Bridging transparency and predictive power: Integrating explainable ML into actuarial modelling

Research Proposal - Big Pitch 2026 Event

Luteijn, J.M., Tam, J. Fan, M.F.

The Big Pitch, March 18th, 2026. Staple Inn, London.

Disclaimer: The views expressed in this publication are those of invited contributors and not necessarily those of the Institute and Faculty of Actuaries. The Institute and Faculty of Actuaries do not endorse any of the views stated, nor any claims or representations made in this publication and accept no responsibility or liability to any person for loss or damage suffered as a consequence of their placing reliance upon any view, claim or representation made in this publication. The information and expressions of opinion contained in this publication are not intended to be a comprehensive study, nor to provide actuarial advice or advice of any nature and should not be treated as a substitute for specific advice concerning individual situations. On no account may any part of this publication be reproduced without the written permission of the Institute and Faculty of Actuaries.

Abstract

Health and care (H&C) actuaries are well positioned to benefit from recent advances in data science as machine learning (ML) techniques have become increasingly transparent and accessible.

The ML developments allow actuaries to detect complex nonlinear patterns and interactions that are difficult to capture using traditional generalised linear models (GLMs), without sacrificing the clarity and governance advantages that make GLMs central to actuarial practice.

Using a large life insurance dataset, we demonstrate and appraise three emerging hybrid approaches: interpretable boosted linear models, XGBoost-informed GLM and an interaction detection workflow. Our findings show that actuaries can improve modelling accuracy, measured by Poisson deviance, by integrating ML insights into traditional modelling techniques, achieving a practical balance of interpretability, expert judgement, and modern analytical innovation.

Correspondence details

Correspondence to: Dr. Michiel Luteijn, Hannover Re UK Life Branch, London, UK. E-mail: michiel.luteijn@hannover-re.com

Keywords

Data science; actuarial science; machine learning; statistics; methodology; review; health insurance; life insurance; health and care;

1. Introduction

Modelling is essential to every aspect of life and health insurance. Pricing and expenses management depend on credible estimates of mortality and morbidity across various factors. Compared to non-life actuaries, life & health datasets pose distinct challenges: mortality and morbidity events are rare (especially at younger ages), exposure at specific segments is thin, and credible factor interactions quickly fragment claim counts into sparsely populated cells. As a result, actuaries face a persistent balancing act between model richness and statistical credibility.

Over time the GLM toolkit has been expanded with shrinkage (e.g. LASSO, Ridge and elastic net) and smoothing techniques (e.g. GAMs and polynomials) to control variance, reduce noise, and avoid over-fitting, yet the fundamental use of GLM has changed little in day-to-day practice. This persistence reflects the transparency, interpretability, and governance readiness of GLMs, which remain central to actuarial validation, communication, and review. However, these strengths come at the cost of a limited ability to capture complex nonlinearities and interactions without substantial manual effort.

In parallel, machine learning (ML) techniques including gradient boosting machines and neural networks have seen rapid development and offer a powerful alternative to GLMs by automatically capturing complex nonlinearities and variable interactions without a-priori specification. Compared to these novel models, various authors have shown that traditional GLMs can leave value on the table. (Bjerre, 2022; Richman & Wüthrich, 2021) However, uptake has been limited due to regulatory scrutiny, difficulties in interpretation and privacy challenges. (CAS Machine Learning Working Party, 2022)

Rather than viewing ML as an alternative to traditional actuarial models, the actuarial literature has largely focused on enhancing traditional models using ML. A wide range of approaches on enhancing GLMs using ML have been proposed. (CAS Machine Learning Working Party, 2022) One strand uses ML for feature engineering, including identifying nonlinearities, clustering or binning. (Dai, 2018; Henckaerts, et al., 2018; Maillart, 2021) A second strand uses ML as a diagnostic or exploratory tool to

detect nonlinearities and/or interactions for a subsequent GLM-like model. (Tam & Luteijn, 2025) A third strand uses ML to directly correct for signal not captured by a baseline model. (Gawlowski & Wang, 2025) These approaches aim to preserve transparency whilst leveraging the predictive power of ML models and are typically evaluated in isolation, using different datasets and performance metrics. As a result, actuaries lack clear guidance on how these hybrid approaches compare in terms of predictive performance, transparency, governance readiness, and modeller control required for practical deployment.

Mortality data provides a useful methodological case study for informing a broader range of life and health actuarial modelling problems. The gap between traditional methods and ML methods is especially evident in health and care insurance where the literature on ML models proposes to replace rather than augment GLM-based frameworks. (Kshirsagar, et al., 2021; Orji & Ukwandu, 2023) The challenges faced in health and care modelling share many aspects with those faced in mortality modelling.

In this paper, we address this disconnect by appraising three hybrid additive modelling approaches that integrate gradient boosting into an additive model framework to support practical adoption in actuarial practice. Specifically, we evaluate three approaches drawn from the latter two strands described earlier: interpretable boosted linear models using XGboost to correct a baseline GLM model (Gawlowski & Wang, 2025), an interaction detection framework in which gradient boosting is used to detect interactions (Tam & Luteijn, 2025) and an XGBoost-informed GLM approach where gradient boosting insights guide manual feature engineering. The focus on gradient boosting is informed by the observation that gradient boosting is state-of-the-art for tabular data. (Grinsztajn, et al., 2022) The contribution of this paper is methodological rather than outcome-specific, focusing on governance-ready additive model structures rather than mortality estimation per se.

2. Methods

2.1. Objective

The objective of this paper is to assess how different hybrid approaches that integrate gradient boosting into additive actuarial models perform across predictive accuracy, transparency, governance readiness, and modeller control. The focus is on approaches that retain an explicit additive structure suitable for actuarial deployment, rather than fully automated black-box models.

To support a fair and reproducible comparison, all approaches are evaluated using a common dataset and consistent performance metrics. Baseline GLM-type models, black-box XGBoost benchmarks, and three hybrid approaches are considered to contextualise performance trade-offs.

2.2. Dataset

We performed analysis on American life insurance experience data, publicly available via the Individual Life Insurance Experience Committee (ILEC) of the Society of Actuaries (SoA) (Individual Life Insurance Experience Committee of the Society of Actuaries, 2024), which we will refer to as the ILEC data.

The ILEC dataset was selected for its public availability (allowing anyone to reproduce our study), large size (273,402 claim counts), documentation and relevance to insurance. The ILEC data covers American life insurance experience over observation years 2012-2019.

2.2.1. Data pre-processing

Inclusion criteria were term business only, attained age between 18-90, issue age between 18-80, issue years 1980 onwards and duration up to 20. Entries with unknown smoking status were excluded as experience was very heavy (197% on the ILEC-supplied expected basis), raising concerns over data

quality. Following these data filters, the dataset contained 273,402 claims with an average claim size of $292,000 over 160.18 million life years.

Amongst the features within the dataset, Observation_Year, Sex, Smoker_Status, Duration, Face_Amount_Band and Attained_Age were considered for mortality modelling. Features excluded for simplicity included features around preferred ratings and product types that may not necessarily translate to UK settings. There were no missing values for any of the features utilised.

We noted that age was reported either on an Age Last Birthday (ALB) or an Age Nearest Birthday (ANB) basis. The life year exposure proportions of ALB and ANB are approximately 44% and 56%, respectively. However, we do not have the granularity to align the two bases in an exact manner. Although we can approximate alignment by adding 0.5 years to the ages of rows with ANB, this comes with undesirable complications as there is material difference between ALB and ANB business not explained by age. For simplicity, no adjustments were made. We do not anticipate any material impact on our analysis.

2.2.2. ILEC VBT 2015

The ILEC dataset comes with expected claim counts, based on the 2015 valuation base table (VBT 2015). VBT 2015 rates were produced by ILEC on 2002-2009 experience data and extended to 2015 using mortality improvement assumptions. (Academy/SOA Valuation Basic Table Team, 2018)

2.2.3. Partitioning

To evaluate model performance, two data partitions were modelled. In the first partition, the dataset was randomly split by row into training (80%) and test (20%) subsets. In the second partition, the final two observation years (2018 and 2019) were test data, whilst 2012-2017 were training data. The random split results in a test dataset more similar to the train data than the observation year partition. Partitioning test data by observation year more closely aligns with actuarial practise of using historic experience to forecast future experience.

For the IBLM XGBoost booster model, the training subset was further divided using random sampling into an internal training set (80%) and a validation set (20%). This secondary split was used for early stopping. The XGBoost models used for interaction detection (models #4-5) were trained using cross-validation so no further partitioning was required.

2.3. Predictive models

Baseline performance was quantified using Poisson GAM, GLM, and LASSO models, whereas performance of modern black-box approaches was quantified using two XGBoost models (**Table 1**, models #1–5). Finally, five models representing the three hybrid GLM-GBM approaches were trained (**Table 1**, models #6-10).

The GAM models (#1, 4 and 8) were trained in Python and the remaining models (#2-3, #5-7 and #9-10) were trained in R. The entire work is publicly available via GitHub. (IFoA Techniques in Data Science in Health and Care, 2026) For readers unfamiliar with some of the terminology and methodology discussed here, we refer to our recently published framework for applying data science techniques to health and care actuarial projects. (Luteijn, et al., 2025)

The two previously published strategies demonstrated here are based the interaction detection approach (Tam & Luteijn, 2025) (models #8-10) and the Interpretable Boosted Linear Model (IBLM), model #6. (Gawlowski & Wang, 2025) The third strategy, model #7, is a more manual approach to leverage SHAP dependence plots for determining GLM regression formulae. The Poisson distribution was used for the sake of simplicity for all ten models.

Table 1: Models evaluated

ID	Model	Purpose	Notes
1	Baseline GAM	Present a baseline of what can be achieved with basic GAM / GLM / LASSO.	Variable interactions were out of scope.
2	Baseline GLM		
3	Baseline LASSO		
4	XGBoost	Present a benchmark of performance that can be achieved with modern black box ML models.	Uses model #1 as starting point
5			Uses model #2 as starting point
6	IBLM (Gawlowski & Beard, 2025; Gawlowski & Wang, 2025)	Demonstrate how GLM can be augmented by XGBoost, whilst retaining interpretability.	Builds basic GLM and enhances using XGBoost-produced SHAP beta corrections.
7	XGBoost informed GLM	Demonstrate a more manual, controlled approach to leverage XGBoost for GLM models.	A GLM built on information achieved by initial XGBoost. The initial XGBoost is not being appraised here.
8	GAM with XGBoost-detected interactions (Tam & Luteijn, 2025)	*The interaction detection approach*. Demonstrate how GLM-type models be augmented by interactions detected by XGBoost.	Variable interactions detected by XGBoost (model #4) were added.
9	GLM with XGBoost-detected interactions		Variable interactions detected by XGBoost (model #5) were added.
10	LASSO with XGBoost-detected interactions		

2.3.1. Baseline models

All baseline models used combinations of polynomials and splines to model nonlinearities within the numeric features. Variable interactions were outside the scope of the baseline models as interactions will be addressed by the boosting machines further downstream. For the GAM, hyperparameters were selected using cross-validated Bayesian tuning.

2.3.2. XGBoost models

Two XGBoost models were trained targeting claim counts, offset by the log of baseline GLM/GAM predictions, respectively (**Table 1**, models #4 and #5). Therefore, the XGBoost effectively targets the remaining signal in the log claim rate unexplained by the GAM/GLM predictions (similar to the internal XGBoost booster model in the IBLM, **Section 2.3.4**).

Cross-validation was used to prevent overfitting. Hyperparameter tuning was performed on model #4 using Optuna (Akiba, et al., 2019) and the calibrated set of hyperparameters were also applied to XGBoost model #5 and the internal IBLM model. Tree depth for the XGBoost models was set to two to restrict variable interactions to first order interactions to support the interaction detection approach (**Section 2.3.3**).

2.3.3. GLM-GBM hybrid approaches

Three hybrid approaches were applied: 1) the Interpretable Boosted Linear Model (Gawlowski & Wang, 2025), 2) the interaction detection approach (Tam & Luteijn, 2025) and 3) an XGBoost-informed GLM. These three approaches selected apply different levels of automation to the process of improving additive model performance using XGBoost. All three approaches produce a final additive model that is suitable for actuarial deployment.

Interpretable Boosted Linear Model

The IBLM (**Table 1**, model #6) combines a traditional GLM model with an XGBoost sequentially. (Gawlowski & Wang, 2025) This is achieved by a sequential approach: first a basic GLM is fitted without feature engineering, or variable interactions. Subsequently, an XGBoost model is tasked with capturing patterns the GLM failed to capture by targeting the residuals of the training data against the GLM predictions. The corrections by the XGBoost model can be approximated by a GLM-like structure by absorbing zeroes and reference levels of categories features into the intercept. The IBLM package code (Gawlowski & Beard, 2025) was updated by adding an offset (life years exposure) to the GLM and weighting the XGBoost training by life years exposure. The IBLM model by default uses a tree-depth of three. To facilitate comparison with models that do not accommodate 3rd interactions, and interpretability of 3rd order interaction is limited, the IBLM was additionally trained using a tree depth of two.

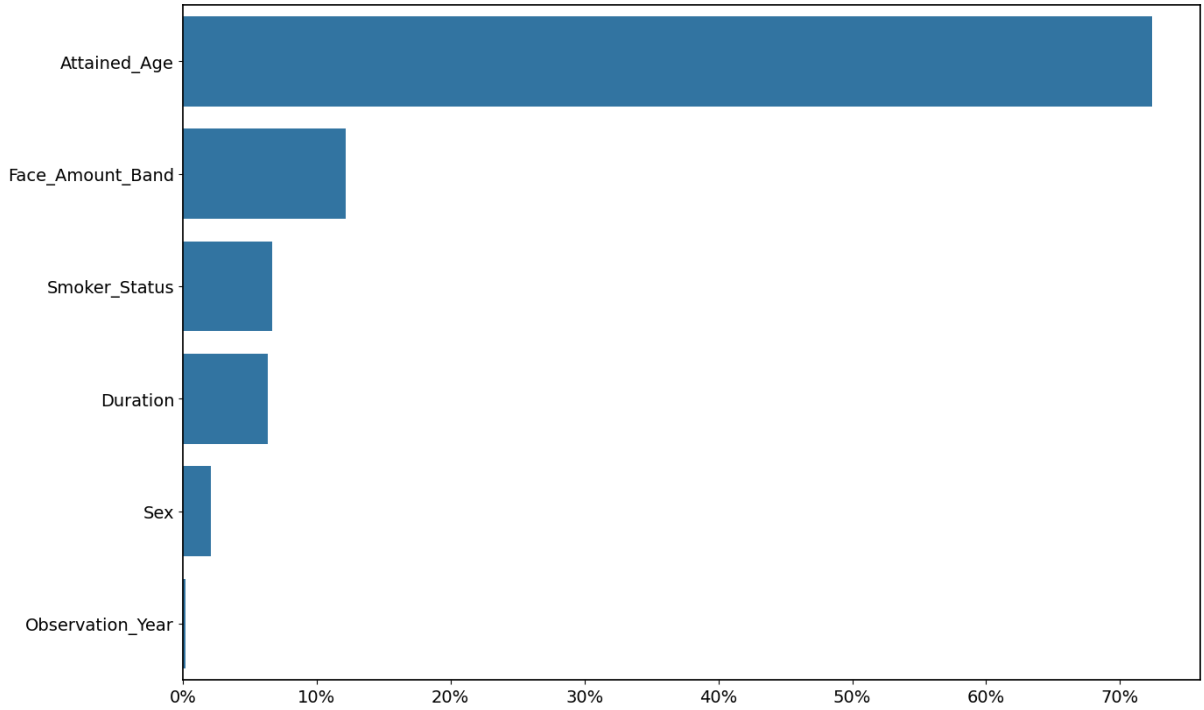
Interaction detection approach

The interaction detection approach (Tam & Luteijn, 2025) involves building a baseline additive model without feature interactions. The residuals of the baseline model are targeted by XGBoost and top ranking feature interactions identified by XGBoost are incorporated back into a final additive model built from scratch (**Table 1**, models #8-10). This strategy was tested with GAM, GLM and LASSO models. Both the GLM and LASSO use a GLM baseline since shrinkage was not considered desirable for setting a starting point.

XGBoost-informed GLM

The XGBoost-informed GLM (**Table 1**, model #7) was trained by first fitting an XGBoost on the training data, targeting claim counts with lives exposure as an offset. The feature importance plot of this XGBoost is included in **Figure 1**. A final GLM was then built using a regression formula informed by XGBoost-derived visuals.

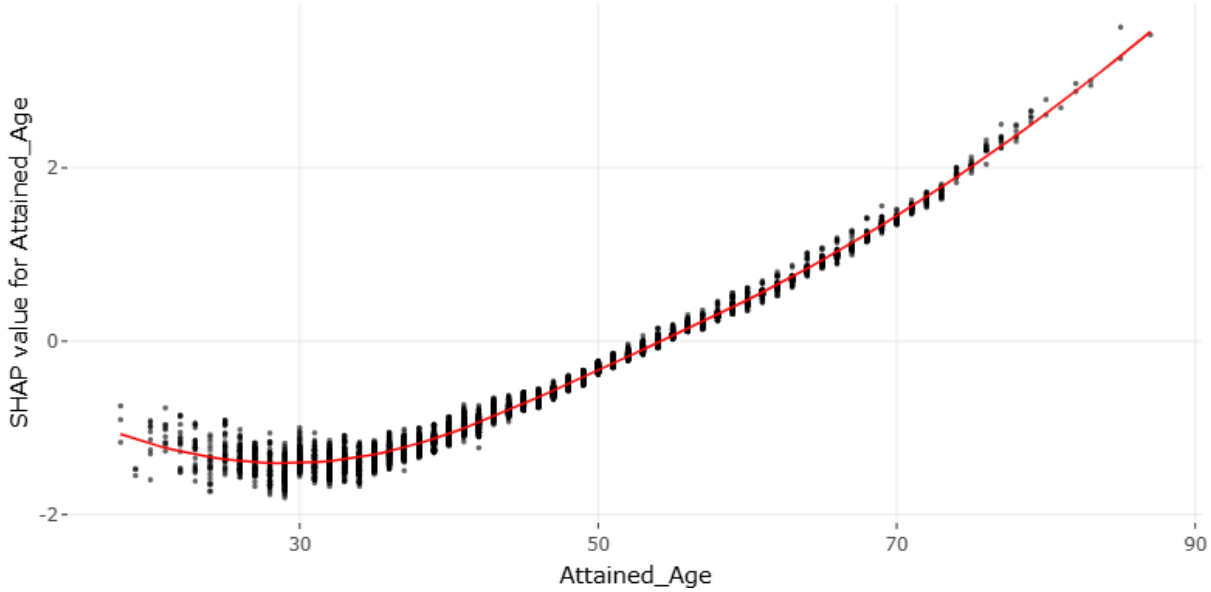
Figure 1: XGBoost normalised feature importance by Gain



For example, Attained age is by far the most important feature in the XGBoost (**Figure 1**), which suggests that using a spline to fit the age curve could be the most impactful. Its SHAP dependence plot shows a smooth but highly nonlinear variation in SHAP values across age (**Figure 2**). This suggests that

a simple linear term would severely underfit this relationship. The SHAP curve shows a dip at approximately ages 25–33, followed by multiple changes in curvature around ages 45, 55, and 70 as the marginal effect of age accelerates at older ages. Placing knots at 25, 33, 45, 55, and 70 thus allows the spline to capture this data-driven curvature, rather than relying on evenly spaced knots or quantile-based placement.

Figure 2: SHAP dependence plot for attained age



2.4. Appraisal framework

Having outlined the three modelling approaches, this section will describe the appraisal framework by which they will be assessed (**Table 2**).

Table 2: Hybrid approach assessment criteria

Assessment criterion	Evaluation metric	Section
Predictive performance	Mean Poisson deviance improvement (relative to benchmark) on test splits	3.1, 4.1
Modeller control	Degree of manual intervention required	4.2
Transparency	Final model additive structure and ability to visualise effects	4.3
Governance readiness	Ease of validation, documentation and stress testing	4.3

We use Poisson deviance to rank the performances of different modelling approaches in Section 3.1. This is consistent with conventions in mortality studies, where death counts are typically assumed to be Poisson distributed (Continuous Mortality Investigation, 2021), and in actuarial machine learning research, where claim frequency models are usually evaluated using deviance loss (Richman, et al., 2025; Richman & Wüthrich, 2023).

Model transparency and governance readiness are closely related assessment criteria. All actuarial models should comply with TAS 100: General Actuarial Standards. The standard document states that technical actuarial work must be fit for purpose, covering comprehensive aspects including defining models' intended uses, understanding model limitations and identification of material biases (Financial

Reporting Council, 2023). Being able to accurately understand the relationships between model predictions and individual features enables scrutiny of model output and thus underpins compliance as well as communication to key stakeholders. This transparency also underpins key governance activities: a model whose effects can be clearly visualised is easier to validate, document and stress test.

The degree of modellers' control over model training and predictions is another important criterion. Given the relatively low claim rate in the life and health & care context, the number of claims may not be high enough to reach full credibility for certain customer segments. Models, including additive ones, can be prone to overfitting to random patterns in the training data. Being able to impart domain knowledge is invaluable in building a robust risk model that can perform well in out-of-time periods. For instance, the underwriting effect fades over time for term assurance. Thus, there is no genuine reason why mortality risk would decrease in certain pockets of duration after accounting for age and mortality improvement. To prevent a model fitting a noisy trend, monotonic constraints can be imposed to, for example, increase model predictions with duration.

3. Results

Model predictive performance is evaluated in **Section 3.1** and the remaining appraisal dimensions are assessed qualitatively in **Sections 4.2–4.4**.

3.1. Predictive performance

We use the ILEC VBT 2015 mortality basis as the baseline benchmark for all the models. The metric used in **Table 3** is the average Poisson deviance improvement against this benchmark.

Across both test datasets, the black-box XGBoost models were the most performant (**Table 3**). Amongst the three hybrid approaches being appraised, the interaction detection approach resulted in the highest improvement against the baseline benchmark. Amongst the interaction-detection-based models, the GAM method performed best both on the random split and on the 2018-2019 test data. The IBLM outperformed all the baseline models and the XGB-informed GLM, but not the interaction detection approaches, on the random split. However, on the 2018-2019 test data, the IBLM underperformed all other models.

Table 3: Improvement of average Poisson deviance above ILEC VBT 2015 basis on test datasets by models

Table 1 ID	**Model**	**Test dataset**			
		Random split		**2018-2019 observation years**	
		Improvement above benchmark	**Rank**	**Improvement above benchmark**	**Rank**
#3	LASSO baseline (lambda min)	0.0462	11	0.0367	9
#2	GLM baseline	0.0466	10	0.0373	7
#1	GAM baseline	0.0469	9	0.0368	8
#7	XGBoost-informed GLM	0.0491	8	0.0378	6
#6	IBLM (depth 2)	0.0492	7	0.0299	11
#6	IBLM (depth 3)	0.0516	6	0.0323	10
#10	LASSO + XGBoost detected interactions (lambda min)	0.0527	5	0.0409	4
#9	GLM + XGBoost detected interactions	0.0537	4	0.0388	5

#8	GAM + XGBoost detected interactions	0.0545	3	0.0410	3
#5	XGBoost (offset by baseline GLM)	0.0553	2	0.0433	2
#4	XGBoost (offset by baseline GAM)	0.0563	1	0.0436	1

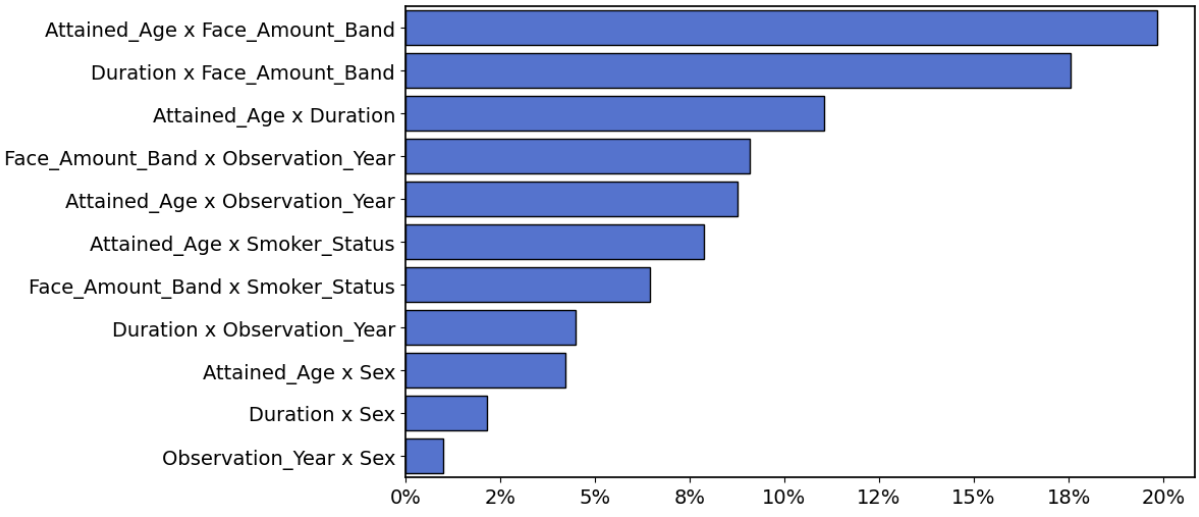
3.2. Analysis insight

3.2.1. Feature importance

Feature importance rankings are broadly similar across different models. Thus, we focus on the XGBoost results shown in **Figure 1**. Attained age is the most important feature, consistent with its well-known exponential relationship with mortality. Face amount band is the second most important feature, ranking above smoker status, duration and sex. This may reflect the wide range of sums assured from $0-9,999 to $10 million+ in the dataset. Unlike typical UK life insurance datasets, over 13% of the life years exposure is for policies in excess of $1 million.

Since the final GAM (model #8, **Table 1**) is the best-performing interaction-detection-based model, we use it to explain the insights from the relative importance of pairwise interaction terms (**Figure 3**). Compared to our CMI study that applied the same methodology (Tam & Luteijn, 2025), face amount band (sum assured) features more prominently in pairwise interactions in the ILEC dataset. The most important interaction is between face amount band and attained age, which ranked second in our CMI study. The interaction between duration and face amount band ranked second on the ILEC data, but 8th in the CMI study. We hypothesise that face amount band is more important in the ILEC data because of the wider range of sums assured, higher levels of income inequality in the US, and the greater variation in underwriting comprehensiveness by sum assured in the US.

Figure 3: Interaction importance for the final GAM (model #8)



3.2.2. Partial dependency

We focus on explaining the insights from the partial dependency plots from GAM, which is the best-performing model under the interaction detection approach.

Baseline GAM

This section describes the impact of the top 4 features (**Figure 4**) in the baseline GAM (model #1) on mortality in the ILEC dataset.

For attained age, mortality is elevated between ages 20 and 30 before reaching the lowest level in the early thirties (**Figure 4**). This effect was also seen in our CMI analysis (Tam & Luteijn, 2025) and is known as the excess mortality hump, which usually takes place between ages 15 and 30 across different countries (Remund, et al., 2021). This effect may be more prominent for an insured population where external causes of death, such as unintentional injuries, suicide and homicide, are more difficult to screen out through underwriting. It could also be that at younger ages, anti-selection is more visible due to lower baseline mortality amongst younger lives in the general population.

The mortality risk of smokers is about 150% higher than that of non-smokers after controlling for other factors (**Figure 4**), which corroborates the regression coefficient of 2.62 for smokers (relative to non-smokers) in the baseline GLM. This excess mortality amongst smokers varies by age, as evidenced by the interaction between attained age and smoking status being the 6^{th}-ranked interaction (**Figure 3**).

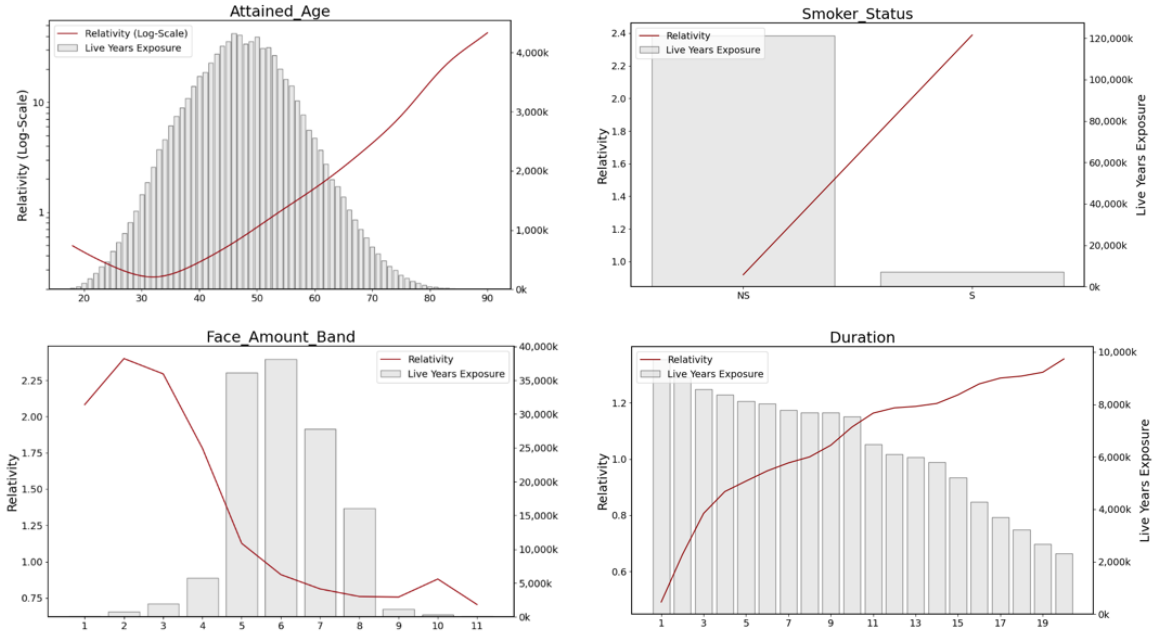
In this model, face amount band is treated as an ordinal variable and the mapping from the bands to ordinal values is shown in **Table 4**. Mortality decreases as the amount bands increase, with the first three bands having similarly high risk before dropping sharply for bands $100-250K and $250-500K. The sharp differential between below $100K policies (which often have weaker underwriting) and policies above $100K often leads to the exclusion of policies below $100K in ILEC experience studies (Individual Life Insurance Experience Committee of the Society of Actuaries, 2024).

The partial dependence plot for duration (**Figure 4**) shows an extraordinarily long select shape of at least 20 years. It is conceivable this is a product of improvements in underwriting and products over time, especially since our models did not incorporate any adjustments for issue year. In the US, products have become progressively sophisticated in relation to number of preferred classes. Since these features were outside of scope of our models, any confounding by product type including number of preferred classes will not have been adjusted for.

Table 4: Ordinal encoding of face amount bands

Face_Amount_Band	Ordinal Value
01: 0 - 9,999	1
02: 10,000 - 24,999	2
03: 25,000 - 49,999	3
04: 50,000 - 99,999	4
05: 100,000 - 249,999	5
06: 250,000 - 499,999	6
07: 500,000 - 999,999	7
08: 1,000,000 - 2,499,999	8
09: 2,500,000 - 4,999,999	9
10: 5,000,000 - 9,999,999	10
11: 10,000,000+	11

Figure 4: Partial dependence plots for the top 4 features in the baseline GAM



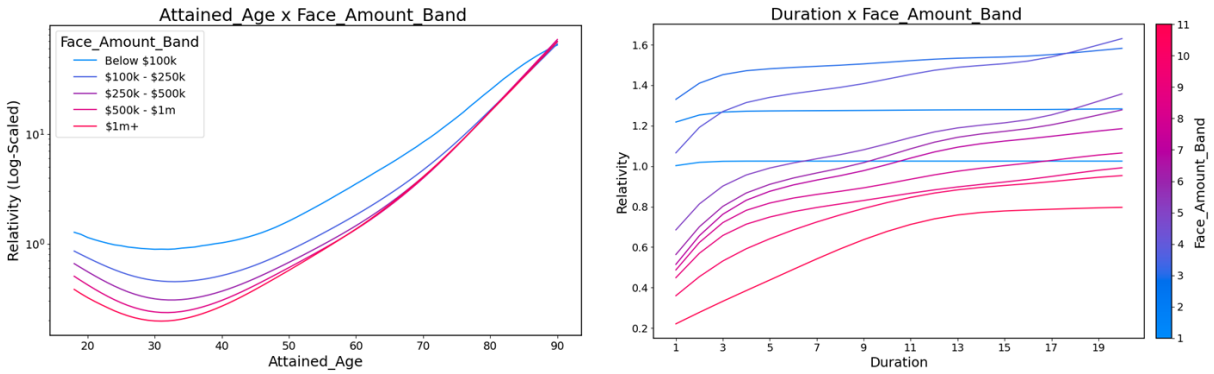
Final GAM with interactions

This section describes the impact of the top interactions (**Figure 5**) in the final GAM (model #8) on mortality in the ILEC dataset.

The highest-ranking interaction is between face amount band and attained age (**Figure 3**). Mortality is especially high for the face amounts below $100K across all ages. The differential in mortality rates by face amount band is especially wide for ages 30-40, following which it converges with increasing age. From age 70 onwards, there is no meaningful mortality difference across face amount bands above $100K, but the mortality excess for policies below $100K remains elevated until age 90.

The second-ranking interaction is between face amount band and duration (**Figure 3**). In addition to the clearly much higher mortality for smaller face amount bands, there is a clear dichotomy between bands 1-4 (below $100K) and bands 5-11 (above). For policies below $100K, there is barely a select effect, presumably due to much lower underwriting requirements for these policies in the US. (Society of Actuaries Research Institute, 2024) For policies above $100K, full underwriting is the default and this is reflected in the presence of a select shape. The mortality differential for face amount bands above $100K does not appear to compress at higher durations.

Figure 5: Top two interaction partial dependence plots for model #8



4. Appraisal of the hybrid approaches

We demonstrated three innovative approaches for using gradient boosting to enhance traditional additive models in the context of mortality modelling. Additionally, we tested variations of the interaction detection approach using additive models (i.e. GLM, LASSO and GAM) as the final output.

4.1. Model performance

All three hybrid approaches (models #6-10) outperformed the baseline models (models #1-3), but underperformed the black-box models (models #4-5). Amongst the three hybrid approaches, the interaction detection approach was the most performant across both test datasets (**Table 3**). The performance of the XGBoost-informed GLM and the interaction detection approach depends upon the modeller's ability to distinguish true signal from noise, perform feature engineering and craft an effective regression formula. This reliance on modeller expertise is especially acute for GLM and LASSO, as GAM auto-calibrates nonlinearities.

IBLM has not previously been applied to life and health insurance datasets and it is conceivable that for life and health datasets the initial GLM leaves a lot of signal on the table for the booster model to pick up. For example, in life and health, the relationship between attained age and mortality is exponential, as can be seen in both **Figure 2** and **Figure 4**. A single linear term in the initial GLM is insufficient to capture its complexity. The performance ranking of IBLM dropped for the out-of-time test data. This is hypothesised to be due to the inability of the internal booster model to extrapolate an observation year effect.

Amongst the interaction detection approaches, GAM outperforms GLM and LASSO on both the random split and the 2018-2019 test data when including interaction terms. This could be due to GAM having more flexible curve fitting, while controlling for overfitting via both L2 regularisation and a roughness penalty on numerical features.

In the random split, the final GAM attains a deviance improvement 16% higher than the baseline GAM. This is about 80% of the improvement attained by the corresponding XGBoost. In the 2018-2019 test data, the final GAM attains 11% improvement, about 62% of the corresponding XGBoost improvement. With sufficient lives exposure, as in the ILEC dataset, there is a trade-off between transparency and accuracy. However, health & care insurers do not usually have sufficient exposure or achieve full claim credibility at the level of customer granularity they require. Under these circumstances, underwriting knowledge and actuarial judgement are key to determining whether risk models capture genuine trends. Having transparent models enables scrutiny and ensures they do not overfit to training data.

4.2. Modeller control

The three approaches are subject to varying levels of automation, with the IBLM being the most automated, and the XGB-informed GLM the least automated. The interaction detection approach and XGB-informed GLM offer high levels of modeller control since in both cases, the XGB is used as a diagnostic tool rather than the final predictive engine. The modeller retains full control over how variables are engineered and what interactions are incorporated in the model. This does come at a cost of increased fragility as predictive performance is more dependent upon the expertise and domain knowledge of the modeller.

The IBLM approach automatically detects nonlinearities and interactions and incorporates these back into the final model predictions through SHAP-derived beta adjustments. This automation reduces manual feature engineering and may improve predictive performance. Unlike the other two approaches, all possible variable interactions are considered by default. The IBLM booster can be controlled by specifying interaction and monotonic constraints and tree depth offering a degree of modeller control. Contrary to the two other approaches (which require a regression formula), IBLM offers strong out-of-the-box performance. Due to the split-based nature of XGB, the IBLM approach

does not naturally produce smooth relationships between continuous predictors and outcomes, which may impair predictive performance, especially where high learning rates have been selected.

All three approaches offer meaningful levels of modeller control. The main difference is that the interaction detection approach and the XGBoost informed GLM allow the modeller to directly specify the functional form of predictor-outcome relationships. For example, logarithmic transformations can enforce proportional effects of sum assured on the outcome, whilst trigonometric functions enforce a seasonal effect. These explicit levels of control allow actuaries to embed domain knowledge into their models and support extrapolation and interpretation (see, for more examples (Luteijn, et al., 2025), Section 3.9.1). The split-based nature of XGBoost does not allow for enforcing these types of effects.

4.3. Transparency and governance readiness

Although TAS 100 does not define a transparency requirement, compliance with principle 5 (models) and 7 (communication) requires a level of understanding and explainability that transparency is, in effect, mandated. (Financial Reporting Council, 2023) Lack of transparency could result in biases against certain customer segments that violate regulatory requirements or produce model outputs changes that cannot be adequately justified to internal stakeholders (e.g. risk managers) and external ones (e.g. customers).

The interaction detection approach and the XGB-informed GLM result in final additive models, including explicitly specified main effects and variable interactions, each with fixed regression coefficients. Although regression coefficients of polynomials can be less intuitive to interpret, individual effects and interactions can be visualised explicitly and scrutinised using domain knowledge, supporting validation and governance review.

In contrast, the final output of IBLM is a GLM model, augmented by a set of SHAP-derived beta adjustments, reflecting the contribution of the internal XGB booster model on the linear predictor scale. These beta-adjustments can be highly observation-specific, subject to the tree-depth of the booster model. Whilst in our study, tree depths of 2 and 3 were modelled, restricting the complexity of the beta adjustments, deeper trees can capture higher-order variable interactions, materially reducing interpretability whilst potentially boosting predictive accuracy. This reduced interpretability for deep IBLM booster models can have negative implications for governance and transparency. By contrast, the IBLM approach will produce consistent results across different modellers and therefore is not exposed to governance risk if interaction selection or feature engineering rationale is poorly documented.

4.4. Applicability to actuarial modelling tasks

The three hybrid approaches assessed aim to improve the ability of additive models to capture nonlinearities and interactions at the structural level. Therefore, the contribution of the paper is methodological, rather than outcome-specific.

Although the dataset analysed contains mortality events, the underlying modelling setup is representative of a broader set of actuarial modelling tasks. Mortality, morbidity, income protection, lapse and fraud detection all involve modelling of rare event counts per unit of exposure. These tasks involve strong nonlinearities and interactions across biometric (e.g. age, sex, smoking), insurance (e.g. sales channel, underwriter loading) and temporal (e.g. duration, calendar year, issue year) features. Residual boosting, interaction detection and SHAP-informed feature engineering are designed to identify and incorporate such nonlinearities and interactions in the linear predictor, independent of whether the underlying outcome represents mortality, morbidity or lapses.

4.5. Next steps

This paper focused on hybrid modelling approaches that produce a final additive model closely aligned with established GLM-centric actuarial workflows. Therefore, various alternative explainable ML models that depart more substantially from standard actuarial workflows were not appraised. This scoping choice was made to preserve comparability across approaches, limit computational and tuning complexity, and maintain a consistent governance and documentation framework.

A natural extension would be to appraise fully automated explainable supervised techniques, including explainable boosting machines (Nori, et al., 2019), neural network models with connectivity restricted to just univariate and pairwise interactions, for example NODE-GAM (Chang, et al., 2022). These models are further removed from traditional actuarial workflows. Comparison against the hybrid approaches would focus on the trade-off between prediction accuracy and model complexity, model complexity and governance. Another more natural extension could be to include approaches leveraging ML to engineer features for downstream additive models. (Dai, 2018; Henckaerts, et al., 2018; Maillart, 2021)

Beyond model classes, future work could extend the appraisal to morbidity, income protection, lapse or fraud detection tasks. Critical illness claims are a particular promising area for application of hybrid approaches since public health interventions such as HPV vaccination and breast cancer screening cause strong nonlinearities in claims incidence by age and birth cohort. Additionally, there is higher risk of anti-selection (compared to mortality claims), which may manifest itself as interactions between sum assured, issue age and/or sales channel. However, health and care datasets may contain additional complicities including partial claims, severity ratings and heterogeneous outcome definitions which may not translate to mortality datasets.

Finally, there is scope to enhance the hybrid approaches themselves. The interaction detection approach could be enhanced by use of diagnostics such as SHAP dependence plots to identify potential poorly captured nonlinearities in the baseline. The IBLM baseline model could be extended to allow bespoke regression formulae into the IBLM baseline model and hyperparameter tuning of the XGBoost model, trading training time for improved predictive performance.

5. Conclusion

To our knowledge, this is the first actuarial study to systematically appraise hybrid GLM–ML approaches along governance-relevant dimensions rather than predictive accuracy alone, using a common dataset and evaluation protocol. Across the appraisal dimensions considered in this paper, no single modelling approach dominates in all respects. Instead, the three hybrid approaches occupy different points along the trade-off between automation, modeller control, and transparency. The fully automated IBLM approach offers strong out-of-the-box performance with minimal manual intervention. Therefore, IBLM is well suited for time constrained modellers and modellers with little domain knowledge to hand-craft regression formulae.

The more manual XGBoost-informed GLM and interaction detection approaches allow experienced modellers to incorporate domain knowledge directly into the modelling process, particularly when capturing nonlinearities and interaction effects. These approaches retain an explicit additive structure and naturally support extrapolation in continuous features, which can be advantageous in actuarial applications requiring forward-looking judgement.

In conclusion, our findings suggest that hybrid modelling frameworks provide a flexible and practical pathway for leveraging state-of-the-art black box models into actuarial practice. For actuaries seeking to improve on purely GLM-based methods, without compromising on transparency, these hybrid approaches represent a pragmatic middle ground.

Acknowledgements
We would like to thank Ramen Marudamuthu for his valuable input and for providing insights from his perspective as a pension actuary.

6. References

Academy/SOA Valuation Basic Table Team, 2018. *2015 Valuation Basic Table Report.* [Online]
Available at: https://www.soa.org/globalassets/assets/files/resources/experience-studies/2018/2015-vbt-report.pdf

Akiba, T. et al., 2019. *Optuna: A Next-generation Hyperparameter Optimization Framework.* s.l., s.n., pp. 2623-2631.

Bjerre, D. S., 2022. Tree-based machine learning methods for modeling and forecasting mortality. *ASTIN Bulletin: The Journal of the IAA,* 52(3), pp. 765-787.

CAS Machine Learning Working Party, 2022. Machine Learning in Insurance. *Casualty Actuarial Society E-Forum,* Winter.pp. 1-16.

Chang, C.-H., Caruana, R. & Goldenberg, A., 2022. NODE-GAM: Neural Generalized Additive Model for Interpretable Deep Learning. *International Conference on Learning Representations.*

Continuous Mortality Investigation, 2021. *Proposed "16" Series term assurance mortality and accelerated critical illness tables,* London: Institute and Faculty of Actuaries.

Dai, J., 2018. Enhancing the Generalized Linear Modeling Approach with Machine Learning Technique. *Casualty Actuarial Society E-Forum,* Volume Spring 2018 – Volume 2, pp. 1-9.

Financial Reporting Council, 2023. *Technical Actuarial Standard 100: Principles for Technical Actuarial Work (Version 2.0),* London: Financial Reporting Council.

Gawlowski, K. & Beard, P., 2025. *R Package: Interpretable Boosted Linear Models. V1.0.2.* London: s.n.

Gawlowski, K. & Wang, P., 2025. Interpretable Boosted GLM. *Available at SSRN: https://ssrn.com/abstract=5751803 or http://dx.doi.org/10.2139/ssrn.5751803,* 30 08.

Grinsztajn, L., Oyallon, E. & Varoquaux, G., 2022. Why do tree-based models still outperform deep learning on tabular data?. *ArXiv,* Volume 2207.08815.

Henckaerts, R., Antonio, K., Clijsters, M. & Verbelen, R., 2018. A data-driven binning strategy for the construction of insurance tariff classes. *Scandinavian Actuarial Journal,* Issue 8, pp. 681-705.

IFoA Techniques in Data Science in Health and Care, 2026. *GitHub.* [Online]
Available at: https://github.com/ckjackytam/ifoa-ds-health-care-wp/tree/main
[Accessed 02 02 2026].

Individual Life Insurance Experience Committee of the Society of Actuaries, 2024. *2019 Individual Life Insurance Mortality Experience Report.* [Online]
Available at: https://www.soa.org/resources/research-reports/2024/ilec-mort-2012-19/

Kshirsagar, R. et al., 2021. Accurate and Interpretable Machine Learning for Transparent Pricing of Health Insurance Plans. *Proceedings of the AAAI Conference on Artificial Intelligence,* 35(17), p. 15535–15543.

Luteijn, J. et al., 2025. A framework for applying data science techniques to health and care actuarial projects. *Authorea,* Issue DOI: 10.22541/au.174733518.83607343/v1.

Maillart, A., 2021. Toward an explainable machine learning model for claim frequency: a use case in car insurance pricing with telematics data. *European Acturial Journal,* 11(2), pp. 579-617.

Nori, H., Jenkins, S., Koch, P. & Caruana, R., 2019. InterpretML: A Unified Framework for Machine Learning Interpretability. *arXiv preprint arXiv:1909.09223.*

Orji, U. & Ukwandu, E., 2023. Machine Learning For An Explainable Cost Prediction of Medical Insurance. *Machine Learning with Applications,* Volume 15, p. 100516.

Remund, A., Camarda, C. G. & Riffe, T., 2021. *Is young adult excess mortality a natural phenomenon?.* [Online]
Available at: https://www.ined.fr/en/publications/editions/population-and-societies/is-young-adult-excess-mortality-a-natural-phenomenon/
Richman, R., Scognamiglio, S. & Wüthrich, M. V., 2025. The credibility transformer. *European Actuarial Journal.*
Richman, R. & Wüthrich, M. V., 2021. A neural network extension of the Lee-Carter model to multiple populations. *Annals of Actuarial Science,* 15(2), pp. 346-366.
Richman, R. & Wüthrich, M. V., 2023. LocalGLMnet: interpretable deep learning for tabular data. *Scandinavian Actuarial Journal,* 2023(1), pp. 71-95.
Society of Actuaries Research Institute, 2024. *2019 Individual Life Insurance Mortality Experience Report.* [Online]
Available at: https://www.soa.org/4a946d/globalassets/assets/files/resources/research-report/2024/ilec-mort-main.pdf
Tam, J. & Luteijn, M., 2025. *A new pathway: A framework for incorporating data science into health and care.* [Online]
Available at: https://www.theactuary.com/2025/07/02/new-pathway-framework-incorporating-data-science-health-and-care