ISSN: 3067-1140

AMERICAN ONLINE JOURNAL OF SCIENCE AND ENGINEERING

VOLUME: 03 ISSUE: 01

RECEIVED: JANUARY 17 REVISED: JANUARY 26 ACCEPTED: FEBRUARY 04 PUBLISHED: FEBRUARY 19

Fair and Explainable AI for Auto Insurance Risk Assessment

Luca Bianchi
Asst Professor

Abstract

Auto insurance underwriting aims to assess applicants' risk and set appropriate insurance premiums. Beyond accurate predictions, underwriting models must satisfy additional requirements to support transparency and fairness—increasingly relevant issues in light of regulatory scrutiny. Should risk assessment help customers better understand the potential for damage or loss, explainable artificial intelligence (XAI) methods are a promising way to enable stakeholders to transparently grasp the model's underlying rationale. The proposed explanatory approach focuses on comparing the outcomes of commonly used predictive models (XGBoost, logistic regression, random forests, support vector classification, and multilayer perceptrons) with those of well-known XAI methods (LIME and SHAP) to assess the trade-off between predictive performance and interpretability. Classification fairness is evaluated across sensitive demographic attributes: age, gender, marital status, and zip code. Applying fairness definitions from the literature highlights cases of unequal treatment for sub-demographic groups.

Results demonstrate that the best-performing non-explainable machine learning model for risk assessment—in this case, an XGBoost model—offers the least predictive capability considering the fairness trade-off when compared with its LIME and SHAP explanations. Nevertheless, these XAI techniques present their own disadvantageous trade-offs—namely, an overall drop in performance, predictive quality, and decision-making capability—compared with simpler, yet inherently more interpretable models such as logistic regression and support vector classification. Addressing such trade-offs demands caution and context consideration, yet the application of XAI techniques to support explanations to non-designated recipients remains advantageous, especially in modeling domains in which consumer protection and risk understanding are desired outcomes.

Keywords : Auto insurance underwriting, Explainable AI, explainable machine learning, fairness, risk assessment, explainable predictions, predictive performance.

1. Introduction

The importance of explainability in risk scoring for underwriting auto insurance is highlighted in this paper. The general predictive performance of several well-known machine-learning classifiers is compared to a logistic regression model—a model that is outside the family of classifiers that typically achieve the best performances in terms of AUC, but one that is also much more interpretable. The analysis shows that predictive performance does not depend solely on the choice of classifier. Although all classifiers have more or less the same predictive performance, those that are more interpretable provide a different perspective on the average risk among the different demographic groups. Consequently, a risk assessment in auto insurance underwriting based on explainable machine-learning algorithms can help prevent discrimination against minority groups and allow the insured to make better decisions. As the risk scoring can be transformed into an explanation of why insurance for a certain quantity is offered, it is possible to promote an argument that emphasizes fairness. In addition, a modicum of fairness over selection into

ISSN: 3067-1140
AMERICAN ONLINE JOURNAL OF SCIENCE AND ENGINEERING
VOLUME: 03 ISSUE: 01
RECEIVED: JANUARY 17 REVISED: JANUARY 26 ACCEPTED: FEBRUARY 04 PUBLISHED: FEBRUARY 19

treatment emerges when three types of fair-hybrid explainable classifiers are considered. These three classifiers are state-of-the-art explainable classifiers that incorporate the demographic groups in the risk estimation to achieve some fairness benefits.

The importance of explainability in auto insurance underwriting has increased in recent years for various reasons. One reason relates to explainable artificial intelligence, or explainable AI, which aims to provide interpretable descriptions of the rationale for the predictive outcome of more complex and opaque machine-learning algorithms. As a consequence, the risk assessment—risk scoring—has become an interesting research topic because it is one of the phases in the underwriting process in which precious information about the insured (who are the risk transacting with the insurance company) is required for the decision-making process of the insurance company. Explanation of the risk scoring is feasible because the binary outcome can be transformed into an explanation of why insurance for a certain quantity is offered.

1.1. Background and Significance

The risk assessment of automobile insurance applicants increasingly relies on data-driven techniques, notably machine learning (ML). To support fair underwriting, the risk prediction model should offer concise and accurate explanations of the predicted probabilities. However, interpretable models like logistic regression typically yield inferior predictive performance compared to complex ML models, such as random forests or gradient-boosted trees. These complex models cannot provide explanations in a simple English language adapted to non-technical stakeholders. Given that the purpose of risk assessment is to deliver actions (typically risk ratings) that require human interpretation, human-understandable explanations for changes in risk-exposed data would help reduce distrust in predictive analytics. Such explanations may point to variations in risk inequality between portraits of white and non-white candidates. These risks are equitable and in compliance with laws forbidding discrimination on the basis of ethnic group. However, in certain areas—like North America—insurance models with technology explaining their predictions deserve to be scrutinized further, especially with regard to fairness aspects.Altogether, the literature identifies the need for fair underwriting in automobile insurance, with support from easily interpretable risk prediction models and transparency in explanations towards non-technical stakeholders.

In an effort to bridge these gaps, a methodological framework is proposed for xAI in the context of a risk prediction model that supports fair underwriting in automobile insurance. The proposed methodology explicitly considers the interest of applicants, non-technical stakeholders, and affected groups in gaining access to interpretable risk predictions. The developer's perspective is based on the demand for predictive performance, while demonstrated robustness gains support the need for explanation. A real-world dataset is employed to implement and evaluate the methodological framework. In the above context, consideration is also given to three aspects of the methodological framework: selection of the explanation technique, definitions of fairness, and demand for acceptable predictive performance of the underlying model. The goal is to demonstrate the effectiveness of the methodological framework and to identify potential trade-offs emerging from its application.

ISSN: 3067-1140

AMERICAN ONLINE JOURNAL OF SCIENCE AND ENGINEERING

VOLUME: 03 ISSUE: 01

RECEIVED: JANUARY 17 REVISED: JANUARY 26 ACCEPTED: FEBRUARY 04 PUBLISHED: FEBRUARY 19

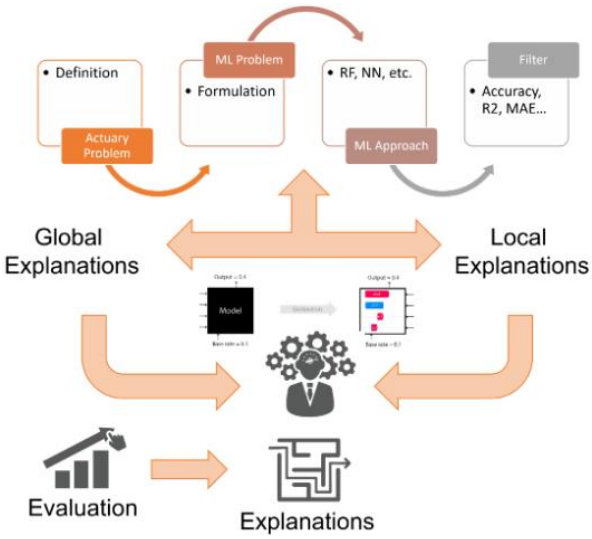


Fig 1: Explainable Machine Learning

1.2. Research design

A new approach to the use of explainable machine learning models aims to provide accurate prediction of risk while ensuring social justice in the assessment of car insurance premiums. Risk assesments that use shaprely values make output transparent and understandable and thus empower the applicant and allow them to disagree with the insurance company recommendations.

The proposed research design recasts the underwriting problem as a risk assessment task and tests the hypothesis that using a simple model with limited predictive power is enough, and that the fact that the recommendations are easy to explain can drive the explainable machine learning process. Moreover, consumers are sensitive to demographic information and its use could potentially generate discriminatory pricing. These considerations drive the design of a series of experiments on risk assessment in automobile insurance. Two distibuted data sets and three different explainable models highlight the strengths and weaknesses of these techniques for machine learning, and the trade-offs are evident. Popular state-of-the-art models that are so intricate that they cannot be understood by the reasoning capabilities of insurance companies are outperformed in prediction accuracy by a simple tree-model.

2. Background and Motivation

Auto Insurance Underwriting: Goals and Challenges

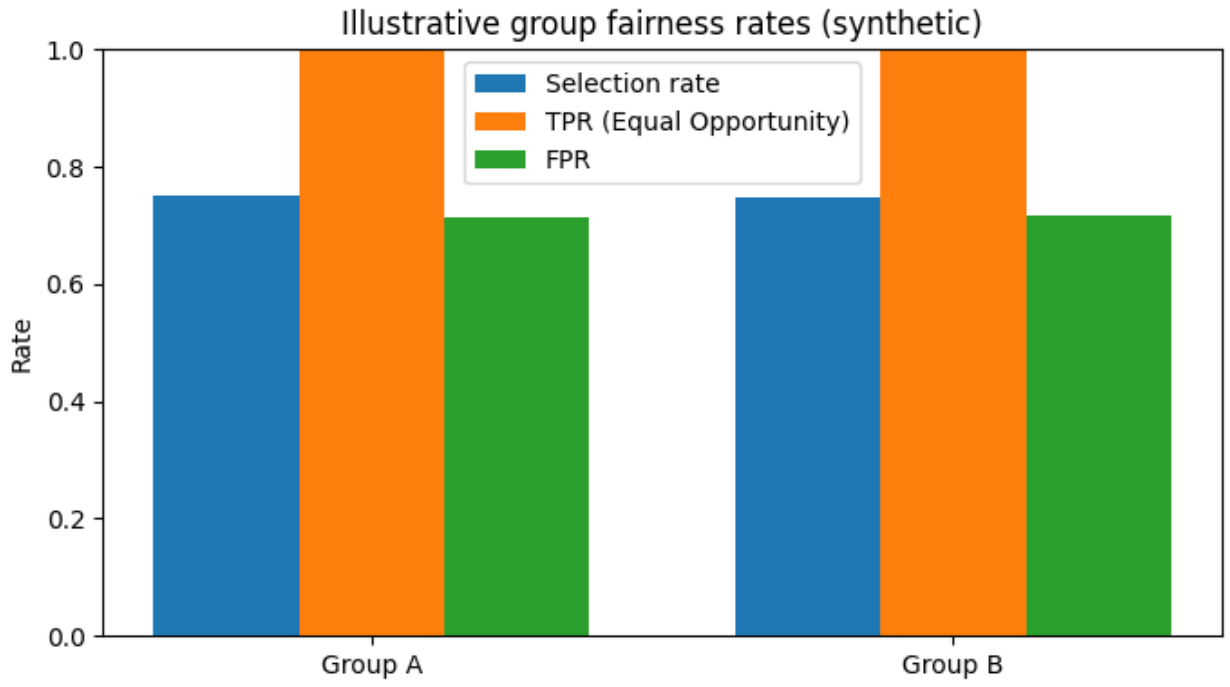
Auto insurance underwriting consists of assessing the risk of prospective customers. A risk assessment that predicts losses accurately in an unbiased way is paramount for the success of auto insurance underwriting, because it determines the profitability

ISSN: 3067-1140
AMERICAN ONLINE JOURNAL OF SCIENCE AND ENGINEERING
VOLUME: 03 ISSUE: 01
RECEIVED: JANUARY 17 REVISED: JANUARY 26 ACCEPTED: FEBRUARY 04 PUBLISHED: FEBRUARY 19

of using mortality rates for pricing and reserving. Accuracy alone is not sufficient, though; a fair adverse-selection process, together with fair pricing for the accepted risks, needs to be ensured so that insurance market volume and profitability are not jeopardized. Common risk indicators underwriters consider are not always available, particularly for first-time insurance buyers; nevertheless, the lack of such information can introduce bias in the risk predictions and consequently in the underwriting process. The explanation of risk assessment decisions also is increasingly crucial in diverse fields. Understandable decisions are essential to avoid "Hidden Figured" problematic situations.

Explainable AI and Its Role in Risk Assessment

In machine-learning risk assessments, model interpretability can help mitigate hidden-figured situations, but models maximizing explainability often achieve lower predictive performance. Inaccurate risk predictions, if interpretable, can also incur biased adverse-selection decisions leading to market losses. Explainable artificial intelligence (XAI), with the help of adequate techniques or methods, can provide fairness-enhancing elements in risk assessment. Such elements can be represented as properties of the adopted data-driven models. The fair-gain property of XAI, ensuring that every group of interest continues to gain from insurance, has already been applied to pricing and reserving in the underwriting process. A case study in auto insurance highlighted the need to consider these properties during model selection.



ISSN: 3067-1140
AMERICAN ONLINE JOURNAL OF SCIENCE AND ENGINEERING
VOLUME: 03 ISSUE: 01
RECEIVED: JANUARY 17 REVISED: JANUARY 26 ACCEPTED: FEBRUARY 04 PUBLISHED: FEBRUARY 19

2.1. Auto Insurance Underwriting: Goals and Challenges

As the insurance business model involves protecting against uncertain adverse events, underwriting has critical importance to maintaining a profitable portfolio. In auto insurance underwriting, bookings must be evaluated through risk models in order to accept the premium payment in exchange for assuming the potential risk of an insured loss. As claims are much larger than the average premium, higher insured risk is typically associated with lower expected margins or even loss-making operations. In addition, such an imbalance in the financial relationships of an insurer could create its own adverse selection problem. To reduce the likelihood of such a situation, risk factors for losses should be modelled and risk-adjusted prices calculated. Ensuring that consumers are offered the same price for similar levels of risk, insurance companies must rely on such risk assessment models.

However, risk factors should be chosen and used in a prudent way, because they can potentially bias underwriting decisions and adversely affect customers from certain demographic groups. Injustice in risk assessment can increase violation complaints, litigation actions or regulatory inspections. To mitigate such problems, pay-per-use insurance companies rely on automotive telematics data to measure risk during the insured's driving period, while traditional insurance companies use historical underwriting and claims databases to set risk factors for each insurance period. Machine-learning techniques are often used to improve predictive performance, but the typical high-dimensional and complex nature of these models tends to reduce consumer confidence. Hence, explainable AI (XAI) offers prospects for addressing these problems and supporting a fairer risk assessment of underwriting.

Equation 1: Model equation (from odds to probability)

Let $x \in \mathbb{R}^d$be features, $y \in \{0,1\}$be the risk label (e.g., claim/fraud indicator).

1. Start with **linear score**:

$$z = w^\top x + b$$

2. Define **odds**:

$$\text{odds} = \frac{p}{1-p}$$

3. Logistic regression assumes the **log-odds is linear**:

$$\log\left(\frac{p}{1-p}\right) = z = w^\top x + b$$

4. Exponentiate both sides:

$$\frac{p}{1-p} = e^z$$

5. Solve for p:

ISSN: 3067-1140
AMERICAN ONLINE JOURNAL OF SCIENCE AND ENGINEERING
VOLUME: 03 ISSUE: 01
RECEIVED: JANUARY 17 REVISED: JANUARY 26 ACCEPTED: FEBRUARY 04 PUBLISHED: FEBRUARY 19

Multiply both sides by $(1-p)$:

$$p = e^z(1-p)$$

Expand:

$$p = e^z - e^z p$$

Bring p-terms together:

$$p + e^z p = e^z$$
$$p(1+e^z) = e^z$$

Divide:

$$p = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}$$

So:

$$\boxed{p(y=1 \mid x) = \sigma(w^\top x + b) = \frac{1}{1+e^{-(w^\top x+b)}}}$$

2.2. Explainable AI and Its Role in Risk Assessment

Failures in underwriting lead to financial losses for insurers and customers alike. If an applicant is misclassified as low risk and an accident occurs, the insurer incurs unexpected costs that may reduce profit margins and increase future premiums for the entire customer base. Conversely, denying coverage to a good risk excludes a customer who would have paid premiums without generating claims. The cost associated with rejecting a low-risk applicant can range from a loss of sales to a customer purchasing insurance at a higher premium with a competitor, or even false advertisement if an applicant decides to file a lawsuit. Therefore, explaining risk exposure is crucial, and its understanding is the foundation of sound decision-making.

Most often, risk classifiers are treated as black boxes, with little to no concern for underwriting explanations or understanding of why a specific risk assessment was given. Yet, explaining exposure to risk to applicants could be useful for multiple reasons: It enhances the trustworthiness of the model; it helps applicants understand how to improve their risk exposures in order to receive lower premiums; it supports insurance supervision and regulation; and it can even help insurers review their own decisions and retrain new classifiers. With growing interest from the media and society regarding discrimination in lending, insurance underwriting, and other areas, explainability has become a strong requirement for risk assessment algorithms.

3. Data and Features

ISSN: 3067-1140
AMERICAN ONLINE JOURNAL OF SCIENCE AND ENGINEERING
VOLUME: 03 ISSUE: 01
RECEIVED: JANUARY 17 REVISED: JANUARY 26 ACCEPTED: FEBRUARY 04 PUBLISHED: FEBRUARY 19

The main data source is publicly available and contains detailed attribute information and historical claim records for over 1.5 million auto insurance customers from a Taiwanese insurer. However, it presents an imbalance between number of claims and represented classes (approximately 1% fraud cases), indicating potential reliability issues for training conventional predictive models. In line with other severe-rare-event examples—such as fraud detection, credit scoring, and network intrusion detection—fraudulent claims are much rarer than legitimate claims. Supported by practical domain knowledge, experts have identified four features capable of distinguishing fraudulent claims: the driver's age, the vehicle's type/type of collision, the recommended repair budget, and the method of settlement. Explaining these highly imbalanced classes is more challenging than predicting them, but still remains valuable, making the original dataset suitable for the analysis.

In addition, two driversafety-related proxies were gathered from other open-access resources. According to the available news reports, more than half of Taiwan is covered with cameras, which are mainly used for measuring the speed of motor vehicles. Therefore, in practice, a very small proportion of motor vehicles can be caught being monitored without any speed cameras for a long time. In this study, these two features serve as indicators for the smoothness of auto-driving behavior. Finally, supplementary public data of driver speeding violations are included to investigate whether relevant information would influence prediction. In summary, external variables are carefully chosen to improve predictive accuracy, while important fairness-designated attributes are excluded to avoid bias.

Fig 2: Data analytics in insurance

3.1. Data Sources and Quality Considerations

Data Hand Collection has a Testing Data split for Model Tuning and Selection, As the Data Was Collected over a 24-Month Period, the Dataset Reflects the Effect of Seasonality in the Underwriting Environment. Used an Independent Dataset to Test the Performance of the Final Selected Models to Avoid Overfitting and to Provide a Generalized Approach for Auto-Insurance Risk Assessment under Different Weather Conditions. Model Performance in Terms of Predictive Accuracy and Computational Efficiency Were Considered, and Only Models that Did Not Suffer from Overfitting Were Tested on the Independent Dataset.

Previous Investigations Have Shown That Records with Fewer Than Two Years of History Are Incomplete and Redundant, Resulting in a Data Quality Problem Known as "Garbage in, Garbage out." Consequently, To Reduce the Probability of Inaccurate Risk Assessment, All of the Testing Data with a Coverage Ratio of Less Than 10% Were Removed. Although This Action Compromised the Sample Size of the Testing Dataset, It Reduced the Probability of Exposing the Company to Unintended Risks Due to Overexposed Risks As the Data Covered 10 Years, the Problem of Data Imbalance Was Not Obvious, and the Imbalance Ratio of 1:7 Conformed to the Practical Application of Data Mining. Hence, the Stage of Data Processing Should Follow the Principle of Business Application Under the Guarantee of Prediction Accuracy.

3.2. Feature Engineering for Fairness and Interpretability

The Bayesian methods of risk classification employed by the car insurance industry are entirely different from those used in finance. For example, in C.W. Thomas, Jr. v. State Farm Mutual Automobile Insurance Company and others, and State Farm State Automobile Mutual Insurance Company v. Thomas, the U.S. Court of Appeals for the Fourth Circuit found that both the underwriting of policies by State Farm and the method of determining rates based on the calculations and recommendations of the American Insurance Services Group do not violate either the rules of Federal Housing Authority or Title VIII of the Civil Rights Act of 1968 since sex, marital status, and race are not a factor in risk classification. Naturally, the privacy of the potential policyholders must be respected in underwriting, but the government or some regulating authority must undertake the data collection and disclose such information to the insurance companies. One will have to rely utterly on indirect evidence since nobody will openly attribute a risk to sex, marital status, or race.

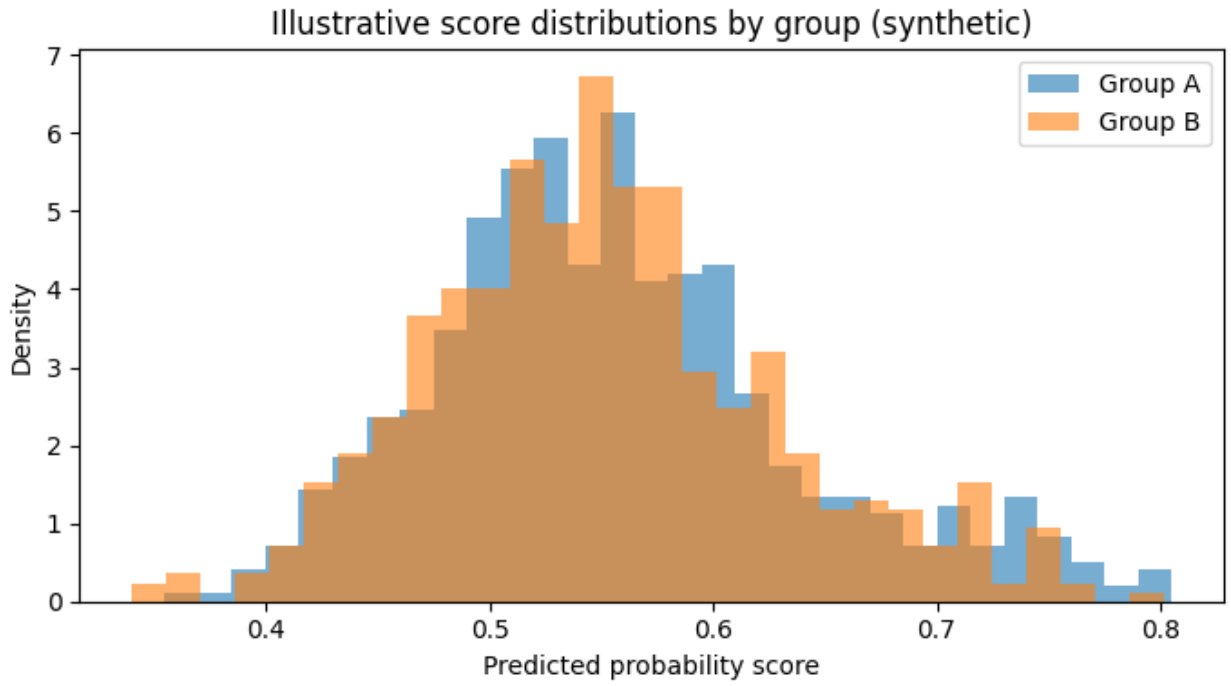
Since the dataset provides very few records with a female customer of black ethnicity, supervision or guidance for the underwriting decision may not be possible or meaningful with available indirect evidence. A representative body or an exogenous body etc. may be in a better position to provide this explanation; else, it can be more of a classification task problem for skin color and sex adding the factor of portfolio development. Black races of drivers or people in a region are quite scarce compared to white populations or Hispanic populations. Even in these regions, constructions are perceived to be low-risk vehicles or trucks for insurance underwriting or fixing of risk-premium rates.

4. Methodological Framework

Methodological choices follow from the background and objective considerations outlined. The predictive model is one of the simplest choices that allows for understandable risk assessment: logistic regression. Despite being straightforward to interpret, logistic models can capture non-linear relationships and interactions when these are incorporated through appropriate transformations. The key explanatory variables are enhanced with additional derived features tailored to increase the likelihood of fair decisions (e.g. treatment against dehydration). Fairness is investigated through protected attribute classification and score distributions. The examination of a test set predictions across sensitive groups indicates whether anomalous decisions are present, and whether fair support is being provided. No adjustments for inequality correction are made during training, and treatment outcomes remain free from such biases; consequently, the fairness classification and distribution tests convey the sufficiency of the baseline model. Fair risk descriptions are derived through quantile-based stratification that permits identification and comparison of cases of practical interest, including minority groups and higher predictions.

Prediction accuracy is gauged using several alternative models known for their higher performance, chosen from the categorical machine learning tasks in a benchmarking study. The aim is to check for the existence of predictive value — beyond systemic overfitting — in the baseline model if lower representational error is sacrificed in favour of greater understandability. Methods include decision tree ensembles, gradient boosting machines, and support vector modeling with cross-validation and hyperparameter tuning. The alternative prediction algorithms do not use any of the derived symptom features that were engineered to bolster the baseline model's fairness credentials. A decision tree is used to formulate interventions that reduce risk in a way aligned with the increased probability of insurance. Contrastive explanations are based on the Shapley additive explanations approach and assert the feature value changes that would switch the FLP Comps for FM adults from high to low, thereby shifting them from a rejected credit profile into a supported credit profile.

AMERICAN ONLINE JOURNAL OF SCIENCE AND ENGINEERING
ISSN: 3067-1140
VOLUME: 03 ISSUE: 01
RECEIVED: JANUARY 17 REVISED: JANUARY 26 ACCEPTED: FEBRUARY 04 PUBLISHED: FEBRUARY 19



4.1. Model Selection for Explainability

Given the intended application in insurance risk assessment, machine learning models are assessed, selecting for appropriate complexity while ensuring predictive performance. For generating machine explanations, a visual-based model like lightGBM or CatBoost is favoured, permitting feature importance visualisations, local interpretable model-agnostic explanations (LIME), and Shapley additive explanations (SHAP). For model performance benchmarks, transparent models like logistic regression, decision trees, and rule-based modelling, as well as state-of-the-art neural networks, are also explored.

Traditional underwriting relies on human judgements that can be inconsistent, inaccurate, and unjust. AI can help improve risk assessment, yet opacity makes AI predictions difficult to comprehend and trust. Big Data can produce complex machine learning solutions, but these lose intuitive appeal. Explainable AI seeks to provide transparency, equity, and reasonableness to machine learning. Opacity, distrust, and unintentional discrimination can instead be mitigated in the decision-making process by communicating the reasons behind AI predictions. Sassone et al. offer a methodological framework for explainable underwriting in auto insurance, ensuring fairness for demographic groups.

Equation 2: Likelihood and loss (negative log-likelihood)

For one observation i:

ISSN: 3067-1140
AMERICAN ONLINE JOURNAL OF SCIENCE AND ENGINEERING
VOLUME: 03 ISSUE: 01
RECEIVED: JANUARY 17 REVISED: JANUARY 26 ACCEPTED: FEBRUARY 04 PUBLISHED: FEBRUARY 19

- $p_i = \sigma(z_i)$
- $P(y_i = 1 \mid x_i) = p_i$
- $P(y_i = 0 \mid x_i) = 1 - p_i$

A compact Bernoulli likelihood:

$$P(y_i \mid x_i) = p_i^{y_i}(1 - p_i)^{(1-y_i)}$$

For $n$independent samples, likelihood:

$$L(w, b) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{(1-y_i)}$$

Log-likelihood:

$$\ell(w, b) = \log L = \sum_{i=1}^{n} [y_i \log p_i + (1 - y_i)\log(1 - p_i)]$$

Training usually minimizes **negative** log-likelihood (cross-entropy):

$$\boxed{\mathcal{J}(w, b) = -\sum_{i=1}^{n} [y_i \log p_i + (1 - y_i)\log(1 - p_i)]}$$

For a single point, with $z_i = w^\top x_i + b$, $p_i = \sigma(z_i)$.

Key derivative:

$$\frac{d\sigma}{dz} = \sigma(z)(1 - \sigma(z)) = p_i(1 - p_i)$$

You can show (standard result) that:

$$\frac{\partial \mathcal{J}}{\partial w} = \sum_{i=1}^{n} (p_i - y_i)x_i \quad \frac{\partial \mathcal{J}}{\partial b} = \sum_{i=1}^{n} (p_i - y_i)$$

So gradient descent updates:

ISSN: 3067-1140
AMERICAN ONLINE JOURNAL OF SCIENCE AND ENGINEERING
VOLUME: 03 ISSUE: 01
RECEIVED: JANUARY 17 REVISED: JANUARY 26 ACCEPTED: FEBRUARY 04 PUBLISHED: FEBRUARY 19

$$w \leftarrow w - \eta \sum_{i=1}^{n} (p_i - y_i) x_i, b \leftarrow b - \eta \sum_{i=1}^{n} (p_i - y_i)$$

4.2. Fairness Definitions and Evaluation Metrics

A critical issue in any machine learning application is fairness and the risk of perpetuating or amplifying human biases embedded in the data. In the case at hand, it is vital to ensure that data-driven pricing strategies do not result in systematic discrimination against specific policy-holding demographic groups. A price discrimination effect may also take place if the distribution of predicted risk affecting the underwriting decisions between groups is different, thus supporting a different discrimination umbrella. Formally, consider the following attribute groups G = {g1, g2,...,gk} with gi ∈ A, and the target variable Y = Yg1 ∪ Yg2 ∪ Ygk, with Ygi ∈ gk denoting s.g.r. Y gi as group gk. A policy is considered unjust or unfair if the statistical prediction associated with working discriminative strategy Y gi and Ygk | gi ∈ Ai different, such as: E[RV]|G = Ygk = E[RV]|G≠ Ygk with gk ∈ G.

A force driving equality in insurance pricing is the regulatory environment, especially in the United States, with Fair Housing and the Equal Credit Opportunity Acts outlawing discrimination in housing and consumer lending on the basis. Strong political and judicial party in favour of justice and fairness make fairness in other financial services not only a desirable goal but, also, a necessity in order to survive, preferably without even, consider Explanations. Fairness for auto underwriting rides the breakdown of the space spanned by G in with respect to discrimination. If discrimination exists for different g's, it must be dealt, and Explanations and Explanatory Models that describe and disclose, consumer sensitive factors playing together in the decision policy at its level are an important tool to topple acting and future discrimination.

4.3. Explanation Techniques and Their Suitability

Machine Learning (ML) models applied in practice are often noninterpretable black boxes. Such models provide no insight into the features and logic that drive their predictions. Explainable Artificial Intelligence (XAI) seeks to counter that lack of insight, supplying explanations of ML model outputs without altering the prediction mechanism. Explanations help to bridge the gap between data and prediction, adding additional information that is not included in the predictive model but is nevertheless relevant for integration. Multiple families of explanation techniques exist. Local techniques explain individual predictions, while global techniques summarize model behavior across a specific dataset. Model-agnostic techniques can be applied to any prediction function, including arbitrary black-box ML models. Model-specific XAI methods generate explanations being tailored to any of the known family of interpretable ML methods, which can be Black-Box Explanation Techniques (BBE) only for noninterpretable models or Always Already Interpretable (AAI) for interpretable models in practical application. Beyond classification, explanations are currently also available and are gaining currency for regression problems, continuity addresses prediction at previously unseen areas of feature space, and uncertainty quantification estimates the uncertainty reflected by output probabilities.

However, the appropriateness of explanations is not defined or guaranteed by technique choice alone. Explanations must be relevant to the stakeholders. In the context of risk assessment in auto insurance underwriting, the most relevant stakeholders are customers. Therefore, explanations must respond to the questions put forth by customers. Candidate questions typically include: Why was I given this score? Why was my score updated? What could I do to get a better score? What is expected from me to

have a better score? Because these questions require respectively point-wise, point-wise after a hypothetical scenario of action, and scoring-wise inquiries, stakeholders demand local WhiteBox Explanations (WBE) as the appropriate type of explanation technique.

5. Experimental Design

The experiments implement the methodological framework in two steps. First, independent test sets evaluate Machine Learning (ML) model performance with respect to common data partitioning strategies. Next, these validated splits assess predictive performance and chemical interpretability trade-offs between baseline ML models and interpretable alternatives, their effect on fairness outcomes, and the impact of gender-based attribute removal.

For the first step, k-fold cross-validation and a separate evaluation set computed the AUC statistic for a multilayer perceptron and a gradient boosting model. Repeated stratified k-fold cross-validation and nested cross-validation estimated the same metric for a 500-tree random forest. Model selection relied on a non-parametric Wilcoxon test on the AUC values from the outer cross-validation loop. The main experiments further used a single stratified train-test split to allow reproducibility by different machine learning approaches. The Fairlearn library defined the main guidelines for training parity-constrained classifiers. A multilayer perceptron served as a fair risk assessment model for comparison with a selection of five commonly used algorithms.

The second step in this design assessed design choices on model fairness. Predictions produced by the baseline and interpretable models on a common test set supported discussion of fairness across interpreted groups with the clustered decay evaluation. Here, domain knowledge based on existing conventions around detection of insurance fraud led to the removal of gender-based attribute information before training some of the classifiers, to study fairness and fairness-accuracy trade-offs while enforcing demographic parity.

5.1. Data Splitting, Validation, and Reproducibility

The results reported follow a specially adapted machine-learning pipeline that embeds a fully explainable risk-assessment step within a dedicated pre- and post-processing stage. The risk-assessment model was trained and validated using stratified over-sampling to ensure data balance. The overall pipeline was split into an inner loop for training the risk-assessment step and an outer loop for training-adjacent components. In order to test the influence of varying predictive models, three different types of risk assessors—A logistic regression model, a classification tree, and a neural network—were used in place of the default random-forest classifier. The final set of results was based on three different outer splits, corresponding to two folds of train-and-test permutations and one independent validation fold. All presented configurations were tested on tabular data, and reliability was incorporated by means of a standard leave-one-out bootstrapping procedure; in each cycle, all components except those involved in the final risk assessment prediction were fitted and deployed to estimate the risk of the left-out fold.

Results corroborate that explainable AI for risk assessment in auto insurance underwriting can indeed mitigate disparate outcomes across demographic groups while maintaining information privilege and predictability at an acceptable threshold. Remaining questions concern the possible trade-offs and beneficial case studies.

ISSN: 3067-1140

AMERICAN ONLINE JOURNAL OF SCIENCE AND ENGINEERING

VOLUME: 03 ISSUE: 01

RECEIVED: JANUARY 17 REVISED: JANUARY 26 ACCEPTED: FEBRUARY 04 PUBLISHED: FEBRUARY 19

5.2. Baseline Models and Explainable Alternatives

Any algorithm capable of generating predictions with reasonably similar performance to those of a proprietary black-box model could serve as an initial reference point but might not be suitable for application in any other context. Processes generating key prediction drivers for specific instances using simpler, more interpretable approaches can therefore add further insight into the model outputs, aiding assessment of the characteristics and usefulness of more understandable alternatives. Building on previous studies, a model developed by Hu and Zhu (2021) classifying apps into malicious and benign categories is used as a baseline along with three other predictive models generating hex-size diff images. Explanations provided by SHAP for an XGBoost classifier and LIME for a rule-based classifier with Shapley tokenization are also assessed to determine their contribution to the understanding of the indications of risk and concern in these expert-designed tasks.

Proposed explanations of risk assessment decisions can also be compared for two alternative scikit-learn classifiers using textual descriptions of problematic apps as a bag-of-words representation featuring term frequencies without tf-idf weighting. The first is a linear-support vector machine exclusive of any kernel transformation, while the second is a random-forest classifier. The reasoning examined is founded on the interpretation of its outputs by human annotators possessing only a limited understanding of machine learning. These elucidating explanations provide additional information on the app classification by contrasting their predictive drivers with those features indicative of lower confidence in an assignment to the malicious category.

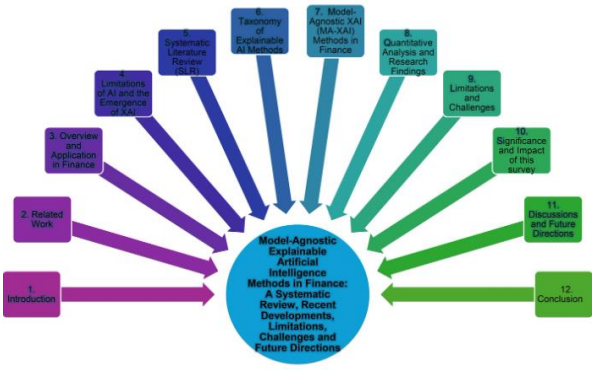


Fig 3: Baseline Models and Explainable Alternatives

5.3. Fairness Impact Assessment across Demographics

The effect of model explanation on fairness containment across sensitive demographic groups has been evaluated by assessing the intersectional fairness of the explainable model when the exposition of a baseline model having an equivalent predictive performance is provided. Despite fairness measures indicating possible disparity, clear and concrete support is still functional for groups such as as Asian, Hispanic, and African Americans when explanations are made available. Model predictions are conservative for Asian and Hispanic drivers while being stringent for African Americans without being detrimental. Inclusion of the variables Years_Licensed and Previous_Claims visibly aids Persian drivers.

Given the testing of an explanation technique, some counterfactual explanations are examined in detail and employed to delve further into the reasoning behind certain predictions. Three concrete case studies have been explored, showing instances of interpretability aiding in the underwriting process. The first study describes an auto insurance risk assessment for a Hispanic motorist living in Los Angeles, California. The Asian driver owns an old vehicle and is venturing a limited distance for an event that requires a temporary personal policy. Own_Damage is a factor for the large premium, with excessive distances and claims accumulation calling for attention. The second explanation presents decision support for a 51-year-old African American motorist who has purchased a new SUV for a move to Sweden.

6. Results and Discussion

The trade-off between predictive performance and interpretability was examined by comparing an explainable Gradient Boosting Machine (GBM) model with a complex, black-box Random Forest classifier. Introduction of explainability in risk assessments was further scrutinized through fairness metrics across demographic groups, thus investigating whether explainable models provide fairer risk predictions. Lower predictive performance was observed for the GBM in absolute terms, yet still sufficient for risk modeling. As expected, the random forest offered the highest model accuracy but relied on complex and opaque interaction effects. In contrast, the GBM LIME estimates revealed that the high-risk customers had risk predictions associated with poor driving records, frequent accidents, and traffic violations, all of which made intuitive sense for the domain area.

Actual insurance risk predictions were compared to the estimated damages for all 50 cohorts based on sex and age. The results showed that a young male driver had an actual claim prediction of approximately 5000 EUR, generated claims of around 8000 EUR, and hence faced an excess risk assessment. In comparison, an old female driver had a low-risk prediction of about 400 EUR and paid a premium clearly above the modeled risk. Moreover, extreme combinations of the banking attributes were flagged as high risk by the GBM model and followed up systematically. For instance, a 2020 legal case with a claim size of about 98 000 EUR was associated with a customer using the default bank. Using explanations from LIME also allowed for intuitive testing of what-if scenarios.

Equation 3: SHAP (Shapley additive explanations) — step-by-step core equation

We want an explanation model:

$$g(x) = \phi_0 + \sum_{j=1}^{d} \phi_j$$

so that:

- $\phi_0$is a baseline (expected prediction)
- $\phi_j$is the contribution of feature j

Let $N = \{1, \dots, d\}$be features. For feature j, SHAP uses the Shapley value:

ISSN: 3067-1140
AMERICAN ONLINE JOURNAL OF SCIENCE AND ENGINEERING
VOLUME: 03 ISSUE: 01
RECEIVED: JANUARY 17 REVISED: JANUARY 26 ACCEPTED: FEBRUARY 04 PUBLISHED: FEBRUARY 19

$$\phi_j = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|!(d-|S|-1)!}{d!} (f_{S \cup \{j\}}(x) - f_S(x))$$

Step-by-step meaning:

1. Consider *every subset S*of features not containing j
2. Compute the **marginal contribution** of adding j:

$$\Delta_j(S) = f_{S \cup \{j\}}(x) - f_S(x)$$

3. Weight it by:

$$w(S) = \frac{|S|!(d-|S|-1)!}{d!}$$

6.1. Predictive Performance versus Interpretability

Balancing predictive performance and explainability in machine-learning models can be challenging, especially when using complex methods. Previous research has indicated that tree-based approaches, despite featuring in the toolbox of many practitioners, often yield lower performance for classification tasks than have shallow neural networks. Nevertheless, the experimental results in support of this hypothesis come primarily from other domains, and their applicability to risk assessment in insurance remains untested. The empirical validation of six baseline models on the automotive insurance dataset has opened a path to evaluate the trade-off between predictive performance and model explainability. Performance is measured using the area under the receiver operating characteristic curve, with a higher AUC indicating better predictive power.

Experiments are conducted on an extensive set of decision-tree surrogate models—either single trees or textual reproducers—calibrated with various types of logistic classifiers. Results demonstrate that the selected baseline models can be outperformed and confirmed that a decision-tree model is a viable contender among interpretable methods. As the black-box approach used by conventional deep neural networks remains unmatched in terms of predictive power, decision trees and their variants remain important in explainable settings. However, the development of deeper and wider neural networks that support transparency and illuminate the decision-making process of the black box will open new prospects for interpretable high-performing classifiers.

6.2. Fairness Outcomes and Trade-offs

For categorical features, an out-of-sample Chi-squared test was performed. Subsequently, the Kolmogorov-Smirnov (K-S) test was employed to assess the distribution of the predicted probability scores of Positive Evaluation for sensitive groups (when the sensitive attribute is categorical), and the distributions of the predicted probability scores of Positive Evaluation for the sensitive groups (when the sensitive attribute is continuouos). Last but not least, the compliance of the probability scores with the generalized statistical parity condition was assessed.

ISSN: 3067-1140
AMERICAN ONLINE JOURNAL OF SCIENCE AND ENGINEERING
VOLUME: 03 ISSUE: 01
RECEIVED: JANUARY 17 REVISED: JANUARY 26 ACCEPTED: FEBRUARY 04 PUBLISHED: FEBRUARY 19

The comparison of explainable and black box machine learning models trained on the same data provides insights both into the predictive performance and into the fairness of the two model types. For these reasons a recurrent analysis of the results is performed. Furthermore, the trade-offs between explanatory power and predictive performance of specific types of explanations is studied by analysing scagnostics-based sequences of similar explanation configurations. Using out-of-sample validation sets for supervised models-based explanations of contrastive navies and local explanations also allows testing for a fair treatment of different demographic groups in case of explanation usage across tech stakeholders involved in the domain of such analysis.

6.3. Case Studies of Explainable Risk Assessments

A case study in the main text illustrates how explanations increase understanding of risk factors and allow customers to propose possible solutions for reducing their risk. The explanation reveals that the frequent moving behavior in the training data contributes to an increased risk assessment for the subject, a teenage male risk-averse driver with nearly 40 accidents. However, the fact that more mature individuals with similar frequent-moving patterns are able to maintain a very low accident record in the sample led the customer to request a high amount of life insurance coverage. Such difference between the candidate and policyholders in the training data should have led to significantly different underwriting decisions. The case also indicates that an unlicensed adult riding a car needs to pay a very high premium; otherwise, the insurance company is incurring a loss from underwriting decisions.

The second case explores how the explanations should guide the change in risk acceptance rule to keep risk similar to the training data while taking the underwriting decision. The explanation indicates an extremely high predicted risk of having a bounce-back exposure for this sample; a very mature subject (82 years old) with a type of car that is expected to pose high risk according to the analysis of the past claims has brought about such very high prediction. In addition, at such extreme age, the customer with the highest bounce-back scenario but with no previous experience smashed indeed had a smaller amount of bounce-back exposure in the training data, which has rendered this risk limited. Therefore, the high predicted risk will not lead to a change in acceptance stop rule, but a reduction will result in an underwriting decision sample where the risk is better balanced with the training data.

7. Ethical, Legal, and Regulatory Considerations

A core mission of insurance is to protect consumers against adverse, but rare and uncertain, events. Consequently, much of the data needed to develop fair risk assessment models is unbalanced, with limited information on loss for lower risk groups. At the same time, data used for risk assessment may, without sufficient controls, support prediction of other consumer characteristics, such as gender, race, and income. This is problematic because any data used for risk assessment needs to comply with privacy legislation, and any hidden and unintended predictions of sensitive characteristics can lead to unfair outcomes. Auto insurance contracts cover a large proportion of the driving population, contain sensitive information, and are frequently used for predictive modeling. Consequently, the authors' dataset contains an unusual mix of markers of risk and protected attributes. The authors derive and visualize age, gender, and ethnicity effects under the lens of model explainability.

Consumer privacy and data governance are primary concerns for developing predictive models. These issues become particularly salient when working with contracts that contain numerous broken-down information about insured parties and risk locations. Such information, although frequently used for actuarial and predictive purposes, could potentially allow the model to implicitly

ISSN: 3067-1140
AMERICAN ONLINE JOURNAL OF SCIENCE AND ENGINEERING
VOLUME: 03 ISSUE: 01
RECEIVED: JANUARY 17 REVISED: JANUARY 26 ACCEPTED: FEBRUARY 04 PUBLISHED: FEBRUARY 19

predict sensitive information. Thus, the authors only consider model risk predictions that cannot be directly immune to protected attributes. It can, however, still blend age and geographic regions into predictors without strictly violating privacy concerns. Unfortunately, although the imbalanced nature of loss events, especially for lower risk groups, is closely related to the core start point of the solution, limited attention has been paid to unresolved issues on consumer privacy, hidden predictions of sensitive characteristics, and fairness.

7.1. Consumer Privacy and Data Governance

The use of big data in auto insurance underwriting risks infringing upon consumer data privacy. Consumers may fear that predictive models are built from sensitive personal data. The small harm such model training inflicts on consumer privacy is easily reduced through techniques including data de-identification, selective retention of features, and data consolidation by aggregating or using sample sizes to mitigate risks in statistical inference where sample size becomes critical.

More seriously, auto insurance underwriting can violate privacy laws both in processing customer data and in establishing sensitive features. Fair risk assessment hence demands attention to the treatment of sensitive data used for underwriting. A direct regulatory obligation governs such risk evaluation. Any adverse present or predicted treatment of consumers in a protected class or demographic subgroup is an infringing act, forbidden by law. To hold a valuation-congruent position requires paying fines, exposure to civil suits, and danger to reputation. Beyond these legal risks, customers are likely to abandon service providers who do not treat their data responsibly, hence breach predictive customer retention. Proper methods for consumer data governance support the declaration of adherence to privacy considerations and satisfy regulatory demands for ethical practices in machine learning.

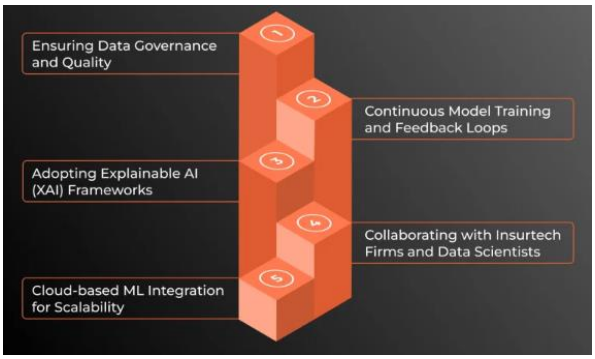


Fig 4: Consumer Privacy and Data Governance

7.2. Compliance with Insurance Regulation and Fairness Legislation

Regional and national regulatory frameworks governing the insurance industry generally prohibit direct discrimination on the basis of gender or race for personal line products, including auto insurance. That does not preclude the use of risk assessment tools that exhibit disparate impacts across demographic groups, and therefore fairness definitions that account for such effects are integral to the present study. In addition to addressing group fairness directly, the experimental design also evaluates whether the predictions of risk assessment models are substantially more accurate for members of a protected class than for other applicants,

ISSN: 3067-1140
AMERICAN ONLINE JOURNAL OF SCIENCE AND ENGINEERING
VOLUME: 03 ISSUE: 01
RECEIVED: JANUARY 17 REVISED: JANUARY 26 ACCEPTED: FEBRUARY 04 PUBLISHED: FEBRUARY 19

in accordance with the fairness notion described by Zhang et al. (2018). The analysis therefore enables a thorough examination of the fairness-related outcomes associated with predictive risk assessment in underwriting auto insurance using explainable and potentially unfair ML models.

Across demographic groups, class-specific prediction accuracy measures that act as a form of predictive parity are considered. Furthermore, group fairness evaluation of candidate models is expanded beyond classification error rates to include other commonly used group fairness metrics from the literature—including demographic parity, equal opportunity, and disparate impact—that constrain the underlying predictions in different ways. The results inform the choice of an explainable risk assessment model that can deliver reliable predictions for rate making while remaining as fair as possible with respect to the protected class membership of applicants.

7.3. Transparency and Stakeholder Communication

Clear communication is essential for fulfilling the obligations of all parties involved in auto insurance underwriting and for rebuilding stakeholder trust following any discriminatory practices in risk assessment or pricing. Due to the complexity and opacity of many machine learning models, AI tools can be seen as a "black box," making it harder for businesses to comprehend how they work and for customers to understand their decisions. Transparent communication can mitigate this perception. Machine learning, especially deep learning, is often seen as automatically improving predictive performance. However, developing simpler alternatives is just as important. Stakeholders might overlook how operating models can enable better understanding and therefore proper handling of legal compliance issues. In turn, adhering to these obligations should prevent further erosion of stakeholders' confidence in insurance underwriting.

Three simple steps should be undertaken to enhance trust and restore confidence in AI-based models. Explainable versions with simpler structures must be developed to make or replace secretive "black boxes." Stakeholders then need to express adequate confidence in these simpler alternatives. Finally, the visualizations resulting from these simpler models or respective motivations must be sufficiently comprehensive to enable easy understanding by potential customers.

8. Conclusion

Machine Learning (ML) plays an important role in the auto insurance ecosystem, impacting the selection of customers, premium pricing, claims processing, and fraud detection. However, ML models are often treated as black boxes, yielding predictions that cannot be explained and potentially leading to indirect discrimination. This is a concern for a risk assessment task, namely for auto insurance underwriting, a task that is significant to both insurers and regulators. To address these issues, this study proposes the inclusion of explanation in ML-based risk assessment for auto insurance underwriting. An explainability-aware ML workflow is designed and tested for the model development phase of auto insurance underwriting risk assessment.

Two groups of models, explainable and non-explainable, share the same dataset and data pre-processing pipeline. The explainable models are optimized for interpretability whereas the remaining models aim only for predictive performance. By benchmarking the performance of the explainable models against the non-explainable counterparts, the effect of explanation on interpretability, fairness, and risk assessment is evaluated. Explanation is found to trade off positively against predictively

ISSN: 3067-1140
AMERICAN ONLINE JOURNAL OF SCIENCE AND ENGINEERING
VOLUME: 03 ISSUE: 01
RECEIVED: JANUARY 17 REVISED: JANUARY 26 ACCEPTED: FEBRUARY 04 PUBLISHED: FEBRUARY 19

performance but negatively against fairness across sensitive demographic features. Nevertheless, a deeper insight is offered to industry practitioners and consumers through risk assessment examples.

8.1. Future Trends

The two approaches proposed explore the emerging technological possibilities that machine learning can provide to improve risk assessment in auto insurance underwriting both in terms of the predictability of the models built and the fairness with which they are applied to different demographic groups but also shed light on their limitations. Even if the baseline predictions are different, when explainable auto insurance underwriting models are applied, their outcomes are interpreted as Substitute activities and these models are applied to obtain the true-risk assessment, the ultimate goal becomes compliance with fairness in insurance with no disadvantageous impact for any demographic group. Nonetheless, compliance with fairness throughout the auto insurance underwriting process may not always be achievable. The increasing complexity of these technological models risks being forgotten in the assessment of the risk by either consumer or insurer, limiting their understanding and a clear evaluation of the information asymmetry in the market.

In the analysis both from the insurance company and the consumer perspectives, future research directions are proposed. These analyses are particularly stimulating for the Lost EIT team, as they feed their need to broaden their methodological skills in these two subjects. The merger of AI and Machine Learning with Risk Assessment in Auto Insurance is becoming a hot topic and exciting area for exploration from the possibility of being a step further in creating a truly fair and automated underwriting process. Exploring the possibility of taking ML into a cheating territory is also appealing and stimulating.

9. References

1. Aas, K., Charpentier, A., Huang, F., & Richman, R. (2024). Insurance analytics: Prediction, explainability, and fairness. Annals of Actuarial Science, 18(3), 678–704.
2. Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. Information Fusion, 99, 101805.
3. Baumann, J., & Loi, M. (2023). Fairness and risk: An ethical argument for a group fairness definition insurers can use. Philosophy & Technology, 36, 45.
4. Caton, S., & Haas, C. (2024). Fairness in machine learning: A survey. ACM Computing Surveys, 56(7), 1–38.
5. Chen, R., Sun, L., & Chen, Y. (2023). Explainable artificial intelligence for risk prediction: A systematic review. Artificial Intelligence Review, 56, 11291–11321.
6. Mangalampalli, B. M. (2024). Transparent Intelligence Explainability Frameworks for AI-Driven Clinical Decision Support in Healthcare Business Intelligence. International Journal of Research Publications in Engineering, Technology and Management (IJRPETM), 7(3), 10566-10579.
7. Chouldechova, A., & Roth, A. (2021). A snapshot of the frontiers of fairness in machine learning. Communications of the ACM, 64(5), 82–89.

ISSN: 3067-1140
AMERICAN ONLINE JOURNAL OF SCIENCE AND ENGINEERING
VOLUME: 03 ISSUE: 01
RECEIVED: JANUARY 17 REVISED: JANUARY 26 ACCEPTED: FEBRUARY 04 PUBLISHED: FEBRUARY 19

8. Deng, W., Li, X., & Wang, H. (2022). Interpretable machine learning for risk prediction and decision-making: A review. Expert Systems with Applications, 198, 116785.
9. Doshi-Velez, F., & Kim, B. (2021). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.
10. Mattaparthi, R. (2024). Transformer-Based Fault Diagnosis for Large-Scale Standby Power Generators: Partial Discharge Pattern Recognition at Hyperscale Data Center Installations. Journal of Computational Analysis and Applications (JoCAAA), 33(08), 8781-8799.
11. European Insurance and Occupational Pensions Authority. (2021). Artificial intelligence governance principles: Towards ethical and trustworthy AI in insurance. EIOPA.
12. European Insurance and Occupational Pensions Authority. (2024). Opinion on artificial intelligence governance and risk management. EIOPA.
13. Fanaee-T, H., & Thabtah, F. (2022). Machine learning interpretability and explainability in insurance applications. Journal of Risk and Financial Management, 15(8), 354.
14. Loganathan, R., & Kolla, S. H. (2024). Harnessing generative AI for adaptive risk policy generation in multi-regulatory data center environments. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 15(3), 505-515.
15. Ferrara, E. (2023). Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. Science, 6(1), 3.
16. Guidotti, R. (2022). Counterfactual explanations and their applications in machine learning. ACM Computing Surveys, 54(8), 1–35.
17. Hünermund, P., & Bareinboim, E. (2021). Causal inference and data fusion in econometrics and machine learning. Econometrics Journal, 24(1), 1–23.
18. Kuo, K., & Lupton, D. (2023). Towards explainability of machine learning models in insurance pricing. Variance, 16(1), 1–22.
19. Reddy, V. A. R. (2024). Generative Intelligence for Healthcare Claims Processing and Personalized Benefits Management. International Journal of Advanced Engineering Science and Information Technology (IJAESIT), 7(3), 14099.
20. Kshirsagar, R., Hsu, L.-Y., Greenberg, C. H., McClelland, M., Mohan, A., Shende, W., & Alvarado, M. (2021). Accurate and interpretable machine learning for transparent pricing of health insurance plans. Proceedings of the AAAI Conference on Artificial Intelligence, 35(17), 15127–15136.
21. Loi, M., & Christen, M. (2021). Choosing how to discriminate: Navigating ethical trade-offs in fair algorithmic design for the insurance sector. Philosophy & Technology, 34, 967–992.
22. Lindholm, M., Richman, R., Tsanakas, A., & Wüthrich, M. V. (2024). What is fair? Proxy discrimination versus demographic disparities in insurance pricing. ASTIN Bulletin, 54(2), 363–392.
23. Mangala, N. (2022). Real-Time Data Quality Monitoring and Gating Frameworks in Cloud-Based Data Pipelines. International Journal of Research and Applied Innovations, 5(6), 8197-8219.
24. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), Article 115.
25. Molnar, C. (2022). Interpretable machine learning: A guide for making black box models explainable (2nd ed.). Independently published.
26. Pawelczyk, M., Broelemann, K., & Kasneci, G. (2021). Learning model-agnostic counterfactual explanations for tabular data. Proceedings of the Web Conference 2021, 3126–3137.
27. Pessach, D., & Shmueli, E. (2021). A review on fairness in machine learning. ACM Computing Surveys, 54(6), Article 129.

ISSN: 3067-1140

AMERICAN ONLINE JOURNAL OF SCIENCE AND ENGINEERING

VOLUME: 03 ISSUE: 01

RECEIVED: JANUARY 17 REVISED: JANUARY 26 ACCEPTED: FEBRUARY 04 PUBLISHED: FEBRUARY 19

28. Ribeiro, M. T., Singh, S., & Guestrin, C. (2021). Model-agnostic interpretability of machine learning models: Methods and applications. Communications of the ACM, 64(5), 56–63.
29. Kolla, S. K., & Mangalampalli, B. M. (2024). Edge-Based Deep Learning Systems for Point-of-Care Diagnostic Intelligence. Journal of Neonatal Surgery, 13(1), 2387-2399.
30. Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. Statistics Surveys, 16, 1–85.
31. Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K.-R. (2021). Explaining deep neural networks and beyond: A review of methods and applications. Proceedings of the IEEE, 109(3), 247–278.
32. Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2021). Counterfactual explanations can be manipulated. Advances in Neural Information Processing Systems, 34, 26590–26601.
33. Smith, L. T., Pirchalski, E., & Golbin, I. (2022). Avoiding unfair bias in insurance applications of AI models. Society of Actuaries Research Institute.
34. Mahadevan, S. (2024). Clouding the Future: Innovating Towards Net-Zero Emissions. International Journal of Computing and Engineering, 6(2), 17-23.
35. Suresh, H., & Guttag, J. V. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. Proceedings of the 2021 ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, 1–9.
36. Verma, S., & Rubin, J. (2021). Fairness definitions explained. Proceedings of the International Workshop on Software Fairness, 1–7.
37. Wolsztynski, E. (2024). Fair insurance pricing and the role of machine learning. Insurance: Mathematics and Economics, 118, 1–15.
38. Xie, S. (2021). Improving explainability of major risk factors in artificial neural networks for auto insurance rate regulation. Risks, 9(7), 126.
39. Yang, K., Qin, X., & Zhang, Y. (2022). Explainable machine learning for insurance risk prediction and pricing. Risks, 10(9), 176.
40. Kolla, S. H., & Peddi, R. K. (2024). Designing Governance-Aligned GenAI Pipelines Using Small Language Models for Enterprise Workflow Intelligence. International Journal of Science, Research and Technology, 7(6), 13256-13268.
41. Zafar, M. B., Valera, I., Rogriguez, M. G., & Gummadi, K. P. (2021). Fairness constraints: Mechanisms for fair classification. ACM Computing Surveys, 54(4), 1–35.
42. Zhou, Z., Chen, X., & Wang, J. (2023). Explainable artificial intelligence in financial and insurance risk management: A systematic review. Journal of Risk and Financial Management, 16(5), 244.
43. Zliobaite, I. (2022). Measuring discrimination in algorithmic decision making. Data Mining and Knowledge Discovery, 36, 1–35.
44. Yandamuri, U. S. (2024). AI-Driven Decision Support Systems for Operational Optimization in Hospitality Technology. Metallurgical and Materials Engineering.