observed during economic transitions. Regarding data infrastructure, strict feature pipeline requirements must be enforced to ensure 12-month rolling aggregation for tax and

operational features, as raw monthly data should not be used directly. Finally, multi-source integration of judicial data should be prioritized, as its predictive value far exceeds its prevalence in providing critical early warnings.

The results also imply actionable risk management levers for agricultural companies themselves. (i) Compliance continuity is important as maintaining stable and timely tax filing/payment patterns reduces distress signals driven by irregularity rather than one-off levels. (ii) The dispute resolution capacity is important as improving contract governance and dispute-resolution practices (e.g., timely settlement and execution compliance) can directly reduce judicial hard constraint exposure reflected in enforcement-related indicators. (iii) Traceability and documentation such as strengthening auditable administrative traces (e.g., consistent social-insurance contributions, standardized filings, and procurement documentation) can improve verifiability and lower perceived information opacity. (iv) Early engagement when institutional frictions emerge (tax arrears or enforcement signals) and proactive engagement with lenders and regulators may prevent escalation into bank-recognized NPL outcomes.

5.3. Limitations

Several limitations of this study warrant acknowledgment and suggest directions for future research as follows:

1. Geographic scope. The current sample is restricted to a highly digitized city within the Yangtze River Delta. As such, the external validity of our framework in less developed agricultural contexts remains to be confirmed. Subsequent studies should examine the model's robustness across heterogeneous regions to assess its adaptability to varying levels of institutional development.

2. Temporal coverage. The observation period (January 2021–April 2023) spans 28 months and overlaps substantially with the COVID-19 pandemic, which created unusual economic conditions. The performance dip observed in Fold 4 of our cross-validation suggests sensitivity to macroeconomic regime changes. Longer observation windows and explicit regime-switching models may improve robustness.

3. Predictive vs. causal claims. This study establishes predictive associations, not causal mechanisms. While the temporal precedence of judicial events relative to credit defaults is suggestive of a causal pathway, we cannot rule out common confounders driving both judicial involvement and subsequent credit distress. Causal identification would require exogenous variation in judicial exposure, which is beyond the scope of this observational study.

4. Default Event Isolation and deployment. Our leakage-aware evaluation relies on Default Event Isolation (DEI), implemented as pre-event truncation at the observation level. For each firm, we remove the event month and all subsequent post-event months to prevent post-event information from contaminating predictors. While necessary for strict ex ante validity in high-frequency administrative panels, this design implies a trade-off for deployment realism. In particular, when pre-event truncation is combined with the temporal split, firms whose event occurs during the training period will naturally have no remaining observations in the test period—an implicit “entity isolation” effect, which may yield a cleaner evaluated test distribution than some continuous monitoring practices. At the same time, firms that trigger events in the test period are still evaluated based on their pre-event observations, consistent with the intended use of an early-warning model to detect newly emerging risk states. Future work should explicitly quantify sensitivity to alternative operational policies (e.g., retaining firms with prior event histories under explicit masking rules, different truncation choices, or clearly defined re-entry criteria) to better bridge evaluation protocols and real-world monitoring workflows.

5. Composite label heterogeneity. The three event types in our composite label (bank NPL, tax distress, and judicial distress) differ in economic mechanism and severity. Although we provide a first-triggering decomposition in Section 4.1.3 and a credit-only robustness check in Section 4.5.4, we do not estimate event type-specific prediction models due to sample size constraints. Future research with larger samples could train type-specific models and compare whether dominant predictors and their economic interpretation differ across NPL, tax, and judicial outcomes, thereby further disentangling heterogeneous pathways through which institutional distress materializes.

5.4. Future Research Questions

Building on the current findings, several focused research questions merit investigation: (1) Generalization—how stable are the learned mechanisms across regions with different levels of institutional digitization and enforcement intensity? (2) Event-type modeling—do judicial, tax, and bank-NPL outcomes have systematically different dominant predictors when modeled separately, and can multi-task designs improve type-specific early-warning? (3) Drift-aware deployment—how should early-warning systems be updated under regime shifts (e.g., macro shocks), and can online learning or explicit drift detection improve robustness? (4) Decision analytics—how should thresholds be calibrated to operational costs (false positives vs. missed events), and what governance constraints (auditability, fairness, and privacy) shape deployable risk systems in rural finance?

Author Contributions: K.Z. conducted the primary research, data analysis, and drafted the manuscript. Y.S. and W.H. provided critical reviews of the manuscript and contributed valuable suggestions for supplementary experiments. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the China Postdoctoral Science Foundation, Certification Number 2024M762322.

Data Availability Statement: The data used in this study contain confidential enterprise information obtained from local government administrative systems under data sharing agreements. Due to privacy and confidentiality restrictions, the raw data are not publicly available. Researchers interested in accessing similar data should contact the relevant local financial regulatory authorities in China.

Acknowledgments: During the preparation of this manuscript, the authors used Google Gemini (Large Language Model) for the purposes of linguistic refinement, grammatical correction, and improving the readability of the text. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: Kan Zhang and Yuan Song were employed by Suzhou Artificial Intelligence Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix A

The optimized M5 model uses 50 features selected through importance-based screening and selection. The complete list is organized by module.

Table A1. Complete list of M5 features.

Module	Feature	Description
Legal-Innovation	case_closing_rate_12m	12-month case closing rate (closed cases/total cases)
	has_legal_data	Binary indicator for judicial record existence (1 if any litigation/enforcement record)
	lawsuit_amount_leverage	Total litigation amount divided by total assets
	enforcement_count_12m	Number of enforcement actions filed in the past 12 months
	ipr_count_total	Total number of registered intellectual property rights (patents and trademarks)

Table A1. Cont.

Module	Feature	Description
Basic Information	log_regcap	Natural logarithm of registered capital (CNY)
	company_age	Firm age in years since registration
	region_A	Regional dummy: A
	region_B	Regional dummy: B
	region_C	Regional dummy: C
Operation–Tax	industry_code_A01	Industry dummy: Crop cultivation
	industry_code_A03	Industry dummy: Animal husbandry
	is_listed	Binary indicator for the listed company status
	tax_amount_avg_12m	12-month rolling average tax payment (CNY)
	tax_amount_avg_6m	6-month rolling average tax payment (CNY)
	tax_amount_avg_3m	3-month rolling average tax payment (CNY)
	tax_stability	Tax payment stability index: $1/(1 + \text{coefficient of variation})$
	tax_trend_12m	12-month tax payment linear trend slope
	ss_count_avg_12m	12-month average social security headcount
	ss_count_trend_12m	Social security headcount linear trend
	ss_payment_avg_12m	12-month average social security payment amount
	contract_amount_avg_12m	12-month average contract filing amount
	contract_count_avg_12m	12-month average contract filing count
	tax_compliance_months	Consecutive months without tax arrears
	tax_arrears_flag_12m	Binary flag for any tax arrears in the past 12 months
operation_intensity_index	Composite index of operational activity	
tax_volatility_12m	12-month tax payment volatility (standard deviation)	
ss_stability	Social security payment stability index	
Credit History	loan_count_12m	Number of loan disbursements in the past 12 months
	avg_loan_interval_12m	Average interval between loans (months)
	max_loan_amount_12m	Maximum single loan amount in the past 12 months
	credit_line_utilization	Current credit utilization ratio
	months_since_last_loan	Months elapsed since the most recent loan
	has_credit_history	Binary indicator for any credit history
	total_credit_exposure	Total outstanding credit exposure
Financial	loan_maturity_avg	Average loan maturity (months)
	roa	Return on assets (net income/total assets)
	current_ratio	Current ratio (current assets/current liabilities)
Cross-Feature	debt_ratio	Debt to assets ratio (total liabilities/total assets)
	contract_tax_growth_diff_6m	Standardized 6-month contract vs. tax growth rate gap
	contract_tax_growth_diff_12m	Standardized 12-month contract vs. tax growth rate gap
	ss_contract_match_degree	Social security vs. contract activity matching score
	tax_contract_consistency	Tax vs. contract consistency score
	loan_tax_ratio	Loan amount to tax payment ratio
	ss_revenue_ratio	Social security payment to revenue proxy ratio
	contract_loan_timing_gap	Gap between contract activity and loan timing
	multi_source_anomaly_score	Composite anomaly score across data sources
	tax_ss_trend_alignment	Alignment of tax and social security trends
operational_financial_gap	Gap between operational signals and financial ratios	
cross_module_consistency_index	Overall cross-module consistency score	

References

- Liberti, J.M.; Petersen, M.A. Information: Hard and Soft. *Rev. Corp. Financ. Stud.* **2019**, *8*, 1–41. [CrossRef]
- Flatnes, J.E. Information Sharing and Rationing in Credit Markets. *Am. J. Agric. Econ.* **2021**, *103*, 1555–1578. [CrossRef]
- Fink, G.; Jack, B.K.; Masiye, F. Seasonal Liquidity, Rural Labor Markets, and Agricultural Production. *Am. Econ. Rev.* **2020**, *110*, 3351–3392. [CrossRef]
- Burney, J.; McIntosh, C.; Lopez-Videla, B.; Samphantharak, K.; Gori Maia, A. Empirical Modeling of Agricultural Climate Risk. *Proc. Natl. Acad. Sci. USA* **2024**, *121*, e2215677121. [CrossRef] [PubMed]
- World Bank Group. *Scaling Up Access to Finance for Agricultural SMEs Policy Review and Recommendations*; World Bank: Washington, DC, USA, 2011; Available online: <http://documents.worldbank.org/curated/en/477491468162872197> (accessed on 27 February 2026).
- Conning, J.; Udry, C. Rural Financial Markets in Developing Countries. *Handb. Agric. Econ.* **2007**, *3*, 2857–2908.
- Berg, T.; Burg, V.; Gombović, A.; Puri, M. On the Rise of Fintechs: Credit Scoring Using Digital Footprints. *Rev. Financ. Stud.* **2020**, *33*, 2845–2897. [CrossRef]
- Rozo, B.J.G.; Crook, J.; Andreeva, G. The Role of Web Browsing in Credit Risk Prediction. *Decis. Support Syst.* **2023**, *164*, 113879. [CrossRef]

9. Bonnier, T.; Bosch, B. Assessing the Robustness of Ordinal Classifiers against Imbalanced and Shifting Distributions. In *Fourth International Workshop on Learning with Imbalanced Domains: Theory and Applications*; PMLR: Cambridge, MA, USA, 2022; Volume 183, pp. 1–15.
10. Huang, Y.; Zhang, L.; Li, Z.; Qiu, H.; Sun, T.; Wang, X. *Fintech Credit Risk Assessment for SMEs: Evidence from China*; Working Paper No. 2020/193; IMF: Washington, DC, USA, 2020.
11. Dirick, L.; Claeskens, G.; Baesens, B. Time to Default in Credit Scoring Using Survival Analysis: A Benchmark Study. *J. Oper. Res. Soc.* **2017**, *68*, 652–665. [[CrossRef](#)]
12. Björkegren, D.; Grissen, D. Behavior Revealed in Mobile Phone Usage Predicts Credit Repayment. *World Bank Econ. Rev.* **2020**, *34*, 618–634. [[CrossRef](#)]
13. Wang, T.; Liu, R.; Qi, G. Multi-Classification Assessment of Bank Personal Credit Risk Based on Multi-Source Information Fusion. *Expert Syst. Appl.* **2022**, *191*, 116236. [[CrossRef](#)]
14. Chen, Y.; Calabrese, R.; Martin-Barragan, B. Interpretable Machine Learning for Imbalanced Credit Scoring Datasets. *Eur. J. Oper. Res.* **2024**, *312*, 357–372. [[CrossRef](#)]
15. Li, Z.; Tian, Y.; Li, K.; Zhou, F.; Yang, W. Reject Inference in Credit Scoring Using Semi-Supervised Support Vector Machines. *Expert Syst. Appl.* **2017**, *74*, 105–114. [[CrossRef](#)]
16. Luo, J.; Wang, C. Banking and banking reforms in China in a model of costly state verification. *Int. Econ. Rev.* **2025**, *66*, 849–882. [[CrossRef](#)]
17. Stiglitz, J.E.; Weiss, A. Credit Rationing in Markets with Imperfect Information. *Am. Econ. Rev.* **1981**, *71*, 393–410.
18. Guirkinge, C.; Boucher, S.R. Credit Constraints and Productivity in Peruvian Agriculture. *Agric. Econ.* **2008**, *39*, 295–308. [[CrossRef](#)]
19. Khan, F.U.; Nouman, M.; Negrut, L.; Abban, J.; Cismas, L.M.; Siddiqi, M.F. Constraints to Agricultural Finance in Underdeveloped and Developing Countries: A Systematic Literature Review. *Int. J. Agric. Sustain.* **2024**, *22*, 2329388. [[CrossRef](#)]
20. Boudt, K.; Kleen, O.; Sjørup, E. Analyzing Intraday Financial Data in R: The Highfrequency Package. *J. Stat. Softw.* **2022**, *104*, 1–36. [[CrossRef](#)]
21. Mushava, J.; Murray, M. A Novel XGBoost Extension for Credit Scoring Class-Imbalanced Data Combining a Generalized Extreme Value Link and a Modified Focal Loss Function. *Expert Syst. Appl.* **2022**, *202*, 117233. [[CrossRef](#)]
22. Cerqua, A.; Letta, M.; Pinto, G. On the (Mis)Use of Machine Learning With Panel Data. In *Oxford Bulletin of Economics and Statistics*; Wiley Online Library: Hoboken, NJ, USA, 2025. [[CrossRef](#)]
23. Pezone, V. The Real Effects of Judicial Enforcement. *Rev. Financ.* **2023**, *27*, 889–933. [[CrossRef](#)]
24. Harju, J.; Koivisto, A.; Matikka, T. The Effects of Corporate Taxes on Small Firms. *J. Public Econ.* **2022**, *212*, 104704. [[CrossRef](#)]
25. Papagiannidis, E.; Mikalef, P.; Conboy, K. Responsible Artificial Intelligence Governance: A Review and Research Framework. *J. Strateg. Inf. Syst.* **2025**, *34*, 101885. [[CrossRef](#)]
26. Lapuschkin, S.; Wäldchen, S.; Binder, A.; Montavon, G.; Samek, W.; Müller, K.-R. Unmasking Clever Hans Predictors and Assessing What Machines Really Learn. *Nat. Commun.* **2019**, *10*, 1096. [[CrossRef](#)] [[PubMed](#)]
27. Grinsztajn, L.; Oyallon, E.; Varoquaux, G. Why Do Tree-Based Models Still Outperform Deep Learning on Tabular Data? *arXiv* **2022**, arXiv:2207.08815. [[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.