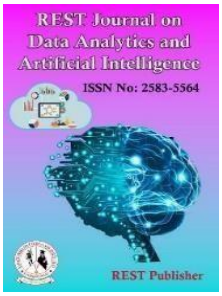
REST Journal on Data Analytics and Artificial Intelligence
Vol: 4(2), June 2025
REST Publisher; ISSN: 2583-5564
Website: http://restpublisher.com/journals/jdaai/
DOI: https://doi.org/10.46632/ jdaai/4/2/3



Fairness-Aware Machine Learning for Bias Mitigation in Vehicle Insurance Pricing

Sai Atmaram Batchu
Jawaharlal Technological University
Corresponding author Email: *sairam.bsa8955@gmail.com*

***Abstract**: As machine learning models become increasingly integrated into vehicle insurance pricing, concerns about fair- ness and discrimination in premium calculations have grown. Traditional risk assessment models may unintentionally reflect historical biases, leading to disparities in insurance costs across different demographic groups. This study explores fairness-aware pricing methodologies to mitigate bias in automobile insurance premium predictions. By applying various pre-processing, in- processing, and post-processing techniques, the research evaluates their impact on both pricing accuracy and fairness metrics. Using real-world vehicle insurance data, the findings demonstrate that bias-aware models can reduce unfair premium discrepancies while maintaining predictive performance. This study highlights the importance of incorporating fairness constraints in actuarial modeling and provides insights into regulatory considerations for ethical insurance pricing.*

1. INTRODUCTION

Motivation: The protection field performs under a distinctive financial structure where premiums are compiled at the commencement of a policy expression in transfer for distribution against unclear forthcoming events. This essential asymmetry stresses the probabilistic nature of coverage appraisal, demanding resilient statistical frameworks to evaluation and lessen risk successfully. Beyond statistical methodologies, high-quality computations indicate a subtle balance of financial longevity, regulatory conformity, and sector capability. With growing contest in the protection environment, both verified businesses and arising sector applicants are forced to polish their tariffing tactics. Regulatory frameworks are uninterruptedly progressing, molding sector circumstances and mandating the integration of elaborate analytical utilities to enhance optimum structures. Firms must position their costing models with corporate purposes whereas validating obedience with developing market standards. Technological improvements, especially in artificial intelli- gence and machine learning, have deeply converted various sectors, featuring guarantee. Augmented computational force and the accessibility of vast datasets have allowed insurers to progress developed forecasting models for risk assessment, pa- tron segmentation, and costing optimization. These technolo- gies serve as essential mechanisms for polishing underwriting mechanisms and amplifying decision-making productivity. Nevertheless, the enhancing commitment on data-driven methodologies has amplified considerations regarding clarity and equity. Unlike advice systems that tailor substance to consumer priorities, prognostic models in guarantee impact financial undertakings and availability to scope. Designated the intense effect of these decisions, guaranteeing equitable treatment across demographic sets is primary. Responding to biases in prognostic models continues an vital facet of ethical decision-making within the field. Regulatory adjustments focused at curbing discriminatory costing practices have been incorporated to support neutrality. Legislative actions aiming inequalities in assurance appraisal, such as limitations on implementing gender as a risk deter- minant, demonstrate the rising highlight on equitable tariffing structures. Therefore, insurers must combine neutrality criteria into their appraisal frameworks, alternatively in response to legal mandates or as fraction of expanded corporate plans. Bias in guarantee models surfaces when reactive

traits unconsciously authority valuation results. Archived situations of algorithmic bias in areas such as criminal righteousness, enlistment, and financial services accentuate the intricacies of alleviating inadvertent discriminatory consequences. During certain techniques pursue to sidestep bias, a further explicit system includes precisely recognizing and dealing with im- balances within prognostic models. A explicit instance of nonlinear bias arises when compo- nents linked with susceptible traits impact appraisal decisions. Moreover, if a model does not precisely use aspects such as gender, interdependent elements may perpetuate inadvertent differences. Traditional calibrations that emphasis uniquely on explicit determinants frequently terminate to statement for these secondary impacts. Offered that demographic parameter such as age and health status effect behavioral patterns and risk awareness, a detailed impartiality assessment persists fundamental in model development and deployment. B. Agenda This exploration aims to furnish actuaries with a organized method for distinguishing, assessing, and lessening the un- foreseen results of susceptible features in coverage tariffing models. Subsection 2 examines key impartiality concepts and sta- tistical methodologies used to appraise bias in forecasting models, specifically in the framework of protection valuation. Chapter 3 presents various procedures constructed to alleviate biases and improve impartiality in rate setting frameworks. Segment 4 illustrates an empirical situation examination on fairness-aware appraisal in automobile assurance, accentuating bias assessment and model refinements. Division 5 explores the productivity of bias amelioration plans and their compre- hensive consequences for assurance rate setting practices. All analytical realizations manipulate Python 3. 8 for computation and validation. C. Notations In this publication, Y symbolizes the target parameter to be expected, which may be absolute or numerical. Designated the principal emphasis on costing, Y is generally a continuous factor, though it may be addressed as binary in explicit illustrative scenarios. The forecasting model is portrayed as m, with non- sensitive illustrative features signified by the vector X = (X1, X2, dots, Xp). Susceptible traits, illustrated as S, cover factors that should not impact appraisal decisions, alternatively clearly or suggestively. These properties, normally definitive, comprise elements such as gender or cultural background. When only one reactive attribute is assessed, S is inferred to be binary, with metrics of 0 and 1. The repercussions of dealing with several responsive traits are reviewed in consecutive sections. Each data instance, associated to a protection contract, hypothesis, or policyholder, is portrayed as (yi, xi, si), where i extents from 1 to n. The subgroups n0 and n1 signify the number of observations where si = 0 and si = 1, sequentially. Uppercase letters denote random parameters, whereas non- capitalized letters exemplify precise witnessed figures. Predictions from an unconstrained model, commonly men- tioned to as a" baseline model" or" unaware model," are indicated by hatyb = mb(x). In comparison, predictions from a fairness-adjusted model integrating boundaries to decrease the contingency of Y on S are symbolized as hatye = me(x). These limitations, articulated in Segment 3, may be executed through feature preprocessing, regularization repercussions, or post-processing amendments targeted at diminishing bias whereas retaining forecasting correctness. Equity analyses in prognostic modeling typically ad- here to two techniques: adjusting the learning operation (in-processing) or adapting the model's productions (post- processing). A bias metric, signified by varepsilon, is presented to assess individual-level contradictions between predictions from unconstrained and fairness-aware models, systematically articulated as: frameworks to characterize neutrality metrics and codify the framework of non-discrimination with respect to responsive features, represented as (S. Various models, illustrated by (m , perform a central part in these conversations. In spite of growth, in that context continues no worldwide concurrence on jargons or methodologies, constructing regulation a obstacle. Numerous specifications of neutrality occur, supporting to the sophistication of analysis within this discipline. This part establishes key equity divisions, with a specific emphasis on cluster neutrality and individual impartiality, reinforcing the functional supremacy of the former. A. Statistical or Group Fairness Collection equity secures that treatment is stable across demographic collections anchored on nominated susceptible traits. Various statistical neutrality metrics leverage the model (m and predictions (hat Y to appraise divergences. The main purpose is to form statistical balance between advantaged and disadvantaged sets though retaining equity. Equity frameworks have been augmented to non-binary goals, combining occasion attributes (more vulnerable bound- aries) or distributional qualities (stronger restrictions). These perspectives adjust with correlation-based neutrality tech- niques and autonomy constraints. Demographic Parity (Independence).: This principle ensures that predictions Yˆ remain statistically independent of the sensitive variable S, expressed as:

$$Y \perp S.$$

This condition implies that the predicted outcome should not be influenced by S, even indirectly. In insurance applications, where S represents age and Y denotes premiums, demographic parity would suggest that insurance rates remain uniform across age groups. However, this condition may be difficult to implement in practice as it disregards underlying risk factors.

Equalized Odds (Separation).: This criterion requires that predictions Yˆ remain conditionally independent of S given the actual outcome Y:

$$\hat{Y} \perp S|Y.$$

It ensures that for identical true outcomes Y, the predicted distributions Yˆ do not vary across groups distinguished by

$$\varepsilon = \hat{y}b - \hat{y}e, \quad (1)$$

where yˆb represents the prediction from an unconstrained model, and yˆe corresponds to the fairness-aware prediction. To beyond assess equity, the function Vs(xi) is presented, which determines the set of k - closest bordering components within the division where S = s for the i - th examination. This framework supports relative analysis among participants with similar properties although assimilating fairness-driven corrections to confirm equitable costing decisions.

2. ASSESSING FAIRNESS AND BIASES

The evaluation of bias has long been a matter of economic analysis, with recent computational enhancements delivering
S. In pricing models, this framework permits variations in predictions based on Y but prevents direct influence from S, promoting fairness without eliminating legitimate predictive signals. Predictive Parity (Sufficiency).: This approach estab- lishes statistical independence between Y and S given the predicted values Yˆ:
$Y \perp\!\!\!\perp S|\hat{Y}$.
This condition ensures that observed differences in predictions are not reflective of disparities in the sensitive attribute, reinforcing fairness in the predictive process. It is often impractical to achieve all fairness criteria si- mutinously, as inherent conflicts exist between them. This necessitates trade-offs in model selection to balance different fairness measures effectively.

3. INDIVIDUAL FAIRNESS

Individual neutrality, commonly named as similarity-based neutrality, verifies that members with comparable traits col- lect similar effects. Unlike category equity, which examines impartiality at a population stage, individual equity unambigu- ously examines the uniformity of model predictions between persons. This notion is quantitatively outlined as adheres to: [Disparate Treatment] A model m exhibits disparate treatment when the predictions Yˆ depend on the sensitive attribute S, for equalized odds, fairness can be evaluated using disparate mistreatment:
$M1|0 = |P(\hat{Y} = 1|Y = 0, S = 1) - P(\hat{Y} = 1|Y = 0, S = 0)|$, $M0|1 = |P(\hat{Y} = 0|Y = 1, S = 1) - P(\hat{Y} = 0|Y = 1, S = 0)|$. (6)
Lower values suggest greater adherence to fairness con- straints. For non-binary cases, correlation measures such as Kendall's tau, Pearson's correlation, and Spearman's rho pro- vide insights into relationships between variables. Another ef- fective measure is Hirschfeld-Gebelein-Renyi (HGR) maximal correlation: [HGR Maximal Correlation] For random variables given explanatory variables X. To satisfy fairness constraints:
$\hat{Y} \perp\!\!\!\perp S|X.$ (2)
This tenet is essential in fields such as guarantee appraisal, U and V:

$$HGR(U, V) = \max_{f \in FU,\, g \in GV} E[f(U)\, g(V)], \quad (7)$$

where persons with indistinguishable features, distinctly from the perceptive attribute, should attain comparable predictions. Nevertheless, the reciprocity between responsive factors and other informative features introduces obstacles in regulating impartiality.

The sophistication of describing neutrality at an individ- ual stage occurs due to the essential control of perceptive properties on other traits. Demographic, health, and behavioral variables assist to alterations in data distributions, creating it difficult to detect genuinely comparable subjects.

A proposed mathematical approach to capture similarity is the Lipschitz constraint:

$dY(\hat{Y}i, \hat{Y}j) < \lambda dX(Xi, Xj), \quad (3)$

where dY and dX represent distance metrics in the prediction and feature spaces, respectively. This formulation ensures that individuals with similar features should have minimal disparities in their predictions. However, implementing this constraint in practice is challenging due to complex variable interactions.

4. QUANTIFYING UNFAIRNESS

Metrics for fairness evaluation must measure both aggregate-level independence (group fairness) and individual- level deviations based on proximity measures.

Extending Beyond the Binary Case

For scenarios with binary sensitive and target variables, fairness assessments often rely on confusion matrices. One commonly used measure is disparate impact:

$P(\hat{Y} = 1|S = 1)$

$P(\hat{Y} = 1|S = 0)$. (4)

Under demographic parity, values approaching one indicate a fair distribution. Another commonly used metric is:

$M1 = |P(\hat{Y} = 1|S = 1) - P(\hat{Y} = 1|S = 0)|. \quad (5)$

where FU and GV are function spaces with zero mean and unit variance constraints. Assessing HGR is arithmetically comprehensive due to the infinite-dimensional function capacity. A functional substitute is its assessment employing kernel-based strategies. Focus on the Case of Pricing When the target element is continuous, disorder matrix- based metrics forfeit their suitability. Alternatively, equity can be evaluated utilizing correlation and distributional evaluations of Y with esteem to S. For this circumstance, inferring S symbolizes gender, Kendall' s tau fulfills as a beneficial correlation metric. In addition, the HGR correlation estimator facilitates a flexible strategy for recognizing linkages across parameter categories. Moreover, impartiality analysis can be executed utilizing probabilistic extents between conditional distributions Y |S. One such measure, Kullback – Leibler (KL) deviation, evalu- ates imbalances in exclusive distributions across gender clus- ters. The Kolmogorov-Smirnov (KS) test supplies another standpoint by analyzing empirical distributions and recogniz- ing significant divergences. These statistical impartiality metrics are enforced employing the Python package SciPy. KL discrepancy is computed uti- lizing corresponding entropy, validating a rigorous framework for equity examination.

Whereas significant research endures on classification neu- trality, individual neutrality, predominantly in cases encom- passing continuous target components, continues fewer exam- ined. Some prior runs have presented k-nearest-neighbors (k- NN) methods to measure proximity-based neutrality.

1) Adaptation of Flip-Test: Building on prior methods, an adapted fairness evaluation technique for continuous variables is introduced. This approach relies on defining an effective similarity metric among individuals.

1) Identify individuals where si = 1.
2) Locate their k nearest neighbors belonging to the oppo- site group.

$$\Delta V_i = \hat{y}_i - \frac{1}{k} \sum_{x_j \in V^{(x_i)}} y_j.$$

Calculate the mean of these differences:

$$\hat{F}T_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} \Delta V_i$$

The procedure is cyclical for members where si = 0 to retrieve F˜T0. A reasosnable model should yield parameters models that enhance fairness without substantially degrading

close none.

The main constraint of this technique remains in the ro- bustness of the adopted adjacency metric. To moderate this, hyperparameter tuning is employed through grid exploration optimization. Key steps encompass:

- Selecting an appropriate distance metric.
- Determining the optimal number of neighbors (k).
- Identifying the most informative features for fairness evaluation.

A detailed grid retrieval validates smallest bias whereas preserving neutrality. Appraisal is implemented on a distinct test dataset to authenticate results. This procedure supplies a programmatically productive substitute to causal inference strategies although adjusting with counterfactual equity con- cepts.

Upon analyzing unfairness, reduction tactics should be investigated to guarantee neutrality without jeopardizing prog- nostic precision

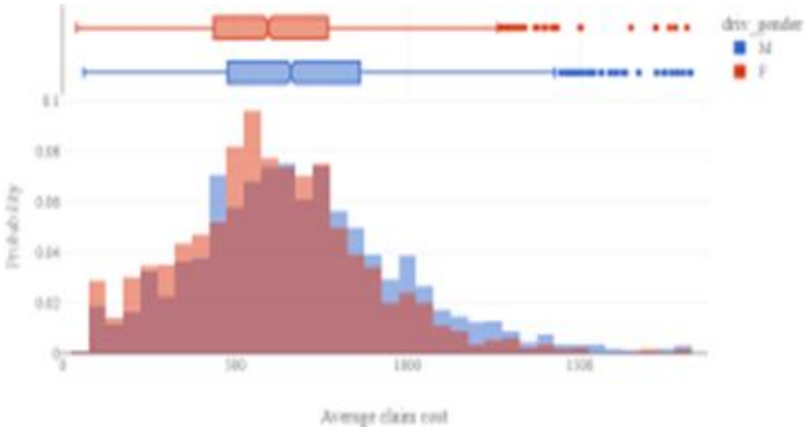


FIGURE 1. Historical average cost and premium distribution by gender.

5. BIAS MITIGATION

Assuring neutrality in prognostic models entails mitigating the unexpected effect of perceptive aspects on effects. Reduction approaches differ relying on the setting, with no widely adopted strategy. Though classification models commonly require binary benchmarks, the priority in this context is on regression-based techniques, predominantly in the assurance field. Bias minimization procedures are mostly sorted into pre- processing, in-processing, and post-processing strategies. Implementing neutrality restrictions during model training frequently results in a trade-off with prognostic functionality Investigations have illustrated the range of these trade-offs, producing it necessary to balance impartiality and exactness optimally. Model assessment is conducted implementing Foundation Mean Square Error (RMSE) and the loss ratio (LR), specified 4) Calculate the mean of these differences: 1 as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} w_i(\hat{y}_i - y_i)^2}, \qquad LR = \frac{\sum_{i=1}^{n} \hat{y}_i}{\sum_{i=1}^{n} y_i} \quad (10)$$

A. Pre-Processing Mitigation

Pre-processing techniques adjust input data to reduce bias while preserving essential patterns. The approaches examined include variable removal, transformation to eliminate depen- dencies, and a tailored synthetic sampling technique. 1) Total Variable Removal: This method eliminates not only the sensitive variable but also other predictors correlated with it, preventing indirect bias transmission. The process involves: • Quantifying dependence between the sensitive variable and each predictor. Defining removal scenarios based on dependency thresh- olds. Evaluating fairness and predictive performance across different removal configurations. Automated threshold-based removal offers an alternative to manually selecting variables for exclusion. 2) Correlation Reduction Transformation: Instead of elim- inating correlated features, this approach modifies them to minimize dependence on the sensitive attribute. One effective method involves using residuals from regression models to decorrelate predictors: $\perp = xj - \alpha S(STS)-1STxj$, (11) where the hyperparameter α controls the degree of adjustment. While this addresses linear dependencies, non-linear associa- tions may persist. 3) Synthetic Sampling Adaptation: Synthetic data gener- ation is used to balance distributions across groups. This method, adapted for regression tasks, involves: • Discretizing the target variable into bins for structured resampling. • Ensuring equal representation within bins across sensitive groups.

• Generating new samples by interpolating between exist- ing instances. The level of binning granularity directly impacts the trade- off between fairness and consistency.

B. In-Processing Mitigation

In-processing techniques incorporate fairness constraints during model training. Methods explored include constrained optimization and parameter search strategies.

1) Fairness-Constrained Optimization: Fairness require- ments can be modeled as linear constraints:

$M\mu(m) \leq c$, (12) where M represents the constraint matrix, c is a control vector, and μ(m) denotes conditional moments. To balance fairness and accuracy, randomized predictions are used:

$\min L(Q)$ subject to $M\mu(Q) \leq c$. (13) 6) Repeat until $\Sigma s < \zeta$.

To maintain consistency, two monitoring metrics are used: • Redistribution integrity: $\max(ye) - \min(ye)$.(18) $\max(y\hat{}) - \min(y\hat{}) \bullet$ Global variation:

$ye - y\hat{}$. (19)

$Q \in \Delta$ i i

An iterative update mechanism adjusts predictions to satisfy constraints while preserving performance.

2) Hyperparameter Search: When optimization-based ap- proaches struggle to converge, structured search methods can identify fair solutions. This method is particularly effective when:

- Gradient-based methods fail to find viable solutions.
- Computational resources limit iterative approaches.
- Discrete tuning of fairness parameters is necessary.

A systematic exploration of parameter space allows for efficient identification of balanced models.

C.post-processing

Post-processing refinements improve model predictions without revising acquired parameters. Common classification perspectives entail threshold tuning, but employing similar tactics to continuous predictions is sophisticated due to the lack of unambiguous decision boundaries.

Substitute approaches manipulate ideal delivery theories to reassign predictions across categories though safeguarding ranking structures.

D.Fair Redistribution

This approach modifies predictions iteratively to align dis- tributions between groups. The process introduces a bias term ε to adjust estimates:

$$ye = y\hat{} - \frac{\varepsilon}{\eta}, \quad (14)$$

where η controls the extent of modification. A higher η retains more of the original estimates, whereas a lower value enforces stronger adjustments.

The redistribution process follows these steps:

1) Initialize η, ζ, and designate the first target group.
2) Adjust predictions for the current group:

$$y\hat{} = y\hat{} - \frac{\varepsilon i}{\eta}, \quad \forall i \text{ in groups.} \quad (15)$$

3) Switch to the next group.
4) Update bias values:

By carefully tuning fairness adjustments, this approach en- sures equitable outcomes while maintaining model reliability. Effective parameter selection plays a critical role in achieving desirable fairness-performance trade-offs.

6. EVALUATING BIAS IN CAR INSURANCE PRICING

Pricing for glass breakage coverage in car insurance requires balancing operational, business, and statistical constraints, including distributional considerations. Table II presents key variables extracted from an insurance database, along with additional vehicle-related characteristics. The dataset outlines variable names, descriptions, values, and statistical summaries. While continuous variables include mean and median values, categorical variables

indicate proportions of the two most common categories. The data underwent preprocessing, involving optimization and discretization, to produce a modeling dataset. The key target elements — frequency, standard cost, and refined high- quality — were inferred from proclamations data. Two ad- ditional variables were established: zoning, which classifies risk into ten levels, and the weight-to-power ratio, which in- dicates automobile properties. Model development assimilated Broadened Linear Models (GLM), Random Forest, and Gra- dient Augmenting, with effectiveness assessed implementing Origin Mean Square Error (RMSE) and Loss Ratio (LR). Interpretability procedures were utilized for optimization and validation. Finally, GLMs were designated for realization due to their lucidity and simple integration into tariffing frameworks. Even though enhanced involved models manifested minor function- ality benefits, their operational hurdles outweighed their mer- its. Yet, all models were analyzed at various stages, affirming that productivity differences endured barely sufficient. To measure equity in rate setting, mention models were developed removing gender as a component. Six discrete association metrics were implemented to measure bias:

- Kendall's tau and mean ratio: Provide an initial under- standing of dependence structures.
- Kolmogorov-Smirnov test p-value and JS divergence: Assess differences in distributional behavior.
- HGR coefficient: Offers a more comprehensive measure

1
ε = yˆ −

yˆ . (16)

of dependency.
i i

5) Evaluate overall bias:

k j
j∈N (i)

Flip-test adaptation: Examines fairness at an individual level. While the first five metrics focus on statistical independence

Σs = εi. (17)
i∈s

for group fairness, the Flip-Test Adaptation provides an alter- native perspective by assessing individual fairness.

TABLE I
SUMMARY OF BIAS MITIGATION TECHNIQUES

Technique	Advantages	Challenges
Preprocessing	Preserves model integrity, computationally efficient	May eliminate useful information, difficult to balance fairness and performance.
Total Variable Removal	Simple approach, detects hidden dependencies	Risk of removing influential predictors, reliance on remaining features.
Correlation Adjustment	Maintains interpretability, mitigates bias effectively.	Requires numerical variables, adjustments necessary post-correction.
FairSMOTE	Generates balanced datasets, customizable sampling	Limited impact on past biases and dependencies.
In-Processing	Balances fairness and predictive accuracy	Complex, resource-intensive, may face convergence issues.
Exponentiated Gradient	Directly integrates fairness constraints	Computationally expensive, requires model adaptation.
Post-Processing	Enhances fairness without retraining	Effectiveness depends on model accuracy, risk of inconsistencies.
Fair Redistribution	Adjusts outputs to achieve fairness	Requires monitoring of premium adjustments, may introduce new constraints.

TABLE II
DATASET VARIABLE OVERVIEW

Variable	Description	Value Range	Statistics
claim_amount	Cost per claim (£)	[0, 1825]	Mean: 580€ — Med: 596€
claim_nb	Number of claims per policy	[0, 5]	0: 97.3%, 1: 2.7%
expo	Policy exposure time (year_pol	Policy issue year	[2015, 2020]
2018: 18%, 2020: 17%			
driv_age	Primary driver age	[18, 77]	Mean: 47, Med: 45
area	Risk category (17 levels)	Categorical	D: 29%, F: 23%
driv_gender	Driver gender	F, M	M: 58.4%
veh_age	Vehicle age (years)	[0, 44]	Mean: 5, Med: 4
veh_power	Engine power (KW)	[13, 220]	Mean: 94, Med: 91
veh_price	Car price (£)	[6.8k, 65k]	Mean: 21.4k€; Med: 15.4k€
claim_hist	Previous claim record	0, 1	0: 91.7%

7. EVALUATING BIAS IN HISTORICAL DATA

Table III showcases various dependence measures used to boundary, validating parallelism in transport employment. Including all available features, including sensitive at- tributes, typically enhances predictive accuracy. However, this introduces fairness concerns as model outputs reveal dispari- ties across demographic groups. Once the sensitive variable is included in the modeling process, its effect on predictions Yˆ is quantified in Table IV.

TABLE 3. Dependence Between Y and S Before Modeling

Y	Kendall	HGR_KDE	KS (p-value)	Div_JS	mean_ratio	Flip-test
Average cost	-0.0796	0.0903	1.94×10^{-7}	0.3217	1.1024	-10.23€
Frequency	-0.0197	0.3031	0.0333	0.8413	1.2489	-2.57%
Premium	-0.0212	0.3106	0.0386	0.8401	1.3450	-7.88€

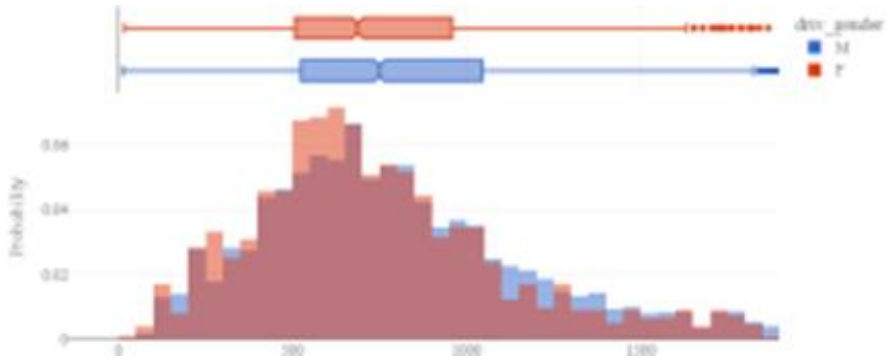


FIGURE 2. Histogram illustrating premium distribution across gender categories. Premium values stem from the product of average cost and frequency, excluding zero cases to enhance visualization.

The Kendall metric signals a weak connection between the responsive attribute and key parameters, though HGR coefficients advocate a minimally stronger attachment. The Kolmogorov-Smirnov test verifies noticeable distributional differences. The flip-test results show that specific policyholder fractions acquire reduce costs by about C 10, adjusting with the undesirable Kendall metrics. Quantities illustrate how these metrics are dissemination across the dataset. Assertion frequency distributions continue mostly uniform Preprocessing Preserves model integrity, computationally efficient Total Variable Removal Simple approach, detects hidden dependencies Correlation

Adjustment Maintains interpretability, mitigates bias ef- fectively Fair SMOTE Generates balanced datasets, customizable sampling May eliminate useful information, difficult to balance fairness and performance. Risk of removing influential predictors, re- liance on remaining features. Requires numerical variables, adjustments necessary post-correction. Limited impact on past biases and dependen- cies. within the [0, 1) range, including 97. 3 The results accentuate that one demographic lean to have enhanced premiums, declaration occurrences, and costs. Nev ertheless, constructing definitive cutoffs for bias in this con- In-Processing Balances fairness and predictive accuracy Complex, resource-intensive, may face con- vergence issues. Exponentiated Gradient Directly integrates fairness constraints Computationally expensive, requires model adaptation. Post-Processing Enhances fairness without retrainin Effectiveness depends on model accuracy, figuration continues difficult. These conclusions stress the historical influence of delicate traits on protection rate setting; Fair Redistribution Adjusts outputs to achieve fairness risk of inconsistencies. Requires monitoring of premium adjust- ments, may introduce new constraints. commonly handled by erasing immediate reliances or post- processing model generations. To management for risk intro- duction differences, the dataset assimilates a uniform mileage

TABLE 4. Dependence between yˆ and s after incorporating the sensitive Variable

$Y^$	Kendall	HGR_KDE	KS (p-value)	Div_JS	mean_ratio	Flip-test
Average cost	-0.1824	0.2241	0.0000	0.3825	1.0960	-3.81€
Frequency	-0.1949	0.3144	0.0000	0.7333	1.2219	-1.09%
Premium	-0.2101	0.3268	0.0000	0.7116	1.3524	-1.44₡

The impact of the perceptive attribute intensifies after model training, as displayed by amplified association marks. Kendall' s tau almost triples for cost calculations, signifying a stronger correlation between predictions and the delicate parameter. Offered statistics visualize these modifications, depicting how non-overlapping distributions increase post-modeling, moreover intensifying inequalities. This exacerbation of bias origins from complex associa tions within the dataset. Many forecasting features illustrate linkages; for instance, automobile authority significantly correlates with price. These links obfuscate neutrality calibrations, as they develop secondary factors of responsive factors on predictions. Furthermore, insignificant initial relationships can have spilling consequences, enhancing variations. Within this dataset, susceptible characteristics illustrate link- ages with various signifiers, encompassing transport standards and steering background. Form 2 highlights how these subsidiary results provide to bias propagation. In cost calculation, the susceptible attribute categories among the four majority influential parameters, moreover bolstering differences through linked forecasters. Behavioral and demographic parameters often underpin these associations. Age can affect risk cognition and guiding conduct, whereas residential sectors and assurance consumption may position with socio-economic attributes. Such aspects exacerbate undertakings to confirm unbiased model generations. A crucial query arises: does neglecting the susceptible attribute completely resolution impartiality issues? In some situations, restriction might be ineffective, as hidden connections in other determinants may nevertheless encrypt demographic patterns. Effective alleviation mandates a ordered framework to bias reduction within prognostic modeling frameworks.

8. BIAS AFTER MODELING WITHOUT GENDER

To mitigate direct discrimination, gender has been excluded from the model, ensuring that predictions are not explicitly influenced by this attribute. However, due to correlations among predictor variables, residual dependencies may still persist. Table V presents the dependence measures between the predicted outcomes and gender, despite its absence in the model.

TABLE 5. Dependence Between Y_b and S After Modeling Without Gender

Y_b	Kendall	HGR_KDE	KS (p-value)	Div_JS	mean_ratio	Flip-test
Average cost	-0.1681	0.2111	9.7460e-39	0.3288	1.0118	-2.71€
Frequency	-0.1309	0.2741	0.0000e+00	0.6493	1.1573	-1.07%
Premium	-0.1733	0.2897	0.0000e+00	0.7226	1.2694	-0.88€

The results signify that even though gender has been extracted from the provision factors, its impact is even so noticeable in the predictions. This implies that other features within the dataset inferredly encrypt gender-related

patterns, upholding variations in the consequences. For case, vehicle- related aspects may illustrate differences across gender sets, unconsciously shaping high-quality assessments. Though insignificant cutbacks in bias are witnessed corre- lated to models expressly assimilating gender, the resilience of secondary determinants features the sophistication of dis- carding bias exclusively. The occurrence of these residual outcomes reinforces the obstacle of attaining accurately neutral predictions, requiring increased elaborate neutrality calibration methods. The analyzed changes in forecasting connections focus on the significance of cautiously examining how recipro- cal relationships among informative elements provide to bias retention in modeling frameworks.

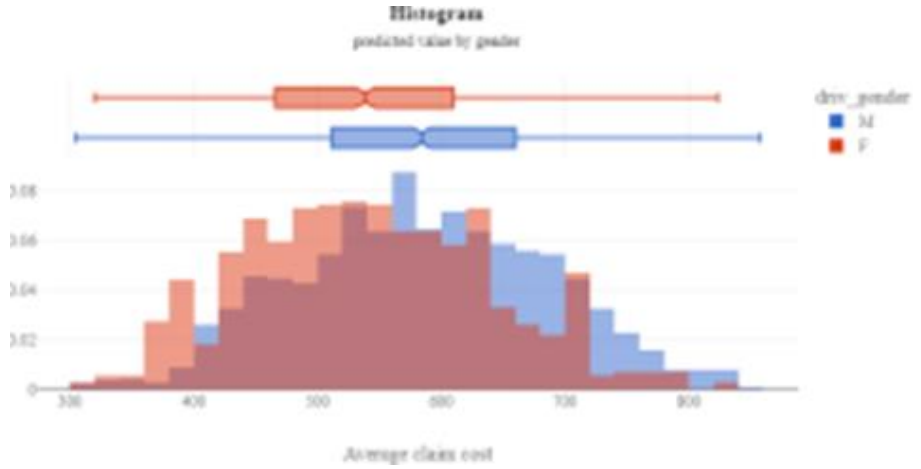


FIGURE 3. Predicted average cost by gender.

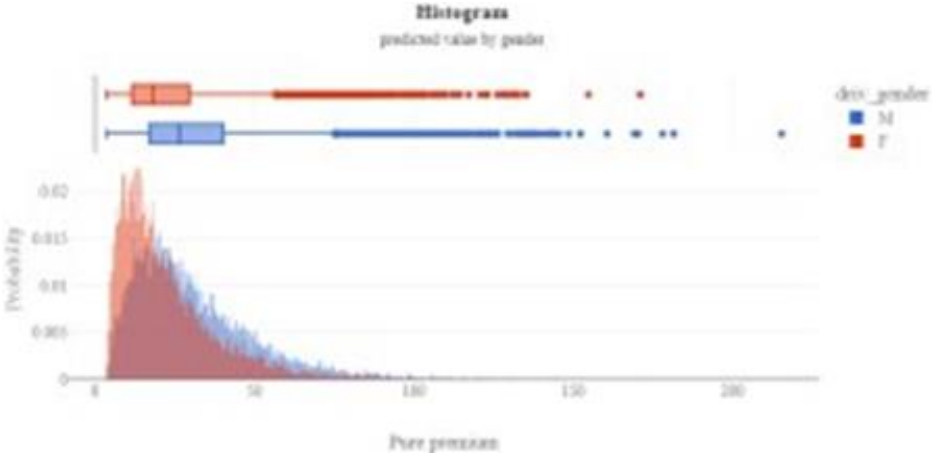


FIGURE 4. Predicted pure premium by gender.

9. BIAS MITIGATION IN INSURANCE PRICING

Standard forecasting models were used to assess impartiality beside productivity. The following point requires deploying alleviation procedures focused at decreasing imbalances al- though upholding prognostic precision. Several frameworks are assessed, as explained beneath.

A. Total Deletion Strategy

To assess the impact of explanatory variables on bias, depen- dencies between features are analyzed. Specific variables are excluded based on dependency thresholds. Figure 7 visualizes the examined scenarios.
Among these scenarios, Scenario 7 (dbe+veh price) offers a balanced trade-off, reducing bias by 37% while maintaining a slight 3.9% reduction in model accuracy. This refined model is further examined for its overall effectiveness.
Performance shifts indicate that reliance on the remaining variables increases, with features such as zoning, driv ly, and veh age playing a more prominent role. However, minor dis- crepancies in equilibrium pricing emerge, requiring additional validation before deployment.

Correlation Remover

Following the methodology in Section 3.1.2, categorical variables remain unaltered, while numerical attributes un- dergo transformation. The selected quantitative features in- cludeveh price, weight kw, veh age, driv yp, and driv ly. A uniform set of 100 values for the hyperparameter α within

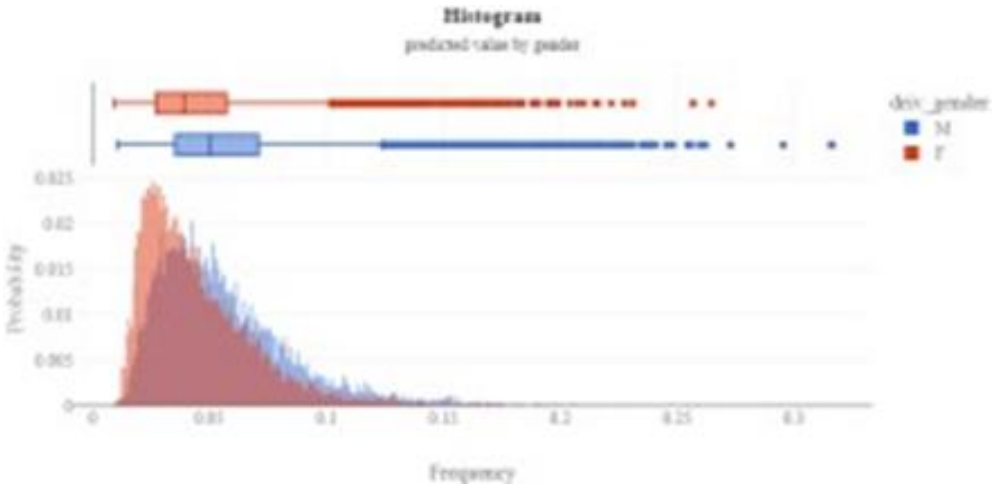


FIGURE 5. Predicted frequency by gender.

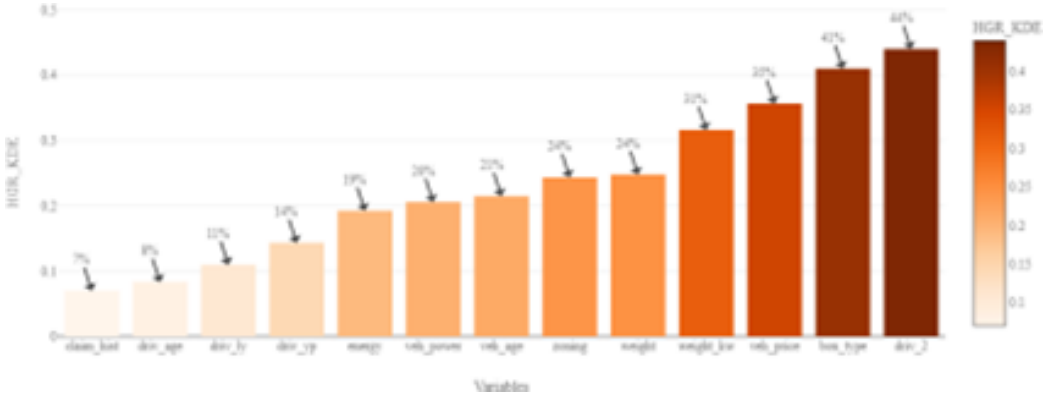


FIGURE 6. Interdependencies between variables and sensitive attributes

the range [0, 1] is tested to account for both extremes. Each setting of α modifies the selected attributes, allowing fairness and performance metrics to be assessed after model training. The analysis yields two primary observations. Firstly, no lin- ear relationship is evident between fairness levels and α. While an inverse relationship between fairness and performance was anticipated, model behavior remains inconsistent. Secondly, specific α values enhance both fairness and accuracy. For instance, setting α = 0.85 achieves an 11% improvement in fairness while marginally boosting predictive performance by 0.5%. Despite these findings, certain constraints emerge within the insurance

pricing context. Pricing models predominantly rely on discretized features rather than continuous values. Ideally, transformations should be applied before discretization to maintain interpretability, yet the residual nature of these transformations complicates direct interpretation. Furthermore, translating transformed customer attributes into final premium calculations poses practical difficulties. Additionally, the ab- sence of a clear relationship between α and model response introduces unpredictability in deployment. These factors sug- gest that the method may be more effective in cases involving continuous sensitive variables.

10. FAIR-SMOTE IMPLEMENTATION

To maintain a balance between bin quantity and subpopu- lation distribution, the segmentation process divides data into seven groups. The bin structure aligns with the observed trends in pre-
mium distribution. After segmentation, resampling is selec- tively applied to S|Y , ensuring that the target distribution

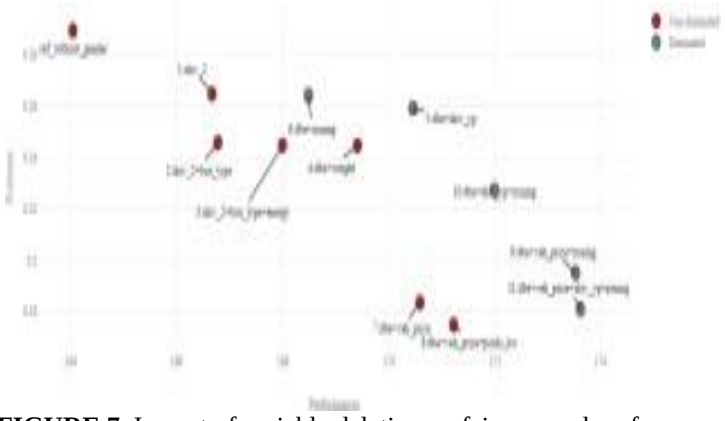


FIGURE 7. Impact of variable deletion on fairness and performance

remains unchanged. Since modifying Y directly impacts the distribution assumptions in GLM models, adjustments are limited to non-parametric models like random forests, were resampling Y leads to inflated premium predictions, reducing model reliability. The Fair-SMOTE technique utilizes threshold values of st = 0.8 and ft = 0.8, maintaining consistency with standard recommendations. The modified distribution after resampling is illustrated in Figure 9.
An additional 9,588 synthetic samples are introduced, in- creasing the training dataset size by 16.1%. This augmentation maintains the core statistical properties of the original data while addressing disparities linked to gender. Box plot analysis reveals improved alignment across groups, mitigating prior imbalances. Model performance assessments are summarized in Table 6.

TABLE 6. Model Comparison Post Fair-Smote Application.

Model	RMSE	HGR_KDE	LR
Baseline Model	164.21	28.97%	99.71%
Fair-SMOTE Model	164.88	28.05%	99.69%

Although the approach minimizes skewness in the training set, test data assessments reveal that residual biases persist. The fairness improvement remains moderate, under 4%, while predictive performance and log-likelihood ratios show negligible deviation from the baseline. These results indicate that balancing S|Y alone does not eliminate disparities, as historical gender-based influences in pricing persist beyond sample representation. Further sensitivity analysis explores refinements to the methodology. Varying hyperparameters ft and st does not yield significant gains in fairness. Similarly, modifying bin segmentation fails to enhance bias mitigation; increasing the number of bins distorts the target distribution, negatively impacting predictive accuracy. A critical deviation from conventional approaches is the decision to resample exclusively within S|Y rather than mod- ifying the overall Y distribution. Attempts at full resampling disrupt structural consistency in Y, generating excessive syn- thetic data without improving fairness outcomes. A partially rebalanced Y approach is explored, with performance com- parisons detailed in Tables 7–10.

TABLE 7. Initial distribution

Y/S	F	H	Total
25 €	250	150	400
35 €	120	170	290
Total	370	320	690

TABLE 8. Balanced on S/Y

Y/S	F	H	Total
25 €	250	250	500
35 €	170	170	340
Total	420	420	840

Expanding the number of synthetic instances by 20% results in an increase in RMSE to 177, while fairness remains at 28.05. This indicates that an excessive amount of artificially generated data can degrade model accuracy. Z uture research could focus on resampling conditioned on multiple explana- tory variables $S|X_1, ..., X_p, Y$ instead f relying solely on $S|Y$. The selection of appropriate variables is critical, as improper choices may reinforce existing biases. Maintaining consistency in premium distributions while accounting for interdependencies among variables remains a key area for further exploration

11. EXPONENTIATED GRADIENT

Optimizing a model to achieve the lowest possible error within the constraint ζ follows a systematic methodology. Initially, the smallest tested value, set at 10−5, is progressively increased in powers of ten until the model converges. The first instance of convergence is observed at $\zeta = 10^{-1}$. Once the optimal solution is identified, reliance on grid search and the M parameter becomes unnecessary. However, for $\zeta < 10^{-1}$, these tools help uncover potentially superior solutions com- pared to those obtained at higher tolerance levels. During experimentation, M values ranging from 0.5 to 500 were analyzed using a grid containing 3000 points. Despite extensive testing, none of the results surpassed those achieved at convergence. The dataset's high dimensionality-imposed constraints on the grid search, influencing these outcomes. Increasing computational resources and allowing for longer processing times might improve model performance. The outcomes for the best-performing model are summarized in Table IX.

TABLE 9. Evaluation of The Exponentiated Gradient Approach.

Models	RMSE	HGR_KDE	LR
Reference Model	164.21	28.97%	99.71%
Model Post-Mitigation	164.30	31.74%	99.70%

Post-mitigation, the model demonstrates a slight reduction in efficiency, accompanied by a noticeable increase in disparity compared to the baseline. While the overall decline in perfor- mance is minimal, the HGR measure shows a 9% uptick after applying mitigation techniques. The imposed error constraint does not effectively enforce statistical independence between Yb and S, highlighting that parity in error rates across gender groups does not inherently ensure equitable outcomes. A key limitation of this method is its inability to accommo- date broader fairness metrics. Additionally, the computational burden intensifies significantly, growing at an exponential rate. Alternative strategies such as grid search and tuning the M parameter offer potential solutions but lead to less optimal results. Bias mitigation within modeling frameworks remains a challenging task, with most existing methodologies tailored for binary classification problems. The substantial level of customization required for adaptation adds to the complexity. Despite these challenges, the adopted approach underscores the intrinsic difficulties of embedding fairness constraints within optimization models. The development of robust and scalable fairness mitigation techniques continues to be a pressing mathematical challenge, particularly in applications involving pricing models.

12. FAIR REDISTRIBUTION

To evaluate individual fairness, an adapted flip test method- ology is utilized. Refinement of the k-nearest neighbors ap- proach through hyperparameter optimization enables the se- lection of optimal variables aimed at minimizing discrepancies among individuals within the test dataset. The primary goal is to construct a model that mitigates bias while preserving similarity in individual characteristics. The chosen variables for this approach encompass: drivyp, driv 2, veh age, veh price, box type, and zoning. A neighbor- hood size of five is adopted, utilizing the Manhattan distance metric ($\ell 1$). The k-nearest neighbors' algorithm in Python, which dynamically selects an appropriate neighbor search technique based on dataset dimensions and structure, produces the best results. For each observation, the mean deviation from opposite- gender neighbors is calculated and integrated into the modeling dataset. An experimental assessment conducted on a test dataset containing 14,000 instances reports total costs of C359,584, while predicted premiums amount to C360,884. The resulting Loss Ratio (LR) of 99.64% closely mirrors the overall LR of 99.7%. On average, premiums assigned to women are C0.8 lower than those assigned to men. Key aggregates derived from the test dataset are summarized in Table X.

TABLE 10. Aggregated statistics post-redistribution

Aggregates	Female	Male	Sum
Total Expenses (€)	154,880	204,704	359,584
Predicted Premiums (€)	154,953	206,931	360,884
Exposure	4776.6	7114.6	11,891.2
Individuals	6069	7931	14,000
Average ε (€)	-0.8	1.2	0.4
Sum of Bias ($\sum s$) (€)	-17,448	22,029	4,581

The cumulative fairness bias assessment reveals that, on average, women incur C17,448 less in costs than their male counterparts with comparable characteristics. Although expected risk differentials align with these findings, residual discrepancies persist due to dataset composition and inherent model limitations. To explore the impact of different parameter choices, a comprehensive grid search was performed, where η varied from $\{2,3,4,5,6,7,8,9,10\}$ and ζ ranged from
$\{2500,2000,1500,1000,500,100,10,1,0.1\}$.

For values where $\zeta \leq 100$, significant modifications result in premium distributions clustering around the mean. While this improves fairness, it comes at the cost of predictive accuracy. A trade-off between preserving distributional structure and limiting undue variation is necessary. Among the 81 tested configurations, only 10 produced viable solutions.
However, five exhibited integrity levels below 25

- $\eta = 6$, $\zeta = 2000$, global variation = C873, integrity = 88
- $\eta = 5$, $\zeta = 2500$, global variation = C771, integrity = 86

The configuration with 88% fidelity is preferred, as it reduces the total disparity to C1,079, representing a 76% reduction from the initial C4,581. Adjustments lead to an average premium difference of C0.071 for women and C0.14 for men, maintaining consistency with prior distributions. Figure displays pre- and post-adjustment histogram overlaps. Following redistribution, premiums maintain a Root Mean Square Error (RMSE) of 164.72, closely matching the baseline value of 164.21. Although total premiums increase by C873, the Loss Ratio (LR) experiences a marginal decline to 99.39%. When $\zeta \leq 100$, computation times remain under five minutes, making execution speed an inconsequential factor in strategy selection. The broader analysis highlights the necessity of mitigation techniques that sustain performance while ensuring fairness. Among the available approaches, variable suppression and redistribution prove particularly effective due to their straightforward implementation and adaptability.

13. CONCLUSION

Achieving fairness in pricing remains a crucial consideration. Whether driven by regulatory requirements or strategic initiatives, addressing disparities is integral to establishing justifiable models. Consequently, this study conducted a

mathematical investigation of fairness, implementing diverse metrics to detect biases at different modeling stages without relying on conventional demographic attributes. A range of mitigation strategies was then explored, involving interventions before, during, and after model development. These methods included data preprocessing, fairness constraints, and direct premium adjustments. While some approaches yielded superior out- comes, each provided valuable insights into bias reduction. Preserving predictive performance while mitigating unfair- ness poses an ongoing challenge. Completely eradicating disparities under the selected fairness measures proved unfeasible without significant performance trade-offs. For example, extending variable exclusion or intensifying fair redistribution iterations could have further reduced bias, yet the stochastic nature of datasets and evolving portfolio compositions suggest that absolute fairness remains elusive. Future work should expand the scope of evaluation by applying these methods to varied datasets and pricing contexts to improve robustness and establish reference benchmarks. Additionally, refining strategies across different pricing granularities such as distinct portfolio segments or coverage categories—could enhance bias mitigation effectiveness. This research establishes a foundational approach for identifying and addressing biases in pricing models. Further enhancements could focus on refining mitigation strategies within the modeling process and incorporating multiple sensitive attributes simultaneously. Advancing these methods will contribute to the continuous evolution of equitable pricing frameworks in the industry.

REFERENCES

[1]. G. S. Becker, *The Economics of Discrimination*, Chicago, IL, USA: University of Chicago Press, 1957.

[2]. A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, "A reductions approach to fair classification," in *Proc. 35th Int. Conf. Machine Learning (ICML)*, 2018, pp. 60–69.

[3]. C. Alycia and X. Wu, "The causal fairness field guide: Perspectives from social and formal sciences," *Frontiers in Big Data*, vol. 5, 2022.

[4]. J. Angwin, J. Larson, L. Kirchner, and S. Mattu, "Machine Bias," *ProPublica*, 2016. [Online]. Available: https://www.propublica.org

[5]. M. Ayuso, M. Guillen, and A. M. Pérez-Marín, "Telematics and gender discrimination: Some usage-based evidence on whether men's risk of accidents differs from women's," *Risks*, vol. 4, no. 2, p. 10, 2016.

[6]. R. Berk et al., "A convex framework for fair regression," *arXiv preprint*, arXiv:1706.02409, 2017.

[7]. R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth, "Fairness in criminal justice risk assessments: The state of the art," *Sociological Methods & Research*, vol. 50, no. 1, pp. 3–44, 2018.

[8]. E. Black, S. Yeom, and M. Fredrikson, "Fliptest: Fairness testing via optimal transport," in *Proc. Conf. Fairness, Accountability, and Transparency (FAT)*, 2020, pp. 111–121.

[9]. S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge, U.K.: Cambridge Univ. Press, 2004. doi:10.1017/CBO9780511804441.

[10]. A. Castelnovo et al., "A clarification of the nuances in the fairness metrics landscape," *Scientific Reports*, vol. 12, 2022. doi:10.1038/s41598-022-07939-1.

[11]. A. Charpentier, *Insurance, Biases, Discrimination and Fairness*, Springer, 2024. doi:10.1007/978-3-031-49783-4.

[12]. A. Charpentier, E. Flachaire, and E. Gallic, "Causal inference with optimal transport," in *Optimal Transport Statistics for Economics and Related Topics*, Springer Verlag, 2023.

[13]. A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big Data*, vol. 5, no. 2, pp. 153–163, 2017.

[14]. S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, "Algorithmic decision making and the cost of fairness," *arXiv preprint*, arXiv:1701.08230, 2017.

[15]. C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," *arXiv preprint*, arXiv:1104.3913, 2011.

[16]. F. Y. Edgeworth, "Equal pay to men and women for equal work," *The Economic Journal*, vol. 32, no. 128, pp. 431–457, 1922.

[17]. E. W. Frees and F. Huang, "The discriminating (pricing) actuary," *North American Actuarial Journal*, vol. 27, no. 1, pp. 2–24, 2023.

[18]. M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 29, pp. 3315–3323, 2016.

.