Time-dependent explanations of neural networks for survival analysis

Kamil Grudzień Krystian Sztenderski Jakub Bednarz

Abstract

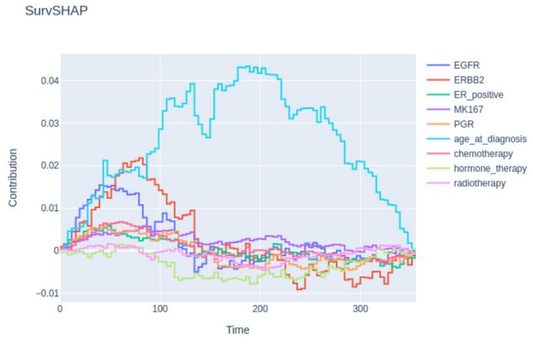
- How well **model-specific explanations** can perform on Deep Neural Network
- Comparison between **model-specific** and **model-agnostic** explanations
- Investigated explanations methods: **DeepLift**, **IG**, **SurvSHAP**, **DeepLiftShap**
- Models: **RSF, CoxPH, DeepHit**
- Experiments performed on **Metabric** dataset (taxonomy of breast cancer)

Results (DeepLift vs. SurvSHAP)

- Evaluation of explanations obtained with SurvSHAP and DeepLift using **Maximum Sensitivity** metric shows that **DeepLift's explanations change more drastically** when the input is varied infinitesimally with on average about 2 times higher sensitivity than SurvSHAP's explanations for the same samples. This suggests that DeepLifts explanations are less accurate, but direct comparisons between these methods show that **it does not have a significant effect on explanations**.

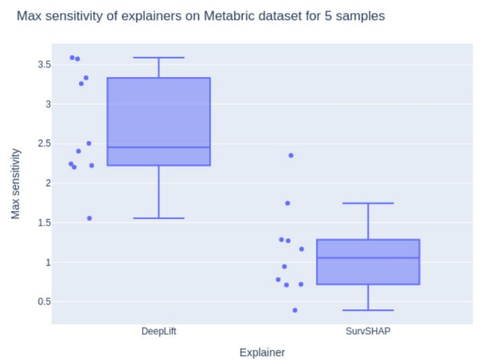
- **Direction and dynamic** of explanations change in time for both explainers **is very similar**. We can expect that for bigger samples the average absolute difference is even more flat. Even more impressing is that the metric value for most of the features is not greater than about 0.007

- The biggest **disadvantage of SurvSHAP is its computational time**. *Table 1* present execution times of different explanation methods. For each method and model we ran the method 10 times on a single sample from the Metabric dataset. We have observed that **SurvSHAP is on average 48 times slower than DeepLift** when explaining neural network DeepHit.

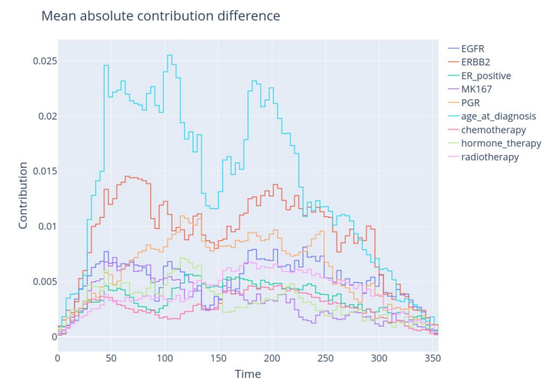


Sample explanations of SurvSHAP for DeepHit model on the METABRIC dataset.

Sample explanations of DeepLift for DeepHit model on the METABRIC dataset.



Maximum Sensitivity of DeepLift and SurvSHAP explanations of DeepHit model predictions for 10 samples from Metabric dataset.



Mean absolute difference between explanations of SurvSHAP and DeepLift for DeepHit model on the METABRIC dataset.

Method	Model	avg time (s)
SurvSHAP	Random Survival Forest	185.59
SurvSHAP	Cox Proportional Hazard	82.39
SurvSHAP	DeepHit	22.79
DeepLiftShap	DeepHit	2.91
Integrated Gradients	DeepHit	0.85
DeepLift	DeepHit	0.47

Table 1: Average (from 10 experiments each) of execution times of explanation methods for different models on a single sample from the Metabric dataset.

You can find more about the project here.