Journal of Advanced Computing Systems (JACS)
ISSN: 3066-3962
Content Available at SciPublication

Going Concern and Bankruptcy Prediction under Extreme Class Imbalance: Cost-Sensitive Learning, Resampling, and Focal Loss with Explainable Financial-Ratio Portraits

Yuanzheng Chen[1], Yitian Zhang[2], Matt Sherman[3]
[1]*Accounting, UIUC, IL, USA*
[2]*Accounting, The University of Wisconsin-Madison (UW-Madison), WI, USA*
[3]*Computer Engineering, Dartmouth College, NH, USA*
chenyuanzheng0920@gmail.com

DOI: 10.69987/JACS.2024.40407

Keywords

going concern; bankruptcy prediction; imbalanced learning; cost-sensitive learning; focal loss; audit sampling; explainable financial ratios

Abstract

Auditors and creditors face a going-concern screening problem: only a small fraction of firms fail, yet failing to identify a distressed firm is substantially costlier than issuing a false alarm. This paper formulates bankruptcy/going-concern prediction as an extreme class-imbalance ranking task and reports a fully reproducible empirical evaluation on the Polish Companies Bankruptcy benchmark distributed by the UCI Machine Learning Repository (five forecasting horizons; 64 financial ratios; 43,405 firm-year records). We compare three imbalance-aware strategy families: (i) cost-sensitive learning via class-weighted objectives, (ii) resampling via random over-sampling, random under-sampling, and SMOTE, and (iii) focal loss to emphasize hard minority examples. Performance is assessed using the area under the precision–recall curve (AUPRC) and Recall@Top-k, where k represents the fraction of firms audited under a constrained sampling budget. Across all five horizons, tree ensembles dominated linear models, and exploiting missingness patterns was critical: augmenting a cost-sensitive random forest with missing-value indicator features increased AUPRC from 0.480 to 0.640 in the most imbalanced 1stYear case. Overall, the cost-sensitive random forest with missing indicators achieved AUPRC values of 0.640, 0.480, 0.459, 0.567, and 0.601 for the 1stYear to 5thYear cases, respectively, under a stratified 70/30 split (seed=42). The audit-oriented metric showed large operational gains: in 1stYear (3.86% bankrupt), auditing only the top 5% of firms ranked by the best model recovered 62.96% of bankruptcies (Recall@Top-5%), compared with 12.35% for standard logistic regression. Finally, we provide an explainable financial-ratio portrait that summarizes the characteristic liquidity, leverage, and profitability patterns of the model-flagged high-risk cohort, bridging predictive ranking with actionable accounting evidence for going-concern planning.

Introduction

Going-concern assessment is a core decision in auditing and credit monitoring. The going-concern principle assumes that an entity will continue operating for the foreseeable future; when substantial doubt exists, auditors must consider whether to modify the audit report and whether to increase procedures focused on liquidity, debt compliance, and viability. Empirical auditing research has shown that the going-concern opinion is strongly linked to financial distress signals and can precede bankruptcy, but the decision is difficult because failure events are relatively rare and because the consequences of a false negative (missing an impending failure) often exceed those of a false positive (investigating a firm that ultimately survives). Early statistical studies modeled the going-concern reporting decision using financial ratios and identified profitability, liquidity, and leverage as key determinants [15]. Subsequent work examined how audit report lag and other practical constraints affect reporting for

bankrupt companies [16]. These findings motivate decision-support tools that can prioritize high-risk engagements under limited review capacity [24-29].

Bankruptcy prediction has a long tradition in accounting and finance. Classic score-based models used a small set of ratios to construct interpretable discriminants. Altman's Z-score [1] demonstrated that a linear combination of ratios can separate bankrupt and non-bankrupt firms, while Ohlson's logit model [2] provided a probabilistic framework that improved flexibility and inference. These models remain influential because they align naturally with accounting reasoning: deteriorating profitability, weak liquidity, and high leverage jointly signal distress. However, modern business environments and data availability motivate more flexible learners that can capture non-linear interactions and handle noisy, high-dimensional features [30-37].

In the machine learning era, bankruptcy and distress prediction has been approached with decision trees, support vector machines, neural networks, and ensemble methods, often with improved predictive accuracy. In particular, tree ensembles are effective on tabular financial data because they can represent non-linear threshold effects and interactions among ratios. The Polish Companies Bankruptcy benchmark used in this paper was originally analyzed with ensemble boosted trees and synthetic feature generation, showing that engineered interactions can materially improve prediction quality [3]. The dataset is publicly available via the UCI Machine Learning Repository [4] and has become a standard benchmark for financial distress modeling under class imbalance [38-44].

A defining characteristic of going-concern prediction is extreme class imbalance. In many real-world portfolios, only a small minority of firms fail, and the rare-event nature of bankruptcy introduces both statistical and operational challenges. Statistically, standard classifiers optimized for overall accuracy are biased toward the majority class, and probability estimates can be poorly calibrated in rare-event regimes. Operationally, the output is often used for triage: auditors, risk managers, or regulators can only investigate a subset of firms, so the relevant objective is to rank the most suspicious cases at the top of a list. These realities make it essential to study algorithms and evaluation metrics that explicitly address imbalance and asymmetric costs [45-57].

Cost-sensitive learning provides a principled approach when misclassification costs are asymmetric. Rather than treating all errors equally, cost-sensitive methods weight mistakes according to their consequences and can be understood as learning with non-uniform loss functions [5]. Practical implementations include class-weighted training, cost-proportionate example weighting, and wrapper methods such as MetaCost [17]. Related work showed that cost-proportionate weighting can make standard learners effectively cost-sensitive while preserving probability estimation [18]. In rare-event settings, logistic regression can also suffer from small-sample bias and separation, which motivates careful evaluation and, in some contexts, specialized rare-events corrections [19]. In auditing, the economic cost of overlooking a distressed firm supports the use of cost-sensitive objectives and budgeted recall as the primary success criteria.

Resampling is another widely used family of imbalance-handling techniques. Random over-sampling (ROS) increases the minority prevalence by duplicating minority observations, while random under-sampling (RUS) reduces the majority class to balance the training distribution. Synthetic over-sampling techniques such as SMOTE interpolate between minority neighbors to generate new synthetic minority instances and can reduce overfitting relative to naive duplication [6]. Surveys on imbalanced learning emphasize that the choice between cost-sensitive learning and resampling depends on classifier bias, noise, and data geometry, and that evaluation must reflect the minority objective [7], [8]. In bankruptcy prediction, resampling is particularly appealing because it can be applied as a preprocessing step and combined with interpretable models such as logistic regression.

More recently, focal loss introduced a loss-shaping approach designed to focus learning on hard examples by down-weighting well-classified negatives [11]. Although focal loss was proposed for dense object detection, the same principle applies to rare-event tabular learning: when negatives dominate, a learner can spend most of its gradient budget on easy negatives unless the objective rebalances attention. For going-concern prediction, focal loss is attractive because it can be used with differentiable models and can be combined with cost-sensitive weighting through its α parameter. However, focal loss also introduces additional hyperparameters and interacts with model capacity, which motivates empirical evaluation on financial data rather than assuming universal benefit.

A critical but sometimes overlooked component is evaluation. Because the positive class is rare, ROC-based metrics can paint an overly optimistic picture of performance by emphasizing true negative rates. Precision–recall (PR) analysis provides a more informative view of minority retrieval because precision directly captures the burden of false positives among the flagged cases. Prior work formalized the relationship between ROC and PR curves [9] and demonstrated that PR curves are more informative under skewed class priors [10]. For audit triage, PR curves and AUPRC capture overall ranking quality, but they still average over budgets that may be unrealistic in practice.

Auditors and risk managers often operate under explicit capacity constraints. Let k denote the fraction of firms

that can be reviewed in depth during a period (e.g., top 1% or top 5% of engagements). A screening model assigns a risk score to each firm; the organization inspects the top-k subset and aims to capture as many true bankruptcies as possible. This setting naturally motivates Recall@Top-k: the proportion of all bankruptcies contained in the inspected subset. Unlike threshold-based recall, Recall@Top-k is directly actionable because k corresponds to resource allocation. It also provides an interpretable baseline: under random ranking, expected recall equals k, so any improvement over this baseline reflects sample-efficiency gains in locating distress.

Interpretability is the second practical requirement. Even when a model delivers strong ranking performance, auditors must justify their judgments with evidence grounded in financial statements and must be able to explain why a firm was flagged. Post-hoc explanation methods such as LIME [13] and SHAP [12] provide local attributions that can support case-level documentation, while feature selection perspectives highlight the importance of identifying stable, meaningful predictors [20]. Interpretable machine learning frameworks further emphasize that explanations should be faithful to the model and usable by stakeholders [21]. In this work, we complement predictive comparisons with an explainable financial-ratio portrait that summarizes the characteristic ratio pattern of the model-flagged high-risk cohort.

This paper provides a reproducible, audit-oriented empirical study of imbalance-aware bankruptcy prediction. We evaluate cost-sensitive learning, three resampling variants (ROS, RUS, SMOTE), and focal loss, using both logistic regression and random forests as base learners. We further study the role of missingness, showing that missing-value patterns themselves carry predictive signal in this benchmark and that adding missingness indicator features can yield substantial gains. Our evaluation reports AUPRC and Recall@Top-k for k in {0.5%, 1%, 2%, 5%, 10%, 20%} and visualizes "Top-k hit rate versus audit sampling cost" curves.

Our contributions are fourfold. First, we provide a consistent experimental protocol across all five forecasting horizons of the Polish benchmark, including explicit handling of missing data and fixed random seeds for reproducibility. Second, we report detailed metric tables and operational audit cost curves that quantify how many bankruptcies are found under realistic review budgets. Third, we clarify the relative effect of cost-sensitive learning, resampling, and focal loss within both linear and non-linear learner families. Fourth, we present an explainable ratio portrait and feature importance analysis that connects model outputs to accounting signals, supporting practical going-concern planning rather than black-box alerting.

Method

A. Dataset and prediction setup. We used the Polish Companies Bankruptcy dataset provided as five ARFF files (1year–5year) [4]. Each file defines an independent supervised learning problem corresponding to a different forecasting horizon. The input X consists of 64 accounting ratios (Attr1–Attr64) derived from financial statements. The target y is binary, where y=1 indicates bankruptcy and y=0 indicates continued operation. The ratio definitions are specified by the dataset documentation (Table 3 provides a subset used in our explainability analysis). Because the five cases differ in horizon and sample size, we treat them as five separate datasets and report results for each case rather than pooling.

B. Data quality and missingness. Financial ratio datasets often contain missing values due to incomplete reporting, division by zero, or differences in accounting practices. In this benchmark, missingness is substantial for several ratios (Table 2). Importantly, missingness can be informative: a missing value may indicate that a quantity was not reported or not applicable, which can correlate with firm distress or disclosure quality. To address this, our primary preprocessing uses median imputation. In addition, we evaluate a missingness-aware variant of the best-performing model by augmenting the feature space with binary missing-value indicators (denoted "+MI"). These indicators allow models to exploit missingness patterns without leaking information from the label, because the indicator is computed from the observed data only.

C. Preprocessing pipeline. All methods used per-feature median imputation fitted on the training split only. Linear models (logistic regression and focal-loss linear model) further used standardization (zero mean, unit variance) fitted on the training split, because ratio magnitudes and variances differ. Tree models did not require scaling. For methods involving resampling, we applied imputation and scaling before resampling to ensure that SMOTE's neighbor computations were well-conditioned. When using missing indicators (+MI), we applied SimpleImputer(add_indicator=True), which appends a binary feature for each ratio that contains missing values in the training data.

D. Experimental protocol and reproducibility. For each forecasting case, we performed a stratified random split into 70% training data and 30% held-out test data using random state=42. All compared methods within a case used the exact same split to ensure paired comparability. We report all metrics on the held-out test set. This protocol matches a standard risk-modeling workflow where historical firm-year data is used to train a screening model and the model is then applied to new engagements. All experiments were implemented in Python using scikit-learn 1.4.2, imbalanced-learn

0.12.4, numpy 1.24.0, pandas 2.2.3, and PyTorch 2.5.1 (CPU).

E. Base learners. We compared two base learner families.

1) Logistic regression (LR): LR provides an interpretable linear baseline. We used LogisticRegression with the liblinear solver and max iter=2000. The output score is the predicted probability of bankruptcy.

2) Random forest (RF): RF is a non-linear ensemble of decision trees that captures interactions and non-linear thresholds in ratios [14]. We used 200 trees with default splitting criteria and feature subsampling. RF outputs a probability estimate based on the fraction of trees voting for bankruptcy. We also evaluated a cost-sensitive RF by applying class weights during tree fitting.

F. Imbalance-handling strategies. We implemented three strategy families and applied them consistently across horizons (hyperparameters summarized in Table 4).

1) Cost-sensitive learning:

LR used class weight="balanced", reweighting the log-loss inversely proportional to class frequency. RF used class_weight="balanced_subsample" so that each bootstrap sample was reweighted.

2) Resampling: We trained LR under ROS, RUS, and SMOTE. ROS duplicates minority examples to balance the class distribution. RUS removes majority examples. SMOTE synthesizes minority points by interpolating between k nearest minority neighbors (k=5) [6].

3) Focal loss: We trained a linear classifier using focal loss [11] with α=0.25 and γ=2.0. Optimization used full-batch Adam (lr=0.01, weight decay=1e-4) for 300 epochs. Focal loss reduces the relative gradient contribution of easy negatives and thus can be viewed as an objective-level analogue of resampling.

G. Evaluation metrics and audit formulation. We view the output of each model as a risk ranking over firms.

1) Precision–recall analysis and AUPRC: For a threshold τ, precision is TP/(TP+FP) and recall is TP/(TP+FN). We computed the precision–recall curve by sweeping τ over all unique score values and measured area under the curve as average precision (AP). AUPRC is recommended for imbalanced learning because it is sensitive to minority retrieval performance [9], [10].

2) Recall@Top-k and Precision@Top-k: Let n be the number of test firms, and let $k \in (0,1]$ be an audit budget. We sort firms by predicted risk score in descending order and select the top $\lceil kn \rceil$ firms. Recall@Top-k is the fraction of all bankrupt firms in the test set that fall inside this audited subset; Precision@Top-k is the bankruptcy rate within the audited subset.

3) Audit cost curve: We interpret k as audit sampling cost (proportional to review workload) and plot Recall@Top-k versus k. Under random ranking, expected Recall@Top-k equals k, providing an intuitive baseline for sample efficiency.

H. Explainability: feature importance and ratio portraits. We report two complementary interpretability views.

1) Global importance: For RF models, we extracted mean decrease in impurity (Gini importance) to identify influential ratios. For the +MI model, we separately examined importances of ratio features and missingness indicators to quantify whether missingness patterns were predictive.

2) Financial-ratio portrait: We constructed a cohort-level portrait for the high-risk cohort ranked by the best model. We selected a small set of influential ratios spanning leverage, liquidity, and profitability and robustly scaled each ratio using the 5th and 95th percentiles of the training data (clipped to [0,1]). We then computed median scaled values for the top 5% highest-risk firms and for the remaining firms. The resulting radar chart summarizes the characteristic accounting profile of model-flagged high risk firms, supporting going-concern audit planning.

I. Implementation details and computational considerations. To keep comparisons stable, all stochastic learners used fixed random_state=42. For the PyTorch focal-loss model, we set the number of CPU threads to 1 to ensure deterministic execution and to avoid runtime variability on shared environments. Random forests were trained with n_jobs=1 for the same reason. While these settings are conservative, they preserve comparability and make the full experiment reproducible with limited resources.

J. Mathematical details of class weighting and resampling. In scikit-learn, class weight="balanced" sets a per-class weight $w_c = n / (|C| \cdot n_c)$, where n is the number of training instances, |C| is the number of classes (2 here), and n_c is the number of instances of class c. Intuitively, the minority class receives larger weight so that a mistake on a bankrupt firm contributes more to the loss. For SMOTE [6], a synthetic minority point is generated as $x_{new} = x_i + \lambda(x_{nn} - x_i)$, where x_i is a minority example, x_{nn} is one of its k nearest minority neighbors, and $\lambda \in [0,1]$ is sampled uniformly. This procedure expands the minority region and can help linear models fit a decision boundary that better covers minority examples.

K. Relationship between audit budget, precision, and recall. Recall@Top-k and Precision@Top-k are linked through the class prevalence π and the audited subset size kn. If the audited subset has precision P k, then the number of bankruptcies found is approximately P_k·kn, and recall is (P k·kn)/(πn)=P k·k/π. Therefore, for fixed prevalence π, high precision at small k is a strong indicator of audit efficiency. This relationship motivates reporting both recall and precision (or lift) for top-k screening.

Table 1. Dataset statistics and imbalance levels for the five forecasting cases.

Dataset	**Instances**	**Bankrupt (1)**	**Non-bankrupt (0)**	**Positive rate (%)**	**Imbalance ratio (0/1)**	**Missing entries**	**Missing %**
1stYear	7027	271	6756	3.86	24.93	5835	1.30
2ndYear	10173	400	9773	3.93	24.43	12157	1.87
3rdYear	10503	495	10008	4.71	20.22	9888	1.47
4thYear	9792	515	9277	5.26	18.01	8776	1.40
5thYear	5910	410	5500	6.94	13.41	4666	1.23

Table 2. Top 10 ratio variables with the most missing values (aggregated across all 43,405 firm-year records).

Feature	Ratio	Missing count	Missing %
Attr37	(current assets - inventories) / long-term liabilities	18984	43.7
Attr21	sales (n) / sales (n-1)	5854	13.5
Attr27	profit on operating activities / financial expenses	2764	6.4
Attr60	sales / inventory	2152	5.0
Attr45	net profit / inventory	2147	4.9
Attr24	gross profit (in 3 years) / total assets	922	2.1
Attr28	working capital / fixed assets	812	1.9
Attr64	sales / fixed assets	812	1.9
Attr53	equity / fixed assets	812	1.9
Attr54	constant capital / fixed assets	812	1.9

Table 3. Accounting meaning of key ratios used in analysis and explainability (definitions from dataset documentation [4]).

Attribute	**Ratio definition (UCI)**	**Category**
Attr1	net profit / total assets	Other

Attr2	total liabilities / total assets	Leverage/Solvency
Attr3	working capital / total assets	Other
Attr4	current assets / short-term liabilities	Other
Attr7	EBIT / total assets	Other
Attr9	sales / total assets	Other
Attr10	equity / total assets	Other
Attr12	gross profit / short-term liabilities	Liquidity/Profitability
Attr13	(gross profit + depreciation) / sales	Profitability (margin)
Attr16	(gross profit + depreciation) / total liabilities	Solvency (coverage)
Attr24	gross profit (in 3 years) / total assets	Profitability (long-term)
Attr26	(net profit + depreciation) / total liabilities	Cash flow / Leverage
Attr27	profit on operating activities / financial expenses	Coverage (operating profit / financial expenses)
Attr29	logarithm of total assets	Size
Attr34	operating expenses / total liabilities	Expense structure / Leverage
Attr46	(current assets - inventory) / short-term liabilities	Liquidity (quick ratio variant)

Table 4. Model configurations and imbalance-handling strategies used in experiments.

Method	**Imbalance handling**	**Classifier**	**Key hyperparameters**
LR	None	LogisticRegression	solver=liblinear, C=1.0, max_iter=2000
CostSensitive-LR	class_weight=balanced	LogisticRegression	solver=liblinear, C=1.0, max_iter=2000
ROS-LR	RandomOverSampler	LogisticRegression	ROS sampling_strategy=auto , random_state=42
SMOTE-LR	SMOTE	LogisticRegression	k_neighbors=5, sampling_strategy=auto , random_state=42
RUS-LR	RandomUnderSampler	LogisticRegression	RUS sampling_strategy=auto , random_state=42

RF	None	RandomForest	n_estimators=200, n_jobs=1, other params default
CostSensitive-RF	class_weight=balanced_subsample	RandomForest	n_estimators=200, n_jobs=1, other params default
CostSensitive-RF+MI	class_weight=balanced_subsample + missing indicators	RandomForest	n_estimators=200, SimpleImputer(add_indicator=True), n_jobs=1
Focal-LR	Focal loss	Linear (PyTorch)	alpha=0.25, gamma=2.0, epochs=300, lr=0.01, Adam, weight_decay=1e-4

Figure 1. End-to-end workflow: preprocessing (including missing indicators), imbalance handling, model training, audit-oriented evaluation, and explainable ratio portrait.

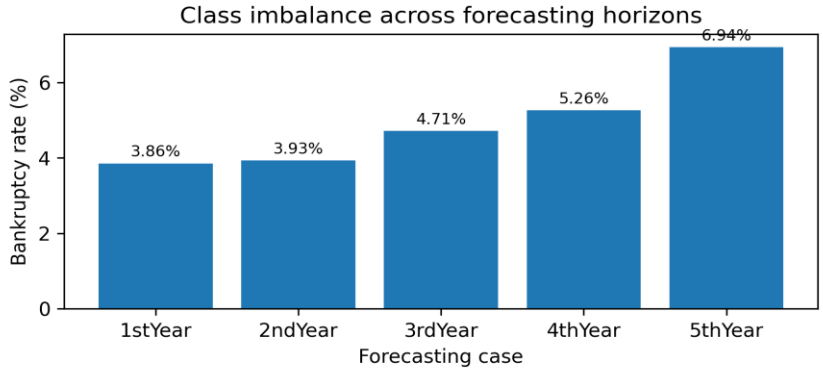

Figure 2. Bankruptcy prevalence across the five forecasting horizons (bankrupt class proportion).

Results and Discussion

A. Data characteristics: imbalance and missingness. Table 1 and Figure 2 quantify the class imbalance across forecasting horizons. The bankruptcy rate is lowest in the longest-horizon 1stYear case (3.86%) and rises to 6.94% in the shortest-horizon 5thYear case. This pattern is consistent with financial distress processes: ratio signals often intensify as bankruptcy approaches. Missing values are also substantial. Table 2 shows that some ratios are missing for a large portion of firms, with Attr37 missing in 43.7% of all rows. Importantly, missingness is not purely nuisance noise. In accounting data, missing ratios can arise from absent line items, division by zero, or reporting irregularities; these mechanisms can correlate with distress and thus constitute predictive signal. This observation motivates our comparison between models that only impute

missing values and models that also include missing-value indicator features (+MI).

B. Global ranking quality (AUPRC). Table 5 reports AUPRC results for all methods across the five horizons. Several conclusions are clear. First, random forests consistently dominate logistic regression in AUPRC, indicating substantial non-linearity and interaction effects among ratios. For example, in 1stYear standard LR achieved AUPRC 0.089, while an unweighted RF achieved 0.463. Second, classic imbalance strategies improve linear models but do not close the gap to non-linear learners. In 1stYear, resampling (ROS and SMOTE) increased LR's AUPRC from 0.089 to 0.109, and cost-sensitive weighting increased it to 0.103. Similar modest gains appear in 3rdYear and 4thYear. Third, cost-sensitive learning interacts with horizon. In 5thYear, unweighted RF achieved the best AUPRC among non-MI models (0.584), while cost-sensitive RF decreased to 0.501. This suggests that when distress signals are strong (short horizon), aggressive reweighting may amplify minority noise and slightly reduce global ranking quality. Fourth, missingness-aware preprocessing produced the largest single improvement in this benchmark. Adding missing-value indicators to the cost-sensitive RF improved AUPRC from 0.480 to 0.640 in 1stYear, from 0.391 to 0.567 in 4thYear, and improved AUPRC in every horizon. This indicates that missingness patterns are informative for bankruptcy risk in this dataset and that purely imputing ratios can discard useful signal.

Figure 3 visualizes representative precision–recall curves for the 1stYear case. The cost-sensitive RF+MI dominates the curve, maintaining high precision at low recall and extending to higher recall levels without severe precision collapse. SMOTE-LR improves over standard LR in the medium recall range, consistent with its AUPRC gain, but remains far below tree ensembles. Because audit triage typically operates at low to moderate recall (reflecting budget constraints), the left portion of Figure 3 is operationally most relevant, and the advantage of RF-based rankers is practically large.

Focal loss (Focal-LR) did not improve AUPRC in our linear configuration. In 1stYear, focal loss achieved AUPRC 0.094 versus 0.089 for standard LR and 0.109 for SMOTE-LR. This suggests that, at least for a linear function class and fixed $(\alpha,\gamma)=(0.25,2.0)$, focal loss behaves similarly to mild reweighting and cannot compensate for missing non-linear structure. In principle, focal loss could be more effective when paired with richer models or when α and γ are tuned per horizon.

C. Budgeted audit performance (Recall@Top-k). The audit perspective can differ from global AUPRC. Tables 6–8 report Recall@Top-1%, Recall@Top-5%, and Recall@Top-10%. These metrics quantify how many bankruptcies an auditor would find if only the top k% of firms were reviewed.

In 1stYear, standard LR recovered 3.7% of bankruptcies at k=1% and 12.3% at k=5%. The best model, cost-sensitive RF+MI, recovered 24.7% at k=1% and 63.0% at k=5%. In other words, under a 5% review budget the best model found about five times as many bankruptcies as the linear baseline. This is not a marginal improvement: it changes the practical feasibility of early-horizon screening.

In 2ndYear–4thYear, the same pattern holds. For example, in 4thYear at k=5%, LR achieved 12.1% recall while RF+MI achieved 60.2%. These consistent gains show that non-linear rankers and missingness-aware features substantially increase sample efficiency in locating rare distressed firms.

D. Audit cost curves and cost efficiency. Figure 4 plots Recall@Top-k versus audit sampling cost k for the 1stYear case. The gap between models is largest at small budgets. Table 9 lists the underlying values: at k=0.5% LR recovered 2.5% of bankruptcies, while cost-sensitive RF recovered 11.1% and RF+MI recovered 12.3%. At k=2%, LR recovered 9.9% while RF+MI recovered 40.7%. These differences can be interpreted as audit cost savings. Using a finer grid of k values, we found that reaching 50% bankruptcy coverage in 1stYear required auditing approximately 2.5% of firms with RF+MI and about 4.0% with cost-sensitive RF, while LR never reached 50% recall even at a 20% review budget. Thus, the best model reduces the review workload required to reach a target coverage level and can enable audit teams to concentrate substantive procedures on a smaller set of engagements.

Precision@Top-k (not tabulated for all k due to space) further clarifies operational yield. In 1stYear, the RF+MI model achieved Precision@Top-1% of 0.909, meaning roughly 91% of firms in the top 1% risk list were truly bankrupt in the held-out test split. Even accounting for test-set uncertainty, such yield is substantially higher than the base bankruptcy rate (3.86%) and indicates that a small audit budget can be converted into a high-density list of truly distressed firms.

E. Explainability: which ratios and which missingness patterns matter? Table 10 and Figure 5 report global ratio importances for the best model (cost-sensitive RF+MI), restricted to the 64 ratio features to preserve accounting interpretability. The most influential ratios are economically plausible and align with classic distress theory: liquidity ratios such as Attr46 and Attr12, leverage and cash-flow coverage such as Attr26 and Attr16, and size/profitability measures such as Attr29 and Attr24. Median differences in Table 10 indicate that bankrupt firms have weaker liquidity and coverage ratios (negative deltas for Attr46,

Attr12, Attr16, Attr26) and tend to be smaller (Attr29), consistent with the intuition that smaller firms with tight liquidity buffers and weak operating coverage are more vulnerable.

A striking result is the predictive role of missingness. Table 11 lists the most important missingness indicator features for the RF+MI model. The top indicator is Missing(Attr27), corresponding to the ratio "profit on operating activities / financial expenses." Its importance suggests that whether this ratio is missing (e.g., due to unavailable or zero financial expense entries) is itself predictive of distress. Other important indicators include Missing(Attr45) (net profit / inventory) and Missing(Attr60) (sales / inventory), both related to inventory-based efficiency. In a going-concern context, these patterns can reflect reporting structure, line-item absence, or abnormal accounting values that lead to undefined ratios. The finding reinforces that data quality and accounting disclosure characteristics can carry risk information beyond the ratio magnitudes.

F. Explainable financial-ratio portrait for high-risk cohorts. Figure 6 provides a cohort-level ratio portrait for the high-risk group (top 5% by RF+MI score) versus the remainder of the 1stYear test set. The portrait aggregates eight ratios spanning leverage (Attr2), size (Attr29), long-term profitability (Attr24), and multiple liquidity/coverage measures (Attr12, Attr16, Attr46, Attr26) plus an expense-structure proxy (Attr34). The high-risk cohort shows substantially higher leverage (higher Attr2) and materially lower liquidity and coverage (lower Attr12, Attr16, Attr46, Attr26). This portrait is actionable for audit planning: it suggests that flagged engagements warrant procedures focused on short-term liquidity stress, debt service capacity, and sustained operating profitability, and it provides a concise narrative that is consistent with the financial logic of classic distress models [1], [2].

Overall, the results highlight three lessons for going-concern triage. First, evaluation must be audit-aligned: Recall@Top-k and audit cost curves reveal large operational differences that can be obscured by single global metrics. Second, non-linear models can deliver major gains in early-horizon screening where ratio signals are subtle. Third, missingness should be treated as potential signal rather than only noise; missing-value indicators can meaningfully improve bankruptcy ranking and can themselves be interpreted as disclosure or accounting-structure cues.

G. Audit yield and lift analysis. While recall measures bankruptcy coverage, audit teams also care about yield: among reviewed firms, how many are truly distressed? One way to summarize yield relative to the portfolio baseline is lift, defined as Precision@Top-k divided by the base bankruptcy rate. Lift captures how strongly a model concentrates bankruptcies near the top of the ranking. Table 12 reports Precision@Top-1% and Lift@1% for LR and RF+MI across horizons. Two patterns stand out. First, LR provides only modest lift at the strictest budget (e.g., 1.6× in 2ndYear), indicating that its top-1% list is only slightly more risky than the overall portfolio. Second, RF+MI produces very large lift values (often above 20×), meaning that the top-1% list is highly enriched with bankruptcies. In operational terms, high lift implies that a small review team can focus on a tiny subset of firms while still encountering a large fraction of all distressed cases.

Lift also helps interpret why Recall@Top-k differs so much across methods. In 1stYear, the base bankruptcy rate is 3.86%. At k=1%, even a perfect oracle could not exceed 100% precision, so the maximum possible lift is about 25.9×. The RF+MI model achieves 23.6× lift at top-1%, approaching this practical ceiling, which explains its strong Recall@Top-1% (24.7%). In contrast, LR achieves 5.9× lift at top-1%, and thus it cannot cover many bankruptcies without expanding the audit budget. These observations reinforce that audit triage models should be evaluated at the specific k values relevant to organizational capacity, rather than only by global summary metrics.

Table 5. AUPRC (average precision) on the held-out test split (seed=42). Higher is better.

Case	LR	CostSensitive-LR	ROS-LR	SMOTE-LR	RUS-LR	Focal-LR	RF	CostSensitive-RF	CostSensitive-RF+MI
1stYear	0.089	0.091	0.109	0.109	0.088	0.087	0.463	0.48	0.64
2ndYear	0.062	0.07	0.071	0.074	0.061	0.051	0.371	0.384	0.48
3rdYear	0.118	0.122	0.12	0.125	0.112	0.111	0.358	0.391	0.459
4thYear	0.11	0.118	0.121	0.121	0.118	0.108	0.375	0.391	0.567
5thYear	0.335	0.349	0.346	0.336	0.282	0.323	0.584	0.501	0.601

Table 6. Recall@Top-1% (fraction of bankruptcies found when auditing the top 1% highest-risk firms).

Case	LR	CostSensitive-LR	ROS-LR	SMOTE-LR	RUS-LR	Focal-LR	RF	CostSensitive-RF	CostSensitive-RF+MI
1stYear	0.062	0.074	0.074	0.074	0.062	0.062	0.185	0.21	0.247
2ndYear	0.017	0.008	0.025	0.008	0.008	0.008	0.192	0.192	0.217
3rdYear	0.034	0.04	0.04	0.04	0.034	0.034	0.114	0.154	0.174
4thYear	0.019	0.013	0.019	0.019	0.026	0.032	0.135	0.129	0.168
5thYear	0.065	0.073	0.073	0.065	0.057	0.073	0.138	0.114	0.114

Table 7. Recall@Top-5% (fraction of bankruptcies found when auditing the top 5% highest-risk firms).

Case	LR	CostSensitive-LR	ROS-LR	SMOTE-LR	RUS-LR	Focal-LR	RF	CostSensitive-RF	CostSensitive-RF+MI
1stYear	0.123	0.123	0.123	0.123	0.123	0.111	0.543	0.568	0.63
2ndYear	0.075	0.067	0.075	0.083	0.058	0.058	0.475	0.475	0.525
3rdYear	0.221	0.188	0.188	0.208	0.154	0.174	0.43	0.423	0.436
4thYear	0.161	0.161	0.161	0.148	0.123	0.123	0.381	0.406	0.542
5thYear	0.325	0.325	0.309	0.333	0.276	0.341	0.439	0.382	0.463

Table 8. Recall@Top-10% (fraction of bankruptcies found when auditing the top 10% highest-risk firms).

Case	LR	CostSensitive-LR	ROS-LR	SMOTE-LR	RUS-LR	Focal-LR	RF	CostSensitive-RF	CostSensitive-RF+MI
1stYear	0.222	0.173	0.222	0.21	0.198	0.185	0.691	0.765	0.778
2ndYear	0.217	0.217	0.217	0.233	0.167	0.15	0.6	0.65	0.625
3rdYear	0.329	0.322	0.322	0.315	0.289	0.295	0.577	0.55	0.624
4thYear	0.265	0.284	0.303	0.303	0.29	0.252	0.581	0.568	0.69
5thYear	0.545	0.545	0.545	0.52	0.488	0.488	0.618	0.626	0.724

Table 9. Audit hit-rate versus sampling cost for the 1stYear case (LR vs RF variants).

Audit rate (%)	LR Recall	CostSensitive-RF Recall	CostSensitive-RF+MI Recall	Random Recall (expected)
0.5	0.025	0.111	0.123	0.005
1.0	0.062	0.210	0.247	0.010
2.0	0.099	0.309	0.457	0.020
5.0	0.123	0.568	0.630	0.050
10.0	0.222	0.765	0.778	0.100
20.0	0.370	0.864	0.901	0.200

Table 10. Top 10 ratio feature importances for CostSensitive-RF+MI in the 1stYear case (ratio features only), with median ratio values on the test set.

Feature	Ratio	Gini importance	Median (bankrupt)	Median (non-bankrupt)	Delta (bankrupt - non)
Attr27	profit on operating activities / financial expenses	0.056	1.273	1.273	0.000
Attr24	gross profit (in 3 years) / total assets	0.038	0.029	0.172	-0.143
Attr26	(net profit + depreciation) / total liabilities	0.032	0.117	0.286	-0.169
Attr34	operating expenses / total liabilities	0.028	0.926	1.738	-0.812
Attr46	(current assets - inventory) / short-term liabilities	0.027	0.625	0.978	-0.353
Attr16	(gross profit + depreciation) / total liabilities	0.027	0.125	0.315	-0.190
Attr13	(gross profit + depreciation) / sales	0.026	0.038	0.082	-0.044
Attr6	retained earnings / total assets	0.018	0.000	0.000	0.000

Attr38	constant capital / total assets	0.018	0.456	0.593	-0.137
Attr12	gross profit / short-term liabilities	0.018	0.053	0.257	-0.204

Table 11. Most influential missingness indicator features in CostSensitive-RF+MI (1stYear), suggesting that missing-value patterns are predictive.

Missing indicator	**Associated ratio**	**Gini importance**
Missing(Attr27)	profit on operating activities / financial expenses	0.079
Missing(Attr11)	(gross profit + extraordinary items + financial expenses) / total assets	0.014
Missing(Attr21)	sales (n) / sales (n-1)	0.010
Missing(Attr45)	net profit / inventory	0.001
Missing(Attr60)	sales / inventory	0.001
Missing(Attr37)	(current assets - inventories) / long-term liabilities	0.001
Missing(Attr53)	equity / fixed assets	0.000
Missing(Attr64)	sales / fixed assets	0.000

Table 12. Audit yield at strict budget: Precision@Top-1% and Lift@1% relative to the base bankruptcy rate.

Case	Base bankruptcy rate (%)	Precision@Top-1% (LR)	Precision@Top-1% (RF+MI)	Lift@1% (LR / base)	Lift@1% (RF+MI / base)
1stYear	3.86	0.227	0.909	5.9x	23.6x
2ndYear	3.93	0.065	0.839	1.6x	21.3x
3rdYear	4.71	0.156	0.812	3.3x	17.2x
4thYear	5.26	0.100	0.867	1.9x	16.5x
5thYear	6.94	0.444	0.778	6.4x	11.2x

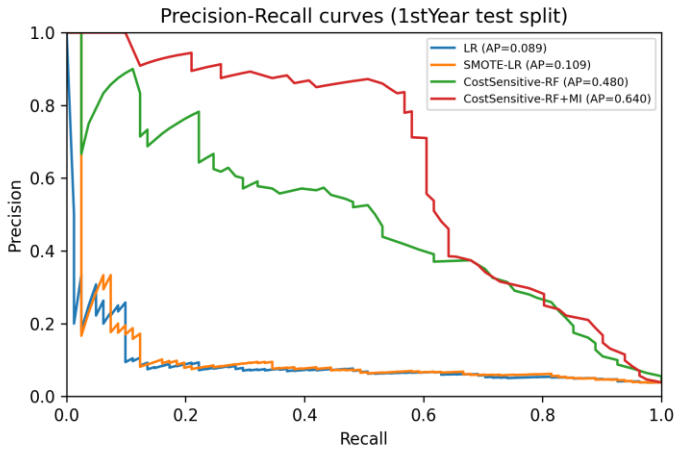

Figure 3. Precision–recall curves for representative methods on the 1stYear held-out test set (including the missingness-aware RF+MI model).

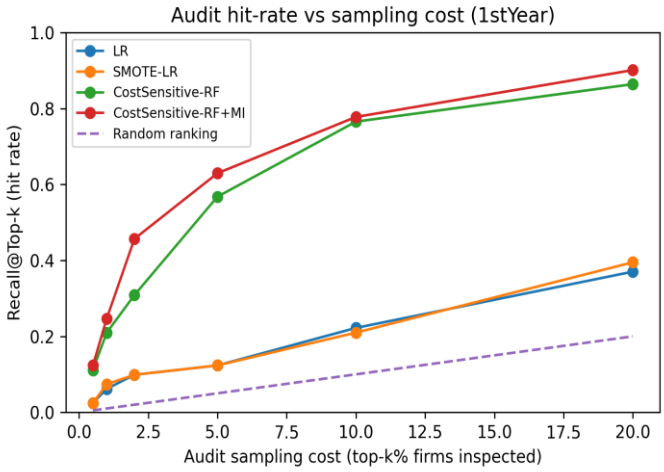

Figure 4. Top-k hit rate (Recall@Top-k) versus audit sampling cost in the 1stYear case (including RF+MI).

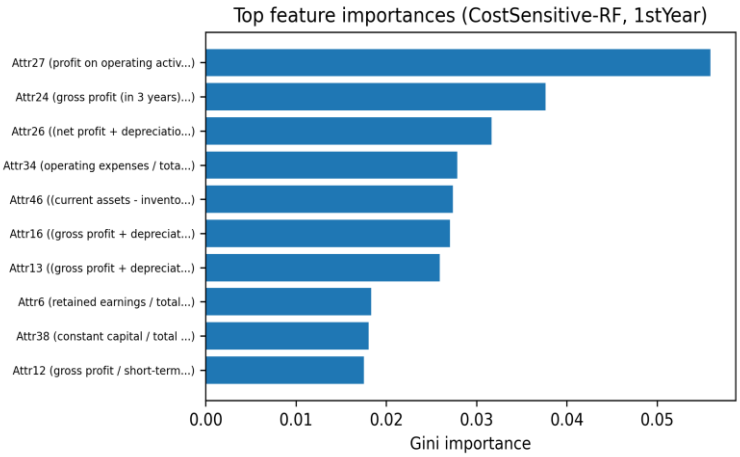

Figure 5. Top ratio feature importances for CostSensitive-RF+MI on the 1stYear case (Gini importance, ratio features only).

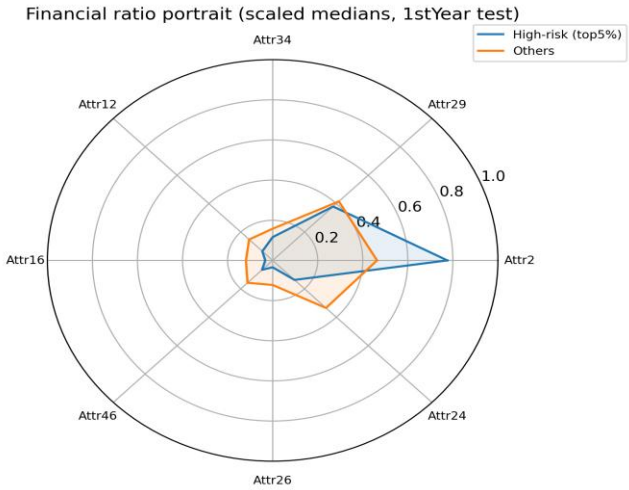

Figure 6. Explainable financial-ratio portrait: robustly scaled median ratios for high-risk (top 5%) versus other firms (1stYear, RF+MI ranking).

Limitations

First, although we evaluated all five forecasting cases, the reported results are based on a single stratified 70/30 split with a fixed random seed to guarantee exact reproducibility. While this mirrors a deployment setting where a model is trained once and then applied prospectively, repeated splits or time-based validation would provide additional uncertainty estimates and could reveal sensitivity to particular firm-year partitions. Second, the benchmark is derived from Polish firms and may not capture sectoral heterogeneity, regulatory environments, or macroeconomic dynamics in other jurisdictions. Therefore, absolute performance values should not be interpreted as universal, and external validation is required before operational deployment.

Third, the dataset is cross-sectional at the firm-year level. It does not explicitly model within-firm temporal trajectories (e.g., the evolution of liquidity over multiple years) and does not include macroeconomic indicators. Temporal models, survival analysis, or sequence learners could better reflect real going-concern assessment where auditors consider trends and subsequent events. Fourth, we restricted base learners to logistic regression and random forests to isolate the effect of imbalance-handling strategies. Stronger tabular learners such as gradient-boosted trees or calibrated ranking models could further improve performance and might interact differently with resampling and focal loss.

Fifth, focal loss was evaluated in a linear configuration with fixed hyperparameters (α=0.25, γ=2.0). Focal loss can be sensitive to tuning and model capacity; pairing focal loss with multi-layer networks or tuning α/γ per horizon may change conclusions about its utility. Finally, explainability was addressed through global importances and cohort-level ratio portraits. These tools align with audit practice and are useful for triage and planning, but they do not replace case-specific documentation required for an individual audit opinion. Future work could integrate local explanations (e.g., SHAP [12]) with cohort portraits and evaluate their effect on auditor decision quality in controlled studies.

Conclusion

This paper studied bankruptcy/going-concern prediction as an extreme-imbalance ranking problem motivated by constrained audit and credit review resources. Using the Polish Companies Bankruptcy benchmark (five forecasting horizons with 64 financial ratios), we conducted a reproducible empirical evaluation of cost-sensitive learning, resampling, and focal loss, reporting both global ranking quality (AUPRC) and operational budgeted recall (Recall@Top-k).

Across all horizons, tree ensembles substantially outperformed linear baselines, indicating meaningful non-linear structure in financial ratios. Classic imbalance techniques (class weighting, ROS, SMOTE) improved logistic regression modestly but did not match the retrieval performance of random forests under small audit budgets. The audit-oriented results were operationally significant: in the most imbalanced 1stYear case, auditing only the top 5% of firms ranked by the best model recovered 62.96% of bankruptcies, compared with 12.35% for standard logistic regression.

A central empirical finding is that missingness is predictive in this benchmark. Augmenting the cost-sensitive random forest with missing-value indicator features increased AUPRC by 0.16 absolute points in 1stYear and improved performance in every horizon.

This suggests that practitioners should treat missingness patterns as potential signal in going-concern datasets rather than only noise to be imputed away. The most important indicators (e.g., Missing(Attr27)) can be interpreted as reflecting reporting structure or undefined ratio regimes that correlate with distress.

Finally, we presented an explainable financial-ratio portrait that summarizes the characteristic ratio pattern of high-risk cohorts. The portrait emphasized higher leverage and weaker liquidity/coverage among flagged firms, aligning with the intuition of classic distress models and offering a concise narrative for audit planning. Overall, the combination of audit-aligned evaluation, imbalance-aware learning, missingness-aware preprocessing, and interpretable ratio portraits provides a practical template for deploying early-warning screening models in going-concern workflows.

References

[1] E. I. Altman, "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy," The Journal of Finance, vol. 23, no. 4, pp. 589–609, 1968.

[2] J. A. Ohlson, "Financial ratios and the probabilistic prediction of bankruptcy," Journal of Accounting Research, vol. 18, no. 1, pp. 109–131, 1980.

[3] M. Zięba, S. Tomczak, and J. M. Tomczak, "Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction," Expert Systems with Applications, vol. 58, pp. 93–101, 2016.

[4] S. Tomczak, "Polish Companies Bankruptcy," UCI Machine Learning Repository, 2016. DOI: 10.24432/C5F600.

[5] C. Elkan, "The foundations of cost-sensitive learning," in Proc. 17th Int. Joint Conf. Artificial Intelligence (IJCAI), 2001, pp. 973–978.

[6] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

[7] H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Trans. Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263–1284, 2009.

[8] S. Japkowicz, "The class imbalance problem: Significance and strategies," in Proc. Int. Conf. Artificial Intelligence (ICAI), 2000, pp. 111–117.

[9] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in Proc. 23rd Int. Conf. Machine Learning (ICML), 2006, pp. 233–240.

[10] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," PLoS ONE, vol. 10, no. 3, e0118432, 2015.

[11] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2017, pp. 2980–2988.

[12] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. 31st Conf. Neural Information Processing Systems (NeurIPS), 2017, pp. 4765–4774.

[13] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should I trust you?": Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

[14] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[15] E. A. Mutchler, "A multivariate analysis of the auditor's going-concern opinion decision," Journal of Accounting Research, vol. 23, no. 2, pp. 668–682, 1985.

[16] M. A. Geiger and K. Raghunandan, "Auditor reporting for bankrupt companies: The role of audit report lag," Auditing: A Journal of Practice & Theory, vol. 21, no. 1, pp. 1–12, 2002.

[17] P. Domingos, "Metacost: A general method for making classifiers cost-sensitive," in Proc. 5th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 1999, pp. 155–164.

[18] B. Zadrozny, J. Langford, and N. Abe, "Cost-sensitive learning by cost-proportionate example weighting," in Proc. 3rd IEEE Int. Conf. Data Mining (ICDM), 2003, pp. 435–442.

[19] G. King and L. Zeng, "Logistic regression in rare events data," Political Analysis, vol. 9, no. 2, pp. 137–163, 2001.

[20] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," Journal of Machine Learning Research, vol. 3, pp. 1157–1182, 2003.

[21] C. Molnar, Interpretable Machine Learning, 2nd ed. 2020.

[22] T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006.

[23] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann, 2011.

[24] T. Shirakawa, Y. Li, Y. Wu, S. Qiu, Y. Li, M. Zhao, H. Iso, and M. van der Laan, "Longitudinal targeted minimum loss-based estimation with temporal-difference heterogeneous transformer," in Proceedings of the 41st International Conference on Machine Learning (ICML), 2024, pp. 45097–45113, Art. no. 1836.

[25] Xinzhuo Sun, Yifei Lu, and Jing Chen, "Controllable Long-Term User Memory for Multi-Session Dialogue: Confidence-Gated Writing, Time-Aware Retrieval-Augmented Generation, and Update/Forgetting", JACS, vol. 3, no. 8, pp. 9–24, Aug. 2023, doi: 10.69987/JACS.2023.30802.

[26] Hanqi Zhang, "DriftGuard: Multi-Signal Drift Early Warning and Safe Re-Training/Rollback for CTR/CVR Models", JACS, vol. 3, no. 7, pp. 24–40, Jul. 2023, doi: 10.69987/JACS.2023.30703.

[27] Meng-Ju Kuo, Boning Zhang, and Haozhe Wang, "Tokenized Flow-Statistics Encrypted Traffic Analysis: Comparative Evaluation of 1D-CNN, BiLSTM, and Transformer on ISCX VPN-nonVPN 2016 (A1+A2, 60 s)", JACS, vol. 3, no. 8, pp. 39–53, Aug. 2023, doi: 10.69987/JACS.2023.30804.

[28] Z. Zhong, M. Zheng, H. Mai, J. Zhao, and X. Liu, "Cancer image classification based on DenseNet model," Journal of Physics: Conference Series, vol. 1651, no. 1, p. 012143, 2020.

[29] Jubin Zhang, "Interpretable Skill Prioritization for Volleyball Education via Team-Stat Modeling", JACS, vol. 3, no. 3, pp. 34–49, Mar. 2023, doi: 10.69987/JACS.2023.30304.

[30] Jinyi Mu, Yifei Lu, and Michelle Smith, "LLM-Assisted Incrementality (Uplift) Modeling for Email Advertising: From Feature Interactions to Interpretable Audience–Creative–Channel Policies ", JACS, vol. 3, no. 1, pp. 31–48, Jan. 2023, doi: 10.69987/JACS.2023.30103.

[31] Siming Zhao, Hailin Zhou, and Daniel Martinez, "LLM-Assisted Causal Attribution of Service Performance Upgrades on Churn and Tenure: Full Evaluation on the IBM Telco Customer Churn Dataset", JACS, vol. 3, no. 2, pp. 18–34, Feb. 2023, doi: 10.69987/JACS.2023.30202.

[32] Daren Zheng, Chenyu Li, and Harvey Davidson, "Continual Red-Teaming for In-the-Wild Jailbreaks via Online Guardrail Updates and Guardrail Distillation", JACS, vol. 3, no. 2, pp. 35–49, Feb. 2023, doi: 10.69987/JACS.2023.30203.

[33] Binghua Zhou, Siming Zhao, and David Chao, "LLM-Guided Energy-Aware A/B Testing for Consolidation and DVFS Policies via Power-Sensitivity Clustering", JACS, vol. 3, no. 4, pp. 12–30, Apr. 2023, doi: 10.69987/JACS.2023.30402.

[34] Jing Chen, Xinzhuo Sun, and Vincent Brown, "Claim-Aware Scientific RAG: Evidence-First Retrieval and Abstention for Scientific Fact Responses on SciFact", JACS, vol. 3, no. 1, pp. 16–30, Jan. 2023, doi: 10.69987/JACS.2023.30102.

[35] K. Xu, H. Zhou, H. Zheng, M. Zhu, and Q. Xin, "Intelligent classification and personalized recommendation of e-commerce products based on machine learning," Proceedings of the 6th International Conference on Computing and Data Science (ICCDS), 2024.

[36] Q. Xin, Z. Xu, L. Guo, F. Zhao, and B. Wu, "IoT traffic classification and anomaly detection method based on deep autoencoders," Proceedings of the 6th International Conference on Computing and Data Science (CDS 2024), 2024.

[37] B. Wang, Y. He, Z. Shui, Q. Xin, and H. Lei, "Predictive optimization of DDoS attack mitigation in distributed systems using machine learning," Proceedings of the 6th International Conference on Computing and Data Science (CDS 2024), 2024, pp. 89–94.

[38] Z. Ling, Q. Xin, Y. Lin, G. Su, and Z. Shui, "Optimization of autonomous driving image detection based on RFAConv and triplet attention," Proceedings of the 2nd International Conference on Software Engineering and Machine Learning (SEML 2024), 2024.

[39] J. Chen, J. Xiong, Y. Wang, Q. Xin, and H. Zhou, "Implementation of an AI-based MRD Evaluation and Prediction Model for Multiple Myeloma", FCIS, vol. 6, no. 3, pp. 127–131, Jan. 2024, doi: 10.54097/zJ4MnbWW.

[40] Xinzhuo Sun, Jing Chen, Binghua Zhou, and Meng-Ju Kuo, "ConRAG: Contradiction-Aware Retrieval-Augmented Generation under Multi-Source Conflicting Evidence", JACS, vol. 4, no. 7, pp. 50–64, Jul. 2024, doi: 10.69987/JACS.2024.40705.

[41] Hanqi Zhang, "Risk-Aware Budget-Constrained Auto-Bidding under First-Price RTB: A Distributional Constrained Deep Reinforcement Learning Framework", JACS, vol. 4, no. 6, pp. 30–47, Jun. 2024, doi: 10.69987/JACS.2024.40603.

[42] Z. S. Zhong and S. Ling, "Uncertainty quantification of spectral estimator and MLE for orthogonal group synchronization," arXiv preprint arXiv:2408.05944, 2024.

[43] Z. S. Zhong and S. Ling, "Improved theoretical guarantee for rank aggregation via spectral method,"

Information and Inference: A Journal of the IMA, vol. 13, no. 3, 2024.

[44] Jubin Zhang, "Tactical Language + AI Tutoring from Structured Volleyball Rally Logs: Reproducible Experiments on NCAA Play-by-Play", JACS, vol. 4, no. 1, pp. 58–66, Jan. 2024, doi: 10.69987/JACS.2024.40105.

[45] Xiaofei Luo, "Semantic Verifier for Post-hoc Answer Validation in Chat Platforms: Claim Decomposition, Evidence Retrieval, NLI, and Traceable Citations", JACS, vol. 4, no. 3, pp. 74–90, Mar. 2024, doi: 10.69987/JACS.2024.40306.

[46] Xiaofei Luo, "Execution-Validated Program-Supervised Complex KBQA: A Reproducible 120K-Question Study with KoPL-Style Programs", JACS, vol. 4, no. 6, pp. 48–63, Jun. 2024, doi: 10.69987/JACS.2024.40604.

[47] Daren Zheng and Chenyu Li, "Behavior-Level Jailbreak Resistance via Multi-Stage Refusal + Utility Preservation", JACS, vol. 4, no. 1, pp. 83–99, Jan. 2024, doi: 10.69987/JACS.2024.40107.

[48] Siming Zhao, Haozhe Wang, and Neil Davison, "Profit-Maximizing Cost-Sensitive Credit Scoring with LLM-Extracted Policy Constraints", JACS, vol. 4, no. 3, pp. 91–108, Mar. 2024, doi: 10.69987/JACS.2024.40307.

[49] Yifei Lu, Jinyi Mu, and Thao Tran, "Uncertainty-Aware Uplift Modeling for Safer Marketing Targeting: Conformal Prediction and Bayesian Calibration with LCB Policies", JACS, vol. 4, no. 5, pp. 84–101, May 2024, doi: 10.69987/JACS.2024.40507.

[50] Jing Chen, Xinzhuo Sun, Qiyou Wu, and Matt Jackson, "Risk-Calibrated Biomedical Search: Calibrated Selection of LLM-Style Query Expansions on BEIR TREC-COVID", JACS, vol. 4, no. 4, pp. 61–79, Apr. 2024, doi: 10.69987/JACS.2024.40406.

[51] Daren Zheng, Boning Zhang, and Julie Geibel, "VerifySafe: Toxicity-Safe Agent Responses under Adversarial Prompts with Evidence-Based Self-Verification", JACS, vol. 4, no. 1, pp. 67–82, Jan. 2024, doi: 10.69987/JACS.2024.40106.

[52] Yushan Chen and Evelyn Chan, "Multimodal UI Representation Learning: Ablation of Screenshot, Wireframe, and View-Hierarchy Proxies on an Uploaded 168-Screen Dataset", JACS, vol. 3, no. 1, pp. 1–15, Jan. 2023, doi: 10.69987/JACS.2023.30101.

[53] Yunhe Li, "Execution-Feedback and Retrieval-Augmented Generation for Conversational Text-to-SQL: From One-Shot Questions to Clarification-Driven Executable Dialogs", JACS, vol. 3, no. 2, pp. 1–17, Feb. 2023, doi: 10.69987/JACS.2023.30201.

[54] Yunhe Li, "Risk-Sensitive Offline Reinforcement Learning for Stable ABR QoE Improvements on Real HSDPA and LTE Traces", JACS, vol. 3, no. 4, pp. 1–11, Apr. 2023, doi: 10.69987/JACS.2023.30401.

[55] Jason Kuhn, Yushan Chen, and Evelyn Chan, "AI-Driven Mobile UI Pattern Recognition and Design Topic Mining on RICO: Semantic Clustering and Screenshot-Based Topic Classification", JACS, vol. 4, no. 5, pp. 67–83, May 2024, doi: 10.69987/JACS.2024.40506.

[56] Yunhe Li, "Findable then Explainable: Retrieval–Summary Integration for Code Intelligence on a Lightweight CodeSearchNet Subset", JACS, vol. 4, no. 7, pp. 65–82, Jul. 2024, doi: 10.69987/JACS.2024.40706.

[57] Yunhe Li, "Test-in-the-loop LLM Repair: Verifiable Automated Program Repair on QuixBugs with a 'Failing Test → Patch → Regression Test' Loop", JACS, vol. 4, no. 2, pp. 62–75, Feb. 2024, doi: 10.69987/JACS.2024.40206.