Academic Journal of Management and Social Sciences
ISSN: 2958-4396 | Vol. 12, No. 3, 2025

A Transfer Learning-Based Model for Predicting Internal Control Deficiencies

Dongjie Lin *

School of Public Finance and Taxation, Central University of Finance and Economics, Beijing 102206, China
*** Corresponding author Email:** rayldj@foxmail.com

Abstract: This paper addresses the inefficiencies and heavy reliance on human judgment in traditional internal control auditing, as well as the challenges posed by sparse sample data. In this context, we propose a transfer learning‐based predictive model for internal control deficiencies. The core idea uses U.S. SOX 404 data as the source domain and Chinese A-share listed companies as the target domain. The model is designed using a "pre-training—transfer—adaptation—calibration" framework, and is developed along four dimensions: the data and labeling system, model pre-training, model fine-tuning, and evaluation metrics. The research provides explainable risk indicators and practical guidelines for regulators, audit practitioners, and corporate governance.

Keywords: Internal Control; Deficiency Prediction; Transfer Learning; Machine Learning.

1. Introduction

Internal control is an essential institutional arrangement designed to balance, constrain, and incentivize agents' behavior, aiming to mitigate information asymmetry, moral hazard, and adverse selection. Section 404 of the Sarbanes–Oxley Act (SOX) mandates that management of listed companies assess the effectiveness of internal control and disclose deficiencies, and requires the engagement of certified public accountants to audit internal control. More than one hundred empirical studies on internal control deficiencies have been published in top international accounting journals.

In China, regulators have implemented a comprehensive internal-control disclosure regime for main-board listed companies, requiring firms to disclose internal control evaluation and audit reports. However, this disclosure often fails to objectively and accurately reflect the true quality of internal control. There are two primary reasons for this issue: First, internal control deficiencies are often omitted or concealed in the disclosures [1]. Second, when disclosing internal control deficiencies, companies tend to downplay significant deficiencies by categorizing them as either important or general deficiencies [2]. As a result, models based on domestic internal control deficiency data often show poor predictive performance. This paper proposes using transfer learning to construct a model that can more accurately predict the internal control quality of Chinese listed companies, thereby enabling a more in-depth evaluation of the true risks of internal control and offering a reliable basis for regulators and investors alike.

2. Transfer Learning Concept and Principles

Transfer learning aims to reuse knowledge between the source and target domains to mitigate the constraints posed by "high noise and distribution drift" on model learning. Its starting point is not simply "copying parameters," but instead involves first learning transferable representations and stable discriminative cues related to the task in the source domain, followed by controlled alignment and calibration in the target domain to enhance the generalization performance and robustness of the model in the target domain. Zhuang et al. (2021) noted that this paradigm has formed a mature theoretical and methodological system in multimodal scenarios such as vision, text, and time series, with the key being to identify which patterns are transferable, how to transfer them, and when not to transfer [3].

In terms of the conceptual framework, transfer learning typically characterizes problems in terms of domains and tasks. A domain reflects the marginal distribution and representation space of the data, while a task reflects the labels and prediction goals. When the feature spaces of the source and target domains are consistent but their distributions differ, this is known as homogeneous transfer. On the other hand, when the feature spaces differ—such as when structured metrics and text are combined—it is referred to as heterogeneous transfer. The methods for addressing these issues can be broadly classified into three categories. First, instance-level transfer, which addresses marginal distribution differences by reweighting or resampling samples; second, representation-level transfer, which learns domain-invariant or domain-controllable intermediate representations to reduce conditional distribution differences; and third, parameter-level transfer, which shares or partially shares model parameters to prevent overfitting in small sample sizes [4].

In terms of the technical path, pre-training and fine-tuning have become the main paradigm in transfer learning. Initially, representations are learned using general or domain-related data, and then limited updates are made on the target domain with a small learning rate. The benefits of pre-training primarily come from the transferability of the representation layers. Pre-trained models generally outperform models trained from scratch or shallow baselines across hundreds of cross-domain transfer scenarios, although the extent of the benefits depends on domain relevance and alignment quality.

However, the effectiveness of transfer learning has its limits. When there is a fundamental mismatch in task semantics between the source and target domains, or when the features primarily reflect language habits rather than task-related mechanisms, transfer learning may lead to negative

transfer. The typical solution to this issue is to explicitly penalize domain differences or constrain invariant features during representation learning, and to establish minimally supervised anchor points in the target domain for threshold/probability calibration and early stopping to control risks [5].

3. Theoretical and Methodological Basis for Building the Prediction Model

The implementation of the U.S. SOX Act has accumulated a systematic record of internal control deficiencies over nearly two decades of stringent regulatory and audit practices. This source domain provides a foundation for understanding the general patterns and stable cues related to internal control failures. Internal control deficiencies are not randomly occurring but are inherently linked to factors such as the effectiveness of governance oversight, the complexity of business processes, and financial pressure. By training a baseline model using data from U.S. listed companies, we can extract the transferable parts of these patterns from empirical facts with minimal prior intervention. Based on this, the learned knowledge is transferred to the target domain of the Chinese capital market, and fine-tuned with local data to adapt to differences in China's accounting standards, market environment, and governance details, while retaining the core discriminative capabilities developed in the source domain. Unlike models that rely entirely on domestic internal control deficiency disclosures, transfer learning places greater emphasis on understanding feature distributions and common mechanisms, thus enabling the formation of a meaningful risk ranking even when the quality of real-world data is limited.

From a theoretical standpoint, the establishment of transfer learning relies on the common mechanism hypothesis. Despite differences in institutional arrangements and disclosure details between China and the United States, the underlying logic that leads to internal control failures is consistent across markets. For example, insufficient oversight at the governance level weakens the control environment, process and system complexity increases the probability of control point failures, and worsening financial conditions induce trade-offs and short-termism. These mechanisms can be observed in both markets. This paper views internal control effectiveness as a latent construct, with financial and governance indicators forming its multidimensional projection. The reuse of knowledge starting from the source domain actually injects stable priors into learning in the target domain. The generalized representations related to internal control deficiencies learned from the source domain are then slightly adjusted in the target domain to make them contextually relevant. This approach ensures that even when there are labeling errors in the target domain's data, the model serves more as a directional calibration tool than as a means of learning the decision boundary from scratch.

In terms of the methodological basis, this study adheres to three guiding principles. First, under the pragmatic principle, a pre-trained base model is selected as the starting point, with the most appropriate initial model determined by the specific research question and data conditions. Second, the research follows a systematic approach, adopting a structured knowledge transfer paradigm that clearly specifies the source, carrier, and application of knowledge. Third, verifiability is ensured by fine-tuning the model in the context of the Chinese capital market and validating it with disclosed data on internal control deficiencies from domestic listed companies. This design enables the model to become a practical tool for both auditing and regulatory decision-making.

4. Construction of the Internal Control Deficiency Prediction Model Based on Transfer Learning

This study follows the general framework of constructing the data and labeling system—pre-training—fine-tuning—evaluation. Given the low probability of internal control deficiency events, their heterogeneous features, and cross-market differences, the limited sample from a single market is insufficient to support high-quality discriminative boundary learning. Transfer learning provides a feasible path that balances generalizability and contextual adaptation. It first acquires stable representations and initial decision-making capabilities related to internal control failures from a source domain with relatively complete information, then adapts these representations gently in the target domain, enhancing sensitivity to deficiency risk and robustness in ranking without expanding the scope of the task.

4.1. Constructing the Data and Label System for Transfer Learning

The prediction model breaks down data requirements into two levels: first, the source domain dataset used for model pre-training; and second, the target domain dataset used for model fine-tuning. The source domain is selected from the U.S. capital market, benefiting from the systematic record of internal control effectiveness provided by the SOX 404 auditing and disclosure system, and from the higher availability and consistency of financial reporting information. The target domain is the sample of A-share listed companies in China. The auditing opinions issued by certified public accountants on internal control are annotated, with unqualified opinions labeled as negative samples, whereas modified opinions (qualified, adverse, or disclaimer) are labeled as positive samples. This results in a structured dataset containing financial ratios and clearly classified labels for the target domain, providing a foundation for the subsequent transfer learning process.

4.2. Pre-training the Internal Control Deficiency Prediction Model

In the pre-training stage, the U.S. capital market serves as the source domain, and a deep-learning base model is trained on financial and governance-structure data. To exploit the information contained in the source domain, the objective is not only to fit labels in that market but to learn generalizable representations that capture cross-firm regularities and stable risk patterns. Once determined, the base model serves as the starting point for designing the prediction model for Chinese listed companies. The initial transfer is not mere model reuse but knowledge initialization: model parameters learned in pre-training provide informative priors. In practice, lower-level layers that encode generic patterns are kept fixed or updated slowly, while task-specific layers are fine-tuned on the target-domain samples with conservative learning rates and regularization to prevent overfitting. Unlike traditional methods starting from random initialization, pre-training is not merely intended to improve single-domain performance but to acquire stable features and transferable decision-

making cues across domains, reducing sample complexity and improving robustness to distribution shift; subsequent calibration (e.g., threshold or probability adjustment) then aligns predictions with target-market base rates and reporting norms, equipping the model with the general ability to identify internal control deficiencies in enterprises.

4.3. Fine-tuning the Internal Control Deficiency Prediction Model

The core of the fine-tuning stage lies in faithful transfer and gentle adaptation. After loading the pre-trained weights, the model is updated by incorporating labeled samples from the target domain. The update strategy emphasizes making limited adjustments to the higher-level parameters closer to the task output layer, while freezing or semi-freezing the lower-level general representations to reduce the risk of catastrophic forgetting. The learning rate is significantly reduced compared to the pre-training phase, and conventional regularization techniques such as early stopping are used to prevent overfitting. Additionally, in response to differences in financial accounting standards, markets, and industries between the U.S. and China, key features undergo gentle reweighting or domain alignment to mitigate systemic bias caused by shifts in marginal and conditional distributions.

4.4. Evaluation Metrics for the Internal Control Deficiency Prediction Model

Given the inherent data imbalance due to the low proportion of internal control deficiency samples in practice, the use of a single accuracy metric may lead to misleading results. This paper adopts a multi-metric evaluation framework centered on ROC-AUC and PR-AUC, combined with precision, recall, and F1 score, to characterize model performance in terms of ranking quality and positive class recognition ability. Additionally, a weak baseline trained solely on target-domain data and traditional machine-learning baselines are included to assess the marginal contribution of transfer. Ablation analysis is used to compare the relative influence of key information sources on model performance.

5. Conclusion

This paper proposes a transfer learning-based framework for predicting internal control deficiencies, focusing on the reuse of knowledge from source to target domains. Using the sample data of U.S. listed companies as the starting point, the model improves the sensitivity and robustness of internal control risk identification and ranking in China's capital market through controlled contextual fine-tuning. Methodologically, the framework forms task-oriented abstract representations using multi-source information and applies gentle calibration in the target domain. In terms of evaluation, a multi-metric system is used to assess model performance, especially in low-probability scenarios, avoiding the one-sidedness of relying solely on accuracy. This research provides a methodological foundation for the introduction of intelligent auditing and regulatory technology.

Based on the findings of this research, the following policy implications are derived. First, for regulatory bodies and exchanges, transfer learning makes the concept of early identification and early intervention in regulation more actionable. Risk scores based on source-domain experience, combined with adjustments for Chinese listed companies' reporting standards, can be used for stratified resource allocation in inquiries, spot checks, and on-site inspections, prioritizing entities with weak governance, complex processes, and significant financial pressure. Second, for the auditing industry, transfer learning-based ranking tools can serve as an important input during the planning phase, helping to determine the sequence of key control points and extended procedures, improving communication efficiency with audited entities and their governance structures. Third, for corporate governance practices of listed companies, the explanatory signals from the model can be used to prioritize corrective actions, focusing on fixing controls with high correlation to deficiencies and significant spillover risks, while using disclosure optimization and process simplification as feasible paths to reduce risk exposure.

Acknowledgments

This work was supported by the Humanities and Social Sciences Research Youth Fund of the Ministry of Education of China (Project title: "Machine Learning for Internal Control Deficiency Prediction: Model Development and Applications"; Grant No. 19YJC790072).

References

[1] Lin Bin, Lin Dongjie, Xie Fan, et al. Research on the Internal Control Index Based on Information Disclosure [J]. Accounting Research (China), 2016, (12): 12–20.

[2] Ding Yougang, Wang Yongchao. Study on the Criteria for Identifying Internal Control Deficiencies in Listed Companies [J]. Accounting Research, 2013, (12): 79–85.

[3] Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, Zhu H, Xiong H, He Q. A Comprehensive Survey on Transfer Learning [J]. Proceedings of the IEEE, 2021, 109(1): 43–76.

[4] Ma Q, Liu Z, Zheng Z, Huang Z, Zhu S, Yu Z, Kwok J T. A Survey on Time-Series Pre-Trained Models [J]. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(12): 7536–7555.

[5] Fang Y, Yap P-T, Lin W, Zhu H, Liu M. Source-Free Unsupervised Domain Adaptation: A Survey [J]. Neural Networks, 2024, 174: 106230.