Elevating Trust in High-Stakes Decisions Using Glass-Box Models and Robust Feature Selection

Matthias Linder Judith C. Schneider Brandon Schwab

Institute for Risk and Insurance, Leibniz University Hannover

London, June 20, 2025

Motivation

The Need for Explainable and Robust AI in High-Stakes Decision-Making

High Predictive Power vs. High Risk

- Modern machine learning methods are extremely powerful for modeling complex, high-dimensional data at scale (Jordan & Mitchell, 2015; LeCun et al., 2015).
- However, in high-stakes applications (e.g. insurance), model performance alone is insufficient. Decisions must be interpretable, transparent, and robust to maintain trust and meet ethical, regulatory, and societal requirements (Doshi-Velez & Kim, 2017; Karimian et al., 2022; Svetlova, 2022).

Limitations of Post-Hoc Explainability

- Popular methods like SHAP and LIME offer approximations, but are often unstable and unfaithful to the model's true decision logic (Lundberg & Lee, 2017; Ribeiro et al., 2016; Rudin, 2019).
- Regulatory frameworks such as the EU AI Act and GDPR demand explanations that are consistent, understandable, and reproducible (EU AI Act, 2024; GDPR, 2016).

Challenges Beyond Explainability: Robustness

Need for Intrinsic Transparency and Robustness

- Even inherently interpretable models (e.g., GAMs, rule-based methods) lose credibility if the chosen features or relationships shift drastically due to minor data perturbations (Hamer & Dupont, 2021; Kalousis et al., 2007).
- In high-stakes applications, domain experts often prioritize a more stable feature selection process over one that yields slightly higher accuracy but exhibits greater variability (Hamer & Dupont, 2021).

Research Contribution: Interpretable Modeling + Robust Feature Selection

Neural Additive Models (NAMs)

- We employ NAMs, which leverage deep learning to learn feature-wise relationships while maintaining an interpretable additive structure (Agarwal et al., 2021).
- Unlike black-box models, NAMs allow for direct visualization and an **exact** explanation of how each feature influences predictions, ensuring faithful and reliable explanations.

Robust Feature Selection

- We extend the Single Feature Introduction Test (SFIT) (Horel & Giesecke, 2022) by integrating a mean-based selection criterion into a forward-selection scheme.
- Using bootstrap aggregation, we identify features that are consistently important across resampled datasets—boosting robustness to data perturbations.

Modeling Framework

Transparent Models + Robust Feature Selection

Model Comparison

- **GLM:** Industry baseline
- **GAM:** Classical transparent model
- **Feed-Forward Neural Net:** Black-box benchmark
- **Gradient Boosting Machine:** Black-box benchmark
- **Neural Additive Model (NAM):** Deep Learning Glassbox Model

Unified Pipeline

- **Step 1:** Apply forward feature selection using modified SFIT across bootstraps
- **Step 2:** Select features that are consistently useful
- **Step 3:** Fit final models on selected features

Robust Feature Selection

Efficient, Stable, and Model-Agnostic

Key Idea:

- Test each feature's value by asking:
  "Does this feature consistently reduce prediction error?"

Our Approach:

- Build on the SFIT framework (Horel & Giesecke, 2022)
- Use a mean-based test (Diebold-Mariano) to assess importance
- Add features forward if they significantly improve model performance
- Repeat on many bootstrap samples to ensure stability

Why It Matters:

- **No need to retrain:** Efficient & scalable
- **Handles interactions** without exhaustive search
- **Robust to data noise:** Final feature set is stable and credible

Experimental Workflow

Datasets

1. Public MTPL Dataset

- 163k policyholders; standard benchmark in actuarial research (Denuit & Lang, 2004; Henckaerts et al., 2021).
- Policy periods range from 1 day to 1 year; static risk factors.

2. Proprietary Motor Dataset

- Real-world dataset from German insurer.
- 10M records with detailed policy & claims data.

Modeling Pipeline Overview

Unified Pipeline for Frequency and Severity Modeling:

1. Train/test split (80/20, stratified)
2. Generate $B = 25$ bootstrap samples
3. Robust feature selection:
   - Select main & interaction effects (keep if selected in $\geq 60\%$ of runs)
4. Tune model hyperparameters via random search
5. Refit final model on selected features (100 bootstraps)

house of insurance
Institute for
Risk and Insurance

Predictive Performance - Poisson Deviance

Table 1: Poisson Deviance Loss by Dataset

Dataset	Model	Mean Loss	Lower 95% CI	Upper 95% CI
Public MTPL Test Set	GLM-Frequency	0.5320	0.5318	0.5323
	GAM-Frequency	0.5319	0.5317	0.5322
	FFNN-Frequency	0.5334	0.5322	0.5344
	NAM-Frequency	0.5316	0.5310	0.5322
	GBM-Frequency	**0.5292**	0.5284	0.5300
Proprietary Test Set	GLM-Frequency	0.2193	0.2193	0.2193
	GAM-Frequency	0.2192	0.2192	0.2193
	FFNN-Frequency	0.2192	0.2190	0.2196
	NAM-Frequency	**0.2192**	0.2191	0.2194
	GBM-Frequency	0.2194	0.2190	0.2209

Results

house of insurance
Institute for
Risk and Insurance

Predictive Performance - Gamma Deviance

Table 2: Gamma Deviance Loss by Dataset

Dataset	Model	Mean Loss	Lower 95% CI	Upper 95% CI
Public MTPL Test Set	GLM-Severity	2.2667	2.2618	2.2733
	GAM-Severity	2.2667	2.2618	2.2733
	FFNN-Severity	2.3742	2.2720	2.4124
	NAM-Severity	2.2605	2.2584	2.2633
	GBM-Severity	**2.2598**	2.2444	2.2774
Proprietary Test Set	GLM-Severity	0.9928	0.9917	0.9938
	GAM-Severity	0.9925	0.9912	0.9936
	FFNN-Severity	0.9966	0.9852	1.0138
	NAM-Severity	**0.9923**	0.9856	1.0010
	GBM-Severity	0.9931	0.9922	0.9948

Results

Interpretability - Feature Selection

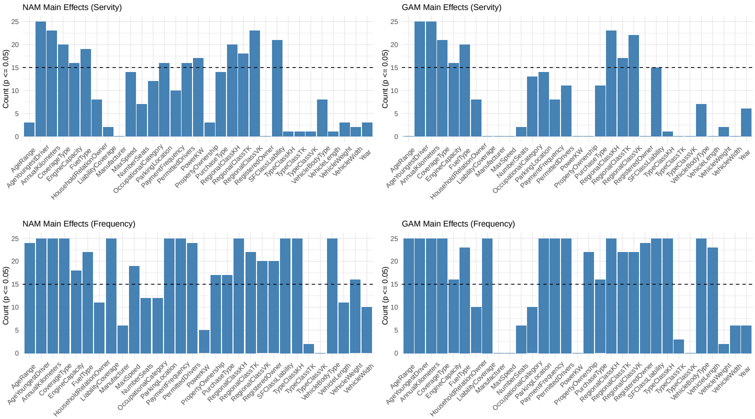
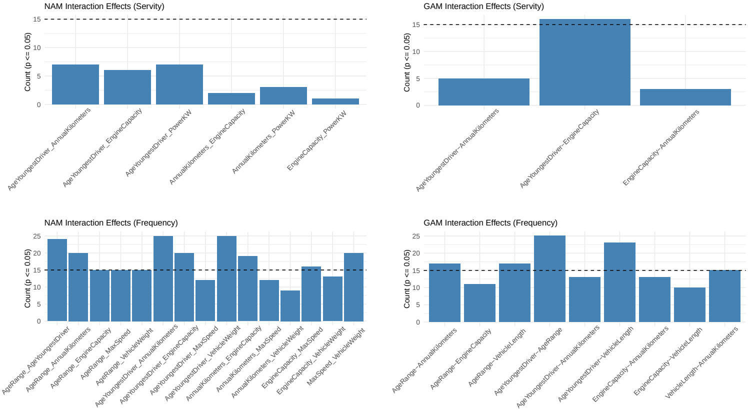


Figure 1: Robust Feature Selection for Main Effects in GAM and NAM.

Results

Interpretability - Feature Selection



Interpretability - Feature Effects

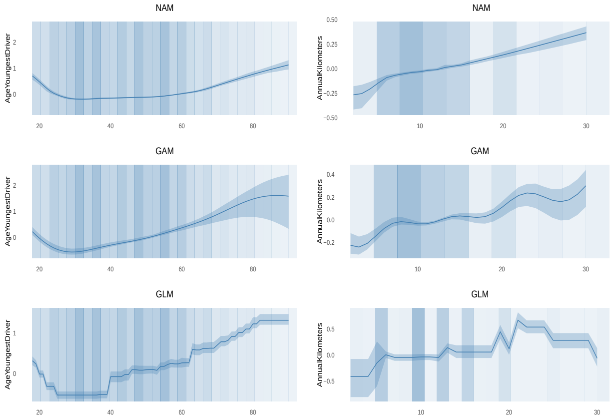


Figure 3: Partial Effect Plots fo the Age of the Youngest Driver and the Annual Kilometers the Frequency Models.

References

References I

Agarwal, R., Melnick, L., Frosst, N., Zhang, X., Lengerich, B., Caruana, R., & Hinton, G. E. (2021). Neural additive models: Interpretable machine learning with neural nets. *Advances in Neural Information Processing Systems*, *34*, 4699–4711.

Denuit, M., & Lang, S. (2004). Non-life rate-making with bayesian gams. *Insurance: Mathematics and Economics*, *35*(3), 627–647.

Doshi-Velez, F., & Kim, B. (2017).Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

EU AI Act. (2024). Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending regulations (ec) no 300/2008,(eu) no 167/2013,(eu) no 168/2013,(eu) 2018/858,(eu) 2018/1139 and (eu) 2019/2144 and directives 2014/90/eu,(eu) 2016/797 and (eu) 2020/1828 (artificial intelligence act) off. *Official Journal of the European Union*, *L series*, 1–144.

References II

GDPR. (2016). Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (general data protection regulation). *Official Journal of the European Union*, *L 119*(1), 1–88.

Hamer, V., & Dupont, P. (2021). An importance weighted feature selection stability measure. *Journal of Machine Learning Research*, *22*(116), 1–57.

Henckaerts, R., Côté, M.-P., Antonio, K., & Verbelen, R. (2021). Boosting insights in insurance tariff plans with tree-based machine learning methods. *North American Actuarial Journal*, *25*(2), 255–285.

Horel, E., & Giesecke, K. (2022). Computationally efficient feature significance and importance for predictive models. *Proceedings of the Third ACM International Conference on AI in Finance*, 300–307.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260.

Kalousis, A., Prados, J., & Hilario, M. (2007). Stability of feature selection algorithms: A study on high-dimensional spaces. *Knowledge and Information Systems*, *12*, 95–116.

Karimian, G., Petelos, E., & Evers, S. M. (2022).The ethical issues of the application of artificial intelligence in healthcare: A systematic scoping review. *AI and Ethics*, *2*(4), 539–551.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should i trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215.

References IV

Svetlova, E. (2022).AI ethics and systemic risks in finance. *AI and Ethics*, *2*(4), 713–725.