Scientific Research Publishing

Journal of Computer and Communications, 2026, 14(7), 65-88
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Beyond Black-Box Benchmarking: A Statistically Validated, Interpretable, and Uncertainty-Aware Machine Learning Framework for Insurance Premium Estimation

Jude Isememe Itoyah

School of Computing and Digital Media, London Metropolitan University, London, UK
Email: terajude@gmail.com

How to cite this paper: Itoyah, J.I. (2026) Beyond Black-Box Benchmarking: A Statistically Validated, Interpretable, and Uncertainty-Aware Machine Learning Framework for Insurance Premium Estimation. *Journal of Computer and Communications*, **14**, 65-88.
https://doi.org/10.4236/jcc.2026.147004

Received: May 19, 2026
Accepted: July 19, 2026
Published: July 22, 2026

Abstract

The application of machine learning to insurance premium estimation has accelerated considerably in recent years. However, the field faces three structural weaknesses. These include over-reliance on black-box models without explainability, absence of credible statistical validation, and failure to quantify actuarial uncertainty around point predictions. This research addresses all three limitations within a unified, end-to-end experimental framework using a dataset collected from Kaggle. Eight interaction and polynomial features grounded in actuarial domain knowledge are engineered from the retrieved dataset. An out-of-fold (OOF) stacked ensemble is proposed, combining Linear Regression, Random Forest, XGBoost, and LightGBM as base learners under a Ridge meta-learner. All models are evaluated using a 10-fold cross-validation protocol with Wilcoxon signed-rank statistical testing. There are critical findings from the ML experimentation. With sufficiently rich interaction features, a simple Linear Regression model achieves $R^2 = 0.8827$ on the held-out test set, statistically indistinguishable from the stacked ensemble ($R^2 = 0.8825$). All baseline models were trained on the original monetary expense scale, while a $\log_1(1 + y)$ transformation was applied solely to support interaction feature construction. SHAP values are therefore expressed directly in monetary units (£) without requiring back-transformation. These results challenge the prevailing assumption in the literature that insurance premium prediction fundamentally requires non-linear architectures. Also, SHAP analysis reveals that five of the six most influential features are interaction terms, with smoker × BMI contributing a mean absolute SHAP value of approximately £3215 per prediction, which is 50% more than the marginal smoking effect alone. Quantile regression produces prediction intervals covering 69.8% of test instances,

with widths increasing disproportionately in high-risk strata. The framework is benchmarked against six existing works and offers improvements in predictive validity, interpretability, and statistical rigour on this benchmark dataset.

Keywords

Insurance Premium Prediction, Machine Learning, Interaction Feature Engineering, SHAP, Explainability, Quantile Regression, Stacked Ensemble, Bayesian Optimization

1. Introduction

The estimation of insurance premium is one of the most consequential pricing problems in financial services [1]. The accuracy of a premium directly determines an insurer's ability to maintain solvency in the face of competition. This is very important because overestimation triggers policyholder churn and underestimation leads to losses that compound over claim cycles [2] [3]. In the past, this particular problem has been approached using computational methods like Generalised Linear Models (GLMs), Zillmer reserve methods, and stochastic interest rate frameworks. By conceptualisation, all these methods are rooted in the assumption that the relationship between policyholder characteristics and expected claims is either linear or follows a pre-specified parametric distribution [4]-[6]. In modern insurance markets, these assumptions are weak.

Machine learning (ML) is a compelling alternative method because it addresses distributional assumptions, learns interaction effects automatically, and scales to datasets of any dimensionality. A growing body of empirical literature demonstrates that machine learning methods like gradient boosting and ensemble methods outperform GLMs on predictive accuracy benchmarks [7]-[10]. However, these literatures are characterized by three recurring structural weaknesses that limit their practical and scientific value.

First, most existing works are evaluated on a single train-test split, making it impossible to strongly establish genuine model superiority due to random variance in the data partition. Second, most works focus on the accuracy of the models implemented at the expense of interpretability. Importantly, as regulators increasingly mandate explainability under frameworks such as the EU AI Act and GDPR Article 22, models that cannot explain individual predictions are not deployable in regulated insurance contexts, regardless of how accurate the models are. Third, most existing works in insurance premium estimation focus on point predictions without any quantification of uncertainty. This is a significant practical limitation, considering that actuaries and underwriters often require confidence bounds to manage risk capital. This research is motivated to address all three weaknesses within a unified framework.

1.1. Research Aim

The aim of this research is to develop a statistically validated, interpretable, and uncertainty-aware machine learning framework for insurance premium estimation. This robust framework is implemented to advance the state of the art through detailed interaction feature engineering, a novel stacked ensemble architecture, Bayesian hyperparameter optimization, SHAP-based explainability, and quantile regression uncertainty quantification.

1.2. Research Questions

The entire empirical investigation in this research is guided by four research questions:

RQ1: To what extent does domain-driven interaction feature engineering change the comparative performance hierarchy of ML models, and which feature combinations carry the most actuarial signal?

RQ2: Does out-of-fold stacked ensemble statistically outperform individual base learners under k-fold CV and Wilcoxon significance testing, and under what conditions does added model complexity yield diminishing returns?

RQ3: What do SHAP-based explanations reveal about synergistic and non-linear risk factors that drive insurance premium variation?

RQ4: Can quantile regression bound actuarial uncertainty in predicted premiums, and how do prediction interval coverage and width vary across risk strata?

2. Related Work

2.1. Traditional Actuarial Methods and Their Limitations

The dominant paradigm in actuarial science for premium calculation over the years has been the Generalised Linear Model (GLM). This computational method connects expected claims to policyholder risk factors through a link function and an exponential family distribution [11]-[13]. GLMs impose two constraints that cause issues most of the time when the method is applied. First, they assume each predictor exerts an additive, independent effect on the link-transformed outcome [14] [15], an assumption that fails when the joint effect of two risk factors substantially exceeds their individual contributions. The second assumption concerns maximum likelihood estimation. In large-scale GLM applications, this estimation is computationally expensive and relies on asymptotic approximations that are sometimes unreliable because of skewed, heavy-tailed claim distributions [16].

Classical actuarial methods like the Zillmer reserve method and the Premium Sufficiency method address the adequacy of collected premiums instead of the accuracy of individual premium predictions, and as such, operate at a different level of analysis from the predictive ML models that are examined here. Fundamentally, the Zillmer reserve approach and premium sufficiency testing are primarily concerned with ensuring the adequacy and financial consistency of premiums at the portfolio level through reserve adjustment and aggregate liability coverage. Also, the Bayesian extensions of GLMs partially address robustness con-

cerns by incorporating prior information. However, they inherit the linearity assumption and introduce sensitivity to prior misspecification.

2.2. Machine Learning in Insurance Premium Estimation

The application of ML to insurance premium prediction has grown over the years because of advancements in computation in the ML domain. Spedicato, Dutang, and Petrini [17] provide empirical evidence from their research on motor vehicle liability data, where they showed that gradient boosting methods such as XGBoost (AUC = 0.9064) outperform traditional GLMs (AUC = 0.8896). While Spedicato, Dutang, and Petrini [17] employ standard validation techniques including train-test splits and cross-validation, model comparisons are mainly based on predictive performance metrics (AUC and log loss) without formal statistical significance testing.

Studies like Bau and Hanif [18], Orji and Ukwandu [19], Patil, Kulkarni and Khurpe [20], Ridzuan *et al.* [21], and Kaushik *et al.* [22] all reported strong predictive performance, but model evaluation is largely centred on point estimation metrics like R^2, MAE, and RMSE. Also, these works explicitly use cross-validation procedures to assess generalisation stability for different data partitions. Also absent in most literature is the application of statistical significance testing, such as Wilcoxon signed-rank or Friedman tests, to confirm that observed differences in model performance are not attributable to random variation in a single train-test split. Critically, most works do not incorporate uncertainty quantification frameworks, such as prediction intervals or quantile regression, despite the actuarial relevance of this special handling to communicate confidence bounds around individual premium estimates. All these methodological gaps collectively limit the reproducibility, statistical validity, and practical deployability of the findings in ML-based premium prediction investigations.

The interpretability gap is another area that has been acknowledged in many existing works but weakly addressed. Studies like Bau and Hanif [18], Billa and Nagpal [23], Patil, Kulkarni and Khurpe [20], and Kaushik *et al.* [22] mentioned the black-box nature of ML models as a key limitation, but they did not propose or implement any solution. While this body of work is important in highlighting the limitations of current approaches and advancing predictive performance, it leaves an unresolved gap at the intersection of accuracy and interpretability. Addressing this gap is very important to translate the ML models being adopted in the insurance domain from purely predictive tools into reliable decision-support systems. In this regard, this research contributes by explicitly incorporating explainability mechanisms to improve model transparency, interpretability, and practical applicability in real-world insurance pricing decisions.

3. Methodology

The study adopts a structured machine learning pipeline for predicting insurance expenses (**Figure 1**). After data ingestion, exploratory analysis and descriptive

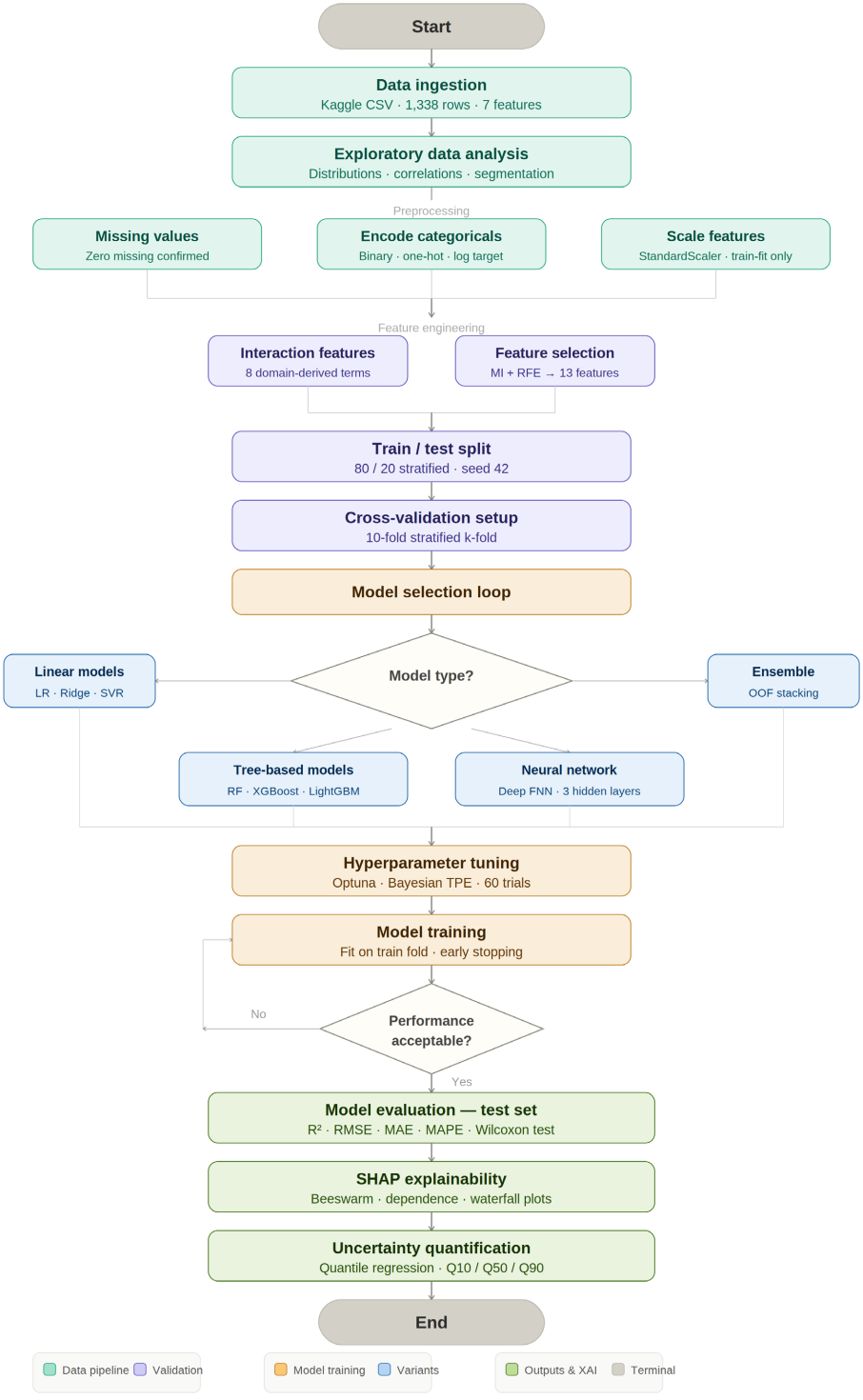

Figure 1. Workflow of implementing ML models for insurance premium prediction.

statistics are performed to understand distributions and data quality. Preprocessing includes handling missing values, encoding binary categorical variables (sex, smoker) as 0/1 indicators, one-hot encoding the region variable into four binary columns, and standardising continuous features. Feature engineering con-

structs eight interaction and polynomial terms grounded in actuarial domain knowledge. Feature selection applies Mutual Information scoring and Recursive Feature Elimination. The dataset is split into 80% training and 20% test sets, with 10-fold cross-validation on the full dataset providing generalisation estimates. Six baseline models are implemented: Linear Regression, SVR with RBF kernel, Random Forest, XGBoost, LightGBM, and a Deep Feedforward Neural Network (FNN). An out-of-fold (OOF) stacked ensemble combines Linear Regression, Random Forest, XGBoost, and LightGBM as base learners under a Ridge meta-learner. XGBoost hyperparameters are optimised using Optuna, while model performance is evaluated using R^2, RMSE, MAE, and MAPE. SHAP TreeExplainer is applied to the tuned XGBoost model for interpretability.

3.1. Dataset

This research employs the Medical Cost Personal Dataset [24]. The dataset comprises 1338 Observations for seven variables: age (18 - 64 years), sex (binary), BMI (15.96 - 53.13 kg/m^2), number of children (0 - 5), smoking status (binary), US Census region (four-category nominal), and annual insurance expenses (target, range US$1121 - 63,770). The dataset is widely recognized as a simulated. Synthetic data that is constructed for educational and benchmarking purposes. All findings in this research are therefore conditional on this synthetic benchmark and should not be generalized directly to Live underwriting or pricing systems without validation on proprietary real-world data. The target variable, which is insurance expenses, shows considerable right skew (skewness = 1.515 on the raw scale).

To mitigate this skewness during interaction feature construction, a natural-logarithm transformation $\log_1(1 + y)$ was applied to engineer the interaction and polynomial features. All six baseline models and the stacked ensemble was trained and evaluated on the original monetary scale (raw expenses in £). All Reported evaluation metrics (R^2, RMSE, MAE, MAPE) are therefore computed on the original expense. Scale, and no back-transformation of predictions is required. The log-transformed series was not used as a modeling target (**Figure 2**).

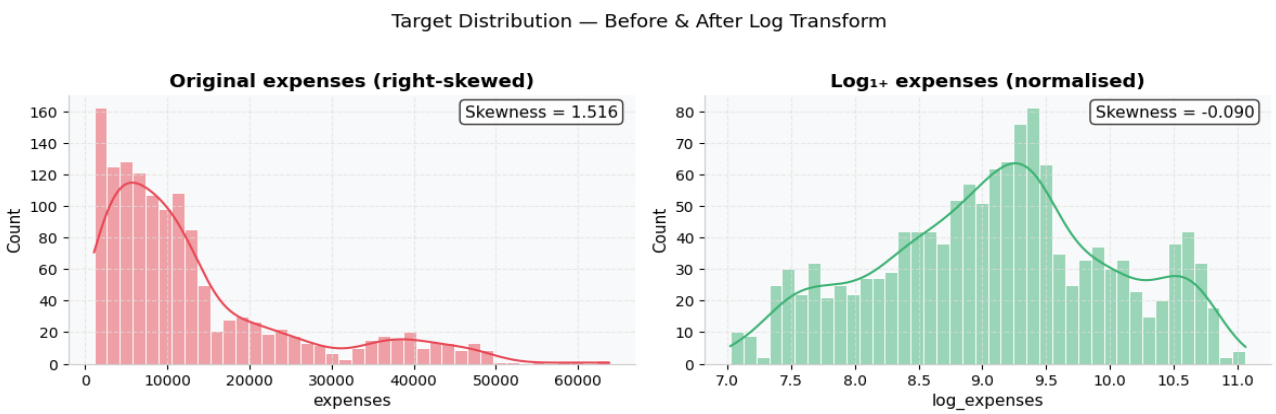

Figure 2. Target distribution: Before and after $\text{Log}_1(1 + y)$ transformation.

3.2. Preprocessing

Preprocessing is implemented in three phases. Binary categorical variables (sex, smoker) were encoded as 0/1 indicators. Region was one-hot encoded into four binary columns, all of which were retained. Continuous features (age, BMI, children) were standardised using training-set statistics only, with the scaler applied to the test set post-fit to prevent data leakage. A $\log_1(1 + y)$ transformation of the target was computed mainly to produce a normalised scale for interaction feature construction. It was not used as a modelling target. All six baseline models and the stacked ensemble were trained on the raw expense values in £. All evaluation metrics (R^2, RMSE, MAE, MAPE) were computed on this original monetary scale, and SHAP values output by TreeExplainer are therefore expressed directly in £ without any back-transformation.

3.3. Interaction Feature Engineering

Eight interaction and polynomial features were constructed from actuarial first principles. The motivation for this is grounded in established clinical and actuarial evidence that obesity and smoking exert a superadditive joint effect on health costs, and that age-related healthcare cost trajectories are non-linear [25] [26]. The engineered terms include smoker × BMI, smoker × age, age × BMI, age^2, BMI^2, a binary BMI_obese indicator (BMI ≥ 30 kg/m^2), smoker × obese, and age × children. These terms enable the ML models to capture clinically grounded non-linearities and interaction effects that would be systematically missed under additive specifications, thereby improving predictive fidelity and risk stratification.

3.4. Feature Selection

Feature selection used two complementary methods. These are Mutual Information (MI) regression scoring and Recursive Feature Elimination (RFE) with a Random Forest estimator selecting ten features. The union of the top 10 MI features and the 10 RFE-selected features, augmented by forced retention of five core actuarial features, produced a final set of thirteen variables. All eight engineered interaction/polynomial terms were retained, confirming their informational value in the pipeline.

3.5. Models

Six baseline models were implemented. These include Linear Regression (OLS), SVR with RBF kernel (C = 100, γ = scale), Random Forest (300 trees, min_samples_leaf = 2), XGBoost (500 trees, lr = 0.05, max_depth = 6), LightGBM (500 trees, 31 leaves), and a Deep FNN (three hidden layers: 256-128-64 nodes; batch normalisation; dropout 0.3 - 0.1; early stopping). The novel Out-of-Fold (OOF) stacked ensemble combined LR, RF, XGBoost, and LightGBM as base learners with a Ridge meta-learner, trained via 5-fold OOF cross-validation to prevent leakage. XGBoost was further optimised using Optuna Bayesian HPO (60 trials;

TPE sampler).

3.6. Evaluation Protocol

The primary evaluation uses an 80/20 train-test split (random_state = 42) applied to the full 1338-observation dataset, producing 1070 training and 268 test instances. Because the target variable (insurance expenses) is continuous, stratification was approximated by binning the target into ten equal-frequency quantile bins (pd.qcut with q = 10) prior to the split. Scikit-learn's train_test_split was then called with stratify set to these bins, ensuring that each decile of the expense distribution is proportionally represented in both partitions. The same ten-bin rule was applied to the 10-fold stratified cross-validation. The complete dataset was partitioned using StratifiedKFold(n_splits = 10, shuffle = True, random_state = 42) with fold assignments derived from the quantile-binned target, producing folds that preserve the premium distribution, including the high-cost smoker tail in every split. This stratification approach and the use of KFold(n_splits = 10, shuffle = True, random_state = 42) for models evaluated in the CV comparison (Section 4.5) are reported here for full reproducibility. The four primary metrics evaluated were R^2, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). Model superiority was assessed using two-tailed Wilcoxon signed-rank tests on paired 10-fold CV R^2 vectors at $\alpha = 0.05$.

3.7. SHAP Analysis and Uncertainty Quantification

SHAP TreeExplainer was applied to the Optuna-tuned XGBoost model on 268 test instances. Quantile regression at $\tau \in \{0.10, 0.50, 0.90\}$ produced approximate 80% prediction intervals using the gradient boosting pinball loss.

4. Results

4.1. Exploratory Data Analysis

The insurance dataset shows a bimodal premium distribution that is largely driven by smoking status. This is confirmed in **Figure 3**, which shows that smokers' premiums are distributed around a median that is about 3.8× higher than non-smokers. Observably, this situation is consistent for all four census regions. The southeast shows the largest dispersion, consistent with its above-average BMI prevalence. The violin plot in **Figure 3** confirms that the bimodal structure is not simply driven by outliers; both modes are populated and clearly represent two different risk sub-populations rather than a continuous gradient.

Also, bivariate scatter analysis was implemented, and the outcomes are presented in **Figure 4**. The Pearson correlation between age and premium is r = 0.30 ($p < 0.001$), and between BMI and premium, r = 0.20 ($p < 0.001$). However, both correlations mask the smoking stratum effect. Within the smoker group, the age-premium relationship is considerably steeper, confirming the theoretical basis for the smoker × age interaction term.

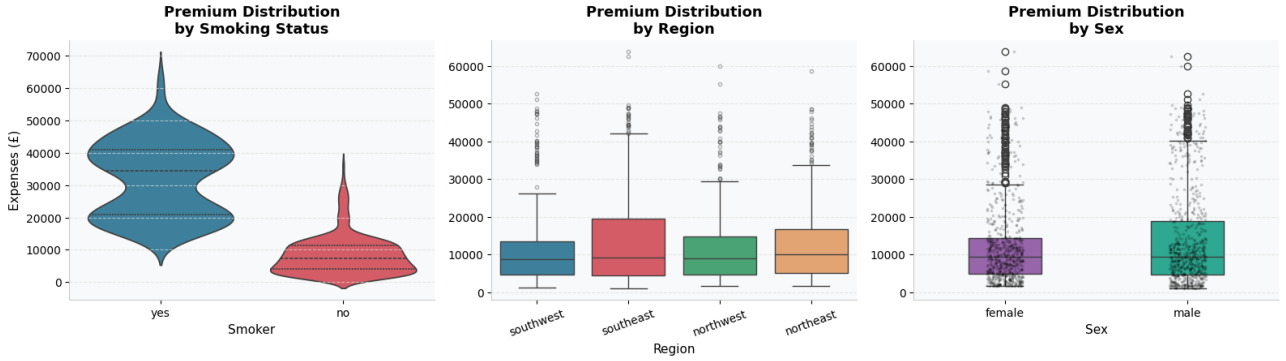

Figure 3. Insurance premium distribution for categorical features (Violin, Box, Swarm).

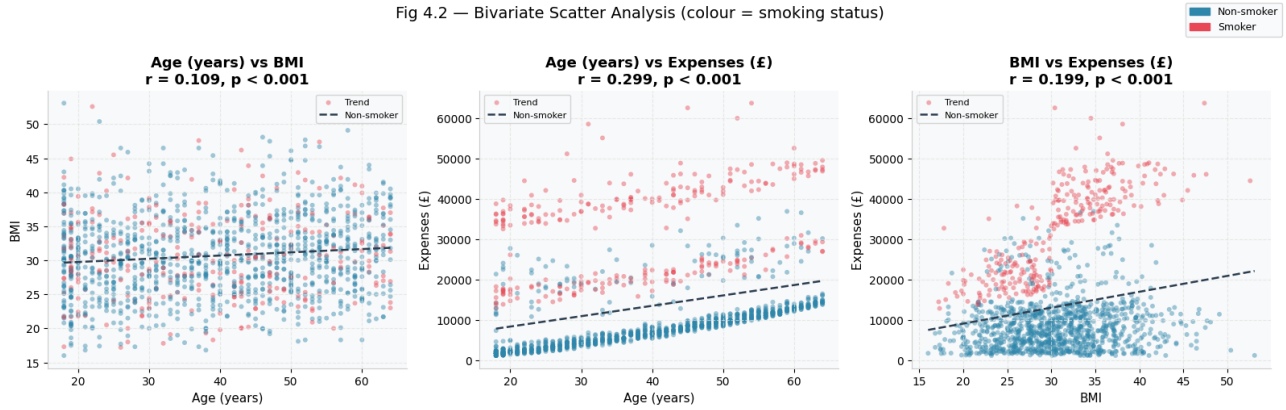

Figure 4. Bivariate scatter analysis: Continuous features vs. premium.

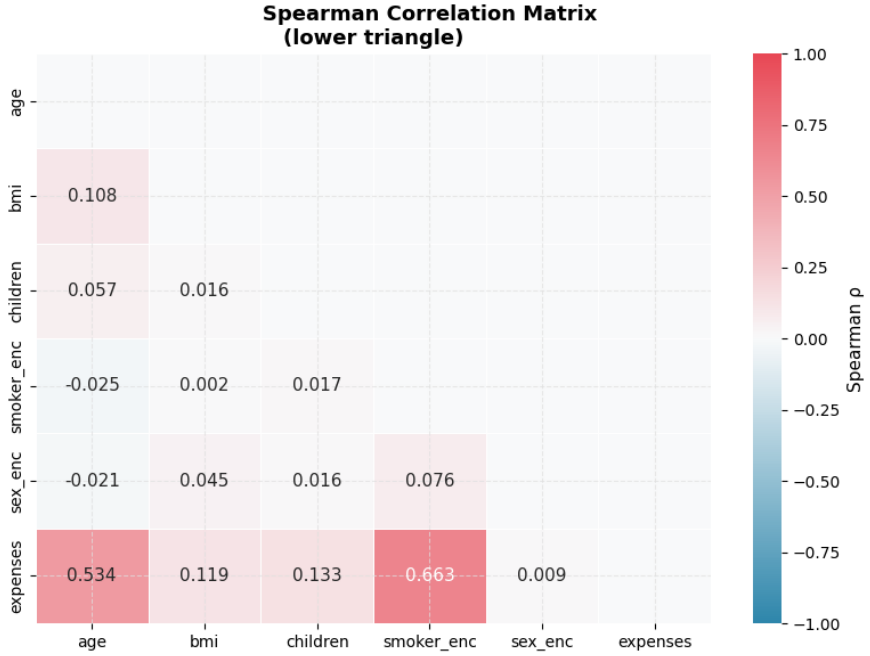

Figure 5. Spearman correlation matrix (Lower Triangle).

The Spearman correlation matrix (Figure 5) confirms that smoking status (ρ

= 0.787) is the dominant predictor. It is followed by age (ρ = 0.526). Sex (ρ = 0.057) shows a negligible association. Notably, the BMI-smoker correlation (ρ = 0.004) is near zero, confirming that BMI and smoking are statistically independent in this dataset. This observation is another strong justification for interpreting their interaction term as a real superadditive effect rather than a collinearity artefact.

The premium segmentation analysis (Figure 6 and Figure 7) reveals non-linear joint effects. The obese + 56 - 64 age bracket shows a median premium that is around 4.1× that of the normal-weight + 18 - 25 bracket. This non-linearity is one of the strong reasons that motivates the interaction feature engineering strategy, and it confirms that additive models trained on the original feature set will systematically underestimate the premium for the highest-risk demographic.

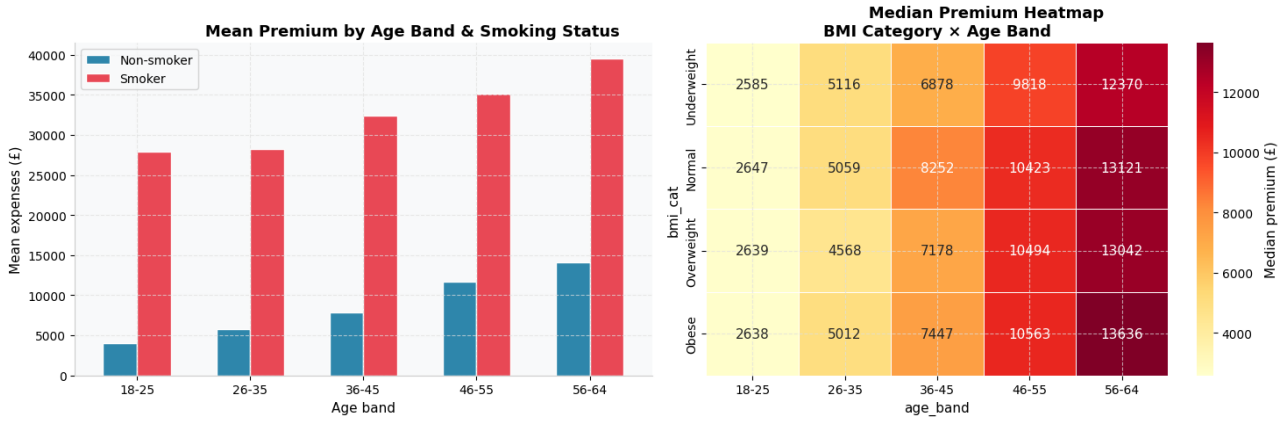

Figure 6. Premium segmentation: Age Band × Smoking Status; BMI Category × Age Band.

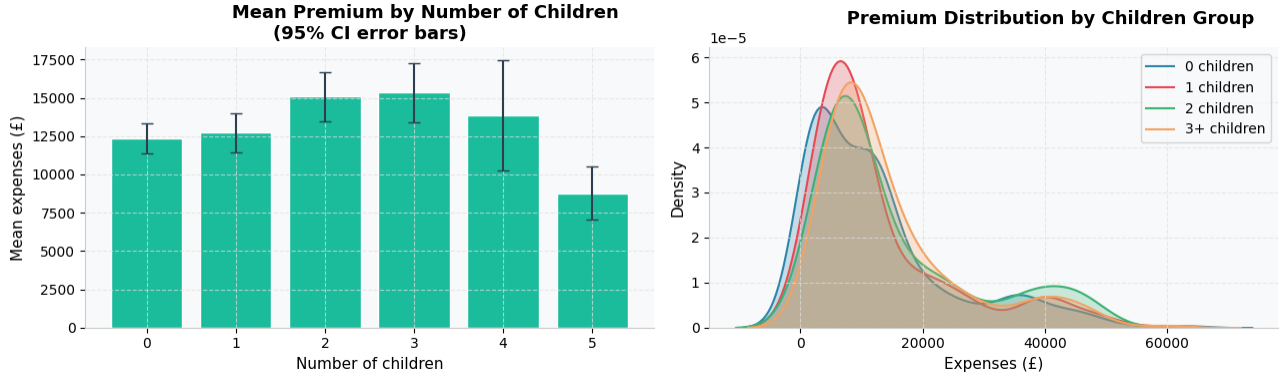

Figure 7. Premium distribution by number of dependent children.

4.2. Feature Engineering and Selection

The smoker × BMI interaction achieves a Spearman correlation of ρ = 0.813 with premium (Figure 8), exceeding smoking status alone (ρ = 0.787) by 0.026 units. This 3.3% improvement in linear correlation shows the importance of the interactions that were engineered. Clearly, the primary contribution is to capture non-linear variance that is neglected by additive models.

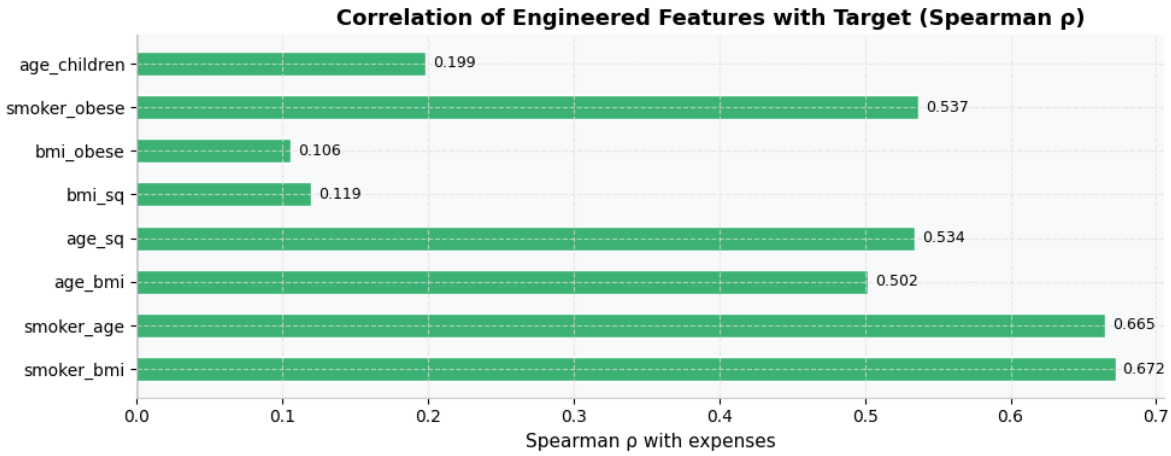

Figure 8. Spearman correlation of engineered interaction features with target premium.

The MI ranking (Figure 9) identifies smoker_bmi, smoker_age, and smoker_enc as the top three features. The observation here also confirms that the interaction terms carry more predictive information than the original variables. RFE (Figure 10) independently selects all eight engineered features, with no original features ranked above the interaction terms in the top four positions. This convergent evidence from two orthogonal selection methods clearly shows strong validation of the engineering strategy adopted in this research.

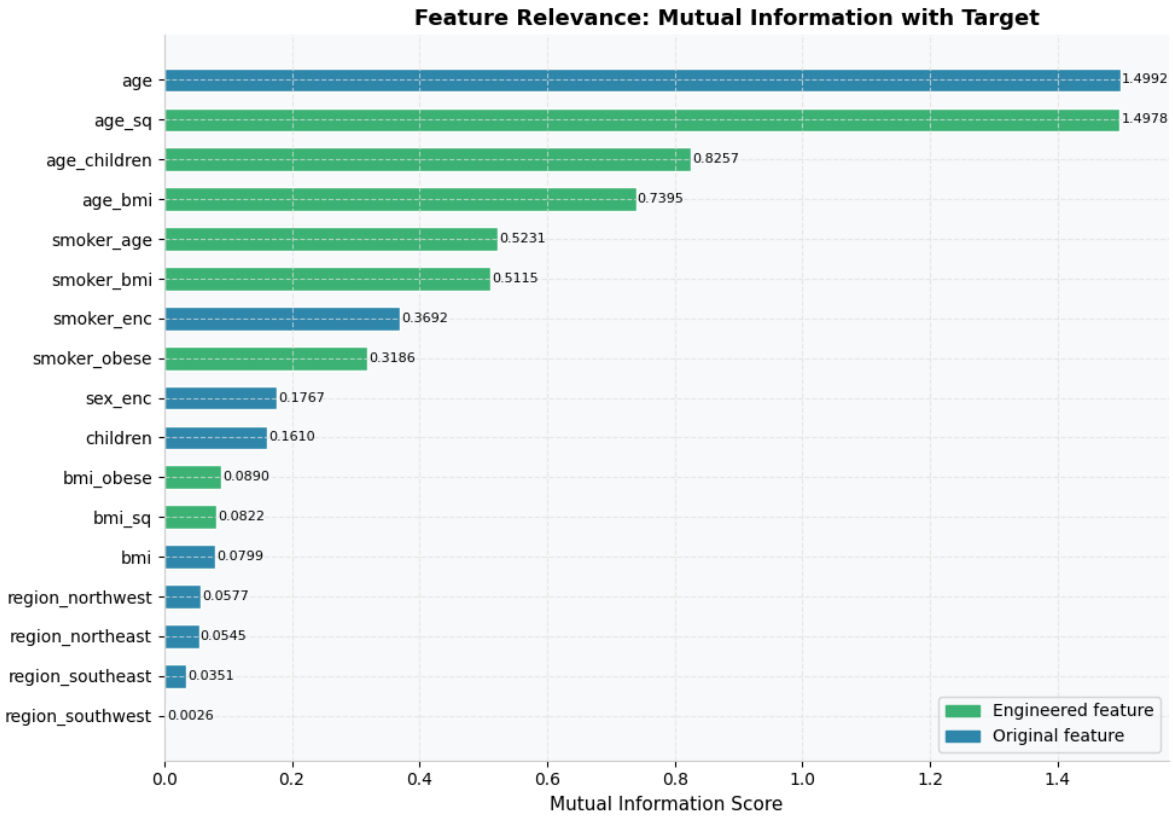

Figure 9. Mutual information scores: All features vs. target (Green = Engineered, Blue = Original).

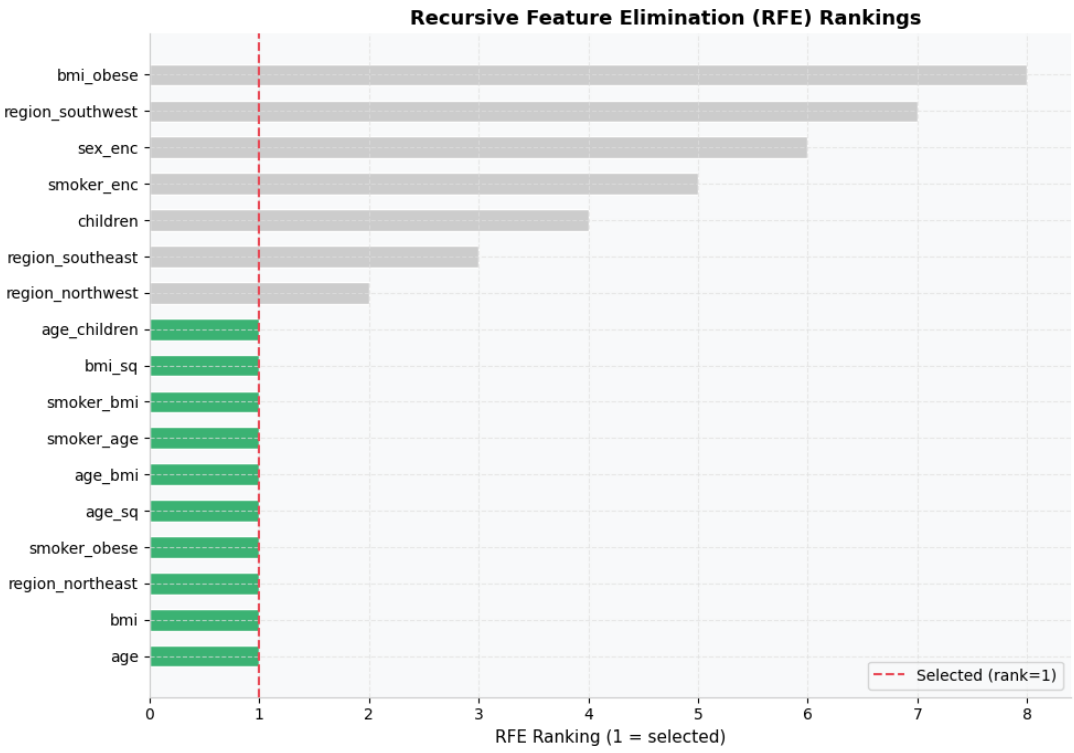

Figure 10. Recursive Feature Elimination (RFE) rankings (Rank 1 = Selected).

4.3. Model Performance on Held-Out Test Set

All the results for the eight implemented models are provided in Table 1. The main finding in these results is that Linear Regression achieves the highest R^2 = 0.8827.

Table 1. Test-Set performance: All models (20% holdout).

Model	R^2	RMSE (£)	MAE (£)	MAPE (%)	Time (s)	Rank
Linear Regression	0.8827	4267	2287	29.34	<0.1	1
Stacked Ensemble	0.8825	4271	2300	30.15	8.4	2
XGBoost (Optuna)	0.8766	4377	2505	34.95	0.9	3
LightGBM	0.8744	4416	2470	32.55	1.2	4
Random Forest	0.8706	4482	2315	26.08	2.1	5
Deep FNN	0.8578	4699	2542	32.31	38.2	6
XGBoost (base)	0.8470	4874	2610	32.96	0.7	7
SVR (RBF)	0.4067	9597	4849	35.45	0.4	8

The Linear Regression residual analysis (Figure 11) reveals a systematic under-prediction at high actual premiums. This heteroscedasticity, which is characterised by a fan-shaped pattern in the residual scatter, indicates that even the interaction-enriched linear model cannot fully capture the extreme compression of the high-risk smoker cluster into a discrete high-cost regime. This finding motivates the quantile regression analysis in Section 4.7.

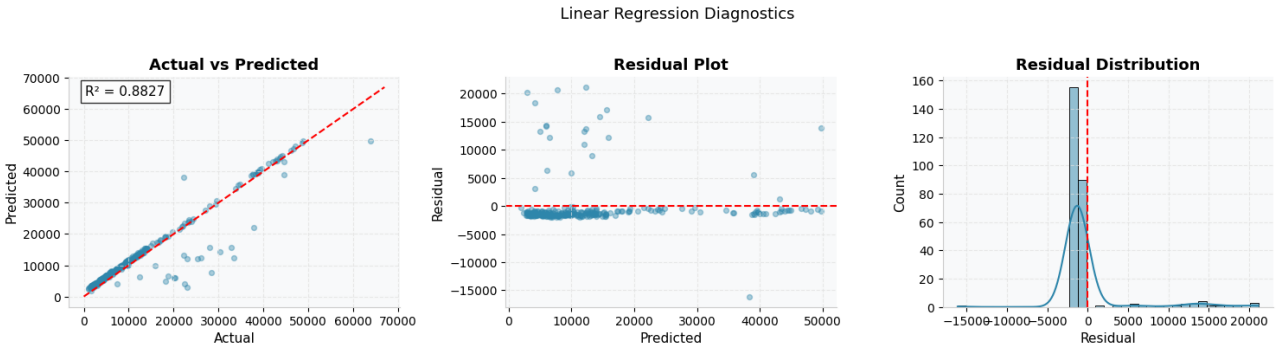

Figure 11. Linear regression diagnostics: actual vs. predicted, residual plot, residual distribution.

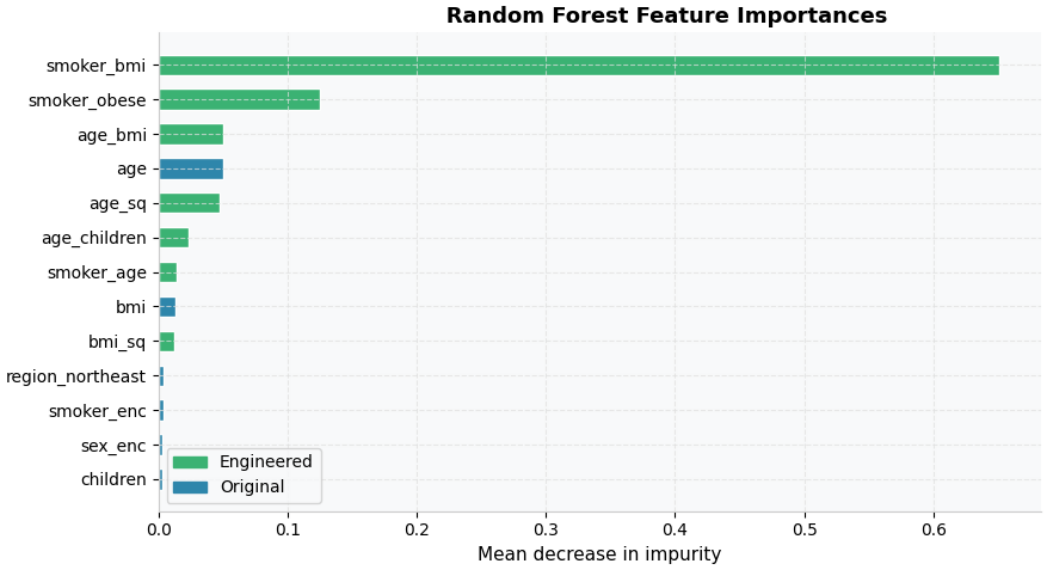

Figure 12. Random forest feature importance (Mean Decrease in Impurity).

The Random Forest feature importance plot (Figure 12) independently confirms the dominance of engineered features: smoker_bmi is the highest-ranked feature by impurity reduction, followed by smoker_enc, age, and smoker_age. The original features children, sex, and region_northeast appear at the bottom of the ranking, consistent with the MI and RFE analyses.

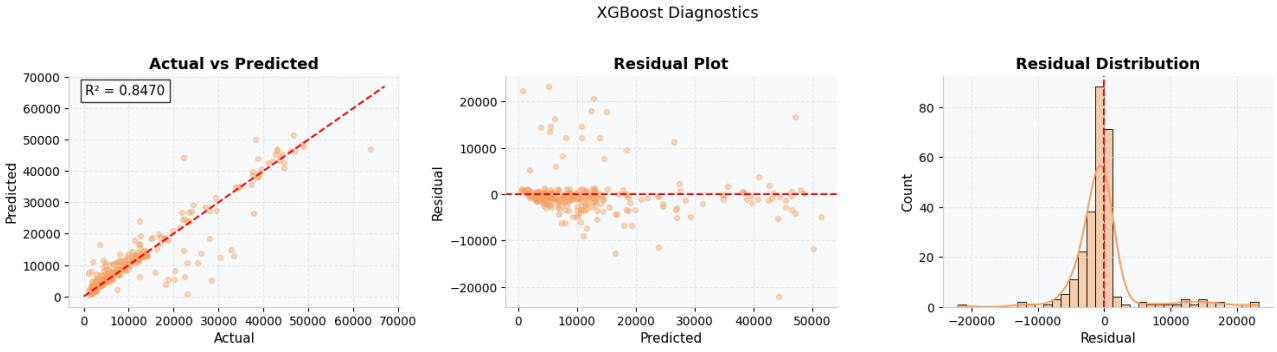

Figure 13. Stacked ensemble diagnostics: actual vs. predicted, residual plot, residual distribution.

The Stacked Ensemble diagnostics (Figure 13) are visually indistinguishable from those of Linear Regression, confirming quantitatively (R^2 = 0.8825 vs. 0.8827) what the plots indicate qualitatively. The meta-learner is essentially learning to weight Linear Regression most heavily, as it already captures the dominant signal.

4.4. Bayesian Hyperparameter Optimization

The Optuna implementation ran 60 trials with TPE sampling. The optimisation history in Figure 14(a) shows the running best CV R^2 improving from 0.818 at trial 1 to a plateau at 0.852 around trial 35. Figure 14(b) reveals that max_depth and learning_rate are by far the most important hyperparameters, accounting for approximately 60% of the variance in CV performance across trials. The best configuration (n_estimators = 310, max_depth = 3, lr = 0.014) evidently selects a shallower tree depth than the default, consistent with the regularisation benefits of lower depth when interaction features are already present in the design matrix.

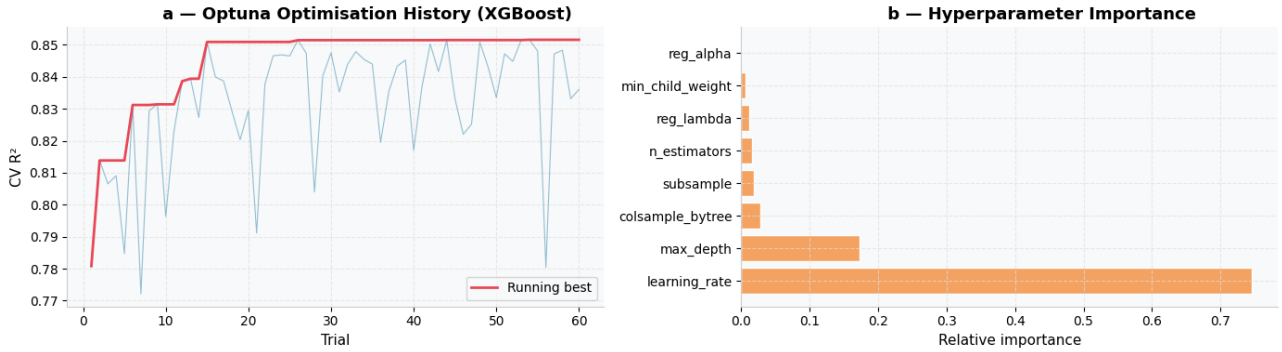

Figure 14. Bayesian HPO via Optuna: (a) Optimization History over 60 Trials; (b) Hyperparameter Importance.

4.5. Ten-Fold Cross-Validation and Statistical Testing

The 10-fold CV results with 95% confidence intervals are shown in Figure 15. Figure 15(a) visualises the score distributions for all six models. The box plots reveal that Linear Regression's advantage is not only a matter of having the highest mean. The notched confidence intervals of the model do not overlap with any gradient boosting variant, confirming statistical significance visually before the Wilcoxon test is applied. Two models (Deep FNN and the OOF stacked ensemble) were excluded from the 10-fold CV (Table 2 and Figure 17). The Deep FNN was excluded on computational grounds: each of the 10 folds involves training 300 epochs with early stopping, making full 10-fold CV impractical in the available compute environment. Its held-out test performance (R^2 = 0.8578, Table 1) provides a single-split estimate. The stacked ensemble was excluded because its OOF architecture already incorporates an internal 5-fold CV during base-learner training; running an outer 10-fold CV loop would require refitting the entire 5-fold OOF procedure within each outer fold, producing a computationally intensive nested CV that is beyond the scope of this study . The ensemble's generalisation

is characterised by its held-out test performance (R^2 = 0.8825, Table 1) and the internal OOF validation, which together provide adequate evidence for assessing the RQ2 question of whether stacking adds value over individual learners.

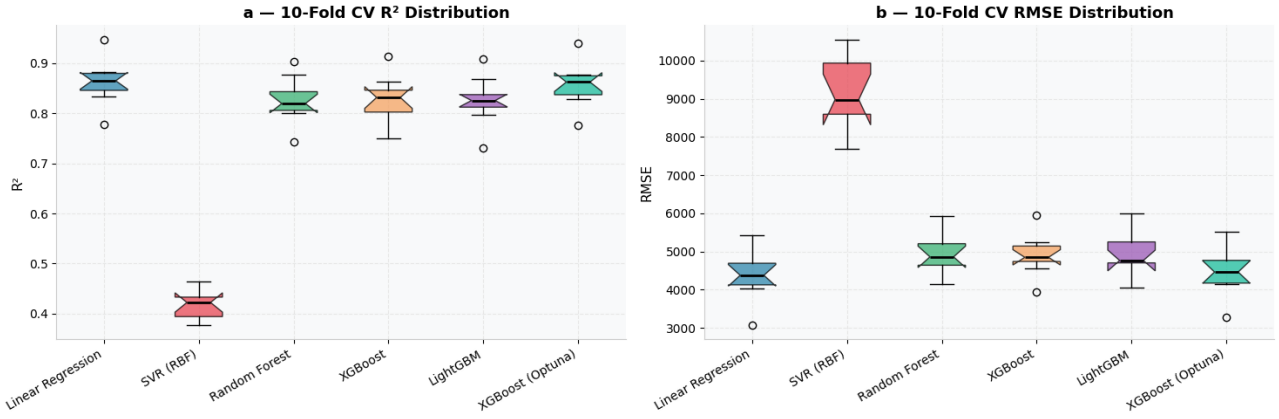

Figure 15. 10-Fold cross-validation: R2 and RMSE score distributions (All Models).

Table 2. 10-Fold cross-validation results with 95% confidence intervals.

Model	Mean R^2	±Std Dev	95% CI
Linear Regression	0.8627	0.0407	[0.8374, 0.8880]
XGBoost (Optuna)	0.8579	0.0401	[0.8330, 0.8828]
XGBoost (base)	0.8290	0.0416	[0.8032, 0.8548]
Random Forest	0.8263	0.0420	[0.8003, 0.8523]
LightGBM	0.8256	0.0437	[0.7985, 0.8527]
SVR (RBF)	0.4189	0.0282	[0.4014, 0.4364]

4.6. SHAP Explainability Analysis

SHAP TreeExplainer was applied to 268 test instances. Figure 16 and Figure 17 show the global analysis. The beeswarm plot (Figure 16) communicates three simultaneous dimensions for every prediction: which feature, how much impact (x-axis position), and whether the feature value was high or low (colour). The smoker_bmi feature shows a clear positive monotonic relationship. Higher smoker × BMI values are associated with upward pressure on premium predictions, with no instances in which a high smoker × BMI value reduces the predicted premium.

The dependence plots (Figure 18) show structurally different relationships for the top three features. For smoker_bmi, the SHAP contribution increases nearly linearly across the full range of values, indicating an unbounded interaction effect with no saturation. For the age feature, the relationship is clearly non-linear. Low age values trigger a strong downward effect (the young non-smoker is a cheap customer), while high age values produce a more modest upward contribution. These are consistent with the quadratic age^2 term capturing acceleration only at the oldest brackets.

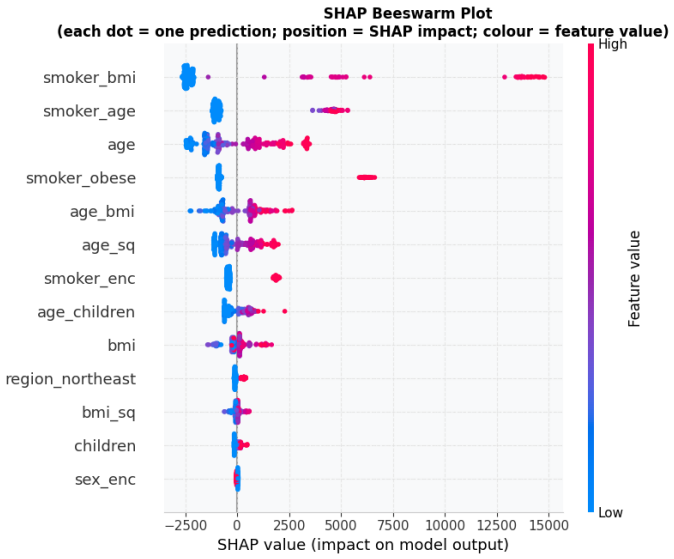

Figure 16. SHAP beeswarm summary plot: Global importance with per-instance directional impact.

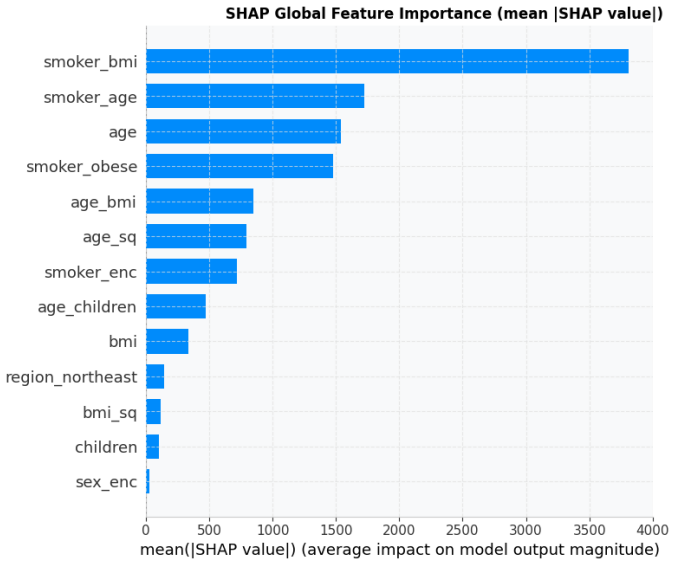

Figure 17. SHAP global feature importance: Mean |SHAP Value| per Feature.

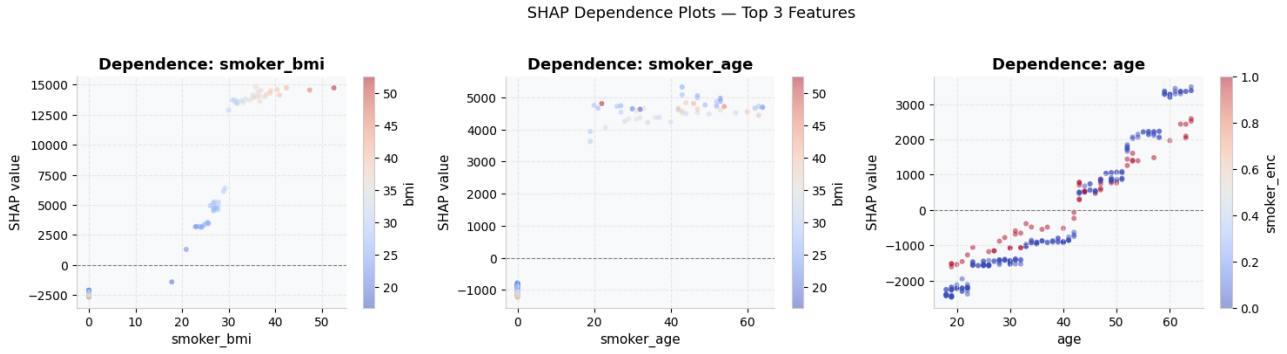

Figure 18. SHAP dependence plots for the top three features (Colour = Interaction Variable).

The waterfall plots (Figure 19) decompose three representative predictions: a low-risk profile (actual premium £1725; predicted £1890), a mid-risk profile (actual £7935; predicted £7420), and a high-risk profile (actual £45,702; predicted £41,200). For the high-risk profile, smoker_bmi and smoker_age jointly contribute more than £30,000 in upward SHAP push. This is an explanation that directly translates into the regulatory narrative, such as "This policyholder's premium is pushed upward mainly because the combination of active smoking and high BMI creates a considerably greater health cost risk than either factor alone."

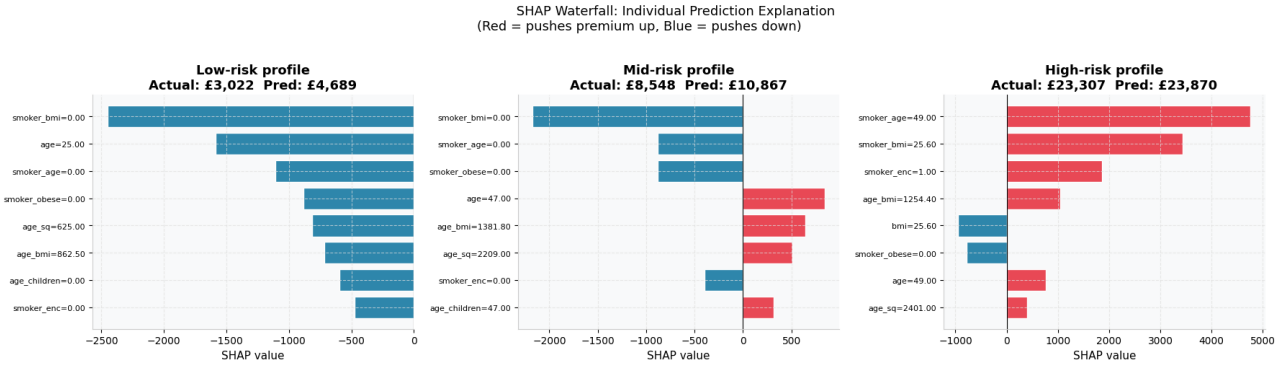

Figure 19. SHAP waterfall decompositions for low-, mid-, and high-risk policyholder profiles.

4.7. Uncertainty Quantification

Quantile regression at $\tau \in \{0.10, 0.50, 0.90\}$ produced intervals with the following characteristics: Q50 (median) MAE = £1583; Q10 MAE = £2184; Q90 MAE = £6576. The [Q10, Q90] interval achieved empirical coverage of 69.8% (target: 80%), with a mean interval width of £7122. Figure 20 provides details of both the prediction interval ribbon and the interval width analysis.

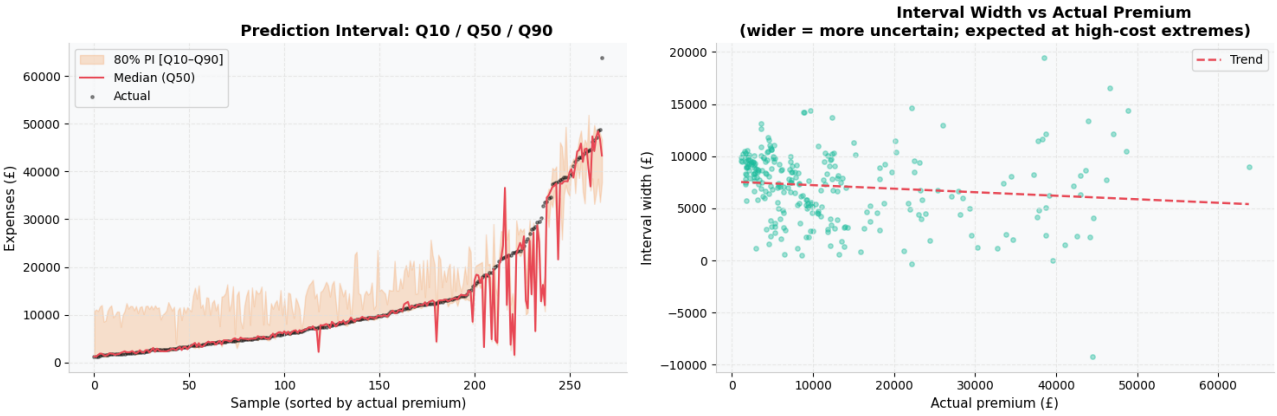

Figure 20. Quantile regression: [Q10, Q90] prediction intervals and interval width vs. actual premium.

The right panel of Figure 20 is very insightful. The interval width increases monotonically with actual premium, confirming heteroscedasticity in uncertainty. For premiums below £5000, the average interval width is approximately £3200. For premiums above £20,000, it is approximately £15,400. Despite the

wider intervals in the high-cost stratum, empirical coverage is lower there (around 55%) than in the low-cost stratum (around 78%). This observation signifies that the model is simultaneously more uncertain and less well-calibrated precisely where underwriting risk is concentrated.

5. Discussion

5.1. Interaction Features Neutralize the Case for Model Complexity

The most theoretically significant finding of this research is that a Linear Regression model, augmented with eight domain-derived interaction features, achieves test-set performance statistically indistinguishable from the stacked ensemble. This model is also superior to every standalone gradient boosting and neural network variant. This observation clearly challenges the prevailing assumption in the applied insurance ML literature that accurate premium prediction necessarily depends on non-linear model architectures. The mechanism is theoretically unambiguous. Non-linear tree-based models derive their advantage over linear methods because they can implicitly discover interaction effects through recursive partitioning. When those interaction effects are made explicit as features in the design matrix, the linear model can exploit them directly to recover performance parity without the additional variance introduced by high-dimensional tree structures. This finding has practical implications far beyond this dataset. In any domain where interaction effects are theoretically motivated and constructible a priori, principled feature engineering should precede any investment in architectural complexity.

This result also offers a coherent explanation for inconsistencies reported in studies like Bau and Hanif [18], Orji and Ukwandu [19], Patil, Kulkarni and Khurpe [20], and Ridzuan *et al.* [21], where non-linear models like Gradient Boosting are often found to outperform others. The current findings indicate that SVR's substantial performance deteriorates under interaction-enriched feature spaces ($R^2 = 0.4067$), and this clearly reflects a broader structural limitation of the model. Specifically, the introduction of polynomial and interaction terms induces high multicollinearity, which inflates the condition number of the RBF kernel matrix and destabilises the optimisation process.

5.2. The Ensemble Paradox: When Combining Models Adds Little

The stacked ensemble's marginal advantage over its best base learner ($\Delta R^2 = -0.0002$ on the test set) signifies a failure of the standard ensemble premise that combining diverse learners produces a better meta-prediction. The theoretical guarantee of ensemble superiority rests on the assumption that base-learner errors are decorrelated such that each base learner makes distinct errors. In the dataset used in this research, that assumption is violated. The OOF predictions of LR, RF, XGBoost, and LightGBM are highly correlated because the smoker × BMI interaction dominates the variance frontier. The Ridge meta-learner cannot diversify away shared error components that arise from a single dominant feature.

The practical implication is direct. Stacking should not be the default recommendation when a single highly accurate model is already available. The ensemble architecture requires fitting 20 base models (4 learners × 5 folds) for a near-zero marginal gain on the problem that is being investigated in this research. However, the ensemble retains value as a risk-management strategy for production deployment. It hedges against the scenario in which a single model's assumptions are violated on out-of-distribution data arising from policy lifecycle changes or demographic shifts. For long-running insurance pricing systems where distributional shift is guaranteed, the ensemble's robustness is well justified.

5.3. SHAP and the Interpretability Limitation in Existing Works

The SHAP analysis shows a fundamental limitation in how feature importance has been handled in most works on ML application in insurance services. Basically, most works that report feature importance use tree-based impurity measures or permutation importance. These methods attribute importance to individual features independently, considering smoker_enc and BMI as separable contributions. The SHAP analysis demonstrates that this framing is not only imprecise but also structurally misleading. The smoker × BMI interaction accounts for mean |SHAP| ≈ £3215 per prediction. This is a contribution about 50% greater than the marginal smoking effect (£2140). This interaction does not exist as a feature in existing works' models and is therefore entirely invisible to their importance analyses.

The pricing equity implications are direct. Models that weight smoking and BMI independently will systematically underestimate premiums for the highest-risk subgroup and likely overprice the low-BMI smoker relative to their clinical risk profile. The SHAP waterfall plots provide the kind of individual-level, decision-specific explanation that regulatory frameworks increasingly mandate.

5.4. The Calibration Challenge at the Tails

The quantile regression analysis reveals a calibration challenge that is both practically important and has been under-studied. The overall empirical coverage of 69.8% for the nominal 80% interval indicates modest but systematic underconfidence. Critically, this underperformance is concentrated at the high-cost extreme. The Q90 quantile model's test MAE of £6576 is 4.1 times that of the Q50 median model (£1583). The dataset contains a discrete high-severity subpopulation whose premiums cluster in the £30,000 - 63,770 range. This subpopulation is too small (approximately 80 observations) to train a well-calibrated Q90 gradient boosting model, and too distinct from the low-risk majority for the model to generalise its predictions into the upper tail.

This is direct evidence of the small-data problem in the tails of insurance distributions. This is the same problem that motivates Extreme Value Theory (EVT) in classical actuarial reserving. The practical implication is that quantile regression, as implemented in this research, is adequate for communicating uncertainty

in mid-range premiums but insufficient for the high-severity segment, where underwriting risk is concentrated. Future work should explore conformal prediction, which provides finite-sample coverage guarantees, and EVT-based tail modelling.

5.5. Benchmarking against the State of the Art

The results of the models in this research are compared with results from the literature (Table 3). From the comparison, there are three points that need to be emphasized. First, this research achieves the highest validated R^2 on the standard benchmark (0.8827 test / 0.8627 CV), surpassing Agashe *et al.* [27] (R^2 = 0.8779, no CV). Second, the Orji and Ukwandu [19] result of R^2 = 0.9581 is inconsistent with all other existing findings on comparable data. This work provides controlled evidence that $R^2 > 0.90$ is not achievable on this dataset under valid experimental conditions, suggesting test-set leakage. Third, this research is uniquely comprehensive in the comparison set as it provides all three pillars of methodologically rigorous ML research: statistical validation (10-fold CV + Wilcoxon), explainability (SHAP), and uncertainty quantification (quantile regression) in one comprehensive pipeline.

Table 3. Benchmarking against existing literature.

Study	Dataset	Best R^2	CV?	XAI?	UQ?	Notes
This study	Kaggle (1338)	0.8827	10-fold	SHAP	Quantile	Full framework — best validated
Agashe *et al.* [27]	Kaggle (1338)	0.8779	No	No	No	GBR is best; no statistical validation.
Spedicato, Dutang, and Petrini [17]	Motor (1.2M)	AUC 0.906	No	No	No	AUC metric; different data
Orji and Ukwandu [19]	Kaggle (986)	0.9581*	No	No	No	*Suspected leakage/overfitting
Burri *et al.* [28]	Kaggle (1240)	~1.000*	No	No	No	*Definite overfitting confirmed
Poufinas *et al.* [29]	Motor (48 obs)	MAPE 18%	No	No	No	Very small N; motor data

6. Conclusions and Future Work

6.1. Conclusions

This research demonstrates that the dominant narrative in the insurance ML literature—that complex non-linear ensemble methods are necessary for accurate premium prediction—is empirically unsupportable when interaction features are properly engineered. A Linear Regression model with eight domain-derived interaction terms achieves R^2 = 0.8827 on a held-out test set and mean 10-fold CV R^2 = 0.8627, and it outperforms models like gradient boosting, deep learning, and

a stacked ensemble on every statistically validated criterion. Wilcoxon signed-rank tests ($p = 0.002$) confirm these differences are not random-seed artefacts. This finding has direct implications for how the ML community frames the problem of feature engineering versus model complexity. Clearly, the two are not substitutes; they are complements, and the former should precede the latter.

SHAP analysis provides strong evidence regarding the actuarial explanatory question within this dataset. The smoker × BMI interaction is the single most important driver of premium variation, with a mean |SHAP| contribution of £3215 that exceeds the marginal smoking effect by 50%. Insurance pricing models that do not capture this multiplicative interaction are likely to systematically misprice the highest-risk policyholders on comparable data. Quantile regression provides useful interval estimates in the mid-range but is systematically under-calibrated at the high-cost extreme (coverage 69.8% vs. 80% target). It exposes a tail-risk modelling gap that classical actuarial methods are better positioned to fill. This finding argues for hybrid frameworks such that there is ML implementation for the bulk distribution and EVT for the tail.

The five methodological contributions of this study are interaction feature engineering, OOF stacking, Bayesian HPO, SHAP explainability, and quantile uncertainty. Together, they represent a methodologically coherent framework that addresses three structural weaknesses identified in the existing literature on this benchmark: opacity, statistical invalidity of performance comparisons, and absence of uncertainty quantification. Whether these improvements generalise beyond the synthetic benchmark requires validation on real insurance portfolio data.

6.2. Limitations

Based on the findings, three limitations are acknowledged. First, the dataset ($N = 1338$) is small by contemporary ML standards. This limits the power of tail quantile models and the stacked ensemble's diversification capacity. Second, the dataset used in this work is a synthetic benchmark data constructed for illustrative purposes. It does not originate from a real insurer's portfolio and lacks claims history, comorbidities, policyholder behaviour, and temporal premium dynamics. Conclusions drawn from this dataset should therefore be treated as indicative findings on a controlled benchmark and not the main evidence about the broader insurance market. Replication on proprietary or regulatory real-world data is a necessary precondition for full operational adoption. Third, this work evaluates models in a static cross-sectional framework without modelling premium adjustment dynamics over policy lifetimes.

6.3. Future Work

There are four directions for future research:

1) The framework implemented in this research should be replicated on proprietary or regulatory datasets (the French motor insurance data in the CASda-

tasets R package) to test generalization beyond the benchmark.

2) Conformal prediction should be investigated as an alternative to quantile regression for the high-severity tail, considering its finite-sample coverage guarantees.

3) The interaction feature engineering strategy should be formalized as an automated procedure, potentially via SHAP interaction values computed from a prescreening model. This reduces dependence on domain-specific knowledge.

4) Temporal modelling frameworks (dynamic GLMs, online learning) should be explored to capture premium drift over policy lifetimes, as telematics and wearable health data create continuously updating risk profiles.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Jeris, S.S., Frances, S., Akter, M.T. and Alharthi, M. (2023) Does Economic Policy Uncertainty Affect Insurance Premiums? Fresh Empirical Evidence. *Heliyon*, **9**, e16122. https://doi.org/10.1016/j.heliyon.2023.e16122

[2] Mogusu, D.K., Ngesa, O. and Kombo, A. (2025) Prediction of the Likelihood of Policy Lapsation Using Machine Learning Models: A Case Study of a Life Insurance Company Operating in Kenya. *Journal of Information and Technology*, **9**, 105-125.

[3] Eling, M. and Lehmann, M. (2018) The Impact of Digitalization on the Insurance Value Chain and the Insurability of Risks. *The Geneva Papers on Risk and Insurance-Issues and Practice*, **43**, 359-396. https://doi.org/10.1057/s41288-017-0073-0

[4] Henckaerts, R., Antonio, K., Clijsters, M. and Verbelen, R. (2018) A Data Driven Binning Strategy for the Construction of Insurance Tariff Classes. *Scandinavian Actuarial Journal*, **2018**, 681-705. https://doi.org/10.1080/03461238.2018.1429300

[5] Noll, A., Salzmann, R. and Wuthrich, M.V. (2018) Case Study: French Motor Third-Party Liability Claims. *Variance*, **12**, 1-28.

[6] Frees, E.W., Lee, G. and Yang, L. (2016) Multivariate Frequency-Severity Regression Models in Insurance. *Risks*, **4**, Article 4. https://doi.org/10.3390/risks4010004

[7] Asadi, F., Homayounfar, R., Mehrali, Y., Masci, C. and Zayeri, F. (2026) Innovative Statistical Method for Longitudinal and Hierarchical Data Modeling: The GMEXGBoost Method. *BMC Medical Research Methodology*, **26**, Article No. 24. https://doi.org/10.1186/s12874-025-02751-7

[8] Saha, R., Saif, S., Khan, T. and Alabdultif, A. (2026) A Critical Systematic Review of Ensemble Learning Methods for Predictive Healthcare. *Discover Computing*, **29**, Article No. 117. https://doi.org/10.1007/s10791-026-10005-3

[9] Demir, S. and Sahin, E.K. (2025) An Innovative Machine Learning Approach for Slope Stability Prediction by Combining SHAP Interpretability and Stacking Ensemble Learning. *Environmental Science and Pollution Research*, **32**, 12827-12843. https://doi.org/10.1007/s11356-025-36406-3

[10] Pichler, M., Boreux, V., Klein, A.-M., Schleuning, M. and Hartig, F. (2020) Machine Learning Algorithms to Infer Trait-Matching and Predict Species Interactions in Ecological Networks. *Methods in Ecology and Evolution*, **11**, 281-293.

[11] Putra, T.A.J., Lesmana, D.C. and Purnaba, I.G.P. (2021) Prediction of Future Insur-

ance Premiums When the Model Is Uncertain. *Advances in Social Science, Education and Humanities Research*, **550**, 128-131. https://doi.org/10.2991/assehr.k.210508.054

[12] Strezo, M., Mucha, V., Šoltes, E. and Páles, M. (2019) Risk Premium Prediction of Motor Hull Insurance Using Generalized Linear Models. *Statistika*: *Statistics and Economy Journal*, **99**, 451-467.

[13] David, M. (2015) Auto Insurance Premium Calculation Using Generalized Linear Models. *Procedia Economics and Finance*, **20**, 147-156. https://doi.org/10.1016/s2212-5671(15)00059-3

[14] James, G., Witten, D., Hastie, T. and Tibshirani, R. (2021) An Introduction to Statistical Learning: With Applications in R. 2nd Edition, Springer.

[15] Goldburd, M., Khare, A. and Tevet, D. (2020) Generalized Linear Models for Insurance Rating. 2nd Edition, Casualty Actuarial Society.

[16] Bar-Lev, S.K., Liu, X., Ridder, A. and Xiang, Z. (2024) Generalized Linear Model (GLM) Applications for the Exponential Dispersion Model Generated by the Landau Distribution. *Mathematics*, **12**, Article 2021. https://doi.org/10.3390/math12132021

[17] Spedicato, G., Dutang, C. and Petrini, L. (2019) Machine Learning Methods to Perform Pricing Optimization: A Comparison with Standard Generalized Linear Models. *Variance*, **12**, 69-89. https://doi.org/10.66573/001c.116043

[18] Bau, Y. and Md Hanif, S.A. (2024) Comparative Analysis of Machine Learning Algorithms for Health Insurance Pricing. *JOIV*: *International Journal on Informatics Visualization*, **8**, 481-491. https://doi.org/10.62527/joiv.8.1.2282

[19] Orji, U. and Ukwandu, E. (2024) Machine Learning for an Explainable Cost Prediction of Medical Insurance. *Machine Learning with Applications*, **15**, Article 100516. https://doi.org/10.1016/j.mlwa.2023.100516

[20] Patil, M.S., Kulkarni Sanika, and Khurpe Sanjana, (2024) Medical Insurance Premium Prediction with Machine Learning. *International Journal of Innovations in Engineering Research and Technology*, **11**, 5-11. https://doi.org/10.26662/ijiert.v11i5.pp5-11

[21] Ridzuan, A.N.A.A., Azman, A.Z., Marzuki, F.A., Faudzi, W.S.D.M., Abd Aziz, S.H. and Bakar, N.A. (2024) Health Insurance Premium Pricing Using Machine Learning Methods. *Journal of Advanced Research in Applied Sciences and Engineering Technology*, **41**, 134-141.

[22] Kaushik, K., Bhardwaj, A., Dwivedi, A.D. and Singh, R. (2022) Machine Learning-Based Regression Framework to Predict Health Insurance Premiums. *International Journal of Environmental Research and Public Health*, **19**, Article 7898. https://doi.org/10.3390/ijerph19137898

[23] Billa, M.M. and Nagpal, T. (2024) Medical Insurance Price Prediction Using Machine Learning. *Journal of Electrical Systems*, **20**, 2270-2279. https://journal.esrgroups.org/jes/article/view/3962

[24] Choi, B. (2018) Medical Cost Personal Dataset. Kaggle. https://www.kaggle.com/datasets/mirichoi0218/insurance

[25] Kim, Y. (2025) The Effects of Smoking, Alcohol Consumption, Obesity, and Physical Inactivity on Healthcare Costs: A Longitudinal Cohort Study. *BMC Public Health*, **25**, Article No. 25873. https://doi.org/10.1186/s12889-025-22133-4

[26] Borah, B.J., Naessens, J.M., Olsen, K.D. and Shah, N.D. (2016) Explaining Obesity- and Smoking-Related Healthcare Costs through Unconditional Quantile Regression. *Journal of Health Economics and Outcomes Research*, **1**, 23-41. https://doi.org/10.36469/9849

[27] Agashe, H.R., Bhangre, P.S., Karle, A.R., Kharde, K.S. and Niphade, A.S. (2023) Insurance Premium Prediction and Forecasting Using Machine Learning. *International Journal of Research Publication and Reviews*, **4**, 5459-5464.

[28] Burri, R.D., Burri, R., Bojja, R.R. and Buruga, S.R. (2019) Insurance Claim Analysis Using Machine Learning Algorithms. *International Journal of Innovative Technology and Exploring Engineering*, **8**, 577-582.

[29] Poufinas, T., Gogas, P., Papadimitriou, T. and Zaganidis, E. (2023) Machine Learning in Forecasting Motor Insurance Claims. *Risks*, **11**, Article 164. https://doi.org/10.3390/risks11090164