ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 02 ISSUE: 03

RECEIVED: JULY 25 REVISED: AUGUST 14 ACCEPTED: AUGUST 28 PUBLISHED: SEPTEMBER 22

AI-Powered Risk Scoring and Premium Optimization in Auto and Property Insurance

Elena Petrova

Asst Professor

***Abstract*—A unified perspective on risk stratification in insurance through the lenses of AI and premiums facilitates optimal risk-based pricing. Risk stratification improves upon traditional segmentation, notably via machine learning, in three areas: identifying and quantifying drivers of claims frequency or severity; reallocating risk among closed groups such that all members share risk characteristics; and estimating risk scores on individual customers based solely on their attribute values. Empirical studies show the added value for stratification of (1) modelling claims as a two-dimensional distribution, (2) incorporating policyholders' driving behaviour, and (3) considering spatial factors, although data quality may determine the feasibility of particular enhancements. Whereas risk stratification is central to allocating premiums properly, premium optimisation considers the entire pricing and underwriting framework: rate-setting, discount/surcharge application, customer lifetime value, and churn mitigation. Joint consideration also allows dynamic pricing—adjusting premiums in line with changing risk over time—while taking precautions against unfair pricing. All formulation dimensions are amenable to algorithmic solutions. Unifying the two perspectives reveals a clear schematic of one's research contribution, underlining that empirical applications in auto and property insurance harness AI-driven risk stratification within broader premium-optimisation frameworks. Successful projects yield novel insights and valuable lessons for future efforts.**

AI has opened exciting avenues for stratifying risk and optimally calibrating premiums in auto and property insurance. Risk stratification assesses underwriting risk for different groups and the entire portfolio, steering loss reserves and capital allocation. AI enhances traditional segmentation by identifying the main drivers of frequency and severity, reallocating risk among closed groups for equitable sharing of risk characteristics, and estimating risk scores for individual consumers—each with wider applicability. Empirical studies highlight the potential for two-dimensional frequency-severity modelling, incorporation of policyholders' driving behaviour, and consideration of spatial factors, while underscoring the paramount importance of data quality. Promoting stratification typically directs research attention towards the technology employed, whether neural networks, external data enrichment, or other areas.

Keywords*—AI-Driven Risk Stratification, Premium Optimization, Risk-Based Pricing, Machine Learning in Insurance, Claims Frequency–Severity Modeling, Individual Risk Scoring, Closed-Group Risk Reallocation, Dynamic Pricing Strategies, Underwriting Analytics, Customer Lifetime Value, Churn Mitigation, Behavioral Telematics, Spatial Risk Modeling, Data Quality in Insurance Analytics, Algorithmic Pricing Frameworks, Fairness in Automated Pricing, Auto and Property Insurance Analytics, Portfolio Risk Management, Capital Allocation Optimization, Intelligent Underwriting Systems.

I. INTRODUCTION (*HEADING 1*)

Rapid advances in artificial intelligence (AI) present opportunities to improve risk stratification and premium optimization in insurance. When designed with an orientation towards Downward Risk and considering market competitiveness, the AI-driven solutions can correspond to insurance regulatory principles that prohibit discrimination. Although risk levels nearly always correlate with underwriting and policy-holding capital set aside to guarantee coverage, the relationship is imperfect in practice due to rating structure, underwriting discretion, data availability, and changing exposures. Consequently, risk stratification can be enhanced by optimizing the bases of premium calculation and, more generally, by moving beyond traditional one-off static pricing towards real-time, multilayer AI-powered mechanisms.

Statistical models estimate the frequencies, severities, and costs of adverse events, while traditional rating factors are design features known before a purchase and are usually externally set by markets. AI methods can determine risk classes for car or property insurance; support recommendation of insurance companies; and decide the amount of money that motor drivers must pay for insurance. AI can optimize both risk stratification and pricing decisions. Stratification becomes more precise with proper use of company-specific data, while pricing becomes fairer and more profitable. Stratification and pricing simulations highlight important market factors that help to ensure the sustainability of a business with a downward Ohm risk appetite.

A. Summary of Findings and Future Perspectives

Core findings highlight that AI-based risk stratification increases model performance predictability and calibration, reduces adverse selection, and aids reserve management. Weak-points analysis for auto insurance indicates that risk-prevention measures should prioritise decreasing claim frequency. Analytical solutions for personalised profitability optimisation confirm operational feasibility and regulatory compliance. External risk models enhance property-insurance risk stratification and premium adjustment; the availability of suitable AI risk models for both lines of business indicates transferability. These results offer valuable implementation perspectives for data-rich insurance companies.

Future research should consolidate emerging capabilities in risk stratification for various insurance categories and explore the complete cycle of risk stratification, dynamic pricing, and risk-transfer management for multiple lines of business, thereby advancing transparency, regulatory compliance, and personalisation. Research gaps persist in the risk analysis of damage-exposed insurance policies and AI-driven pricing policy adjustment. Addressing these needs would enable the application of related risk models beyond insurance companies and caution against their use as stand-alone pricing or risk-transfer strategies.

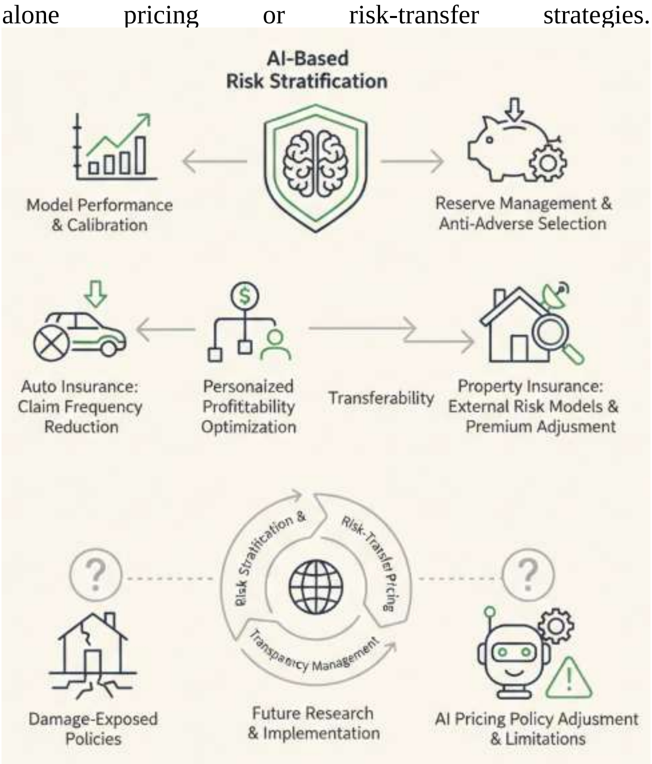


Fig 1: AI-Driven Risk Stratification in Insurance: Optimizing Predictability, Regulatory Compliance, and Dynamic Pricing Lifecycle

II. FUNDAMENTALS OF RISK STRATIFICATION IN INSURANCE

Risk stratification refers to the classification of customers, or more generally exposure units, according to their risk profiles. Technically, three types of risk stratification should be distinguished: risk segmentation, risk scoring, and risk clustering. These approaches differ primarily in the way that risk is represented: categorically, through risk labels, or through risk amounts (or exposure).Risk stratification is a fundamental step in the pricing, underwriting, and capital allocation processes. Prices are primarily based on the expected losses, but the importance of risk analysis extends well beyond premium estimation. The actual reserves that insurers must set aside, the capital that they need to allocate for solvency purposes, the potential discount or surcharge customers can receive (or pay) if dynamic pricing policies are applied, and the corresponding Customer Lifetime Value (CLV) are closely linked to the correct fulfillment of this task.

Historically, auto insurance pricing has involved a “black box” approximation of accident risk rather than a thorough investigation of its sources, making emphasis on the choice of the rating factors rather than on the relations between risk and accidents. Consequently, premium—fraud—discriminatory attitudes—price competition, and results were largely used to pick up the most appropriate level. Descriptive analysis has identified the main characteristics of risk allocation, although the combinations that really drive risk have remained elusive. While pricing is conducted by regulatory (or internally) imposed rules, an experimental perspective has provided partially different focus. Actual claims theory dynamics has examined interactions in frequency-severity structures, insurance services by branches, continuous-time models, and, more recently, claims migrations.

Equation 1) Claims as a two-dimensional distribution (Frequency–Severity)

1.1 Random variables and total loss

For a single policy over one exposure period (e.g., 1 year):

- N = number of claims in the period (frequency)

- $X_1, X_2, \dots, X_N$ = claim severities (costs), i.i.d. given a claim happens
- Total loss:

$$L = \sum_{i=1}^{N} X_i$$

1.2 Derive expected total loss $E[L]$

Use the **law of total expectation** conditioning on N:

$$E[L] = E\big[E[L \mid N]\big]$$

Now compute the inner expectation:

If $N = n$, then:

$$L \mid N = n = \sum_{i=1}^{n} X_i$$

By linearity of expectation:

$$E[L \mid N = n] = \sum_{i=1}^{n} E\,[X_i] = n\,E[X]$$

Plug back:

$$E[L] = E\big[N\,E[X]\big] = E[X] \cdot E[N]$$

So the **pure premium / expected loss cost** is:

$$\boxed{E[L] = E[N]\ E[X]}$$

This is the core actuarial engine behind "expected claims plus loading" mentioned in the premium framework.

1.3 Common frequency model: Poisson with exposure

Let exposure be e (e.g., policy-years). A standard model is:

$$N \sim \text{Poisson}(\lambda e)$$

Then:

$$E[N] = \lambda e$$

So:

$$E[L] = (\lambda e)\,E[X]$$

Per unit exposure ($e = 1$):

$$\boxed{\text{Pure premium} = \lambda\ E[X]}$$

1.4 Two-part (hurdle) decomposition

Sometimes severity is conditional on claim occurrence and frequency handles the "claim vs no claim" mechanism. If:

- $p = P(\text{at least one claim})$
- $E[L \mid \text{claim}]$ is the conditional expected cost

Then:

$$E[L] = p \cdot E[L \mid \text{claim}] + (1 - p) \cdot 0$$
$$\boxed{E[L] = p\ E[L \mid \text{claim}]}$$

A. Historical Approaches and Baseline Models

Risk assessment in insurance dates back centuries, with an extensive literature covering steady developments. Traditional actuarial techniques for risk stratification and premium optimization have become the basis for several industry-leading commercial software solutions. For auto insurance, a joint probability distribution of risks presented in matrix form has been the standard departure point for identifying and selecting rating factors and detecting interaction effects.

Development phases have included refinements in the predictive ability of the underlying models and methodological choices in the data-processing and analysis steps. Yet the historical approaches remain fundamentally the same, relying on the identification of relevant variables and a two-step procedure whereby the main mechanism is first analyzed separately and subsequently combined with other components. Both risk stratification and premium calculation have been treated independently. Recent publications explore a few primary risks of insured assets, structural pricing, and/or dynamic pricing. All present considerable scope for improvement.

Within this historical evolution, the insurance industry has employed fundamental actuarial models to guide practice continuously, thus serving as useful baseline models for comparison. In auto insurance, most of the pioneering research came from the United States. Its early foundations centered on controlling frequency and severity via predictive analysis. Rating assignment was subsequently achieved, initially through ad hoc procedures, followed by generalized regression techniques. Important explanatory factors in severity became increasingly recognized—first vehicle exposure, then driving behavior, the vehicle's accident record, and finally other environmental characteristics. With progressively larger datasets now available, rapid developments in machine learning tools enable exploration of these factors' conjoint effects and previously unconsidered variables.

B. Key Risk Factors in Auto Insurance

Risk aspects specific to auto insurance encompass basic attributes of the insured structures that precede the occurrence of claim causative events. When tackling the frequency-severity nature of auto claims, the target is the count—a ratio exposed to crash with additional factors that promote the occurrence of claims, even if those claims are not high-cost ones. These factors come from the driver (behavior of apprehension, motivation, attention), the vehicle (its innate characteristics and level of protection it offers), and the context of the road (physical context and risk environment). The relevant questions are: Do these factors matter? Do they interact with other elements in the process? Is there sufficient clean data for these dimensions? The outcome of data cleaning and quality assessment will determine any eventual gradual inclusion/consideration of proxied aspects.

Auto insurance can be examined under the prism of a frequency-severity dyad. Frequency is the more delicate issue at auto handling. Distribution-wise, it is usually analyzed with a Count Data Regression. The analysis should center directly on the count of claims and its determinism, while the additional severity factors add weight and cost to the premium assignment. Any model aiming to tackle the claim number should dissect the role of the different dimensions that govern this frequency attitude of driving. The more natural risk ratio exposed would lean toward claim-number determinants rather than claim-cost determinants. Therefore, a lower-risk interior segment would not hold point relevance if viewed under a pure exhaustible-cost prism.

Segment	Exposure (policy-years)	E[Frequency] λ	E[Severity] (₹)
Low risk	50000	0.03	45000
Medium risk	35000	0.06	65000
High risk	15000	0.11	90000

III. ARTIFICIAL INTELLIGENCE METHODS FOR RISK ASSESSMENT

Risk stratification in insurance may be supported through AI-supported methodologies. Risk stratification is the task of assessing how exposure units (e.g., drivers, properties, companies) differ in terms of either the probability distribution of risk incidents (e.g., claims) or of severity/cost, taking into account an appropriate measure of their apparent similarity. It can take three different forms: segmentation, scoring, or clustering (i.e. the use of an unsupervised algorithm to assign blocks of exposure units to clusters associated with higher or lower risks relative to the general average). Each has distinct purposes within insurance pricing, underwriting, investment, and reserving.

Data for AI-supported risk stratification come typically from different sources and need to be probed, cleaned, and normalized before entering the model-building phase. Deriving new features from existing ones is a key part of the process and the considered (real or proxy) variables need to be often down-selected. Supervised, unsupervised, ensemble, and deep-learning models fit different types of tasks. Besides predictive capabilities, selection should reflect external factors such as certification, auditability, or fairness considerations. Performance assessment on validation/test sets requires dedicated metrics and calibration procedures. The proportion of accidents predicted properly is seldom the most critical measure. When the event under examination is rare, like for the overwhelming majority of claims in auto or property insurance, predictive capabilities need to be assessed by means of different metrics that take into account the difficulty of predicting the minority class.

A. Data Sources and Feature Engineering

Data provenance, cleaning, normalization, feature construction, and handling of missingness, as well as potential proxies and dimensionality reduction.

Actuarial datasets are typically both rich and diverse, originating from a multitude of departments and systems. Their size raises data management and engineering challenges endemic to these areas of actuarial practice. In particular: during the design of machine learning or artificial intelligence models, processes including cleaning, normalization, denormalization, and redaction must occur before the model's execution. These steps are often formalized in a data pipeline. Furthermore, the act of modifying data is often accompanied by feature construction and name conception processes. Feature engineering can lead to feature sets that are large, too strongly correlated, or beyond the threshold of discrimination power needed for artificial intelligence models. As such, it is important to apply it judiciously by either reducing the number of features directly or creating an upstream process to select the most appropriate set, thus reducing the model's dimensionality.

Fig 2: Optimizing Actuarial Intelligence: A Multi-Stage Framework for Data Provenance, Feature Engineering, and Latent Risk Estimation

Data engineering continues after a model has been completed owing to the need to manage missingness in a data set. Missingness can result from factors such as data capture methods and sample size. Missing values can necessitate precautionary measures, such as the introduction of artificial intelligence models into the data pipeline as a curtain to fill in missing values. These preparations can be particularly beneficial in situations involving regression-based estimators of merged tails and segmentation approaches requiring a response variable for each observation. The choice of proxies is critical in the construction of an estimator of risk. These proxies attempt to identify the variable that the analyst cannot measure. Their use is a way of attempting to model the unobservable, reducing the size of the latent variable. Nevertheless, it is sometimes not possible to obtain a satisfactory variable for a given proxy due to the depression-consumption-homelessness cycle; a more suitable alternative at a greater geographical distance from the population can mitigate the problem, serving a higher level of insurance. Beyond proxies, dimensionality reduction remains an important step.

B. Machine Learning Models for Risk Stratification

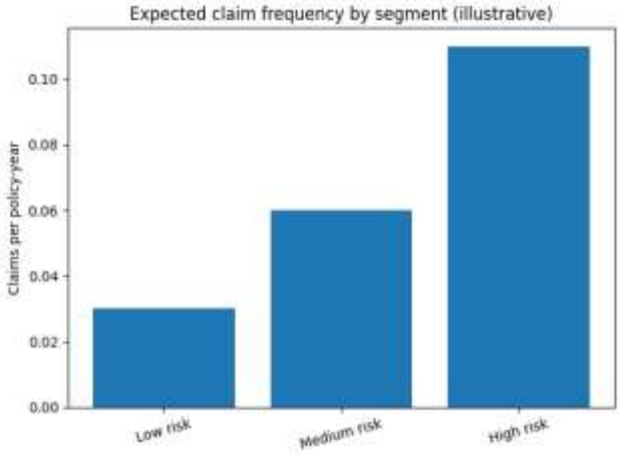
Supervised, unsupervised, ensemble, and deep learning approaches also constitute suitable candidates for risk stratification in insurance. Supervised approaches segment the underlying risk landscape, supervised and partially supervised approaches calibrate risk scores onto probabilities, semi-supervised and unsupervised methods mine latent structures in actual claims, while semi-supervised, unsupervised, and ensemble methods account for potential legitimacy concerns and fairness issues.

Model selection is driven primarily by classification performance, while evaluation considers operational reliability, susceptibility to attacks, scalability to high-dimensional data with differing sample density and data distributions, interpretability, and facilitation of quasi-automatic risk stratification. Typical considerations for commercial implementation include visual validation of results by domain experts, monitoring of real-world performance for drift detection, compliance with applicable regulation, and auditability of fairness and transparency for end customers.

IV. PREMIUM OPTIMIZATION FRAMEWORKS

The actuarial foundation for premium calculation resides in the construction of statistical models for claims frequency and severity prediction. Risk-based premium assignment corresponds directly to the actuarial logic, and it naturally affects reserve evaluation. For an insurance line covering many independent loss distributions, the application of the law of large numbers ensures that the loss ratio converges to the expectation as the exposure volume increases. Hence, the underlying premise of risk-based pricing is that a diverse portfolio containing various risks forms the optimal strategy for an insurer. In many jurisdictions, the fundament of risk stratification is further shrunk to a condition falls under pure premium, where the parameters involved in the construction of the model for operation can be only risk-based prices that done in sound actuarial approach. The only points of divergence from an insurance point of view are that the bases for pricing is done with sound actuarial basis and the pricing is not for maximization but for the purpose of regulation. The actual regulation is then remain only in the re-examining periodical whether they are explicitly excessive in the industry or not and the only result after a lot of book-keeping k are their approval to be rate filings under the legal requirement.

The majority of the suppliers of price changes the ground for which cascade down discounts or surcharges, treatment of shift or conceded premium already assumed into these function territories. This additional dimension for which the pricing changes in a dynamic and personal aspect is that of customer lifetime value recommendation. This breaks the partition of different departments and requires a continual looping validation between marketing, actuarial , claim and underwriting. With the integrated approach on-line model management become much important for insurance company, also from a marketing point of view any pricing recommendation require prior evaluation merging the area of loss minimizqation and increase of insurance naturality (churn). In this sense even if it is marketing that suggest a Dynamic Pricing change it is not marketing that manage the process but off-line separate loop that guarantee also a fair design of the price recommendation.



Equation 2) Risk stratification math: segmentation, scoring, clustering

Let policyholder features be a vector **x** (age, vehicle, location, telematics, etc.).

2.1 Segmentation (categorical risk labels)

Define a mapping:

$$g(\mathbf{x}) \in \{1,2,\dots,K\}$$

Each segment k has its own parameters, e.g.:

$$\lambda_k, \quad E[X]_k$$

Then:

$$E[L \mid g(\mathbf{x}) = k] = E[N \mid k]\ E[X \mid k]$$

and premium uses that segment's expected loss.

2.2 Risk scoring (continuous risk amount)

A **risk score** is a real-valued function:

$$s(\mathbf{x}) \in \mathbb{R}$$

Often it is linked to a probability or expected loss via a link function.

Example (GLM-style) for frequency:

$$\lambda(\mathbf{x}) = \exp(\beta_0 + \beta^\top \mathbf{x})$$

Example for severity mean (log-link):

$$E[X \mid \mathbf{x}] = \exp(\gamma_0 + \gamma^\top \mathbf{x})$$

Then:

$$\boxed{E[L \mid \mathbf{x}] = \lambda(\mathbf{x})\ E[X \mid \mathbf{x}]}$$

2.3 Clustering (unsupervised "closed groups")

Clustering assigns:

$$c(\mathbf{x}) \in \{1,\dots,K\}$$

without using claim labels directly. You then compute **empirical risk** inside each cluster, e.g.:

$$E[\widehat{L \mid c = k}] = \frac{\sum_{i:c_i=k} L_i}{\sum_{i:c_i=k} e_i}$$

A. Actuarial Foundations and Regulatory Considerations

Premium optimization frameworks rest on sound actuarial foundations and regulatory requirements. Essentially, each insurance policy generates future cash flows amounting to the premium income (net of commissions) received at inception less the claims payments and expenses incurred during the policy term. Thus, risk-based pricing should be set at the expected claims plus a loading, enabling the insurer to cover all associated costs. This conservative approach guarantees a positive reserve position, permitting the company to honour its obligations. However, actuarial pricing is just one of many possibilities. Effective optimization enables the insurer to satisfy multiple objectives with a single premium-setting mechanism. Risk-based pricing remains the default—for now—but different pricing strategies may reward socially desirable customer behaviours such as seeking loyalty or using data to disclose latent risks. Dynamic or algorithmic pricing approaches find altered risk rates for individual insurance policies, automatically adjusting prices to market conditions, types, and even variants of behaviour. Such calculations for risk-based premium rates, discounts, and surcharges seek to maximise the insurer's lifetime value, while allocating sufficient resources to retain at-risk policyholders. Other factors may also enter the pricing process, applying predictive analytics to customer churn.

Dynamic, algorithmic, and personalised pricing methods differ from those prevailing in auto/health insurance. Whereas regulated cases often require full risk-based pricing, algorithmic discounting differs from those employed by e-commerce platforms only in the reality of detecting and rewarding customers presenting low-cost insurance. Personalised pricing moves further, abandoning the core principle of a sole pricing mechanism for each contract and instead devising different prices for every customer—a risky strategy for even the best-regulated banks or financial firms. Nevertheless, algorithmic or data-mined approaches should not escape scrutiny: the risk of unaffordable pricing remains present, and fairness issues persist, especially when training a pricing model on historical data. For these reasons, automated pricing loops must be monitored regularly—fulfilling auditability requirements—by different actors to diminish the risk of discriminatory practices. Indeed, commitments to transparency, compliance, and customers' best interests ensure that any algorithm becomes just a sophisticated way to implement and document what good practice dictates.

B. Dynamic Pricing and Personalization

Dynamic premium personalization is a key benefit associated with advanced AI-driven risk stratification that can be exploited to create new business value. Thus, beyond standard risk-based pricing, refining the commercial motivation for dynamic pricing and leveraging risk-based discounts, surcharges, or both represent useful avenues. Optimal premium calculation should further consider customer lifetime value, churn risk, distribution costs, and retention strategy. Once identified, the pricing algorithm enables personalized offers by dynamically matching premium products to single customers based on a set of profit-maximising criteria.

Within this framing, the pricing loop described above generates risk-adjusted premiums that can vary freely over time according to changes in individual risk, portfolio profitability, or competition. Discrimination is inherently fair if premium differentiation is fully justified on risk and affordability grounds. It is then essential to check that pricing rules respect zero-profit requirements in the considered risk assessment subspace. Insurance companies are also expected to operate with the lowest possible expenditure for achieving a deterministic lifetime discount, while nevertheless guaranteeing an adequate level of reserve for covering insurance claims.

V. IMPLEMENTATION CHALLENGES AND OPERATIONAL CONSIDERATIONS

Application of AI by insurance companies has to expand beyond fundamental risk assessment and premium optimization capabilities to encompass their operational facets as well. Consequently, while both the current discussion and the preceding sections on premium optimization have concentrated chiefly on the AI components of risk stratification and premium calculation, the subsequent exposition shifts attention towards considerations that influence the operationalization of AI methods. The first challenge relates to data privacy and cybersecurity concerns since the AI algorithms often rely on personally identifiable information. The second operational aspect deals with compliance and governance issues.

IT security policy recommendations highlight a comprehensive array of data protection strategies, which focus on establishing restricted access to sensitive information, employing strong cryptography, maintaining audit trails of data usage, and creating incident response plans that cover the entire data life cycle. During the planning and operational stages of AI projects, attention should therefore be paid to these recommendations, especially when handling data or potential proxies related to protected attributes. Additionally, planning and data handling include obtaining explicit user consent and a clear justification for retention or processing of personally identifiable information that is not directly relevant to the predictive models while adhering to the principle of data minimization.

A. Data Privacy and Security

Data privacy concerns are paramount throughout the predictive modeling process, including data collection, storage, and usage. Automated diagnostic procedures should indicate the need for humanly comprehensible explanations for training data and model predictions. Classes must be balanced to the greatest reasonable extent, in addition to other straightforward considerations such as non-sensitivity and abundance of missing values. Consistent access control policies should be implemented by means of dedicated authentication infrastructure, with regular changes of authentication credentials. Data at rest and in transit should be encrypted according to compliance requirements. Automation of audit trail generation and incident response planning is necessary.

Public information from the Internet of Things (IoT) and other connected devices needs no user's consent, according to the rules of the General Data Protection Regulation (GDPR), nor does the collection of data for cybersecurity and fraud prevention. Potential proxies for legally sensitive attributes—such as age, gender, or race—should be properly validated. Privacy-preserving dimensionality reduction techniques are impositioned to further alleviate risk. Data minimization, taking into account the minimum amount of personal data required, should be carried out, determining whether both joint control of the personal data of the individuals concerned and data processing by a third party are lawful and necessary.

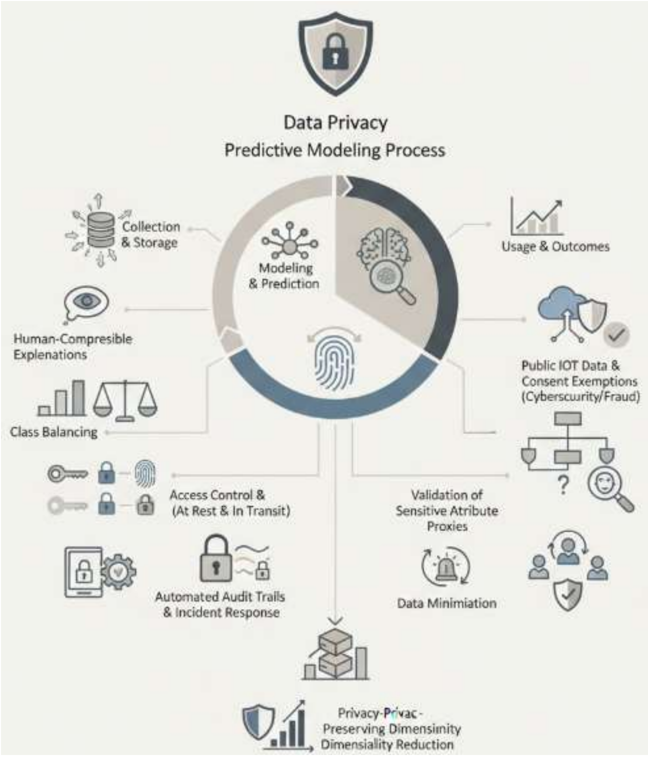


Fig 3: A Multi-Layered Framework for Privacy-Preserving Predictive Modeling: Integrating GDPR Compliance, Automated Auditability, and Explainable AI in IoT Ecosystems

B. Compliance and Governance

Effective governance is crucial for deploying AI models in practice. Factors such as data selection, labeling and preparation, development and training capabilities, application and embedding in corporate user workflows, and system-wide integration influence model risk and potential gains. Good governance, which assigns clear responsibilities and defines appropriate oversight, facilitates success. Moreover, thorough documentation of what is done, why it is done, and how it could fail enhances transparency. Without proper documentation, even the best-designed and best-governed models are likely to fail.

Model risk management is essential. Independent management functions need to review developments to ensure they operate within the objectives, strategy, and risk appetite of the organization. Model validation helps identify problems and encourages improvement. Monitoring confirms that models respond as expected and that performance follows predicted trends. Models should be auditable. Proper documentation supports monitoring, validation, and model risk management; excessive, unnecessary documentation hinders it.

VI. CASE STUDIES AND EMPIRICAL EVIDENCE

Empirical insights bolster the credibility and relevance of a method or concept. In this instance, the focus is on demonstrated benefits in risk stratification for auto and property insurance using AI technologies, reflecting ongoing interest among insurance entities.

Auto Insurance Risk Stratification

At the crossroads of machine learning and insurance, AI-driven risk stratification constitutes a promising area for application. Car insurance risk assessment methods have remained stable for decades, relying on risk factors such as the driver’s age and gender, the vehicle's age and model, and the insured area. However, estimates of the share of traffic accidents explained by these variables indicate that the predictive power of such segmentation is limited. AI should allow for more credible risk stratification based on the available data of insured drivers while enabling alternative approaches to risk management.

The proposed method uses AI to stratify the risk of different auto insurance agents, making it particularly relevant for insurance products whose annual claims are low, such as auto insurance. The data come from a Brazilian insurance company and reflect the risks associated with more than 100,000 policies, several aspects of which make it suitable for stratification. The data set has a long time history, which is essential for risk assessment. It includes a wide variety of risk factors—without the common limitations of few variables deployed in the stratification—and inspection of the pattern of claims and premiums highlights a dynamic nature, characteristic of insurance segments with low frequency and high severity. The results indicate that the risk stratification is significantly enhanced when using AI.

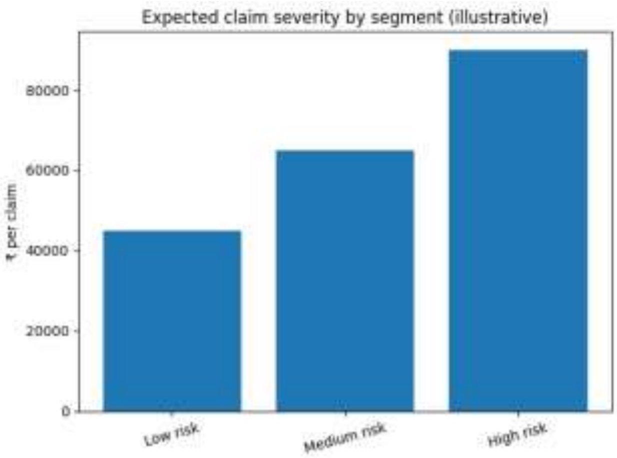
The performance of the AI-led auto insurance risk stratification is evaluated through various methods, encompassing calibration tests based on the expected version of the Brier score, Z-test for risk-forecasting superiority of different agents, and econometric swapping tests. The AI capability of requiring a minimum amount of data to achieve good predictions makes it convenient for real-world implementation, as it allows for the development of segmented models in periods of restricted historical claims record. Data-driven testing confirmed the risk-forecasting inequality of the two classes of drivers, which corresponds to differing crime rates of distinct assortments of automobiles. The innovation of the stratification model contributes to insurance fraud risk management.

Mean predicted p	Observed event rate	Count
0.009793158278297048	0.00816326530612245	490
0.025306128384052516	0.025495750708215296	706
0.041480083421760496	0.038510911424903725	779
0.0581570280749267	0.06095791001451379	689
0.07510049844291686	0.0576	625

A. Auto Insurance Risk Stratification

Automobile insurance is a mature insurance line in most developed markets, characterized by broad coverage of numerous and diverse customers. Accurate assessment of accident likelihood and expected loss is critical for maintaining financial stability, ensuring reasonable customer treatment, and ultimately promoting sustainable growth. Therefore, the importance of risk stratification cannot be overstated. Nevertheless, the majority of companies still rely on classical actuarial approaches based on a limited number of rating factors. In contrast, many businesses in other industries leverage machine learning methods for risk assessment; however, the potential of these techniques for risk stratification in automobile insurance has yet to be fully realized. Empirical research using real data from a demand-sensitive company fills this gap and reveals surprising results: risk stratification based on a broad set of variables, including detailed proxies for driving behavior, do not improve performance compared to conventional models using simple, interpretable factors.

Supervised classification trees, random forests, and their gradient-boosted extensions are commonly employed for generic classification tasks, supporting applications such as churn prediction and response modeling. Nevertheless, such classification models are not directly deployable as calibration factors in a risk-based pricing regime, nor do they provide solutions to disparate-treatment concerns. When rating factors are calibrated using a predefined pricing corridor, related techniques such as isotonic regression or more elaborate supervised-calibration frameworks may be applied. By contrast, when the objective is to develop pricing discounts (or surcharges) that are guaranteed not to lead to losses for the insurer, unsupervised classification and clustering algorithms are typically preferred.



Equation 3) Premium calculation: expected claims + loading

3.1 Basic premium formula

Let:

- Pure premium $PP = E[L]$
- Loading factor θ (expenses + risk margin + profit)

Then indicated premium:

$$P = PP(1 + \theta)$$

So:

$$\boxed{P = E[L]\,(1 + \theta)}$$

3.2 Break-even / "zero-profit" condition on a group

For a subgroup G (segment/cluster), break-even is:

$$E[P - L \mid G] \geq 0$$

If P is fixed within G, then:

$$P_G \geq E[L \mid G]$$

B. Property Insurance Risk Stratification

Deployment of AI-driven risk stratification in property insurance has broad implications for premium valuation systems, underwriting and capital allocation decisions. Other property lines, including personal, commercial, and catastrophe insurance, may yield similar gains. Outcomes for home insurance are particularly significant since many companies in the Americas simply rely on actuarial rate-making formulas to define their premiums, exposing them to important commercial risks and cracking their account profitability. Finally, the proposed models can be transferred to government-run monopolies that manage national catastrophe insurance funds. For any line, the ultimate risk groupings, risk clusters and estimated rates can become useful proxies for their segmentation, scoring and clustering.

Embedded AI-driven stratification in a property premium pricing mechanism. Premium calculation is based on the best available estimates of expected losses, and the standard actuarial logic reasoning process to achieve desirable reserve levels. Risk groups emerge naturally from an unsupervised algorithm to define the basic price buckets at which capital allocation is attracted and the expected losses covered. Standard regulatory compliance, fairness and anti-discrimination constraints must be incorporated into dynamic price optimization methodology.

VII. CONCLUSION

The findings confirm that artificial intelligence enhances risk stratification for both auto and property insurance lines. The importance of transparency and the potential transferability of models across regions and products have direct implications for commercial practices. Despite the significant contributions, several research areas remain unexplored. These include the implementation of an AI-based premium optimization engine capable of formulating dynamic and personalized offers, the investigation of risk models built using unsupervised learning techniques that serve as inputs for artificial intelligence–driven pricing algorithms, the systemic inclusion of churn considerations in premium optimization, the extension of results to other lines of business, and the assurance of equity and fairness via a much wider set of protected attributes.

The advent of AI at the core of operational processes enables local, national, and multinational insurance companies to stratify their portfolios with unprecedented precision and granularity. Risk assessment moves beyond conventional actuarial principles and classical limits of model building, opening the door to other domains of artificial intelligence and machine learning. In this way, companies can price more accurately, apply more targeted behavioral campaigns, execute smarter underwriting, manage wider external dynamics (e.g. climate change) with greater realism, allocate capital with reduced risk of sufficiency error, and, ultimately, also trigger a reallocation to the better. The operational suitability of AI-based risk stratification, however, remains untested across the insurance industry as a whole. By covering the lines of business most affected by data privacy issues, defining a complete process from data acquisition to real-world implementation, and providing transparently assessable AI stratification packages, the effort replicably advances a research agenda for AI risk stratification and opens the door to premium optimization models spiced up with churn-driven considerations.

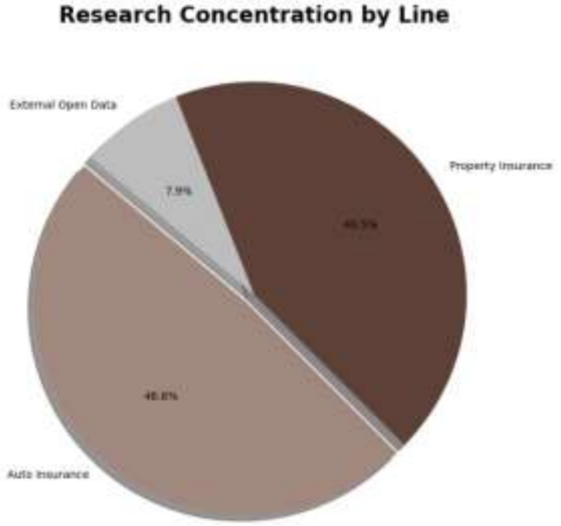


Fig 4: Research Concentration by Line

A. Summary and Future Directions

The emerging dominance of artificial intelligence across diverse economic sectors raises the question of its availability and applicability to less data-rich segments. Greater-than-expected skews from actuarial models, still in their infancy, have sparked new interest in the area of risk assessment, where much more substantial datasets have long been available but not fully exploited. Risk stratification, the task of identifying groups with distinct losses, entails distinct yet overlapping concepts of segmentation, scoring, and clustering. Core performance matrices emerge from the literature, although alternative lenses exist, including those of fairness, explainability, and internal models. As well as responding to practically oriented research questions, the present work forms part of a doctoral thesis on the application of artificial intelligence to risk stratification and premium optimization.

Presented herein are two tightly integrated AI-driven proposals for risk stratification in auto and property insurance, valid on real data and addressing gaps in the literature. AI methods demonstrate superior predictive capability to the traditional approach, although the performance difference diminishes with improved property data quality. Despite the absence of transferability between lines, the results highlight the significant impact of risk stratification on insurance operations and underline the value of enhanced property data quality. Prior review of the full practical-development cycle indicates that negative comments on the implementation of purely technology-driven strategies are overstated. Life cycle elements worthy of particular consideration include the acquisition of dimensionality-reduced open-data proxies for external-influence factors and the definition of fairness conditions for rating-group determination.

References

1. Blier-Wong, C., Cossette, H., Lamontagne, L., & Marceau, E. (2021). Machine learning in P&C insurance: A review for pricing and reserving. Risks, 9(1), 4.
2. Eling, M., Nuessle, D., & Staubli, J. (2022). The impact of artificial intelligence along the insurance value chain and on the insurability of risks. The Geneva Papers on Risk and Insurance—Issues and Practice, 47, 205–241.
3. Davuluri, P. S. L. (2023). AI-Augmented Sanctions Screening: Enhancing Accuracy and Latency in Real Time Compliance Systems. AI-Augmented Sanctions Screening: Enhancing Accuracy and Latency in Real Time Compliance Systems (December 15, 2023).
4. Gao, G., Wang, H., & Wüthrich, M. V. (2022). Boosting Poisson regression models with telematics car driving data. Machine Learning, 111, 243–272.
5. Xie, S. (2021). Improving explainability of major risk factors in artificial neural networks for auto insurance rate regulation. Risks, 9(7), 126.
6. Guillen, M., Nielsen, J. P., Pérez-Marín, A. M., & Densgill, J. (2021). Near-miss telematics in motor insurance. Journal of Risk and Insurance, 88(4), 1047–1072.
7. Kolla, T. (2024). Graph Neural Networks for HCC Risk Adjustment and Interoperability. International Journal of Science, Research and Technology, 7(6), 13244-13255.
8. Henckaerts, R., & Antonio, K. (2022). The added value of dynamically updating motor insurance prices with telematics collected driving behavior data. Insurance: Mathematics and Economics, 105, 79–95.
9. Henckaerts, R., Côté, M.-P., Antonio, K., & Verbelen, R. (2021). Boosting insight in insurance pricing with tree-based machine learning models. North American Actuarial Journal, 25(sup1), S22–S45.
10. Inala, R. (2023). AI-powered investment decision support systems: Building smart data products with embedded governance controls. Journal for ReAttach Therapy and Developmental Diversities, 6(10), 2251-2266.
11. Wüthrich, M. V. (2021). Bias-variance analysis for stochastic gradient boosting. Computational Statistics & Data Analysis, 157, 107158.
12. Lindholm, M., Richman, R., Tsanakas, A., & Wüthrich, M. V. (2022). Discrimination-free insurance pricing. ASTIN Bulletin: The Journal of the IAA, 52(1), 55–89.
13. Lindholm, M., Richman, R., Tsanakas, A., & Wüthrich, M. V. (2024). A multi-task network approach for calculating discrimination-free

insurance prices. European Actuarial Journal, 14(2), 329–369.
14. Maillart, A. (2021). Toward an explainable machine learning model for claim frequency: A use case in car insurance pricing with telematics data. European Actuarial Journal, 11, 1–25.
15. Kolla, S. K., & Reddy, V. A. R. (2024). Evaluating Cloud-Native vs. Hybrid Architectures for Health Benefit Administration Systems. International Journal of Medical Toxicology and Legal Medicine, 27(5), 1042-1053.
16. Xie, S. (2024). Analyzing the influence of telematics-based pricing strategies on traditional rating factors in auto insurance rate regulation. Mathematics, 12(19), 3150.
17. Usman, F., Chan, J. S. K., Makov, U. E., Wang, Y., & Dong, A. X. D. (2024). Claim prediction and premium pricing for telematics auto insurance data using Poisson regression with Lasso regularisation. Risks, 12(9), 137.
18. McDonnell, K., Murphy, F., Sheehan, B., Masello, L., & Castignani, G. (2023). Deep learning in insurance: Accuracy and model interpretability using TabNet. Expert Systems with Applications, 217, 119543.
19. Moosavi, S., & Ramnath, R. (2023). Context-aware driver risk prediction with telematics data. Accident Analysis & Prevention, 192, 107269.
20. Gottimukkala, V. R. R. (2024). Federated Learning Approaches for Fraud Detection in International Payment Systems. https://www. jisem-journal. com/download/118_JISEM. pdf.
21. Poufinas, T., Gogas, P., Papadimitriou, T., & Zaganidis, E. (2023). Machine learning in forecasting motor insurance claims. Risks, 11(9), 164.
22. Kurniawati, R., & Choiruddin, A. (2024). Optimizing claim assessment processes in property insurance: A case study. Procedia Computer Science, 234, 520–526.
23. Vorobyev, I. (2024). Fraud risk assessment in car insurance using claims graph features in machine learning. Expert Systems with Applications, 251, 124109.
24. Byrne, A. (2024). Pricing risk: An XAI analysis of Irish car insurance premiums. In L. Longo, I. Lapuschkin, & C. Seifert (Eds.), Explainable Artificial Intelligence (pp. 315–330). Springer.
25. Jaiswal, R., Gupta, S., & Tiwari, A. K. (2024). Big data and machine learning-based decision support system to reshape the vaticination of insurance claims. Technological Forecasting and Social Change, 209, 123829.
26. Inala, R., & Somu, B. (2024). Agentic ai in retail banking: Redefining customer service and financial decision-making. Journal of Artificial Intelligence and Big Data Disciplines, 1(1), 1-19.
27. Finger, D., Albrecher, H., & Wilhelmy, L. (2024). On the cost of risk misspecification in insurance pricing. Japanese Journal of Statistics and Data Science, 7, 1111–1153.
28. Fahrenwaldt, M., Furrer, C., Hiabu, M. E., Huang, F., Jørgensen, F. H., Lindholm, M., Loftus, J., Steffensen, M., & Tsanakas, A. (2024). Fairness: Plurality, causality, and insurability. European Actuarial Journal, 14(2), 317–328.
29. Lemaire, J., & Wang, X. (2022). Three-step risk inference in insurance ratemaking. Insurance: Mathematics and Economics, 105, 1–13.
30. König, D., & Loser, F. (2024). Claim frequency modeling in insurance pricing using GLM, deep learning, and gradient boosting. European Actuarial Journal, 14, 1–20.
31. De Virgilis, M., Lupton, D., McGrath, L., Qazvini, M., Roby, S., & Vernon, L. (2022). Machine learning in insurance. CAS Forum, Winter E-Forum, 1–45.
32. Kuo, K., & Richman, R. (2021). Actuarial applications of machine learning: A review and research agenda. ASTIN Bulletin: The Journal of the IAA, 51, 1–30.
33. 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.
34. Lindholm, M., Richman, R., Tsanakas, A., & Wüthrich, M. V. (2023). Antidiscrimination insurance pricing: Regulations, fairness criteria, and models. North American Actuarial Journal, 27(4), 1–24.
35. 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.
36. Kurniawati, R., & Choiruddin, A. (2024). Machine learning approaches for property insurance claim prediction and risk profiling. Procedia Computer Science, 234, 520–526.
37. Lindholm, M., Richman, R., Tsanakas, A., & Wüthrich, M. V. (2022). A multi-task network approach for calculating discrimination-free insurance prices. European Actuarial Journal, 14(2), 329–369.