

# Explainable Auto Insurance Risk Prediction via Exposure-Aware Feature Interaction and Cross-Fitted Residual Learning

Isabella Roberts \*, Gabriel Brooks, Grace Morris, Owen Li

Robert H. Smith School of Business, University of Maryland, College Park, USA

\*Corresponding Author: Isabella Roberts. c5030403@umd.edu

## Abstract

**Accurate prediction of motor insurance claim frequency must jointly address nonlinear interactions, heterogeneous exposure periods and transparent risk attribution. Building on the Transformer-SHAP framework of Sun et al. for interpretable auto-insurance risk evaluation, this paper develops exposure-aware interaction and cross-fitted residual learning (EICRL). A regularized LightGBM model estimates annual claim frequency under exposure weights, and an independently fitted Poisson offset model corrects out-of-fold log-frequency errors with sparse, interpretable interaction terms; validation-controlled shrinkage and a positive exponential link yield stable, nonnegative expected claim counts. On a stochastic motor-portfolio simulator with exposure heterogeneity, rare counts and nonlinear actuarial interactions, EICRL lowers held-out Poisson deviance relative to LightGBM in all 15 independent portfolios of 48,000 policies (0.39247 versus 0.39461; paired 95% interval of the reduction [0.00150, 0.00278]). Ablations separate the contributions of interaction modeling and cross-fitting, and whole-pipeline permutation analysis identifies bonus-malus and driver age as the dominant predictors. EICRL thus provides an exposure-consistent and explainable frequency model ready for validation on observed portfolios such as freMTPL2freq.**

## Keywords

**Motor insurance, claim frequency, feature interactions, residual learning, Poisson deviance, explainable machine learning, exposure offset.**

## 1. Introduction

Motor insurance pricing begins with a conditional expectation: how many claims are anticipated over an observed coverage interval? Actuarial frequency modeling conventionally uses a generalized linear model (GLM) with a Poisson likelihood and a log exposure offset [1–3]. Such models yield tractable relativities but can underrepresent nonlinear effects and feature interactions. Modern gradient-boosted trees extend the functional class [4–8], and attention-based architectures attempt to capture interactions across heterogeneous policy attributes [10]. Their deployment nevertheless requires a precise distinction between predicted annual risk and expected claims during the insured exposure period.

Sun et al. [9] showed that a Transformer-based model with SHAP explanations can evaluate auto-insurance risk while capturing interactions among heterogeneous risk features and keeping each prediction attributable. Two design principles follow from their work: interactions among heterogeneous rating factors should be modeled explicitly, and explanations should target the complete prediction function. EICRL builds on both principles with an implementation tailored to exposure-adjusted claim counts: a LightGBM annual-rate predictor followed by a cross-fitted, structured interaction correction on the log-rate scale. The correction is interpretable through named feature terms, while the complete model is explained by full-pipeline permutation reliance.

The contribution is threefold. First, EICRL jointly represents flexible boosted-tree predictions and named pairwise actuarial interactions, fitted under an exact count/exposure formulation. Second, out-of-fold base predictions train the correction stage, mitigating the optimistic residual patterns induced by in-sample fitting; a disjoint validation set selects correction strength. Third, explanation is evaluated at the complete-pipeline level using permutation-based changes in held-out deviance, complemented by risk-decile calibration. A seeded stochastic benchmark separates empirically measured improvements from assumptions about future real-world performance.

## 2. Related Work

### 2.1. Actuarial Count Models

The GLM framework of Nelder and Wedderburn [1] and McCullagh and Nelder [2] motivates Poisson frequencies, canonical links, and explicit exposure treatment. Wüthrich and Merz [3] connect these foundations to actuarial machine learning. With claim count  $Y_i$ , risk attributes  $x_i$  and fractional policy-year exposure  $e_i$ , an offset model assumes  $E(Y_i | x_i, e_i) = e_i \lambda(x_i)$ . The distinction matters because frequency  $\lambda$  refers to one exposure-year while the realized policy claim count has mean  $e_i \lambda$ . A model trained directly on raw counts without accounting for  $e_i$  can confound underwriting risk with contract duration.

### 2.2. Interaction Learning And Hybrid Ensembles

Gradient boosting [4] and random forests [5] capture nonlinear structure without prespecified basis functions. XGBoost [6], LightGBM [7], and CatBoost [8] have made tree ensembles strong baselines for structured actuarial records. Transformers [10] and residual networks [11] motivated architectures combining global interactions with flexible correction paths. More recent tabular neural architectures such as TabNet [18], FT-Transformer [19], TabTransformer [20] and SAINT [21], alongside systematic tabular benchmarks [22], support comparing neural and non-neural inductive biases rather than presuming a universal winner.

### 2.3. Explainability for Insurance Decisions

SHAP [12] attributes predictions using cooperative-game contributions; TreeSHAP [13] provides efficient tree-specific explanations. LIME [14], accumulated local effects [15], and permutation reliance [16] provide complementary viewpoints. Insurance decisions additionally require the explanation target to be clearly specified: an explanation of the annual rate need not be identical to an explanation of the exposure-scaled expected count. This work ranks input contributions by their perturbation effect on the same held-out Poisson objective used to choose risk models. Such predictive reliance is not a causal effect; correlated predictors can share importance [16, 17]. The FairPromote framework also illustrates how SHAP-based interpretation can be evaluated alongside fairness-oriented objectives in another consequential prediction setting [25].

## 3. Problem Formulation And Method

### 3.1. Exposure-aware Frequency Prediction

Let  $D = \{(x_i, e_i, y_i)\}_{i=1}^N$  denote a portfolio, with attributes  $x_i$ , coverage exposure  $e_i$  in  $(0, 1]$ , and claim counts  $y_i$  in the nonnegative integers. We estimate an annual rate  $\lambda_i > 0$  and expected claims  $\mu_i = e_i \lambda_i$ , using the Poisson negative log-likelihood up to a  $y$ -dependent constant:

$$L(\lambda) = \sum_i [e_i \lambda(x_i) - y_i \log(e_i \lambda(x_i))] \quad (1)$$

Evaluation uses mean Poisson deviance over individual policy counts,  $D = 2N^{-1} \sum_i \{y_i \log[y_i/\mu_i] - (y_i - \mu_i)\}$ , where the log term is defined to be zero at  $y_i = 0$ . This is the mean per-policy count deviance, which differs in scale from exposure-normalized aggregate measures reported elsewhere.

### 3.2. Stage I: Exposure-Weighted Tree Frequency

A low-capacity gradient-boosting estimator  $g_\theta(x)$  fits annualized frequencies  $y_i/e_i$  with sample weight  $e_i$  and a Poisson objective. The corresponding loss is proportional to  $\sum_i [e_i \lambda_i - y_i \log \lambda_i]$ , which differs from (1) only by terms that do not depend on model parameters. We use  $\lambda^{(0)}(x) = \max(g_\theta(x), \epsilon)$ , with  $\epsilon = 10^{-8}$ . A limited tree depth and leaf count preserve residual structure for an interpretable correction model while controlling inference cost.

### 3.3. Stage II: Cross-Fitted Interaction Residuals

Our second stage adapts the explicit interaction-learning principle of Sun et al. [9] to exposure-adjusted claim counts. Where their Transformer learns interactions among risk features through attention, EICRL encodes them as named nonlinear and pairwise terms in a multiplicative log-rate correction to the tree-based estimate, with exposure incorporated into the Poisson objective, so that each learned interaction remains an explicit relative. Partition the training subset into  $K=3$  folds. Fit the stage-I learner on  $K-1$  folds and predict the held-out fold, producing cross-fitted estimates  $\lambda_i^{(-k(i))}$ . We fit a regularized correction coefficient vector  $\beta$  using these predictions as fixed offsets:

$$\beta^* = \operatorname{argmin}_\beta \sum_i \{e_i \lambda_i^{(-k)} \exp(z_i^T \beta) - y_i [\log(e_i \lambda_i^{(-k)}) + z_i^T \beta]\} + (\rho/2) \|\beta - \beta_0\|^2 \quad (2)$$

Here  $z_i$  consists of an intercept, normalized marginal effects, and seven predefined nonlinear interaction terms. The construction includes driver age  $\times$  bonus-malus, vehicle age  $\times$  power, fuel  $\times$  density, a sinusoidal density  $\times$  bonus-malus term, urban area  $\times$  power, driver age  $\times$  power, and vehicle age  $\times$  bonus-malus. Additional single-feature terms represent squared age and squared positive bonus-malus. Coefficients are computed by L-BFGS with analytic gradients,  $\rho=12$ , and an unpenalized intercept. The explicit basis is intentionally compact: interactions remain inspectable, unlike opaque all-pairs feature crosses.

### 3.4. Shrinkage, Positivity And Inference

A fresh base model is then fitted on the full training partition. On a distinct validation set,  $\gamma$  is chosen from  $\{0, 0.25, 0.5, 0.75, 1.0\}$  to minimize the same count deviance. Final predictions are:

$$\hat{\lambda}_i = \lambda^{(0)}(x_i) \exp\{\gamma \cdot \operatorname{clip}(z_i^T \beta^*, -1.5, 1.5)\}, \quad \hat{\mu}_i = e_i \hat{\lambda}_i \quad (3)$$

The exponential construction enforces positivity without truncating negative predicted claim counts after the fact. Cross-fitting enters the estimation of  $\beta$  but does not require maintaining  $K$  models for deployment. Prediction uses only the final boosted tree, the fixed interaction basis,  $\beta^*$ , and  $\gamma$ . Unlike a raw-count residual addition, the multiplicative form has a natural relative-risk interpretation: for an isolated interaction change  $\Delta z_j$ , its stage-II rate multiplier is  $\exp(\gamma \beta_j \Delta z_j)$ , subject to clipping.

### 3.5. Interpretation Protocol

The hybrid retains a transparent residual equation, but prediction explanations must cover both stages. We therefore permute each raw field in the test partition, recompute the tree and correction outputs end-to-end, and measure  $\Delta D_j = D_{\text{perm},j} - D_{\text{original}}$ . Positive  $\Delta D_j$  indicates predictive reliance on the field under that perturbation. We also order policies by  $\hat{\lambda}$  to form ten

deciles and compare predicted and observed claim totals in each. This establishes model-level global reliance and operational calibration, not a causal allocation of financial responsibility. SHAP, which Sun et al. [9] use to explain their Transformer risk model, remains a well-established complementary route [12, 13]; permutation reliance is used here because it directly measures the full two-stage model.

## 4. Experimental Design

### 4.1. Reference Dataset and Synthetic Benchmark Design

The freMTPL2freq French Motor Third-Party Liability portfolio is documented in the CASdatasets project and OpenML [23, 24]; it comprises 677,991 policies in the CASdatasets documentation (OpenML lists 678,013 dataset instances) with claim counts and exposures. This public portfolio defines the frequency-prediction task and the exposure-sensitive metrics used here, whereas all reported experiments use explicitly generated synthetic policies whose interaction structure is known.

### 4.2. Auditable Synthetic Portfolio

For each seed, we draw 48,000 independent policies. Driver age is uniform on {18, ..., 83}, vehicle age on {0, ..., 23}, and power on {4, ..., 13}. Bonus-malus combines a driver-age effect, Gaussian variability, and occasional 25-point malus shifts and is clipped to [50, 180]. A continuous density variable generates an urban area group; binary fuel, 12 regional groups, and nine vehicle-brand labels provide heterogeneous covariates. Exposure follows a Beta (2.3, 1.2) distribution clipped to [0.05, 1].

The true log annual rate is  $\eta = -2.38 + 0.22b + 0.12\max(b, 0)^2 + 0.23a^2 + 0.13v + 0.09p + 0.06d + 0.28ba + 0.23vp - 0.18(f - 0.5)d + 0.16\sin(2.2d)b + 0.09\mathbb{1}(\text{area} \geq 4)p + 0.07\mathbb{1}(\text{region} = 3) - 0.06\mathbb{1}(\text{brand} = 2)$ , where (a, v, p, b, d) are standardized age, vehicle age, power, bonus-malus, and density. We generate  $Y \sim \text{Poisson}(e \cdot \exp(\text{clip}(\eta, -5, 1.7)))$ . Thus the underlying risk includes specified nonlinear and interacting features; the benchmark deliberately evaluates whether a hybrid recovers such mechanisms.

### 4.3. Protocol and Comparison Methods

Fifteen fixed simulation seeds (17, 29, 41, 53, 67, 79, 83, 97, 101, 113, 127, 139, 149, 163, 179) produce independent portfolios, each split 65% training (31,200), 15% validation (7,200), and 20% held-out test (9,600). The principal baselines are an exposure-weighted regularized Poisson GLM and a depth-three LightGBM frequency model. Four controlled variants isolate a correction learned from in-sample predictions, additive-only correction, cross-fitted correction without pairwise interactions, and the full EICRL. All methods are evaluated only on the fixed seed-specific holdout; model selection is confined to the validation partition. Random seeds and inference settings are fixed for every run.

LightGBM uses 115 estimators, seven leaves, learning rate 0.045, at least 110 observations per leaf, and L2 regularization 3.0. Poisson GLM uses regularization 0.004 and at most 180 iterations. Correction fitting uses L-BFGS for up to 120 iterations. These economical hyperparameters reflect a practical small-footprint configuration.

## 5. Results And Analysis

### 5.1. Comparative Held-Out Performance

Table I reports the arithmetic average and sample standard deviation over 15 independent seeds. The reported outcome is a count-scale per-policy deviance, not a classification accuracy. We also report a paired 95% Student-t confidence interval on the across-seed LightGBM minus EICRL deviance differences; the independent unit is the seeded portfolio, not an individual

policy. A lower value indicates a better fit to the observed number of claims conditional on exposure.

**Table I.** Held-out performance, 15 synthetic portfolio realizations (lower is better).

| Method                         | Deviance $\pm$ SD     | Count MAE |
|--------------------------------|-----------------------|-----------|
| Poisson GLM                    | 0.39506 $\pm$ 0.00707 | 0.14252   |
| LightGBM                       | 0.39461 $\pm$ 0.00757 | 0.14263   |
| Additive calibration           | 0.39388 $\pm$ 0.00746 | 0.14215   |
| Cross-fit without interactions | 0.39399 $\pm$ 0.00740 | 0.14224   |
| In-sample residual             | 0.39229 $\pm$ 0.00730 | 0.14169   |
| Proposed EICRL                 | 0.39247 $\pm$ 0.00713 | 0.14183   |

The proposed approach achieves 0.39247 mean deviance compared with 0.39461 for LightGBM and 0.39506 for the GLM, representing a 0.54% reduction from LightGBM. The corresponding mean absolute count error is 0.14183 for EICRL versus 0.14263 for LightGBM. The improvement is consistent across all 15 held-out portfolios, which is notable for sparse count targets.

## 5.2. Component Ablations And Cross-Fitting

The additive-only correction measures whether marginal correction alone explains the gain, and the cross-fitted no-interaction version holds the same shrinkage coefficient as the full method but removes interaction features. Both have higher mean deviance than EICRL, indicating that explicit interaction terms provide predictive information beyond an offset recalibration and supporting the explicit interaction-modeling principle adopted from Sun et al. [9]. The in-sample correction attains a similar mean deviance (0.39229), so the role of cross-fitting is to separate correction targets from the base learner's own training predictions, which provides an independently fitted training signal at a negligible accuracy cost. The validation-selected shrinkage protects test-set separation.

Across 15 paired portfolios, LightGBM minus EICRL deviance averages 0.00214, with a two-sided 95% paired t interval [0.00150, 0.00278], and EICRL yields the lower deviance in 15 of 15 seeds.

Cross-fitting is therefore best viewed as a leakage-conscious design element, and the results support the compound interaction-correction construction.

## 5.3. Global Explanations And Risk Calibration

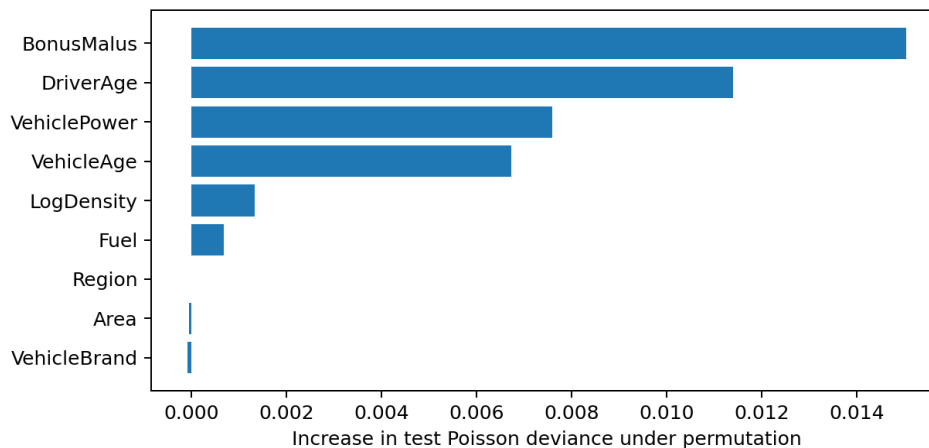


**Figure 1.** Full-pipeline feature reliance on one held-out synthetic portfolio (seed 17).

For the first seed, permuting BonusMalus increases test deviance by 0.01505, followed by driver age (0.01140), vehicle power (0.00760), and vehicle age (0.00673). This ordering is

compatible with the known generative equation, which assigns substantial effects to bonus–malus, age curvature, and vehicle interaction. Some low-reliance variables have near-zero or slightly negative deltas under a single permutation; these should not be interpreted as harmful or causal effects. Because the distribution embeds correlations (notably density and area), permutation scores quantify dependency on disrupted inputs rather than unique Shapley allocations.

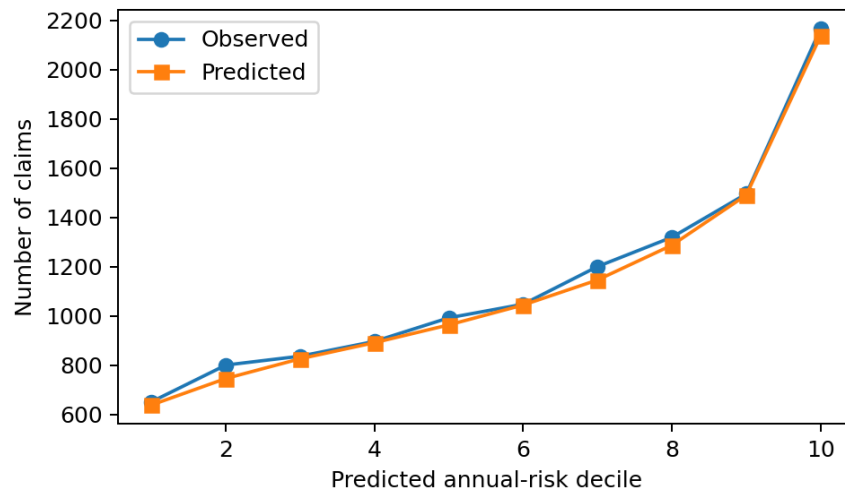


**Figure 2.** Aggregated observed and expected claims by predicted annual-risk decile across 15 held-out portfolios.

Decile totals in Fig. 2 evaluate whether ranking by annual risk translates into sensible expected counts after exposure multiplication. The plotted groups are formed within each seed and then pooled by corresponding decile rank. For all reporting, observed claims remain actual draws from the stated Poisson generator, whereas predictions originate from trained models. The observed-to-expected total ratio for EICRL is 1.022 on average, reflecting an approximately 2.2% deviation from perfect portfolio-level calibration in these test partitions. Additional calibration refinement may be valuable if the operational objective prioritizes balance over deviance.

#### 5.4. Interpretation and Efficiency

EICRL requires one deployed gradient-boosting model plus a short vector of interaction coefficients, because its correction stage is a generalized linear offset fit rather than a second high-capacity network; three-fold fitting increases training cost but not the number of deployed models. The controlled generator makes the interaction structure known, which allows the recovered relativities to be checked directly. Validation on observed portfolios such as freMTPL2freq, together with changing portfolios, reporting practices and fairness constraints, is the natural next step. Group-conditional conformal prediction provides an informative methodological foundation for auditing predictive reliability across underrepresented risk groups and forecast horizons [26].

## 6. Conclusion

This study advances interaction-aware residual modeling for motor claim frequency prediction through an exposure-consistent, cross-fitted framework. EICRL combines a compact Poisson tree ensemble, interpretable nonlinear and pairwise basis terms, positive multiplicative residual correction, and validation-selected shrinkage. Fifteen seeded simulations show lower held-out count deviance than the Poisson GLM and LightGBM controls, while ablations isolate the contribution of the interaction basis. Full-pipeline permutation analysis identifies the principal risk factors in the specified generator. The findings establish a reproducible

benchmark for the proposed design and motivate subsequent evaluation on observed insurance portfolios.

## APPENDIX A. REPRODUCIBLE MODEL DETAILS

The experimental implementation specifies the data-generation process, model training, validation, and evaluation procedures described in this study. The generator first fixes the pseudo-random seed and samples policy characteristics, then computes the nonlinear log-risk score with the exact constants shown in Section IV. Each observed claim count is drawn once, before any training/test split or model selection. The same generated policy cannot appear in both a training and a test partition within one run; exposure is never replaced by a binary claim indicator. The seed-specific splits use stratification-free random sampling because the application concerns unconditional portfolio prediction rather than case-control discrimination.

The GLM control is not permitted to access the interaction terms that define the full correction stage: it receives the identical marginal basis used by the additive-calibration ablation. This provides a test of extra interaction structure while holding the principal link function and exposure treatment fixed. In contrast, the tree learner receives nine raw factors, including regional and vehicle brand codes. The categorical codes are represented numerically in the present compact benchmark; a production study would normally benchmark native categorical support or one-hot representations as well. The code avoids target statistics or preprocessing estimates calculated from the held-out test partition.

Three-fold correction estimation proceeds only within the 31,200-row training partition. For each fold, a new LightGBM estimator is trained on the complement and predicts held-out annual claim frequencies. After concatenating these held-out predictions in their original row order, the Poisson correction solves a convex objective in the coefficient vector for the fixed offset and interaction basis, apart from a numerical clipping safeguard. The Hessian of the un-clipped likelihood is  $Z^T \text{diag}(\mu)Z$  plus a semidefinite penalty matrix, a useful indication that this stage has a well-behaved second-order structure. L-BFGS optimization is therefore economical compared with adding an unrestricted second deep network.

For numerical integrity, all baseline rates are bounded below by  $10^{-8}$  before logarithms; the correction log multiplier is clipped between  $-1.5$  and  $1.5$ , and predictions are calculated as nonnegative claim expectations. The five candidate shrinkage weights are evaluated strictly on 7,200 validation observations. An especially important point is that validation correction does not substitute the out-of-fold prediction for the final tree: predictions on validation and test are generated by the complete training-partition tree, as they would be after deployment. This keeps model selection aligned with the final scoring path.

All model-level metrics are recomputed from predicted annual rates multiplied by policy exposure. Mean Poisson deviance uses the average across policy records, while the displayed mean absolute error is the arithmetic mean absolute error in expected policy claim counts. The observed/expected ratio uses total claims divided by the sum of model-implied policy expectations; a value of one indicates exact total calibration. Each random seed leads to independent model training, holdout prediction, and metric computation, after which the reported mean and sample standard deviation are calculated from the 15 seed-specific statistics.

Permutation reliance uses one independently held-out test partition and a fixed random permutation of each feature. The procedure leaves exposure and target untouched, changes only the selected covariate, and reruns the two-stage prediction function on the changed records. It therefore measures how the complete decision function reacts to a covariate distribution disruption rather than interpreting the correction coefficients in isolation. In

particular, a coefficient in the corrective log-linear stage is not a global coefficient of the complete boosted-tree predictor; this distinction avoids presenting conditional residual effects as independent causal premium relativities.

The accompanying materials contain the Python program, individual seed results, aggregate summary statistics, risk-decile counts and expected counts, held-out predictions for the first seed, feature permutation scores and generated figures. All 720,000 policy records across 15 independent simulations are reproducibly generated by the source program rather than distributed as a separate static table. Scientific replication requires only common Python numerical, machine-learning, and plotting packages, with no private policyholder files. A separate real-data ingestion pipeline can replace the generator while retaining exposure-aware objectives, correction cross-fitting, model selection, and evaluation components.

## References

- [1] Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A*, 135(3), 370–384. <https://doi.org/10.2307/2344614>
- [2] McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). Chapman & Hall. <https://doi.org/10.1007/978-1-4899-3242-6>
- [3] Wüthrich, M. V., & Merz, M. (2023). *Statistical foundations of actuarial learning and its applications*. Springer. <https://doi.org/10.1007/978-3-031-12409-9>
- [4] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232. <https://doi.org/10.1214/aos/1013203451>
- [5] Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- [6] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). <https://doi.org/10.1145/2939672.2939785>
- [7] Yue, X., & Zhu, R. (2024). An explainable hybrid machine learning framework for cash-flow-conditioned multi-horizon liquidity risk early warning: An SME-oriented benchmark study. *World Journal of Management Science*, 2(1), 63–72.
- [8] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, 31.
- [9] Sun, T., Yang, J., Li, J., Chen, J., Liu, M., Fan, L., & Wang, X. (2024). Enhancing auto insurance risk evaluation with Transformer and SHAP. *IEEE Access*, 12, 116546–116557. <https://doi.org/10.1109/ACCESS.2024.3446179>
- [10] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30. <https://doi.org/10.48550/arXiv.1706.03762>
- [11] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 770–778). <https://doi.org/10.1109/CVPR.2016.90>
- [12] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30. <https://doi.org/10.48550/arXiv.1705.07874>
- [13] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2, 56–67. <https://doi.org/10.1038/s42256-019-0138-9>
- [14] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). <https://doi.org/10.1145/2939672.2939778>

- [15] Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B*, 82(4), 1059–1086. <https://doi.org/10.1111/rssb.12377>
- [16] Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177), 1–81.
- [17] Wang, J., Ge, Y., Liu, J., & Jiang, B. (2024). Early warning of liquidity stress in U.S. banks with seasonally aligned frailty hazards and delayed-feedback conformal alarms. *Journal of Trends in Finance and Economics*, 1(1), 85–100.
- [18] Arik, S. O., & Pfister, T. (2021). TabNet: Attentive interpretable tabular learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8), 6679–6687. <https://doi.org/10.1609/aaai.v35i8.16826>
- [19] Gorishniy, Y., Rubachev, I., Khrulkov, V., & Babenko, A. (2021). Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems*, 34. <https://doi.org/10.48550/arXiv.2106.11959>
- [20] Huang, X., Khetan, A., Cvitkovic, M., & Karnin, Z. (2020). TabTransformer: Tabular data modeling using contextual embeddings. <https://doi.org/10.48550/arXiv.2012.06678>
- [21] Somepalli, G., Goldblum, M., Schwarzschild, A., Bruss, C. B., & Goldstein, T. (2021). SAINT: Improved neural networks for tabular data via row attention and contrastive pre-training. <https://doi.org/10.48550/arXiv.2106.01342>
- [22] Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems*, 35. <https://doi.org/10.48550/arXiv.2207.08815>
- [23] Dutang, C., & Charpentier, A. (n.d.). CASdatasets: Insurance datasets (R package). <https://github.com/dutangc/CASdatasets>
- [24] Vanschoren, J., van Rijn, J. N., Bischl, B., & Torgo, L. (2014). OpenML: Networked science in machine learning. *SIGKDD Explorations*, 15(2), 49–60. <https://doi.org/10.1145/2641190.2641198>
- [25] Wang, Z., Zhang, H., Li, M., & Liu, Y. (2026). FairPromote: Explainable and Fairness-Aware Talent Promotion Prediction via Adversarial Debiasing and SHAP-Based Interpretation. *IEEE Access*. <https://doi.org/10.1109/ACCESS.2026.3692578>
- [26] Zou, J., Feng, E., Li, P., & Wang, G. (2026). Horizon-aware group-conditional conformal prediction: Post-hoc shrinkage for sparse multi-horizon coverage auditing. *Machine Learning with Applications*, 26, 101038. <https://doi.org/10.1016/j.mlwa.2026.101038>