Institute and Faculty of Actuaries

Understanding AI - eXplainable AI (XAI) techniques in practice

Karol Gawłowski

#GiroConf22

Agenda

Motivation

1. ICE – Individual Conditional Expectation
2. PDP – Partial Dependence Plot
3. ALE – Accumulated Local Effects
4. XAI in
5. SHAP – Shapley Values
    5.1. Adapting Shapley values to XAI
    5.2. Observation level explanation
    5.3. Variable level explanation
    5.4. Model level explanation
    5.5 SHAP in

Further reading and observations

Introduction – Motivation

Accuracy

Automation

FCA / PRA

Product Innovation

Differentiation

Data capture

AI Regulations

Portfolio manag.

Accountability

GDPR

New data types

eXplainable AI

Interpretable ML

Transparency

Interpretable AI

Responsible AI

Fairness

Institute
and Faculty
of Actuaries

Introduction – Motivation

Performance – interpretability trade-off

		Interpretability	
		yes	no
Predictive power	+		Machine Learning
	-	Statistical Modelling	

XAI

- Interpretability – the degree to which a human can consistently predict the model's result[1]
- Predictive power – degree to which a mathematical model is useful in determining results of some process

Why XAI?

- No more trade-off between models' power and explainability
- XAI is essential in debugging black box models
- Black box models can serve as benchmarks on parameter level, not just performance
- Modelers can assess fairness and bias
- Most popular XAI techniques are model agnostic

1. B. Kim et. al. "Examples are not enough, learn to criticize! Criticism for interpretability." Advances in Neural Information Processing Systems (2016)

Institute and Faculty of Actuaries

Introduction – a black box model

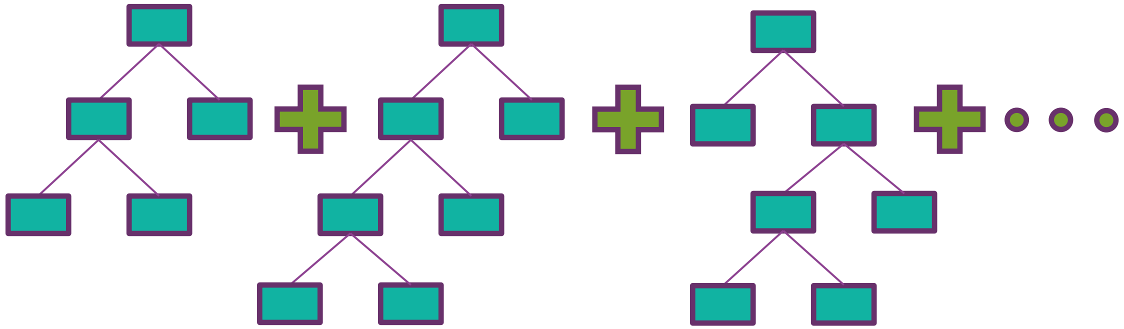
Boosted trees with `library(xgboost)` and `CASdatasets::data(freMTPL2freq)`

The data

- 677,991 Motor Third Party Liability policies
- Exposure and claim data
- 7 features:
  - Power (categorical)
  - VehAge (cont.)
  - DrivAge (cont.)
  - Brand (cont.)
  - Gas (Diesel/Regular)
  - Region (categorical)
  - Density (cont.)

The model

- Predict claim frequency (poisson)
- Gradient Boosting Machine
- Parameters – Grid search



github.com/Karol-Gawlowski/GIRO_2022 The Actuary

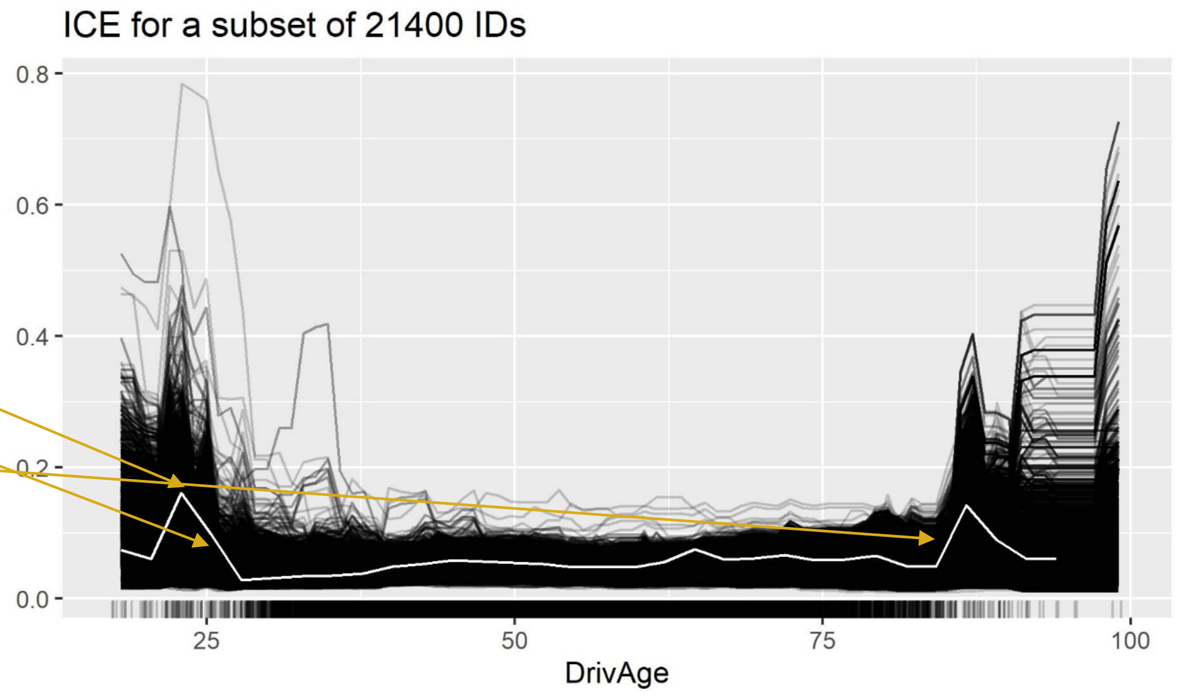
Institute and Faculty of Actuaries

1. ICE – Individual Conditional Expectation

ID	Area	…	DrivAge	Pred
123	D	…	9	0.475

ID’	Area	…	DrivAge	Pred
123.1	D		20	0.060
123.2	D		23	0.161
123.3	D		25	0.098
…	…	…	…	…
123.32	D		84	0.049

Repeat for a sample of policies

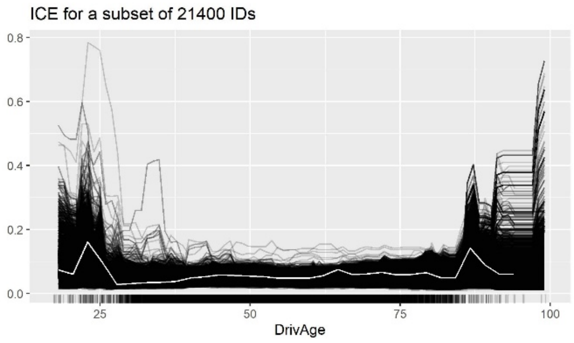


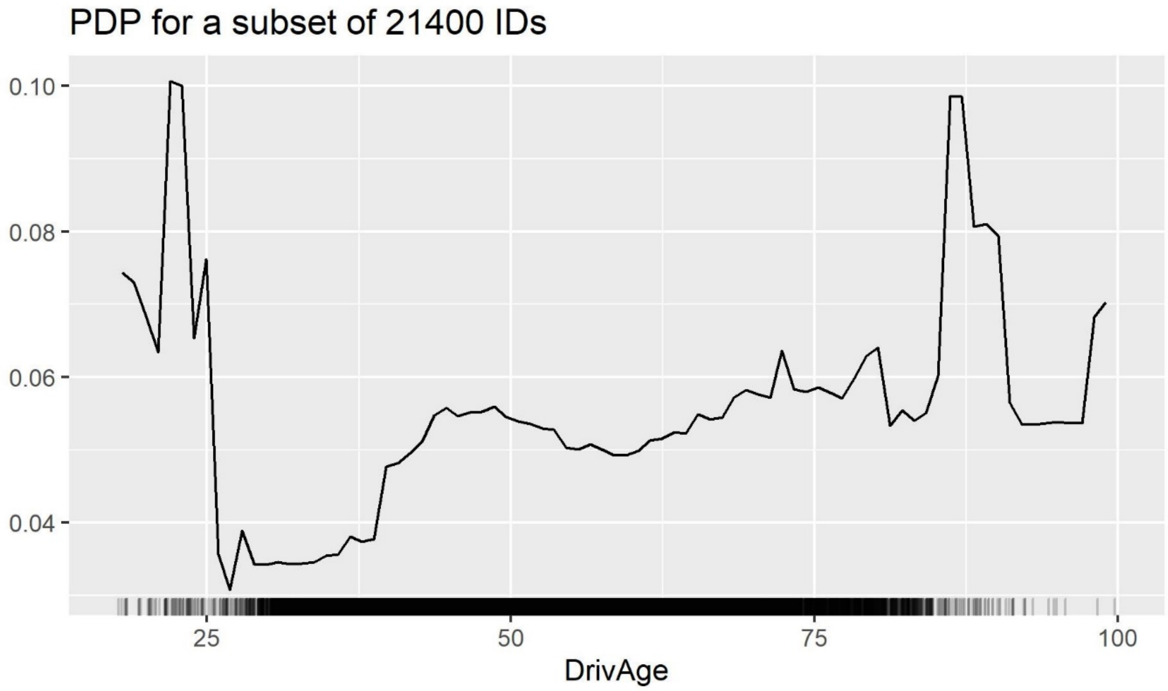
- Observation/variable level explanation
- Alter DrivAge keeping other variables constant and obtain the prediction for the artificial points

The IDs chosen are for drivers with lowest Bonus Malus score owning 5-year-old cars

Institute and Faculty of Actuaries

2. PDP – Partial Dependence Plot





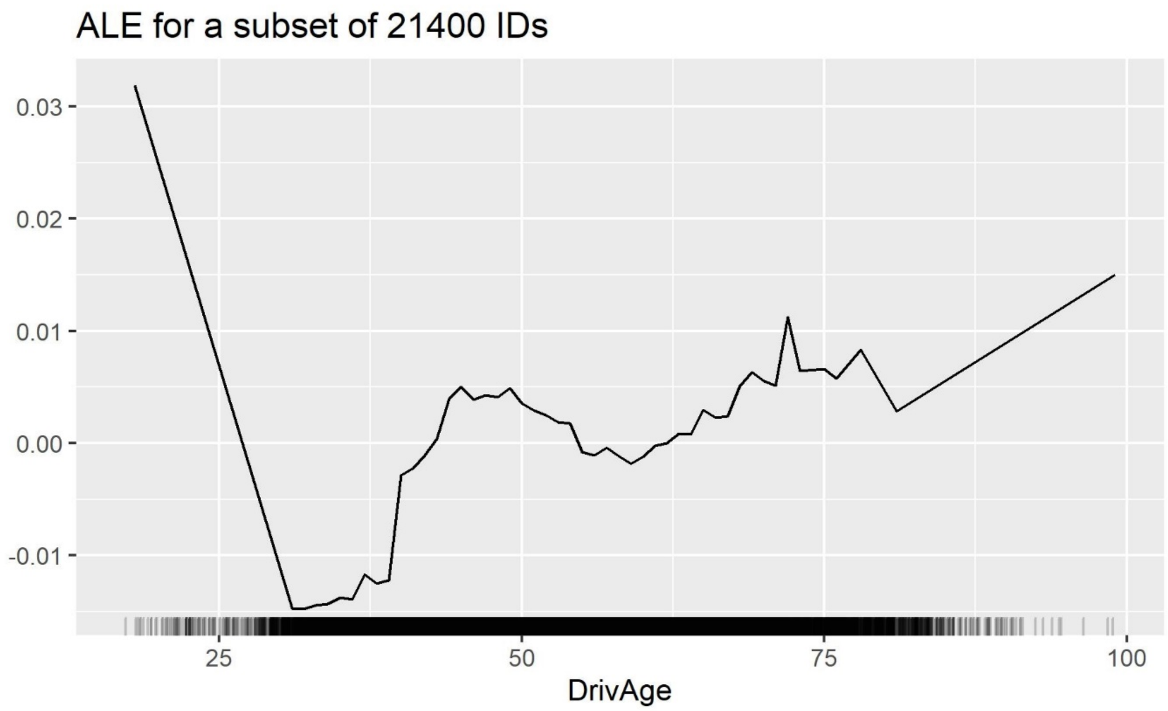
- Aggregated ICE output – model level explanation
- Shows a global influence/impact
- Other variants that handle correlation/dependence:
  - ALE plots / M-plots

The IDs chosen are for drivers with lowest Bonus Malus score and 5-year-old cars

Institute and Faculty of Actuaries

3. ALE – Accumulated Local Effects

- Variable level explanation
- Solution for strongly correlated features
- Shows models sensitivity to variation, for a specific value window of a feature
- Alternative: M-plots



The IDs chosen are for drivers with lowest Bonus Malus score and 5-year-old cars

Institute and Faculty of Actuaries

4. XAI in R

ICE, PDP, ALE with `library(iml)`

- A predictor object holding the model, data and a predictor function if needed
- iml produces ggplot objects
- Methods: ice, pdp, ale, pdp+ice
- Two-dimensional outputs supported
- iml CRAN vignette

```r
PR = iml::Predictor$new(model= xgb,
                        data = sample,
                        predict.function = f)

plt = iml::FeatureEffects$new(predictor = PR,
                              feature = "VehAge",
                              method = "ice",
                              grid.size = 32)+
ggtitle("Individual Conditional Expectation")
```

Institute and Faculty of Actuaries

5. SHAP – Shapley values

The set up:

1. A cooperative *game* involving two or more participants
2. Specified payout function – mapping (sub)sets of participants (called coalitions), to an expected numerical game outcome

Shapley values:

- Assign contributions to players in each coalition
- Distribute the obtained coalition payout as individual contributions among the players, so it sums up to the total amount

Note:

- Shapley values are *the only* way to distribute gains *fairly* among coalition members
- The formula below shows a contribution for player i out of N, in a game with payout v for possible coalitions S, including an empty coalition with payout $v(\emptyset) = 0$

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \underbrace{\frac{|S|!\,(n - |S| - 1)!}{n!}}_{weight} \left(v(S \cup \{i\}) - v(S)\right)$$

Institute and Faculty of Actuaries

5.1 SHAP – Adapting Shapley values to XAI

The set up

1. We treat input variables as participants in a game
2. The payout function is simply the black box model to be explained
3. As an explanation of models' prediction for a given observation, we obtain an additive decomposition of its output, distributed among the inputs

The problem

If our model is built on a set S of input variables, how to obtain an output for $v(S\backslash \mathrm{k})$ or even $v(\emptyset)$?

↓

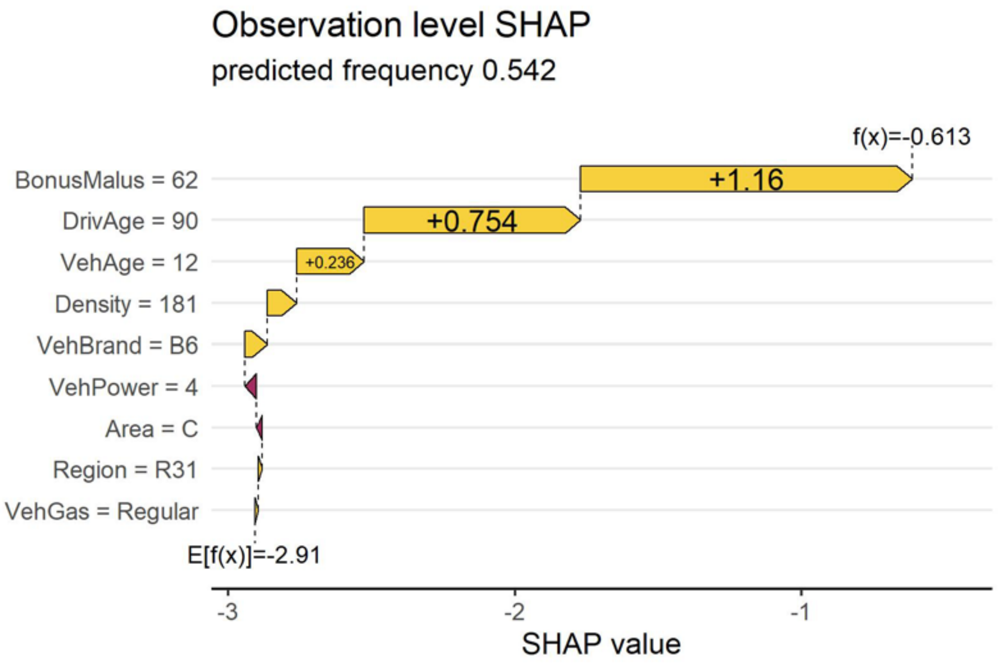
Remove the effect of feature k by averaging through its range of possible values.

Scott M. Lundberg, Su-In Lee – A Unified Approach to Interpreting Model Predictions

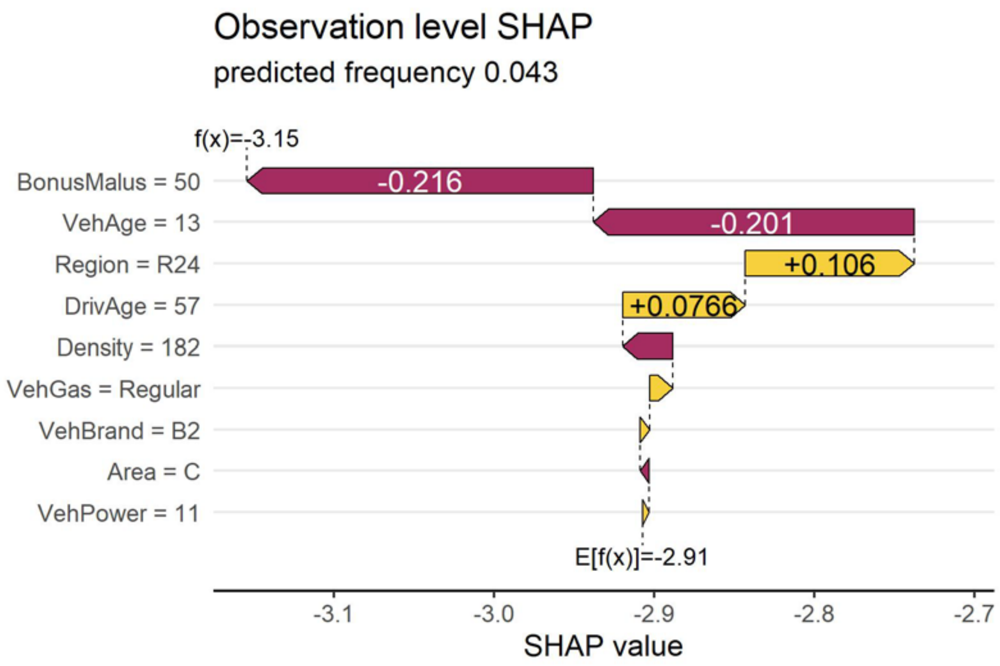
Institute and Faculty of Actuaries

5.2 SHAP – observation level explanation

Example A



Example B

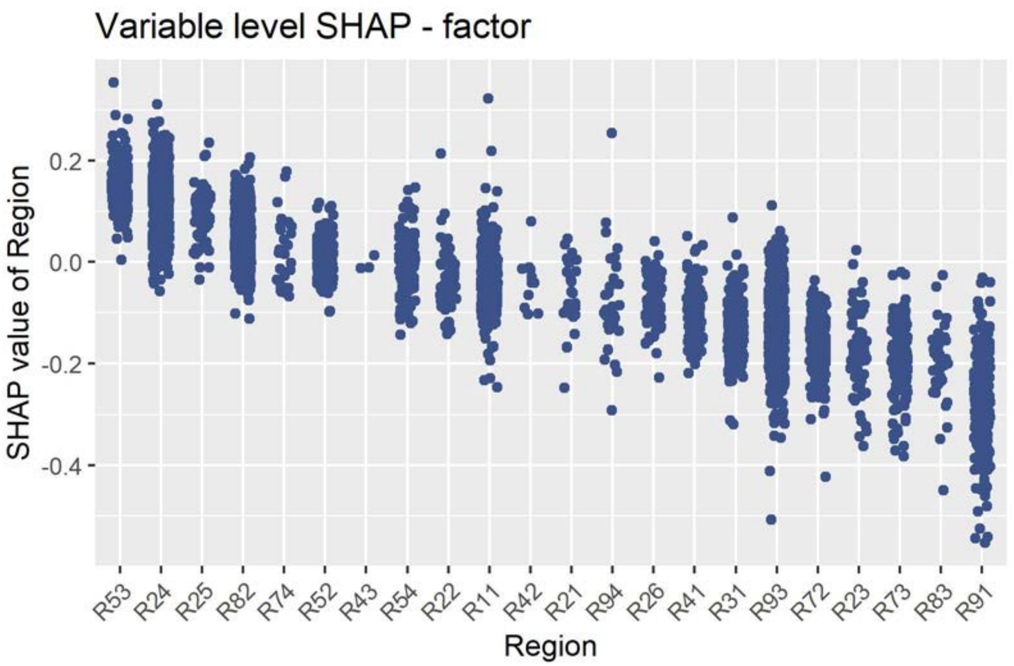


Note: The SHAP values shown are relativities, so the predicted frequency is $\exp(f(x))$

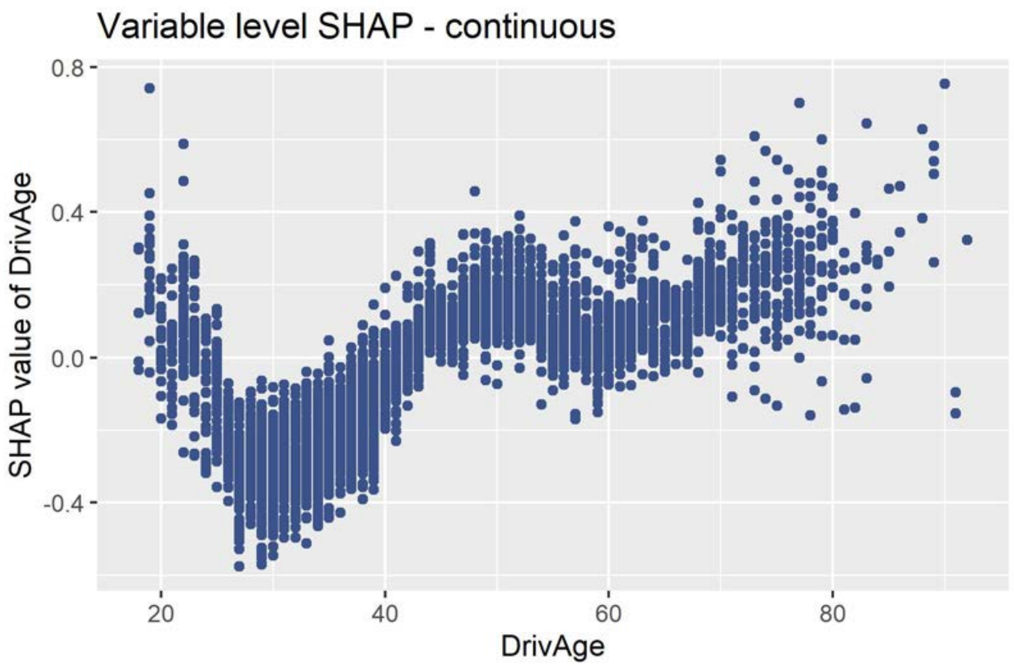
Institute and Faculty of Actuaries

5.3 SHAP – variable level explanation

Example A



Example B



Note: The SHAP values shown are relativities, so the predicted frequency is $\exp(f(x))$

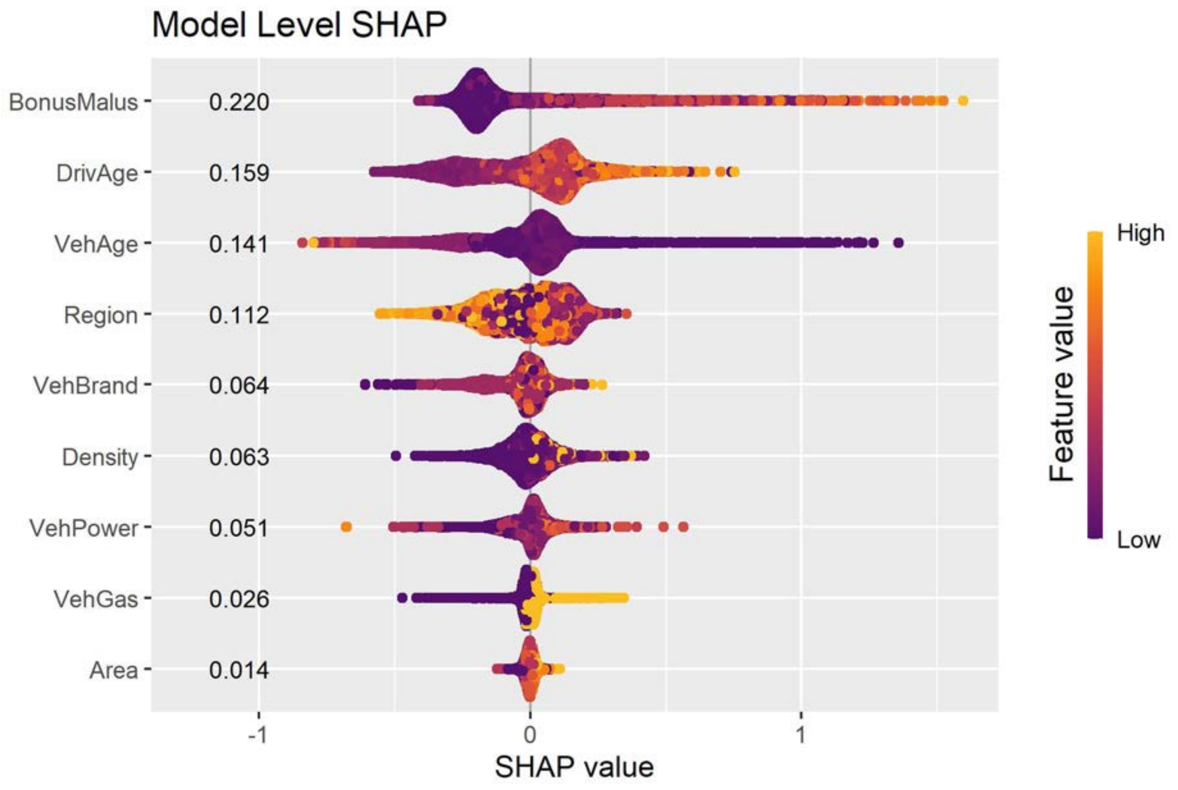
Institute and Faculty of Actuaries

5.4 SHAP – model level explanation

- High/low feature value can both increase/decrease models' output
- SHAP is model agnostic (e.g. kernel-SHAP)
- There are model specific SHAP algorithms that significantly speed up the calculations (e.g. Tree-SHAP)
- Some SHAP algorithms account for correlated variables (e.g. XGBoost or shapr)

More on kernel-SHAP:

The Actuary – All clear: How Shapley values make opaque models more transparent



Note: The SHAP values shown are relativities, so the predicted frequency is $\exp(f(x))$

Institute and Faculty of Actuaries

5.5 SHAP in R

Visualization with `library(shapviz)`

- A predictor object holding the model, data and a predictor function if needed
- shapviz produces ggplot objects
- Supports outputs of:
  - XGBoost
  - LightGBM
  - h2o
  - shapr
  - fastshap

```
viz = shapviz(object = xgb,
              X = data,
              X_pred = data %>% encode())

sv_waterfall(viz, row_id = 1) +
ggtitle("Observation Level Explanation")

sv_dependence(viz,v = "VehAge") +
ggtitle("Variable Level Explanation")

sv_importance(viz, kind = "beeswarm") +
ggtitle("Model Level Explanation")
```

Institute and Faculty of Actuaries

6. Further reading

- IFOA XAI WP
- Christoph Molnar – iml author
- Christian Lorentzen – shapviz author
- Scott Lundberg – SHAP author
- Przemyslaw Biecek / mi2 lab
- Explainable AI Methods - A Brief Overview
- xxAI – Beyond Explainable AI
- Explanatory Model Analysis

Institute
and Faculty
of Actuaries

https://ifoadatascienceresearch.github.io/

linkedin.com/in/karol-gawlowski/

github.com/karol-gawlowski

karol.gawlowski@bayes.city.ac.uk

#GiroConf22

Q&A

Institute
and Faculty
of Actuaries

Thank you

#GiroConf22