Machine Learning-Based Claims Fraud Triage in Guidewire ClaimCenter: A Comparative Validation Study

Garima Sharma

Independent Researcher, Arlington Heights, Illinois, USA

DOI: https://doi.org/10.51584/IJRIAS.2026.11100135

Received: 07 November 2026; Accepted: 12 November 2026; Published: 19 November 2026

ABSTRACT

Insurers that operate Guidewire ClaimCenter increasingly attach machine-learning fraud scores to first notice of loss (FNOL), yet published benchmarks rarely test whether the reported accuracy holds up under validation that respects the data's structure. This study evaluates claim-fraud triage on a 2024 open U.S. auto insurance claims dataset and asks three questions: how far conventional cross-validation overstates performance, whether complex learners outperform parsimonious and rule-based scoring, and what a score is worth to a special investigation unit (SIU). An audit showed that the 8,000 labeled claims are 1,000 distinct claim profiles, each replicated eight times with perturbed numeric fields and an identical label. Random five-fold cross-validation produced an area under the ROC curve (AUC) of 0.9999 for gradient boosting, whereas profile-grouped validation reduced it to 0.862 (95% CI 0.843–0.881), an inflation of 0.138. Logistic regression (0.852), random forest (0.849), and an isotonic-calibrated boosting model (0.858) performed similarly, and a two-variable logistic model using only incident severity and insured hobby reached 0.885 (95% CI 0.868–0.902).

All learned models outperformed an a priori rule-based score (AUC 0.771). On a 1,600-claim hold-out, referring 20% of claims captured 59–63% of fraudulent claims and 61–65% of fraudulent claim value, against 18% and 20% for random selection. Across 20 repeated holdouts, capture at 20% referral averaged 47–59%, and dropping the hobby field cut AUC from 0.865 to 0.739. Referring nearly every claim paid best only when costs were low and fraud was common; with higher costs or a fraud rate of 5–10%, a ranked queue of 20–40% of claims earned more. The audit also found that 99.5% of records violate claim-component additivity, which supports FNOL validation rules. The findings concern a replicated benchmark and should not be read as production performance.

Keywords: Claims fraud detection; Guidewire ClaimCenter; Gradient boosting; Group-aware validation; Cost-sensitive triage

INTRODUCTION

Insurance fraud is a persistent cost to the U.S. insurance sector. The Coalition Against Insurance Fraud has estimated its annual cost to the U.S. economy at roughly US$308.6 billion across all lines of business [6]. Property and casualty carriers rely on core claims platforms to enforce workflow at FNOL, and Guidewire ClaimCenter is a widely deployed platform of this kind [10]. A claims platform manages intake, exposure creation, assignment rules, and referral to an SIU, and a predictive score computed at FNOL can inform each of these decisions. Machine learning for fraud scoring has been studied since early comparisons of classifiers on automobile claims [23], and gradient-boosted trees are now the default choice for tabular data [5, 9, 14].

Two problems limit how far the resulting literature helps platform owners. The first concerns validation: when records are not independent, random cross-validation inflates reported metrics [12, 13, 19]. The second concerns the decision context: a score creates value through the referral policy it supports, and that policy depends on investigation cost and unit capacity rather than on AUC alone [8, 24]. This study addresses both problems with a recent open dataset published in 2024 [16]. Three research questions guide the work. RQ1 asks how much conventional random cross-validation overstates fraud-detection performance when the data

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue X October 2026

contain replicated claim profiles. RQ2 asks whether gradient boosting outperforms logistic regression, a random forest, a calibrated boosting model, a two-variable parsimonious model and a transparent rule-based score once validation is grouped by profile. RQ3 asks what a fraud score is worth for SIU referral when the unit has limited capacity, and how that value changes when capacity is unconstrained.

The contributions are fourfold: a documented data-integrity audit of the dataset; a quantification of validation leakage that separates apparent from credible performance; evidence that a two-variable model is competitive with complex learners on this benchmark; and a capacity-based referral analysis together with a ClaimCenter integration design whose validation rules derive from the audit. Generative AI and large language models (LLMs) are increasingly proposed for claims handling [4], but the dataset contains no free-text narratives.

This study, therefore, evaluates supervised machine learning on structured FNOL attributes and treats LLM-based extraction as an extension (Section 6.4). The study also assesses how stable the results are across repeated splits, how much the models depend on the hobby field and on fields that may not be apparent at first glance, and how the value of a score changes with referral capacity, cost, and fraud rate (Sections 4.5 and 5.7 to 5.9).

LITERATURE REVIEW

Fraud Scoring in Automobile Claims

Viaene et al. [23] benchmarked a range of classification techniques on expert-labeled automobile claims and reported that several techniques, including comparatively simple ones, reached similar performance. Šubelj et al. [22] showed that relational information, analyzed with social network methods, can reveal organized automobile fraud that attribute-level models miss. Boosted ensembles [9], and their scalable implementations [5, 14], have since become standard baselines.

Hand [11] argued that the advantage of sophisticated classifiers over simple ones is frequently smaller in practice than benchmark comparisons suggest, and Rudin [20] argued that interpretable models deserve preference in high-stakes decisions. These positions motivate the explicit inclusion of parsimonious and rule-based comparators in this study.

Evaluation Validity and Leakage

Kaufman et al. [13] define leakage as the introduction of information about the prediction target that should not legitimately be available, and Kapoor and Narayanan [12] document how leakage undermines reproducibility across machine-learning-based research. Roberts et al. [19] show that data with hierarchical, spatial, or temporal structure require blocked cross-validation because random folds place dependent observations on both sides of a split. Replicated claim profiles are a hierarchical structure of exactly this kind: a replica in the training fold carries the label of its sibling in the test fold.

Cost-Sensitive Decisions, Calibration, and Governance

Elkan [8] shows that optimal decisions under asymmetric costs follow from calibrated probabilities and a cost-determined threshold, and Vickers and Elkin [24] frame the same idea as net benefit. Precision–recall analysis is more informative than ROC analysis under class imbalance [21], and the quality of the probability estimates can be assessed with the Brier score [3] and improved by post hoc calibration [17].

On the governance side, the National Association of Insurance Commissioners has issued a model bulletin on insurers' use of algorithms and predictive models [15], and Barocas and Selbst [1] describe how seemingly neutral features can act as proxies for protected characteristics. No prior study located by this research combines a leakage audit, grouped validation, and a capacity-based referral analysis on a 2024 U.S. auto claims dataset, which defines the gap addressed here.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue X October 2026

DATA AND AUDIT

Source and Structure

The analysis uses the open Zenodo record titled "Auto Insurance Claims," published in August 2024 [16]. The dataset holds 10,000 rows and 40 columns, one of which is empty.

Attributes cover the policy (state, combined single limit, deductible, premium, umbrella limit), the insured (demographics, occupation, hobby), the incident (type, collision type, severity, authorities contacted, state, hour, vehicles, property damage, bodily injuries, witnesses, police report), the claim (total, injury, property and vehicle amounts), the vehicle (make, model, year) and the binary label fraud_reported. The record does not document how the data were generated.

Table 1 summarizes the audit. Of the 10,000 rows, 1,000 are blank except for a fraud label, and 1,000 carry attributes but no label, leaving 8,000 labeled claims, of which 1,976 (24.7%) are fraudulent. Within the labeled rows, the 17 categorical attributes form exactly 1,000 distinct combinations, each occurring exactly eight times, and no combination carries conflicting labels.

Numeric fields and dates differ between replicas. This study refers to each combination as a claim profile. Only 31 policy numbers are duplicated, so de-duplication on policy numbers would not reveal the structure. The replication is an empirical observation, and no claim is made about its origin.

Table 1. Audit summary of the Zenodo auto insurance claims file [16].

Item	Value
Rows × columns in file	10,000 × 40
Blank rows carrying only a label	1,000
Rows with attributes but no label (excluded)	1,000
Labelled claims analyzed	8,000
Fraudulent claims (share)	1,976 (24.7%)
Distinct claim profiles × replicas each	1,000 × 8
Profiles with conflicting labels	0
Duplicated policy numbers	31
Incident years	2016–2024
Features after engineering	43 (17 categorical, 26 numeric)

Data-Quality Violations

Table 2 lists rule violations among the 8,000 labeled claims. The most pervasive is the failure of claim-component additivity: in 99.5% of records, the sum of injury, property, and vehicle claims differs from the total claim by more than US$1, with a standard deviation of the gap of roughly US$700-US$800. A further 16.4% of records report a vehicle model year later than 2025. Such defects are precisely what FNOL validation rules in a claims platform are designed to catch, and Section 6.2 maps them to rule candidates.

Table 2. Rule violations among 8,000 labeled claims.

Check	Records	Share
Claim components do not sum to total (>US$1)	7,961	99.5%
Vehicle model year later than 2025	1,312	16.4%
Vehicle model year earlier than 1950	39	0.5%
Negative umbrella limit	8	0.1%
Incident hour recorded as 24	3	<0.1%
Incident state differs from policy state	7,928	99.1%
Property damage recorded as "?"	2,880	36.0%
Police report availability recorded as "?"	2,744	34.3%
Incident dated before policy bind date	0	0.0%

METHODOLOGY

Preparation and Features

The 1,000 blank and 1,000 unlabeled rows were removed. The policy number, ZIP code, incident street location, and raw timestamps were excluded as identifiers. Thirty-three original attributes were retained, and ten features were engineered: days from policy bind to incident, claim-to-premium ratio, vehicle and injury shares of the total claim, vehicle age at incident, a night-incident flag (22:00–05:59), net capital, the absolute component gap, incident month, and day of week. The token "?" and the absent authority value were treated as categories. Profiles were identified by the joint value of the 17 categorical attributes and used as the grouping key in every validation step.

Models

Five scorers were compared. Logistic regression (L2 penalty, C = 0.3, balanced class weights) and a random forest (60 trees, depth 12, minimum leaf 3) used one-hot encoding. Histogram gradient boosting (learning rate 0.05, 120 iterations, 15 leaves, minimum leaf 30, L2 = 1.0, native categorical handling) followed the leaf-wise design of modern boosting libraries [14]. A calibrated variant fitted isotonic regression [17] on out-of-fold boosting scores from an inner three-fold profile-grouped split.

A parsimonious logistic model used only incident severity and insured hobby. A transparent rule-based score, specified a priori from common SIU red flags and not tuned on labels, added points for major damage (+3), total loss (+1), unconfirmed police report (+0.5), claim above the 75th percentile (+1), claim-to-premium above the 75th percentile (+0.5), night incident (+0.5) and zero witnesses (+0.5). All models were implemented with scikit-learn 1.8 [18].

Validation Design

Conventional validation used a stratified random five-fold cross-validation. Credible validation used stratified group five-fold cross-validation repeated twice with different seeds (ten folds), so that all eight replicas of a profile fall in the same fold. A single profile-grouped 80/20 hold-out (1,600 claims, 200 profiles, 392 fraudulent claims) supported the referral analysis. Uncertainty in the hold-out was estimated using a cluster

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue X October 2026

bootstrap [7], which resamples profiles (300 resamples). Fold-level differences were tested with the Wilcoxon signed-rank test [25]; with ten folds, the smallest attainable two-sided p-value is 0.002.

Metrics and Referral Economics

Discrimination was measured using AUC, average precision (PR-AUC) [21], lift in the top decile, and recall in the top 20% of scores; probability quality was measured using the Brier score [3]. Referral value was assessed in two ways. The capacity view sets the referral rate to 10%, 20%, or 30% and reports the share of fraudulent claims and fraudulent claim dollars captured. The net-benefit view [8, 24] assumes an investigation cost C of US$400 per referral and that a correctly referred fraudulent claim avoids a fraction ρ of its amount A (ρ = 0.5 in the base case). A claim is worth referring when its fraud probability p exceeds the break-even level $p^* = C / (\rho A)$. The referral rate maximizing net benefit was chosen based on out-of-fold training scores, then applied to the hold-out set, and a sensitivity grid varied ρ over {0.3, 0.5, 0.7} and C over {US$200, US$400, US$800}. These cost parameters are illustrative assumptions, not carrier data.

Added Checks on Stability, Fairness, and Cost

Four checks were added to assess how much the main results depend on a single split, on the copying in the file, and on a single set of cost figures. First, one claim was drawn at random from each of the 1,000 profiles. This yielded 1,000 claims (247 fraudulent), with no duplicate profiles. Models were scored on these claims with two repeats of stratified five-fold cross-validation, and the draw was repeated five times. This check does not replace a second dataset, but it shows whether the copying itself drives the results. Second, the profile-grouped 80/20 hold-out was repeated 20 times with different seeds, and the mean, standard deviation, and range of each metric were recorded. Third, boosting was refitted with groups of fields removed. The hobby field was dropped alone. Then all nine personal fields were dropped (sex, education, occupation, hobby, relationship, age, capital gains, capital loss, and net capital). Then the nine fields that may be known only after the claim is reviewed were dropped (incident severity, the four claim amounts, and four ratios or gaps built from them). Last, both groups were dropped together. Which fields are available at first notice varies by carrier, so this grouping is an assumption. For the full, no-hobby, and no-personal models, the share of each sex and age group sent to referral at 20% capacity was recorded, along with the share of each group's fraudulent claims that was caught. The link between the hobby field (chess or CrossFit, compared with all others) and the other personal fields was measured using Cramér's V for one claim per profile. Fourth, the cost analysis was widened. Referral capacity ranged from 2% to 100% of claims. Five cost cases were used. They were a flat US$400 per referral (the base case), a flat US$1,500, a flat US$800 plus US$300 for each legitimate claim that is referred (to stand for delay and customer handling), US$250 plus 0.5% of the claim amount plus the same US$300, and a flat US$2,500 plus US$500. The share of a fraudulent claim that is saved, ρ, was set to 0.3 and 0.5. The fraud rate of 24.7% in this file may be higher than in real books of business, so fraudulent claims were also given lower weights to bring the rate to 10% and 5%. Net benefit per claim is the money saved on fraudulent claims, minus investigation costs, divided by the number of claims. Results were averaged over the 20 hold-outs, and the best capacity was read from the average curve, so the gains are best-case figures. All cost figures are assumptions and were not taken from a carrier.

Reproducibility

All steps can be repeated from the public file. Rows with no policy number (1,000) or no fraud label (1,000) were dropped. The text "?" and the empty authority field were kept as categories. The ten added features follow Section 4.1, and profiles were defined from the 17 categorical fields. Categorical fields were one-hot encoded for logistic regression and random forests, passed as category codes to boosting, and numeric fields were scaled for logistic regression only. Random seeds were fixed for every split. The scripts, a README with the order in which to run them, the software versions, and the result files are listed in the Data Availability section. The packaged scripts, run with scikit-learn 1.9.1 instead of 1.8, reproduce the reported AUC values to within about 0.01 and the capture shares to within about one percentage point.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue X October 2026

RESULTS

Descriptive Structure of Fraud

Fraud prevalence varies sharply with incident severity: 60.5% for major damage (2,208 claims), 10.7% for minor damage (2,832), 12.9% for total loss (2,240), and 6.7% for trivial damage (720). Among hobbies, chess (82.6%) and cross-fit (74.3%) stand out. Table 3 shows that the two variables together nearly determine the label: major-damage claims are about 60% fraudulent regardless of hobby, whereas non-major claims are 2–5% fraudulent unless the insured lists chess or cross-fit, in which case 80–92% are fraudulent. Fraudulent claims average US$60,279, compared with US$50,288 for legitimate claims, and fraud accounts for US$119.1 million of the US$422.0 million claimed (28.2%). This near-deterministic structure is unusual for operational claims data.

Table 3. Fraud rate (%) by incident severity and hobby group (n in parentheses).

Severity	Chess or cross-fit	Other hobbies
Major damage	60.0 (160)	60.5 (2,048)
Minor damage	92.3 (208)	4.3 (2,624)
Total loss	80.0 (240)	4.8 (2,000)
Trivial damage	80.0 (40)	2.4 (680)

Validation Leakage (RQ1)

Table 4 compares random and profile-grouped cross-validation. Random folds yielded an AUC of 0.9999 for gradient boosting and 0.998 for the random forest because replicas of a test claim are present in the training data; grouped folds reduced these to 0.862 and 0.849, respectively.

Logistic regression, which cannot memorize replicas as easily, was inflated less (0.933 versus 0.852). PR-AUC fell from 0.9999 to 0.656 for boosting. The reported accuracy of this file, therefore, depends mainly on the validation scheme.

Table 4. Random versus profile-grouped cross-validation (AUC and PR-AUC).

Model	AUC random	AUC grouped	AUC inflation	PR-AUC grouped
Gradient boosting	0.9999	0.862	+0.138	0.656
Random forest	0.998	0.849	+0.150	0.561
Logistic regression	0.933	0.852	+0.081	0.622

Model Comparison Under Grouped Validation (RQ2)

Table 5 reports grouped ten-fold results with 95% t-intervals across folds, which are descriptive because folds share data. Boosting led the full-feature models (AUC 0.862) but exceeded logistic regression by only 0.010 (Wilcoxon $p = 0.084$) and the random forest by 0.013 ($p = 0.049$). Isotonic calibration improved the Brier score from 0.132 to 0.119 but reduced PR-AUC by 0.044 ($p = 0.002$), reflecting tied scores after the stepwise mapping. The two-variable logistic model achieved an AUC of 0.885 (95% CI 0.868–0.902) and a PR-AUC of 0.687 using the same folds, so its interval overlaps those of the full models; the supportable conclusion is

that added complexity brought no measurable gain, not that the small model is superior. Every learned model beat the rule-based score (AUC 0.771; boosting minus rule 0.091, p = 0.002). Figure 1 shows the pooled out-of-fold ROC curves and reliability diagram.

Table 5. Profile-grouped cross-validation, 2 × 5 folds (mean; 95% CI for AUC).

Model	AUC (95% CI)	PR-AUC	Brier	Lift@10%	Recall@20%
Logistic regression	0.852 (0.834–0.869)	0.622	0.143	2.74	0.533
Random forest	0.849 (0.836–0.861)	0.561	0.130	2.22	0.482
Gradient boosting	0.862 (0.843–0.881)	0.656	0.132	2.75	0.539
Boosting, calibrated	0.858 (0.839–0.876)	0.612	0.119	2.65	0.530
Rule-based score	0.771 (0.756–0.785)	0.503	—	2.42	0.472
Two-variable logistic	0.885 (0.868–0.902)	0.687	—	—	—

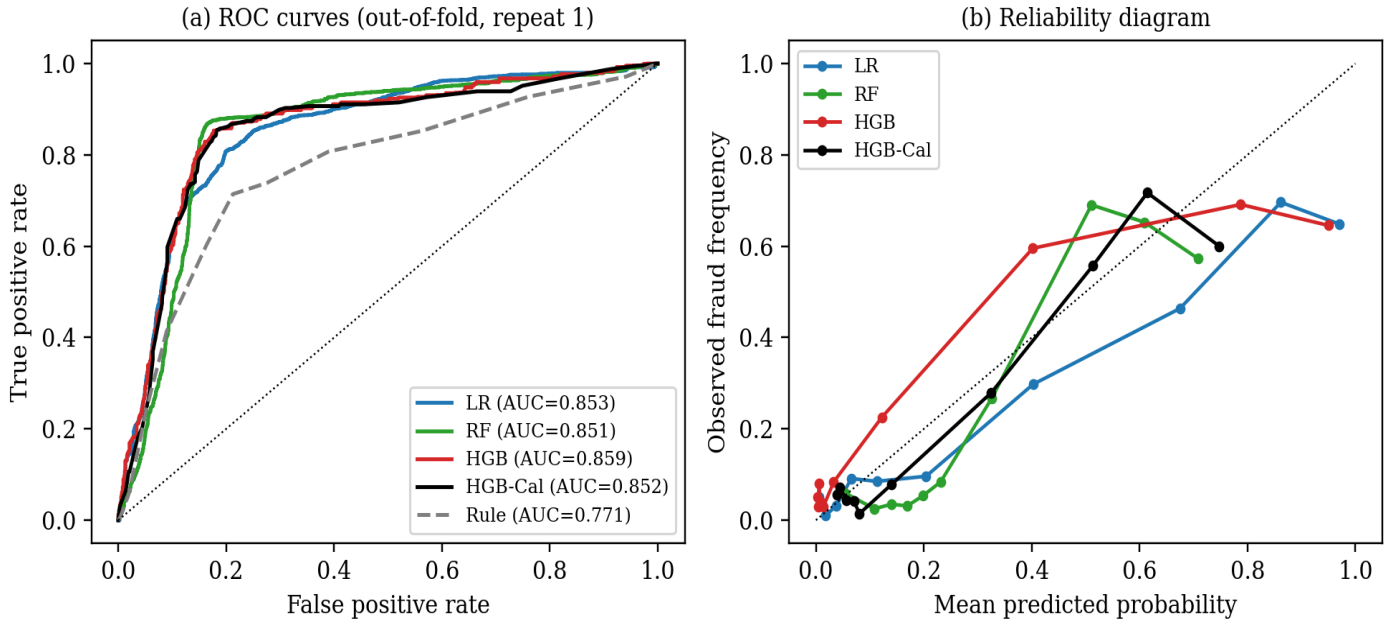


Figure 1. Pooled out-of-fold ROC curves (a) and reliability diagram (b) under profile-grouped validation, first repeat.

Feature Importance

Permutation importance on the hold-out (eight repeats) is concentrated in two variables (Figure 2). Shuffling incident severity lowers AUC by 0.285 and shuffling hobby by 0.140, whereas all other features lower it by less than 0.002, including claim amount, witnesses, relationship, and education.

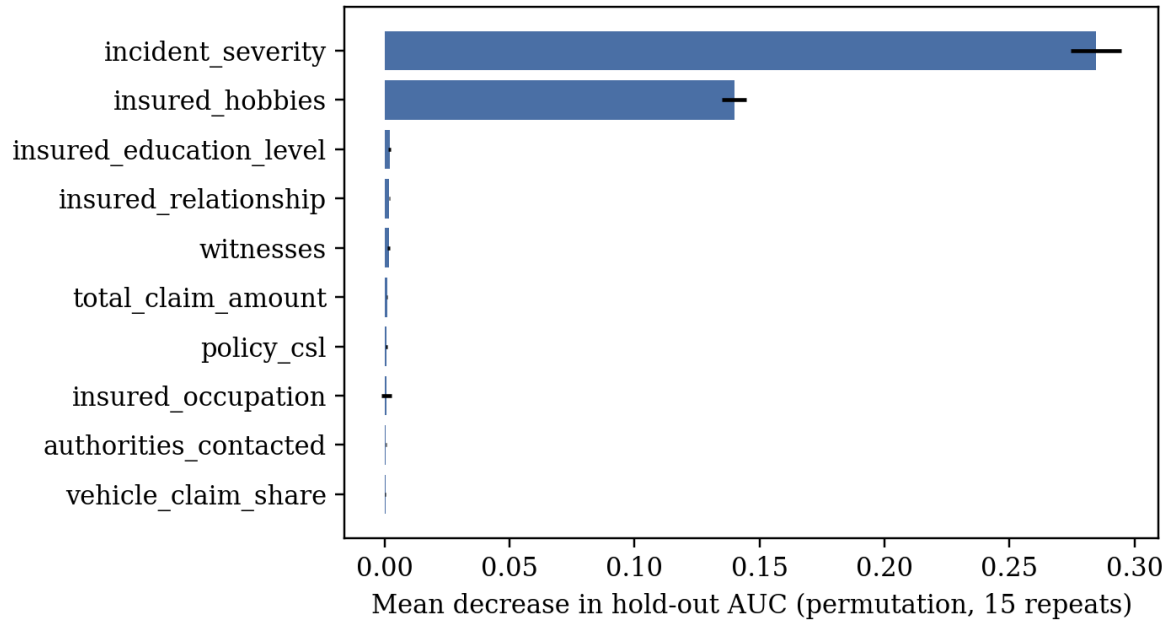


Figure 2. Permutation importance of the ten most influential features (hold-out, boosting).

Hold-Out and Capacity-Constrained Referral (RQ3)

On the hold-out, AUC was 0.870 for the calibrated boosting model (cluster-bootstrap 95% CI 0.796–0.924), 0.879 for uncalibrated boosting, 0.891 for logistic regression (0.816–0.943), and 0.770 for the rule score (0.685–0.848); the difference between calibrated boosting and logistic regression was −0.022 (−0.059 to 0.012). Table 6 shows capacity-constrained gains. Referring 20% of claims captured 59.2% of fraudulent claims with calibrated boosting and 63.0% with logistic regression, against 47.4% for the rule score and 18.1% for random selection. The bootstrap gain of boosting over the rule score at 20% was 12.0 percentage points in claim count (95% CI 1.6 to 22.9) but only 5.8 points in claim dollars (−7.4 to 19.1), because the rule score deliberately favors large claims. Figure 3 gives the full gain curves.

Table 6. Share (%) of fraudulent claims and of fraudulent claim dollars captured at fixed referral rates (hold-out, 1,600 claims).

Model	Claims 10%	Claims 20%	Claims 30%	US$ 10%	US$ 20%	US$ 30%
Boosting, calibrated	30.4	59.2	83.4	30.3	60.6	86.6
Boosting	30.6	59.2	81.6	31.5	60.6	85.1
Logistic regression	33.9	63.0	83.2	36.3	65.1	85.2
Rule-based score	21.4	47.4	70.2	29.5	55.2	78.1
Random selection	7.7	18.1	29.3	8.8	19.7	31.2

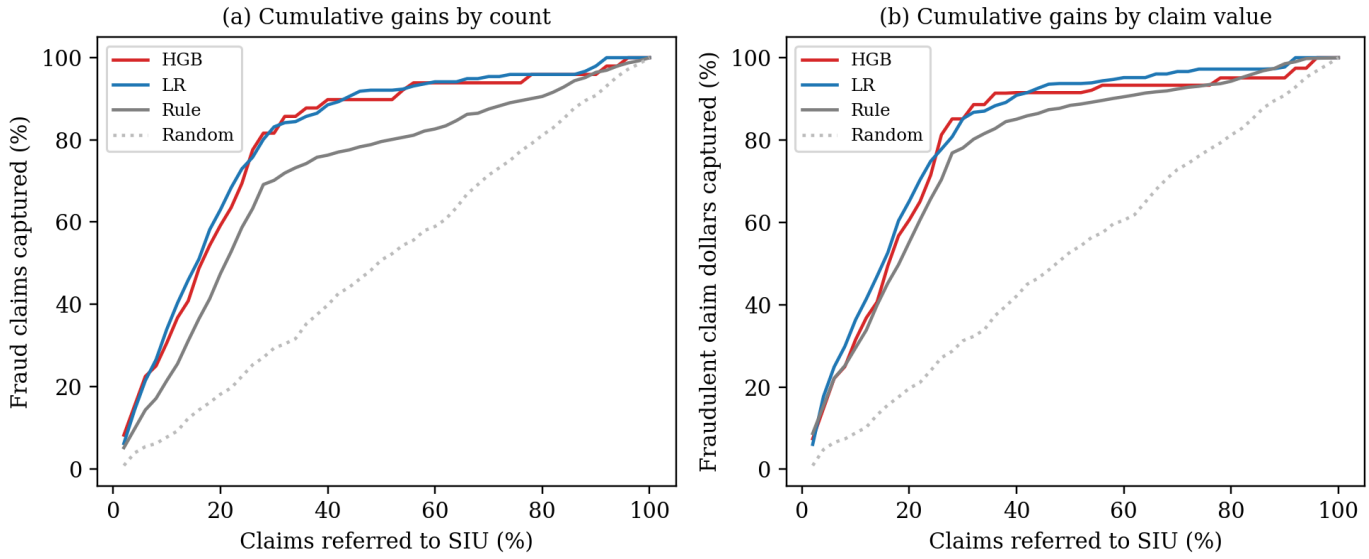


Figure 3. Cumulative gains by claim count (a) and by claim value (b) on the hold-out set.

Net Benefit when Capacity is Unconstrained

With a mean claim of US$52,756, ρ = 0.5, and C = US$400, the break-even probability is p* = 1.5%, below nearly every predicted probability, so the cost-optimal policy refers to almost every claim. The rate selected on training scores was 94%. On the hold-out, the net benefit at that rate was US$10.10 million for calibrated boosting, US$10.35 million for the rule score, US$10.38 million for logistic regression, US$10.34 million for referring all claims, and US$9.72 million for random selection.

Table 7 shows that across all nine cost scenarios, scoring-based triage at the cost-optimal rate did not exceed the cost of referring everything. A score, therefore, earns its value when the SIU cannot investigate every claim. This holds only for the low cost and high fraud rate used here; Section 5.9 shows that higher costs or a lower fraud rate change the answer.

Table 7. Hold-out net benefit (US$ million) at the cost-optimal referral rate, by recovery fraction ρ and investigation cost C.

ρ	C (US$)	Calibrated boosting	Rule-based score	Refer all claims
0.3	200	6.12	6.27	6.27
0.3	400	5.82	5.97	5.95
0.3	800	5.22	5.37	5.31
0.5	200	10.40	10.65	10.66
0.5	400	10.10	10.35	10.34
0.5	800	9.49	9.75	9.70
0.7	200	14.68	15.04	15.05
0.7	400	14.38	14.74	14.73
0.7	800	13.77	14.13	14.09

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue X October 2026

Stability: One Claim per Profile and Repeated Hold-Outs

Table 8 shows two checks on stability. When one claim was drawn from each profile, the 1,000 independent claims gave AUC values of 0.866 for boosting, 0.854 for logistic regression, 0.880 for the two-variable model, and 0.770 for the rule score. These are close to the grouped cross-validation values in Table 5 (0.862, 0.852, 0.885, and 0.771). The copying in the file therefore inflates scores under random folds but does not change the results once profiles are kept apart. Over 20 repeated hold-outs, boosting had a mean AUC of 0.865 with a standard deviation of 0.028, and its recall at 20% referral ranged from 45.7% to 64.3%. The first hold-out (59.2%) was on the favorable side of the average (53.8%). Boosting had a higher AUC than the rule score in all 20 splits and than logistic regression in 16, but the two were level on recall (boosting was higher in 9 of 20). The two-variable model had a higher AUC than boosting in 19 of 20 splits. The splits reuse the same 1,000 profiles, so these counts are consistent and not independent evidence.

Table 8. Stability checks: one claim per profile (1,000 claims) and 20 repeated profile-grouped hold-outs (1,600 claims each).

Model	AUC, one claim per profile (95% CI)	Recall at 20%	AUC, 20 hold-outs (SD)	Recall at 20%, 20 hold-outs (SD) [range]
Boosting	0.866 (0.858–0.873)	0.549	0.865 (0.028)	0.538 (0.046) [0.457–0.643]
Logistic regression	0.854 (0.846–0.862)	0.530	0.851 (0.029)	0.531 (0.048) [0.418–0.612]
Two-variable logistic	0.880 (0.873–0.887)	0.574	0.888 (0.025)	0.590 (0.050) [0.500–0.691]
Rule-based score	0.770 (0.760–0.780)	0.466	0.760 (0.045)	0.467 (0.038) [0.406–0.543]
Random selection	0.500	0.200	0.500	0.200

Dependence on Personal and Late-Review Fields

Table 9 shows the effect of boosting when groups of fields are removed. Dropping only the hobby field lowered mean AUC from 0.865 to 0.739 and recall at 20% from 53.8% to 44.5%. Dropping all nine personal fields gave almost the same result (0.745 and 44.9%), so no other personal field carried much of the signal. Dropping the nine fields that may be known only after review (severity, the four claim amounts, and four fields built from them) lowered AUC to 0.660 and recall to 34.7%. With both groups removed, AUC was 0.540 and recall was 19.4%, which is no better than random selection (20%). In this file, nearly all the usable signal is contained in two fields: one that describes the person and one that may not be apparent at first glance.

Table 9. Boosting with groups of fields removed (20 repeated profile-grouped hold-outs; mean and standard deviation).

Fields used in boosting	Mean AUC (SD)	Recall at 20% (SD)	Fraud dollars caught at 20%
All 43 fields	0.865 (0.028)	0.538 (0.046)	0.542
Without hobby (42 fields)	0.739 (0.039)	0.445 (0.044)	0.476
Without the 9 personal fields (34)	0.745 (0.045)	0.449 (0.044)	0.474
Without the 9 late-review fields (34)	0.660 (0.038)	0.347 (0.044)	0.335
Without both groups (25)	0.540 (0.057)	0.194 (0.059)	0.200
Random selection	0.500	0.200	0.200

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue X October 2026

Group-wise results at 20% capacity are as follows. With all fields, the model sent 17.3% of women and 23.3% of men to referral and caught 49.7% and 57.9% of their fraudulent claims, a gap of 8.2 points. Without the hobby field, the figures were 18.9% and 21.1% referred and 43.1% and 45.8% caught, a gap of 2.7 points. Among the holdouts, fraud was slightly more common among men (26.6%) than among women (23.0%), so a gap is expected. Age told a different story. The 46-and-over group was referred more often (24.8%) than the 31-to-45 group (18.0%), and this gap hardly changed when the hobby field (23.8% vs. 18.4%) or all personal fields (23.7% vs. 18.4%) were removed. Months as customer moves closely with age in this file (correlation 0.92), which is a likely reason. The hobby field was not tied to sex, education, or relationship status: chess or cross-fit appeared in 8.4% of women's profiles and 7.8% of men's (Cramér's V = 0.007), and the links to education (V = 0.063) and relationship status (V = 0.056) were weak. The link to the age band was small but nonzero (V = 0.082, p = 0.034). Among profiles, 79.0% of those with chess or CrossFit were fraudulent, compared with 19.9% of the rest.

Capacity and Cost Scenarios

Table 10 and Figure 4 widen the cost analysis. When the cost was US$400 per referral, ρ was 0.5, and the fraud rate was 24.7%; as in Section 5.6, referring every claim was best, and the ranked queue gained nothing. The answer changed as costs rose or the fraud rate fell. Take a fraud rate of 10%, a cost of US$800 plus US$300 for each legitimate claim referred, and ρ = 0.3. Referring all claims earned US$723 per claim. Sending only the top 30% by boosting score earned US$1,342, while a random 30% earned US$217.

At a 5% fraud rate, referring all claims lost US$189 per claim, and the top 20% earned US$546. A unit that could handle only 5% of claims would still earn US$374 per claim at a 10% fraud rate and US$205 per claim at a 5% fraud rate. Logistic regression and the rule score also beat referring everything in these cases, with smaller gains (US$521 and US$399 per claim at the 10% fraud rate). With ρ = 0.5, the gains were of similar size (US$521 per claim for the same cost case).

The curves in Figure 4 are fairly flat near the top: at a 10% fraud rate, any capacity from 20% to 50% of claims earned between US$1,161 and US$1,342 per claim, so an SIU can choose a capacity close to its staffing level without losing much. Because the best capacity was read from the average of the 20 hold-outs, the gains in Table 10 are best-case values.

Table 10. Gain of boosting over referring every claim (US$ per claim, ρ = 0.3) at three fraud rates. The share of claims referred at the best point is in brackets. A dagger marks cases in which referring every claim lost money.

Cost case	Fraud rate 24.7%	Fraud rate 10%	Fraud rate 5%
Flat US$400	0 (100%)	133 (30%)	215 (30%)
Flat US$1,500	545 (40%)	902 (30%)	1,057 (20%) †
US$800 plus US$300 per legitimate claim referred	298 (40%)	619 (30%)	735 (20%) †
US$250 plus 0.5% of claim plus US$300 per legitimate claim referred	118 (50%)	401 (30%)	486 (20%)
US$2,500 plus US$500 per legitimate claim referred	1,433 (40%)	1,952 (20%) †	2,253 (20%) †

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue X October 2026

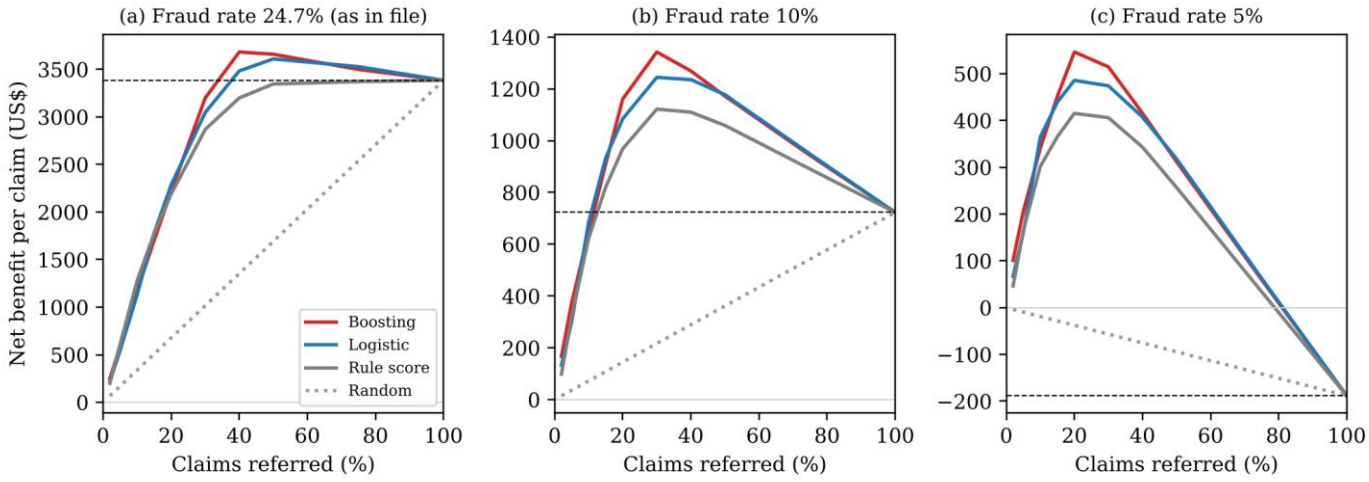


Figure 4. Net benefit per claim by referral capacity at fraud rates of 24.7% (a), 10% (b), and 5% (c), for a cost of US$800 plus US$300 per legitimate claim referred and ρ = 0.3. The dashed line marks the result of referring every claim.

DISCUSSION

Interpretation

The central result is a pair of calibrations. Random cross-validation overstates credible AUC by 0.08 to 0.15 on this file, and the ordering of models flips when validation is corrected, since the most flexible learners gain most from leakage. Under grouped validation, boosting, logistic regression, and a random forest are close, and a model with two variables is at least as good, which echoes the cautions of Hand [11] and Viaene et al. [23]. The explanation is visible in Table 3: the label is almost a deterministic function of severity and hobby, leaving little for additional capacity to learn. Such sharp structure is atypical of real claims and is consistent with a synthetic label-generation process, so absolute performance here cannot be transferred to production books of business. The one-claim-per-profile check (Table 8) gave nearly the same results as grouped validation, so the copying explains the inflated scores under random folds but not the strong link between the label and the two fields.

Integration with Guidewire ClaimCenter

This study was not conducted in a Guidewire environment; the following is a design proposal. At FNOL, a claim created in ClaimCenter would trigger a call through the platform's integration layer to an external scoring service, which returns a fraud probability and reason codes for the activity or note attached to the claim. Referral would be capacity-driven: claims in the top k% of daily scores, with k set by SIU staffing, enter the investigation queue, a policy supported by Table 6 and the break-even analysis. The audit also yields validation rules that can run before scoring: component additivity within a tolerance, model year not later than the loss year plus one, incident hour between 0 and 23, and non-negative limits. Violation rates of 99.5%, 16.4%, 0.04% and 0.1% in Table 2 show why such rules matter, because a score computed from malformed inputs is unreliable. Production validation should group records by claimant, vehicle, or policy, and by time, in line with [19].

The proposed process has five steps (Table 11). On Guidewire Cloud, ClaimCenter events can be sent to other systems through the Application Events Service, which is built on Apache Kafka [26]. A scoring service can also be called through the Integration Gateway or the Cloud API using REST and JSON [27]. In older installations, event messaging, controlled by business rules, determines which messages go to which destinations [26]. The exact set-up depends on the Guidewire version and on how the carrier has configured it, so this design must be checked with the carrier's own integration team.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue X October 2026

Table 11. Proposed steps for scoring claims inside ClaimCenter and the practical issues at each step.

Step	What happens	Practical issue
1. First notice	A claim is created in ClaimCenter and an event with the claim fields is sent out.	Only fields already filled in at that moment can be used. Severity and claim amounts may come later.
2. Data checks	Rules flag impossible values, such as components that do not add up to the total, a model year after the loss year plus one, hour 24, or negative limits.	In the audited file 99.5% of claims fail the sum rule, so these rules should flag at first and block only later.
3. Scoring call	The scoring service returns a fraud score and the main reasons, through the Integration Gateway or the Cloud API.	Response time, outages, and model version control. If the service does not answer, the claim should follow the normal path.
4. Referral	Claims in the top share set by SIU staffing are added to the investigation queue as an activity or note.	The share must be reset when staffing or the fraud mix changes. Handlers should read a referral as a reason to look closer.
5. Feedback	SIU outcomes are written back to the claim and used to retrain the model.	Only referred claims are investigated, so outcomes lean toward what the model already flagged. A small random sample of referrals gives a fair check.

Data availability is the first practical problem. The strongest field in this study, incident severity, and the four claim amounts may be set or revised only after the first review. When boosting was refitted without the nine late-review fields, AUC fell from 0.865 to 0.660 and recall at 20% fell from 53.8% to 34.7% (Table 9). Without those fields and without the personal fields, the score was close to a coin flip (AUC 0.540). A score at first notice, therefore, needs fields that are fully filled in at that point, and a second score after the first adjuster review is likely to be needed. Missing values are also common: property damage and police report were recorded as "?" in 36.0% and 34.3% of claims (Table 2). System compatibility is the second problem. The fields in the file are generic and would have to be mapped to ClaimCenter's own claim, exposure, and contact records. Code lists, such as the severity levels, would have to be matched one-to-one, and the scoring service would have to be tested on the carrier's own version and extensions before use. A model trained on one carrier's claims cannot be assumed to work on another's.

Workflow is the third problem. SIU staff need the reasons behind a score, a clear way to record the outcome, and a rule for what happens to a claim while it waits in the queue. A claim that is referred and found legitimate has already been delayed, which is why a delay cost was included in Section 5.9. Because no test was run inside ClaimCenter, a staged test is proposed. The first stage replays past claims offline. The second runs in shadow mode, where scores are stored but not acted on. The third is a limited live trial with a fixed referral share and a small random sample sent to the SIU as a control. The main measures would be fraud found per hundred referrals, money saved per claim, delay, and referral rates by group.

Governance and Fairness

The second most important predictor is a lifestyle attribute, the insured's hobby. In an operational setting, a hobby is a questionable indicator of fraud and a potential proxy for protected characteristics [1]. Insurers subject to the NAIC model bulletin [15] would need to provide documented justification, undergo bias testing,

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue X October 2026

and be able to explain referrals. The two-variable result also shows how transparent a competitive scorer can be, which supports the argument for interpretable models in high-stakes settings [20].

Section 5.8 shows how much of the score rests on this one field. With the hobby field removed, boosting lost 0.126 of AUC and 9.3 points of recall at 20%, but the gap between women and men in the share of fraud caught fell from 8.2 points to 2.7 points. The field therefore buys accuracy while widening the gap between groups. Whether that trade is acceptable is a decision for the carrier, its regulator, and the public, and it cannot be settled from this file. The hobby field was not tied to sex, education, or relationship status here, but that says little about real claims, where a hobby can stand in for age, income, disability, or neighborhood. The age result is a warning: removing the age field did not eliminate the age gap because another field (months as customer) tracks age. Fairness checks therefore have to look at what the model does, and not only at the list of fields it uses [28, 29].

Four steps are advised for any real use. Fields that describe the person instead of the claim should be left out unless there is a written reason and a test that supports them. Referral rates and the share of fraud caught should be reported each month by sex, age band, and other protected groups. SIU staff should see the main reasons for each score and be free to overrule it. The model version, its fields, and its test results should be kept on record for regulatory review, in line with the NAIC bulletin [15]. Fairness has several definitions that can pull in different directions, so the choice should be made in advance and written down [29].

Extension with Generative AI

The dataset has no adjuster notes or FNOL narratives, so no LLM was evaluated. A plausible extension is to use an LLM [4] to convert free-text narratives into structured flags that feed into the same scorer, leaving the validation and referral framework unchanged. The leakage lesson applies equally to such pipelines, since narratives drawn from the same claimant or incident must remain within a single fold.

LIMITATIONS

The effective sample is about 1,000 independent profiles, of which roughly 247 are fraudulent, so confidence intervals are wide and fold-level tests have limited power. The near-deterministic dependence on severity and hobby suggests synthetic labeling, which restricts external validity. Hyperparameters were fixed rather than extensively tuned, and tuning could narrow or widen the reported gaps. The main hold-out is one profile-grouped split of 200 profiles, and 20 repeated splits show that its recall at 20% referral (59.2% for boosting) was above the average of 53.8%, so the first hold-out figures are on the favorable side. Investigation cost and recovery fraction are assumptions, and the rule-based score reflects the analyst's priors. The wider cost cases in Section 5.9 are also assumptions, not carrier data, and the grouping of fields into personal and late-review fields is an assumption as well. The timestamps were perturbed within profiles, so a temporal validation could not be performed. No second dataset was tested. The one-claim-per-profile check removes the copying but not the unusual link between the label and the fields, so it is not a substitute for independent claims with realistic fraud patterns. The study also did not use Guidewire software or proprietary data, so the integration in Section 6.2 and its staged test have not been tried.

CONCLUSION

On a 2024 open U.S. auto claims dataset, conventional validation reported near-perfect fraud detection that vanished once replicated claim profiles were kept together, with boosting falling from an AUC of 0.9999 to 0.862. Under credible validation, complex learners were statistically indistinguishable from logistic regression, and a two-variable model matched or beat them, while all learned models outperformed a rule-based score. With a 20% claim capacity, scoring captured 53%-59% of fraud on average across 20 repeated holdouts (47% for the rule score), compared with about 20% for random selection. When the investigation cost was low and fraud was common, as in the base case, referring every claim was as valuable as scoring. With higher costs or a fraud rate of 5% to 10%, a ranked queue of 20% to 40% of claims earned clearly more. For Guidewire ClaimCenter operators, the practical implications are to validate by claim profile, benchmark

against simple models, size referrals by SIU capacity, and encode data-quality rules at FNOL. Fields that describe the person, such as hobby, should be left out or tested for fairness first, and the score should rely on fields that are really known at first notice. Future work should repeat the design on a second dataset of independently recorded claims with realistic fraud patterns, test it in shadow mode inside ClaimCenter, add narrative features through LLM-based extraction, and compare fairness measures using real carrier cost data.

Ethical Approval

This research used only a publicly available, anonymized dataset and involved no human participants or animals; formal ethical approval was therefore not required.

Conflict of Interest

The authors declare no conflict of interest. This research received no external funding and had no affiliation with Guidewire Software, Inc.

Data Availability

The dataset analyzed is publicly available on Zenodo (Ndinguri, 2024): Auto Insurance Claims, https://doi.org/10.5281/zenodo.13381118 [16]. The analysis code, the data preparation steps, and the result files are shared in a public repository at [repository link to be inserted before publication]. The repository holds a README that gives the order in which to run the scripts, the software versions, and the random seeds.

License

This article is published under the Creative Commons Attribution License (CC BY 4.0).

REFERENCES

1. Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104(3), 671–732.
2. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.
3. Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78(1), 1–3.
4. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., … Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.
5. 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). ACM.
6. Coalition Against Insurance Fraud. (2022). The impact of insurance fraud on the U.S. economy. https://insurancefraud.org
7. Efron, B., & Tibshirani, R. J. (1994). An introduction to the bootstrap. Chapman & Hall/CRC.
8. Elkan, C. (2001). The foundations of cost-sensitive learning. In Proceedings of the 17th International Joint Conference on Artificial Intelligence (pp. 973–978).
9. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189–1232.
10. Guidewire Software, Inc. (2024). Guidewire ClaimCenter [Product information]. https://www.guidewire.com
11. Hand, D. J. (2006). Classifier technology and the illusion of progress. Statistical Science, 21(1), 1–14.
12. Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. Patterns, 4(9), 100804.
13. Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. ACM Transactions on Knowledge Discovery from Data, 6(4), 1–21.

14. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., … Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30, 3146–3154.
15. National Association of Insurance Commissioners. (2023). Model bulletin: Use of algorithms, predictive models, and artificial intelligence systems by insurers.
16. Ndinguri. (2024). Auto Insurance Claims [Data set]. Zenodo. https://doi.org/10.5281/zenodo.13381118
17. 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). ACM.
18. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
19. Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., … Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography, 40(8), 913–929.
20. 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.
21. 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.
22. Šubelj, L., Furlan, Š., & Bajec, M. (2011). An expert system for detecting automobile insurance fraud using social network analysis. Expert Systems with Applications, 38(1), 1039–1052.
23. Viaene, S., Derrig, R. A., Baesens, B., & Dedene, G. (2002). A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection. Journal of Risk and Insurance, 69(3), 373–421.
24. Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. Medical Decision Making, 26(6), 565–574.
25. Wilcoxon, F. (1945). Individual comparisons by ranking methods. Biometrics Bulletin, 1(6), 80–83.
26. Guidewire Engineering. (n.d.). Guidewire Application Events Service. Guidewire Engineering Blog. https://medium.com/guidewire-engineering-blog/guidewire-application-events-service-d1b0bee685b3
27. Xtivia. (n.d.). Integrating Guidewire with DataRobot for smarter insurance decisions. https://www.xtivia.com/blog/integrating-guidewire-with-datarobot-for-smarter-insurance-decisions/
28. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in Neural Information Processing Systems, 29, 3315–3323.
29. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), Article 115.