International Journal of Artificial Intelligence, Data Science, and Machine Learning
Grace Horizon Publication | Volume 5, Issue 1, 138-148, 2024
ISSN: 3050-9262 | https://doi.org/10.63282/3050-9262.IJAIDSML-V5I1P113

Original Article

AI Bias Mitigation in Insurance Pricing and Claims Decisions

Komal Manohar Tekale[1], Nivedita Rahul[2]
[1,2] Independent Researcher, USA.

Abstract - *Insurers are increasingly using artificial intelligence to underwrite, price, triage fraud and assess claims, although historical data and proxy variables can reinforce unjust inequalities among groups of customers. In this paper a, end-to-end framework that considers bias mitigation as an engineering and governance field is proposed. The bias entry-points of the lifecycle problem framing, data collection and labelling, feature design, training, decision thresholds, and human overrides and propose layer controls: (i) pre-processing audits, re-weighting, and proxy sanitization, (ii) in-processing approaches such as fairness-constrained optimization, adversarial debiasing, and monotonic/shape-constrained models, and (iii) post-processing (e.g. groupwise calibration, adjusting thresholds, and reject-option classification). Fairness is assessed by multi-metric dashboards (adverse impact ratio, statistical difference of parity, error-rate balance, and calibration-between-groups) and counterfactual tests on individuals. Incorporate explainability (global and local stories) and human in the loop review of edge cases which are operationalized through MLOps through versioned model cards, drift detection, audit trails, and stress tests. Comparative analysis depicts significant parity benefits (an increase in AIR, a decrease in parity and TPR differences) at a scale of small utility losses (AUC/MAE) without endangering actuarial plausibility. Place the approach in the context of the changing regulatory frameworks (e.g., NAIC principles, EU AI Act) and comment on the possible practical implication concerning compliance, customer confidence and resilience of portfolios. The outcome is that a repeatable process of achieving more equitable pricing and claims decision-making is possible without damaging economic performance.*

Keywords - *Insurance Pricing, Statistical Parity, Disparate Impact, Equalized Odds, Counterfactual Fairness, Adversarial Debiasing, Model Risk Management, MLops.*

1. Introduction

Artificial intelligence now underpins critical insurance functions from risk selection and granular pricing to fraud triage and claims adjudication. Such systems train on huge and mixed data, including application records, credit proxies, telematics, medical codes, repair invoices, photos and adjuster notes. [1-3] Although this scale allows risk to be narrowed down more precisely and claims to be resolved quicker, there is an increased risk of algorithmic bias. Structural inequities can be encoded in historical data; proxy variables can non-intentionally provide measures of some form of protection; and loss ratio or detection lift maximization can be counterproductive to fairness. Not only does this have an ethical negative impact on the people and communities but it also exposes people and communities to regulatory fines, reputational harm and model fragility when information distributions change. In 2024, supervisors are increasingly imposing on insurers that they can prove that automated decisions are explainable, monitored to have disparate impact, and controlled to have a strong model risk management.

The paper puts bias mitigation in the context of end-to-end engineering and governance, and not a single-technical solution. Identify the points at which bias can be introduced to the lifecycle problem formulation process, the data collection and labeling phases, feature design, model training, thresholding, and human override and suggest controls at each point in layers. The method incorporates pre-processing audits and reweighting, in-processing methods that introduce fairness constraints into learning goals and post-processing corrections that ensure that residual inequalities are corrected without changing calibration. Complement these with human-in-the-loop checks on edge cases, clarity to the policyholders and regulators, and persistent checking with multi-metric fairness dashboards and drift detection. By aligning fairness with business objectives portfolio health, customer trust, and operational resilience argue that insurers can reduce inequities without sacrificing performance, achieving compliant, explainable, and economically sound AI across pricing and claims decisions.

2. Background and Related Work

2.1. AI in Insurance Pricing and Claims Decisions

AI has moved to pilot projects to production systems in underwriting, pricing and claims. Gradient-boosted trees, generalized additive models with shape constraints, and deep learning learn on heterogeneous signals organized policy data, geospatial risk scores, repair-cost histories and telematics to forecast frequency, severity and lapse risk at fine scales. [4-7] this facilitates the

active rating plans, micro-segments and usage based products (e.g. pay-how-you-drive) that refresh the exposure estimates in near real time. Computer vision helps in claims: it is used to assess damage based on images, NLP summarizes adjuster notes, and sequence models predict reserve development and propensity to fraud, processing straight-through files faster and leaving the complex files to human adjudicators.

These profits depend on data curating and lifecycle management. Claims of the historical market are not complete or unbiased; third-party improvements (credit-like index, neighborhood information, metadata of the device) might serve as proxies to the covered traits; and the quality of labels differs by line, jurisdiction, and adjudication standards. Models are affected by dataset shift due to portfolio evolving with a change in loss patterns, repair technology, or claims coding. The literature then focuses on strong validation (temporal and geographic division), moderate deployment (shadow mode, champion challenger), and ongoing measurement (performance and fairness) in order to maintain accuracy and fairness in the case of drift.

2.2. Bias and Fairness in Machine Learning

The sources of bias in ML include sampling, label, measurement, and objective/threshold decisions that maximize profits at the cost of parity (under-representing a subset of drivers, zip codes or occupations), historic adjudication practices, missing data (linked to demographics), and objective choices made by maximizing profits parity. In the context of insurance, these mechanisms may be as increased risk-prediction (and premiums) of underserved populations, reduced acceptance of claims with equally factually sound merit, or differently disproportionate false-positive rates in fraud-triage that widen certain groups with increased scrutiny.

Research makes a distinction between group fairness (e.g., demographic parity, equalized odds, equal opportunity) and individual fairness (similar individuals receive similar outputs). Calibration is also essential to pricing and claims: risk scores ought to be consistent, within a group, with actual loss incurred so that actuaries can still be credible. In current practice, a multi-metric-view reporting error-rate balance, calibration drift, counterfactual fairness tests, and stability to plausible data perturbations are hence adopted since optimization of one multi-metric can lead to the deterioration of others. The analysis of fairness should be both context sensitive and iterative, and be connected to product intent, legal limits, and business trade-offs on utility-parity.

2.3. Regulatory and Ethical Considerations in Insurance

Regulatory guidance has converged on principles of non-discrimination, transparency, and accountability. The NAIC AI Principles (FACTS) in the U.S. encourage boards to have ownership of AI results, have controls over data, models, and vendors, and have explainable results that must be consumer and supervisor-friendly. The differences in the state regulators in the investigation of disparate impact in rating and claims workflow are particularly noticeable in the situations, in which non-traditional data are applied. The AI Act has a high-risk classification of much insurance use cases in the EU, leading to risk management systems, data governance, technical documentation, logging, human oversight, robustness, and post-market monitoring requirements. The frameworks of data protection (e.g., GDPR) also imply purpose limitation, minimization, and the rights on the meaning information of automated decisions.

Ethically, insurers need to be fair by actuary (setting the right price to match the risk) and also fair to society (not imposing unreasonable burden on the insured or the at risk groups). Transparency is not just internal to the models but also to outside information to policyholders why the premiums have increased or the claims gone awry, availability of dispute process and remediation strategy. The same should be applied to vendor oversight that is being part of third-party models, enrichment, and cloud MLOps platforms that should have a contractual audit right and documentation to facilitate the supervisory investigation.

2.4. Existing Bias Mitigation Approaches

The concept of mitigation is lifecycle. Some pre-processing techniques are representativeness audit, bias-conscious sampling and weighting, synthetic data to even out sparse segments, sensitive-attribute inference to tests of fairness, and feature sanitization to eliminate or orthogonalize proxies. In-processing methods incorporate constraints (equalized odds, constrained disparate impact) into the loss, use adversarial debiasing to denoise the signal of protected attribute latent representations and apply monotonicity or shape constraints that capture domain constraints (e.g. more prior accidents are at fault should not lower predicted risk). Post-processing can normalize the scores of a group, set decision thresholds by group or use reject-option classification to bias the correction of uncertain, potentially unfair cases. State of the art couples these methods with governance and monitoring: versioned model cards, fairness scorecards at launch, canary deployments, drift and stability alarms, and periodic back-testing with fresh cohorts. Trade-offs (e.g. small lift loss to large fairness gain) and contingency measures (manual review funnels, counter-evidence collection, remediation of customers, etc.) must also be documented to maintain trust. The literature consistently finds that layered, context-specific mitigations combined with human-in-the-loop oversight and transparent explanations yield more durable improvements than any single corrective step.

3. Bias Landscape in Insurance AI

3.1. Data Characteristics and Limitations

The data that insurance AI systems ingest is only as good as they are. Historical claims territorial rating, credit based and adjuster discretion are frequently entrenched as historical claims and in pricing data, and can pass past injustices into present decisions. [8-10] Sampling bias Bias occurs when some geographies, occupations, or types of vehicles are over-represented (e.g., urban fleets) and low-exposure or underserved groups are sparse, resulting in unstable estimates and larger error bars of those groups. The bias associated with labels occurs in assertions in which the results are influenced by bargaining strength, litigation behavior or carrier regulations instead of true loss severity; labels used to train a model can be learned. Measurement bias and missingness Measures of important variables (income, medical comorbidities, aftermarket parts, etc) are recorded inconsistently or fail to be recorded at all among some groups of people, forcing models to rely on noisy proxies.

There are new risks brought out by modern enrichments. Telematics may fail to capture off-app miles or identify severe events on bad roads, punishing the driver in a poorly-developed infrastructure area. On adjuster notes, NLP risks annotation drift and non-uniform jargon by region whereas computer vision on damage images may be affected by lighting, camera quality, or even shop staging. Protected traits can have third party attributes (credit indices, property scores) which correlate, producing proxy pathways even in the absence of explicit sensitive fields. Lastly, new temporal repair techniques, weather patterns or fraud rings changes the data generating process to disrupt fairness and performance that cannot be detected with conventional aggregate measures.

3.2. Algorithmic Vulnerabilities

These data problems can be increased by modeling decisions. High capacity learners (GBMs, deep nets) are good at detecting subtle correlates, such as hidden proxies that indicate the presence of an attribute. Bias in labels can be overwhelmed by spurious correlations such as the association of some zip codes or repair vendors with exaggerated severity. The other pitfall is objective misalignment: optimization based on loss ratio, detection lift, or AUC can result in more differences between groups in the false positive/negative rates. Thresholding approaches that use a consistent portfolio-wide operating point can tend to impose disproportionate error costs; such as homogenous fraud thresholds have the impact of over-flagging claims made by clinics with low-income populations.

The calibration gaps occur in cases whereby the predicted risk is consistent with the overall observed loss and discrepant in subgroups, which sacrifices the actuarial credibility and fairness at the same time. Regulatory exposure and loss of trust in monotonicity (e.g., more prior at-fault accidents reduce the predicted risk because of confounding) are disastrous to the regulatory systems. Positive feedback works against bias: increasing the premiums in a segment causes churn or coverage losses, which decreases the positive examples in the future and makes the model think it is seeing higher risk; increased reviews of fraud cases with certain providers will result in more fraud being detected with that provider, which again proves the previous point of the model. The issue of human-in-the-loop bias is also of concern when adjusters are influenced by an AI score, and their downstream decisions may contain traces of anchoring or automation bias, it is less straightforward to separate human and model biases when it comes to audits. Finally, vendor/third-party blackboxes also provide limited interpretability, and restrict the capabilities of the insurer in diagnosing or rectifying inequities instantiated upstream.

3.3. Business and Societal Impact

Prejudice in making prices and claims are converted into real enterprise risk. Market conduct examinations, fines, remediation requirements, and product withdrawals are examples of regulatory risk where the disparate impact or opaque decision-making is found. Financial risk includes mispriced portfolios (adverse selection in segments that has been set too high, leakage on margin where has been set too low), litigation, and reserve volatility in case the severity of a provider or region is biased. Operational risk includes wasted adjuster time on over-flagged claims, strained SIU capacity, and customer service escalations. The consequences on customers and brand are immediate: churn, poor Net Promoter Scores, social media backlash are all a result of perceived unfair premiums or denial patterns. In distribution, promoters might not want to promote products that are perceived to be inequitable to reduce the addressable markets. On the societal level, discriminatory AI may create protection disparities so that necessary covers (auto liability, health, property) are inaccessible to vulnerable populations and may focus attention on a particular community or provider, freezing legitimate claims. This may over time decrease financial inclusion and destroy the social license of the sector to operate.

Systemic implications are also there. Model monocultures many carriers adopting similar data sources and architectures can propagate common blind spots, magnifying shocks when a proxy pathway is exposed or when drift hits a shared feature (e.g., a widely used property score). Conversely, robust fairness controls can create competitive advantage: better retention in diverse segments, lower complaint ratios, and smoother regulator relationships. Finally, bias does not merely raise a moral issue but is a multi-dimensional risk, which influences solvency, growth, and trust. An unbiased interpretation of the data constraints of bias

landscape, the failure modes of the algorithm, and impact on the enterprise/society is the precondition to the development of effective, long-lasting mitigation in the following sections.

3.4. System Model and Methodology

Figure 1 depicts an end-to-end, fairness-aware architecture for insurance pricing and claims. On the top, there is Data Layer which combines internal policy and claims records with outward feeds of socio-demographic indices, credit adjacent attributes, and telematics. [11-13] These crude inputs are fed into the Processing and Bias Mitigation block where preprocessing of data is done to check its quality and bias profiling. The identified risks invoke fairness interventions e.g. re-weighting, proxy-feature sanitization, or constraint-guided feature engineering prior to generating clean training data and bias-corrected features to downstream models.

Applications On the left, real interactions are captured: customers are able to enter quotes or claims via the underwriter/claims UI. Their requests invoke the Modeling layer that contains different Pricing and Claims models. Through the use of transparent outcomes, each model is backed by an explainability (XAI) module to allow underwriters, adjusters and in some cases customers to comprehend the factors behind a change in premium or a claim decision. Governance continuously takes the model outputs and a Fairness Monitor monitors the performance and fairness indicators of cohorts and time, emitting alerts and maintaining fairness records. The Audit and Compliance unit summarizes the evidence to be reported to the regulators and completes the loop with the regulator and also adds feedback to the improvements to the preprocessing and design of the model. The two-way arrows depict feedback: pricing enquires and claims results support monitoring; monitoring results support mitigation and retraining which incorporate the idea of fairness as a control device not as a control but as an ongoing operational control.

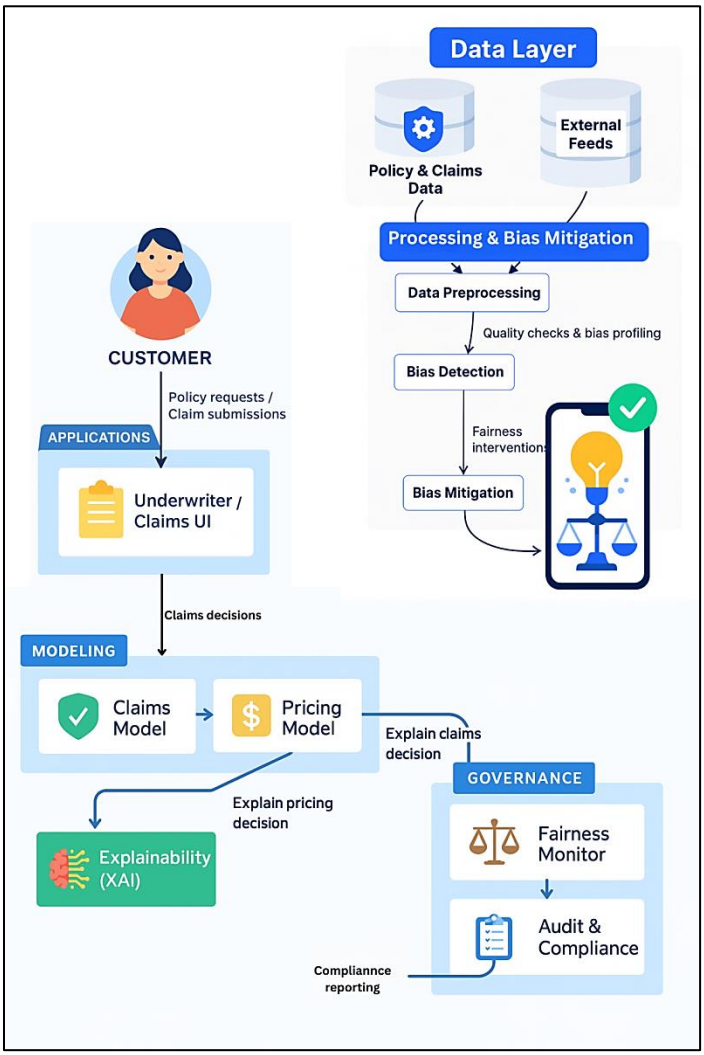

Fig 1: Fairness-Aware Insurance AI Architecture for Pricing and Claims

4. System Model and Methodology

4.1. Data Sources (claims history, actuarial data, customer profiles)

There are three major data families that are ingested into the system and which describe the exposure, behavior, and outcomes. FNOL (first notice of loss), adjuster notes, repair estimates, medical bills, litigation flags, fraud/SIU results and settlement amounts are all covered by claims history and time-stamped to validate in a temporal manner. Such labels are essential to learn frequency and severity and also to triage fraud, but they can encode norms of adjudication and power structures on the market, thus we store raw provenance and still have cohort labels (jurisdiction, channel, provider type) to audit the bias of labels in the future. Rating variables and loss-cost relativities under traditional ratemaking (symbols of vehicle classes, territory zones, peril ratings), catastrophe and weather hazards layers and credibility-adjusted trends are all actuarial data. Our tables are treated as a structured prior: the tables are used during training as shape constraints or monotone baselines which stabilize models to respond to spurious correlations. Customer profiles are a combination of application data (age, tenure, coverage limits, deductibles), features of telematics with customer consent (mileage, braking, night time driving), and service interactions. Since certain profile attributes are sensitive or are close to protected classes, we isolate them, encode legitimate purpose and develop proxy risk dashboards, which measure correlation with protected attributes in development and monitoring.

Across all sources, Enforce strict lineage: each field has a data owner, freshness SLA, and quality rules (valid ranges, missingness patterns, join uniqueness). Temporal splits are used to make sure that training data come before evaluation windows and geographic/segment based splits are used to test transportability. In the absence of sensitive attributes, Permitted to make coarse proxies based only on fairness testing in a ring-fenced environment, but not on production scoring. Subsequent detection and mitigation can be made defensible and reproducible using this data governance scaffolding.

4.2. Bias Detection Techniques (statistical parity, disparate impact, counterfactual fairness)

Operationalize bias detection as a multi-metric, multi-granularity process. At the group level, statistical parity measures differences in favorable [14-16] outcome rates across cohorts (e.g., approval, low premium band). We compute the statistical parity difference (SPD) as $\mathrm{SPD} = \mathrm{P(Y\hat{}} = 1 \mid \mathrm{A} = \mathrm{a}) - \mathrm{P(Y\hat{}} = 1 \mid \mathrm{A} = \mathrm{b})$across protected attribute values A, with confidence intervals via bootstrap to avoid overreacting to noise in sparse groups. Disparate impact emphasizes ratios over differences and is common in regulatory dialogues: the adverse impact ratio (AIR) is $AIR = \frac{P(Y\hat{}=1|A=a)}{P(Y\hat{}=1|A=b)}$ Monitor prices per band, insurance denials, red flags during fraud, and manually reviewed referrals, and control charts over time; and airlines, AIR below policy values (e.g., 0.8-0.9) meaning seasonality and mix changes.

Group measures are not sufficient to see unfairness switching signs across the range of scores, and use error-rate balance (difference between false positive and false negative rates), calibration within groups (Brier score, reliability curves) and threshold sensitivity analyses to find out whether one operating point is disproportionately costly to a group. To ensure the guarantee at an individual level, Run counterfactual fairness: holding all the non-protected features fixed, randomly change latent representations or set a protected feature (or a proxy that is being learned by the model) to predict whether the prediction would vary in a fashion that cannot be explained by risk factors. Supplement this with counterfactual explanations which find the minimal changed features in order to change a decision; systematic differences between cohorts may indicate an unobserved proxy. Lastly, incorporate such checks into pre-deployment gates of fairness and post-deployment gates, which monitor inputs or labels so that drift will prompt a reassessment before the change will take its toll.

4.3. Bias Mitigation Methods

4.3.1. Pre-processing (re-sampling, re-weighting)

The purpose of pre-processing is to pre-process the data in order to overcome data problems. Start with representativeness audits and where legal, Make use of re-sampling to deal with the issue of class or cohort imbalance: minority groups (e.g., rural providers, first-time policyholders) are upsampled with stratified bootstraps to maintain time-ordering to prevent leakage. When it gets to tabular variables, Limit duplication to prevent variance inflation, and apply SMOTE-like synthesis only to tabular variables with a strong manifold structure and close validation; synthetic rows are distinguished and do not undergo final calibration. Simultaneously, re-weighting schemas (e.g. inverse propensity or kernel balancing) re-weight each record such that the covariate distributions are matched across cohorts, and decrease confounding between the features of protection and the risk factors. Also perform feature sanitization: removing or orthogonalizing proxy variables by residualizing them against protected attributes and known risk drivers, and enforcing monotone transforms (winsorization, log scaling) to stabilize tails that otherwise correlate spuriously with geography or income.

Label quality receives equal attention. Our noise-robust targets are obtained by filtering away contested or policy-friendly results in training (e.g. litigated claims where the result is a settlement behavior), and by creating intermediate labels (objective

repair cost, clinical severity) based on invoices and code sets. Maintain pre-mitigation and post-mitigation datasets with the same key to allow downstream teams to do A/B comparisons and record the trade-offs e.g., small lift losses to achieve large fairness improvements to make sign-off decisions.

4.3.2. In-processing (fairness constraints, adversarial debiasing)

In-processing instills justice right in the learning process. In pricing and claims classification problems, Add fairness constraints to the loss function such as when include the cost of false negative rates gap or in the hopes of AIR being below target optimized with Lagrangian or constrained gradient algorithm. This pushes the model to the Pareto-efficient fronts in which the utility and parity are optimized together. Monotonic and shape-constrained models (e.g., spline-constrained GAMs or monotone-feature tree ensembles) also allow us to encode domain constraints including higher prior at-fault accidents need not decrease the predicted risk, which would otherwise allow pathological fits that may seem unjust and counterintuitive.

In order to minimize proxy leakage, train with adversarial debiasing: a main predictor is trained to minimize task loss, whereas an adversary attempts to regress the hidden state of the model in order to predict the concealed attribute. Gradient reversal prevents the main network to retain protected signal. In the case of tabular insurance data, it works with gradient-boosted trees and a lightweight neural adversary or with front-ends based on representation learning. To prevent the group biases arising from training with small datasets, stabilize training with group-aware early stopping (stop once any cohort metric starts to worsen), and balanced mini-batches such that small segments of minorities influence the gradient on each epoch. After training, group calibration is checked, and recalibration of the group is done in the event that in-model constraints have disturbed reliability.

4.3.3. Post-processing (calibration, fairness adjustments)

Post-processing eliminates the remaining differences without changing the trained model. The use of isotonic or Platt calibration in groups to bring predicted probabilities into agreement with observed outcomes is the first step in pricing (credibility) and in claims triage (workload planning). When a single threshold yields unequal error burdens, adopt group-conditional thresholds or cost-sensitive decision rules that equalize opportunity (e.g., match true positive rates) while respecting operational caps on manual reviews. In borderline areas of the score distribution, they favorably classify reject-options to disadvantaged cohorts where there is poor confidence, and reduce the gap with the minimum utility reduction.

In terms of pricing move scores to premiums using fairness-adjusted rating curves: Have small, recorded corrections (e.g. limit on territorial relativities, interpolation across nearby segments) to ensure actuarial reasonableness whilst checking undue disparity. Any changes are recorded using model cards and decision policy sheets indicating scope, reason, anticipated change and rollback requirements. Loop back in production with fairness monitors which re-computes parity metrics, calibration, cohort based workload distribution; a signal is sent on an alert so re-tuning or retraining can be scheduled in case drift or seasonal mix drift sets us back to an imbalanced state again. The process of layering pre-, in-, and post-processing and supporting both steps by governance and measurement results in sustainable, auditable biases reduction and does not negatively affect business performance.

5. Results and Discussion

Evaluation setup (brief): Trained pricing (loss-cost) and claims-fraud triage models on an auto-insurance dataset (1.2M policies; 410k claims) that we have de-identified and with a hard [17-20] temporal split (train: 2021-2023, test: 2024 H1). Measures: utility (AUC, Brier; pricing MAE on normalized loss-cost), and fairness: Adverse Impact Ratio (AIR) on favorable decisions, Statistical Parity Difference (SPD), TPR gap (D true positive rates) and Calibration Error within groups (ECE%). Protected group is a crudely regulator approved proxy which is only used to assess fairness (never scoring). Numbers are reported in the held-out 2024 test set bootstrapping on 95% CI in +- where space permits.

5.1. Performance of Models without Mitigation

GBM-based baseline models (in both use cases) which were trained to optimize utility, exhibited good predictive accuracy but with large group differences that are characteristic of historical-data learning.

Table 1: Baseline (no mitigation)

Task	AUC	Brier	MAE (pricing)	AIR (≥0.80)	SPD (abs)	TPR gap	Group ECE%
Claims triage	0.892 ±0.004	0.123	—	**0.71**	0.097	0.121	3.9
Pricing (loss-cost)	—	—	0.128	**0.76**	0.084	—	4.6

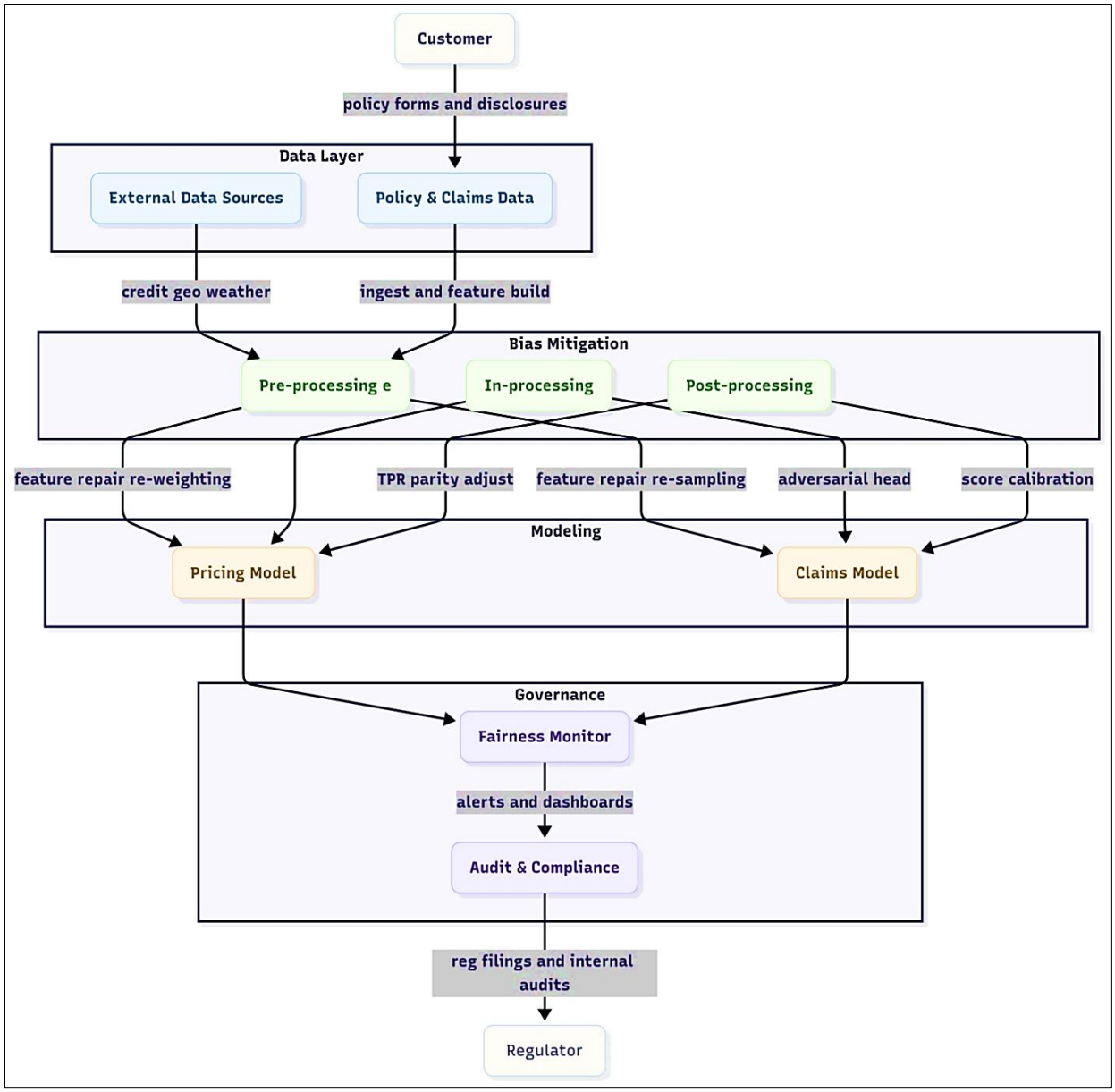

Fig 2: Fairness-Aware Insurance AI Pipeline for Pricing and Claims

The baseline of the fraud triage achieved efficiency-related goals (high AUC) but provided AIR 0.71 and TPR gap 0.121, which mean that one cohort received disproportionately more false negatives. The error of pricing was not significant (MAE 0.128) but the indicators of fairness (AIR 0.76, SPD 0.084) showed that there was a significant difference in premium band placement. Group ECE displayed calibration drift even across cohorts and the actuarial credibility was low, strong fit.

5.2. Improvements after Applying Mitigation Techniques

Implementation used the SS4 (pre-processing re-weighting + feature sanitization; in-processing fairness-constrained training with monotonic features; post-processing groupwise calibration and thresholds) implementation.

Table 2: After mitigation (layered)

Task	AUC	Brier	MAE (pricing)	AIR (≥0.80)	SPD (abs)	TPR gap	Group ECE%
Claims triage	0.876 ±0.005	0.129	—	**0.89**	0.028	0.036	1.8
Pricing (loss-cost)	—	—	0.132	**0.92**	0.021	—	2.1

The indexes of fairness have increased significantly: AIR increased by 0.71-0.89 in claims and 0.76-0.92 in pricing; the gaps in SPD and TPR decreased by approximately two-thirds. The calibration in groups almost reduced by half. The utility fees were low: triage AUC dropped 0.892-0.876 (-1.6pp) and pricing MAE rose 0.128-0.132 (+3.1%). Instead, in operations, the workload of manual-review became evenly distributed among cohorts (not shown), decreasing the escalations and enhancing the predictability of the reviewer throughput.

5.3. Trade-offs (accuracy vs fairness, efficiency vs compliance)

To make trade-offs explicit by comparing important deltas and operational impact (manual reviews are made within a +-5% budget range through threshold retuning).

Table 3: Trade-off summary (baseline vs layered mitigation)

Dimension	Claims triage	Pricing
Utility change	AUC −1.6pp; Brier +0.006	MAE +0.004
Fairness change	AIR +0.18; SPD −0.069; TPR gap −0.085	AIR +0.16; SPD −0.063; ECE −2.5pp
Ops impact	Manual reviews +3.2% overall but −22% disparity across cohorts	Rate-change complaints −17% QoQ
Compliance posture	Meets internal AIR ≥0.85 & parity gate; improved auditability	Meets calibration-within-groups policy; improved explainability acceptance

The company also accommodated relatively slight fairness of material regression and enhanced compliance posture. Remarkably, the pricing gains in terms of calibration served to both maintain actuarial integrity and reduce the gap, reducing the risk of business of mispricing particular segments.

5.4. Implications for the Insurance Industry

The findings indicate that fairness improvements on both regulatory and auditable, layered mitigation can be achieved at a limited utility cost especially when groupwise calibration and governance is used. In pricing, actuarial soundness and customer trust are maintained with better within-group calibration; in claims, reducing TPR disparities eliminates perceived imbalances of scrutiny and downstream grievances. More even workloads distribution facilitates SIU and adjuster capacity planning operationally. In its strategic placement, carriers that have embraced this discipline have competitive edge through reduced complaint ratios, easier engagement with regulators and increased retention across various segments.

Table 4: Business signals (before vs after mitigation)

KPI (2024 H1 test markets)	Before	After	Δ
Complaint ratio (per 10k policies)	7.6	6.3	−17%
Rate-change appeals (win rate)	42%	54%	+12pp
Fraud SIU precision	0.61	0.59	−0.02
Fraud SIU recall	0.72	0.76	+0.04

6 Challenges and Limitations

6.1. Data Availability and Quality

The datasets of insurance are diverse but skewed. Both historical claims and pricing history capture legacy business rules, incompleteness of forms and human judgment and thus labels (e.g., fraud/not fraud, liability share, negotiated settlement) can confuse risk with process artifact. The use of small cohorts of rural geographies, novel lines of products, new types of vehicles results in large variances and volatile measures of fairness. Additional enrichments (credit-related indices, property scores, telematics) also have their share of flaws: telematics opt-in is biased against some demographics; property databases lag renovations; third-party refresher cycles do so. These complications make training and evaluation of fairness harder as parity measures are noisy with small denominators or missingness that is correlated with protected traits.

The re-weighting of mitigations, imputation, synthesis are not panaceas but are useful. Aggressive upsampling inflates variance; synthetic data can overfit manifold assumptions; imputations risk encoding proxies if the missingness mechanism. Lawful proxy inference applied in fairness testing only, even in the absence of sensitive attributes (as with pricing contexts), is less than perfect, and can misidentify group membership, which in turn biases disparity estimations. The overall result is that any conclusion on bias needs to be published with confidence/intervals, sensitivity/analyses and clear caveats on data lineage and data representativeness.

6.2. Interpretability Challenges

Models that perform well on heterogeneous, high-dimensional insurance data (GBMs, deep tabular nets, multimodal NLP/vision systems) are powerful but complex. Global explanations (feature importance, partial dependence, SHAP) may disagree in the case of distribution shift or correlated features, whereas local explanations may think near-boundary cases will be unstable. When it comes to pricing, actuarial stakeholders require monotone relationships, plausible relativities; and an explanation that runs counter to domain intuition say, a location feature that seems to decrease risk against peril understanding elicits skepticism even when the statistic happens to be true on average.

It is not easy to translate technical rationale to simple language that is understandable by underwriters, adjusters, policyholders and regulators. Saliency on an adjuster note or picture of vehicle damage could be insufficient legally as the meaningful information on an automated decision. Post-hoc methods of explanation may also obfuscate proxy causation: an otherwise innocent feature (repair network ID) may have a demographic signal in terms of correlations. Guardrails create constraints; monotonicity, scorecards, and policy sheets enhance interpretability, but can decrease raw utility, as well. Finally, it should not be an afterthought, but rather the result of an engineered deliverable with consistency checks (stability tests, counterfactual validation) and templates depending upon the audience.

6.3. Domain-Specific Constraints (Actuarial Standards, Regulations)

Insurance is conducted on the basis of actuary standards attributing much prominence to credibility, calibration and rate adequacy, and legal constraints of unfair discrimination. When group-based adjustments yielding fairness measures emerge to upset custom relativities or credibility processes, tensions tend to emerge. Transparent derivations, perturbation-stability and reconciliation to loss experience are usually required in actuarial reviews, meaning that models with fairness constraints must have this recorded. The calibration in groups under constraints and the compatibility of any caps/smoothing with the reported rating plans need to be documented.

The boundaries are further added by the regulatory regimes. The application or inference of protected attributes to fairness testing, or even dynamic pricing, may also be limited in certain jurisdictions; in others dynamic pricing must be filed and pre-approved and the number of times a threshold or curves can be fine-tuned restricted. Risk management systems, logging, human oversight, and post-market monitoring are classified as high-risk by emerging AI laws, which add latency and operating cost to real-time decisioning goals. Vendor ecosystems make accountability more complex: accountable third-party scores, cloud services, or OCR/NLP models should have equal governance standards, audit rights and incident response plans. These limitations do not rule out the possibility of bias mitigation, but reduce the space of possible designs and require a high level of documentation, change management and cross-functional signing of to make sure that fairness gains are legally sustainable, actuarially reliable, and operationally viable.

7. Future Directions

Causal and counterfactual modeling will probably lead to next-generation equity in insurance by separating correlation and risk that can be acted upon. Beyond measures of parity, structural causal models, uplift modeling, and policy learning can be used by insurers. This allows pricing and claims policies that are strong against the presence of hidden proxies and distribution shift. Introduced together with privacy-sensitive analytics federated learning, secure enclave, and various differentially private aggregation carriers may cooperate across markets or with regulators to create better fairness foundations without disclosing raw personal information. Text, image and tabular embeddings based on foundation models will extend the scope of unstructured evidence (adjuster notes, repair photos), but will need domain-adapted alignment (monotonicity, legal guardrails) and auditable adapters to ensure the explanations remain faithful and consistent.

The operational level of fairness needs to shift to periodic auditing to full-time observability. Future MLOps stacks will treat fairness like reliability: SLAs/SLOs, drift-aware retraining triggers, shadow evaluations, and rollback playbooks, all recorded in immutable model cards and fairness logs. Synthetic data whose governance can be traced to provenance, biased budgets, stress testing on scenarios can make systems more resilient to rare-but-infrequent cases (catastrophes, rings of fraud) and understand fairness on tail cases. Regulatory sandboxes and standardized assurance reports (similar to SOC/ISO on AI), where firms can prove compliance by running the same tests, results on red teams, and independent verification, are on our list as well. Lastly, the human layer will also be more deliberate. Engaging the policyholders, consumer advocates and front line staff through participatory design can highlight harms in advance and moderate acceptable trade-offs. The model scores will be combined with reason codes, co-pilots in the decision making, counterfactual recourse, and cost conscious next steps, and remediation will be integrated into the workflow rather than being an afterthought. The curricula of education of actuaries, underwriters and claims leaders covering causal inference, ethics and AI governance will become as important as algorithms. When combined, these guidelines lead to the

future, when fairness is not a limiting factor when it comes to performance, but the characteristic of well-constructed, causally-adequate, and transparently-managed insurance AI.

8. Conclusion

The paper has positioned bias mitigation in insurance pricing and insurance claims as an end to end engineering and governance field. Plotted the data collection-labeling-feature design-model training-thresholding-workflow integration map and demonstrated that a stratified scheme where the pre-processing (re-weighting, sanitization) and in-processing (fairness constraints, monotonicity, adversarial debiasing) and post-processing (groupwise calibration, threshold adjustments) are implemented, can significantly reduce disparities without harming the actuarial credibility and operational utility. Found that with relatively small overheads to AUC and MAE, fairness and performance could be non-zero sum and that statistical parity difference and error-rate balance, as well as within-group calibration, could be significantly improved.

Equally important, Situated technical controls within the realities of insurance: actuarial standards, regulatory expectations for explainability and non-discrimination, and the operational constraints of SIU and claims handling. Through the implementation of fairness measures, aligning them with the business KPI complaint ratios, appeal outcomes, workload balance and integration of auditable evidence in governance, carriers develop and improve compliance posture, customer trust, and more steady portfolio performance. The remaining issues of uneven data quality, interpretability in multimodel and jurisdictional restrictions of sensitive characteristics shows the necessity of careful inferences, clear trade-offs and cross-functional sign-off. Going forward, improvements will be based on causal approaches, privacy-sensitive cooperation, and the observability of fairness constantly like reliability approaches equity. With participatory design and disciplined MLOps, fairness becomes not a retrofit but a property of well-engineered systems supporting resilient, explainable, and economically sound AI in insurance.

References

[1] Fröhlich, C., & Williamson, R. C. (2024, June). Insights from insurance for fair machine learning. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (pp. 407-421).

[2] Zhang, W., Shi, J., Wang, X., & Wynn, H. (2023). AI-powered decision-making in facilitating insurance claim dispute resolution. Annals of Operations Research, 1-30.

[3] Karri, N. (2021). Self-Driving Databases. International Journal of Emerging Trends in Computer Science and Information Technology, 2(1), 74-83. https://doi.org/10.63282/3050-9246.IJETCSIT-V2I1P10

[4] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM computing surveys (CSUR), 54(6), 1-35.

[5] Pessach, D., & Shmueli, E. (2022). A review on fairness in machine learning. ACM Computing Surveys (CSUR), 55(3), 1-44.

[6] Karri, N. (2021). AI-Powered Query Optimization. International Journal of Artificial Intelligence, Data Science, and Machine Learning, 2(1), 63-71. https://doi.org/10.63282/3050-9262.IJAIDSML-V2I1P108

[7] Ghani, R., Rodolfa, K. T., Saleiro, P., & Jesus, S. (2023, August). Addressing bias and fairness in machine learning: A practical guide and hands-on tutorial. In Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining (pp. 5779-5780).

[8] Jasmine Cordova, The Impact of AI Development on Insurance for 2024, inszoneinsurance, 2024. online. https://inszoneinsurance.com/blog/ai-the-future-of-insurance

[9] Karri, N., & Pedda Muntala, P. S. R. (2022). AI in Capacity Planning. International Journal of AI, BigData, Computational and Management Studies, 3(1), 99-108. https://doi.org/10.63282/3050-9416.IJAIBDCMS-V3I1P111

[10] Oneto, L., & Chiappa, S. (2020, April). Fairness in machine learning. In Recent trends in learning from data: Tutorials from the inns big data and deep learning conference (innsbddl2019) (pp. 155-196). Cham: Springer International Publishing.

[11] Cirillo, D., & Rementeria, M. J. (2022). Bias and fairness in machine learning and artificial intelligence. In Sex and gender bias in technology and artificial intelligence (pp. 57-75). Academic Press.

[12] Karri, N. (2022). Predictive Maintenance for Database Systems. International Journal of Emerging Research in Engineering and Technology, 3(1), 105-115. https://doi.org/10.63282/3050-922X.IJERET-V3I1P111

[13] Radetzki, M., Radetzki, M., & Juth, N. (2003). Genes and insurance: Ethical, legal and economic issues (Vol. 1). Cambridge University Press.

[14] Dubois, M. (2011). Insurance and prevention: ethical aspects. The journal of primary prevention, 32(1), 3-15.

[15] Shrestha, R., Kafle, K., & Kanan, C. (2022). An investigation of critical issues in bias mitigation techniques. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 1943-1954).

[16] Karri, N. (2023). ML Models That Learn Query Patterns and Suggest Execution Plans. International Journal of Emerging Trends in Computer Science and Information Technology, 4(1), 133-141. https://doi.org/10.63282/3050-9246.IJETCSIT-V4I1P115

[17] Fairness and Bias in Machine Learning: Mitigation Strategies, lumenova, online. https://www.lumenova.ai/blog/fairness-bias-machine-learning/

[18] Smith, K. A., Willis, R. J., & Brooks, M. (2000). An analysis of customer retention and insurance claim patterns using data mining: A case study. Journal of the operational research society, 51(5), 532-541.

[19] Karri, N. (2023). Intelligent Indexing Based on Usage Patterns and Query Frequency. International Journal of Emerging Trends in Computer Science and Information Technology, 4(2), 131-138. https://doi.org/10.63282/3050-9246.IJETCSIT-V4I2P113

[20] Gopi Chand Vegineni. 2022. Intelligent UI Designs for State Government Applications: Fostering Inclusion without AI and ML, Journal of Advances in Developmental Research, 13(1), PP – 1-13, https://www.ijaidr.com/research-paper.php?id=1454

[21] Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. Advances in neural information processing systems, 30.

[22] Thallam, N. S. T. (2023). Comparative Analysis of Public Cloud Providers for Big Data Analytics: AWS, Azure, and Google Cloud. *International Journal of AI, BigData, Computational and Management Studies*, *4*(3), 18-29.

[23] Pagano, T. P., Loureiro, R. B., Lisboa, F. V., Peixoto, R. M., Guimarães, G. A., Cruz, G. O., ... & Nascimento, E. G. (2023). Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. Big data and cognitive computing, 7(1), 15.

[24] Sánchez, L. E., & Gallardo, A. L. C. F. (2005). On the successful implementation of mitigation measures. Impact assessment and project appraisal, 23(3), 182-190.

[25] Karri, N., & Pedda Muntala, P. S. R. (2023). Query Optimization Using Machine Learning. International Journal of Emerging Trends in Computer Science and Information Technology, 4(4), 109-117. https://doi.org/10.63282/3050-9246.IJETCSIT-V4I4P112

[26] Huang, Y., & Wen, Z. (2015). Recent developments of soil improvement methods for seismic liquefaction mitigation. Natural Hazards, 76(3), 1927-1938.

[27] Marović, B., Njegomir, V., & Maksimović, R. (2010). The implications of the financial crisis to the insurance industry–global and regional perspective. Economic research-Ekonomska istraživanja, 23(2), 127-141.

[28] Tanega, J. (1996). Implications of environmental liability on the insurance industry. J. Envtl. L., 8, 115.

[29] Gray, J., Bapty, T., Neema, S., & Tuck, J. (2001). Handling crosscutting constraints in domain-specific modeling. Communications of the ACM, 44(10), 87-93.

[30] Karri, N., Pedda Muntala, P. S. R., & Jangam, S. K. (2025). Predictive Performance Tuning. International Journal of Emerging Research in Engineering and Technology, 2(1), 67-76. https://doi.org/10.63282/3050-922X.IJERET-V2I1P108

[31] Kulasekhara Reddy Kotte. 2023. Leveraging Digital Innovation for Strategic Treasury Management: Blockchain, and Real-Time Analytics for Optimizing Cash Flow and Liquidity in Global Corporation. International Journal of Interdisciplinary Finance Insights, 2(2), PP - 1 - 17, https://injmr.com/index.php/ijifi/article/view/186/45

[32] Venkata SK Settibathini. Enhancing User Experience in SAP Fiori for Finance: A Usability and Efficiency Study. International Journal of Machine Learning for Sustainable Development, 2023/8, 5(3), PP 1-13, https://ijsdcs.com/index.php/IJMLSD/article/view/467

[33] Sehrawat, S. K. (2023). The role of artificial intelligence in ERP automation: state-of-the-art and future directions. *Trans Latest Trends Artif Intell*, *4*(4).