The Institute of Actuaries of Japan
Think the Future, Manage the Risk

Maximizing Interpretability of Black-Box Models with {**midr**}

Ryoichi Asashiba
T&D Holdings Inc.
Convention A | Asia, March 24, 2026

Software Implementations

{midr}

- The core **R** package for MID.
- The **core engine** for the decomposition utilizing {**RcppEigen**}, a C++ based library for performing linear algebra.

- CRAN (latest release)
  https://CRAN.R-project.org/package=midr
- GitHub (latest development version)
  https://github.com/ryo-asashi/midr

{midnight}

- **R** package to integrate {**midr**} to the {**tidymodels**} ecosystem.
- Enhances some features of {**midr**} by defining S3 methods for MID objects.

- GitHub (latest development version)
  https://github.com/ryo-asashi/midnight

{midlearn}

- **Python** library to integrate {**midr**} to the {**scikit-learn**} ecosystem.
- Depends on {**rpy2**} for integration, and {**plotnine**} for visualization.

- PyPI (latest release)
  https://pypi.org/project/midlearn/
- GitHub (latest development version)
  https://github.com/ryo-asashi/midlearn

Dataset: French Motor Insurance

- We use the **French Motor Third-Party Liability dataset** for claim frequency.
- It contains detailed policyholder, vehicle, and driver information for thousands of French car insurance policies, including claim counts, and has been used widely for actuarial modeling,

Objectives and Evaluation Metrics

- Our goal is to fit regression models that predict **Claim Frequency** = Claim Numbers / Exposure. Then, we aim to interpret these predictive models by constructing surrogate models.
    - Predictors: policyholder/vehicle information such as **Driver Age**, **Vehicle Power**, **Vehicle Age**, **Vehicle Brand**, **Fuel Type**, as well as geographical factors such as **Region** (the policy region where the policyholder resides) and **Population Density** (inhabitant density of the policyholder's city).

- The loss metric is the **Exposure-Weighted Mean Poisson Deviance**:

$$\text{Loss}(\mathbf{y}, \hat{\mathbf{y}}, \mathbf{w}) = \frac{\sum_{i=1}^{n} w_i d(y_i, \hat{y}_i)}{\sum_{i=1}^{n} w_i}, \quad d(y_i, \hat{y}_i) = 2[y_i(\log y_i - \log \hat{y}_i) - (y_i - \hat{y}_i)]$$

- For surrogate models, we assess the **Uninterpreted (Variation) Ratio** (**UR**) to measure how closely the MID model g replicates the black-box model f.

$$\mathbf{UR}(f, g; \mathbf{X}) = \frac{\mathbf{E}\left[\left(f(\mathbf{X}) - g(\mathbf{X})\right)^2\right]}{\mathbf{Var}[f(\mathbf{X})]} \approx \frac{\sum_{i=1}^{n}\left(f(\mathbf{x_i}) - g(\mathbf{x_i})\right)^2}{\sum_{i=1}^{n}\left(f(\mathbf{x_i}) - \bar{f}\right)^2}$$

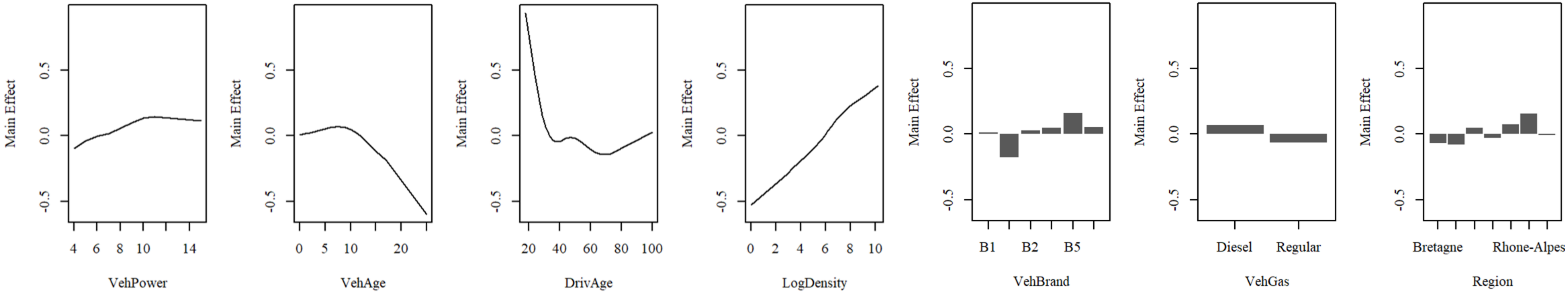
Interpretable Baseline: GAM

- **Generalized Additive Model** (**GAM**) has the simple additive structure *by design*:

$$\hat{\mathbf{y}}_{\text{gam}} = f_{\text{gam}}(\mathbf{X}) = \exp\left(\beta_0 + \sum_j f_j(X_j)\right)$$

  - Poisson Deviance: $\text{Loss}_{\text{test}}(\mathbf{y}, \hat{\mathbf{y}}_{\text{gam}}, \mathbf{w}) \approx 0.4679$

  - Feature Effects:



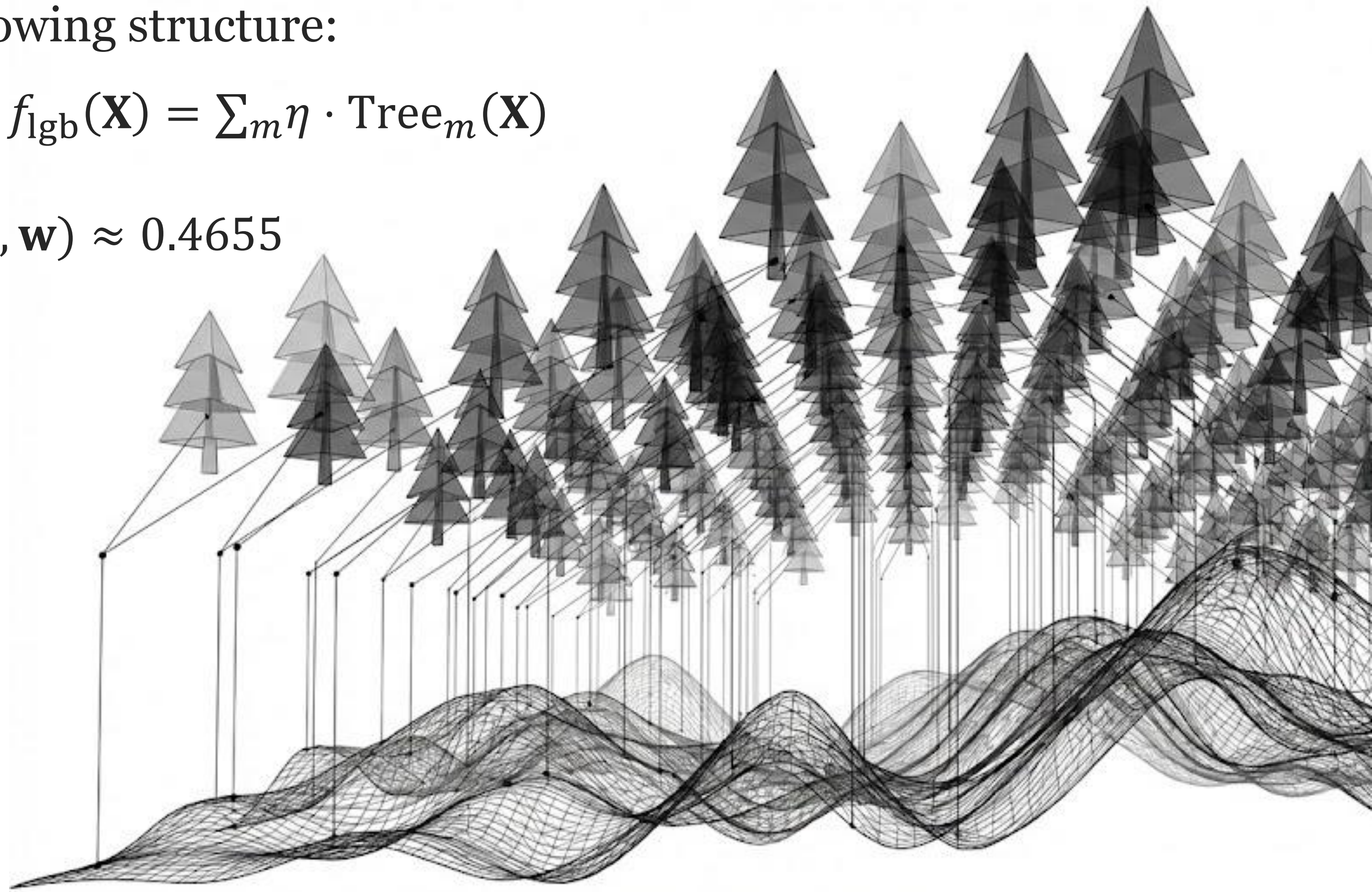
Black-Box Model: LightGBM

- **LightGBM** (**LGB**) is a gradient boosting framework that uses tree-based learning algorithms.

  We fit a LightGBM regressor of the following structure:

$$\hat{\mathbf{y}}_{\text{lgb}} = f_{\text{lgb}}(\mathbf{X}) = \sum_m \eta \cdot \text{Tree}_m(\mathbf{X})$$

  - Poisson Deviance: $\text{Loss}_{\text{test}}(\mathbf{y}, \hat{\mathbf{y}}_{\text{lgb}}, \mathbf{w}) \approx 0.4655$
  - Feature Effects: N/A

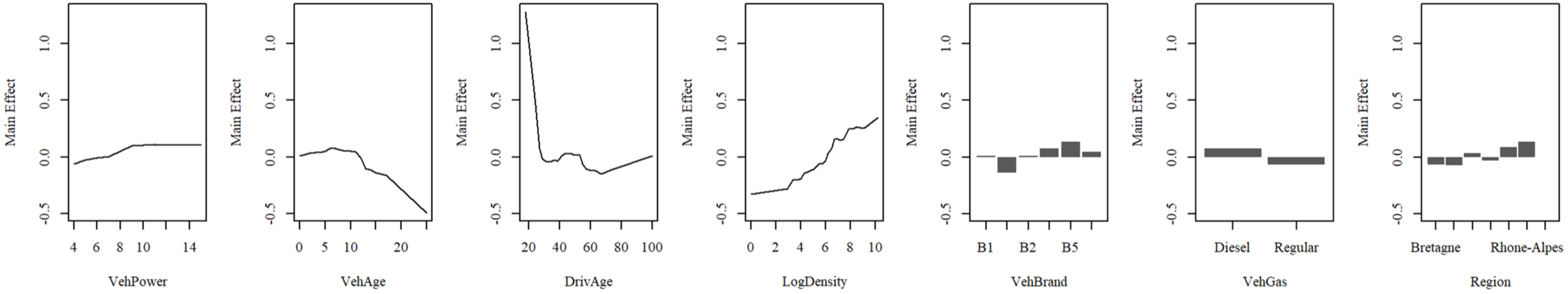


Surrogate Model: Second-order MID

- With {**midr**}, we fit a Second-order MID model of the LightGBM regressor:

$$\log\left(f_{\mathrm{lgb}}(\mathbf{X})\right) \approx \log(\hat{\mathbf{y}}_{\mathrm{mid}}) = g_{\emptyset} + \sum g_j(X_j) + \sum g_{jk}(X_j, X_k)$$

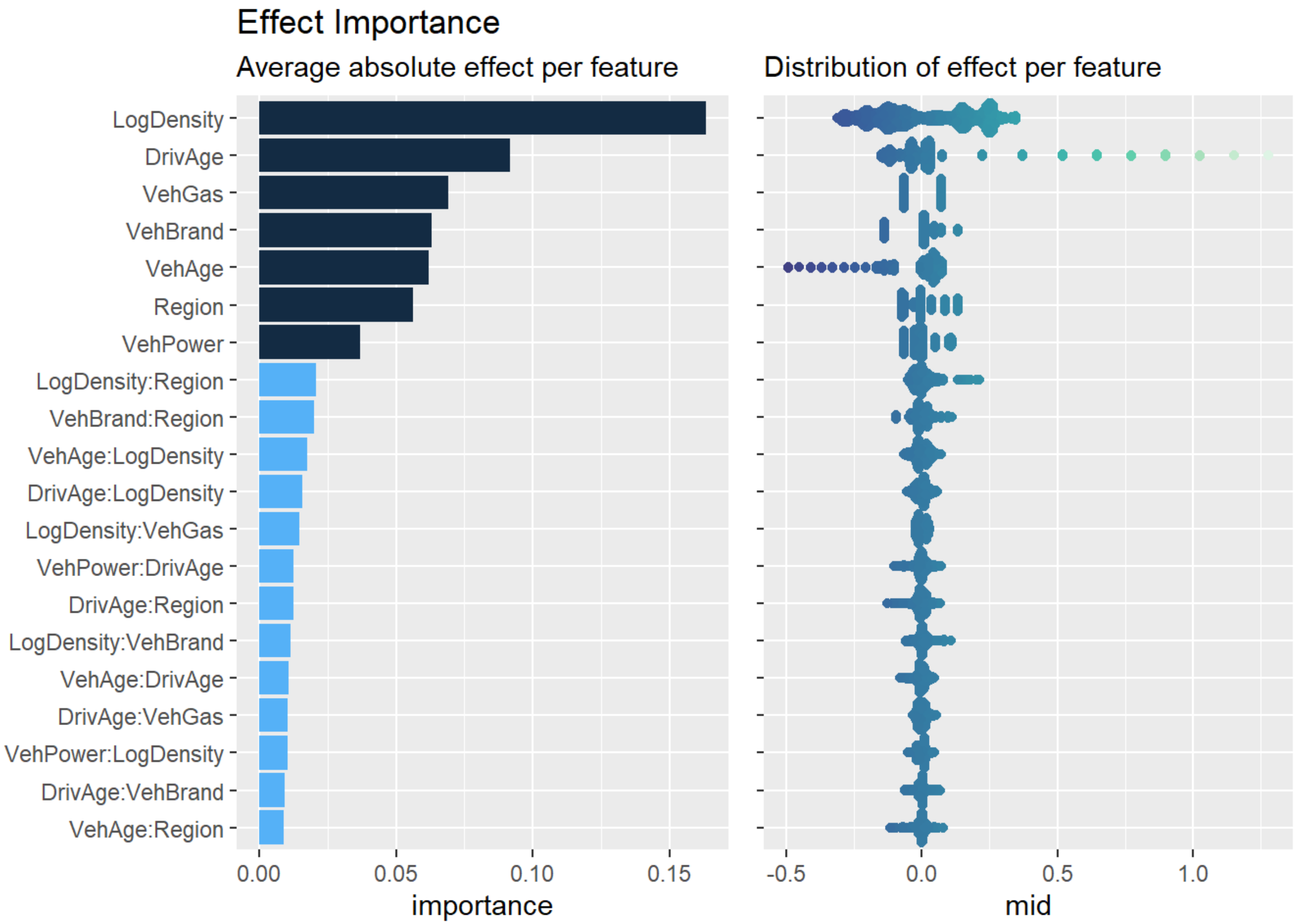
  - Model Fidelity: $\mathbf{UR}_{\mathrm{train}}\left(\log \circ f_{\mathrm{lgb}}, g\right) \approx 7\%$
  - Poisson Deviance: $\mathrm{Loss}_{\mathrm{test}}(\mathbf{y}, \hat{\mathbf{y}}_{\mathrm{mid}}, \mathbf{w}) \approx 0.4671$
  - Feature Effects (Main Effects):



Feature Effect Importance

- We can measure and visualize the **Feature Effect Importance** of each component function g_S:

$$\text{Importance}(g_S) = \mathbf{E}\left[\left|g_S(X_j)\right|\right] \approx \frac{1}{n}\sum_{i=1}^{n}\left|g_S(x_{ij})\right|$$



Interaction: the Commuter Effect in Paris

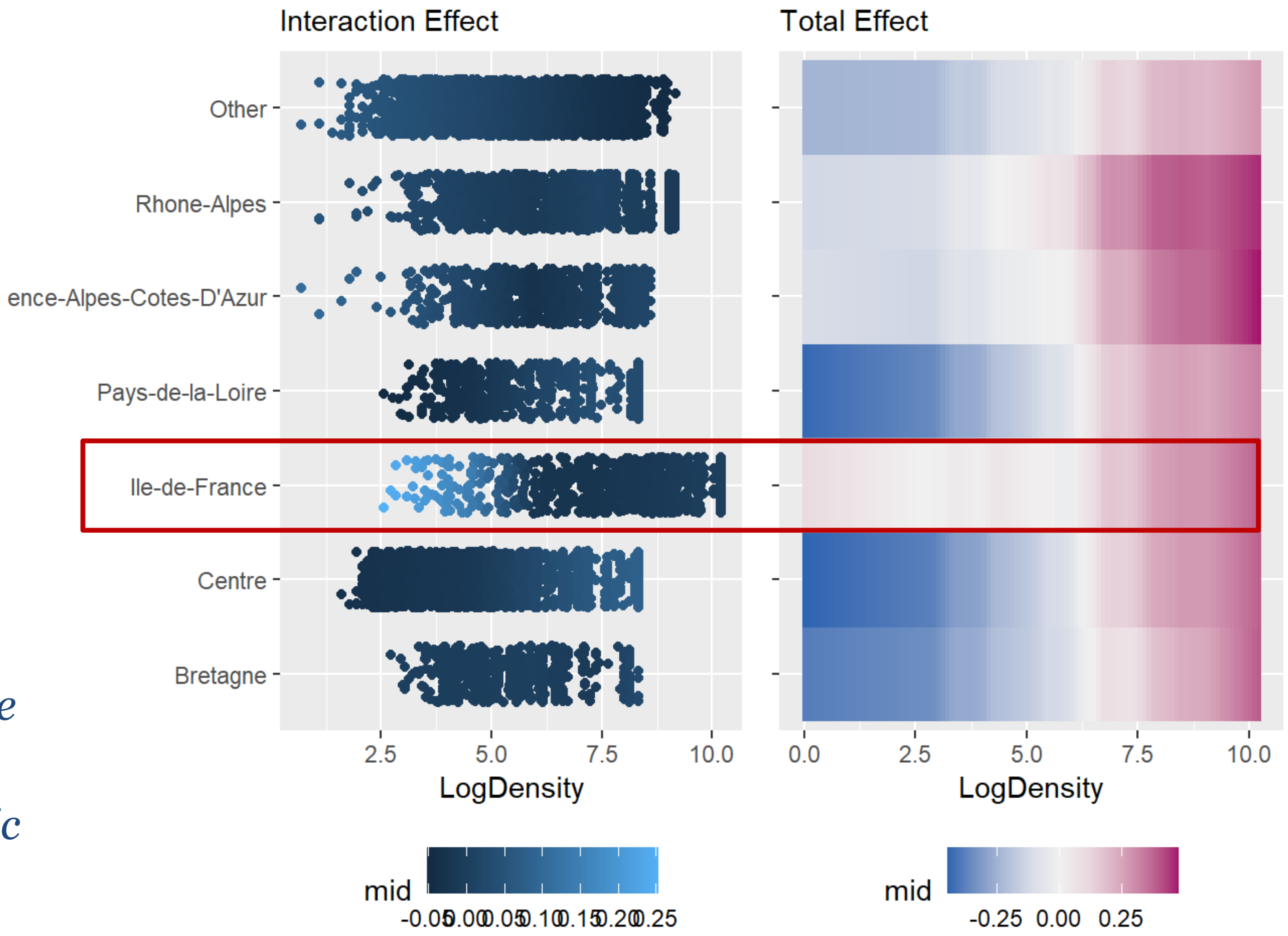
- We focus on the most important interaction: **LogDensity-Region**

- While the main effect of LogDensity is increasing, the joint effect reveals high risks even in low-density areas in *Île-de-France*.

- A possible explanation:
  the Commuter Effect in IDF region

  Even residents of low-density areas in Île-de-France typically commute into the dense urban core and spend substantial time on heavily congested roads.

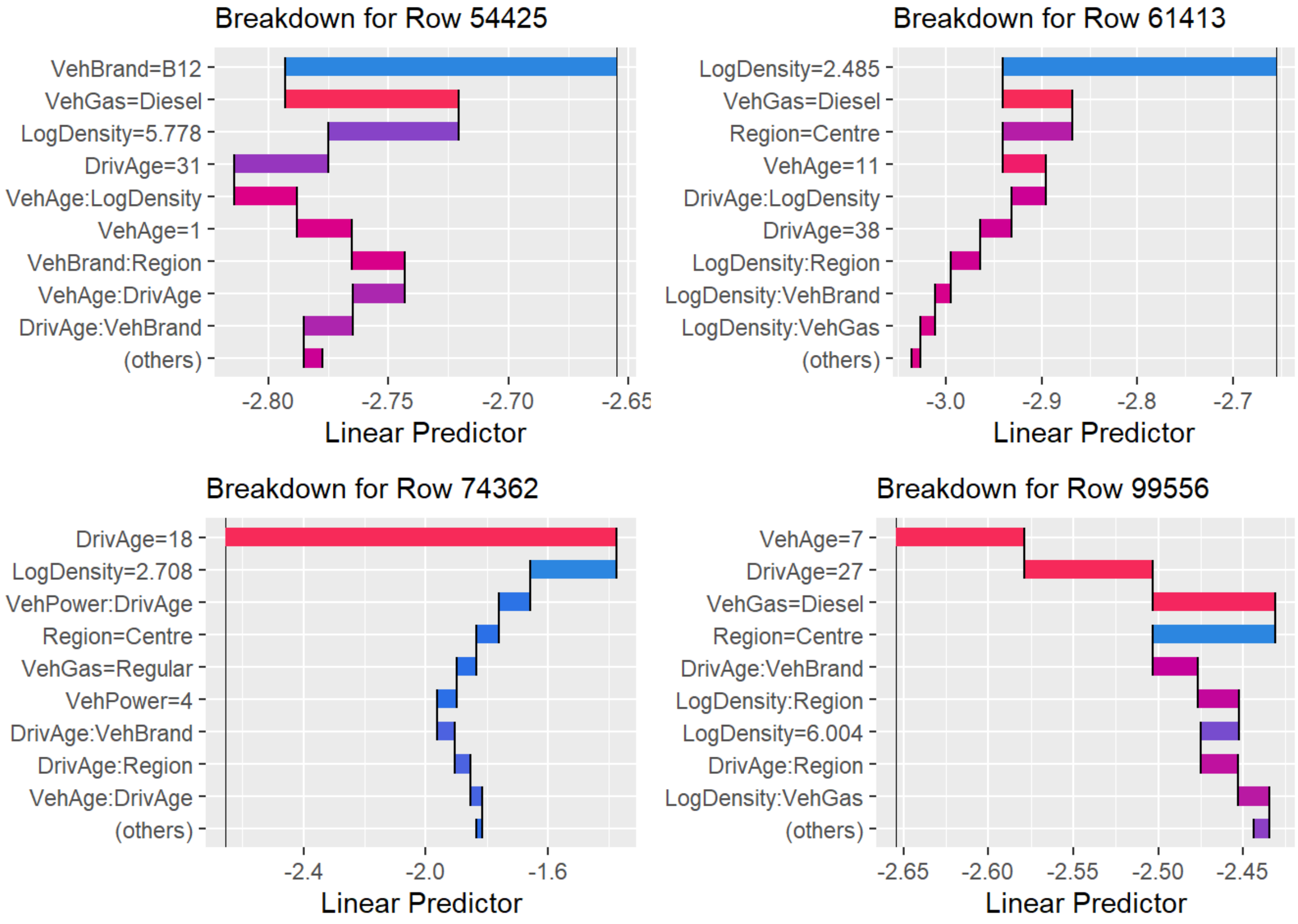
  As a result, their exposure to high-risk traffic conditions remains elevated despite low residential density.



Explaining Individual Policy Risks

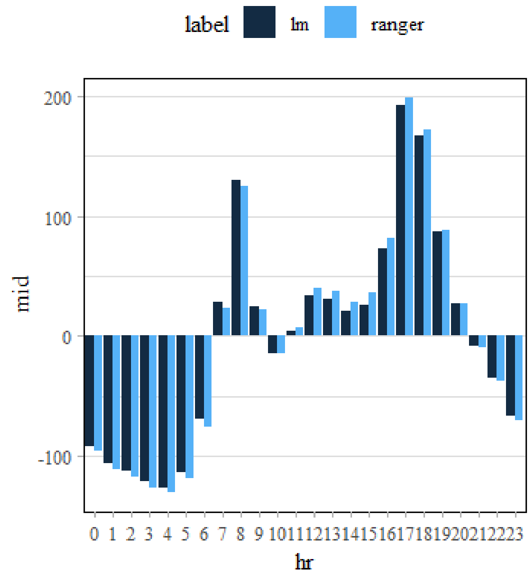
- We can perform instance-level explanation through **Additive Breakdown** of each prediction:

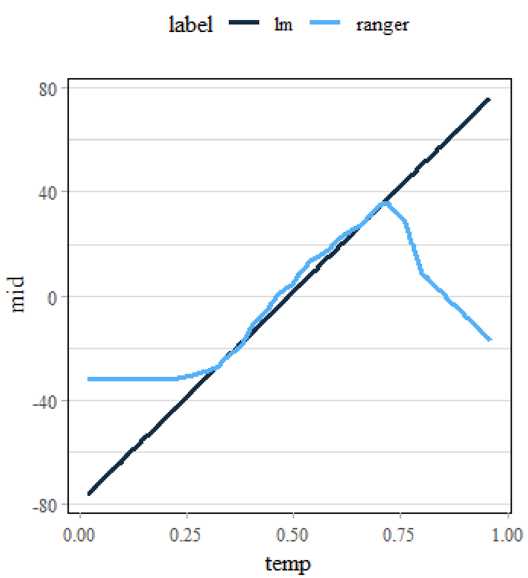
$$g(\mathbf{x}) = g_{\emptyset} + \sum_{j} g_j(x_j) + \sum_{j<k} g_{jk}(x_j, x_k)$$

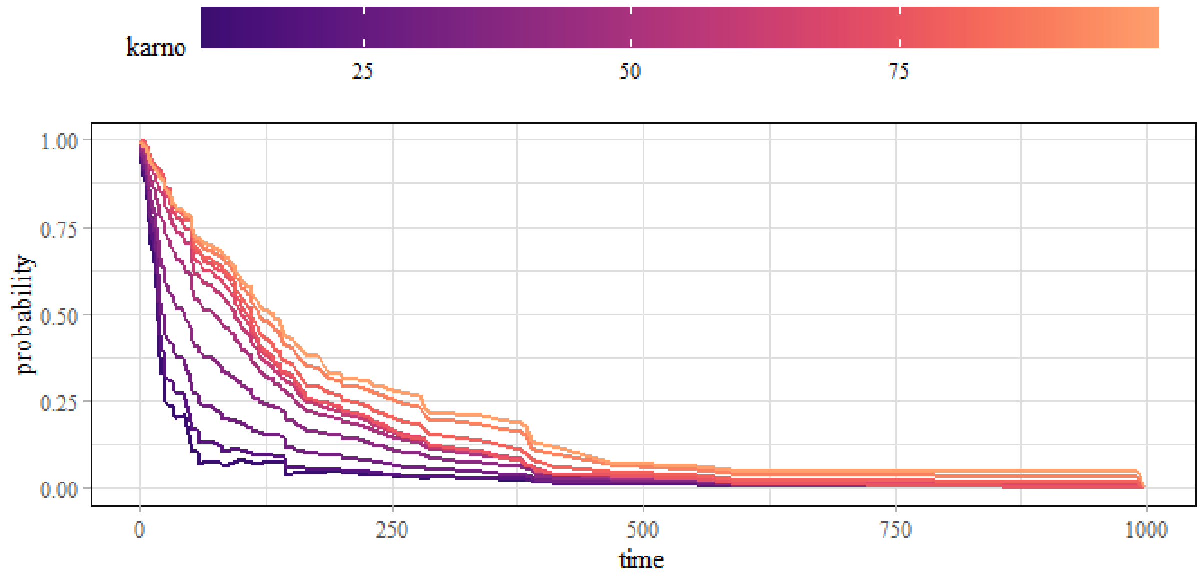


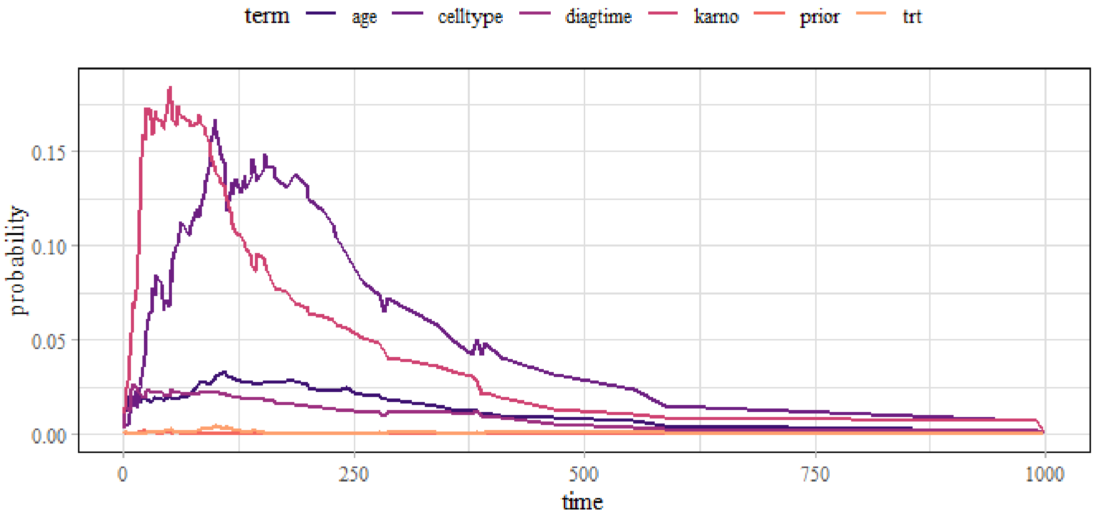
Looking Ahead: Future Directions

- New functionalities of **{midr}**:
    - ➢ **Unified Model Comparison**
    - ➢ **Interpretable Survival Analysis**









Conclusion

- **Bridge to Transparency**:
  MID transforms complex Black-Box models into transparent, additive structures.

- **Complexity Diagnostics**:
  The Uninterpreted (Variation) Ratio quantifies model complexity.
  If the ratio is high, the model relies on complex effects beyond second-order interaction.

- **Justifiable Decisions**:
  We can now justify model complexity for both performance and regulatory needs.
  No more "operating in the dark."

➢ R/Python Examples: https://ryo-asashi.github.io/Convention_A_Asia/

Thank you for joining!