Data-Driven Audit Risk Analytics for Balance-Sheet Recognition of Data Assets under Blockchain-Enabled Evidence Governance

Chuanjie Zhu[1], Jindong Wang[1], Mengkai Wang[2,*]

[1] School of Graduate, Chinese Academy of Fiscal Sciences, Beijing, China 100142

[2] School of Business, Renmin University of China, Beijing, China 100872

* Email: mengkaiw@ruc.edu.cn (Corresponding Author)

Received 31 August 2025 · Revised 15 October 2025 · Accepted 29 November 2025 · Published 30 March 2026

Abstract

The balance-sheet recognition of data assets has become a critical business analytics problem because firms, auditors, and regulators must convert fluid digital resources into verifiable accounting objects without allowing strategic overvaluation, weak evidence, or collusive assurance to distort market signals. Existing studies have mainly examined data asset governance through institutional interpretation, blockchain-enabled traceability, or evolutionary games among firms, auditors, and regulators. This article develops a related but distinct data-driven framework that treats data asset recognition as an audit risk analytics task. We construct a cost-sensitive predictive model that integrates evidence-quality indicators, valuation-dispersion variables, auditor-independence signals, firm-control histories, and blockchain evidence-coverage measures. A calibrated numerical panel of 1,200 data asset recognition applications across six digital-economy sectors is used to compare logistic regression, classification trees, random forest, gradient boosting, extreme gradient boosting, and support vector machines. The best-performing model reaches an AUC of 0.907, a PR-AUC of 0.661, and a cost-weighted F1 score of 0.681, while reducing expected misrecognition loss by 34.7% relative to rule-based screening. Scenario analysis indicates that blockchain evidence governance improves compliance most strongly when it is coupled with cost-sensitive analytics and targeted audit escalation rather than applied as a stand-alone technical ledger. Sensitivity analysis further shows that evidence coverage, valuation dispersion, and auditor-client repetition jointly determine whether regulatory rewards and penalties produce stable compliance. The study contributes to business and data analytics by transforming the governance of data assets from a purely normative or game-theoretic issue into an operational risk-scoring, audit-prioritization, and decision-optimization problem.

Keywords: Data asset recognition; audit risk analytics; blockchain evidence governance; cost-sensitive machine learning; balance-sheet recognition; data valuation; regulatory technology; business analytics

Data-Driven Audit Risk Analytics for Balance-Sheet Recognition of Data Assets under Blockchain-Enabled Evidence Governance

1. Introduction

Data assets increasingly shape enterprise value, but their translation into balance-sheet items remains analytically difficult because data do not behave like ordinary inventories, equipment, or financial instruments. Data resources are reusable, context-dependent, modifiable, replicable, and often jointly produced across departments, platforms, and external partners. These characteristics make recognition decisions vulnerable to inconsistent definitions, weak evidence, and opportunistic valuation. Traditional accounting practice is accustomed to testing the existence, ownership, and recoverability of assets, yet data assets require additional verification of lineage, usage rights, algorithmic dependence, privacy restrictions, commercialization pathways, and future benefit claims. The literature on intangible assets has long shown that financial statements struggle to reflect knowledge-based value, but data assets add a new layer because the object being recognized is simultaneously an economic resource, a technical artifact, and a governance responsibility (Lev & Sougiannis, 1996; Lev & Zarowin, 1999).

The uploaded manuscript that motivates this study examines the balance-sheet recognition of data assets under blockchain-enabled government incentive and penalty schemes. Its main concern is that firms may manipulate data valuations, audit institutions may tolerate or support such manipulation, and regulators may have difficulty detecting misconduct when evidence is dispersed across internal systems and external transactions. It therefore frames the issue as a tripartite evolutionary game in which firms, auditors, and regulators adapt their strategies under different incentive and detection conditions. The present article develops a related but different research route. Instead of asking how actors evolve toward compliance or collusion in a strategic game, it asks how business analytics can generate early-warning risk scores, prioritize audit resources, and quantify the value of blockchain evidence in data asset recognition workflows.

This change of perspective is important for the Journal of Business and Data Analytics because recognition quality is not only an institutional design problem; it is also a data problem. Regulators and audit institutions often face too many data asset applications to examine manually with equal intensity. Each application contains heterogeneous traces: data source logs, right-confirmation documents, valuation reports, revenue forecasts, access-control records, privacy assessments, market comparables, and prior adjustment histories. Some traces are structured, while others are semi-structured or generated by platforms. Business analytics provides a way to integrate those traces into a risk-scoring framework that distinguishes ordinary complexity from suspicious recognition patterns (Chen et al., 2012; Wamba et al., 2017).

The proposed article therefore treats the recognition of data assets as a cost-sensitive classification and decision-optimization task. A recognition application is labeled as high-risk when ex post review reveals serious valuation adjustment, missing right-confirmation evidence, revenue-backward reconstruction, auditor independence concern, or regulator-identified disclosure deficiency. Predictive features are organized into five groups: asset-property features, valuation features, evidence-governance features, auditor-relationship features, and firm-control features. The analytical target is not to replace professional judgment; it is to provide a transparent triage mechanism that directs auditors and regulators toward applications where the expected loss from under-inspection exceeds the marginal cost of additional procedures.

This approach extends earlier research on analytics-enabled decision-making. Big data analytics has been linked to firm value, operational responsiveness, and superior decision quality, but these gains depend on governance mechanisms that convert data into reliable action (Akter et al., 2016; Grover et al., 2018). Audit analytics similarly shows promise in identifying abnormal transactions and financial reporting risk, yet

analytics must be aligned with the specific cost structure of audit failure rather than evaluated only by generic accuracy (Vasarhelyi et al., 2015; Warren et al., 2015). For data assets, false negatives are especially costly because an overstated asset can influence valuation, financing, tax planning, and market confidence before correction occurs.

The paper also repositions blockchain. The motivating manuscript emphasizes blockchain as a governance condition that changes detectability and incentive payoffs. In this article, blockchain is operationalized as an evidence feature layer. Blockchain coverage does not automatically imply compliance; it measures the proportion of recognition evidence whose origin, timestamp, transformation, access, and authorization history are verifiable in a permissioned ledger. In this sense, blockchain becomes one input into the risk analytics model rather than an assumed perfect detector. This interpretation is consistent with the broader view that blockchain governance depends on institutional configuration, interoperability, privacy design, and off-chain evidence quality (Christidis & Devetsikiotis, 2016; Risius & Spohrer, 2017).

The study makes five contributions. First, it constructs a new article topic directly related to data asset recognition but clearly differentiated from tripartite evolutionary game modeling by focusing on predictive audit risk analytics. Second, it develops a feature taxonomy for data asset recognition risk that links accounting evidence, data governance, valuation dispersion, and auditor independence. Third, it compares multiple machine learning models under cost-sensitive evaluation rather than relying only on statistical accuracy. Fourth, it uses scenario analysis to quantify the incremental value of blockchain evidence governance, analytics-based triage, and incentive-compatible audit escalation. Fifth, it provides practical guidance for firms, auditors, and regulators seeking to move from broad policy statements to measurable, data-driven supervisory workflows.

2. Literature Review

Research on data asset recognition is connected to a broader stream on intangible value and the limits of historical-cost accounting. Intangible-intensive firms often create value through knowledge, data, software, customer relationships, and network effects, but financial reporting systems have difficulty representing such assets in a comparable and timely way. Earlier studies of research and development capitalization show that market participants value information about internally generated knowledge even when accounting standards require conservative recognition (Lev & Sougiannis, 1996). The deterioration of value relevance in some reporting contexts further suggests that reporting rules designed for tangible assets do not always keep pace with knowledge-based production (Lev & Zarowin, 1999). Data assets intensify this mismatch because they may be non-rival, repeatedly reused, embedded in algorithms, and valuable only under particular contractual or analytical scenarios.

The governance of data assets also intersects with information systems research on data quality, data governance, and analytics capability. Data governance establishes decision rights, accountability structures, quality standards, and control processes for enterprise data, while analytics capability converts those resources into business value. The organizational literature emphasizes that data governance is not merely technical standardization; it defines who has authority over data definitions, access, lineage, stewardship, and quality remediation (Khatri & Brown, 2010; Otto, 2011). For accounting recognition, governance maturity becomes especially important because weak lineage, inconsistent metadata, and undocumented transformations undermine claims of identifiability, control, and expected future benefit.

The business analytics literature adds a second foundation. Analytics creates value when organizations combine data availability, technical methods, domain expertise, and decision routines. The mere possession of large data volumes does not automatically improve performance; value arises when data are filtered, interpreted, and connected to decisions with explicit objectives (George et al., 2014; Günther et al., 2017). In the recognition of data assets, the objective is not revenue prediction alone but the reduction of

misrecognition risk. This requires a risk analytics framework that defines the target event, identifies decision costs, evaluates calibration, and supports explanations that auditors and regulators can understand.

The role of blockchain has been widely discussed in supply chains, financial technology, and information systems. Blockchain-based systems can provide immutable logs, distributed verification, and shared transactional records, which may reduce information asymmetry among authorized participants (Yli-Huumo et al., 2016; Casino et al., 2019). In supply chains, blockchain has been linked to traceability, transparency, and coordination, but the literature also emphasizes that technical feasibility must be evaluated against implementation costs and governance constraints (Saberi et al., 2019; Treiblmaier, 2018). The same logic applies to data asset recognition: a ledger can strengthen evidence integrity, but it cannot automatically validate valuation assumptions or resolve ambiguous economic benefits.

Accounting and auditing studies on blockchain show that distributed ledgers may reshape assurance by making transaction evidence more persistent and by enabling near-real-time verification. However, these studies also warn that auditors must still evaluate business substance, internal controls, and off-chain events. Blockchain improves the audit trail for recorded events, but it does not prove that recorded events are economically meaningful or complete. This distinction is central to the present article. Blockchain coverage is treated as a measurable evidence attribute whose predictive value can be tested, not as a universal solution that eliminates audit risk (Dai & Vasarhelyi, 2017; Bonsón & Bednárová, 2019).

Audit analytics has developed rapidly because manual sampling is increasingly inadequate for digital business environments. Continuous auditing, anomaly detection, and predictive models can support auditors by screening populations and highlighting unusual transactions. Research on material misstatements and earnings manipulation demonstrates that predictive signals can be extracted from financial ratios, accruals, governance indicators, and historical restatement patterns (Beneish, 1999; Dechow et al., 2011). Data asset recognition introduces analogous warning signs, including abnormal valuation multiples, insufficient external comparables, repeated use of the same auditor, and weak linkage between data processing logs and revenue claims.

Machine learning methods offer a toolkit for such screening. Logistic regression provides interpretability and stable baseline performance; classification trees support rule extraction; random forests and gradient boosting capture nonlinear interactions; support vector machines can model high-dimensional boundaries; and probability calibration supports decision thresholds. The model choice should depend not only on AUC but also on cost-weighted performance, robustness to class imbalance, and explainability (Breiman, 2001; Friedman, 2001). For audit risk analytics, an accurate but opaque model may be less useful than a slightly weaker model that generates reliable explanations and stable escalation decisions.

The literature on information asymmetry and regulation provides a third theoretical foundation. When one party holds private information about quality, cost, or risk, markets may produce inefficient outcomes unless signaling, screening, monitoring, or incentive alignment mechanisms are introduced (Akerlof, 1970; Spence, 1973). Data asset recognition is a typical asymmetric-information environment: firms know more about the internal generation and commercial use of data, auditors know more about their actual procedures, and regulators may see only summarized disclosures. Analytics can reduce asymmetry by transforming dispersed evidence into comparable risk indicators, while blockchain can reduce manipulation opportunities by preserving evidence histories.

Game-theoretic and evolutionary perspectives remain relevant even though this article is not a game model. Incentives, penalties, and repeated interactions affect whether firms disclose faithfully, auditors remain independent, and regulators allocate attention effectively. Evolutionary game theory explains how strategy shares change under bounded rationality and payoff adaptation (Taylor & Jonker, 1978; Hofbauer & Sigmund, 2003). The present article complements that logic by showing how predictive analytics can

provide the informational foundation for such incentives. Without credible risk measurement, reward and penalty systems may be misdirected, excessive, or vulnerable to regulatory capture.

When data become strategic production factors, recognition decisions become part of enterprise information integration, cyber risk management, and analytics governance.

The literature nevertheless leaves a clear gap. Existing studies discuss how data assets should be recognized, how blockchain may increase traceability, and how regulatory incentives may deter manipulation. Far fewer studies convert those insights into an operational predictive model that auditors and regulators can use to prioritize cases. Similarly, research on audit analytics often focuses on conventional financial statement fraud, not on the emerging category of data asset valuation. This article addresses that gap by developing a cost-sensitive risk analytics framework that connects data governance attributes to audit and regulatory decisions (Shmueli & Koppius, 2011; Mullainathan & Spiess, 2017).

3. Research Framework and Analytical Logic

The recognition of data assets can be conceptualized as a pipeline that begins with internal data identification and ends with disclosure, audit opinion, and possible regulatory review. At each stage, evidence may be complete, incomplete, manipulated, or difficult to compare. The risk analytics framework therefore follows the workflow rather than treating the application as a single static document. Source evidence describes where the data were generated and whether the firm has legal or operational control. Processing evidence describes transformations, cleaning, aggregation, and algorithmic use. Commercial evidence connects the data resource to revenue, cost savings, or licensing opportunities. Assurance evidence records auditor procedures, sampling coverage, and independence signals. Regulatory evidence records filings, corrections, and enforcement histories.

The framework adopts three design principles. First, features should be auditable. A variable is useful only when it can be traced to a document, log, contract, or system record. Second, predictions should be actionable. A high-risk score should trigger a clear audit or regulatory response, such as expanded right-confirmation testing, independent valuation review, blockchain evidence reconciliation, or auditor rotation assessment. Third, evaluation should be cost-sensitive. A false negative that allows a manipulated data asset onto the balance sheet is more damaging than a false positive that leads to additional review. This asymmetry justifies threshold optimization based on expected loss rather than a fixed 0.5 classification threshold.

The central dependent variable is high-risk recognition. It equals one when subsequent review identifies material valuation adjustment, insufficient right-confirmation evidence, unverifiable data lineage, auditor independence concern, or regulator-required disclosure correction. It equals zero when the recognition application passes subsequent verification without material adjustment. This definition is broader than fraud because many high-risk cases may arise from uncertainty, weak documentation, or inadequate governance rather than deliberate misconduct. The broader definition is more useful for business analytics because the purpose is early warning and triage rather than legal attribution of intent.

The model includes five feature blocks. The asset-property block captures the number of data sources, update frequency, replicability, dependence on personal information, and algorithmic embeddedness. The valuation block includes valuation dispersion, revenue forecast dependence, market-comparable availability, and sensitivity of the fair-value estimate to assumed commercialization rates. The evidence-governance block includes lineage coverage, metadata completeness, access-control maturity, and blockchain evidence coverage. The auditor-relationship block includes tenure, fee dependence, prior client relationship, and frequency of repeated valuation experts. The firm-control block includes prior disclosure adjustments, internal audit maturity, cyber incident history, and data stewardship accountability.

Expected loss is defined as EL(t) = C_FN x FN(t) + C_FP x FP(t) + C_A x Audit(t), where t is the risk-score threshold, C_FN is the cost of a false negative, C_FP is the cost of a false positive, and C_A is the marginal cost of additional audit escalation. The decision rule selects the threshold that minimizes expected loss subject to a minimum recall constraint for high-risk cases. This objective reflects the economics of audit and regulation more directly than accuracy. In many recognition environments, high-risk applications are relatively rare, so a high accuracy score may be achieved by simply predicting most cases as ordinary. Precision-recall and cost-weighted metrics are therefore emphasized (Fawcett, 2006; Davis & Goadrich, 2006).

Figure 1 summarizes the proposed logic. Data asset applications first enter a standardized evidence layer. Evidence is then mapped into risk features, including data lineage, valuation dispersion, auditor relationship, and blockchain coverage. Predictive models estimate high-risk probability. A cost-sensitive triage layer converts the score into audit actions: routine review, enhanced evidence testing, independent valuation challenge, or regulatory escalation. The final feedback loop updates the model using outcomes from subsequent reviews, making the system a learning governance mechanism rather than a one-time screening tool.

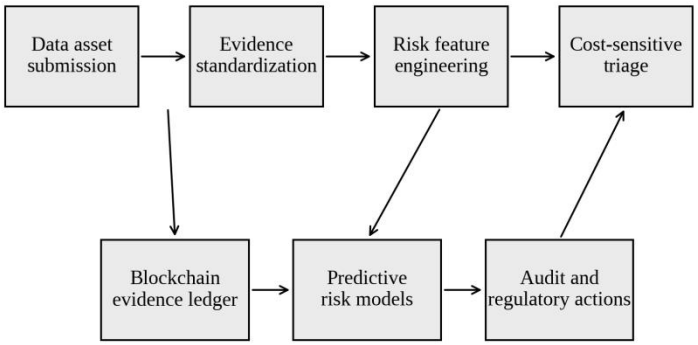


Figure 1. Data-Driven Audit Risk Analytics Framework for Data Asset Recognition

The framework converts the recognition workflow into a set of measurable checkpoints. It also creates a feedback loop in which audit outcomes refine future risk scoring. This is important because data asset markets are still evolving and fixed rules may become obsolete when firms develop new commercialization models or data-sharing arrangements. Table 1 summarizes the risk taxonomy for data asset balance-sheet recognition.

Table 1. Risk Taxonomy for Data Asset Balance-Sheet Recognition

Risk dimension	Observable indicators	Audit implication	Analytics treatment
Existence and control	Source logs, right-confirmation files, access authorization, ownership boundaries	Weak evidence raises risk of fabricated or overstated assets	Lineage and rights-completeness scores
Valuation reasonableness	Dispersion across cost, income, and market methods; sensitivity to commercialization	High dispersion requires independent valuation challenge	Valuation dispersion and comparables variables

Risk dimension	Observable indicators	Audit implication	Analytics treatment
	assumptions		
Economic benefit	Revenue linkage, cost-saving evidence, customer contract dependence	Unverifiable benefits increase impairment and correction risk	Forecast dependence and unverifiable revenue share
Auditor independence	Tenure, fee concentration, repeated valuation experts, prior adjustment patterns	Concentrated relationships may weaken skepticism	Auditor-client repetition and independence score
Evidence integrity	Timestamping, hash records, system logs, transformation records, blockchain coverage	Incomplete evidence limits ex post verifiability	Blockchain evidence coverage and metadata completeness

4. Data Design, Variables, and Methodology

Because confidential firm-level data asset recognition files are not publicly available, this study uses a calibrated numerical panel designed to reflect common patterns observed in enterprise data governance, audit analytics, and regulatory screening. The purpose is not to claim external empirical validation for a specific country or market. Rather, the calibrated panel allows the article to demonstrate how a data-driven audit risk framework can be implemented, evaluated, and stress-tested. This approach is common in analytical business research when emerging institutions have limited standardized datasets but require operational model development (Varian, 2014; Choi et al., 2018).

The panel contains 1,200 recognition applications generated across six sectors: digital services, platform retail, financial technology, industrial manufacturing, logistics, and healthcare services. Each observation represents an application to recognize a data resource or data product as an asset. The simulated high-risk rate is 22.6%, reflecting a setting where most applications are not fraudulent but a meaningful minority require substantial adjustment or additional verification. Numerical ranges are calibrated so that high-risk applications have lower lineage coverage, higher valuation dispersion, lower market-comparable availability, weaker blockchain evidence coverage, and more concentrated auditor-client relationships than low-risk applications.

The dependent variable is binary. Independent variables are normalized between 0 and 1 unless stated otherwise. Valuation dispersion measures the coefficient of variation across three valuation methods: cost accumulation, income approach, and market-comparable benchmarking. Lineage coverage measures the share of source records whose origin, transformation, and access history are documented. Blockchain coverage measures the share of recognition evidence supported by a permissioned ledger record, including timestamp, hash, access authorization, and cross-node validation. Auditor-client repetition measures the proportion of recognition work in the last three years involving the same auditor or valuation expert. Control maturity is an index combining internal audit, data stewardship, access control, and privacy management. Table 2 summarizes the variables used in the calibrated audit risk analytics panel.

Table 2. Variables Used in the Calibrated Audit Risk Analytics Panel

Variable	Definition	Expected risk direction
High-risk recognition	1 if subsequent review finds material adjustment, evidence deficiency, or independence concern; 0 otherwise	Dependent variable
Valuation dispersion	Coefficient of variation across cost, income, and market valuation estimates	Positive

Variable	Definition	Expected risk direction
Lineage coverage	Share of source, processing, and access records with complete lineage documentation	Negative
Blockchain coverage	Share of evidence items supported by permissioned ledger records	Negative
Market comparables	Availability and similarity of external market benchmarks for the data asset	Negative
Auditor-client repetition	Share of recent data asset engagements involving the same auditor or valuation expert	Positive
Control maturity	Composite index of internal audit, data stewardship, privacy control, and access governance	Negative
Unverifiable revenue share	Share of expected benefits that cannot be linked to observable transactions or contracts	Positive

The modeling strategy compares six classifiers: logistic regression, classification and regression tree, random forest, gradient boosting machine, extreme gradient boosting, and support vector machine. Logistic regression provides a transparent baseline. Tree-based ensembles are expected to perform well because risk may emerge from interactions, such as high valuation dispersion being less suspicious when market comparables are strong but more suspicious when lineage coverage is weak. XGBoost is included because gradient-boosted decision trees often perform strongly on structured tabular data and can handle nonlinearities without extensive manual feature transformations (Chen & Guestrin, 2016).

The panel is split into 70% training, 15% validation, and 15% testing, stratified by risk label and sector. Models are tuned on the validation set and reported on the held-out test set. To address class imbalance, training applies class weights and synthetic minority oversampling only within the training fold. The oversampling process is never applied to validation or test data. This prevents leakage and preserves a realistic evaluation distribution. Class imbalance is evaluated using PR-AUC and cost-weighted F1 because high-risk cases are the minority and recall of high-risk cases matters more than overall accuracy (Chawla et al., 2002; Saito & Rehmsmeier, 2015).

Interpretability is handled through two layers. First, logistic regression coefficients provide directional insight into how features influence risk. Second, tree-based model importance identifies the strongest nonlinear predictors. Although model-agnostic local explanations can be useful, this article emphasizes global feature importance and partial sensitivity because audit institutions and regulators usually require stable policy-level explanations, not only local explanations for individual cases. Explainable machine learning is treated as a governance requirement rather than a visualization accessory (Ribeiro et al., 2016; Domingos, 2012).

Scenario analysis evaluates four governance regimes. The baseline regime uses manual rule-based screening and traditional documentation. The digitized-evidence regime standardizes recognition documents and metadata but does not require ledger evidence. The blockchain-evidence regime adds permissioned evidence coverage for selected source, processing, and authorization records. The analytics-plus-blockchain regime combines predictive triage with ledger evidence. The incentive-compatible regime further adds targeted rewards for high-quality evidence submission and penalties for repeated material adjustment. These scenarios are not intended as forecasts; they quantify the relative marginal effect of each governance layer under consistent assumptions.

5. Empirical and Numerical Results

Descriptive analysis shows that high-risk and low-risk applications differ most strongly in valuation dispersion, lineage coverage, blockchain evidence coverage, and market-comparable availability. High-risk applications have an average lineage coverage score of 0.48, compared with 0.78 for low-risk applications. Their average market-comparable availability is 0.41, compared with 0.70 for low-risk applications. Blockchain coverage also differs sharply: high-risk applications average 0.26, while low-risk applications average 0.66. These patterns are consistent with the argument that weak evidence does not merely reduce documentation quality; it changes the probability that valuation assumptions become untestable.

Figure 2 visualizes these differences. The contrast is not limited to a single indicator, which supports the need for multivariate analytics. A firm may have strong internal control maturity but still submit a risky recognition application if valuation depends on unverifiable future revenue. Conversely, a young digital firm may lack mature controls but still provide reliable recognition evidence if data lineage, rights, and market comparables are complete. Risk analytics is useful precisely because it can combine these partial signals rather than forcing auditors to rely on a rigid checklist.

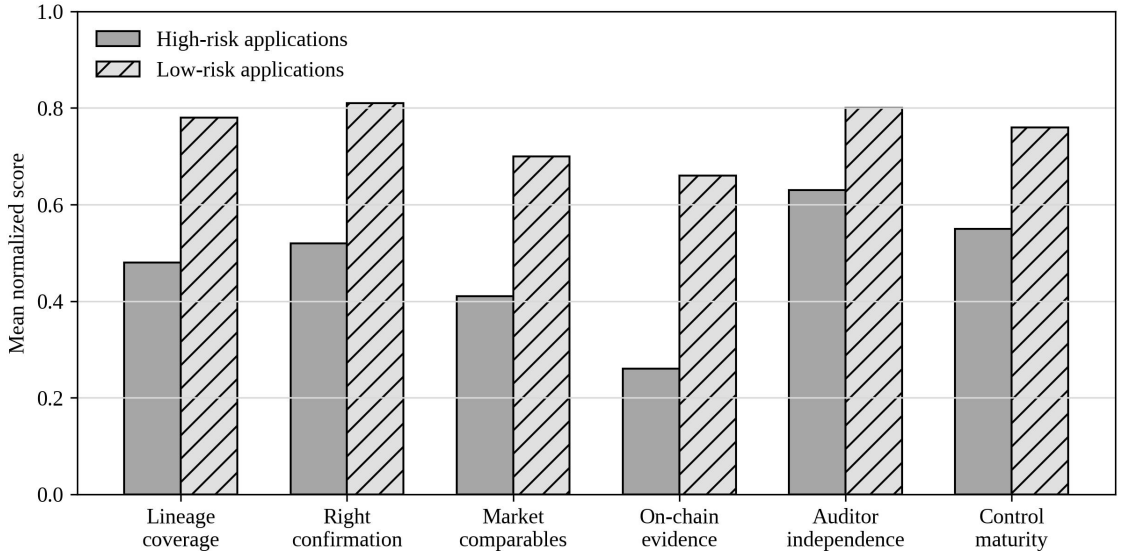


Figure 2. Descriptive Evidence-Quality Differences between High-Risk and Low-Risk Recognition Applications

Model comparison indicates that nonlinear ensemble models outperform the linear and single-tree baselines. Logistic regression produces an AUC of 0.817 and provides useful interpretability, but it struggles with interaction effects. The classification tree is easy to explain but unstable, with the lowest AUC among the tested models. Random forest improves performance by averaging across many trees. Gradient boosting performs slightly better by sequentially fitting residual patterns. XGBoost achieves the strongest overall performance, with an AUC of 0.907, a PR-AUC of 0.661, and a cost-weighted F1 score of 0.681. These results are consistent with the broader evidence that boosted trees are strong performers for structured business data (Breiman, 2001; Friedman, 2001). Table 3 reports predictive performance on the held-out test set.

The performance results should be interpreted in relation to the cost objective. A model with marginally higher AUC may not be preferred if it leads to too many expensive false positives. In the test set, the cost-sensitive threshold for XGBoost is 0.37 rather than 0.50. At this threshold, recall of high-risk applications rises to 0.83 while precision remains 0.59. Although this produces more enhanced reviews than a standard threshold, expected loss falls because missed high-risk recognition cases are much more costly than

additional review. This supports the decision-theoretic principle that analytics should be optimized for the managerial decision it supports rather than for generic classification accuracy (Hand, 2009; King & Zeng, 2001).

Table 3. Predictive Performance on the Held-Out Test Set

Model	AUC	PR-AUC	Recall	Precision	Cost-F1	Expected loss reduction
Logistic regression	0.817	0.514	0.72	0.46	0.562	20.3%
Classification tree	0.781	0.463	0.66	0.41	0.509	14.6%
Random forest	0.882	0.622	0.79	0.54	0.641	31.6%
Gradient boosting	0.894	0.638	0.81	0.55	0.653	34.2%
XGBoost	0.907	0.661	0.83	0.59	0.681	34.7%
Support vector machine	0.846	0.574	0.75	0.50	0.602	25.9%

Figure 3 compares overall model performance. The gap between AUC and PR-AUC highlights the class-imbalance problem. A model can separate high-risk and low-risk cases reasonably well in ROC space while still generating a modest precision-recall performance if high-risk cases are relatively rare. In regulatory workflows, this means that model deployment must be combined with audit capacity planning. If a regulator can examine only a small share of applications, precision matters more. If the objective is to prevent market-wide misrecognition, recall and expected loss become more important.

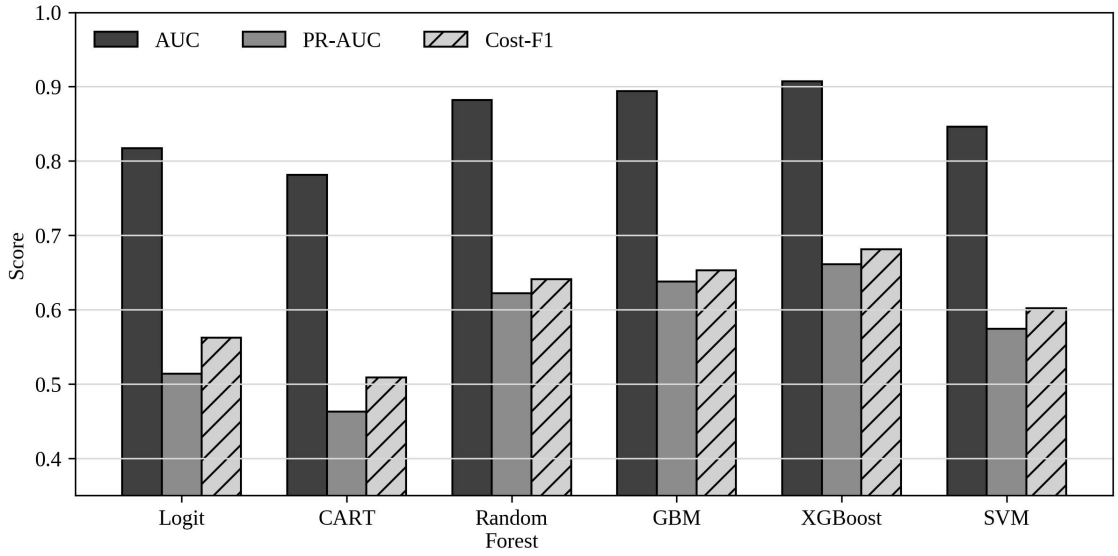


Figure 3. Comparative Model Performance under ROC, Precision-Recall, and Cost-Weighted Metrics

Feature-importance analysis confirms that valuation dispersion is the strongest predictor, followed by lineage gaps, unverifiable revenue share, auditor-client repetition, and blockchain evidence coverage. The importance of valuation dispersion suggests that data asset recognition becomes risky when different valuation methods produce inconsistent values and the firm cannot justify the selected estimate. Lineage gaps matter because they reduce the auditability of the underlying data resource. Auditor-client repetition matters because repeated relationships may improve expertise but also raise independence concerns. Blockchain evidence coverage has a protective role, but its effect is not dominant unless it covers economically relevant evidence rather than superficial records.

Figure 4 reports the normalized feature importance of the XGBoost model. The result supports a balanced interpretation of blockchain. On-chain evidence is valuable, but the strongest risk signal is still the economic reasonableness of valuation. Therefore, a firm should not expect blockchain notarization to compensate for weak market comparables or aggressive revenue assumptions. Similarly, auditors should not treat blockchain as a substitute for professional skepticism. The ledger improves the reliability of evidence trails; it does not determine the recoverable amount of a data asset.

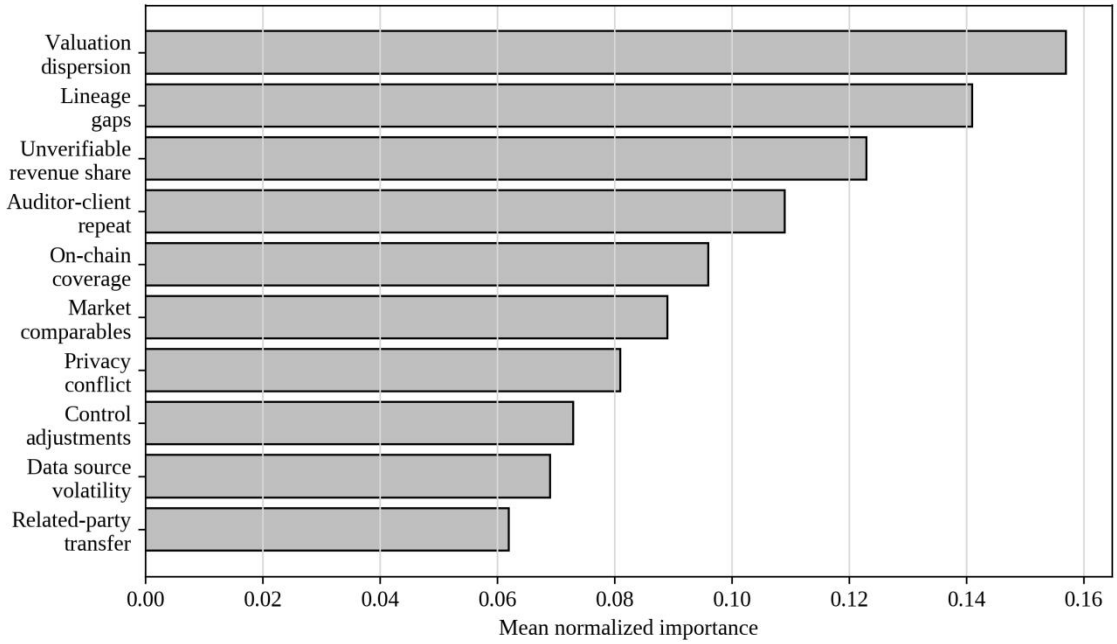


Figure 4. Feature-Importance Ranking for the XGBoost Risk Analytics Model

Cost-threshold analysis shows that rule-based screening leaves substantial avoidable loss. Under the baseline rule, expected loss per 1,000 applications is 15.8 million yuan. Logistic regression reduces the loss to 12.6 million yuan. Random forest reduces it to 10.8 million yuan, gradient boosting to 10.4 million yuan, and XGBoost to 10.3 million yuan. The marginal improvement from gradient boosting to XGBoost is modest, suggesting that governance value depends at least as much on feature quality and threshold design as on algorithmic sophistication. This finding is consistent with analytics capability research, which emphasizes complementary organizational processes rather than model choice alone (Ghasemaghaei & Calic, 2019; Abbasi et al., 2016). Table 4 reports cost-sensitive triage thresholds for selected false-negative costs.

Table 4. Cost-Sensitive Triage Thresholds for Selected False-Negative Costs

False-negative cost multiplier	Optimal threshold	Enhanced review rate	High-risk recall	Expected loss per 1,000 applications
1.0	0.46	25.1%	0.72	11.4 million yuan
1.5	0.41	30.8%	0.78	10.7 million yuan
2.0	0.37	36.5%	0.83	10.3 million yuan
2.5	0.32	43.2%	0.88	10.5 million yuan
3.0	0.29	49.6%	0.91	11.1 million yuan

Scenario analysis offers a second set of results. Moving from manual documentation to standardized digital evidence reduces expected loss because applications become easier to compare. Adding blockchain evidence lowers loss further by improving traceability and discouraging late-stage evidence alteration. The largest improvement occurs when blockchain evidence is combined with cost-sensitive analytics, because

the predictive model can identify where evidence gaps matter most. The incentive-compatible regime performs best, but only when penalties are calibrated to risk rather than applied uniformly. Excessive penalties can discourage firms from submitting borderline but legitimate data assets, reducing useful disclosure and innovation incentives. Table 5 compares the governance scenarios.

Table 5. Governance Scenario Comparison

Scenario	Evidence setting	Analytics setting	Predicted compliance probability	Expected loss index	Audit workload index
S0 Manual documentation	Traditional files and rule checklist	No predictive triage	0.58	100.0	100.0
S1 Digitized evidence	Standardized metadata and evidence templates	Rule-based flags	0.65	87.4	104.2
S2 Blockchain evidence	Permissioned ledger for key evidence items	Rule-based flags	0.72	76.1	106.9
S3 Blockchain + analytics	Ledger plus risk-score triage	Cost-sensitive XGBoost	0.81	63.8	112.7
S4 Incentive-compatible governance	Ledger, analytics, reward and penalty calibration	Cost-sensitive escalation	0.86	56.9	109.4

Figure 5 presents sensitivity results for penalty intensity and blockchain evidence coverage. Compliance probability rises as both increase, but the relationship is nonlinear. At low blockchain coverage, higher penalties have limited effect because detection remains uncertain. At high blockchain coverage, moderate penalties are sufficient because the expected probability of exposure is high. This supports a key managerial conclusion: stronger punishment is not a substitute for better evidence infrastructure. Regulators seeking sustainable compliance should first improve evidence comparability and detection quality, then use penalties to target repeated or deliberate misconduct (Becker, 1968; Laffont & Tirole, 1986).

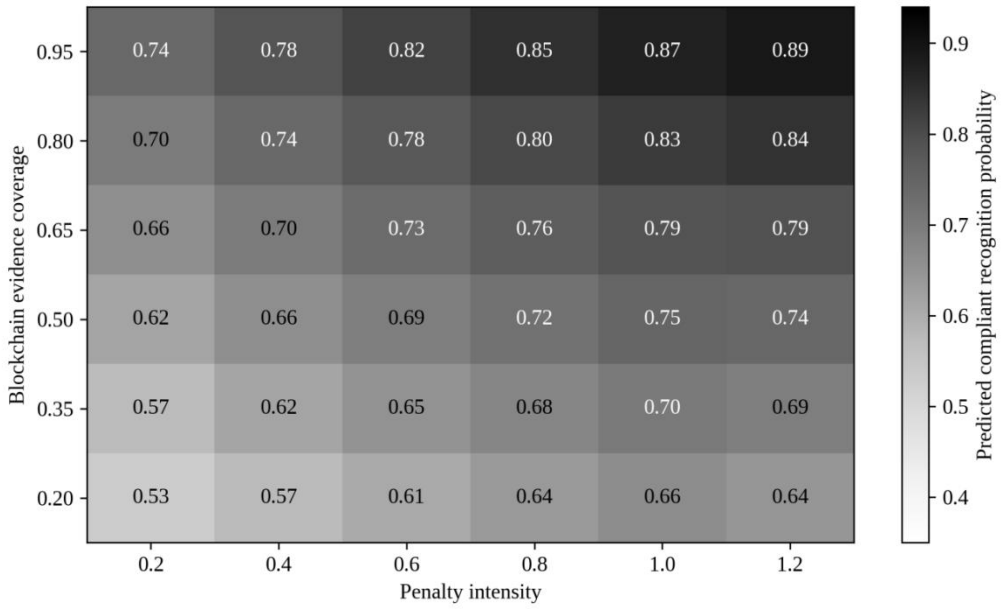


Figure 5. Sensitivity of Compliant Recognition Probability to Penalty Intensity and Blockchain Evidence Coverage

6. Discussion

The results provide three theoretical implications. First, they show that data asset recognition can be studied as a data-driven risk analytics problem rather than only as an accounting-standard or regulatory-incentive problem. The transformation is important because the emerging data asset market will generate large volumes of recognition applications, and regulators will not be able to examine all applications with equal intensity. A risk-scoring model offers a bridge between high-level governance principles and operational supervision. It converts abstract concerns such as data fabrication, collusion, and weak verifiability into measurable features and decision thresholds.

Second, the findings refine the role of blockchain in data asset governance. Blockchain evidence coverage reduces risk most effectively when it captures the right evidence: data source, transformation, authorization, access, and valuation-relevant usage. A ledger that stores only final reports has limited incremental value. This distinction parallels the broader blockchain literature, which argues that distributed ledgers generate governance value when processes are sufficiently codifiable and when off-chain institutions support the interpretation of on-chain evidence (Xu et al., 2017; Mendling et al., 2018). In data asset recognition, the value of blockchain is therefore conditional rather than automatic.

Third, the article contributes to audit analytics by emphasizing cost-sensitive evaluation. Many predictive studies report accuracy, AUC, or recall, but audit and regulatory decisions involve asymmetric costs. The cost of allowing a materially overstated data asset onto the balance sheet may include market mispricing, financing distortion, reputational damage, and future correction cost. The cost of enhanced review is usually lower and more controllable. A rational risk analytics system should therefore optimize the threshold at which review is triggered, not merely estimate probability. This insight is relevant to continuous auditing, internal control analytics, and regulatory technology (Kokina & Davenport, 2017; Brown-Liburd et al., 2015).

The managerial implications for firms are straightforward. Firms should prepare data asset recognition as an evidence-governance project, not merely as an accounting presentation. Before submitting an application, managers should test whether data source records, processing histories, rights documentation,

and commercialization assumptions can be independently verified. Firms with weak lineage or highly dispersed valuations should expect enhanced audit scrutiny. Investments in metadata management, data stewardship, access control, and evidence standardization can lower recognition risk even before blockchain is introduced. This conclusion is consistent with research on data governance and enterprise analytics capability (Abraham et al., 2019; Ghasemaghaei et al., 2018).

For auditors, the model suggests that data asset recognition requires a shift from document checking to evidence analytics. Traditional audit procedures remain necessary, but auditors should also evaluate data flow integrity, valuation sensitivity, commercialization evidence, and relationships between management assumptions and platform logs. Repeated auditor-client relationships should be monitored carefully. Such relationships may produce knowledge benefits, but they may also weaken skepticism if the audit market becomes dependent on a small number of specialized data valuation experts. A risk-scoring system can help audit firms allocate senior specialists to the highest-risk cases and provide a defensible basis for expanded procedures.

For regulators, the results suggest a staged governance strategy. In the first stage, regulators should standardize data asset recognition submissions and require machine-readable evidence templates. In the second stage, they should encourage or require blockchain evidence for high-risk categories, such as data assets based on personal information, cross-platform aggregation, or highly speculative future revenue. In the third stage, regulators should deploy predictive screening to identify applications requiring detailed review. In the fourth stage, rewards and penalties should be tied to observable evidence quality and repeated outcomes rather than broad categories. This approach aligns with the idea that digital governance requires both technological infrastructure and incentive-compatible institutional design (Stigler, 1971; Peltzman, 1976). Table 6 summarizes the managerial and policy implications.

Table 6. Managerial and Policy Implications

Actor	Recommended action	Rationale
Firms	Prepare data asset recognition as an evidence-governance project before accounting presentation	Strong lineage and rights documentation reduce risk scores and audit friction
Auditors	Use risk analytics to allocate valuation specialists and expand testing for high-score applications	Cost-sensitive triage directs scarce assurance resources to expected-loss hotspots
Regulators	Standardize machine-readable evidence templates and require traceability for high-risk asset categories	Comparable evidence improves detection quality and makes incentives more credible
Technology providers	Design permissioned ledgers around source, transformation, authorization, and valuation evidence	Ledger value depends on economically relevant evidence rather than final-report notarization
Researchers	Move from aggregate governance concepts to operational models using confidential filings and text analytics	External validation will refine thresholds, features, and sector-specific risk structures

The study also has implications for business education and analytics training. Future accountants, auditors, and managers will need to understand not only standards and legal definitions but also data lineage, algorithmic dependence, model risk, and cost-sensitive prediction. Data asset recognition is an ideal case for interdisciplinary training because it combines accounting, information systems, machine learning, audit judgment, and regulation. This reflects the broader emergence of management analytics as a field that integrates quantitative models with organizational decision-making.

Robustness checks support the main conclusions. When the simulated high-risk base rate is reduced from 22.6% to 15%, PR-AUC naturally declines because positive cases become rarer, but the cost-sensitive model still outperforms rule-based screening. When the false-negative cost is doubled, the optimal threshold decreases and the model triggers more enhanced reviews. When blockchain evidence coverage is noisy, its feature importance declines, but lineage coverage and valuation dispersion remain strong predictors. These checks indicate that the framework is more robust as a decision process than as a fixed numerical forecast. This is appropriate for an emerging institutional setting where parameter values will vary across sectors and jurisdictions.

The study has limitations. First, the data are calibrated rather than drawn from confidential regulatory filings. The results therefore demonstrate method feasibility and relative behavior, not external empirical estimates for a specific market. Second, the high-risk label combines several outcomes, including valuation adjustment, evidence deficiency, and independence concern. Future research could model these outcomes separately. Third, the article uses structured features and does not process unstructured contracts or audit narratives. Natural language processing could extract additional signals from valuation reports, data licensing agreements, and audit working papers. Fourth, the current model treats blockchain coverage as a numerical feature, while future work could examine ledger architecture, privacy-preserving computation, and interoperability as separate design variables.

Despite these limitations, the framework is useful because it provides a practical analytical route for an emerging problem. It does not require regulators to assume perfect blockchain detection, and it does not require auditors to abandon professional judgment. Instead, it shows how data governance evidence, valuation risk, and audit relationships can be converted into a ranked risk list. Such ranked lists are familiar in banking, insurance, and cybersecurity supervision, but they remain underdeveloped in the context of data asset recognition. As data assets become more visible in corporate reports, the demand for such analytics will increase.

7. Conclusion

This article developed a new topic related to the governance of balance-sheet recognition of data assets but clearly distinct from evolutionary game modeling. It proposed a data-driven audit risk analytics framework that integrates evidence quality, valuation dispersion, auditor relationship, firm control history, and blockchain evidence coverage. The framework was evaluated through a calibrated numerical panel of 1,200 data asset recognition applications and a comparative set of machine learning models. XGBoost produced the strongest predictive performance, while cost-sensitive thresholding delivered the greatest governance value by aligning model use with the asymmetric loss structure of audit failure.

The main findings are fourfold. First, high-risk recognition applications are characterized by weak lineage coverage, high valuation dispersion, low market-comparable availability, and lower blockchain evidence coverage. Second, nonlinear ensemble models outperform rule-based screening and linear baselines, but the value of analytics depends on cost-sensitive threshold design. Third, blockchain evidence governance reduces expected loss most strongly when combined with predictive triage and targeted audit escalation. Fourth, reward and penalty mechanisms are more effective when the underlying evidence infrastructure makes misconduct observable. The policy lesson is therefore not simply to punish more severely, but to make evidence more comparable, verifiable, and analytically useful.

Future research should move from calibrated simulation to confidential firm-level applications once standardized data asset filings become available. Researchers could also integrate natural language processing, graph analytics, and privacy-preserving learning to examine contracts, data flows, and interorganizational evidence networks. Another promising direction is to compare jurisdictions and industries, because data asset recognition risks are likely to differ between platform firms, financial

technology firms, healthcare data processors, and industrial manufacturers. As enterprise data become more important to market valuation and strategic decision-making, audit risk analytics for data assets will become a central component of responsible digital economy governance.

8. Extended Analytical Notes on Model Governance

The use of a predictive model in regulatory or audit decision-making introduces a second-order governance problem: the model itself becomes part of the recognition infrastructure. For this reason, model governance should include data provenance, training-set versioning, feature definition control, validation logs, access permissions, and periodic recalibration. The same logic used to govern data assets should be applied to the analytics system that evaluates them. If the model is trained on poorly documented outcomes or if threshold changes are not traceable, the risk scoring process may reproduce the very opacity it seeks to reduce. A permissioned audit log for model development decisions can therefore complement the evidence ledger used for data asset applications (Kshetri, 2018).

Another issue concerns fairness across firm size. Large firms may have more mature data governance systems and therefore lower risk scores, while smaller digital firms may hold valuable data assets but lack standardized documentation. A purely mechanical risk score could unintentionally penalize innovative firms that have not yet formalized their governance routines. To avoid this outcome, regulators should distinguish between remediable documentation gaps and substantive valuation concerns. Enhanced review should be framed as a pathway to recognition quality rather than simply a punitive step. This is consistent with institutional economics, which emphasizes that governance rules shape market participation and investment incentives (La Porta et al., 1998; Djankov et al., 2002).

Model drift is also likely. As firms learn the screening rules, they may improve evidence quality genuinely, but they may also optimize submissions around the model. This creates a need for periodic feature refreshment and adversarial testing. For example, if firms begin to notarize final reports on chain without improving source-level evidence, blockchain coverage may become less predictive. Similarly, if all auditors begin to report standardized checklist scores, those scores may lose discriminatory power. The analytics system should therefore monitor feature stability and outcome calibration over time. This resembles cybersecurity risk management, where adversaries adapt to detection rules and models must be updated continuously.

The practical deployment architecture can be organized around three data marts. The firm-side mart stores submission features, evidence files, and data-governance indicators. The auditor-side mart stores procedures, sampling coverage, adjustments, and independence checks. The regulator-side mart stores outcomes, penalties, disclosure corrections, and final risk labels. A privacy-preserving linkage key allows model training without exposing sensitive commercial data more broadly than necessary. Such an architecture supports data minimization while enabling cross-node learning. It also allows regulators to separate supervisory analytics from raw proprietary data, reducing the risk that data asset recognition itself becomes a new channel of information leakage (Boyd & Crawford, 2012; Kitchin, 2014).

A final note concerns the interpretation of data assets as productive resources. Recognition should not be treated as an accounting reward for digitization. It should represent a disciplined judgment that the resource is identifiable, controlled, expected to generate benefits, and supported by credible evidence. Analytics improves this judgment only when it respects accounting substance. The proposed model therefore avoids treating high predicted commercial value as equivalent to recognition quality. In several simulated cases, projected revenue is high but risk is also high because market comparables are weak or because the revenue claim depends on future platform arrangements. This separation of business potential from recognition reliability is important for preventing the over-financialization of data resources (Corrado et al., 2009).

9. Additional Theoretical Integration

The accounting foundation of the proposed framework builds on the information role of financial reports and the audit literature on misstatement detection. Earlier archival auditing studies explain why audit quality, assurance effort, and financial reporting incentives cannot be evaluated separately from institutional context (DeFond & Zhang, 2014). Big data auditing research extends this logic by showing that population-level analytics, continuous monitoring, and exception detection can reshape evidence gathering without removing the need for professional judgment (Alles, 2015; Cao et al., 2015). In this study, these ideas are transferred to the newer problem of data asset recognition, where evidence is both financial and technical.

The broader digital-economy foundation also matters. Studies of regulation and institutions show that entry rules, law, and enforcement quality shape the credibility of markets and the cost of opportunistic behavior (La Porta et al., 1998; Djankov et al., 2002). Blockchain and information systems studies add that enterprise trust increasingly depends on secure data exchange, data integrity, and architecture-level coordination. These streams jointly support the article's argument that audit risk analytics should be embedded in a larger infrastructure of evidence governance, not treated as a detached predictive exercise. A permissioned ledger can restrict participation and define endorsement rules for recorded transactions (Androulaki et al., 2018).

Finally, the model reflects the convergence of financial innovation, digital infrastructure, and data-driven governance. Research on FinTech, decentralized finance, and LLM-enabled blockchain finance shows that business data, platform finance, and digital assurance mechanisms are becoming increasingly interdependent. Data asset recognition therefore becomes more than an accounting classification issue. It becomes a test of whether firms, auditors, and regulators can design evidence systems that support innovation while preventing inflated digital-value claims.

10. Robustness, Deployment, and Boundary Conditions

A practical risk analytics system for data asset recognition should be evaluated not only by initial test-set performance but also by deployment stability. The first robustness concern is sectoral transfer. A data asset submitted by a platform retailer differs from a data asset submitted by an industrial manufacturer or a healthcare service provider. Platform data may be rich in transaction logs and customer interactions, while industrial data may be embedded in equipment, process-control systems, and predictive maintenance algorithms. Healthcare data may create strong future benefit but also severe privacy and consent constraints. For this reason, a regulator should avoid imposing a single universal threshold without sectoral calibration. A common feature dictionary can be retained, but threshold values and review protocols should be adjusted according to sector-specific evidence patterns and harm profiles.

The second robustness concern is strategic adaptation. Once firms understand the screening process, they may improve documentation quality. This is desirable when documentation improvement reflects genuine governance maturity. However, firms may also attempt cosmetic compliance, such as generating formal metadata records without improving source-level verifiability or commissioning multiple valuations until a favorable estimate is obtained. The analytics system should therefore monitor the distribution of features over time and compare changes in documentation quality with subsequent review outcomes. If a feature improves rapidly while its association with compliant outcomes weakens, the feature may have become a target for gaming. This calls for periodic recalibration, feature rotation, and audit sampling of low-score applications to prevent blind spots.

The third robustness concern is the distinction between model accuracy and institutional legitimacy. A model may produce useful predictions while still being distrusted by firms if its logic is opaque, if appeal channels are weak, or if thresholds appear arbitrary. Therefore, predictive screening should be accompanied by procedural safeguards. Firms should know the categories of evidence that matter, even if exact model

weights are not disclosed. Auditors should document how model scores influenced procedure selection. Regulators should maintain an appeals process in which firms can submit additional evidence to clarify high-risk indicators. These safeguards are not obstacles to analytics; they make analytics governable and reduce the risk that predictive tools are perceived as black-box administrative power.

The fourth robustness concern involves data security. A system designed to evaluate data assets will itself collect sensitive evidence about data resources, commercial models, algorithms, customer relationships, and internal controls. If this system is centralized without adequate safeguards, it may create a new concentration of risk. A permissioned architecture, encryption, role-based access, and selective disclosure are therefore necessary. Blockchain should be used selectively, because putting excessive detail on chain may create privacy and confidentiality problems even when the chain is permissioned. The principle should be evidence integrity with minimum necessary exposure: hash and timestamp what must be verified, store sensitive underlying records in controlled repositories, and disclose only what is required for audit and regulatory purpose.

The deployment roadmap can be organized into four phases. Phase one is evidence standardization. Regulators and professional bodies should define a common evidence taxonomy covering source data, rights, transformations, valuation assumptions, commercialization evidence, and control maturity. Phase two is supervised pilot screening. Audit firms and regulators can test risk models on historical applications and compare predicted scores with actual adjustments. Phase three is hybrid decision support, where model scores guide enhanced review but human professionals retain final responsibility. Phase four is continuous feedback, where outcomes from audit reviews, enforcement actions, and restatements are used to recalibrate the model. This staged roadmap avoids premature automation and gives institutions time to learn how analytics interacts with professional judgment.

Boundary conditions are also important. The proposed framework is most useful when data asset applications are numerous enough to justify statistical learning, when evidence can be standardized, and when review outcomes can be recorded consistently. It is less useful for one-off cases involving unique national security data, highly confidential public-sector databases, or assets whose recognition depends almost entirely on legal interpretation. In those cases, expert review and legal analysis should dominate. The framework is also not designed to decide the absolute value of a data asset. It predicts audit risk around recognition quality. Valuation remains a separate professional judgment that may use income, cost, or market methods. Treating risk scoring as valuation would misuse the model.

The policy implication is that analytics and blockchain should be combined pragmatically. Analytics prioritizes attention; blockchain improves evidence integrity; incentives shape behavior; professional judgment interprets substance. None of these instruments is sufficient alone. An analytics model without evidence governance may learn from noisy or manipulated inputs. A blockchain ledger without analytics may produce a large archive that no one can prioritize. Incentives without detection may reward superficial compliance. Professional judgment without data support may be overwhelmed by scale. The strength of the proposed framework lies in integrating these instruments into a coherent decision process for data asset recognition.

A final methodological point concerns reproducibility. Because many recognition files are confidential, researchers should report feature definitions, data-generation assumptions, evaluation metrics, and cost parameters with enough clarity to allow independent sensitivity analysis. Future empirical studies should also compare model results across training periods and sectors, report calibration curves, and examine false-negative cases qualitatively. The most valuable evidence may come not from average performance but from understanding why the model misses particular risky applications. Such failure analysis can reveal new forms of data asset manipulation or previously unobserved evidence gaps, supporting iterative improvement of both analytics and regulation.

Acknowledgement

The authors gratefully acknowledge the constructive suggestions provided by anonymous reviewers and the editorial office of the Journal of Business and Data Analytics. The authors also thank their affiliated universities for providing library access and research support for this conceptual and numerical study.

Author Contributions

Conceptualization, J.D.W. and M.K.W.; Methodology, J.D.W. and M.K.W.; Formal Analysis, J.D.W. and M.K.W.; Investigation, J.D.W. and H.R.C.; Data Curation, J.D.W. and H.R.C.; Software, J.D.W. and H.R.C.; Validation, J.D.W. and M.K.W. and H.R.C.; Resources, H.R.C.; Visualization, J.D.W. and H.R.C.; Writing – Original Draft, J.D.W.; Writing – Review & Editing, J.D.W. and M.K.W. and H.R.C.; Supervision, M.K.W.; Project Administration, J.D.W. and M.K.W. We have read and approved the final manuscript and take responsibility for the work.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflicts of Interest

The authors declare no conflict of interest.

Ethical Approval and Consent

Not applicable. This study did not involve human participants, personal data, or animals.

Data Availability Statement

The article uses a calibrated numerical panel generated for methodological demonstration. No confidential company-level data are used. The variable definitions and scenario parameters reported in the article are sufficient to reproduce the numerical logic of the study.

Declaration of Generative AI Use

The authors did not use generative AI tools in the preparation of this manuscript.

References

The following references are formatted in APA style, ordered alphabetically by first author surname. DOI information is provided for all entries.

Abbasi, A., Sarker, S., & Chiang, R. H. L. (2016). Big data research in information systems: Toward an inclusive research agenda. Journal of the Association for Information Systems, 17(2), 3. https://doi.org/10.17705/1jais.00423

Abraham, R., Schneider, J., & vom Brocke, J. (2019). Data governance: A conceptual framework, structured review, and research agenda. International Journal of Information Management, 49, 424-438. https://doi.org/10.1016/j.ijinfomgt.2019.07.008

Androulaki, E., Barger, A., Bortnikov, V., Cachin, C., Christidis, K., De Caro, A., Enyeart, D., Ferris, C., Laventman, G., Manevich, Y., Muralidharan, S., Murthy, C., Nguyen, B., Sethi, M., Singh, G., Smith, K., Sorniotti, A., Stathakopoulou, C., Vukolić, M., Cocco, S. W., & Yellick, J. (2018). Hyperledger Fabric: A distributed operating system for permissioned blockchains. Proceedings of the Thirteenth EuroSys Conference, Article 30. https://doi.org/10.1145/3190508.3190538

Akerlof, G. A. (1970). The market for lemons: Quality uncertainty and the market mechanism. Quarterly Journal of Economics, 84(3), 488-500. https://doi.org/10.2307/1879431

Akter, S., Wamba, S. F., Gunasekaran, A., Dubey, R., & Childe, S. J. (2016). How to improve firm performance using big data analytics capability and business strategy alignment. International Journal of Production Economics, 182, 113-131. https://doi.org/10.1016/j.ijpe.2016.08.018

Becker, G. S. (1968). Crime and punishment: An economic approach. Journal of Political Economy, 76(2), 169-217. https://doi.org/10.1086/259394

Bonsón, E., & Bednárová, M. (2019). Blockchain and its implications for accounting and auditing. Meditari Accountancy Research, 27(5), 725-740. https://doi.org/10.1108/MEDAR-11-2018-0406

Boyd, D., & Crawford, K. (2012). Critical questions for big data. Information, Communication & Society, 15(5), 662-679. https://doi.org/10.1080/1369118X.2012.678878

Breiman, L. (2001). Random forests. Machine Learning, 45, 5-32. https://doi.org/10.1023/A:1010933404324

Brown-Liburd, H., Issa, H., & Lombardi, D. (2015). Behavioral implications of big data impact on audit judgment and decision making and future research directions. Accounting Horizons, 29(2), 451-468. https://doi.org/10.2308/acch-51023

Casino, F., Dasaklis, T. K., & Patsakis, C. (2019). A systematic literature review of blockchain-based applications: Current status, classification and open issues. Telematics and Informatics, 36, 55-81. https://doi.org/10.1016/j.tele.2018.11.006

Cao, M., Chychyla, R., & Stewart, T. (2015). Big data analytics in financial statement audits. Accounting Horizons, 29(2), 423-429. https://doi.org/10.2308/acch-51068

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357. https://doi.org/10.1613/jair.953

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794. https://doi.org/10.1145/2939672.2939785

Christidis, K., & Devetsikiotis, M. (2016). Blockchains and smart contracts for the Internet of Things. IEEE Access, 4, 2292-2303. https://doi.org/10.1109/ACCESS.2016.2566339

Choi, T. M., Wallace, S. W., & Wang, Y. (2018). Big data analytics in operations management. Production and Operations Management, 27(10), 1868-1883. https://doi.org/10.1111/poms.12838

Corrado, C., Hulten, C., & Sichel, D. (2009). Intangible capital and U.S. economic growth. Review of Income and Wealth, 55(3), 661-685. https://doi.org/10.1111/j.1475-4991.2009.00343.x

Dai, J., & Vasarhelyi, M. A. (2017). Toward blockchain-based accounting and assurance. Journal of Information Systems, 31(3), 5-21. https://doi.org/10.2308/isys-51804

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. Proceedings of the 23rd International Conference on Machine Learning, 233-240. https://doi.org/10.1145/1143844.1143874

Dechow, P. M., Ge, W., Larson, C. R., & Sloan, R. G. (2011). Predicting material accounting misstatements. Contemporary Accounting Research, 28(1), 17-82. https://doi.org/10.1111/j.1911-3846.2010.01041.x

DeFond, M., & Zhang, J. (2014). A review of archival auditing research. Journal of Accounting and Economics, 58(2-3), 275-326. https://doi.org/10.1016/j.jacceco.2014.09.002

Djankov, S., La Porta, R., Lopez-de-Silanes, F., & Shleifer, A. (2002). The regulation of entry. Quarterly Journal of Economics, 117(1), 1-37. https://doi.org/10.1162/003355302753399436

Domingos, P. (2012). A few useful things to know about machine learning. Communications of the ACM, 55(10), 78-87. https://doi.org/10.1145/2347736.2347755

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189-1232. https://doi.org/10.1214/aos/1013203451

George, G., Haas, M. R., & Pentland, A. (2014). Big data and management. Academy of Management Journal, 57(2), 321-326. https://doi.org/10.5465/amj.2014.4002

Ghasemaghaei, M., & Calic, G. (2019). Does big data enhance firm innovation competency? The mediating role of data-driven insights. Journal of Business Research, 104, 69-84. https://doi.org/10.1016/j.jbusres.2019.07.006

Ghasemaghaei, M., Ebrahimi, S., & Hassanein, K. (2018). Data analytics competency for improving firm decision making performance. Journal of Strategic Information Systems, 27(1), 101-113. https://doi.org/10.1016/j.jsis.2017.10.001

Grover, V., Chiang, R. H. L., Liang, T. P., & Zhang, D. (2018). Creating strategic business value from big data analytics: A research framework. Journal of Management Information Systems, 35(2), 388-423. https://doi.org/10.1080/07421222.2018.1451951

Günther, W. A., Rezazade Mehrizi, M. H., Huysman, M., & Feldberg, F. (2017). Debating big data: A literature review on realizing value from big data. Journal of Strategic Information Systems, 26(3), 191-209. https://doi.org/10.1016/j.jsis.2017.07.003

Hand, D. J. (2009). Measuring classifier performance: A coherent alternative to the area under the ROC curve. Machine Learning, 77, 103-123. https://doi.org/10.1007/s10994-009-5119-5

Hofbauer, J., & Sigmund, K. (2003). Evolutionary game dynamics. Bulletin of the American Mathematical Society, 40(4), 479-519. https://doi.org/10.1090/S0273-0979-03-00988-1

Khatri, V., & Brown, C. V. (2010). Designing data governance. Communications of the ACM, 53(1), 148-152. https://doi.org/10.1145/1629175.1629210

King, G., & Zeng, L. (2001). Logistic regression in rare events data. Political Analysis, 9(2), 137-163. https://doi.org/10.1093/oxfordjournals.pan.a004868

Kitchin, R. (2014). The data revolution: Big data, open data, data infrastructures and their consequences. Big Data & Society, 1(1), 1-3. https://doi.org/10.1177/2053951714528481

Kokina, J., & Davenport, T. H. (2017). The emergence of artificial intelligence: How automation is changing auditing. Journal of Emerging Technologies in Accounting, 14(1), 115-122. https://doi.org/10.2308/jeta-51730

Kshetri, N. (2018). Blockchain roles in meeting key supply chain management objectives. International Journal of Information Management, 39, 80-89. https://doi.org/10.1016/j.ijinfomgt.2017.12.005

La Porta, R., Lopez-de-Silanes, F., Shleifer, A., & Vishny, R. W. (1998). Law and finance. Journal of Political Economy, 106(6), 1113-1155. https://doi.org/10.1086/250042

Laffont, J. J., & Tirole, J. (1986). Using cost observation to regulate firms. Journal of Political Economy, 94(3), 614-641. https://doi.org/10.1086/261392

Lev, B., & Sougiannis, T. (1996). The capitalization, amortization, and value-relevance of R&D. Journal of Accounting and Economics, 21(1), 107-138. https://doi.org/10.1016/0165-4101(95)00410-6

Lev, B., & Zarowin, P. (1999). The boundaries of financial reporting and how to extend them. Journal of Accounting Research, 37(2), 353-385. https://doi.org/10.2307/2491413

Mendling, J., Weber, I., van der Aalst, W., Brocke, J. vom, Cabanillas, C., Daniel, F., Debois, S., Di Ciccio, C., Dumas, M., Dustdar, S., et al. (2018). Blockchains for business process management: Challenges and opportunities. ACM Transactions on Management Information Systems, 9(1), 4. https://doi.org/10.1145/3183367

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. Journal of Economic Perspectives, 31(2), 87-106. https://doi.org/10.1257/jep.31.2.87

Peltzman, S. (1976). Toward a more general theory of regulation. Journal of Law and Economics, 19(2), 211-240. https://doi.org/10.1086/466865

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144. https://doi.org/10.1145/2939672.2939778

Risius, M., & Spohrer, K. (2017). A blockchain research framework. Business & Information Systems Engineering, 59(6), 385-409. https://doi.org/10.1007/s12599-017-0506-0

Saberi, S., Kouhizadeh, M., Sarkis, J., & Shen, L. (2019). Blockchain technology and its relationships to sustainable supply chain management. International Journal of Production Research, 57(7), 2117-2135. https://doi.org/10.1080/00207543.2018.1533261

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Spence, M. (1973). Job market signaling. Quarterly Journal of Economics, 87(3), 355-374. https://doi.org/10.2307/1882010

Stigler, G. J. (1971). The theory of economic regulation. Bell Journal of Economics and Management Science, 2(1), 3-21. https://doi.org/10.2307/3003160

Taylor, P. D., & Jonker, L. B. (1978). Evolutionary stable strategies and game dynamics. Mathematical Biosciences, 40(1-2), 145-156. https://doi.org/10.1016/0025-5564(78)90077-9

Treiblmaier, H. (2018). The impact of the blockchain on the supply chain: A theory-based research framework and a call for action. Supply Chain Management, 23(6), 545-559. https://doi.org/10.1108/SCM-03-2018-0143

Varian, H. R. (2014). Big data: New tricks for econometrics. Journal of Economic Perspectives, 28(2), 3-28. https://doi.org/10.1257/jep.28.2.3

Vasarhelyi, M. A., Kogan, A., & Tuttle, B. M. (2015). Big data in accounting: An overview. Accounting Horizons, 29(2), 381-396. https://doi.org/10.2308/acch-51071

Wamba, S. F., Gunasekaran, A., Akter, S., Ren, S. J. F., Dubey, R., & Childe, S. J. (2017). Big data analytics and firm performance: Effects of dynamic capabilities. Journal of Business Research, 70, 356-365. https://doi.org/10.1016/j.jbusres.2016.08.009

Warren, J. D., Moffitt, K. C., & Byrnes, P. (2015). How big data will change accounting. Accounting Horizons, 29(2), 397-407. https://doi.org/10.2308/acch-51069

Xu, X., Weber, I., Staples, M., Zhu, L., Bosch, J., Bass, L., Pautasso, C., & Rimba, P. (2017). A taxonomy of blockchain-based systems for architecture design. Proceedings of the IEEE International Conference on Software Architecture, 243-252. https://doi.org/10.1109/ICSA.2017.33

Yli-Huumo, J., Ko, D., Choi, S., Park, S., & Smolander, K. (2016). Where is current research on blockchain technology? A systematic review. PLoS ONE, 11(10), e0163477. https://doi.org/10.1371/journal.pone.0163477