ART AND TECHNOLOGY PRESS INC
(United States)

2026 Volume 2, Issue 1
Economic Development Studies

Research on the Construction of Audit Risk Identification Model Based on Big Data Technology

Feiya Duan
Dianchi College, Kunming 650000, Yunnan, China

Abstract: With the improvement of enterprise business complexity and the surge in data scale, traditional audit risk identification methods are facing bottlenecks such as insufficient full-data processing and delayed risk response. This paper focuses on the construction of an audit risk identification model based on big data technology, combining the advantages of big data in full-volume analysis, real-time processing and intelligent mining, and clarifies the core logic and technical support for model construction. Firstly, it sorts out the core elements of audit risk identification and the adaptability of big data technology, then constructs a model framework from the data layer, technology layer and application layer, and focuses on expounding the design points of multi-source data integration mechanism, intelligent risk feature mining and dynamic risk assessment system. The research shows that the model can break through the limitations of traditional experience-driven methods, realize accurate identification, dynamic tracking and forward-looking early warning of audit risks through data-driven approach, and effectively improve the efficiency and accuracy of audit risk identification. The research of this paper provides a feasible model reference for the digital transformation in the audit field, and is of great significance for promoting the development of audit risk identification towards intelligence and systematization.

Keywords: Big Data Technology; Audit Risk; Identification Model; Data Integration; Feature Mining

Online publication: January 20, 2026

1. Introduction

As the core component of the economic supervision system, audit's risk identification ability directly determines audit quality and supervision efficiency. At present, with the intensification of global competition and the deepening of digital transformation, the enterprise operating environment presents complex and dynamic characteristics, and the forms of audit risks are more concealed. Traditional risk identification methods based on sampling audit and empirical judgment have been difficult to adapt to the processing needs of massive multi-source data, with problems such as blind spots in risk identification and delayed response. Big data technology, with its core characteristics of large data volume, diverse types, fast processing speed and low value density, provides a new technical path for audit risk identification. Through full-data collection, intelligent algorithm analysis and dynamic correlation mining, it can realize accurate capture and early warning of potential risks. In this context, constructing an audit risk identification model based on big data technology has become a key measure to break through the bottlenecks of traditional auditing and improve the level of audit intelligence. Based on the core advantages of big data technology and combined with the business logic of audit risk identification, this paper

systematically constructs a model framework and expounds the key implementation links, aiming to provide scientific technical support for audit practice, promote the transformation of audit risk identification from experience-driven to data-driven, and help the high-quality development of the audit field.

2. Foundation for the construction of audit risk identification model based on big data technology

2.1. Core elements of audit risk identification and adaptability with big data

The core elements of audit risk identification include risk data sources, characteristic indicators, identification logic and evaluation standards. Traditional methods rely on structured data and manual judgment, with limitations such as single data dimension, incomplete feature extraction and rigid logic. Big data technology is highly adaptable to these elements and can break through the limitations from multiple dimensions: at the data source level, it integrates internal and external multi-source heterogeneous data to achieve panoramic coverage of risk data; at the feature extraction level, it mines implicit correlations and abnormal patterns through intelligent algorithms to improve the pertinence of indicators; at the identification logic level, it supports dynamic update of rules to adapt to the risk evolution trend; at the evaluation standard level, it accurately defines risk levels through quantitative analysis, replacing fuzzy empirical evaluation. This adaptability provides core support for model construction, promotes the transformation of risk identification from”sampling judgment”to”full-volume insight”and”static analysis”to”dynamic tracking”, and improves the comprehensiveness and accuracy of identification.

2.2. Core technical support system for model construction

Model construction relies on four major technical modules: data processing, feature mining, risk assessment and dynamic optimization. The data processing module uses RPA, ETL tools and stream processing technology to realize automatic collection, cleaning and integration of multi-source data, and ensures data adaptability through data standardization and quality evaluation; the feature mining module integrates machine learning and deep learning algorithms, uses clustering and isolation forests to identify abnormal behaviors, captures temporal risk trends through RNN and LSTM, extracts unstructured text risk clues with NLP, and constructs risk knowledge graphs; the risk assessment module adopts analytic hierarchy process(AHP), fuzzy comprehensive evaluation method and Bayesian network to realize risk quantitative assessment and factor correlation modeling, and improve the scientificity of evaluation; the dynamic optimization module optimizes algorithm parameters and generalization ability through model iteration, rule update and transfer learning, ensuring that the model adapts to the risk evolution law. The four modules work together to ensure the efficiency and accuracy of risk identification[1].

2.3. Core principles and objective orientation of model construction

The construction of audit risk identification model based on big data technology needs to follow four core principles to ensure the scientificity, practicality and scalability of the model. Firstly, the data-driven principle: the model construction is based on full-volume multi-source data, getting rid of excessive dependence on manual experience, extracting risk rules through data mining, and realizing the objectivity and accuracy of risk identification; secondly, the systematic principle: the model needs to cover the whole process links such as data collection, feature extraction, risk identification, evaluation and early warning, balance the collaborative adaptation of each link, and form a complete risk identification closed loop; thirdly, the dynamic adaptation principle: the model needs to have real-time update and iteration capabilities, and can dynamically optimize risk characteristic indicators and identification rules according to business process changes, policy adjustments and risk evolution trends; fourthly, the compliance principle: in the process of model construction, it is necessary to strictly follow the relevant laws and regulations on data security and privacy protection, adopt technologies such as data desensitization and federated learning, and prevent data security risks while ensuring data sharing and

utilization. The core objective orientation of the model is to construct an audit risk identification system of"full coverage, intelligent identification, dynamic evaluation and accurate early warning", which specifically includes three levels: first, to realize the comprehensiveness of risk identification, break through the limitations of traditional data dimensions, and cover potential risk points in the whole process of enterprise operation; second, to improve the timeliness of risk identification, realize early discovery and early warning of risks through real-time data processing and dynamic monitoring; third, to strengthen the scientificity of risk identification, improve the accuracy and reliability of risk identification through quantitative evaluation and intelligent algorithms, provide solid data support for audit decisions, and ultimately reduce the risk of audit failure and improve audit supervision efficiency[2].

3. Construction of audit risk identification model framework based on big data technology

3.1. Design of multi-source data integration and standardization processing layer

The multi-source data integration and standardization processing layer is the basic support layer of the audit risk identification model, and its core function is to realize comprehensive collection, cleaning and integration, and standardization processing of risk data, so as to provide high-quality data resources for subsequent risk identification. In the data collection link, a "internal+external" two-dimensional data collection system is constructed. Internal data collection covers structured data such as enterprise financial systems, business operation systems, ERP systems and internal control logs, as well as unstructured data such as contract texts, meeting minutes and management discussion and analysis. Automatic collection is realized through RPA robots and API interfaces to ensure the real-time and completeness of data collection; external data collection focuses on industry dynamics, policies and regulations, tax information, public opinion data, etc. , which are obtained through web crawlers and third-party data interfaces to make up for the limitations of internal data and construct a panoramic risk data view. In the data cleaning link, intelligent cleaning algorithms are adopted to accurately process problems such as data missing, duplication and abnormality, and improve data quality through operations such as missing value completion, duplicate data elimination and outlier detection; at the same time, data lineage tracking technology is introduced to record data sources, processing processes and circulation paths, ensuring data traceability and providing guarantee for the effectiveness of audit evidence. In the data integration and standardization link, a distributed storage architecture is adopted to realize centralized storage and efficient scheduling of multi-source heterogeneous data, and different formats and dimensions of data are integrated and associated through ETL tools to construct a unified data model; unified data standardization specifications are formulated to clarify data coding, format requirements and indicator definitions, realize consistent docking of data from different sources, and eliminate data islands; at the same time, a data quality evaluation system is constructed to conduct dynamic evaluation from dimensions such as accuracy, completeness, consistency and timeliness, and establish a data quality early warning mechanism to ensure that the data sources entering the model meet the identification needs. The design of this layer can effectively break through the bottlenecks of single data dimension and uneven quality in traditional auditing, and provide comprehensive and high-quality data support for subsequent risk identification and evaluation[3].

3.2. Construction of intelligent risk feature mining and identification layer

The intelligent risk feature mining and identification layer is the core functional layer of the model, and its core goal is to mine risk features and identify abnormal patterns through big data technology to achieve accurate capture of audit risks. This layer mainly includes three core modules: risk feature system construction, intelligent feature mining and abnormal risk identification. In terms of risk feature system construction, based on the core elements of audit risk and combined with the risk characteristics of different industries and business scenarios, a multi-dimensional risk feature indicator system is constructed, covering four dimensions: financial risk characteristics, business risk characteristics, internal control risk characteristics and external environment risk characteristics. Financial risk characteristic indicators include cash flow

volatility, abnormal asset-liability ratio, abnormal changes in income and profit, etc.; business risk characteristic indicators include abnormal transaction frequency, related transaction ratio, business process compliance, etc.; internal control risk characteristic indicators include abnormal authorization and approval, lack of post balance, process execution deviation, etc.; external environment risk characteristic indicators include industry policy changes, negative market public opinion information, industry risk early warning, etc. In the intelligent feature mining module, machine learning and deep learning algorithms are integrated to construct a multi-algorithm fusion feature mining model. Unsupervised learning algorithms are used to train historical data, automatically mine implicit risk features and association rules in the data, and identify potential risk patterns; natural language processing technology is used to conduct semantic analysis and entity recognition on unstructured text data, extract risk clues from them, and convert them into quantifiable risk feature indicators; graph neural networks(GNN)are used to model complex correlations between data[4], construct risk transmission path maps, and reveal the internal connections and transmission mechanisms between different risk points. In the abnormal risk identification module, based on the constructed risk feature system and mined risk patterns, a dynamic anomaly detection model is built to identify abnormal transactions, abnormal behaviors and abnormal indicators by real-time comparing the deviation between data and normal patterns; a dynamic threshold adjustment mechanism is introduced to real-time optimize the anomaly detection threshold according to business environment changes and risk evolution trends, so as to improve the adaptability of the model to different risk scenarios; at the same time, a risk identification rule base is constructed to integrate industry experience and historical risk cases, and combine rule-based risk features with algorithm-mined features to realize dual identification of”rules+algorithms”, which not only ensures the accuracy of identification, but also improves the comprehensiveness of identification, and effectively avoids blind spots in risk identification[5].

3.3. Design of dynamic risk assessment and early warning layer

The dynamic risk assessment and early warning layer is the output and application layer of the model, and its core function is to conduct quantitative assessment, grade definition and real-time early warning on the identified risks, so as to provide scientific basis for audit decisions. In the dynamic risk assessment module, a multi-dimensional risk assessment system is constructed, integrating quantitative assessment and qualitative analysis methods to realize accurate definition of risk levels. Quantitative assessment uses analytic hierarchy process(AHP)to determine the weight of each risk characteristic indicator, and combines fuzzy comprehensive evaluation method to conduct quantitative scoring on risk indicators and calculate comprehensive risk scores; qualitative analysis modifies the quantitative assessment results combined with industry expert experience and business scenario characteristics to ensure the scientificity and rationality of the assessment results. Based on the comprehensive risk score, audit risks are divided into three levels: high, medium and low, and the definition standards and influence scope of risks at different levels are clarified to provide basis for the allocation of audit resources. In the real-time risk early warning module, a dual early warning mechanism of “indicator early warning+pattern early warning” is constructed. Warning thresholds are set for high-risk indicators, and early warning is triggered immediately when the indicators exceed the thresholds; at the same time, for the mined risk patterns, the matching scenarios in the data are monitored in real time, and early warning signals are sent immediately once potential risks are found. Early warning information is pushed in real time through a visual platform, clarifying risk points, risk levels, influence scope and potential consequences, and providing accurate risk disposal guidelines for auditors. In the dynamic optimization module, a model iteration mechanism is established to regularly optimize risk characteristic indicators, identification rules and evaluation standards combined with audit practice feedback and risk evolution data; transfer learning technology is introduced to integrate cross-industry and cross-scenario risk data to improve the generalization ability of the model in different audit objects; at the same time, a model effect evaluation system is constructed to conduct dynamic evaluation from dimensions such as identification accuracy, early warning timeliness and risk coverage, and continuously optimize model parameters according to the evaluation results to ensure that the model always maintains good identification performance. The design of this layer can realize dynamic tracking and forward-looking early warning of audit risks, promote the transformation of audit risk identification from”post-event verification”to”pre-event prevention

and in-event control", and greatly improve the audit risk management and control ability.

4. Conclusion

This paper carries out research around the construction of audit risk identification model based on big data technology, and forms a three-in-one audit risk identification model system of "data layer-technology layer-application layer "by sorting out the model construction foundation and designing the model framework. The research shows that the high adaptability between big data technology and audit risk identification can effectively break through the limitations of traditional audit risk identification, realize panoramic coverage of risk data through multi-source data integration, achieve accurate extraction of risk features with the help of intelligent algorithm mining, and realize real-time risk management and control relying on dynamic evaluation and early warning. The data-driven, systematic, dynamic adaptation and compliance principles followed in the model construction ensure the scientificity and practicality of the model, while the collaborative design of the multi-source data integration and standardization processing layer, intelligent risk feature mining and identification layer, and dynamic risk assessment and early warning layer constitutes a complete risk identification closed loop, realizing the comprehensiveness, accuracy and timeliness of audit risk identification. The construction of this model not only provides a feasible technical scheme for the digital transformation in the audit field, but also promotes the transformation of audit risk identification from experience-driven to data-driven, from static analysis to dynamic tracking, and from sampling judgment to full-volume insight. It is of great significance for improving audit quality, reducing audit risks and strengthening audit supervision efficiency.

Disclosure statement

The author declares no conflict of interest.

References

[1] Ma C, 2025, Research on Internal Audit Risk Identification and Prevention Measures of Enterprises Based on Big Data Technology. *Finance & Audit*, (2): 38-40.

[2] Qian M, 2024, Analysis of Audit Process Based on Big Data Audit Risk Model. *Research on Economic and Social Development*, (24): 0047-0049.

[3] Yuan HY, 2022, Research on Medical Insurance Fund Audit Based on Big Data Technology. *Chinese Agricultural Accounting*, (11): 45-48.

[4] Shi D, 2019, Research on Audit Risk Identification in the Era of Big Data. Shenyang Jianzhu University.

[5] Xie PJ, Jin X, Duan HR, 2018, Research on the Construction of Audit Risk Model Based on Data Mining. *Journal of Hunan University of Finance and Economics*, 34(4): 8.