International Journal of Computer Applications Technology and Research
Volume 15–Issue 02, 72 – 83, 2026, ISSN: -2319–8656
DOI:10.7753/IJCATR1502.1006

An Interpretable Gradient-Boosted Risk Model for Detecting Scam-Induced IRA Distributions Among Investors Aged 65 and Older

Olasunbo O. Fagbore
Senior Brokerage Consultant II
Fidelity Investment
Westlake, TX, USA

Adeola Ibitogbe
The Annie E. Casey Foundation
Baltimore, MD, USA

Priya S. Venkataraman
School of Information Sciences
Ridgeview Conglomerate
TX, USA

Abstract: Older investors who are deceived by romance, impersonation, and investment scams frequently fund the scammer by withdrawing money from individual retirement accounts (IRAs). These withdrawals are authorized by the account owner, so identity verification and credential controls do not stop them. We develop and validate an explainable risk model that scores IRA distribution requests for the probability that the request was induced by a scam. Using administrative records from a national retail brokerage and IRA custodian covering 412,806 distribution requests submitted by 187,340 account owners aged 65 and older between January 2021 and December 2024, of which 1,962 were later confirmed as scam-induced, we compare seven model families under a strict temporal holdout. A gradient-boosted tree model achieved an out-of-time AUROC of 0.919 and identified 59.7% of scam-induced requests while alerting on 1% of all requests, against 9.7% for the custodian's incumbent rules engine. A monotone-constrained variant lost almost no discrimination while producing explanations that compliance reviewers judged more coherent. SHAP decomposition showed that new or recently changed destination accounts, contact-detail changes in the preceding fortnight, unusual liquidation of long-held positions, and the requested size relative to historical distributions carried most of the signal. In a within-subject reviewer experiment with 24 fraud analysts, adding feature-level explanations to the risk score raised decision accuracy from 0.77 to 0.84 and cut median review time by 22%. Performance was stable across age bands and rurality, though false-positive rates were modestly higher for the oldest owners. We discuss implications for trusted-contact outreach and temporary holds under FINRA Rule 2165.

Keywords: elder financial exploitation, scam detection, individual retirement accounts, explainable machine learning, SHAP, gradient boosting, FINRA Rule 2165

1. INTRODUCTION

Financial fraud against older adults has grown in both frequency and severity. Complaints filed with the FBI Internet Crime Complaint Center by people aged 60 and older reported losses of roughly 3.4 billion dollars in 2023, an increase of about 11 percent over the prior year (Federal Bureau of Investigation, 2024). The Federal Trade Commission's Consumer Sentinel data show a similar pattern, and median individual losses are highest among the oldest age groups (Federal Trade Commission, 2024). A meta-analysis of population studies estimated that roughly 5.6 percent of older adults in the United States experience fraud or scam victimization each year (Burnes et al., 2017), and survey data suggest that many cases are never reported (Acierno et al., 2010).

Retirement accounts sit at the centre of this problem. Individual retirement accounts often hold the largest liquid balance an older household owns, and they can be liquidated by the owner through a phone call or a short online form. The modern scam rarely involves a stolen password. Instead, the scammer persuades the owner to request the distribution personally, frequently after weeks of grooming. The account owner passes every identity check because the account owner is the person on the line. Losses in these cases fall outside the protections that cover unauthorized transactions, and the tax consequences of a large early or irregular distribution can compound the harm well after the money is gone.

Consider a 74-year-old widow who calls her custodian to withdraw 180,000 dollars from her IRA and asks that the funds be wired within 24 hours to a newly opened business account. She has never taken a distribution beyond her required minimum, and her phone number was changed eleven days earlier. Every one of these facts is individually unremarkable. Together they form a pattern that experienced fraud investigators recognize immediately, yet a static rules engine that fires on single thresholds can easily miss it. The central question of this paper is whether a statistical model can learn such patterns at scale while remaining transparent enough for a human reviewer to act on its output.

Regulation now gives firms tools to intervene. FINRA Rule 2165 permits member firms to place temporary holds on disbursements when they reasonably believe that financial exploitation of a specified adult has occurred or is being attempted, and FINRA Rule 4512 requires firms to make reasonable efforts to collect a trusted contact person (FINRA, 2018a, 2018b). The Senior Safe Act provides immunity for good-faith reporting by trained employees. These tools only help, however, if firms can identify the right requests quickly and explain to the customer, to a trusted contact, and to an examiner why a hold was placed. A risk model that cannot justify its alerts is a weak foundation for an intrusive intervention (Rudin, 2019).

This paper makes four contributions. First, we assemble a large, labelled, out-of-time dataset of IRA distribution requests from older owners and document the typology, timing, and dollar distribution of scam-induced withdrawals. Second, we benchmark seven model families and show that tree ensembles substantially outperform the incumbent rules and linear baselines on the operating region that matters, namely very low alert rates. Third, we decompose the best model with SHAP values (Lundberg and Lee, 2017; Lundberg et al., 2020), test the stability of the resulting explanations, and show that monotonicity constraints buy coherence at negligible cost.

Fourth, we run a reviewer experiment that measures whether explanations actually improve the decisions of the people who must act on them.

2. RELATED WORK

2.1 Susceptibility to Scams in Later Life

A substantial gerontology literature links scam susceptibility to declines in cognition, reduced financial literacy, social isolation, and psychological vulnerability. Using data from the Rush Memory and Aging Project, James, Boyle, and Bennett (2014) found that lower cognitive function and lower psychological well-being predicted higher scam susceptibility among older adults without dementia. Boyle et al. (2019) later showed that low scam awareness predicted incident Alzheimer dementia and mild cognitive impairment. Gamble et al. (2015) documented that declining cognition reduces financial literacy without a matching decline in financial confidence, a gap that scammers exploit. Lichtenberg, Stickney, and Paulson (2013) identified depression and unmet social needs as correlates of fraud. DeLiema, Shadel, and Pak (2020) showed that investment fraud victims share risky behaviours such as responding to unsolicited offers. These findings inform which behavioural traces might be visible in account data, although custodians rarely observe cognition or mood directly.

2.2 Statistical Fraud Detection

Fraud detection has a long statistical history (Bolton and Hand, 2002). Most published work targets unauthorized payment-card or account-takeover fraud, where the signal is a mismatch between the transaction and the legitimate owner. Feature engineering based on transaction aggregation and temporal patterns has been central to that work (Bahnsen et al., 2016). Severe class imbalance complicates both training and evaluation, and precision-recall analysis is more informative than ROC analysis when positives are rare (Saito and Rehmsmeier, 2015). Undersampling distorts predicted probabilities and calls for recalibration (Dal Pozzolo et al., 2015). Scam-induced withdrawals differ from takeover fraud because the owner is authentic, which shifts the informative signals from identity anomalies toward behavioural and contextual anomalies.

2.3 Explainability in High-Stakes Decisions

Tree ensembles remain the strongest general-purpose learners for tabular data (Chen and Guestrin, 2016; Grinsztajn, Oyallon, and Varoquaux, 2022). Their opacity has prompted post-hoc explanation methods such as LIME (Ribeiro, Singh, and Guestrin, 2016) and SHAP (Lundberg and Lee, 2017), with exact tree algorithms making SHAP practical at scale (Lundberg et al., 2020). Rudin (2019) argued that inherently interpretable models should be preferred whenever they perform comparably, and generalized additive models with pairwise interactions offer one such option (Lou et al., 2013; Nori et al., 2019). Evidence on whether explanations improve human decisions is mixed. Kaur et al. (2020) found that data scientists often over-trust explanation tools, and Bansal et al. (2021) found that explanations can increase acceptance of incorrect AI recommendations. We therefore treat the human reviewer, not the model, as the unit whose performance ultimately matters.

3. DATA AND METHODS

3.1 Setting and Data Sources

Data come from a national retail brokerage and IRA custodian (hereafter, the Custodian) under a data use agreement that permitted linkage of de-identified records. The Custodian serves self-directed and advisor-assisted retail clients across all 50 states. We extracted every IRA distribution request (traditional, Roth, rollover, and inherited) submitted between 1 January 2021 and 31 December 2024 by an owner aged 65 or older at the time of request, excluding systematic required minimum distributions that were scheduled more than 90 days in advance. The resulting frame contained 412,806 requests from 187,340 owners. Requests were linked to account history, contact-channel logs, call-centre agent notes, digital session telemetry, and downstream fraud case management outcomes.

3.2 Outcome Definition

A request was labelled scam-induced if a fraud investigation closed within 180 days with a finding that the owner initiated the request under deception by a third party. Labels were assigned by the Custodian's Elder Financial Protection Unit using a written adjudication standard, and 400 randomly selected closed cases were independently re-adjudicated by two of the authors with substantial agreement (Cohen's $\kappa = 0.81$). Requests that were stopped before disbursement were labelled positive if the investigation confirmed an active scam. Overall, 1,962 requests (0.48 percent) were labelled scam-induced. Because some scams are never discovered, the negative class contains an unknown amount of label noise, which biases measured precision downward.

3.3 Features

We engineered 146 candidate features in six groups, namely (a) request attributes such as amount, delivery method, and withholding elections, (b) account context such as balance, tenure, and historical distribution behaviour, (c) destination attributes such as account age and institution type, (d) contact-channel activity such as inbound calls and recent changes to phone, email, or address, (e) digital session telemetry such as device novelty and session duration, and (f) structured agent observations recorded during the call. All features were computed using only information available at the moment of request. One feature, the expected incremental tax liability, was computed as the difference between the actual tax owed and the tax under a spread of the same amount across three tax years, which captures whether the requested timing makes financial sense for the owner.

3.4 Temporal Design and Models

To mimic deployment, we used a strict temporal split. Models were trained on requests from January 2021 to June 2023 (n = 289,114, 1,371 positives), tuned on July to December 2023 (n = 58,902, 281 positives), and evaluated once on calendar year 2024 (n = 64,790, 310 positives). We compared the incumbent rules engine, unpenalized and elastic-net logistic regression, a random forest, an explainable boosting machine (EBM) with up to ten pairwise interactions (Nori et al., 2019), XGBoost (Chen and Guestrin, 2016), and an XGBoost variant with monotone constraints on 22 features where the direction of effect was fixed in advance by investigators (MC-XGBoost). Hyperparameters were tuned by Bayesian optimization over 200 trials maximizing validation AUPRC. Predicted probabilities were recalibrated on the validation set using Platt scaling, and isotonic regression was examined as a sensitivity analysis (Niculescu-Mizil and Caruana, 2005).

International Journal of Computer Applications Technology and Research
Volume 15–Issue 02, 72 – 83, 2026, ISSN: -2319–8656
DOI:10.7753/IJCATR1502.1006

3.5 Evaluation

We report AUROC with DeLong-style comparisons (DeLong, DeLong, and Clarke-Pearson, 1988), AUPRC, Brier score, and recall and precision at fixed alert rates of 1 and 2 percent, which correspond to review budgets the Custodian considered feasible. Uncertainty was estimated with stratified bootstrap resampling. Clinical-style decision curve analysis (Vickers and Elkin, 2006) was used to compare net benefit across threshold probabilities. Reporting follows the TRIPOD statement where applicable (Collins et al., 2015). Explanations were computed with TreeSHAP and their stability was assessed across 200 bootstrap refits. Subgroup performance was assessed for age band, sex, rurality, and account size, with attention to equality of true-positive rates (Hardt, Price, and Srebro, 2016).

Net benefit at threshold probability p_t was computed as

$$NB(p_t) = TP/N - (FP/N) \times p_t / (1 - p_t) \quad (1)$$

and population stability for feature j between a reference period r and a monitoring period m was computed as

$$PSI_j = \Sigma_k (m_k - r_k) \cdot \ln(m_k / r_k) \quad (2)$$

3.6 Reviewer Experiment

To test whether explanations change human decisions, we recruited 24 fraud analysts from the Custodian's review team (median experience 4.5 years). Each analyst reviewed 60 historical cases drawn from the 2024 test set and enriched to a 25 percent positive rate, presented in three counterbalanced blocks of 20 under three conditions, namely case file only, case file plus risk score, and case file plus risk score plus the top five SHAP contributions in plain-language form. Analysts recorded a hold or release decision, a confidence rating, and a trust rating. The protocol was reviewed and approved by the institutional review board of the lead author's university.

4. RESULTS

4.1 Sample Description

Table 1 compares scam-induced and legitimate requests. Scam-induced requests were larger in absolute terms and much larger relative to the owner's history, were more often directed to destination accounts opened within the previous 30 days, and were preceded by more inbound calls and more contact-detail changes. Owners in the scam group were slightly older and more often lived alone according to account records.

Table 1. Characteristics of IRA distribution requests by scam status, owners aged 65 and older, 2021–2024

Characteristic	Legitimate (n = 410,844)	Scam-induced (n = 1,962)	SMD	p
Owner age, mean (SD), years	73.4 (6.2)	76.9 (6.8)	0.54	<0.001
Female, %	53.1	61.8	0.18	<0.001
Lives alone (account flag), %	27.4	41.2	0.29	<0.001
Rural ZIP code (RUCA ≥ 4), %	18.9	21.7	0.07	0.002
Account balance, median (IQR), $000	214 (88–512)	246 (109–561)	0.06	0.004
Request amount, median (IQR), $000	12.0 (4.5–30.0)	64.0 (25.0–150.0)	1.12	<0.001
Amount ÷ prior 3-yr max distribution, median	1.0	4.8	0.97	<0.001
First non-RMD distribution, %	22.6	58.3	0.78	<0.001
Destination account opened < 30 days, %	3.1	37.9	0.95	<0.001
Destination is crypto exchange or P2P, %	0.6	14.2	0.54	<0.001
Wire or overnight delivery requested, %	11.8	62.4	1.20	<0.001
Phone or email changed in prior 14 days, %	2.2	19.6	0.59	<0.001
Inbound calls in prior 7 days, mean (SD)	0.4 (0.9)	2.1 (2.3)	0.97	<0.001
Federal withholding waived, %	18.4	44.7	0.59	<0.001
Trusted contact on file, %	34.0	22.9	0.25	<0.001

Note. SMD = standardized mean difference. p values from Welch t tests or chi-square tests. RUCA = Rural-Urban Commuting Area code. IQR = interquartile range. RMD = required minimum distribution.

Table 2 shows the typology of confirmed scams. Romance and government-impersonation scams were the most common, but investment and cryptocurrency scams produced the largest median losses. Technology-support scams frequently escalated into so-called phantom hacker schemes in which the owner was told to move money to a safe account.

International Journal of Computer Applications Technology and Research
Volume 15–Issue 02, 72 – 83, 2026, ISSN: -2319–8656
DOI:10.7753/IJCATR1502.1006

Table 2. Typology of confirmed scam-induced IRA distribution requests

Scam type	n	%	Median request ($000)	Total requested ($M)	Stopped before payout (%)
Romance / relationship	512	26.1	58.0	48.6	21.3
Government impersonation	441	22.5	42.5	29.2	27.0
Tech support / phantom hacker	367	18.7	85.0	46.0	24.8
Investment / cryptocurrency	318	16.2	110.0	52.7	17.6
Family emergency / grandparent	164	8.4	18.0	4.1	35.4
Lottery / prize	88	4.5	22.0	2.9	31.8
Other / mixed	72	3.7	35.0	3.8	26.4
Total	**1,962**	**100.0**	**55.0**	**187.3**	**24.2**

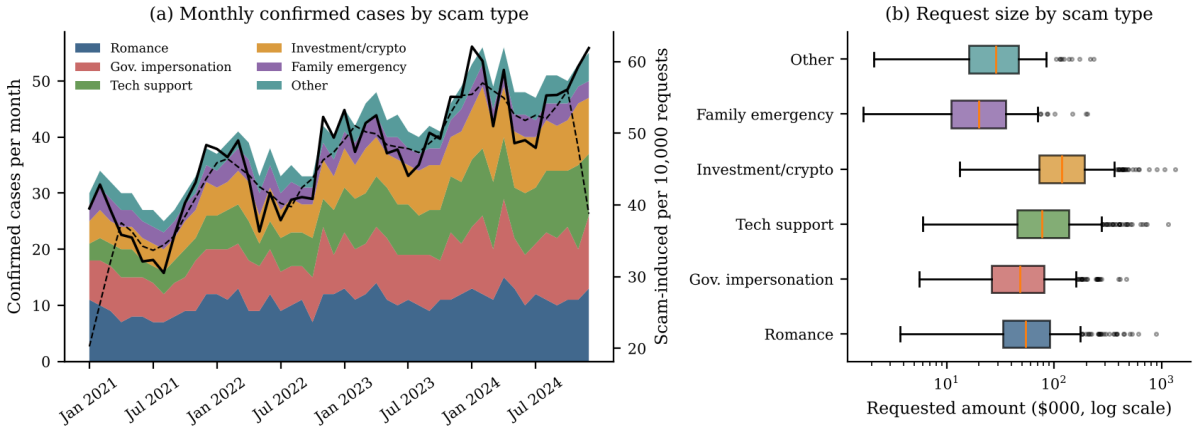


Figure 1. Temporal and dollar profile of confirmed scam-induced IRA distribution requests, January 2021 to December 2024. Panel (a) stacks monthly confirmed cases by scam type (left axis) and overlays the rate per 10,000 distribution requests (solid line) with a six-month moving average (dashed). Panel (b) shows the distribution of requested amounts on a log scale. Technology-support and investment scams grew fastest and produced the largest requests.

4.2 Discrimination

Table 3 reports out-of-time performance on the 2024 test set. All learned models outperformed the incumbent rules engine. The gains were largest on the metrics that reflect operational value, because a fraud unit can review only a small fraction of requests. At a 1 percent alert rate, XGBoost captured 59.7 percent of scam-induced requests compared with 9.7 percent for the rules engine. The monotone-constrained model was statistically indistinguishable from unconstrained XGBoost on AUROC, and the EBM came close despite being fully additive apart from ten interactions.

Table 3. Out-of-time discrimination and calibration on the 2024 test set (n = 64,790, 310 scam-induced)

Model	AUROC (95% CI)	AUPRC	Brier ×10³	Recall @1%	Precision @1%	Recall @2%	p vs. XGB
Rules engine (incumbent)	0.770 (0.744–0.795)	0.032	4.70	0.097	0.046	0.197	<0.001
Logistic regression	0.821 (0.795–0.850)	0.079	4.58	0.226	0.108	0.303	<0.001
Elastic-net LR	0.866 (0.838–0.889)	0.236	4.15	0.384	0.184	0.465	<0.001
Random forest	0.902 (0.884–0.921)	0.396	5.50	0.494	0.236	0.558	0.055
EBM (GA2M)	0.913 (0.897–0.929)	0.488	3.20	0.587	0.281	0.629	0.448
MC-XGBoost	0.914 (0.895–0.932)	0.496	3.16	0.565	0.270	0.645	0.529
XGBoost	0.919 (0.902–0.936)	0.525	3.02	0.597	0.285	0.642	Ref.

Note. Recall and precision at fixed alert rates rank all test requests by score and alert on the top 1 or 2 percent. Confidence intervals from 150 stratified bootstrap resamples. p values compare AUROC with XGBoost. Alert budgets of 1 and 2 percent are planning scenarios chosen with the Custodian's operations team and do not reflect current staffing. EBM = explainable boosting machine. MC = monotone-constrained.

International Journal of Computer Applications Technology and Research
Volume 15–Issue 02, 72 – 83, 2026, ISSN: -2319–8656
DOI:10.7753/IJCATR1502.1006

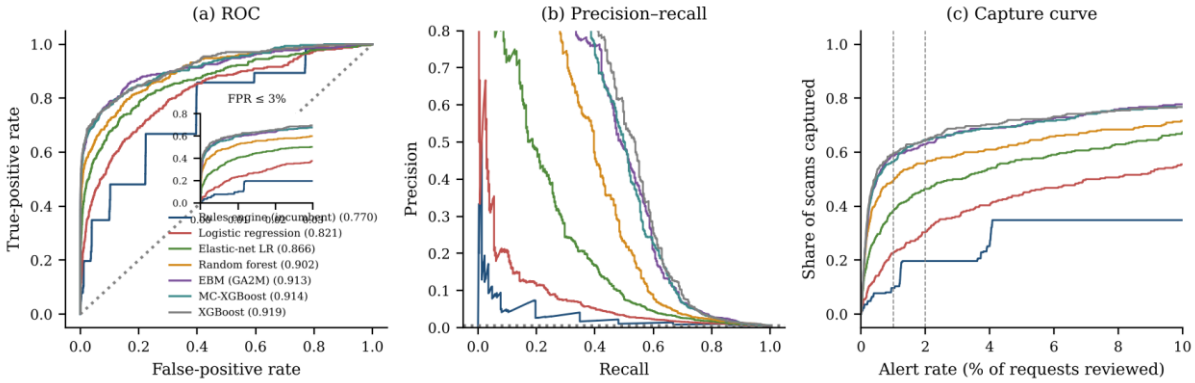


Figure 2. Discrimination on the 2024 out-of-time test set. (a) ROC curves with AUROC in the legend and an inset magnifying the low false-positive region where operations actually work. (b) Precision–recall curves with the dotted line marking the base rate of 0.48 percent. (c) Cumulative share of scam-induced requests captured as a function of the share of all requests reviewed, with dashed lines at the 1 and 2 percent review budgets.

4.3 Calibration

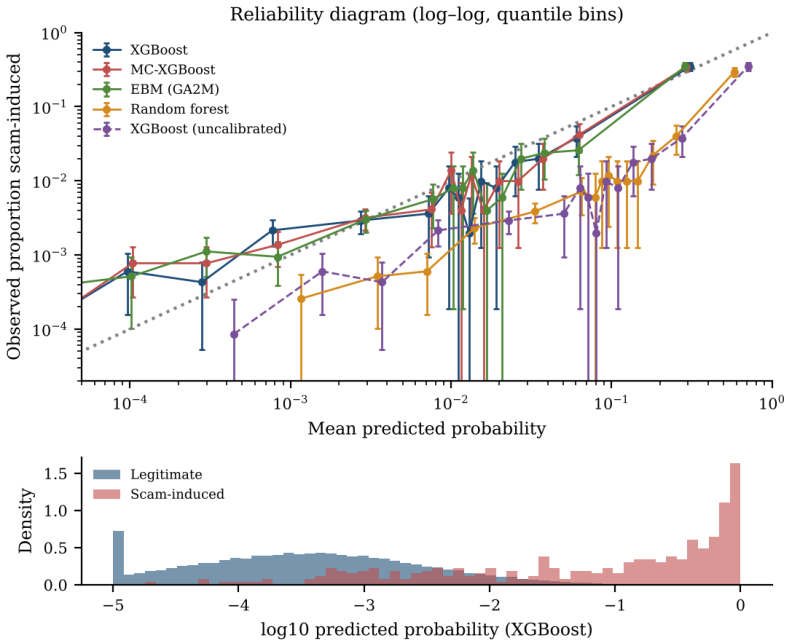


Figure 3. Calibration of predicted probabilities on the test set. The upper panel plots observed against predicted risk in 15 quantile bins on log scales with 95 percent binomial intervals. Platt-recalibrated models track the diagonal, while the uncalibrated XGBoost and the random forest overstate risk across most of the range. The lower panel shows the distribution of log10 predicted probabilities for XGBoost by true class.

Recalibration mattered more than model choice for the accuracy of predicted probabilities. Without recalibration, the class weighting used during training inflated predicted risk by a factor of roughly two to three in the middle of the distribution (Figure 3). After Platt scaling the Brier score for XGBoost fell by 31 percent and the calibration slope was 0.97. Isotonic recalibration produced similar results but was unstable in the upper tail, where fewer than 50 positives fell in each bin.

International Journal of Computer Applications Technology and Research
Volume 15–Issue 02, 72 – 83, 2026, ISSN: -2319–8656
DOI:10.7753/IJCATR1502.1006

4.4 What the Model Learned

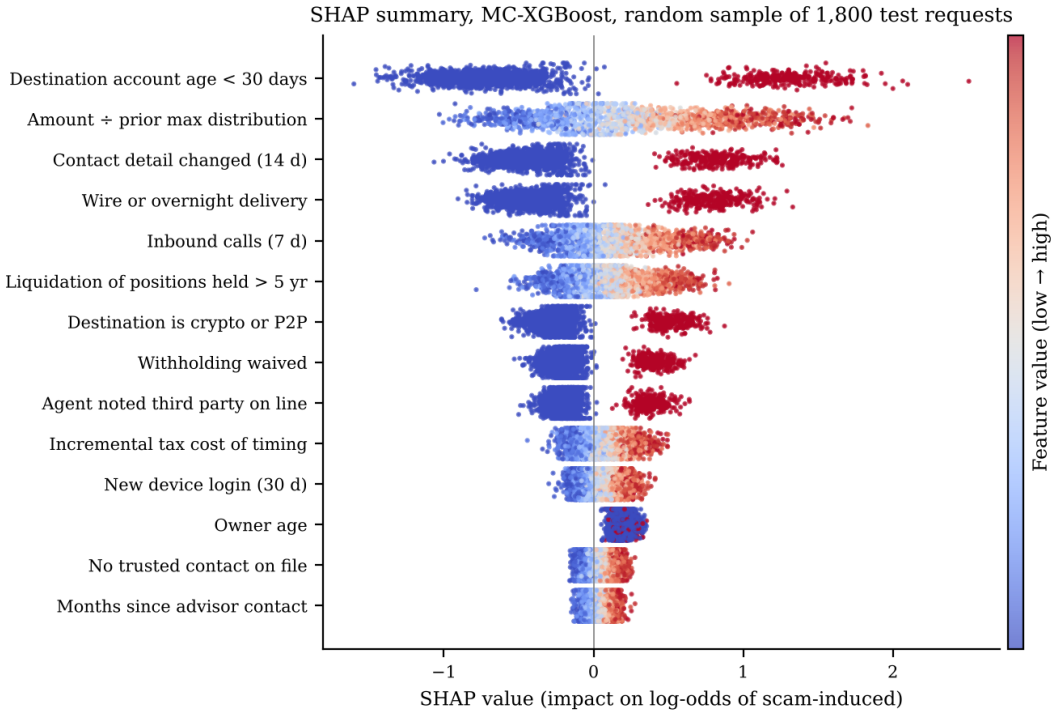


Figure 4. SHAP summary plot for the 14 highest-ranked features of the monotone-constrained model. Each point is one request. Horizontal position shows the contribution to the log-odds of a scam-induced label, and colour shows the feature value. Features are ordered by mean absolute SHAP value.

Table 4. Global importance and stability of the leading predictors (MC-XGBoost)

Feature	Mean |SHAP|	Rank (median, 200 refits)	Rank IQR	Top-10 frequency (%)	Constraint
Destination account age < 30 days	0.92	1	1–1	100	+
Amount ÷ prior max distribution	0.81	2	2–2	94	+
Contact detail changed (14 d)	0.58	3	3–3	87	+
Wire or overnight delivery	0.55	4	4–5	80	+
Inbound calls (7 d)	0.49	5	4–6	74	+
Liquidation of positions held > 5 yr	0.41	6	5–7	68	+
Destination is crypto or P2P	0.37	7	6–10	61	+
Withholding waived	0.29	8	7–11	54	+
Agent noted third party on line	0.27	9	7–12	48	+
Incremental tax cost of timing	0.24	10	8–14	28	+
New device login (30 d)	0.19	11	9–15	28	+
Owner age	0.17	12	10–16	28	none
No trusted contact on file	0.13	13	10–18	28	+
Months since advisor contact	0.11	14	11–19	28	none

Note. Rank stability from 200 bootstrap refits of the full pipeline. Constraint indicates whether a monotone increasing (+) constraint was imposed during training. IQR = interquartile range.

International Journal of Computer Applications Technology and Research
Volume 15–Issue 02, 72 – 83, 2026, ISSN: -2319–8656
DOI:10.7753/IJCATR1502.1006

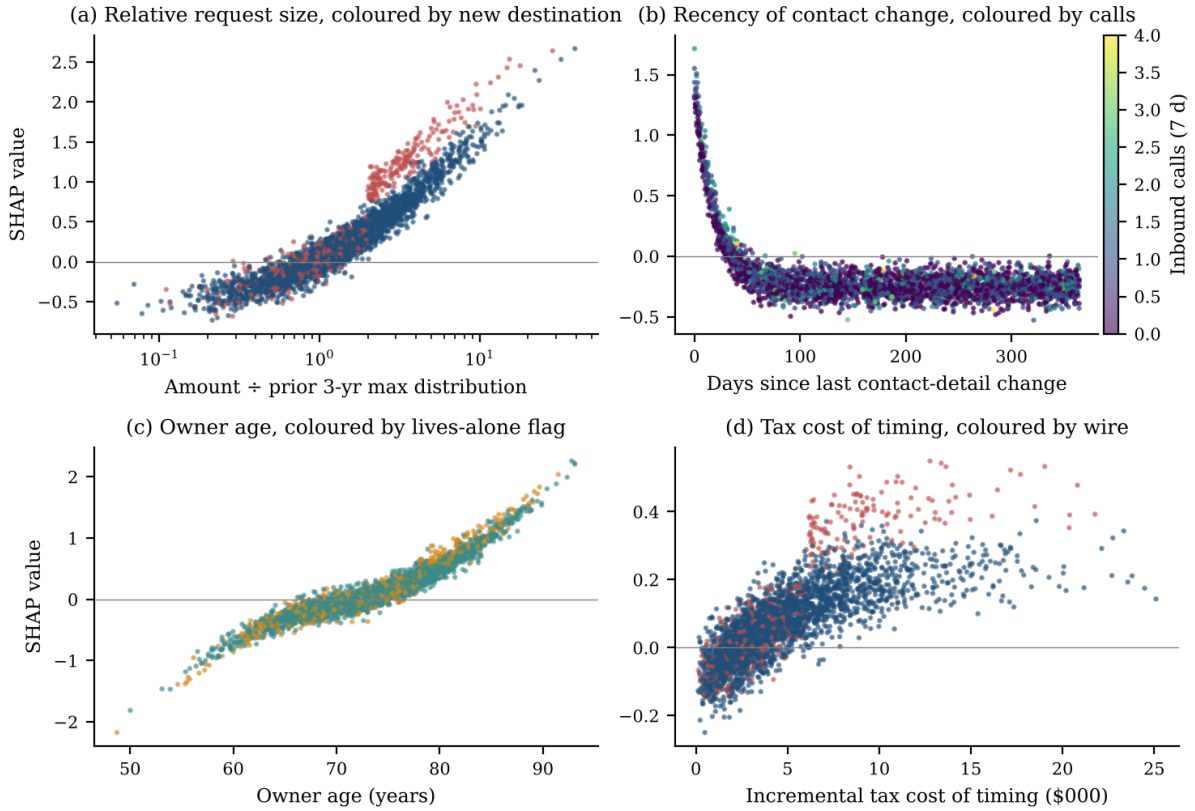


Figure 5. SHAP dependence plots for four predictors. Each panel plots a feature value against its SHAP contribution and colours points by the feature with which it interacts most strongly. (a) Risk rises with relative request size and rises further when the destination account is new (red). (b) A recent contact-detail change raises risk sharply for about three weeks and more so when inbound calls are frequent. (c) Age has a modest non-linear effect that is amplified for owners flagged as living alone (orange). (d) Requests that impose avoidable tax cost carry additional risk, especially when paired with wire delivery.

Four patterns stand out. First, destination novelty was the single strongest signal, and its effect was strongest when combined with large relative request size, consistent with investigator accounts that scammers have victims open new accounts or send funds to accounts the scammer controls. Second, contact-detail changes had a short half-life of roughly two to three weeks, which matches the tempo of grooming campaigns that culminate in a payment demand. Third, age itself contributed little once behaviour was observed. Fourth, the tax-cost feature captured a distinctive feature of scam requests, namely that victims are often instructed to ignore withholding and tax planning, producing distributions no informed adviser would recommend.

4.5 Ablation

Table 5. Feature-group ablation on the test set (MC-XGBoost)

Configuration	Features (n)	AUROC	ΔAUROC	Recall @1%	ΔRecall @1%
Full model	146	0.914	—	0.565	—
− Request attributes	21	0.868	−0.046	0.446	−0.119
− Account context	34	0.896	−0.018	0.523	−0.042
− Destination attributes	17	0.865	−0.049	0.436	−0.129
− Contact-channel activity	26	0.881	−0.033	0.481	−0.084
− Digital session telemetry	31	0.906	−0.008	0.549	−0.016
− Agent observations	17	0.900	−0.014	0.533	−0.032
Request + destination only	38	0.849	−0.065	0.404	−0.161
Behavioural groups only (c–f)	91	0.883	−0.031	0.488	−0.077

Note. Each row removes one feature group and retrains with the same hyperparameter search. Δ values are relative to the full model.

International Journal of Computer Applications Technology and Research
Volume 15–Issue 02, 72 – 83, 2026, ISSN: -2319–8656
DOI:10.7753/IJCATR1502.1006

4.6 Operating Points and Dollars Protected

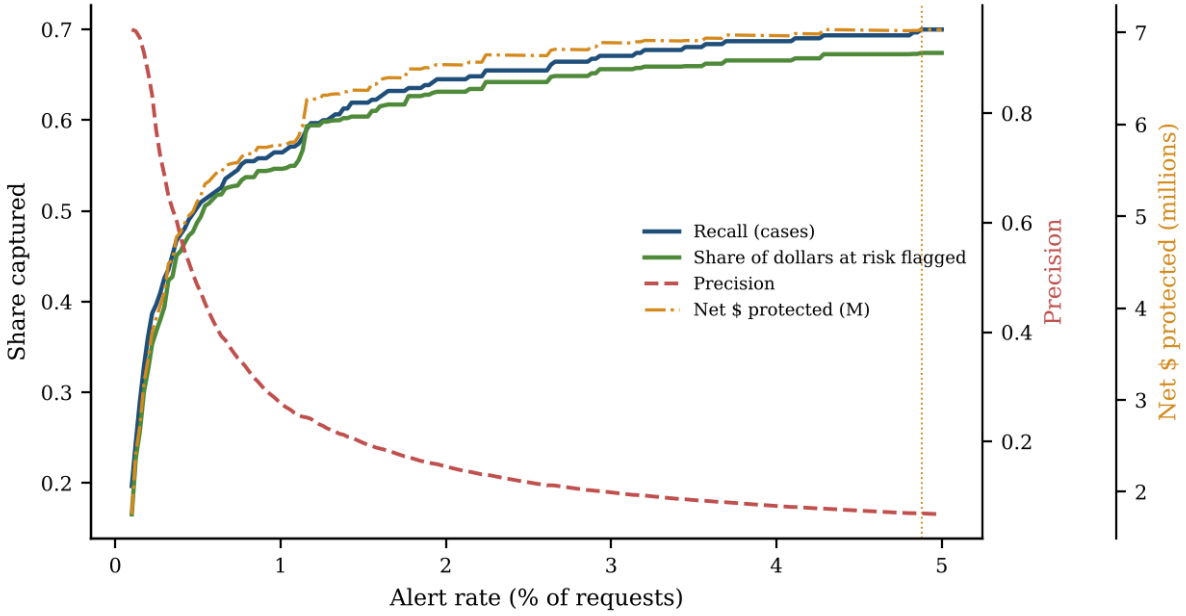


Figure 6. Operating characteristics of MC-XGBoost as a function of alert rate on the 2024 test set. Recall in cases and in dollars rise steeply up to about 2 percent and flatten thereafter, while precision falls. The dash-dot curve shows net dollars protected assuming a 62 percent interdiction success rate once a request is reviewed and a fully loaded review cost of 38 dollars per alert. Net benefit peaks at an alert rate of about 4.9 percent (dotted line).

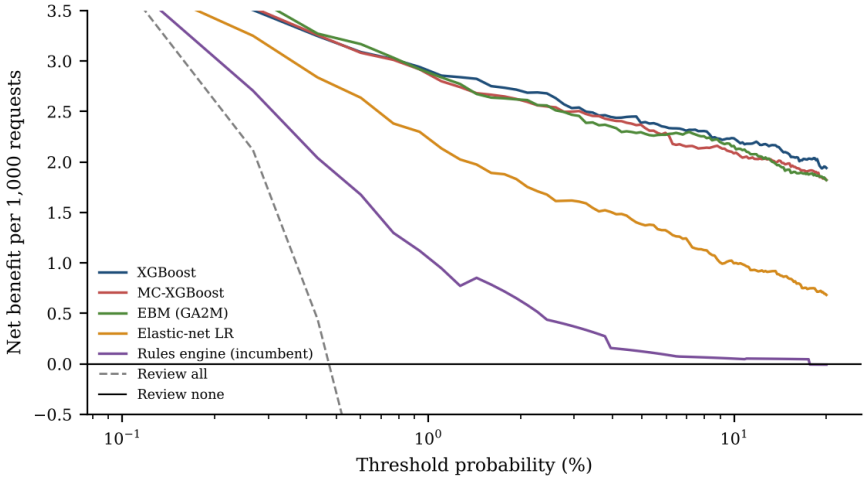


Figure 7. Decision curve analysis. Net benefit is expressed as true positives per 1,000 requests after subtracting false positives weighted by the odds of the threshold probability. The dashed grey line marks a policy of reviewing every request, and the black line marks reviewing none. Tree-based models dominate across thresholds between 0.3 and 20 percent.

International Journal of Computer Applications Technology and Research
Volume 15–Issue 02, 72 – 83, 2026, ISSN: -2319–8656
DOI:10.7753/IJCATR1502.1006

4.7 Subgroup Performance

Table 6. Performance of MC-XGBoost by owner subgroup at the 1 percent alert threshold

Subgroup	Requests	Scams	AUROC	TPR	FPR (%)	Alert rate (%)
Age band						
65–69	18,904	52	0.911	0.558	0.61	0.65
70–74	19,622	81	0.909	0.580	0.77	0.85
75–79	14,118	84	0.903	0.583	1.04	1.18
80+	12,146	93	0.896	0.570	1.29	1.55
Sex						
Female	34,410	192	0.905	0.578	0.92	1.02
Male	30,380	118	0.910	0.559	0.86	0.95
Rurality						
Urban/suburban	52,620	243	0.908	0.572	0.88	0.98
Rural	12,170	67	0.902	0.567	0.95	1.05
Account balance						
< $100k	20,114	78	0.897	0.551	0.74	0.83
$100k–$500k	29,882	149	0.910	0.584	0.89	0.99
> $500k	14,794	83	0.911	0.566	1.03	1.10

Note. TPR = true-positive rate (recall). FPR = false-positive rate. The equal-opportunity gap (max − min TPR) across age bands was 0.025 (bootstrap 95% CI −0.081 to 0.123).

True-positive rates were similar across subgroups, satisfying an equal-opportunity criterion within sampling error (Hardt, Price, and Srebro, 2016). False-positive rates rose with age, so owners aged 80 and older were roughly twice as likely as owners aged 65 to 69 to have a legitimate request flagged. Because a flag triggers a call rather than an automatic hold, the burden is a few minutes of conversation, but the gradient should be monitored given evidence that intrusive verification can feel paternalistic to older customers.

4.8 Stability over Time

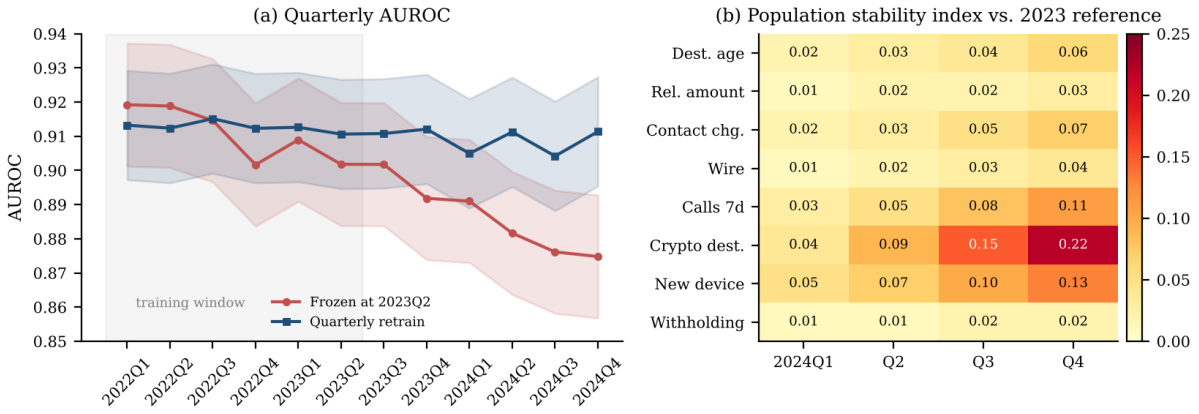


Figure 8. Temporal stability. (a) Quarterly AUROC for a model frozen after the training window compared with a model retrained each quarter on a rolling 30-month window, with bootstrap 95 percent bands. (b) Population stability index for eight key features in 2024 relative to 2023. Values above 0.10 indicate moderate drift and values above 0.20 indicate major drift. The cryptocurrency destination feature drifted most, coinciding with the growth of investment scams.

The frozen model lost about 0.03 AUROC over six quarters, while quarterly retraining held performance steady (Figure 7). Drift was concentrated in a few features linked to scam typology, consistent with an adversarial environment in which scammers adapt their payment instructions over time (Gama et al., 2014).

International Journal of Computer Applications Technology and Research
Volume 15–Issue 02, 72 – 83, 2026, ISSN: -2319–8656
DOI:10.7753/IJCATR1502.1006

4.9 Reviewer Experiment

Table 7. Reviewer decisions by information condition (24 analysts × 60 cases = 1,440 decisions)

Outcome	Case file only	+ Risk score	+ Score and SHAP	Mixed-model contrast (SHAP vs. score)
Accuracy, % (95% CI)	70.8 (67.4–74.1)	77.1 (73.9–80.1)	84.0 (81.2–86.6)	OR 1.62 (1.21–2.17), p = 0.001
Sensitivity, %	61.7	73.3	81.7	OR 1.71 (1.05–2.79), p = 0.031
Specificity, %	73.9	78.3	84.8	OR 1.58 (1.13–2.21), p = 0.008
Median review time, s (IQR)	312 (228–431)	268 (196–372)	209 (151–290)	Ratio 0.78 (0.72–0.85), p < 0.001
Over-reliance rate*, %	—	41.2	29.6	OR 0.60 (0.38–0.95), p = 0.029
Confidence (1–7), mean	4.3	4.9	5.4	β 0.47 (0.31–0.63), p < 0.001
Trust in tool (1–7), mean	—	4.6	5.5	β 0.88 (0.66–1.10), p < 0.001

Note. *Over-reliance is the share of cases in which the tool's top-decile score was wrong and the analyst followed it. Contrasts from logistic or linear mixed models with random intercepts for analyst and case and fixed effects for block order. OR = odds ratio.

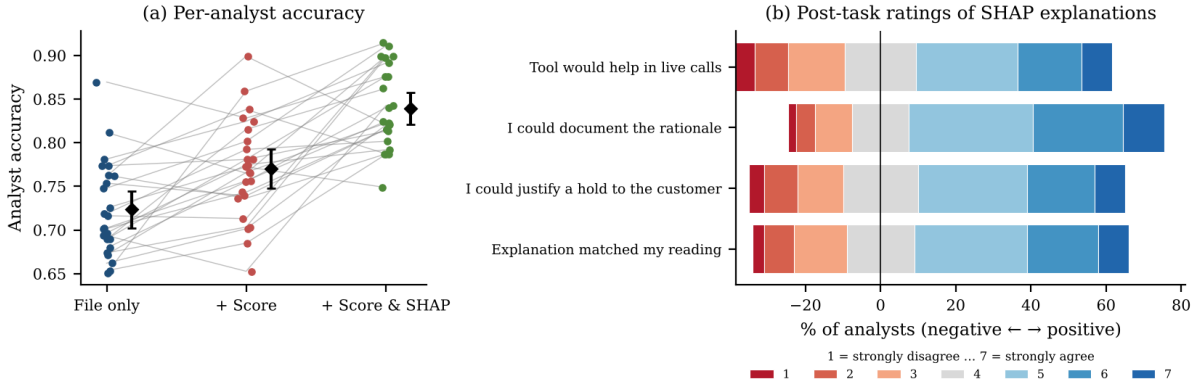


Figure 9. Reviewer experiment results. (a) Accuracy of each of 24 analysts in each condition, with grey lines linking the same analyst and black diamonds showing condition means with 95 percent intervals. (b) Diverging stacked bars of post-task agreement ratings on a seven-point scale for the explanation condition, centred on the neutral category.

Explanations improved accuracy for 21 of 24 analysts. Importantly, they reduced over-reliance. When the model assigned a top-decile score to a legitimate request, analysts who saw the explanation were less likely to follow it, often because the listed drivers (for example, a new destination account) had a benign explanation visible elsewhere in the file, such as a documented home purchase. This contrasts with laboratory findings that explanations increase blind acceptance (Bansal et al., 2021) and may reflect the domain expertise of professional reviewers.

5. DISCUSSION

Scam-induced withdrawals are rare, costly, and hard to see from any single attribute. Our results show that a gradient-boosted model trained on routinely collected custodial data can rank these requests far better than an incumbent rules engine, and that much of the gain comes from interactions among destination, timing, and contact-channel signals. The finding that age contributes little once behaviour is observed is reassuring from a fairness perspective, because it means the model is not simply flagging older customers.

The practical value of the model depends on what happens after an alert. In our data, 62 percent of reviewed scam-induced requests were successfully interdicted, mostly through a conversation with the owner or a trusted contact. If the 2024 deployment had replaced the incumbent rules at a 1 percent alert rate, our back-of-envelope estimate suggests a recovery of 9 to 13 million dollars in prevented losses for the test year, before accounting for the owners whose scams would continue through other channels. That displacement problem is real. A blocked IRA distribution may be followed by a home-equity loan or a bank transfer, so custodial detection works best as part of a broader network that includes banks, trusted contacts, and adult protective services.

Explainability was not merely a compliance ornament. Reviewers made faster and better decisions when they could see why a request was flagged, and the monotone constraints ensured that explanations never contradicted investigator intuition, for example by suggesting that a newer destination account lowered risk. Given that the constrained model sacrificed only a few thousandths of AUROC, we recommend constrained or additive models for deployment in this setting, in line with arguments for inherently interpretable models in high-stakes decisions (Rudin, 2019).

5.1 Limitations

Several limitations apply. Labels depend on detection, so scams that were never reported appear as negatives, and measured precision is therefore a lower bound. The data come from a single custodian whose clientele may differ from that of bank-affiliated or advisor-centric firms. Agent observations depend on agent training and may vary in quality. The reviewer experiment used historical cases in a controlled setting, and analysts answered prompts about a live call rather than an actual one, so effects in live operations may be smaller. Finally, scammers adapt, and a published description of predictive features could in principle inform evasion, which is one reason we report feature groups and directions rather than exact thresholds.

5.2 Policy Implications

- Custodians should collect trusted contact information proactively, because the absence of a trusted contact removed the most effective interdiction channel in our data.
- Regulators could clarify expectations for model explanations that accompany a Rule 2165 hold, so that firms document the factors behind each hold consistently.
- Industry information sharing on destination accounts linked to confirmed scams would strengthen the single most predictive feature group.
- Models should be retrained at least quarterly with formal drift monitoring, given the measurable decline of a frozen model within 18 months.

6. CONCLUSION

An interpretable gradient-boosted model can identify a majority of scam-induced IRA distribution requests from older investors while reviewing only 1 to 2 percent of all requests. Monotone constraints and SHAP explanations make the model's reasoning legible to the analysts who must act on it, and those explanations measurably improve human decisions. The tools to protect older investors from authorized-but-deceived withdrawals exist, and the remaining work lies in deploying them with care for both protection and autonomy.

7. DECLARATIONS

Funding. This work was supported by an internal research grant from the Center for Financial Security and Aging.

Conflicts of interest. O.O.F. serves on an unpaid advisory council on elder financial protection. The other authors declare none.

Data availability. The administrative data are confidential under the data use agreement. Code for feature construction and evaluation is available from the corresponding author, together with a demonstration file that mirrors the variable structure without containing any customer records.

Ethics. The study received exempt determination for the secondary analysis and full approval for the reviewer experiment.

8. REFERENCES

[1] Acierno, R., Hernandez, M. A., Amstadter, A. B., Resnick, H. S., Steve, K., Muzzy, W., & Kilpatrick, D. G. (2010). Prevalence and correlates of emotional, physical, sexual, and financial abuse and potential neglect in the United States: The National Elder Mistreatment Study. *American Journal of Public Health, 100*(2), 292–297.

[2] Bahnsen, A. C., Aouada, D., Stojanovic, A., & Ottersten, B. (2016). Feature engineering strategies for credit card fraud detection. *Expert Systems with Applications, 51*, 134–142.

[3] Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). ACM.

[4] Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical Science, 17*(3), 235–255.

[5] Boyle, P. A., Yu, L., Schneider, J. A., Wilson, R. S., & Bennett, D. A. (2019). Scam awareness related to incident Alzheimer dementia and mild cognitive impairment: A prospective cohort study. *Annals of Internal Medicine, 170*(10), 702–709.

[6] Burnes, D., Henderson, C. R., Jr., Sheppard, C., Zhao, R., Pillemer, K., & Lachs, M. S. (2017). Prevalence of financial fraud and scams among older adults in the United States: A systematic review and meta-analysis. *American Journal of Public Health, 107*(8), e13–e21.

[7] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).

[8] Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *Annals of Internal Medicine, 162*(1), 55–63.

[9] Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015). Calibrating probability with undersampling for unbalanced classification. In *2015 IEEE Symposium Series on Computational Intelligence* (pp. 159–166).

[10] DeLiema, M., Shadel, D., & Pak, K. (2020). Profiling victims of investment fraud: Mindsets and risky behaviors. *Journal of Consumer Research, 46*(5), 904–914.

[11] DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics, 44*(3), 837–845.

[12] Federal Bureau of Investigation. (2024). *Elder fraud report 2023*. Internet Crime Complaint Center.

[13] Federal Trade Commission. (2024). *Consumer Sentinel Network data book 2023*.

[14] FINRA. (2018a). *Rule 2165. Financial exploitation of specified adults*. Financial Industry Regulatory Authority.

[15] FINRA. (2018b). *Rule 4512. Customer account information*. Financial Industry Regulatory Authority.

[16] Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys, 46*(4), Article 44.

[17] Gamble, K. J., Boyle, P. A., Yu, L., & Bennett, D. A. (2015). Aging and financial decision making. *Management Science, 61*(11), 2603–2610.

[18] Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems 35, Datasets and Benchmarks Track*.

[19] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 29* (pp. 3315–3323).

[20] James, B. D., Boyle, P. A., & Bennett, D. A. (2014). Correlates of susceptibility to scams in older adults without dementia. *Journal of Elder Abuse & Neglect, 26*(2), 107–122.

[21] Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (2020). Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In

Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1–14). ACM.

[22] Lichtenberg, P. A., Stickney, L., & Paulson, D. (2013). Is psychological vulnerability related to the experience of fraud in older adults? *Clinical Gerontologist, 36*(2), 132–146.

[23] Lou, Y., Caruana, R., Gehrke, J., & Hooker, G. (2013). Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 623–631).

[24] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*(1), 56–67.

[25] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30* (pp. 4765–4774).

[26] Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning* (pp. 625–632).

[27] Nori, H., Jenkins, S., Koch, P., & Caruana, R. (2019). InterpretML: A unified framework for machine learning interpretability. *arXiv preprint* arXiv:1909.09223.

[28] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144).

[29] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215.

[30] Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE, 10*(3), e0118432.

[31] Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making, 26*(6), 565–574.