AAE | ACTUARIAL ASSOCIATION OF EUROPE

Explainable Artificial Intelligence

Claudio Senatore Reso, *Vice-Chairperson AI-Data Science Working Group*

24 June 2025

ADVISING • ACHIEVING • ENGAGING

ACTUARIAL ASSOCIATION OF EUROPE

Status and Usage of XAI

The XAI research field can be split in two[1]:

Research
Explore
Debug

responsiBle
Legal
trUst
Ethics

RED XAI: Model Validation Oriented Explanations primarily designed for model developers.
BLUE XAI: Human Values Oriented Explanations primarily designed for final users of a model.

1. "*Position: Explain to Question not to Justify*" by Przemyslaw Biecek and Wojciech Samek

Status and Usage of XAI

Research
Explore
Debug

Audience

Experts who trains, audits, debug, check and mantain AI models.

Accessibility

Access to internal model parameter, training data or ready trained model.

Technical Knowledge

High level of technical knowledge.

Status and Usage of XAI

responsiBle
Legal
trUst
Ethics

Audience	Accessibility	Technical Knowledge
Final users of a model: policy holder, bank customer, patient.	No or partial access to model and data.	Usually low or no technical knowledge.

Status and Usage of XAI

There is no single key to unlock all doors

The key is the right model for the right audience:

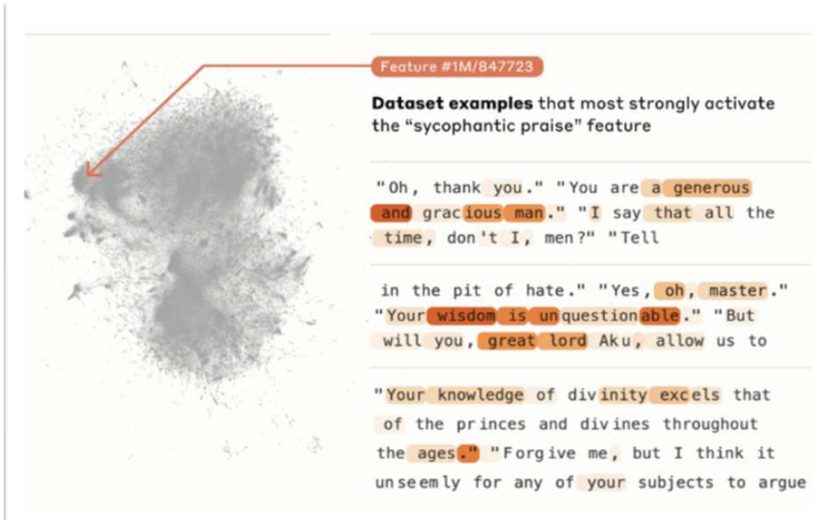
- Who is the end user?
- What is the aim?
- What is the interface that can be used?
- What type and level of model/data access is it necessary?

Status and Usage of XAI

AAE | ACTUARIAL ASSOCIATION OF EUROPE

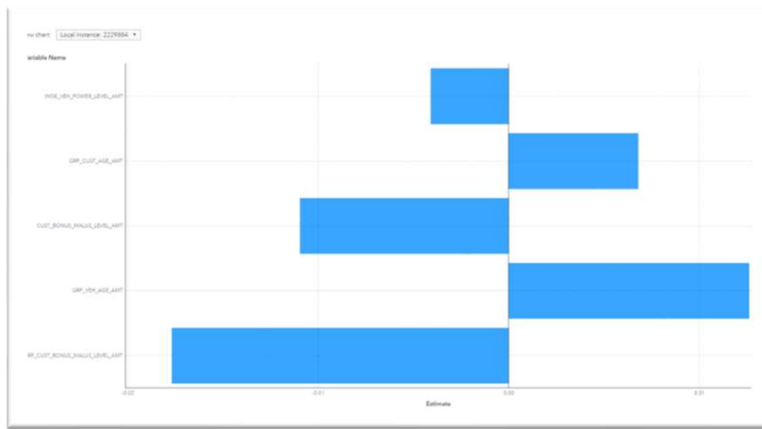
Mechanistic interpretability

RED

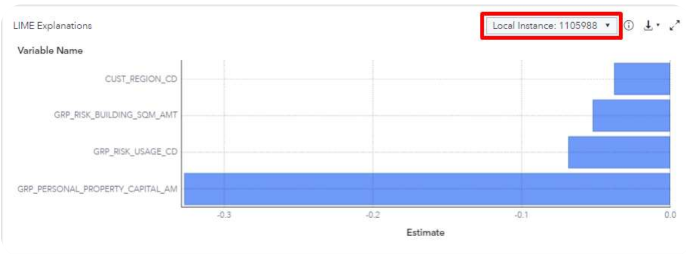


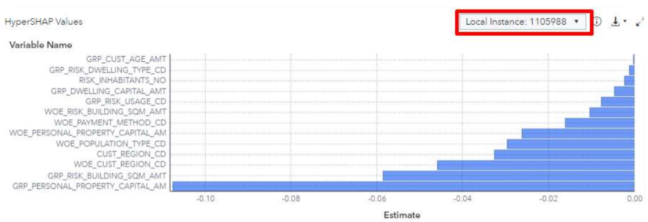
BLUE

Local/Global Meaure



XAI and the Actuarial World

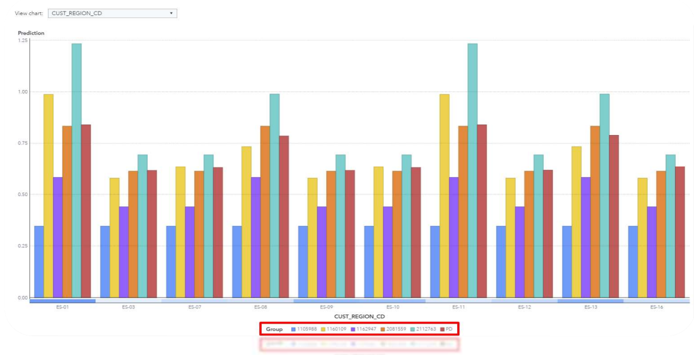




1 LIME*

Local Interpretable Model-Explanations (LIME) provide a list explanatory variables that specific predictions, regardless of model used. This helps drivers of results and aids in making.

2 ICE

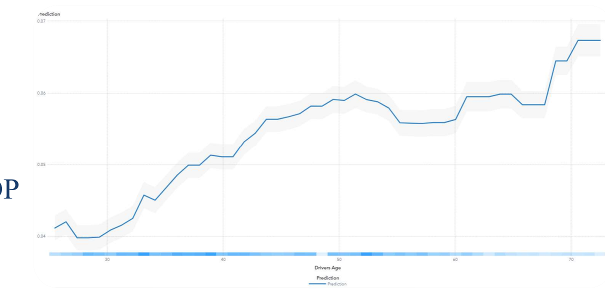


Individual Conditional Expectation (ICE) enables conducting a series of "what if" analyses by examining how predictions would change if certain risk factors were altered, facilitating scenario analysis.

3 Shapley Value

Shapley Value provides a local decomposition of the marginal impact impact of explanatory variables on specific predictions, helping understand understand which variables and how how they impacted those values.

4 PDP



*Alternative: local decision tree or ridge regression. Can be noisy, data sensitive.

Beyond Individual Methods: The Power of Mixed Explanations*

What Is the "Mixed" Approach?

Beyond Individual Methods: The Power of Mixed Explanations

Definition – A mixed explanation **integrates outputs of several XAI methods** into a single, richer narrative or score.

Simple formula:

$$MixedScore = w_{PD} \quad PD + w_{LIME} \quad LIME + w_{SHAP} \quad SHAP + \ \ldots$$

How it's done

1. Compute individual metrics (faithfulness, complexity, stability).
2. Convert each into **rank scores** (best = 1).
3. Feed ranks into a decision-making algorithm (TOPSIS) to obtain one closeness coefficient.
4. Use weighted sum (WSUM) to present a single "mixed" indicator.

*Chatterjee S., Colombo, E. R., & Raimundo, M. M. (2025). Multi-criteria Rank-based Aggregation for Explainable AI.

Why Mix? Shortcomings of Single Methods

AAE | ACTUARIAL ASSOCIATION OF EUROPE

Faithfulness Gap

PD averages away outliers; LIME may oversimplify; SHAP can over-focus on locality.

High fidelity means removing or perturbing the features flagged as important **changes the model output accordingly**.

Simplicity vs. Detail

Simple plots hide subtle interactions; detailed plots drown business users.

Counts how many features rank highly and how evenly importance is distributed.

Stability Issues

Some methods swing wildly with minor data changes.

Guards against "one-hit wonders" that look good on one sample but fail under slight drift.

Mixed Remedy

Combining PD's global context with LIME's local nuance **improves faithfulness while keeping explanations simple and stable**.

Added Value: Deeper Insights

The Benefits of Blending

Holistic Perspective

Simultaneously answers "How does the model work in general?" and "Why this specific prediction?"

Robustness

Aggregation dampens noise; paper reports top-2 ranking on 2⁄3 metrics across five datasets.

Interaction Discovery

Mixed dashboards highlight where local effects diverge from global trends.

Stakeholder Alignment

One composite score for executives + drill-down views for data scientists.

Regulatory Goodness

Single, auditable KPI meets multi-metric transparency clauses (EU AI Act Art. 15).

Limitations & Challenges

1 Information Loss

One great number can mask that fairness is poor while accuracy is stellar.

2 Weight Tuning

Deciding *w*'s is subjective; equal weights in the paper may not fit every domain.

3 Domain Sensitivity

Healthcare might value stability > simplicity; marketing the reverse.

4 Compute Cost

Running SHAP + LIME + noise-robust metrics is resource-intensive.

Practical Safeguard: Always accompany the mixed KPI with a **dashboard of the raw metrics** and a short human-written summary.

Conformal Prediction – Distribution-free, Instance-level Confidence

Validation tells us how a model scored on past data. Conformal Prediction tells us how much to trust today's single prediction.

Five essential components

1 **Calibration slice** – reserve 5–15 % fresh data **after** training.

2 **Non-conformity score** – quantify error per case (e.g., $|\hat{y} - y|$, $1-p_{pred}$).

3 **Quantile lookup** – take the $(1 - \alpha)$ score quantile → $\varepsilon\alpha$.

4 **Prediction wrapper** – at inference, return:

- Regression: $[\hat{y} \pm \varepsilon\alpha]$
- Classification: {labels with score $\leq \varepsilon\alpha$}

5 **Coverage guarantee** – $\Pr\{\text{truth} \in \text{set}\} \geq 1 - \alpha$ for every future case, *no distribution assumptions*.

Key properties

Model-agnostic: works unchanged for GLM, XGBoost, CNN, transformer.

Finite-sample exact: 90 % means 90 %, even with 100 calibration points.

Per-instance delivery: interval arrives **before** the true label, ready for decision-time use.

Classification and Regression Applications

Classification – Fraud Flagging

Business question

Will this claim be fraudulent?

Base model

Gradient-boosted tree fraud classifier (probability output)

Plain output (before CP)

p(fraud) = 0.27 → "No action" if threshold = 0.5

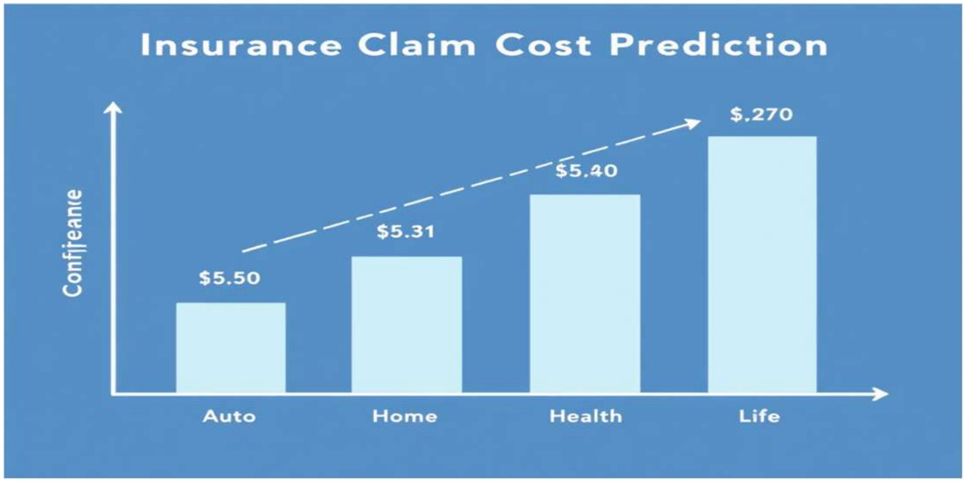
CP-enhanced output

Prediction set = {"legit", "fraud"} at 90 % coverage → handler triggers **manual review**

Decision impact

Focus on cases where CP reveals ambiguity

Regression – Loss Severity



Business question

What will the ultimate cost of this bodily-injury claim be?

Base model

Deep claims-severity network predicting $ amount

Plain output (before CP)

Point estimate = €12 400

CP-enhanced output

Prediction interval = €9 100 – €18 300 (90 % coverage)

Decision impact

Reserves booked at prudently high end; capital models use full interval

Conformal Prediction vs. Classical Prediction Intervals

Question you’re trying to answer	Classic train / test split	Conformal prediction (CP)
When is the information available?	*After* you already know the true labels of the test set.	*Before* you know the label of each new case you must act on.
Granularity	One (or a few) numbers that summarise the whole test set (RMSE, AUC, accuracy …).	A tailor-made interval / label-set for **every single new prediction**.
What it guarantees	“The model averaged RMSE = 12 000 € on houses like these.” (An estimate that can vary if you drew another test set.)	“For this house, the true price will fall inside **[$\hat{y}$ ± ε]** with ≥ 90 % probability.” (Finite-sample, distribution-free guarantee.)
Dependence on model assumptions	Interval formulas exist only if you can write down a distribution (e.g., Normal errors).	Works unchanged for linear regression, gradient boosting, or a black-box transformer—no distribution assumption.
Cost at inference time	None (you just quote yesterday’s metric).	Adds a small wrapper computation (quantile ε), no model re-fit.

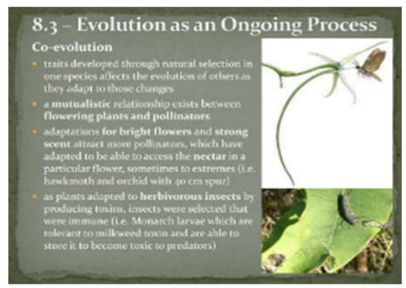
Limitations and Mitigations

Limitation (what can break the guarantee)	Impact in actuarial settings	Mitigation / best practice
Exchangeability assumption (future ≈ calibration data)	Coverage degrades if portfolio mix, legislation, or macro-economy shifts	Monitor drift; re-calibrate when shift detected
Marginal, not conditional, coverage	Rare sub-segments (e.g. high-value life policies) may be under-covered	Produce **slice-level coverage reports**; adjust with weighted or group-CP variants
Interval width vs. usability	High confidence → wide ranges that impede pricing precision	Choose confidence level aligned with risk appetite; investigate better non-conformity scores
Calibration data cost	5–15 % of data diverted from training	Use cross-conformal or jackknife-plus to recycle data
Many-class classification	Motor-claim cause code (≈50 classes) might yield large label sets	Top-k or soft CP variants, accepting approximate guarantees
Temporal dependence (claims triangles)	Standard CP ignores lag structure	Use specialised time-series CP or sliding-window calibration
Adversarial or strategic behaviour	Fraudsters may game inputs to escape intervals	Combine CP with adversarial-robust training and business rules

Conclusion

Use Multiple Indicators

Leverage a range of indicators and methods to gain a comprehensive understanding of model interpretability.



Actuarial Consensus

Acknowledge the lack of consensus and continue to refine best practices as the field progresses.

Seek Practical Guidance

Stay informed as industry-specific guidelines and standards continue to take shape.

Thank you!

Annex

- “*Position: Explain to Question not to Justify*” by Przemyslaw Biecek and Wojciech Samek
- *Chatterjee, S., Colombo, E. R., & Raimundo, M. M. (2025). Multi-criteria Rank-based Aggregation for Explainable AI.*
- *Angelopoulos, A. N., & Bates, S. (2022). A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification.*
- *Il Idrissi, M., Fernandes Machado, A., Gallic, E., & Charpentier, A. (2024). Unveil Sources of Uncertainty: Feature Contribution to Conformal Prediction Intervals.*