

Machine-Learning Approaches to Survival Analysis of Clinical Data

Phuse EU Connect Birmingham

06 Nov 2023

The Speakers

Dorothee Childs



Senior Data Scientist,
Project Manager

Christoph Bergen



Senior Data Scientist,
Project Manager

@ HMS Analytical Software

HMS Analytical Software



1989 established

150+ employees

TURNING DATA INTO VALUE

HMS Analytical Software GmbH is providing highly specialized services to (regulated) industries in the area of data science and software engineering.

Setting the scene

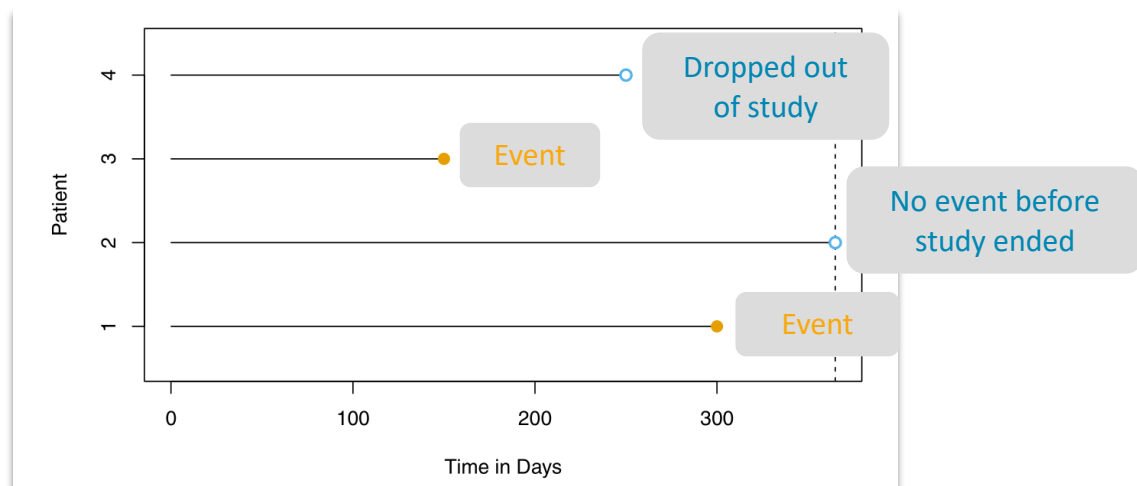
We aim to introduce a machine learning approach for modeling time-to-event data.

For instance, this allows **predicting** time-to-event endpoints, for individuals right from the baseline.

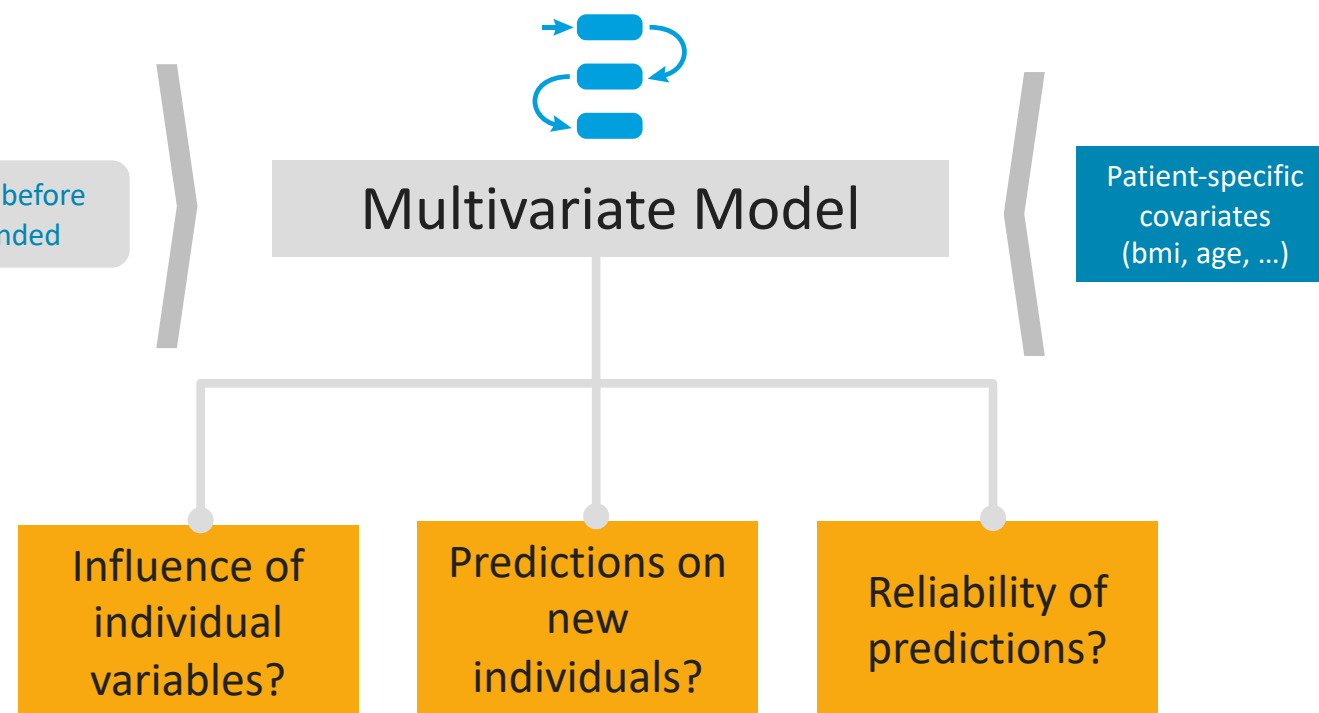
You will learn:

1. Basic overview of the setting, data, and established statistical methodology
2. Machine Learning approaches for time-to-event data
3. A Machine Learning framework in python
4. The Machine Learning model zoo
5. Explainable AI techniques to elucidate black-box models.

Modelling time-to-event data

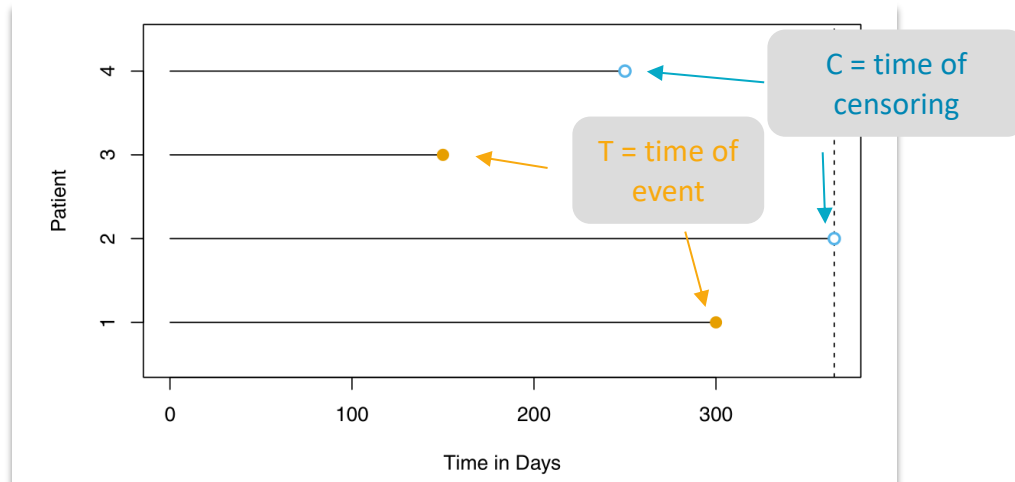


^[1]James et al. (2023). An Introduction to Statistical Learning with Applications in Python. Springer Texts in Statistics (STS)



Classical statistical modelling approaches

Cox Proportional Hazard Regression



^[1]James et al. (2023). An Introduction to Statistical Learning with Applications in Python. Springer Texts in Statistics (STS)

Observations:

$$Y = \min(T, C)$$

Model^[1]:

$$h(t|x_i) = h_0(t) \exp \left(\sum_{j=1}^p x_{ij} \beta_j \right)$$

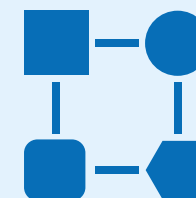
- $h(t|x_i)$: hazard at time t
- $h_0(t)$ baseline hazard
- x_{ij} : covariates (e.g. bmi, age in the medical context)
- β_j : coefficients
- Analysis often focused on hazard ratios without explicitly estimating h_0

What is the idea?

Classical models like Cox proportional hazard aims at **interpretability**. Models are fitted to understand the effect of covariates on the hazard.

ML models:

- **Aim to predict** outcomes for individual data instances
- Offer high flexibility
- Can fit non-linear effects in data
- Often viewed as "black box" models
- Interpretability can be challenging or limited



ML models do not aim at replacing tried and tested statistical methods. Instead they look at the data from a different angle:

- Check predictive capability of features at baseline to discover unknown patterns
- Derive new insights which can be utilized in future trials (exploratory approach)

Machine learning approaches for time-to-event data

- Time-to-event data affects model training and evaluation
- Applying standard ML regression models would lead to biases
- The scikit-survival^[1] package implements customized adaptations of classical ML algorithms for time-to-event data and is fully integrated into the established scikit-learn^[2] pipeline



THE Standard ML Framework in Python

+ Survival Analysis = **scikit-survival**

The sksurv model zoo

Penalized
linear ML
models:

CoxPHSurvivalAnalysis(alpha) -> alpha = Ridge / L^2 penalization parameter
CoxnetSurvivalAnalysis(l1_ratio=1.0) -> LASSO, full L^1 penalization
CoxnetSurvivalAnalysis(l1_ratio=0.8) -> l1_ratio controls the ratio between L^1 and L^2 penalization.

Tree-based
learners:

RandomSurvivalForest() -> Analog to RandomForest, uses different criteria to grow trees

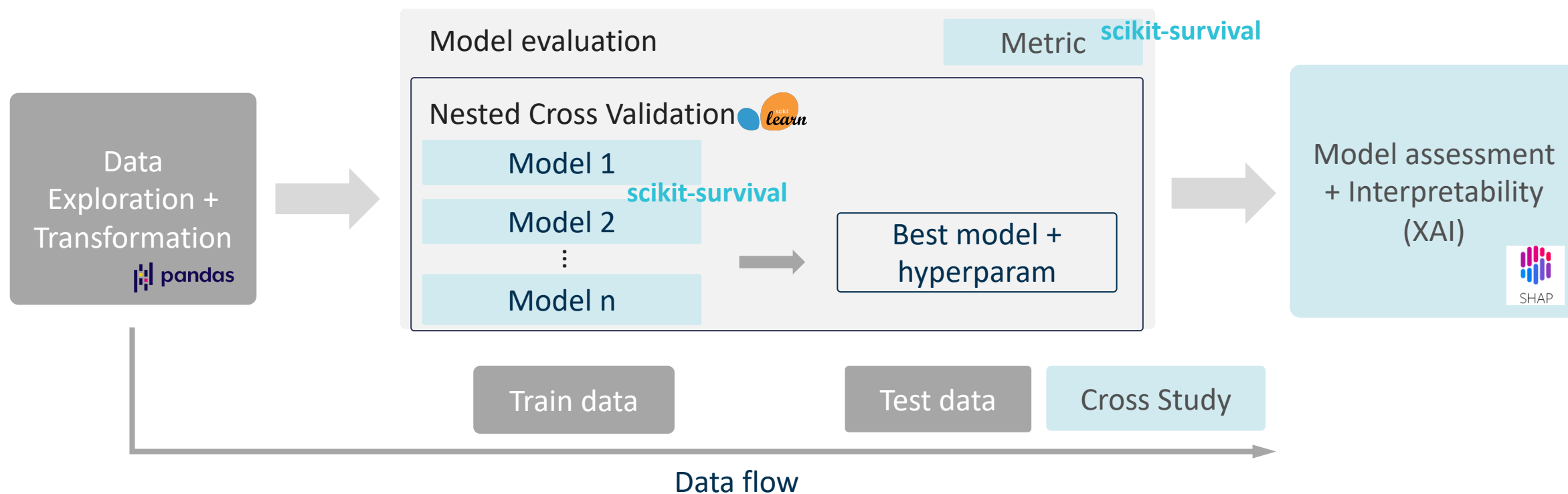
Boosting:

GradientBoostingSurvivalAnalysis() -> Trees as base learner
ComponentwiseGradientSurvivalAnalysis() -> Component-wise least squares as base learners
xgboost("objective": "survival:cox") -> Standard XGB (does not work with sklearn API out-of-the-box)

Kernel:

FastKernelSurvivalSVM()

Looking into the different ML components



The concordance index (Harrells C-index)

- Typical ML metrics like AUC/ RMSE (classification/ regression) not applicable for censored data
- Harrell's C-index:

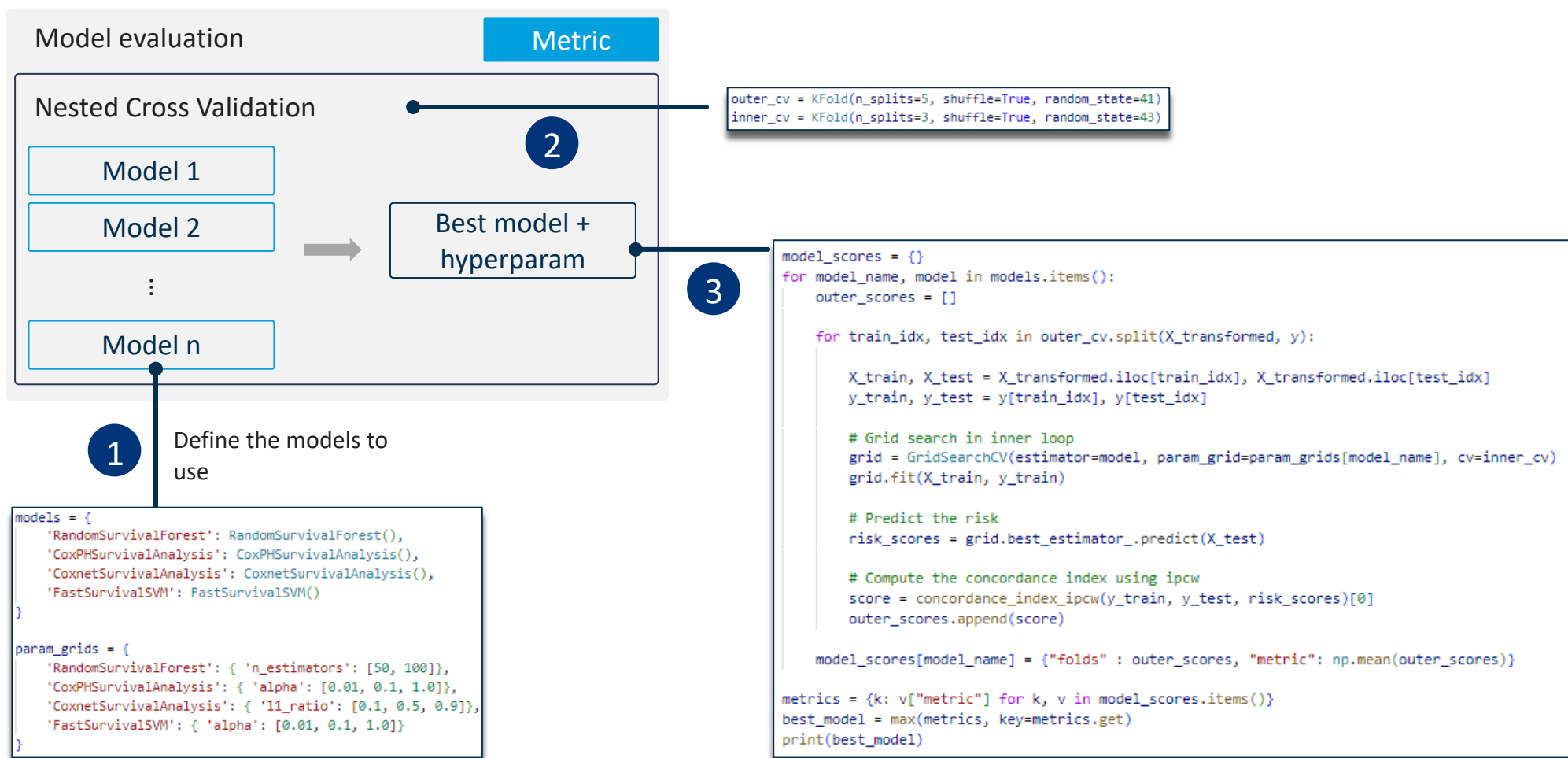
$$\text{C-index} = \frac{\sum_{i,j} 1_{T_j < T_i} \cdot 1_{\eta_j > \eta_i} \cdot \delta_j}{\sum_{i,j} 1_{T_j < T_i} \cdot \delta_j}$$



Similar interpretation as AUC: C-index = 1 -> optimal model;
C-index=0.5 -> random model

- Consider all possible pairs of individuals in your dataset.
 - Case 1: event (e.g., disease progression) combined with censoring -> not informative, ignore.
 - Case 2: two events, higher risk at individual with first event -> pair is concordant.
 - Case 3: two events, higher risk at individual with later event -> pair is discordant.
 - Case 4: equal risk to both -> pair is tied.

Sksurv in action – Survival ML modelling with a few commands



Explainable Machine Learning

We want to **explain both the model** (which variables are important for the model globally) and the **individual prediction** (what is the contribution of a variable value to the predicted value)

Shapley Approach/SHAP package:

- Based on Shapley values from cooperative game theory.
- Interprets the output of machine learning models.
 - Model agnostic, needs access to the model object
 - Can be computational expensive
- Allocates contribution of each feature to a specific prediction.
- Guarantees a fair distribution of feature contributions based on certain axioms.

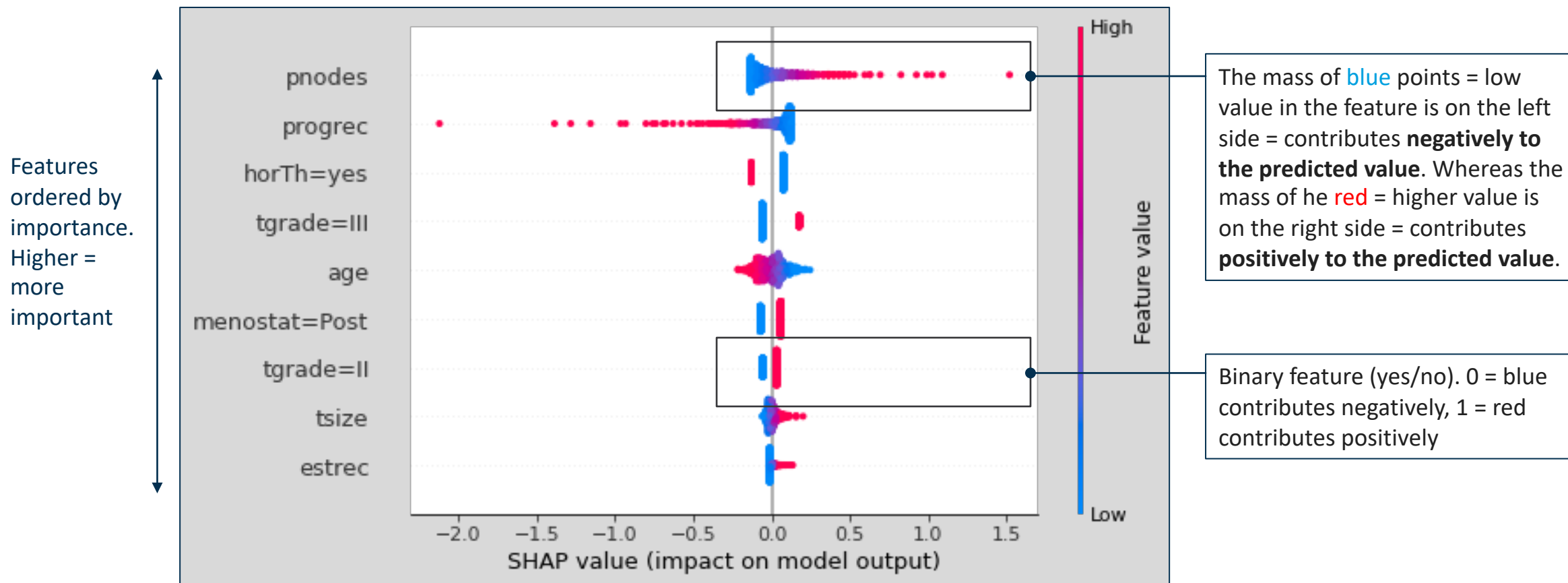


<https://shap.readthedocs.io/en/latest/>

➤ SHAP develops to be the “gold” standard for XAI with tabular data

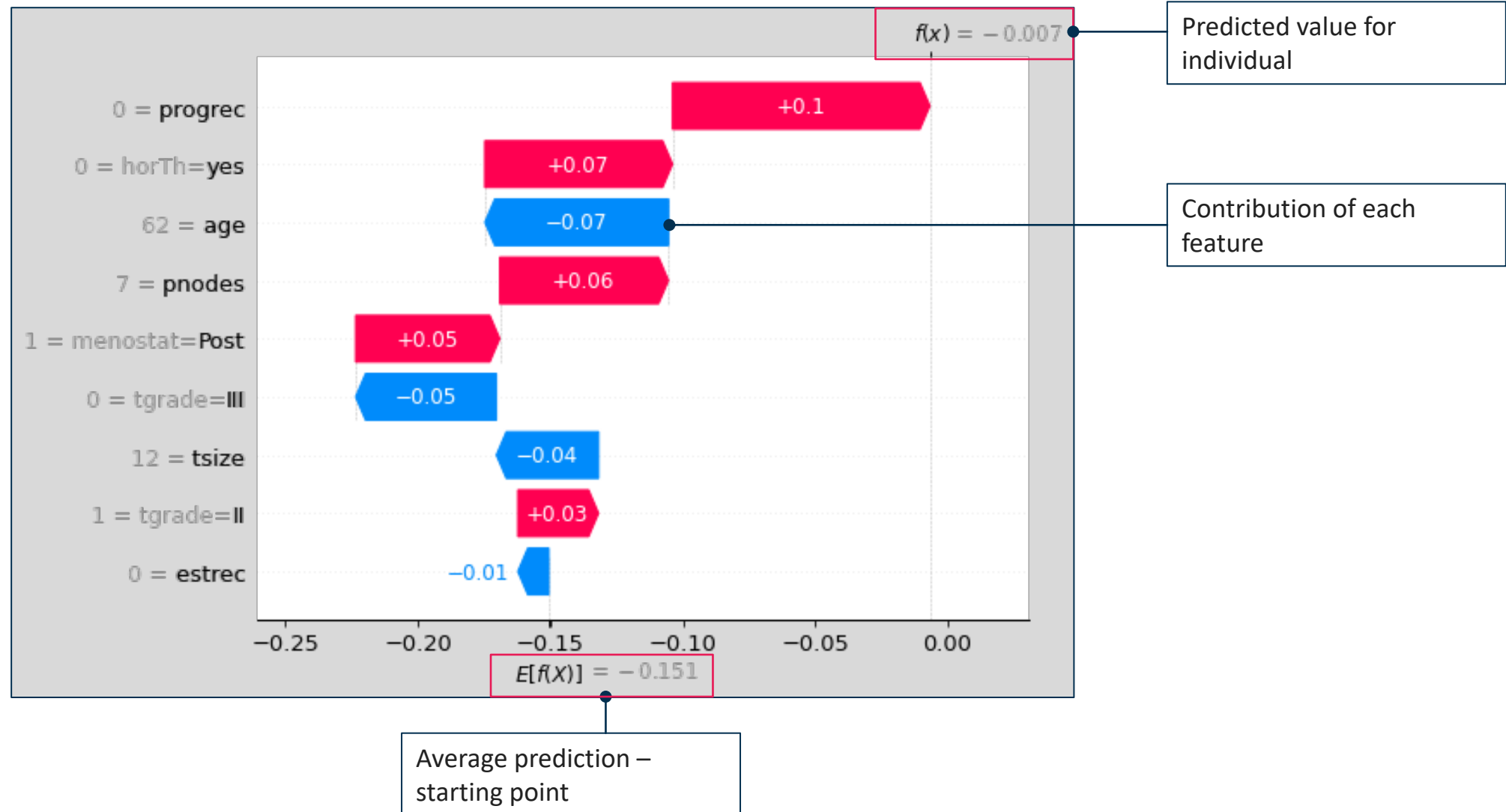
Explainable AI (XAI) with SHAP - Beeswarm

The contribution of a feature for each individual prediction is plotted (= one dot)



Explainable AI (XAI) with SHAP - Waterfall

Explanation of a individual prediction



Thank you very much for your attention!

Contacts:



Dorothee Childs
Dorothee.Childs@analytical-software.de



Christoph Bergen
Christoph.Bergen@analytical-software.de