Accounting Fraud Detection Project Report

By: Bini Teklehaimanot

# Problem statement

Can we develop a fraud detection framework that generalizes across financial ecosystems; specifically accounting related financial reporting of publicly traded US companies by leveraging structured data and multi layers feature interactions using deep neural network (DNN)?

# Data Wrangling

The project began with the raw dataset sourced from the [FraudDetection GitHub repository](https://github.com/JarFraud/FraudDetection), which contains financial statement data from publicly traded U.S. firms alongside confirmed misstatement labels. The raw CSV was inspected for structure, missingness, and formatting inconsistencies.

Key preprocessing steps included:

* **Missing value handling**: Features with excessive missingness were dropped, while others were imputed using domain-appropriate strategies (e.g., filling with zero or industry averages).
* **Feature pruning**: Redundant or ID-based columns (e.g., gvkey, fyear) were removed to prevent data leakage and reduce dimensionality.
* **Derived features**: Several financial ratios were engineered from raw variables to enhance interpretability and predictive value.
* **Column normalization**: Continuous features were standardized using a ColumnTransformer with appropriate scalers (RobustScaler, StandardScaler) depending on feature distribution.

The cleaned dataset retained over 40 numeric financial indicators ready for EDA and modeling.

A screenshot of a computer screen

AI-generated content may be incorrect.

The custom pairplot above illustrates density and scatter comparisons of key engineered features by fraud label (misstate). Notably, fraudulent filings tend to cluster around specific ranges of reoa and dpi, indicating these features may hold predictive value—an insight that helped prioritize features for downstream modeling.

# Exploratory Data Analysis

The dataset consists of structured financial statement data from publicly traded U.S. firms, covering fiscal year 2023, and includes labeled cases of confirmed accounting misstatements. The EDA phase focused on identifying potential red flags and distributional shifts between fraudulent and non-fraudulent financial reports. Since this was a supervised classification problem with the target variable misstate (1 = fraud, 0 = non-fraud), all explorations were conducted with label awareness in mind.

One of the most informative tools during this phase was the **lower-triangle correlation heatmap**, which helped uncover multicollinearity and potential feature relationships. The matrix below shows modest correlations between misstate and several financial variables - particularly rect (accounts receivable), ni (net income), and pstk (preferred stock) - highlighting their potential relevance for fraud detection models.

A graph of a number of red and blue squares

AI-generated content may be incorrect.

To better understand class-specific behavior, we also visualized the distribution of key engineered features—such as the book-to-market ratio (bm)—separated by fraud label. The histogram below shows a clear difference in the density curve of bm between misstated and non-misstated cases. While overlapping exists, this kind of separation in distributions guided the modeling effort by emphasizing which features showed class-discriminative power.

A graph with a line graph

AI-generated content may be incorrect.

These findings laid the foundation for feature selection and informed downstream modeling decisions—including how to handle class imbalance and which transformations to apply for better model interpretability.

Treatment of Outliers:

During exploratory data analysis, several features exhibited extreme values and long-tailed distributions. These outliers were not removed or capped. Given that the dataset represents publicly traded U.S. companies spanning a wide spectrum of industries, sizes, and reporting practices, such variation is expected. Excluding these values could suppress meaningful signals from large-cap firms or those with atypical financial profiles—both of which are critical to the integrity of fraud detection modeling. Preserving the full distribution ensures that the model reflects the diversity and complexity inherent in real-world financial reporting.

# Model Preprocessing with feature engineering

The modeling process began by designating the misstate column as the target variable (1 = misstatement, 0 = no misstatement). Features used for prediction were derived from structured financial data sourced from publicly traded U.S. companies. After confirming consistency between the training and evaluation splits, we adopted a stratified 80/20 train-test split to maintain class balance and support generalizability.

To improve model interpretability and performance, we applied a mixed-scaler ColumnTransformer to handle varying feature distributions. This included using a StandardScaler for normally distributed variables and a RobustScaler for features with outliers.

Feature importance was a critical part of our strategy. Using CatBoost's built-in feature ranking, we identified the top variables driving predictions. The plot below shows the top 20 features, highlighting attributes like operating cash flow ratio (ch\_cm), accounts payable (ap), change in working capital (dch\_wc), and retained earnings (re) as dominant fraud indicators.

📊 **Top 20 CatBoost Feature Importances**

A chart of a catboost

AI-generated content may be incorrect.

To understand how sensitive fraud probability is to changes in top features, we simulated controlled perturbations of the four most important variables. We adjusted each one independently from -20% to +20% in 5% increments and measured the corresponding effect on predicted fraud probability using our final blended model (CatBoost + DNN).

One illustrative result is shown below for the **Days Payable Index**. As the index increases, the model predicts a rising likelihood of fraud—suggesting that firms delaying payments to suppliers might signal red flags.

📈 **Simulation Results for Days Payable Index**

A graph with blue lines and dots

AI-generated content may be incorrect.

These simulations not only validate the model’s directional logic, but also help financial analysts identify leverage points for risk mitigation.

# Fraud Detection findings and recommendations

**📍 Current Position: Accounting Fraud in Public Companies**

Our modeling analysis confirms that financial misstatements in publicly traded U.S. companies exhibit detectable patterns within structured accounting data. By integrating a Deep Neural Network (DNN) with a tree-based model (CatBoost), we developed a hybrid fraud detection framework that demonstrates strong predictive capability despite pronounced class imbalance. This framework establishes a scalable foundation for identifying accounting irregularities across diverse financial reporting environments.

**📈 Key Modeling Insights: How Access Could Improve**

The blended model revealed several high-impact financial indicators linked to increased fraud risk. Among the top-ranked features were:

* **Operating Cash Flow Ratio**
* **Accounts Payable**
* **Change in Working Capital**
* **Retained Earnings**

Simulation analysis showed that even modest changes (±5–20%) in these features significantly altered the model’s predicted fraud probability. For instance, a decrease in the **Operating Cash Flow Ratio** corresponded to a measurable increase in fraud likelihood, signaling potential liquidity manipulation. Likewise, irregular shifts in **Working Capital** or **Retained Earnings** flagged accounting inconsistencies tied to earnings management.

Importantly, the DNN model alone was less responsive to class imbalance, but when blended with CatBoost, the ensemble benefited from DNN’s complex pattern recognition and CatBoost’s calibrated probabilities and feature interpretability.

These insights suggest that fraud detection models should account not only for traditional financial ratios but also for how they fluctuate over time and in relation to one another.

**🚀 Strategic Approach for Business Leadership**

This model equips financial executives, audit committees, and compliance teams with a data-driven tool for proactive fraud surveillance. Rather than relying solely on retrospective audits or whistleblower reports, organizations can integrate this model into existing financial reporting workflows to prioritize high-risk filings for further review.

Key strategic applications include:

* **Early Warning System**: Use the model to flag abnormal financial behavior before formal audits begin, helping to allocate internal resources more efficiently.
* **Risk-Based Audit Planning**: Enhance audit strategies by identifying specific accounts—such as retained earnings or payables—where risk is elevated.
* **Compliance Monitoring**: Integrate model outputs with internal controls systems to ensure consistent fraud screening at scale.
* **Investor Relations and Governance**: Demonstrate commitment to financial transparency and risk mitigation by adopting advanced fraud detection protocols.

By combining transparency, scalability, and adaptability, the framework offers a pathway to operationalize predictive risk scoring within routine finance and audit cycles.

**🛡️ Governance & Risk Applications**

This fraud detection framework presents a valuable opportunity for improving financial governance and strengthening internal risk management programs. Its ability to surface subtle irregularities across structured accounting data enables organizations to enhance oversight before financial misstatements escalate into regulatory violations or reputational damage.

Key applications include:

* **Risk-Based Controls Enhancement**  
  Integrate model outputs into SOX 404 compliance reviews to dynamically assess risk across accounts. Features like changes in working capital or cash flow ratios can trigger targeted testing.
* **Audit Committee Reporting**  
  Use the model's insights as part of quarterly audit committee briefings to highlight emerging patterns, outlier behavior, and shifts in fraud risk posture.
* **Internal Audit Prioritization**  
  Refine internal audit planning by using predictive flags to prioritize high-risk business units or accounts for deeper forensic analysis.
* **Ethics & Compliance Alignment**  
  Link fraud probability scores with whistleblower intake channels and policy compliance logs to identify correlations between data-driven risk signals and employee-reported concerns.
* **Continuous Monitoring**  
  Establish a near real-time monitoring system where financial statements are automatically scored against the fraud risk model during reporting cycles, creating a closed-loop alerting mechanism.

Incorporating machine learning into governance practices doesn't replace human judgment—it enhances it. This model empowers compliance leaders to shift from reactive detection to proactive prevention, elevating the strategic role of internal controls in enterprise risk frameworks.

**🔮 Future Improvements: Which Scenarios Should Be Studied Further?**

While the current framework demonstrates strong predictive power, several avenues exist for expanding its scope and impact:

* **Temporal Modeling**  
  Incorporating time-series features—such as year-over-year changes or trend velocities—could enhance detection of gradual manipulation or irregular fluctuations not visible in single-period snapshots.
* **Textual Disclosures & Unstructured Data**  
  Augmenting the model with NLP-based analysis of 10-K filings, management discussion sections, or audit opinion language could provide critical context and catch tone-based or linguistic fraud indicators.
* **Industry-Specific Customization**  
  Sector-level calibration (e.g., financial vs. manufacturing) could improve precision by accounting for differences in accounting norms, seasonal cash flow behavior, or regulatory complexity.
* **Explainability Enhancements**  
  Integrating SHAP (SHapley Additive exPlanations) or LIME could make the DNN component more interpretable, offering clearer justification for flagged risks—critical for adoption by compliance and audit professionals.
* **Expanded Datasets**  
  Leveraging international filings, private company data (where available), or merging with ESG or executive compensation datasets could expose broader fraud dynamics and incentivized behaviors.

These enhancements would not only improve detection accuracy but also increase stakeholder trust in model outputs. Future iterations could also be deployed as part of automated policy simulations, scenario testing, or fraud scoring APIs accessible across enterprise risk systems.

# Conclusion: A Scalable, Interpretable Framework for Modern Fraud Detection

This project demonstrates that financial fraud is not only detectable—it is quantifiable through machine learning models grounded in structured accounting data. By combining the interpretability of CatBoost with the pattern recognition strength of DNNs, we created a hybrid model capable of flagging subtle misstatement patterns with high precision, even under severe class imbalance.

Beyond technical success, the framework is designed for operational deployment. It integrates naturally with internal audit workflows, supports SOX compliance, and can guide risk-based decision-making at both tactical and strategic levels. Its modular architecture allows for adaptation to different industries, jurisdictions, or regulatory contexts.

While the model is already valuable in its current state, future work—including integration of textual disclosures, time-series behavior, and enhanced explainability—will push it closer to a real-time, enterprise-ready fraud intelligence platform.

In an era of increasing regulatory scrutiny and data availability, this approach positions compliance teams, auditors, and financial executives to move from reactive investigation to proactive detection—safeguarding shareholder trust and reinforcing market integrity.

# Future scope of work

**Data Limitations & Deficiencies**

While the model delivered meaningful predictive performance, several limitations in the underlying data warrant caution and highlight areas for refinement.

First, the dataset reflects a single fiscal year snapshot (2023), limiting visibility into temporal patterns or year-over-year manipulations that often characterize fraudulent reporting. Without multi-period data, the model cannot detect trends such as earnings smoothing or gradual asset overstatement.

Second, the target variable—fraudulent misstatement—is binary and derived from labeled cases. These labels are valuable but may underrepresent the true scope of fraud, especially when undetected or unreported misconduct exists. This introduces a risk of **label bias**, where only known fraud instances shape the model.

Third, while financial statement data offers a robust foundation, it lacks contextual variables such as executive incentives, auditor reputation, or restatement history, all of which can affect fraud likelihood.

Finally, feature correlation and multicollinearity among accounting metrics may obscure individual variable contributions unless explicitly addressed through dimensionality reduction or regularization -- approaches reserved for future iterations.

These limitations, while not fatal, underscore the importance of interpreting model results as directional risk signals rather than definitive fraud judgments.

**Filling the Gaps: Key Data to Strengthen the Model**

To improve accuracy, interpretability, and generalizability, several types of financial and contextual data could be incorporated into future versions of the model:

* **Multi-Year Financial Statements**  
  Access to historical filings would enable trend-based fraud detection. Modeling deltas and rolling averages across time can highlight manipulation strategies that occur gradually rather than in isolation.
* **Restatement History and Audit Flags**  
  Including whether a company has restated financials or received a qualified audit opinion would enrich the signal strength around risky behaviors.
* **Executive Compensation and Insider Trading Data**  
  Linking financial incentives to fraud risk by incorporating stock-based compensation structures or unusual insider transactions could uncover motive-based dimensions.
* **Governance and Audit Committee Attributes**  
  Board independence, audit committee experience, or auditor tenure are often cited in fraud literature as structural risk factors. Adding these elements could improve model granularity.
* **Textual Disclosures**  
  Management’s Discussion and Analysis (MD&A), risk factors, and footnote narratives may contain subtle linguistic cues—an opportunity for NLP-based augmentation.

By expanding the data landscape, future iterations of the model can more holistically represent fraud dynamics and move closer to replicating the decision-making frameworks used by forensic auditors.

**Unexpected Insights for Financial Leaders**

While the model's primary goal was fraud detection, several insights emerged that may challenge traditional assumptions held by executives and financial leadership:

* **Fraud Risk Is Highly Sensitive to Subtle Shifts**  
  Perturbation analysis revealed that even minor changes (as small as ±5%) in key features—such as retained earnings or the operating cash flow ratio—can meaningfully alter fraud probability. This suggests that potential red flags may emerge from ordinary fluctuations rather than large, obvious anomalies.
* **Liquidity Metrics Signal More Than Solvency**  
  Features like the operating cash flow ratio and accounts payable were more predictive of fraud than some profitability ratios. This implies that fraud risk may be more tightly linked to cash flow pressure and short-term liabilities than to net income figures.
* **Threshold Tuning Matters**  
  The model's sensitivity to fraud improved significantly when prediction thresholds were calibrated, highlighting the risk of relying on default 0.5 cutoffs. This has implications for internal risk models that often treat outputs in binary terms.
* **CatBoost Outperformed Deep Learning—Until Blended**  
  While DNNs alone struggled with class imbalance, blending with CatBoost provided superior performance. This reinforces the value of hybrid approaches that balance interpretability with complex pattern recognition.

These insights challenge conventional thinking that large, sudden changes or bottom-line profit shifts are the only meaningful indicators of fraud. Instead, they point to a broader and more nuanced risk profile that can be operationalized.

**Practical Use Cases for Enterprises**

The fraud detection model is designed not just as a proof of concept but as a practical tool for real-world enterprise adoption. Organizations can apply it in several impactful ways:

* **Prioritized Risk Reviews**  
  Integrate model predictions into quarterly close or audit cycles to automatically flag filings or subsidiaries with elevated risk scores for additional scrutiny.
* **Internal Control Optimization**  
  Use feature importance insights to refine internal control testing. For example, if changes in working capital are highly predictive, controls related to accrual adjustments may warrant closer attention.
* **M&A Due Diligence**  
  Apply the model to financials from acquisition targets to identify hidden risk signals prior to investment decisions.
* **Training and Scenario Analysis**  
  Use the model in training programs for audit staff and compliance teams to demonstrate how subtle accounting shifts impact fraud probability, enabling more informed investigative instincts.
* **Dashboard Integration for Continuous Monitoring**  
  Connect the model to existing enterprise data pipelines or business intelligence tools to automate monitoring and reporting over time.

These use cases help bridge the gap between theoretical modeling and enterprise risk management, allowing businesses to detect, investigate, and ultimately prevent accounting fraud with greater precision.

**Enabling Analyst-Driven Exploration**

To unlock the full value of the model across teams, it must be made accessible to business analysts, auditors, and compliance professionals who may not have deep machine learning expertise. This calls for thoughtful packaging and deployment:

* **Interactive Dashboards**  
  A dashboard built using tools like Streamlit, Power BI, or Tableau could allow users to explore fraud risk predictions interactively—adjusting inputs or thresholds to simulate various scenarios.
* **Scheduled Reporting Pipelines**  
  Automating monthly or quarterly fraud reports, including flagged entities, evolving risk trends, and top contributing features, ensures actionable insights are routinely surfaced to decision-makers.
* **Lightweight APIs**  
  Exposing the model via a simple API would allow internal systems or analyst tools to submit financial data and receive fraud risk scores in return—enabling scalable, self-service access across departments.
* **Model Documentation and Interpretability Tools**  
  Including feature importance charts, threshold tuning guides, and explanations of model logic (e.g., SHAP plots) can help demystify the underlying algorithms and build organizational trust.

Making the model accessible transforms it from a data science artifact into a repeatable operational asset—empowering teams to ask better questions, act faster, and uncover fraud risks before they escalate.

**🎯 Conclusion: Next Steps for Business Adoption**

This project establishes a strong foundation for the practical application of machine learning in financial fraud detection. The hybrid CatBoost–DNN model demonstrates that, even with imbalanced data, it is possible to surface early signals of accounting misstatements through structured financial indicators alone.

1. To translate this work into business value, the following steps are recommended:
2. Pilot Integration: Begin with a limited deployment in internal audit or compliance teams to validate fraud flags against past incidents and restatement histories.
3. Stakeholder Alignment: Present findings to finance leadership, audit committees, and risk officers to build consensus around model adoption, refinement, and ethical use.
4. System Integration: Connect the model to data pipelines from ERP or financial consolidation tools, enabling real-time or batch scoring during reporting periods.
5. Feedback Loop Development: Collect analyst feedback and investigation outcomes to retrain and improve the model—further reducing false positives and increasing interpretability.
6. Scalability Planning: Determine the feasibility of expanding the model across business units, geographies, or even portfolio companies in investment firms.

By embedding this model within audit and risk frameworks, businesses can shift from post-incident reviews to proactive fraud mitigation—creating a smarter, more resilient financial reporting environment.

**Data Source and Reference**:

The primary dataset was sourced from the [FraudDetection GitHub repository](https://github.com/JarFraud/FraudDetection), which accompanies a peer-reviewed paper published in the *Journal of Accounting Research*.

* **Dataset URL**: [data\_FraudDetection\_JAR2020.csv](https://raw.githubusercontent.com/JarFraud/FraudDetection/refs/heads/master/data_FraudDetection_JAR2020.csv)
* **Codebook**: [docs/codebook.md](https://github.com/BTExpress1/accounting-fraud-detection/blob/main/docs/codebook.md)

Yang Bao, Bin Ke, Bin Li, Julia Yu, and Jie Zhang (2020). *Detecting Accounting Fraud in Publicly Traded U.S. Firms Using a Machine Learning Approach*. Journal of Accounting Research, 58(1): 199–235. [Read the paper](https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12292)