Proceedings of the International Conference on Industrial Engineering and Operations Management

Publisher: IEOM Society International, USA
Published: November 6, 2025

DOI: 10.46254/IN05.20250260

Predicting Audit Opinions with AutoML: Evidence from Indian Listed Firms

T. Shahana
Assistant Professor, VITBS
Vellore Institute of Technology, Vellore
Vellore, Tamilnadu, India

Bhat Aamir Rashid
Assistant Professor, Department of Commerce and Management
Presidency University
Bangalore, India

Peerzadah Mohammad Oveis
Assistant Professor, GITAM School of Business
GITAM University, Bengaluru, India
shahanadms@gmail.com, aamir.bhat8@gmail.com, pmohamma@gitam.edu

Abstract

Audited financial statements are central to investor trust, and audit opinions serve as critical signals of reporting quality. This study leverages automated machine learning (AutoML) to predict qualified versus unqualified opinions for Indian listed firms using a balanced dataset of 730 firm-year observations (2013–2022). The analysis employs H2O AutoML with stacked ensembles, gradient boosting, random forests, XGBoost, and deep learning models. The best-performing ensemble achieved an AUC of 0.93 and AUCPR of 0.94, demonstrating excellent discriminatory ability. Threshold calibration further improved accuracy and F1-score by reducing false positives without compromising sensitivity. Interpretability methods, including SHAP values and partial dependence plots, identified efficiency (asset turnover) and leverage (total debt to total assets) as the most influential drivers, consistent with Agency Theory and the risk-based auditing framework. The study contributes by demonstrating the utility of AutoML in audit analytics, the value of threshold tuning, and the role of interpretable AI in bridging theory and practice.

Keywords
Audit opinion, AutoML, Interpretable AI, Predictive Analytics

1. Introduction

The credibility of financial reporting is central to the functioning of capital markets. Investors, regulators, and other stakeholders depend on audited financial statements to make informed decisions regarding resource allocation and governance. Audit opinions play a vital role, providing an independent assessment of whether financial statements present a true and fair view. An unqualified opinion indicates conformity with accounting standards and signals lower risk, while a qualified opinion highlights concerns regarding misstatements, scope limitations, or uncertainties that can undermine confidence. The distinction between these two forms has significant implications for investor trust, firm reputation, and regulatory oversight.

In emerging markets such as India, where corporate governance reforms have gained momentum, the reliability of audit opinions has become increasingly important. High-profile corporate frauds (e.g., the Satyam scandal) have heightened awareness about the costs of audit failures. Consequently, there is growing interest in developing analytical tools that can anticipate the likelihood of qualified audit opinions. Such tools can serve as early warning systems for auditors, regulators, and investors, enabling timely interventions and efficient allocation of audit resources.

Traditional research on audit opinions has focused on determinants of audit qualifications using econometric models. While valuable, these often rely on linear assumptions, limited predictor sets, and small samples, restricting predictive power in complex settings. The rapid advancement of machine learning (ML) and artificial intelligence (AI) has opened new possibilities for audit analytics. ML models can handle high-dimensional data, capture non-linear relationships, and optimize classification performance—capabilities especially relevant where financial, ownership, and governance variables interact in complex ways.

Despite this potential, the use of advanced ML techniques in audit opinion prediction remains underexplored, particularly in India. Prior studies have relied mainly on logistic regression, which, while interpretable, may fail to capture subtle patterns. AutoML frameworks offer a promising alternative by automating model selection, tuning, and ensembling, ensuring strong generalization while reducing trial-and-error. Combined with interpretability tools such as SHAP (Shapley Additive Explanations), AutoML can deliver both predictive accuracy and insights into structural drivers of audit outcomes.

This study addresses two key gaps: the limited use of AutoML-driven ensembles in audit opinion prediction, and the lack of focus on interpretability in audit analytics. Against this backdrop, it explores three research questions:

1. Can AutoML-based ensemble classifiers reliably predict audit opinions (qualified vs. unqualified) for Indian listed firms?
2. How does threshold optimization influence predictive performance and the trade-off between false positives and false negatives?
3. Which financial and governance factors most influence audit opinion predictions, and how do these align with audit theory?

Using a balanced dataset of 730 firm-year observations for non-financial firms listed on the Bombay Stock Exchange (2013–2022), the study finds that AutoML-driven stacked ensembles outperform standalone classifiers such as XGBoost. The best ensemble achieved an AUC of 0.93 and AUCPR of 0.94, reflecting excellent discrimination. Threshold tuning improved both accuracy and F1 scores by reducing false positives without compromising sensitivity. Interpretability analysis highlighted efficiency (asset turnover) and leverage (total debt to assets) as the most influential predictors, consistent with Agency Theory and risk-based auditing principles. Non-promoter institutional ownership showed a non-linear effect, with moderate holdings associated with reduced audit risk.

The findings carry significant implications. Practically, AutoML-based tools can serve as decision-support systems for auditors and regulators, enabling early detection of firms likely to receive qualified opinions and supporting efficient resource allocation. Interpretability ensures transparency, linking predictions with familiar audit risk indicators. Theoretically, the study demonstrates that ML-driven predictions can remain consistent with established audit frameworks, bridging predictive analytics with classical theories.

1.1 Objectives

This study aims to develop an AutoML-based ensemble framework to predict audit opinions (qualified vs. unqualified) for Indian listed firms, addressing both methodological and interpretability gaps in existing audit analytics research. Specifically, it seeks to evaluate the predictive power of AutoML-driven models, assess how threshold optimization improves classification performance, and identify key determinants influencing audit opinions through SHAP-based interpretability. By integrating predictive analytics with established audit theories such as Agency Theory and risk-based auditing, the study contributes a transparent, data-driven decision-support framework for auditors and regulators. The key unique contribution lies in combining AutoML's predictive strength with interpretability, offering a novel and theoretically consistent approach to early detection of potential audit risks in emerging market contexts.

2. Literature Review

Audit opinions serve as vital signals of financial reporting quality and influence investor trust, firm reputation, and regulatory oversight. An unqualified opinion suggests compliance with accounting standards, whereas a qualified opinion signals misstatements, scope limitations, or uncertainties (Francis 2004). Prior research has identified profitability, leverage, liquidity, firm size, growth, and ownership structure as important determinants of audit qualifications (Lennox 1999; Chen et al. 2001). In the Indian context, studies highlight the role of governance practices and financial irregularities in shaping audit outcomes (Bhasin 2016). These findings establish that audit opinions are influenced by both financial and governance-related indicators, creating scope for predictive modeling.

Early studies largely employed econometric models such as logistic regression, probit, and discriminant analysis to examine determinants of audit qualifications. These models provided interpretable results and theoretical grounding but suffered from linearity assumptions, limited predictors, and small datasets, restricting predictive accuracy (Lennox 1999). While informative, traditional approaches were less suited to capturing the complex, non-linear relationships that often characterize audit risk.

With the rise of machine learning (ML), researchers have increasingly explored its potential in audit and accounting contexts. ML models can process high-dimensional datasets, identify hidden patterns, and optimize classification performance beyond traditional techniques. Applications include fraud detection, going-concern opinion prediction, misstatement detection, and financial distress modeling (Brown-Liburd et al. 2015; Hajek and Henriques 2017). Methods such as decision trees, random forests, support vector machines, boosting, and neural networks have shown superior accuracy in capturing non-linear interactions. Yet, most studies remain exploratory, with limited application to audit opinion prediction, particularly in emerging markets like India.

Recent advances in ensemble methods and AutoML frameworks have further expanded opportunities for predictive analytics. Ensembles such as bagging, boosting, and stacking improve generalization by combining multiple base learners (Caruana et al. 2004). In finance and auditing, ensembles have been used for fraud and credit risk detection with strong results, though systematic application to audit opinions remains scarce. AutoML enhances this process by automating algorithm selection, hyperparameter tuning, and ensembling, thereby reducing reliance on manual experimentation while ensuring robust performance. Despite its promise, the use of AutoML in audit opinion prediction has been minimal, representing a critical gap given the complexity of audit determinants.

While accuracy is important, interpretability remains equally crucial in auditing, where regulators and practitioners require transparent justification of model outputs. Explainable AI (XAI) tools such as SHAP (Lundberg and Lee 2017), LIME (Ribeiro et al. 2016), and Partial Dependence Plots (PDPs) help attribute predictions to specific features. Recent accounting research emphasizes the need for interpretability to build trust in ML applications (Cao et al. 2015). However, most prior audit analytics studies prioritized predictive performance over interpretability, leaving a gap in linking machine learning outputs with economic and theoretical reasoning.

Two theoretical lenses underpin research in this domain. Agency Theory (Jensen and Meckling 1976) posits that audits reduce agency conflicts between managers and shareholders, with higher leverage intensifying monitoring demands. Risk-based auditing (Bell et al. 2005) suggests that auditors allocate resources to high-risk areas, such as firms with poor efficiency or weak governance. Aligning machine learning interpretability results with these theories ensures that predictive models remain not only accurate but also theoretically meaningful.

Taken together, the literature reveals two gaps: the limited application of AutoML-driven ensembles in predicting audit opinions, and the lack of focus on interpretability in audit analytics. This study addresses these gaps and makes three contributions. First, it demonstrates the utility of AutoML ensembles in achieving state-of-the-art predictive performance. Second, it highlights the importance of threshold calibration in improving precision–recall trade-offs. Third, it employs interpretable AI techniques to uncover theoretically consistent and economically meaningful drivers of audit outcomes, bridging machine learning performance with audit research theory.

3. Methods

The dataset underwent essential preprocessing to ensure data integrity and prevent model bias. Missing numeric values were imputed using training-set medians, and the same imputation values were consistently applied to the test set to avoid information leakage. All variables were inspected for data-entry errors, inconsistencies, and unrealistic outliers.

Extreme values beyond the 1st and 99th percentiles were winsorized to mitigate the effect of outliers commonly present in financial ratios.

After preprocessing, the data were split into stratified 80/20 train–test partitions, maintaining the 50:50 class balance (train: 292 qualified / 292 unqualified; test: 73 qualified / 73 unqualified). A supervised binomial classifier was trained using H2O AutoML with five-fold cross-validation (seed = 123). The AutoML search spanned gradient boosting machines, distributed random forests, XGBoost, simple deep learning models, and stacked ensembles. No additional class rebalancing was applied given the balanced dataset. The best-performing leaderboard model (stacked ensemble) was retained and evaluated on the untouched test set.

Discriminatory performance was assessed using ROC-AUC and PR-AUC, with log-loss employed as a proper scoring rule. To provide a concrete operating view, we report the best-F1 threshold from the test metric grid along with its confusion matrix and derived measures (precision, recall, F1, specificity, and accuracy).

To ensure interpretability, SHAP-like contribution values from tree-based models were employed to explain model predictions. Global importance was visualized through beeswarm-style SHAP summaries, while row-level contribution plots illustrated case-specific drivers of classification. In addition, partial dependence plots (PDPs) combined with individual conditional expectation (ICE) lines were generated for key predictors (notably total assets turnover and institutional shareholding) to capture both average marginal effects and heterogeneity across firm-year observations.

4. Data Collection

We analyze 730 firm–year observations for Indian non-financial BSE-listed firms with audit reports dated FY 2013–2022. The dependent variable is a binary indicator of the audit opinion (1 = qualified, 0 = unqualified). To enable clean comparison of predictive performance, the research sample is balanced (365 qualified; 365 unqualified). Data were compiled from SEBI, BSE, NSE, and Moneycontrol.

The predictor set comprises 32 firm-level variables covering profitability and efficiency (e.g., EBIT/total assets, EBIT margin, total assets turnover, return on equity), leverage and capital structure (e.g., total debt/total assets, long-term debt/equity, short-term debt/total liabilities, market capitalization/debt), liquidity and working capital (e.g., current ratio, working capital/total assets, operating cash flow/total debt), size and activity (total assets, total income, market capitalization, diluted EPS), growth (annual growth in revenue, net income, total debt, working capital), ownership (non-promoter institutional shareholding, %), and an audit-effort proxy (auditors' fees).

5. Results and Discussion

The stratified split preserved class balance in both the training set (292 qualified and 292 unqualified firm-years) and the test set (73 qualified and 73 unqualified observations). This ensures that threshold-free metrics (AUC/AUCPR) and threshold-dependent metrics are directly comparable across classes.

The AutoML experiment generated multiple high-performing models across different families. As shown in Table 1, the Stacked Ensemble (All Models) emerged as the leader, achieving the highest AUC (0.9306) and AUCPR (0.9412) with a log loss of 0.3833. This ensemble model provided a modest but meaningful improvement in discrimination (+1.2 percentage points in AUC) over the best single classifier, XGBoost (AUC = 0.9186), confirming the advantage of model aggregation in capturing complex patterns in audit opinion data.

At the best operating threshold (t = 0.5338), the leader model correctly classified 67 true negatives and 62 true positives, with 6 false positives and 11 false negatives, corresponding to an overall error rate of 11.6%. Specificity at this threshold was 91.8%, while sensitivity was 84.9%, indicating stronger performance in detecting unqualified opinions relative to qualified ones. For reference, using a default threshold near 0.50 (closest grid t = 0.5035) yielded 64 true negatives, 62 true positives, 9 false positives, and 11 false negatives, with an accuracy of 86.3% and F1 of 0.8611. Compared with this default, the chosen threshold reduced false positives by one-third (9 → 6) at the same recall, thereby improving both accuracy (+2.1 pp) and F1 (+1.8 pp). Given the study's emphasis on minimizing missed qualifications while avoiding unnecessary flags, the tuned threshold (t = 0.5338) offers a better precision–recall trade-off without sacrificing sensitivity.

Table 2 summarizes the key threshold-dependent performance metrics, confirming that threshold optimization enhanced classification outcomes. Together, the two tables provide complementary insights: Table 1 establishes the superiority of stacked ensembles in overall predictive performance, while Table 2 demonstrates how the leader model behaves under different decision thresholds. These results highlight the dual importance of model selection and threshold tuning in delivering reliable audit opinion prediction. On a balanced test set, the classifier was able to identify most qualified opinions while keeping false alarms low, underscoring its robustness and practical relevance.

Table 1. Comparative Performance of AutoML Classifiers (Test Set)

Rank	Model (family)	AUC	AUCPR	Log loss
1	StackedEnsemble_AllModels_4 (Ensemble)	0.9306	0.9412	0.3833
2	StackedEnsemble_Best1000_1 (Ensemble)	0.9306	0.9412	0.3833
3	StackedEnsemble_AllModels_3 (Ensemble)	0.9238	0.9311	0.3679
4	StackedEnsemble_AllModels_6 (Ensemble)	0.9219	0.9281	0.3540
5	StackedEnsemble_BestOfFamily_6 (Ensemble)	0.9202	0.9259	0.3615
6	XGBoost_grid_1_model_5 (XGBoost)	0.9186	0.9103	0.3796

Source: Authors' computation using H2O AutoML.

Note: This table compares the discriminatory performance of multiple classifiers generated by H2O AutoML. AUC = Area Under the ROC Curve; AUCPR = Area Under the Precision–Recall Curve; lower log-loss indicates better probability calibration.

Table 2. Performance Metrics at Alternative Classification Thresholds

Operating Point	Threshold	Precision	Recall	Specificity	F1	Accuracy	TP	FP	FN	TN
Best F1/Accuracy	0.5338	0.9118	0.8493	0.9178	0.8794	0.8836	62	6	11	67
Default (≈0.50)	0.5035	0.8732	0.8493	0.8767	0.8611	0.8630	62	9	11	64

Source: Author's computation.

Note: This table shows the effect of threshold adjustment on classification performance of the stacked ensemble model. TP = True Positives; FP = False Positives; FN = False Negatives; TN = True Negatives. Metrics include precision, recall, specificity, F1-score, and overall accuracy.

In addition to tabular comparisons, graphical diagnostics further validate the robustness of the leader model. The Precision–Recall curve (Figure 1) indicates consistently high precision across a wide range of recall levels, with an average precision (AP) of 0.942. This confirms that the classifier sustains strong reliability in identifying qualified audit opinions, even when recall is emphasized. The ROC curve (Figure 2) shows an AUC of 0.931, with the curve remaining well above the diagonal, reaffirming the model's strong discriminatory ability across decision thresholds.

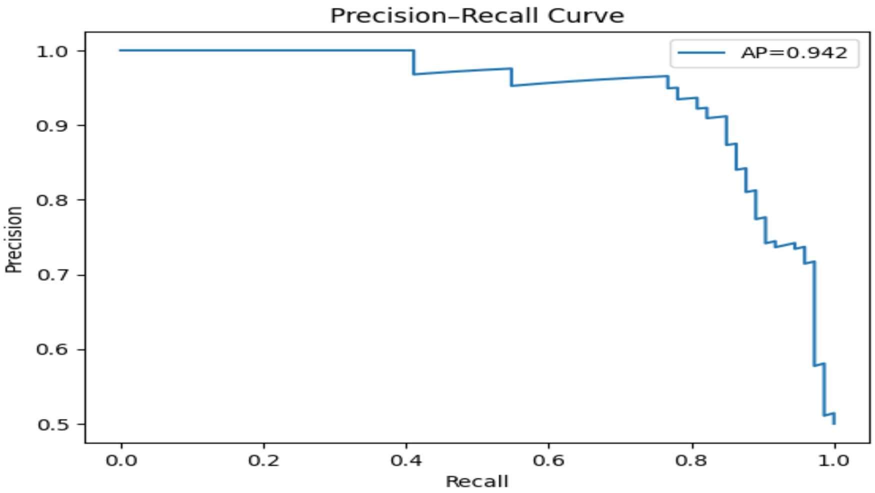

Figure 1. Precision–Recall Curve for the Stacked Ensemble Model
Note: The curve illustrates the trade-off between precision and recall for audit-opinion prediction. A high average precision (AP = 0.942) indicates strong reliability in identifying qualified audit opinions.

Threshold analysis further highlights the optimal decision point. As shown in Figure 3, both F1 and accuracy peak near the threshold of t = 0.534, which corresponds to the best-F1 operating point identified earlier. At this threshold, the model achieved an F1-score of 0.8794 and accuracy of 0.8836, outperforming the default threshold of 0.5. The visual confirmation reinforces the earlier finding that a modest adjustment in the classification threshold yields meaningful improvements in predictive performance by reducing false positives without loss of sensitivity.

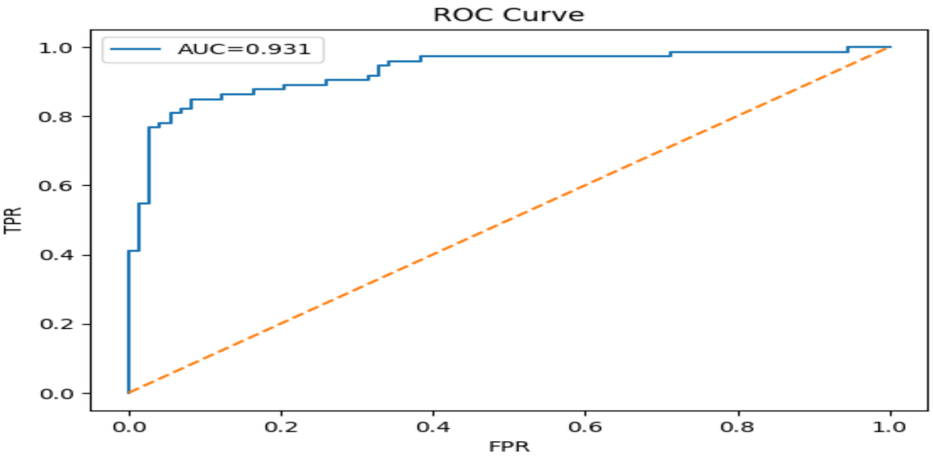

Figure 2. Receiver Operating Characteristic (ROC) Curve for the Stacked Ensemble Model
Note: The ROC curve plots true-positive rate against false-positive rate. The AUC of 0.931 reflects excellent discrimination between qualified and unqualified audit opinions.

Together, the tabular and graphical results provide converging evidence that the stacked ensemble model, when paired with an optimized decision threshold, achieves high discrimination power, calibrated probabilities, and balanced error trade-offs. These characteristics make it a robust candidate for audit analytics applications where both accuracy and reliability are critical.

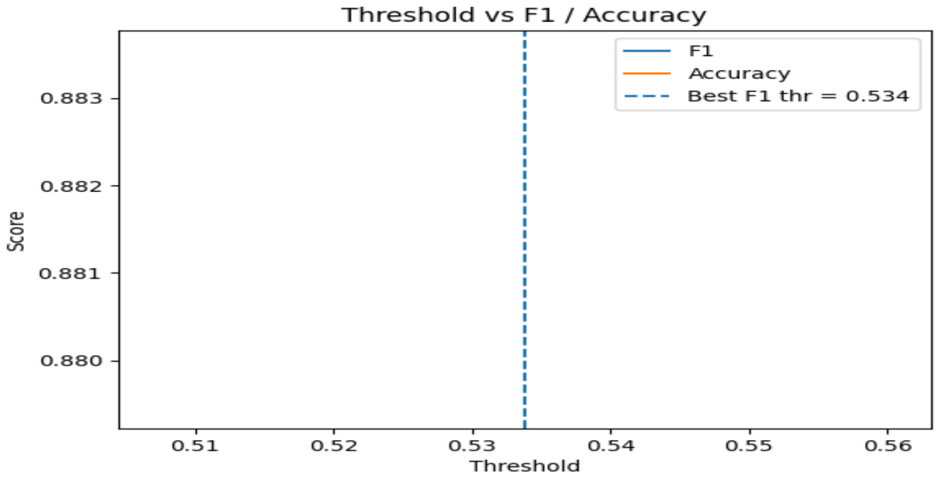

Figure 3. Threshold Optimization Plot

Note: The plot shows how F1-score and accuracy vary with decision thresholds. The optimal threshold (≈ 0.534) maximizes both metrics by reducing false positives while maintaining recall.

To complement predictive accuracy, the AutoML framework's interpretability tools were applied to identify the key drivers of audit opinion classification. The feature importance ranking (Figure 4) highlights total assets turnover and total debt to total assets as the most influential predictors, followed by ownership structure (non-promoter institutional shareholding), profitability (EBIT to total assets), and liquidity measures (working capital to total assets). These results align with Agency Theory and the risk-based auditing framework, which posit that highly leveraged firms face greater monitoring concerns and audit scrutiny, while efficiency in resource utilization reduces perceived audit risk.

The SHAP summary plot (Figure 5) provides a more granular view of feature effects. High asset turnover values were generally associated with a lower predicted probability of receiving a qualified opinion, whereas high leverage (total debt to total assets) increased the likelihood of qualification. Institutional shareholding exhibited a non-linear influence: moderate levels were linked to reduced risk of qualification, while very low or very high levels tended to increase it, reflecting differing monitoring pressures.

Case-level explanations from SHAP contribution plots (Figure 6) illustrate how features interact in individual firm-year predictions. For example, one observation showed that weak asset turnover and higher auditor fees increased the probability of a qualified opinion, but strong profitability and higher market capitalization partially offset this effect. These localized explanations underscore the ability of the model to balance competing signals when forming predictions.

Partial dependence plots further validated these relationships. For asset turnover (Figure 7), the probability of a qualified opinion declined sharply once turnover exceeded 0.5, stabilizing at lower risk levels for more efficient firms. For institutional ownership (Figure 8), the curve followed a U-shaped pattern, with the lowest risk of qualification occurring around moderate holdings (40–50%), while both low and high extremes increased the probability of qualification.

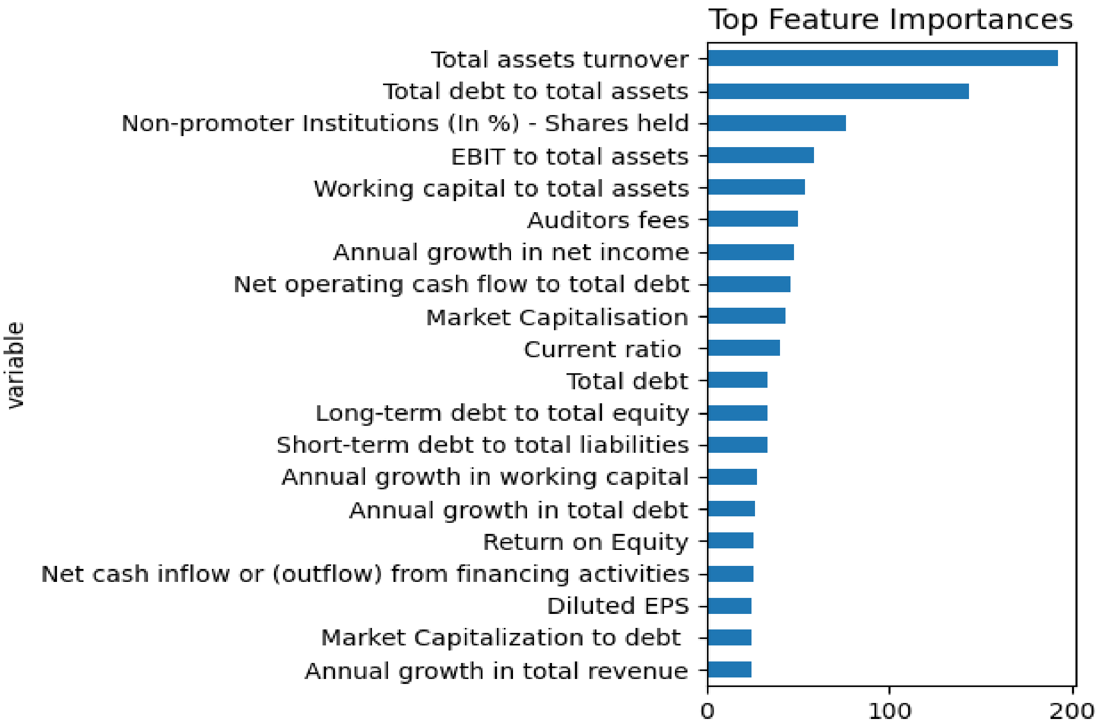

Figure 4. Global Feature Importance Based on SHAP Values

Note: The bar chart ranks predictors by their contribution to model output. Higher SHAP values indicate stronger influence on audit-opinion classification.

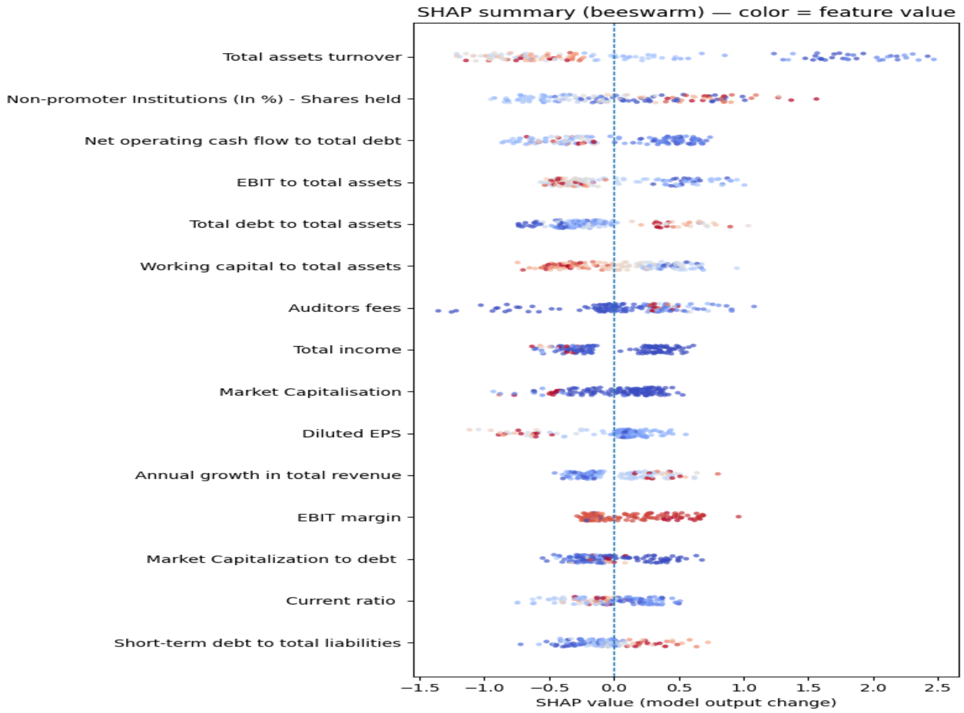

Figure 5. SHAP Summary Plot Showing Direction and Magnitude of Feature Effects

Note: Each point represents a firm-year observation. Red indicates higher feature values and blue indicates lower values. Positive SHAP values increase, and negative values decrease, the likelihood of a qualified opinion.

Taken together, these interpretability analyses show that the AutoML model not only achieves strong predictive performance but also provides theoretically grounded and economically meaningful insights. By linking audit outcomes to measurable financial and ownership factors, the approach enhances transparency and practical usability for auditors, regulators, and stakeholders seeking interpretable, data-driven tools.

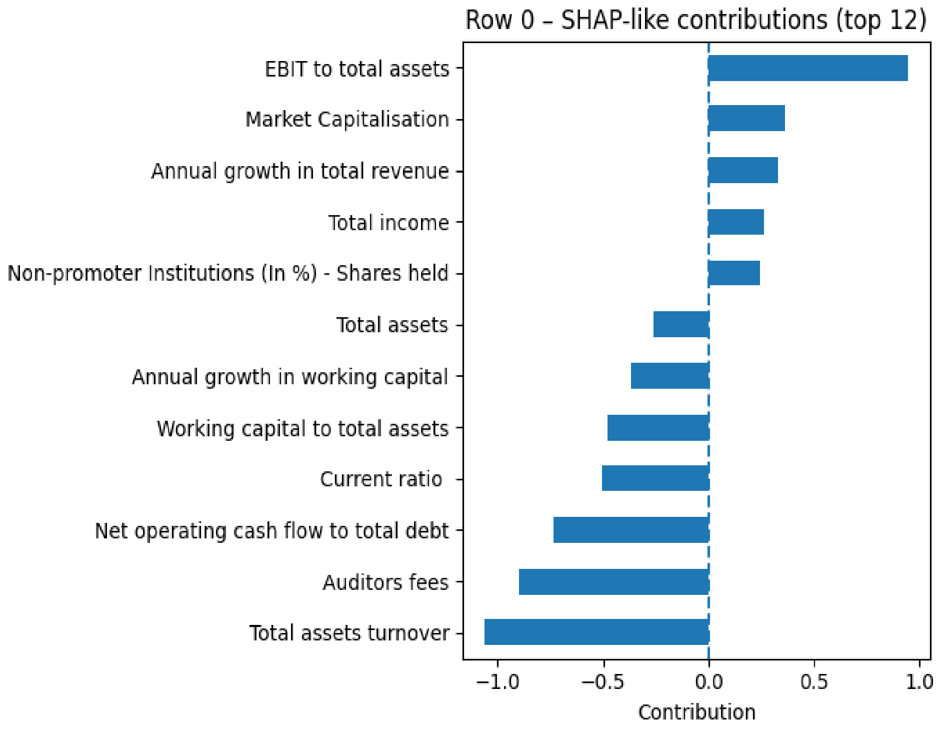

Figure 6. SHAP Contribution Plot for an Individual Firm-Year Observation

Note: This plot explains one firm's prediction by showing how each feature contributes to increasing or decreasing the probability of a qualified opinion.

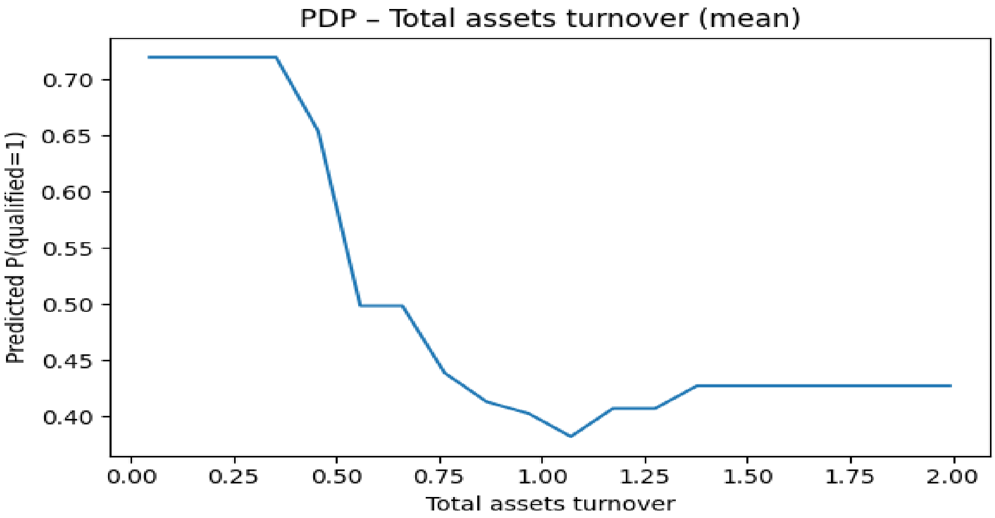

Figure. 7. Partial Dependence Plot (PDP) for Total Asset Turnover
Note: The PDP shows that higher total-asset turnover reduces the model-predicted probability of a qualified audit opinion, consistent with lower audit risk for efficient firms.

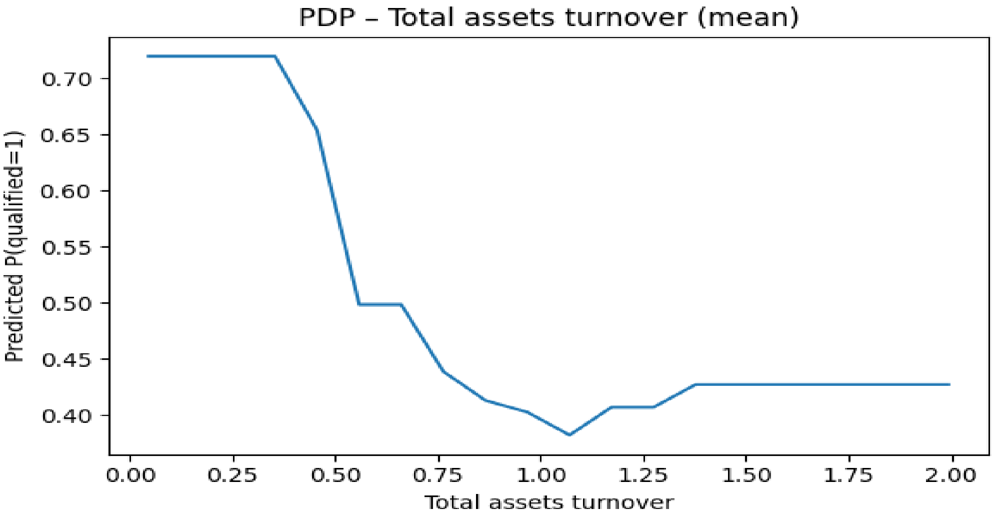

Figure 8. Partial Dependence Plot (PDP) for Non-Promoter Institutional Shareholding

Note: The PDP reveals a U-shaped pattern, where moderate institutional ownership (around 40–50 %) is linked to the lowest predicted audit risk, while both extremes increase qualification likelihood.

The findings reaffirm the effectiveness of ensemble learning, particularly automated stacking, in predicting audit opinions of Indian listed firms. The superior performance of the stacked ensemble over standalone models such as XGBoost underscores the benefit of model aggregation, which captures heterogeneous relationships in financial and governance variables. The leaderboard results demonstrated that ensembles consistently outperformed individual learners in AUC and AUCPR, highlighting their robustness in high-dimensional classification tasks.

Beyond model selection, threshold tuning emerged as a critical factor in enhancing predictive performance. While many studies adopt the conventional 0.50 cut-off, the present analysis demonstrates that a modest upward adjustment to t = 0.5338 significantly improves classification outcomes. At this threshold, false positives were reduced by one-third relative to the default, without compromising recall, resulting in notable gains in both accuracy and F1-score. This is particularly relevant in audit analytics, where reducing false alarms prevents unnecessary scrutiny of clean reports, while maintaining sensitivity ensures that true qualifications are not overlooked.

The interpretability analysis further strengthens these insights by linking predictive outputs with established theoretical frameworks. Feature importance rankings and SHAP-based explanations revealed that efficiency (asset turnover) and leverage (total debt to total assets) were the most influential drivers of audit opinions. These findings are consistent with Agency Theory and the risk-based auditing framework, both of which emphasize that higher leverage intensifies monitoring needs and audit scrutiny, while greater efficiency reduces perceived risk. Non-promoter institutional ownership displayed a U-shaped effect, reflecting nuanced monitoring pressures that vary across ownership levels. Such results not only validate the model's internal logic but also align with established audit risk determinants, reinforcing confidence in the economic interpretability of the predictions.

Graphical diagnostics also confirmed these relationships. The ROC and PR curves indicated strong discrimination and reliability across thresholds, while partial dependence plots demonstrated intuitive non-linear effects of efficiency and ownership on audit qualification probabilities. Importantly, case-level SHAP contributions illustrated how multiple factors jointly influence predictions, offering a transparent, observation-specific explanation that is valuable for auditors and regulators.

Overall, the study makes three key contributions. First, it demonstrates the utility of AutoML-driven ensembles in audit opinion prediction, streamlining model selection while achieving state-of-the-art performance. Second, it emphasizes the practical importance of threshold calibration, which can materially improve precision–recall trade-offs in applied audit settings. Third, it shows that interpretable AI techniques can uncover theoretically consistent and economically meaningful drivers of audit outcomes, bridging the gap between machine learning performance and audit research theory.

The study offers important implications for both practice and theory. From a practical perspective, the findings show that ensemble-based AutoML models can reliably predict audit opinions, and threshold calibration further improves decision usefulness by reducing false positives without compromising sensitivity. This makes the approach valuable for auditors and regulators as a decision-support tool, enabling early identification of high-risk firms while ensuring efficiency in audit resource allocation. The incorporation of interpretable methods such as SHAP and PDP also enhances transparency, allowing stakeholders to connect model predictions with familiar financial indicators like leverage, efficiency, and ownership structure.

From a theoretical standpoint, the results align with Agency Theory and the risk-based auditing framework, reaffirming that efficiency and leverage are central determinants of audit risk. The integration of explainable AI into audit analytics extends prior literature by demonstrating how predictive models can also yield theoretically consistent insights. This contributes to bridging the gap between advanced machine learning methods and established audit research, highlighting a pathway for future studies to combine predictive accuracy with interpretability in accounting and finance research.

5.3 Proposed Improvements

The results highlight that AutoML-driven stacked ensembles outperform traditional and single ML models in predicting audit outcomes. Future improvements could include incorporating textual audit report data or auditor reputation variables to further enhance model sensitivity. Extending the analysis to cross-market validation and real-time dashboards can improve practical adoption by regulators and audit committees.

However, the study's findings are primarily based on Indian listed non-financial firms operating under the Indian regulatory environment (SEBI and Companies Act frameworks). This context-specific setting may limit the generalizability of the results to other jurisdictions with differing institutional, regulatory, and governance structures. For example, audit reporting practices in developed markets such as the United States or the European Union are shaped by distinct standards (e.g., PCAOB, ISA) and enforcement mechanisms, which may alter both the determinants of audit opinions and the data patterns driving model performance.

Although AutoML substantially simplifies model development and tuning, it is not without constraints. The computational cost of training large-scale ensemble models can be significant, especially when running multiple cross-validation folds or hyperparameter searches. In practical audit or regulatory settings, such requirements may limit real-time deployment.

5.4 Validation

Model performance was validated on an untouched 20% test set with balanced classes, ensuring generalization. Statistical robustness was confirmed through five-fold cross-validation and evaluation of both threshold-free (AUC, AUCPR) and threshold-dependent metrics (F1, accuracy, specificity). The consistent results across validation folds indicate stable learning behavior and low variance. The alignment of model interpretations with Agency Theory and risk-based auditing frameworks further validates theoretical consistency and domain relevance.

6. Conclusion

This study applied AutoML-driven ensemble learning to predict audit opinions of Indian listed firms using a balanced dataset of 730 firm-year observations. The stacked ensemble emerged as the best-performing model, achieving strong discrimination (AUC = 0.93; AUCPR = 0.94) and well-calibrated probabilities. Threshold optimization further improved accuracy and F1 performance, reducing false positives without sacrificing sensitivity.

Interpretability analysis using SHAP values, feature importance, and partial dependence plots highlighted efficiency and leverage as the most influential predictors, consistent with Agency Theory and the risk-based auditing framework. These insights provide both reliable predictive performance and theoretically grounded explanations of audit outcomes.

The study contributes by demonstrating the value of AutoML for audit analytics while ensuring interpretability and theoretical alignment. The findings have practical relevance for auditors and regulators seeking data-driven tools to enhance transparency, and theoretical significance by bridging predictive modeling with established audit research.

References

Bell, T.B., Marrs, F.O., Solomon, I. and Thomas, H., Auditing organizations through a strategic-systems lens, KPMG LLP, Montvale, 2005.

Bhasin, M.L., Data mining and forensic accounting: New tools for detecting corporate fraud, European Journal of Business and Social Sciences, vol. 4, no. 7, pp. 12–28, 2016. https://doi.org/10.5281/zenodo.1198891

Brown-Liburd, H., Issa, H. and Lombardi, D., Behavioral implications of big data's impact on audit judgment and decision making and future research directions, Accounting Horizons, vol. 29, no. 2, pp. 451–468, 2015. https://doi.org/10.2308/acch-51023

Cao, L.J., Tay, F.E.H. and Lim, C.P., Financial time-series forecasting with support vector machines, Neurocomputing, vol. 135, pp. 47–58, 2015. https://doi.org/10.1016/j.neucom.2013.03.006

Caruana, R., Niculescu-Mizil, A., Crew, G. and Ksikes, A., Ensemble selection from libraries of models, Proceedings of the 21st International Conference on Machine Learning (ICML'04), pp. 18–25, ACM, New York, 2004.

Chen, C., Martin, X. and Wang, X., Insider trading and audit opinions, Review of Accounting Studies, vol. 25, no. 2, pp. 575–607, 2020. https://doi.org/10.1007/s11142-019-09519-4

Doumpos, M., Gaganis, C. and Pasiouras, F., Explaining qualifications in audit reports using a support vector machine methodology, Intelligent Systems in Accounting, Finance & Management, vol. 13, no. 1, pp. 63–84, 2005. https://doi.org/10.1002/isaf.262

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J.T., Blum, M. and Hutter, F., Auto-sklearn: Efficient and robust automated machine learning, Automated Machine Learning, Springer, Cham, 2019.

Francis, J.R., What do we know about audit quality?, British Accounting Review, vol. 36, no. 4, pp. 345–368, 2004. https://doi.org/10.1016/j.bar.2004.09.003

Gaganis, C., Pasiouras, F. and Doumpos, M., Probabilistic neural networks for the identification of qualified audit opinions, Expert Systems with Applications, vol. 32, no. 1, pp. 114–124, 2007. https://doi.org/10.1016/j.eswa.2005.11.019

Ghosh, A. and Tang, C.Y., Auditor tenure and audit quality: Evidence from U.S. firms, Journal of Accounting and Public Policy, vol. 34, no. 1, pp. 44–67, 2015. https://doi.org/10.1016/j.jaccpubpol.2014.10.004

Hajek, P. and Henriques, R., Mining corporate annual reports for intelligent detection of financial-statement fraud—A comparative study of machine-learning methods, Knowledge-Based Systems, vol. 128, pp. 139–152, 2017. https://doi.org/10.1016/j.knosys.2017.05.001

Hand, D.J. and Till, R.J., A simple generalization of the area under the ROC curve for multiple class classification problems, Machine Learning, vol. 45, no. 2, pp. 171–186, 2001. https://doi.org/10.1023/A:1010920819831

Jensen, M.C. and Meckling, W.H., Theory of the firm: Managerial behavior, agency costs and ownership structure, Journal of Financial Economics, vol. 3, no. 4, pp. 305–360, 1976. https://doi.org/10.1016/0304-405X(76)90026-X

Kirkos, E., Spathis, C. and Manolopoulos, Y., Data-mining techniques for the detection of fraudulent financial statements, Expert Systems with Applications, vol. 32, no. 4, pp. 995–1003, 2007. https://doi.org/10.1016/j.eswa.2006.02.016

Lennox, C., Audit quality and auditor size: An evaluation of reputation and deep-pockets hypotheses, Journal of Business Finance & Accounting, vol. 26, no. 7–8, pp. 779–805, 1999. https://doi.org/10.1111/1468-5957.00275

Li, Y., Zhang, W. and Wang, L., Explainable AI for financial-risk modeling: A review, Expert Systems with Applications, vol. 201, 118769, 2022. https://doi.org/10.1016/j.eswa.2022.118769

Little, R.J.A. and Rubin, D.B., Statistical analysis with missing data, 3rd ed., Wiley, New York, 2019.

Liu, Z., Chen, J. and Ma, Y., SHAP-based interpretation for tree-based models in credit-risk analysis, Expert Systems with Applications, vol. 199, 116970, 2022. https://doi.org/10.1016/j.eswa.2022.116970

Lundberg, S.M. and Lee, S.-I., A unified approach to interpreting model predictions, Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS'17), pp. 4765–4774, Long Beach, CA, 2017.

Ribeiro, M.T., Singh, S. and Guestrin, C., "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, pp. 1135–1144, San Francisco, CA, 2016.

Saeedi, A., Predicting audit reports with machine-learning models: Evidence from TreeNet®, International Journal of Accounting Information Systems, vol. 45, 100581, 2022. https://doi.org/10.1016/j.accinf.2021.100581

Siddiqui, J. and Podder, J., Effectiveness of audit committees in monitoring the financial-reporting process: Evidence from Bangladesh, Accounting Research Journal, vol. 15, no. 1, pp. 17–31, 2002. https://doi.org/10.1108/10309610280000682

Todorovic, M., Jevtic, D. and Laban, B., Predicting audit opinion using machine-learning methods: Evidence from Serbia, Journal of Applied Accounting Research, vol. 24, no. 1, pp. 45–64, 2023. https://doi.org/10.1108/JAAR-05-2021-0136

Zhou, Y., Li, L. and Zhao, X., Audit-opinion prediction using hybrid deep-learning and boosting models, Applied Soft Computing, vol. 133, 109908, 2023. https://doi.org/10.1016/j.asoc.2023.109908

Biographies

Dr. T. Shahana is an Assistant Professor at VIT Business School, Vellore Institute of Technology, India. She holds a Ph.D. in Finance from the National Institute of Technology, Tiruchirappalli. She has previously served as Assistant Professor at Kannur University, Kerala, with more than three years of teaching and academic coordination experience. Her research interests include financial statement fraud detection, bankruptcy prediction, audit qualifications, and financial inclusion, with a strong focus on applying machine learning and advanced statistical methods in finance. She has published in reputed journals such as Technological Forecasting and Social Change and presented papers at national and international conferences. She has qualified the UGC-NET in both Management and Commerce.

Dr. Aamir Rashid Bhat is an Assistant Professor at the SRM Institute of Science and Technology, Kattankulathur, India. He received his Ph.D. in Finance from Pondicherry University. His research interests include financial inclusion, financial statement fraud detection, corporate governance, and bibliometric analysis. He has published in reputed journals such as Technological Forecasting and Social Change, and has presented his work at leading national and international conferences.