MI²DataLab Seminar

Extensions of the SHAP method:

using Shapley value to interpret models with dependencies and survival analysis models

Mateusz Krzyziński

May 9th, 2022

Outline

1. SHAP

Lundberg, Lee. *A unified approach to interpreting model predictions*
(NeurIPS 2017)

a. **overview of the method**
b. **why is it (considered) so good?**

2. Shapley Flow

Wang, Wiens, Lundberg. Shapley Flow: A Graph-based Approach to Interpreting Model Predictions
(AISTATS 2021)

a. **what was the motivation?**
b. **overview of the method**

3. survSHAP

to be prepared
(???)

a. **a brief introduction to XAI in survival analysis domain**
b. **presentation of the method**

SHAP
theory

- **SHAP** = **SH**apley **A**dditive ex**P**lanations
- Shapley value (Lloyd Shapley, 1953)
- Additive feature attribution method

$$g(z') = \phi_0 + \sum_{i=1}^{p} \phi_i z'_i$$

- Contribution of a variable *i*:

$$\phi(i) = \sum_{S \subseteq \{1,\dots,p\}/\{i\}} \frac{|S|!(p-1-|S|)!}{p!} \left(e_{S\cup\{i\}} - e_S\right)$$

subset of variable indexes

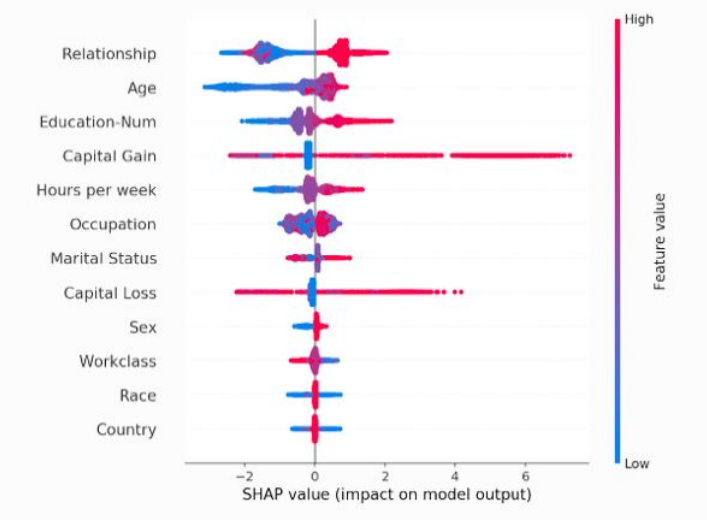
expected value for a conditional distribution

$$e_S = E[f(x)|x_S = x^*_S]$$

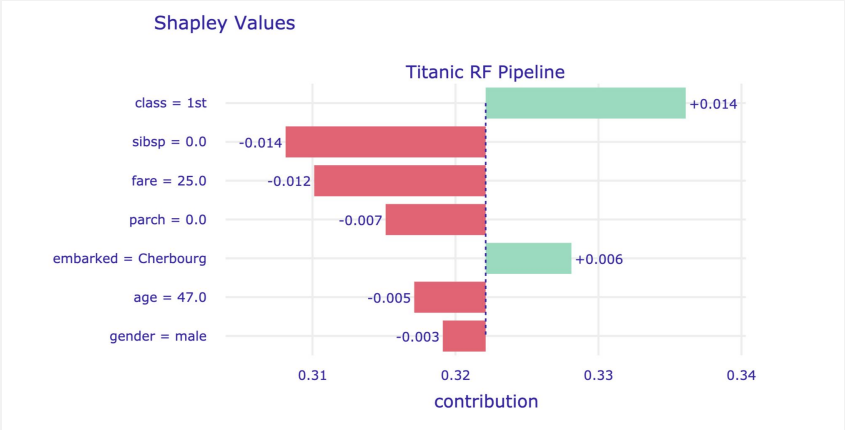
References:

- Lundberg, Lee. *A unified approach to interpreting model predictions* (2017)
- Štrumbelj, Kononenko. *An efficient explanation of individual classifications using game theory* (2010)
- Štrumbelj, Kononenko. *Explaining prediction models and individual predictions with feature contributions* (2014)

SHAP
practice

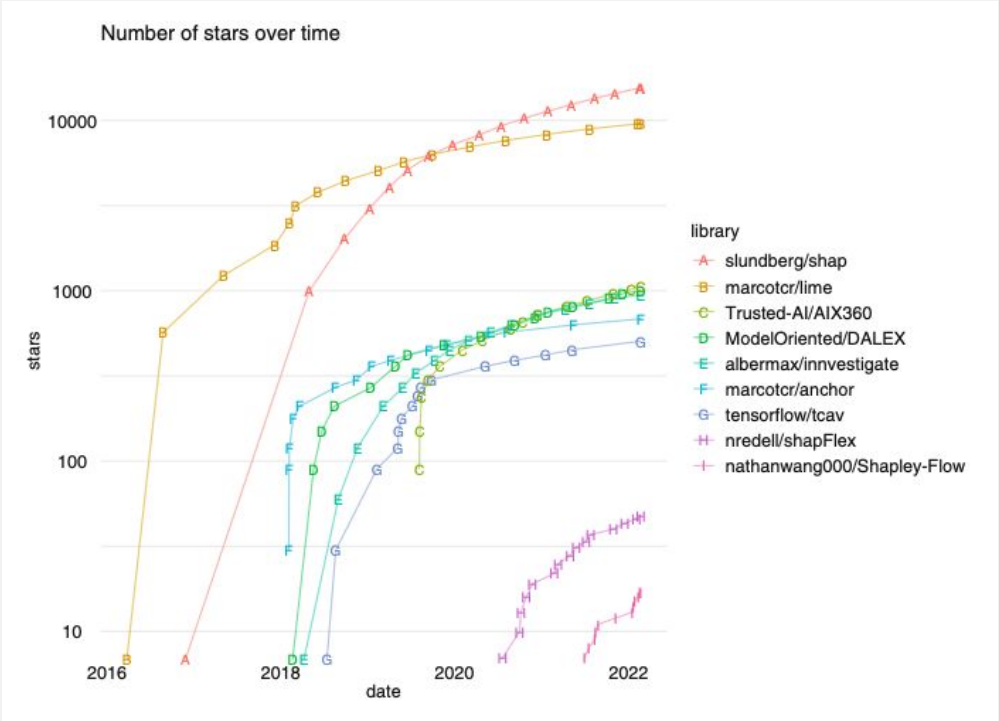


SHAP Summary Plot



Shapley values plot from DALEX

SHAP
popularity



Holzinger et al. *Explainable AI methods-a brief overview*

SHAP
properties

- **Local accuracy**
  the sum of Shapley values is equal to the model's prediction

- **Missingness**
  if feature is missing in input, then its Shapley value is 0 (have no impact)

- **Consistency**
  if a model changes so that the marginal contribution of a feature value increases or stays the same, then the Shapley value also increases or stays the same

Linearity,
Dummy,
Symmetry

Kernel SHAP

From reviewer: *Kernel SHAP is a significantly superior way of approximating Shapley values compared to classical Shapley sampling – much lower variance vs. number of model evaluations.*

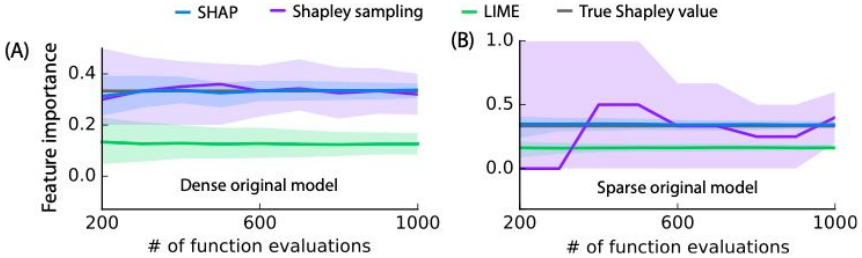


Figure 3: Comparison of three additive feature attribution methods: Kernel SHAP (using a debiased lasso), Shapley sampling values, and LIME (using the open source implementation). Feature importance estimates are shown for one feature in two models as the number of evaluations of the original model function increases. The 10th and 90th percentiles are shown for 200 replicate estimates at each sample size. (A) A decision tree model using all 10 input features is explained for a single input. (B) A decision tree using only 3 of 100 input features is explained for a single input.

Shapley Flow
motivation

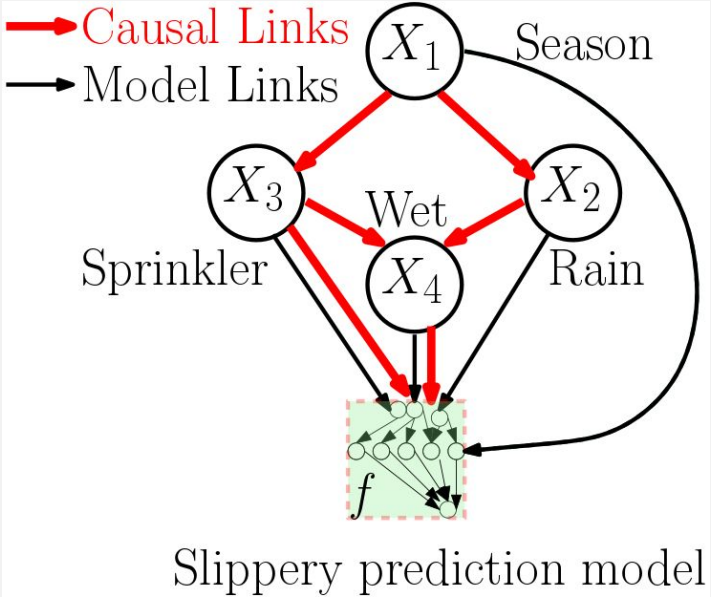
Explaining a model's predictions by assigning importance to its inputs (i.e., feature attribution) is critical to many applications in which a user interacts with a model to either make decisions or gain a better understanding of a system.

However, correlation among input features presents a challenge when estimating feature importance.

A causal graph, which encodes the relationships among input variables, can aid in assigning feature importance.

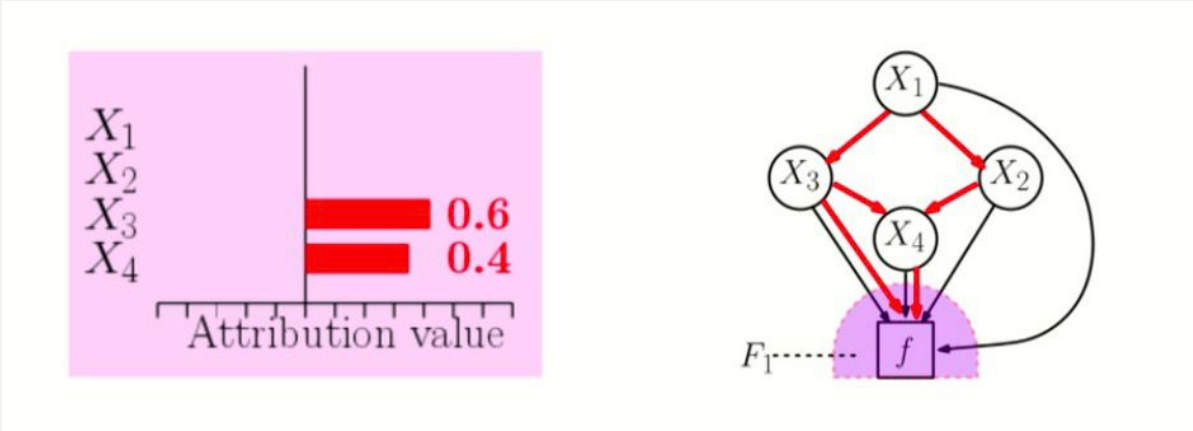
However, current approaches that assign credit to nodes in the causal graph fail to explain the entire graph.

Example
causal graph



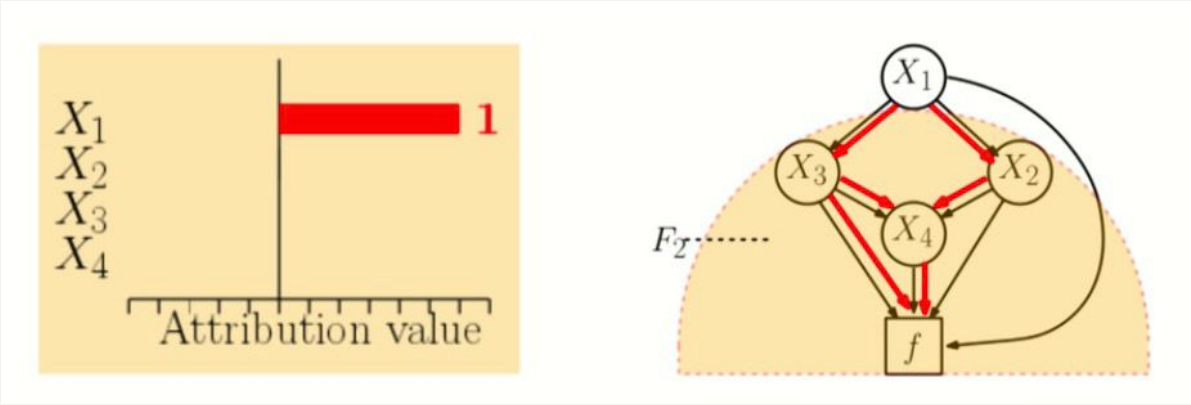
Pearl. *Causality*

Example
(independent) SHAP explanation



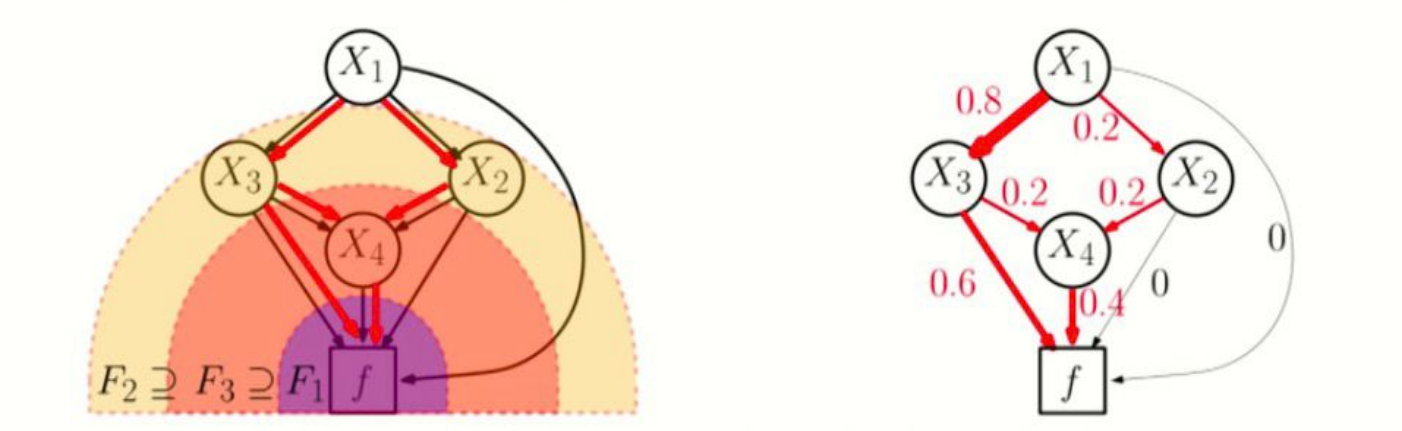
Example

Asymmetric Shapley Values (ASV) explanation



Frye et al. *Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability*

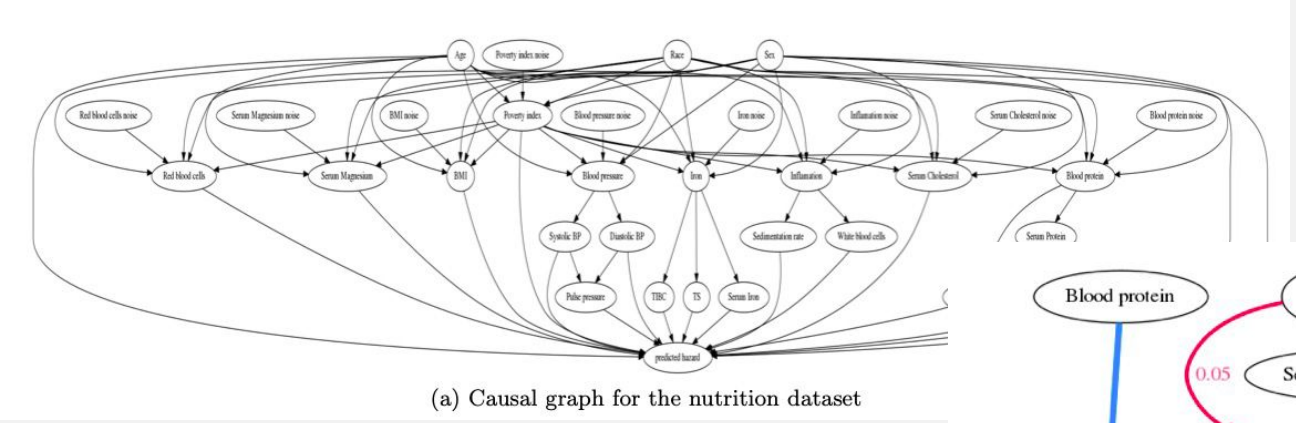
Example
Shapley Flow explanation



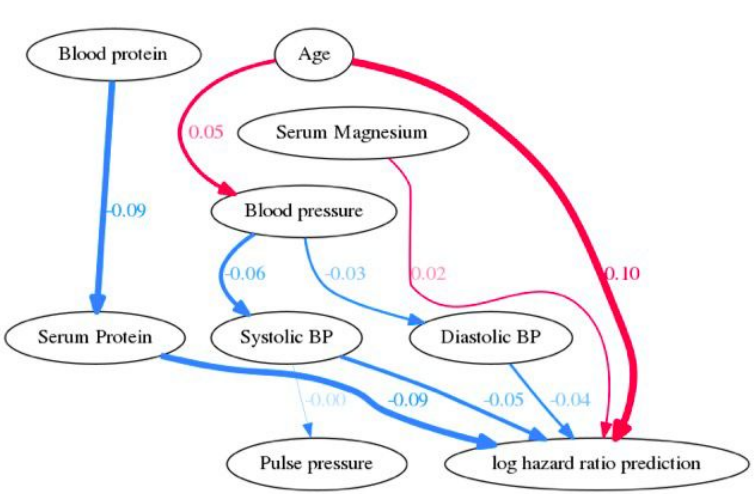
- An edge is important if removing it causes a large change in the model's prediction.
- The Shapley Flow value for an edge is the difference in model output when removing the edge averaged over all histories that are *boundary consistent*.

Drawback

the requirement to know the dependency structure



(a) Causal graph for the nutrition dataset

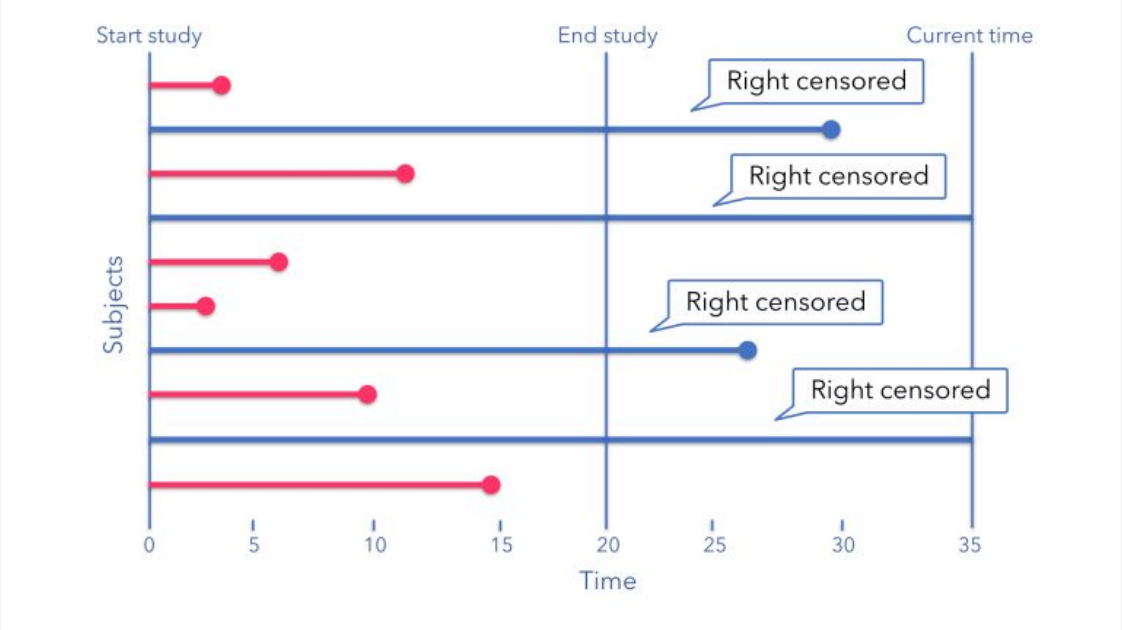


(a) Shapley Flow

Future work

- Can Shapley Flow work with partially defined causal graphs?
- How to explore Shapley Flow attribution when the causal graph is complex?
- Can Shapley Flow be useful for feature selection?

Survival Analysis
problem definition



Stainer. *The Notion of Censoring in Survival Analysis*

Survival Analysis

problem definition

Survival function

$$S(t) = \mathbb{P}(T > t) = 1 - \mathbb{P}(T \leq t)$$

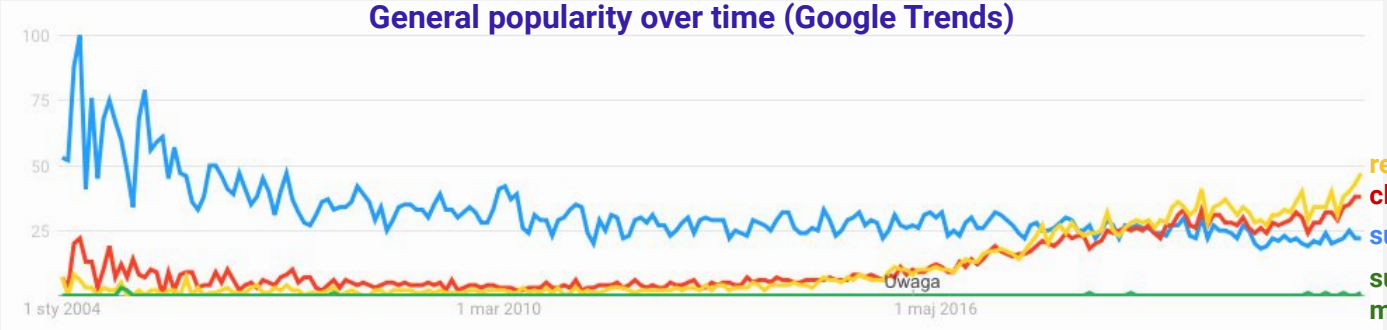
Hazard function

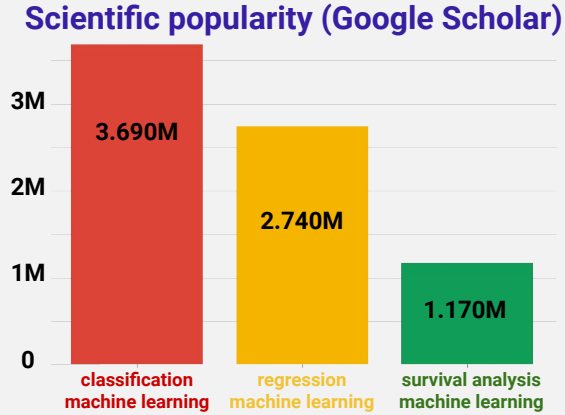
$$h(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

$$h(t) = \frac{-S'(t)}{S(t)}$$

Survival Analysis

state of ML





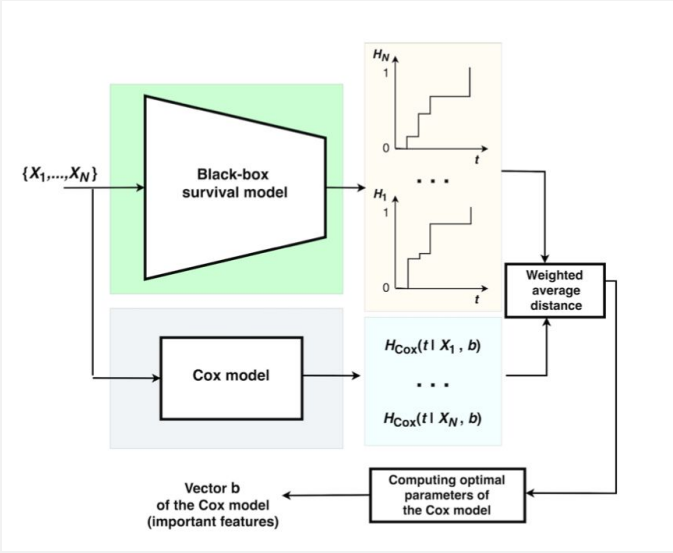
Survival Analysis

state of XAI

- **SurvLIME family**
  - SurvLIME (L_2-norm for the distance between CHFs)
  - SurvLIME-Inf (L_∞-norm)
  - SurvLIME-KS (Kolmogorov-Smirnov bounds)

- **Counterfactual explanations**

- ✗ **no open-source implementation**



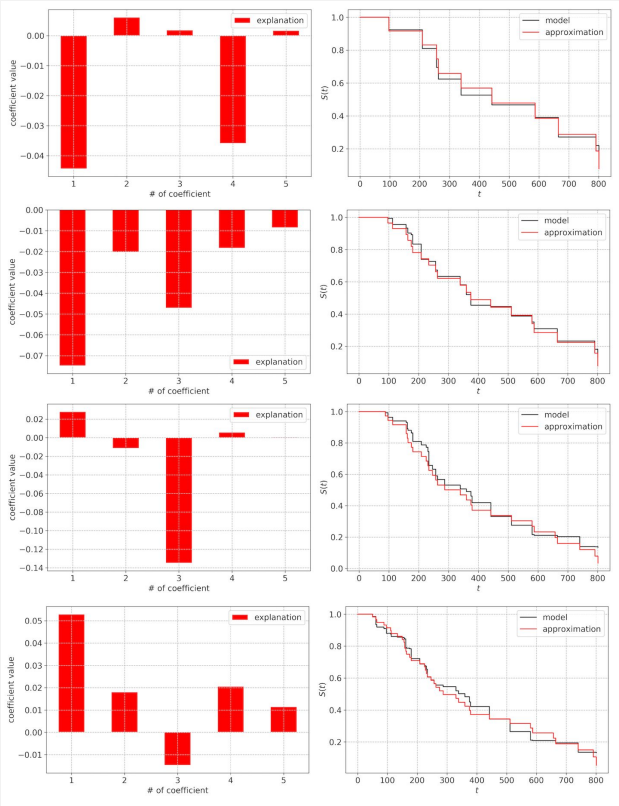
SurvLIME idea: apply the Cox proportional hazards model to approximate the black-box survival model

References:

- Kovalev, Utkin, Kasimov. *SurvLIME: A method for explaining machine learning survival models* (2020)
- Utkin, Kovalev, Kasimov. *SurvLIME-Inf: A simplified modification of SurvLIME for explanation of machine learning survival models* (2020)
- Kovalev, Utkin. *A robust algorithm for explaining unreliable machine learning survival models using the Kolmogorov-Smirnov bounds* (2020)
- Kovalev, Utkin. *Counterfactual explanation of machine learning survival models* (2021)

SurvLIME

explanations

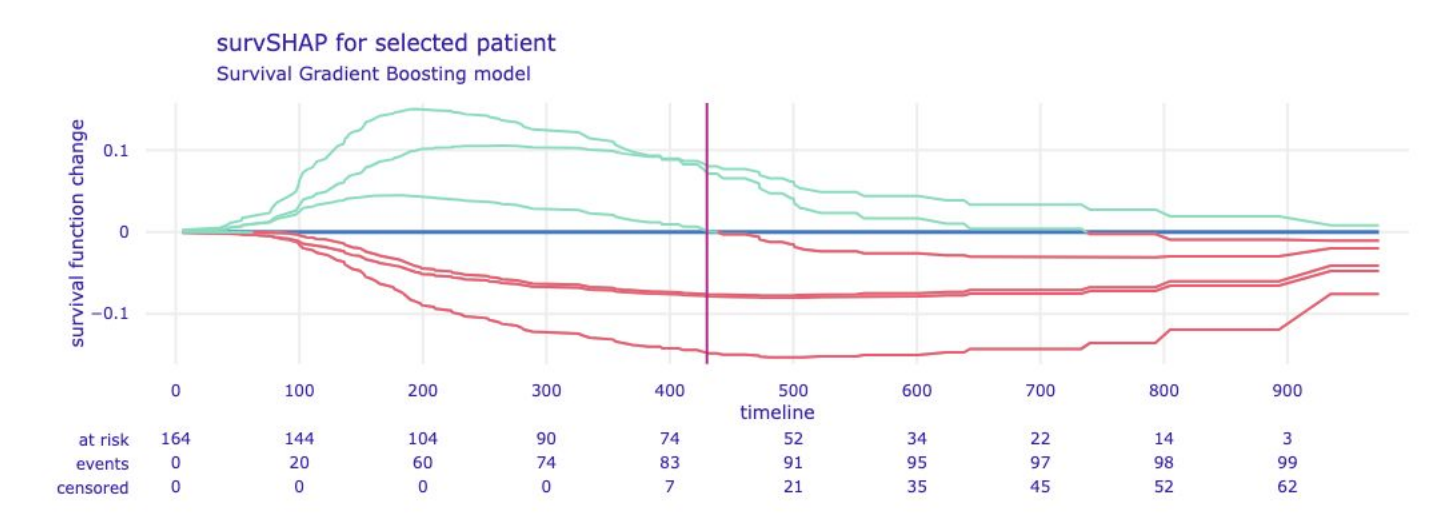


survSHAP
ideas, contribution

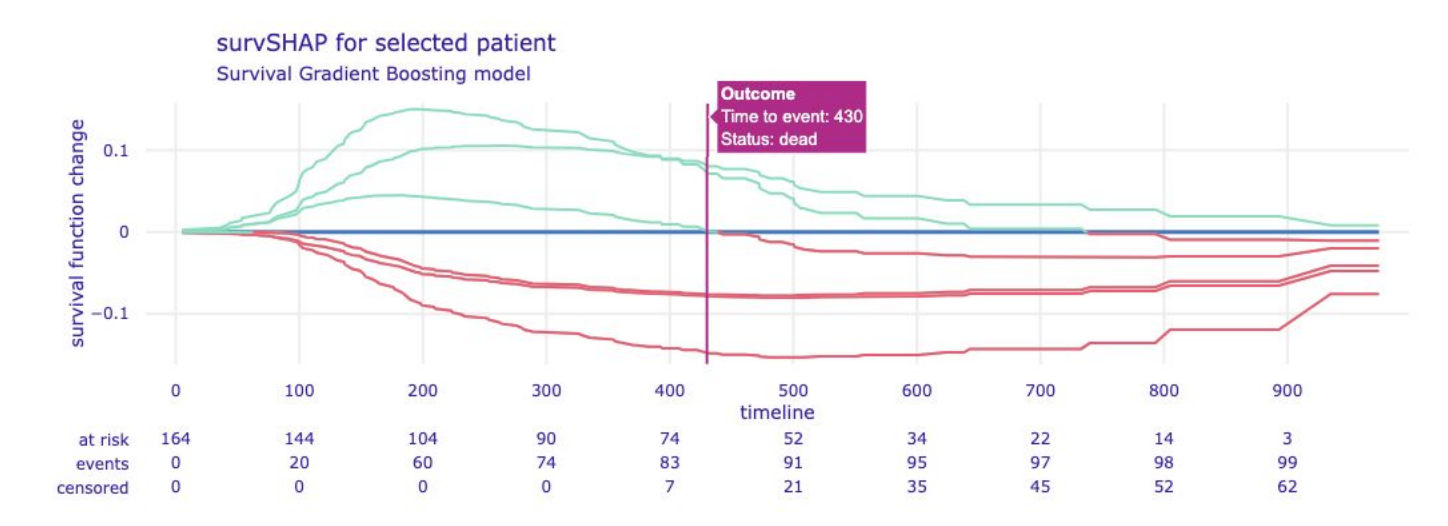
- Generalization of the SHAP method for the case of survival analysis models
- The use of concepts directly related to the area of survival analysis –
  explanations related to the survival function
- Creating an entire framework —
  family of visual explanations based on various aggregations and ways of interpretation
- Creating open-source implementation of the developed methods
  (in the DrWhy family with the use of dalex explainer)

survSHAP

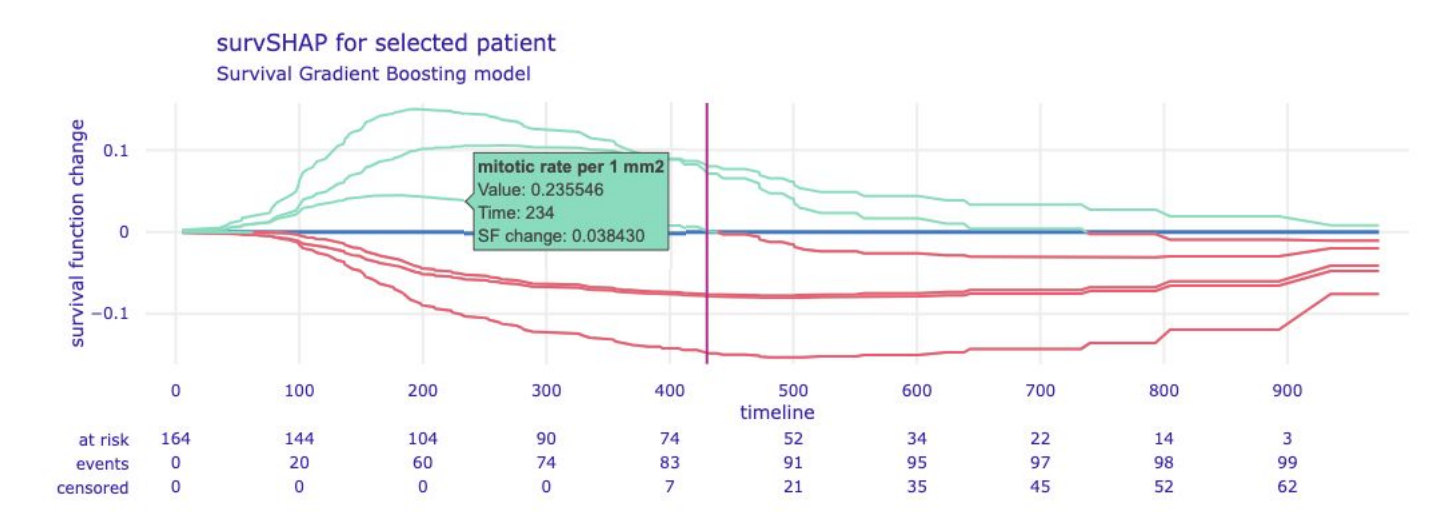
base example



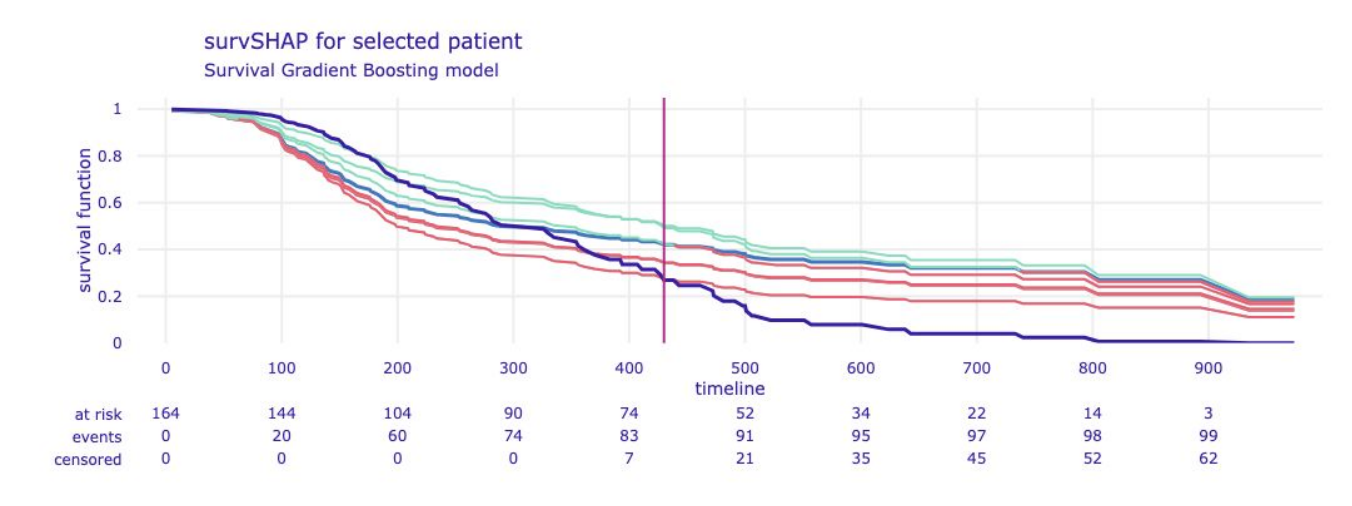
survSHAP
base example



survSHAP
base example



survSHAP
base example

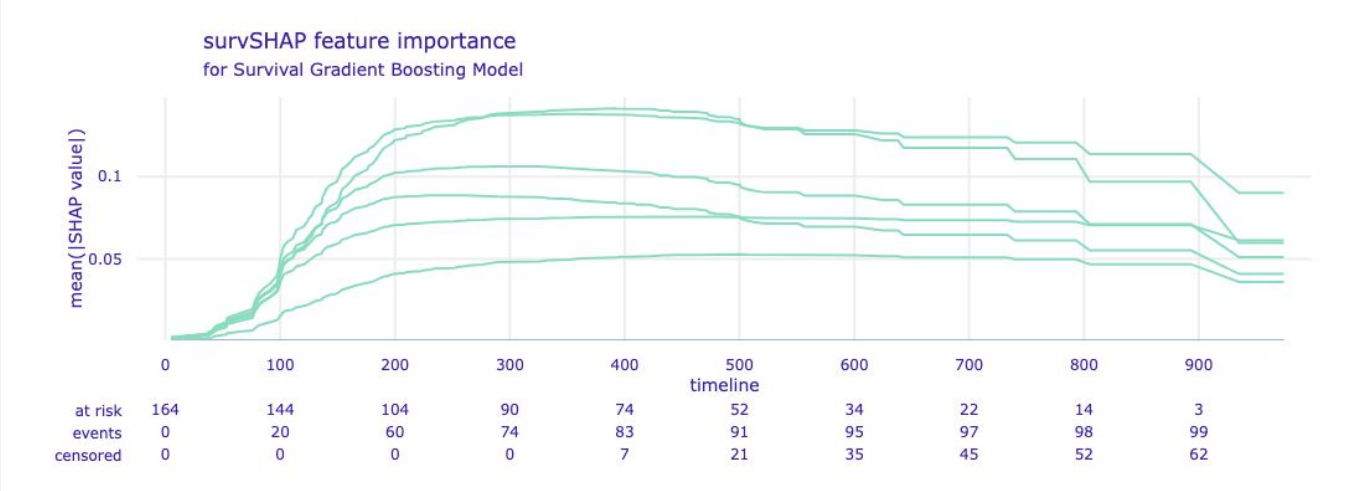


survSHAP
computations

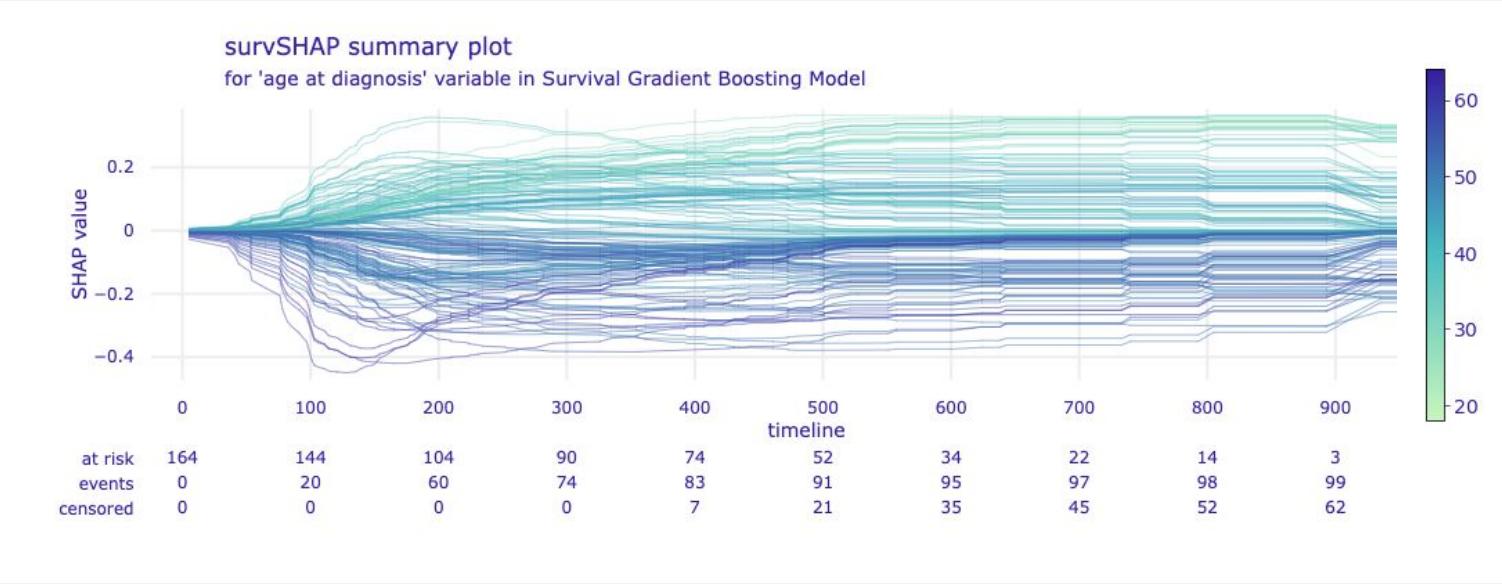
- Shapley values estimation (discrete time)
  - sampling
  - Kernel SHAP
    - using functional regression in SC-FR scenario (scalar covariate, functional response)
    - using functional LASSO

- ranking the importance of variables
  - Shapley values aggregation (sum of squares, max/mean absolute value over time)
  - functional depth (Band Depth, Modified Band Depth)

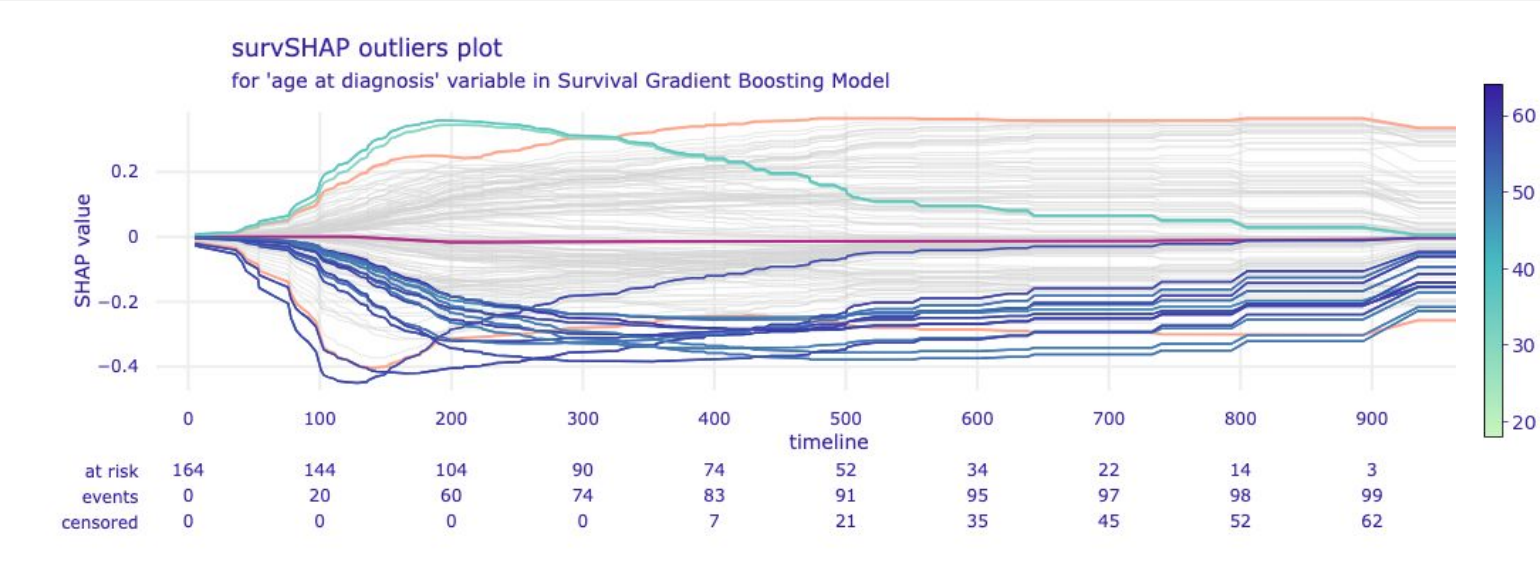
survSHAP
feature importance



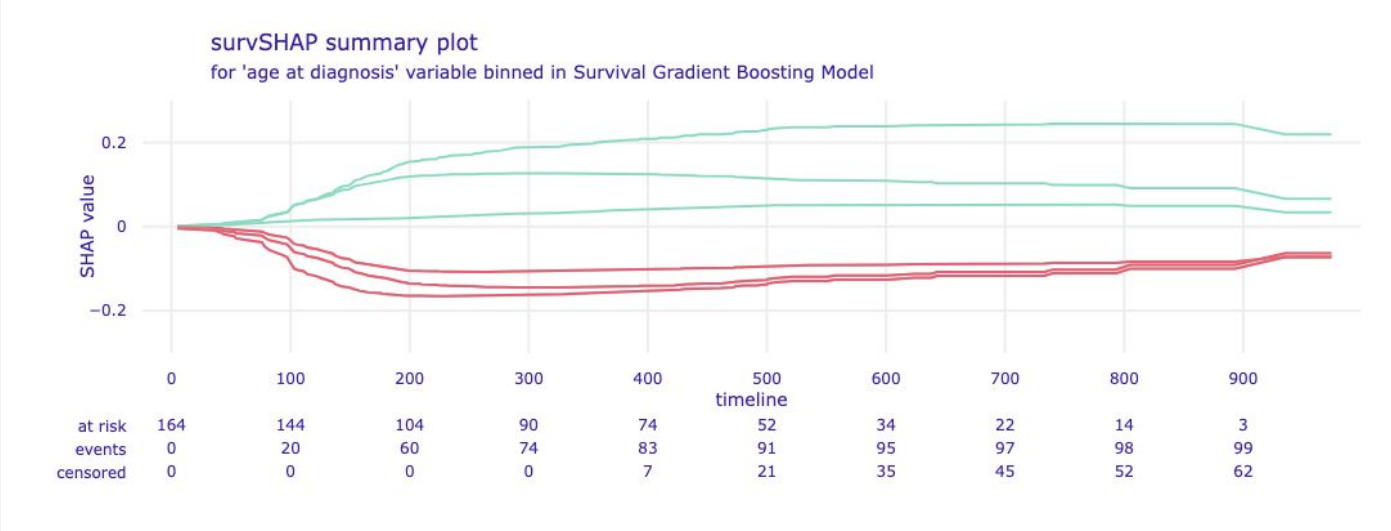
survSHAP
summary plot



survSHAP
summary plot



survSHAP
summary plot



survSHAP
future work

- implementation of other visual explanations
- the use of the functional data ANOVA –
  for analysis of variance of Shapley's values for individual covariates
- time-varying covariates support
- including the developed methods in the R survxai package