Explainable Machine Learning

EAA e-Conference on
Data Science & Data Ethics

29 June 2021

Education

- *Msc. Engineering (Applied Mathematics)*
- *Msc. Actuary: IA|BE qualified actuary (IA|BE = Institute of Actuaries in Belgium)*
- *Master in Management*

Function

- *CEO Reacfin*
- Expert in Non-Life and Health insurance (pricing, product development, reserving and risk management).

ABOUT ME

Xavier
Maréchal

Reacfin

Education

- *Msc. Engineering (Applied Mathematics)*
- *Msc. Actuary: IA|BE qualified actuary (IA|BE = Institute of Actuaries in Belgium)*

Function

- *Director – Head of Non-Life Reacfin*
- Expert in Non-Life and Health insurance (pricing, product development, reserving and risk management).

ABOUT ME

Samuel Mahy

Reacfin

GOALS OF THIS PRESENTATION

The problem

- Whereas advanced Machine learning techniques (e.g. random forest or neural networks) usually have a better predictive power than statistical techniques (e.g. GLM), their main drawback is that they are black-box and their results are difficult to **understand/interpret**

Two different strategies to use ML for practical applications

- There are basically 2 strategies to use ML techniques in predictive modelling
  1. **Replacing** traditional models (e.g. GLM) by ML models
  2. **Combining** the pros of traditional and ML models to improve predictive modelling

- The goals of this presentation are therefore to
  - Present several techniques that have been developed in order to **better understand the results** of machine learning techniques
  - Explain how these **interpretation techniques** can be used to implement the 2 strategies presented above and improve predictive modelling

Agenda

1. Adding complexity means increasing need for interpretability

2. An introduction to ML interpretation tools

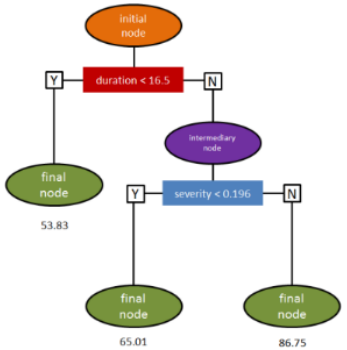
3. Conclusions: how to make the most of ML techniques

SOME MACHINE LEARNING TECHNIQUES ARE BLACK BOXES AND INTERPRETATION OF THE RESULTS CAN BE QUITE DIFFICULT

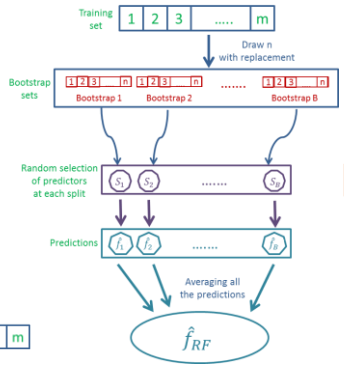
Increasing complexity to boost predictive power often means decreasing the interpretability of the results

Complexity

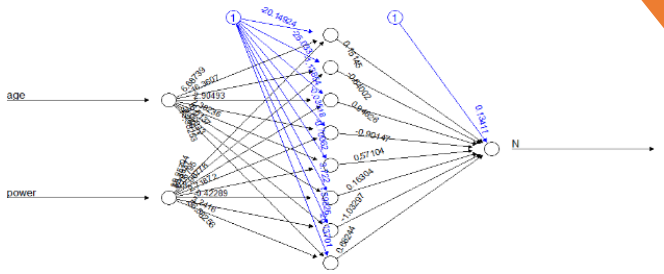
In the case of regression trees, understanding how the model predicts the target variable for new data points is not a problem, as it is very intuitive

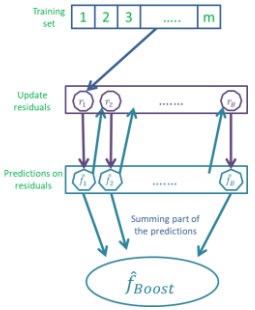


In the case of more complex methods such as Bagging and Random forests, even understanding how the model predicts values for new data points is rather difficult



Things may be even worse for GBM and NN





Interpretability

UNDERSTANDING THE RESULTS OF ML MODELS IS NEVERTHELESS KEY FOR SOUND BUSINESS DECISION-MAKING AS MANY STAKEHOLDERS USE THE RESULTS OF THE MODELS

Quant (Actuaries, data scientists,…)

- Are able to understand the technical details
- Trust its outputs based on cross-validation, error measures and assesment plots

Other stakeholders

- Not necessarily « quantitative people »
- Should nevertheless understand and trust results to take decisions

Machine learning techniques usually improve predictive power but at the expense of a certain loss of interpretability ➔ Find trade-off between

- Predictive power
- Capacity to understand the results
- Ability to take sound decisions based on the results

High-end questions

Who will use the results? For what purpose? With which impact?

Agenda

1. Adding complexity means increasing need for interpretability

2. An introduction to ML interpretation tools

3. Conclusions: how to make the most of ML techniques

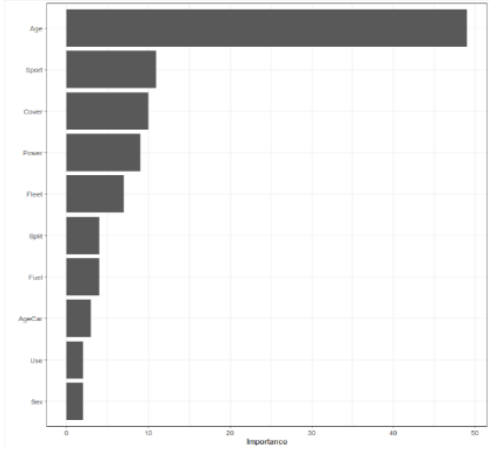
GLOBAL VS LOCAL INTERPRETABILITY OF ML TECHNIQUES

- **Global Model Interpretability**
  - How does the trained model make predictions?
    - ✓ Which features are **important** and what kind of **interactions** between them take place?
    - ✓ Global model interpretability helps to understand the **distribution of the target outcome based on the features**
    - ✓ Global model interpretability is very difficult to achieve in practice → Any model that exceeds a handful of parameters or weights is difficult to understand
    - ✓ Some models are interpretable at a parameter level :
      - For linear models, the interpretable parts are the weights,
      - For trees interpretable parts are the splits (selected features plus cut-off points) and leaf node predictions.

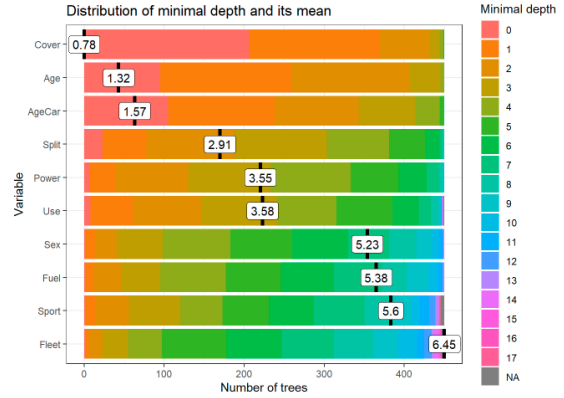
- Global Interpretation tools
  - ✓ Interpretable Models by nature (eg. Linear models, Regression Tree)
  - ✓ Feature Importance
  - ✓ Partial Dependant Plot (PDP) and Individual Conditional Expectation (ICE)
  - ✓ Interaction Measures (H-statistic).

GLOBAL MODEL INTERPRETATION FEATURES IMPORTANCE

- **Features Importance**
  - In a tree-based method : Go through all the splits for which the feature was used and measure how much it has reduced the Loss Function (eg. Gini, MSE, Poisson Deviance,…) compared to the parent node
  - The sum of all importance measures is scaled to 100
  - This means that each variable importance can be interpreted as share of the overall model importance



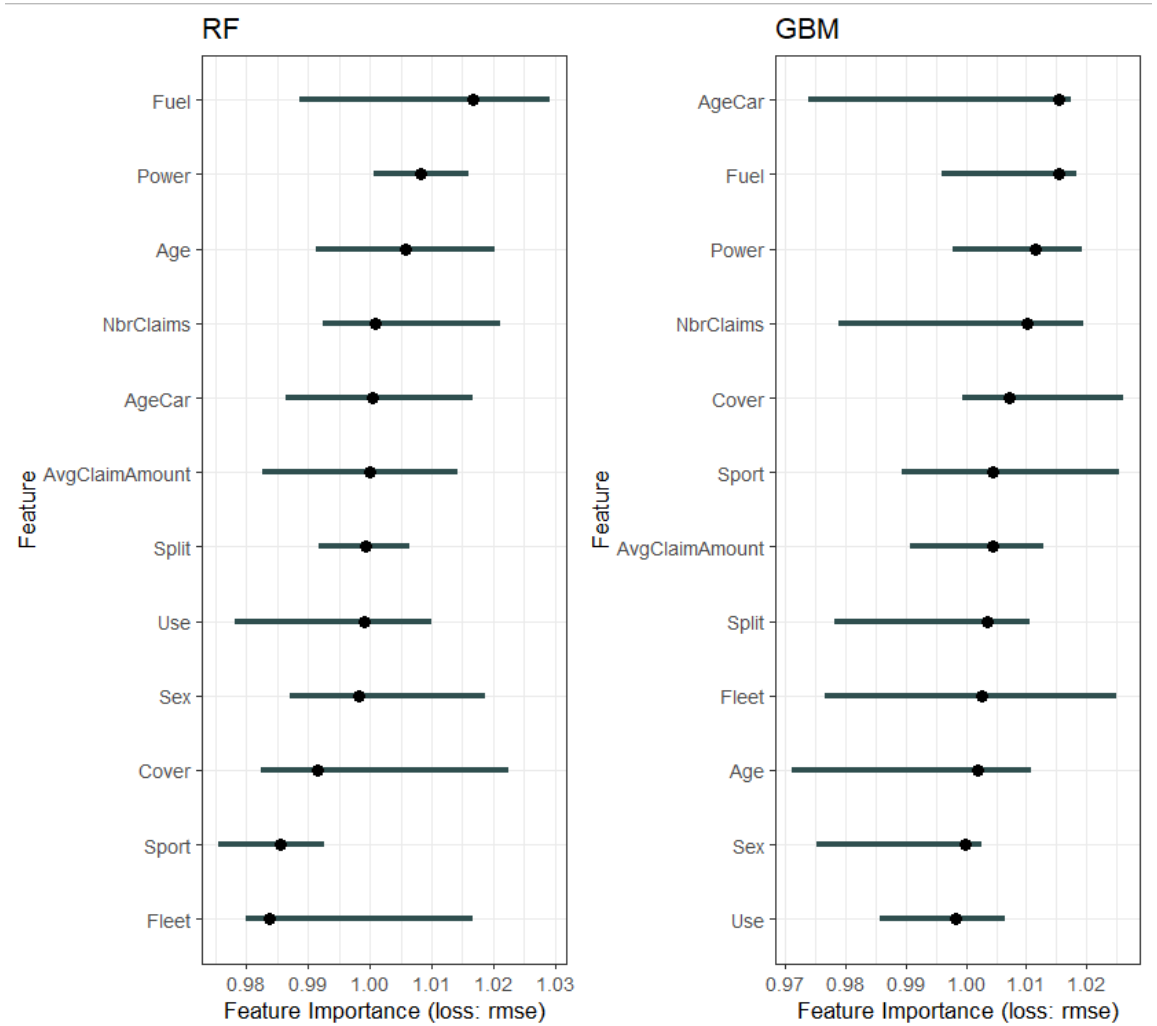
- One can get additional measures such as:
  - Minimal depth and its mean :
    - Which variables were the most often on the top of the tree
    - Mean depth of first split



- Features Importance can be used as a **features' selection tool**
  - Goal: Identify the **most relevant variables**
  - Pay attention: when some variables are correlated, their **global impact can be spread** between them, therefore reducing individual importance of each variable.

GLOBAL MODEL INTERPRETATION PERMUTATION IMPORTANCE

- **Permutation Importance**
  - Model agnostic → can be applied to all types of models !
  - Measure the importance of a feature by calculating the increase in the model's prediction error after permuting the feature.
  - A feature is "important" if shuffling its values increases the model error, because in this case the model relied on the feature for the prediction.
  - A feature is "unimportant" if shuffling its values leaves the model error unchanged, because in this case the model ignored the feature for the prediction.



GLOBAL MODEL INTERPRETATION
PARTIAL DEPENDENCE PLOT

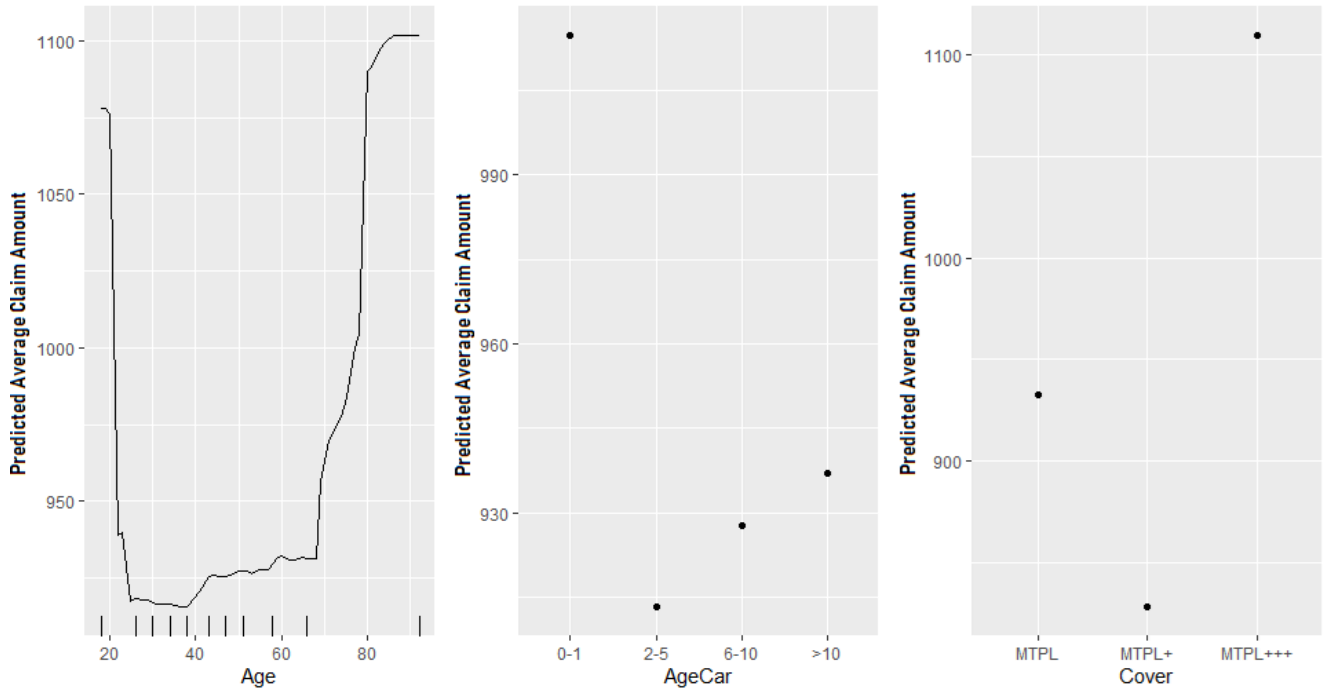
- **Partial Dependence Function/Plot**
  - Partial dependence plot (short PDP or PD plot) shows the **marginal effect one or two features** have on the predicted outcome of a machine learning model
  - Partial dependence plot can show whether the **relationship between the target and a feature** is linear, monotonic or more complex. It can be computed as

$$PD_{age}(age) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}\left(age, agecar^{i}, cover^{i}, \dots\right)$$

  - In this formula, $agecar^{i}$,… are actual features' values from the dataset for the features in which we are not interested, $\hat{f}$ is the trained model and n is the number of instances in the dataset
  - So we marginalize model outputs over the distribution of the features we are not interested in (e.g. agecar, cover, …)
    - → the function shows the relationship between the feature age we are interested in and the predicted outcome
    - → By marginalizing over the other features, we get a function that depends only on features age, interactions with other features included.

GLOBAL MODEL INTERPRETATION PARTIAL DEPENDENCE PLOT

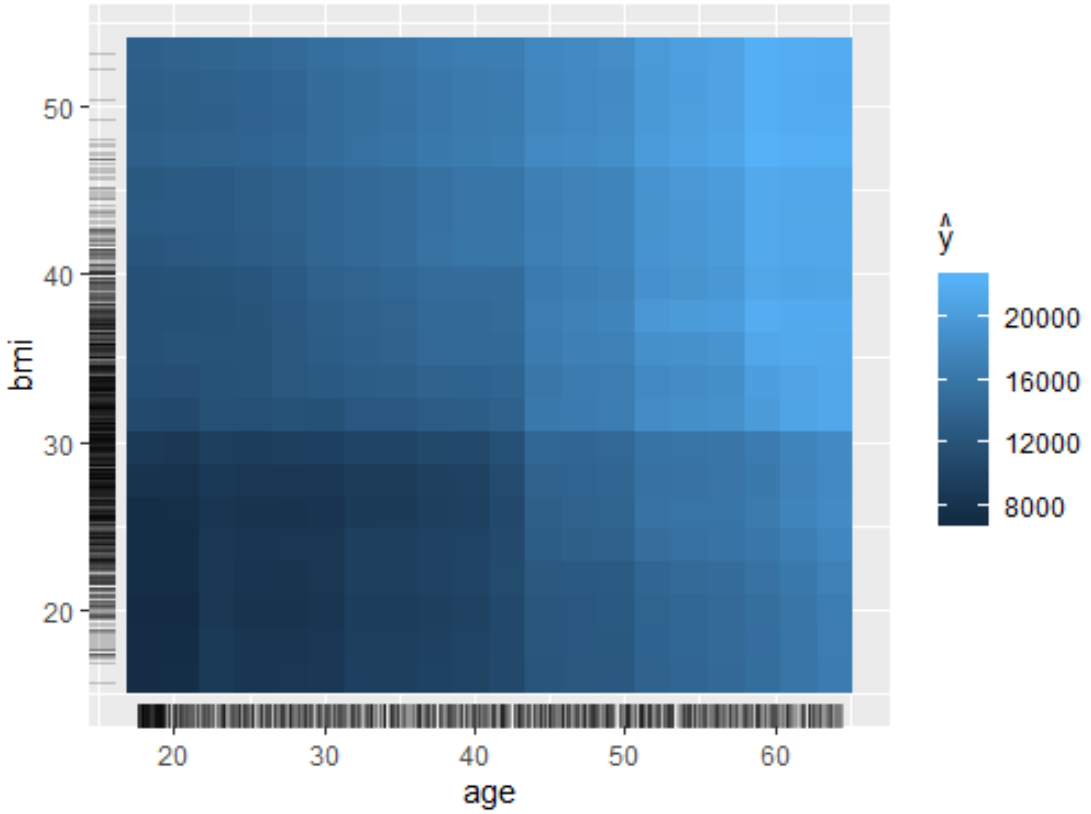
- **Example of Partial Dependence Plot (1D) on Average Claim Amount :**



- Partial dependence plot can be used as **a features' impact explanation tool**
  - It allows to better understand the marginal impact of a variable on the prediction
  - It is very similar to the interpretation of the multiplicative factors we obtain in a GLM or GAM model.

GLOBAL MODEL INTERPRETATION
PARTIAL DEPENDENCE PLOT

- **Example of Partial Dependence Plot (2D) :**
  - PD can be generalized to more than one feature
  - PDP -2D can be very useful to highlight interactions.



GLOBAL MODEL INTERPRETATION BASIC EXAMPLE WITH ONLY TWO FEATURES

Original data features		
Age of the driver	License Age	Predicted Frequency
20	1	6,2%
35	10	5,1%
20	3	5,5%
55	32	4,2%
60	40	4,3%

PDP Age of the driver		
Age of the driver	License Age	Predicted Frequency
20	1	6,2%
20	10	3,0%
20	3	5,5%
20	32	1,0%
20	40	0,5%
	PDP Age(20)=	**3,2%**

Partial Depend Plot

- PDP compute what the model predicts on average when **each data** instance has the value 20 for driver age.
- Weird instance are created during the calculation process (see yellow rows)
- **Marginal distribution** is used so **all instance** in the data set enter in the calculation for each driver age computation.
- Computation time can be huge with large dataset.

GLOBAL MODEL INTERPRETATION PARTIAL DEPENDENCE PLOT

- **Attention point with Partial Dependence Plot**
  - Correlated features :
    - ✓ With correlated features, computation of a PDP involves averaging predictions of artificial data instances that can be unlikely in reality.
    - ✓ E.g. “Age of the driver” and “License Age” in motor insurance : we don’t expect a 20 years old policyholder with 10 years of license whereas PDP computation process will consider this type of instance…
  - 1D Flat PDP does not imply that the feature has no influence!
    - ✓ Interaction effect might still be there
    - ✓ E.g. half of the instance have a positive impact on the prediction and the other half has a negative impact. Both effects could cancel each other in the PDP.
    - ✓ These interactions effects can be observed in Individual Conditional Expectations (ICE – see further)

Nice alternative to PDP are Accumulated local effect plot (ALE)

GLOBAL MODEL INTERPRETATION EXAMPLE

Original data features		
Age of the driver	License Age	Predicted Frequency
20	1	6,2%
35	10	5,1%
20	3	5,5%
55	32	4,2%
60	40	4,3%

M-Plot Age of the driver		
Age of the driver	License Age	Predicted Frequency
20	1	6,2%
35	10	5,1%
20	3	5,5%
55	32	4,2%
60	40	4,3%
	M-Plot Age(20)=	**5,9%**

M-Plot

- M-plot (marginal plot) computes what the model predicts on average for policyholders that are **close to 20** years old.
- **Conditional distribution** is used so only instance where driver age is close to 20 are used in the calculation
- **Attention point:** The effect observed in M-Plot could be due to that feature, but also due to another correlated features (like License Age in our example)

ALE-Plot

- Based on **Conditional distribution** (like M-Plot) but use **the sum incremental effects** of the feature of interest in order to avoid effects of correlated features.
- Calculation out-of-the scope of this presentation see *(Daniel W. Apley and Jingyu Zhu 2019*)

GLOBAL MODEL INTERPRETATION
INDIVIDUAL CONDITIONAL EXPECTATION

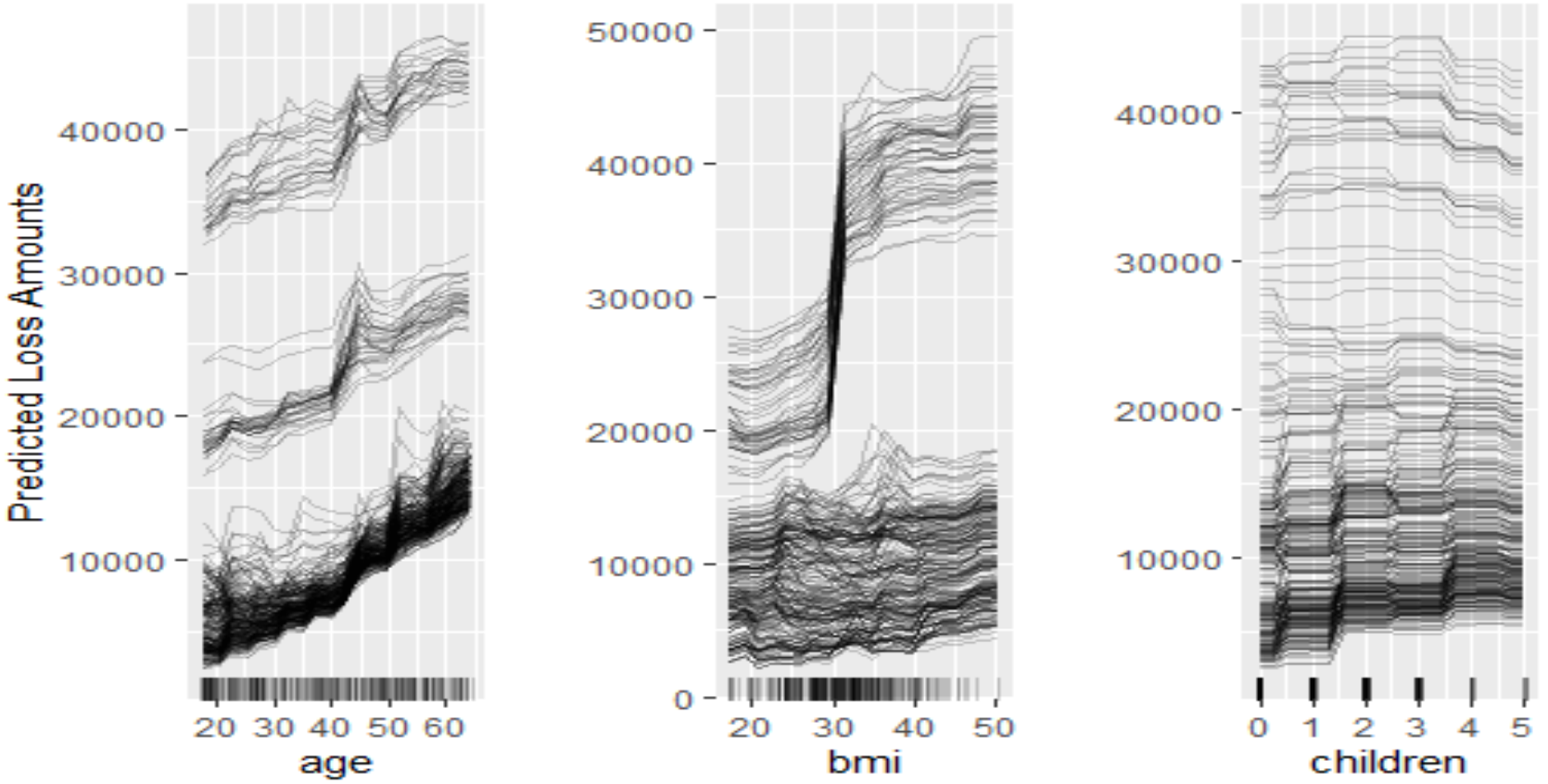
- **Individual Conditional Expectation (ICE) :**
  - One line per instance that shows how the instance's prediction changes when a feature changes
  - An ICE plot visualizes the dependence of the prediction on a feature for each instance separately → one line per instance compared to one line overall in PDP.
  - A PDP is the average of the lines of an ICE plot.

  - Advantage over PDP :
    - ✓ In case of interactions, the ICE plot will provide much more insight.

  - How to compute ICE ?
    - ✓ Creating variants of an observation by replacing the feature of interest value with values from a grid
    - ✓ Keeping all other features the same
    - ✓ Make predictions with the black box model for these newly created observations.
    - ✓ The result is a set of points for an original observation with the feature value from the grid and the respective predictions.

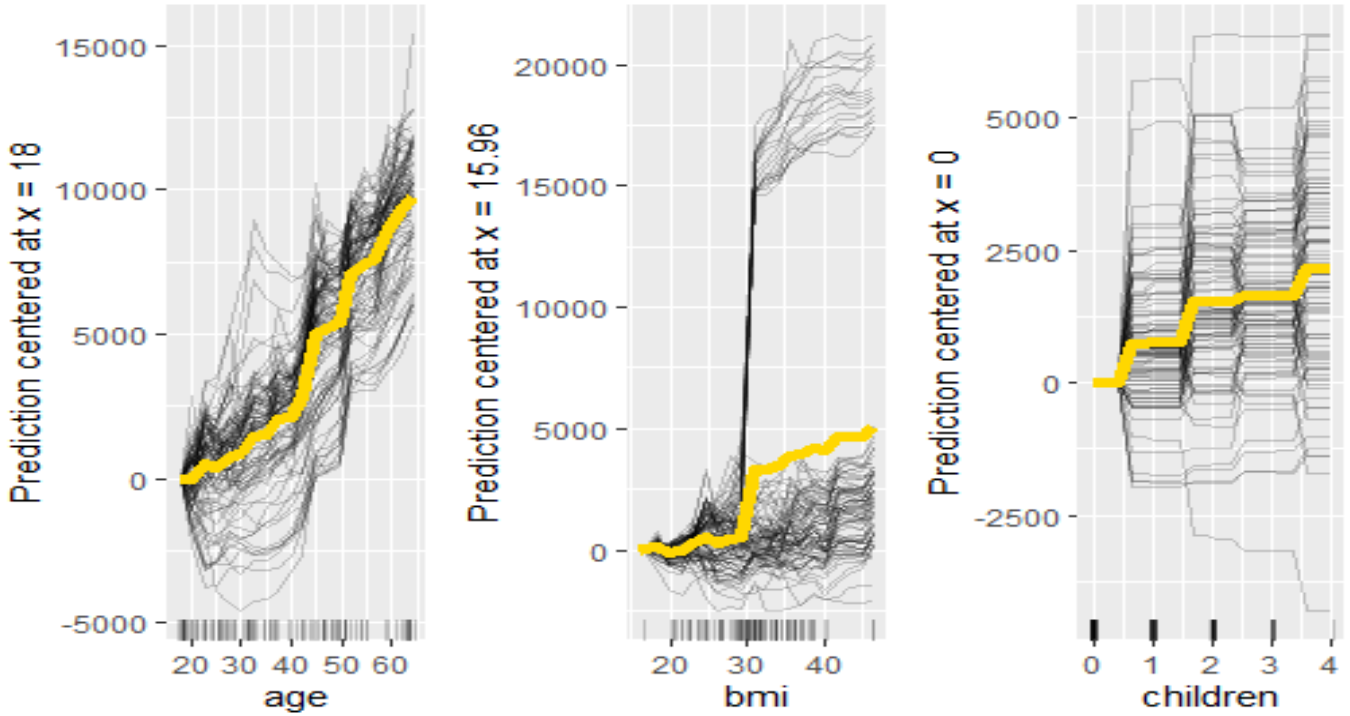
GLOBAL MODEL INTERPRETATION
INDIVIDUAL CONDITIONAL EXPECTATION

- **Individual Conditional Expectation (ICE) :**
  - Do you notice the interaction?



GLOBAL MODEL INTERPRETATION CENTERED ICE

- **Individual Conditional Expectation (ICE) :**
  - It can be hard to tell whether the ICE curves differ between individuals because they start at different predictions.
  - A simple solution is to center the curves at a certain point of the feature and display only the difference in the prediction to this point.



GLOBAL MODEL INTERPRETATION
DETECTION OF INTERACTION BETWEEN VARIABLES WITH H-STATISTICS

- **Interaction Measures (H-Statistics)**
  - In case of interaction, prediction cannot be expressed as the sum of the feature effects, because the effect of one feature depends on the value of the other feature
  - **How to measure the level of interaction between two features?**

  → Have a look at H-Statistic. The main idea is:

  ✓ If two features do not interact, we can decompose the partial dependence function

  $$PD_{age,power}(age, power) = PD_{age}(age) + PD_{power}(power)$$

  ✓ Measure the difference between the observed partial dependence function and the decomposed one without interactions.

  $$H^2 = \sum_{i=1}^{n}\left[PD_{age,power}(age^i, power^i) - PD_{age}(age^i) - PD_{power}(power^i)\right]^2 \Big/ \sum_{i=1}^{n} PD_{age,power}{}^2(age^i, power^i)$$

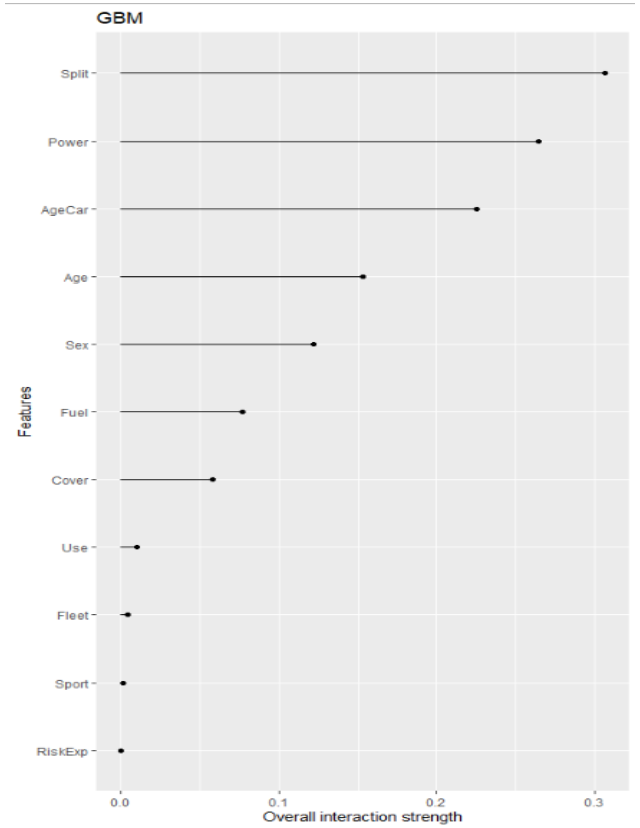
  ✓ *H is 0 if there is no interaction at all*

  ✓ *A value H of 1 between two features means that each single PD function is constant and the effect on the prediction only comes through the interaction.*

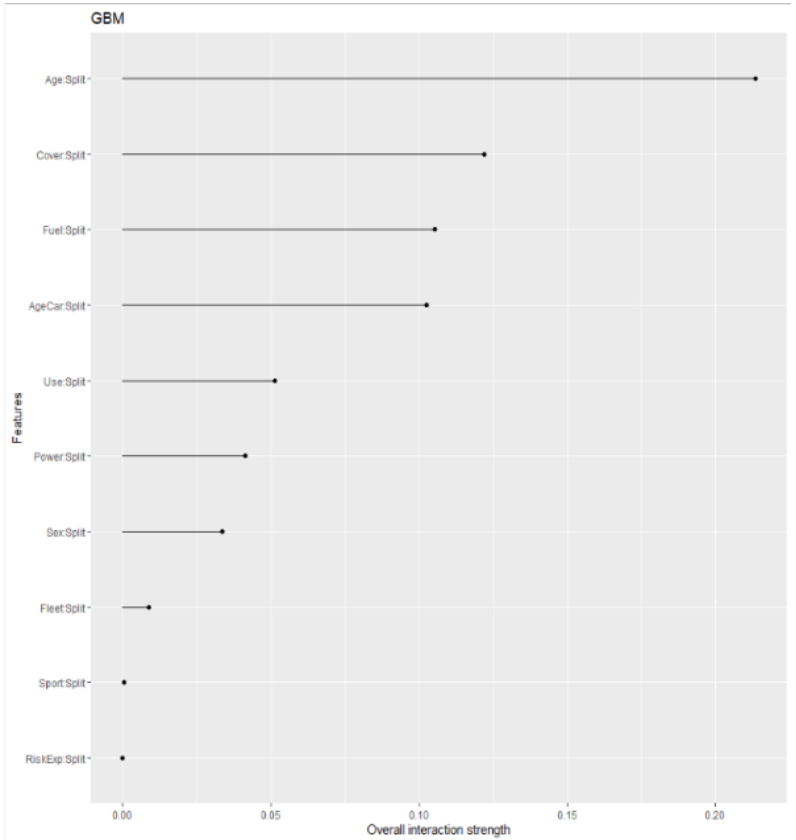
  - It is also possible to measure the **total interaction** of a feature which tells us **whether and to what extent a feature interacts** in the model **with all other features.**

GLOBAL MODEL INTERPRETATION
DETECTION OF INTERACTION BETWEEN VARIABLES WITH H-STATISTICS

Total interaction for each feature with all other features



2-way interactions between the split feature and the other features



- H-Statistics can be used as a **features' interaction identification tool**
  - It allows to identify features strongly interacting with other features
  - It can then be used for **features engineering** (e.g. creating a new feature as an interaction between 2 features).

GLOBAL VS LOCAL INTERPRETABILITY OF ML TECHNIQUES

- **Local Interpretability for a Single Prediction**
  - Why did the model make a certain prediction for an instance?
    - ✓ If you look at an individual prediction, the behavior of the otherwise complex model might behave more pleasantly
    - ✓ You can zoom in on a single instance and examine what the model predicts for this input, and explain why
      - Shapley Value
      - Breakdown.

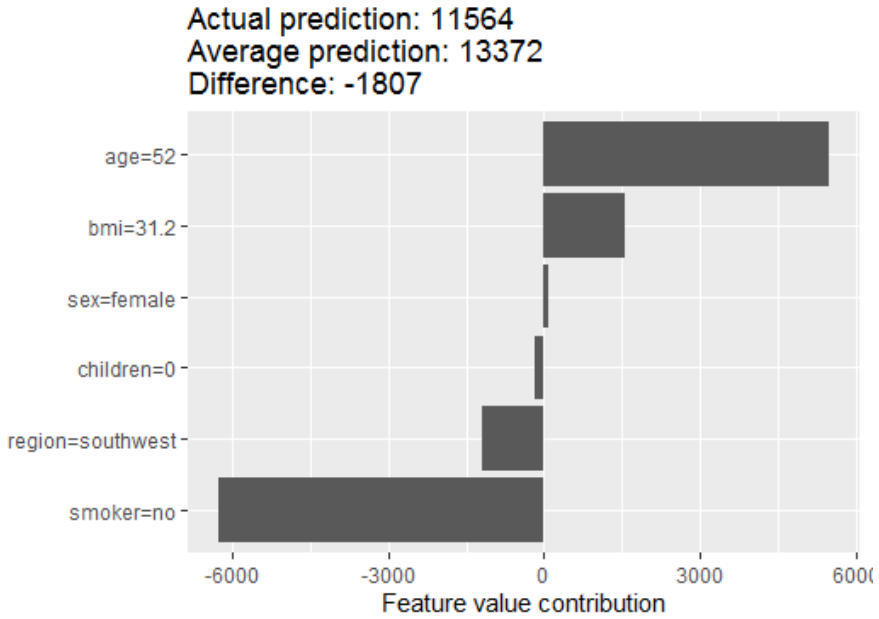
- **Local Interpretability for a Group of Predictions**
  - Why did the model make specific predictions for a group of instances?
    - ✓ Model predictions for multiple instances can be explained either with global model interpretation methods or with explanations of individual instances
    - ✓ The global methods can be applied by taking the group of instances, treating them as if the group were the complete dataset, and using the global methods with this subset
      - LIME (Local Interpretable Model-agnostic explanations)
      - LIVE.
    - ✓ The individual explanation methods can be used on each instance and then listed or aggregated for the entire group.

LOCAL INTERPRETABILITY FOR A SINGLE PREDICTION

- **Shapley Value :**
  - The shapley value measures for a single prediction how much each specific feature value will contribue to make the instance prediction different from the overall prediction
  - The computation time increases exponentially with the number of features.

From Game Theory

- *The Shapley value is the average marginal contribution of a feature value across all possible coalitions (= sets composed of different number of features).*
- *For each of these coalitions we compute the prediction with and without the feature value of interest and take the difference to get the marginal contribution.*
- *The Shapley value is the (weighted) average of marginal contributions across all the coalitions.*



Agenda

1. Adding complexity means increasing need for interpretability

2. An introduction to ML interpretation tools

3. Conclusions: how to make the most of ML techniques

HOW TO MAKE THE MOST OF ML TECHNIQUES?

Two different strategies

1. Replacing traditional models (e.g. GLM) by ML models
2. Combining the pros of traditional and ML models to improve predictive modelling.

Replacing traditional models by ML models

- The main drawback of this approach is the black-box effect of the ML results
- There is therefore a strong need in using interpretations tools
  - **Feature importance** to select the most relevant variable (e.g. if we have too many variables available and/or we want to limit the number of modelled variables)
  - **PDP and/or H-Statistics** to understand the impact of the selected variables on the prediction and identify the potential interactions
  - **Shapley value** to better understand the prediction of specific profiles.

Combining traditional and ML models

- ML methods would then be used to perform **features extraction, features selection and/or features engineering**
  - **Feature extraction** = reducing the dimensionality of too voluminous datasets (in terms of # features)
  - **Feature selection** = selecting the most relevant variables to our problem
  - **Feature engineering** = identifying the best representation of the sample data to learn a solution to your problem (e.g. interactions).
- The selected/engineered variables could then be **introduced in a simpler model (e.g. GLM) in order to obtain easily interpretable results.**

EAA e-Conference on
Data Science & Data Ethics

29 June 2021

Contact

Samuel Mahy *& Xavier Maréchal*

samuel.mahy@reacfin.com & *xavier.marechal@reacfin.com*

+32 498 04 23 90 *& +32 497 48 98 48*