Time-dependent explanations of neural networks for survival analysis

Maurycy Moczulski mf.moczulski@student.uw.edu.pl
Jakub Skrajny j.skrajny@student.uw.edu.pl
Maciej Domaradzki m.domaradzki@student.uw.edu.pl
University of Warsaw, Poland

Abstract

SurvSHAP(T) presented in 'SurvSHAP(T): TIME-DEPENDENT EXPLANATIONS OF MACHINE LEARNING SURVIVAL MODELS' is the first time-dependent explanation that allows for interpreting survival black-box models. The proposed methods aim to enhance precision diagnostics and support domain experts in making decisions (Krzyziński, 2023). Gradient-based explanations such as Deeplift (Shrikumar, 2017) or Integrated Grandients (Sundararajan, 2017) were not made specific for survival models but they has got much lover complexity. Gradient-based cannot be used interchangeably with SurvSHAP(T) due to inferior performance.

1. Introduction

In this work our goal is comparison of Gradient-based explanations (Deeplift and Integrated Grandients) with SurvSHAP(T).

2. Methodology

In this work we explained deep neural network (DeepHit) using SurvSHAP(T), Deeplift and Integrated Gradients. We trained network on two datasets and compared computation times of explanations.

2.1 Datasets

We used two datasets. first dataset is "METABRIC EXPERIMENT - The Molecular Taxonomy of Breast Cancer International Consortium" with 1428 records including medical information, age and some binary variables such as hormone treatment indicator, radiotherapy indicator or chemotherapy indicator. Second dataset is ' SUPPORT EXPERIMENT - Study to Understand Prognoses Preferences Outcomes and Risks of Treatment' with 6654 records and 14 variables such as age, sex, presence of diabetes, presence of dementia, presence of cancer, heart rate etc. Those datasets contain also information about event (survived or not) and duration (time).

2.2 Models

First of all we trained and explained random forest, coxph and DeepHit using SurvSHAP(T) to see if our implementation of DeepHit is correct. Then we explained DeepHit using Deeplift, Integrated Grandients and SurvSHAP(T) and computed explanations time.

DeepHit handles competing risks; i.e. settings in which there is more than one possible event of interest. Comparisons with previous models on the basis of real and synthetic datasets demonstrate that DeepHit achieves large and statistically significant performance improvements over previous state-of-the-art methods. (Lee, 2018)

2.3 Explanation methods

SurvSHAP(T) is a model-agnostic explanation method designed to explain survival models and can be applied to all models with functional output (Krzyziński, 2023).

Deeplift is an explanation method for decomposing the output prediction of a neural network on a specific input by backpropagating the contributions of all neurons in the network to every feature of the input. (Shrikumar, 2017)

Integrated Grandients is an explanation method designed to explain neural networks. which satisfy Sensitivity and Implementation Invariance. (Sundararajan, 2017)

3. Experimental results

We achieved train score equal 0.89, test score equal 0.62 for random forest; train score equal 0.65, test score equal 0.6 for coxph and train score equal 0.68, test score equal 0.66 for DeepHit trained on Metabric dataset.

We achieved train score equal 0.83, test score equal 0.61 for random forest; train score equal 0.57, test score equal 0.56 for coxph and train score equal 0.63, test score equal 0.61 for DeepHit trained on Support Experiment dataset.

3.1 Comparison of time computation

Explainer working time for SurvSHAP(T) on DeepHit for Metabric dataset is equal 11 seconds, for Deeplift 0.9 sec. and for Integrated Gradient 1.4 sec..

Explainer working time for SurvSHAP(T) on DeepHit for Support Experiment dataset is equal 1276 seconds, for Deeplift 37 seconds and for Integrated Gradient 4 seconds.

4. Conclusion

There was a little misunderstanding in our team. because of this misunderstanding no one has implemented metrics to check the quality of explanations. But we learned that the results of Gradient-based explanations are much worse even if the computational time is significantly less they cannot be used as faster alternatives for SurvSHAP(T)

5. Explanation results

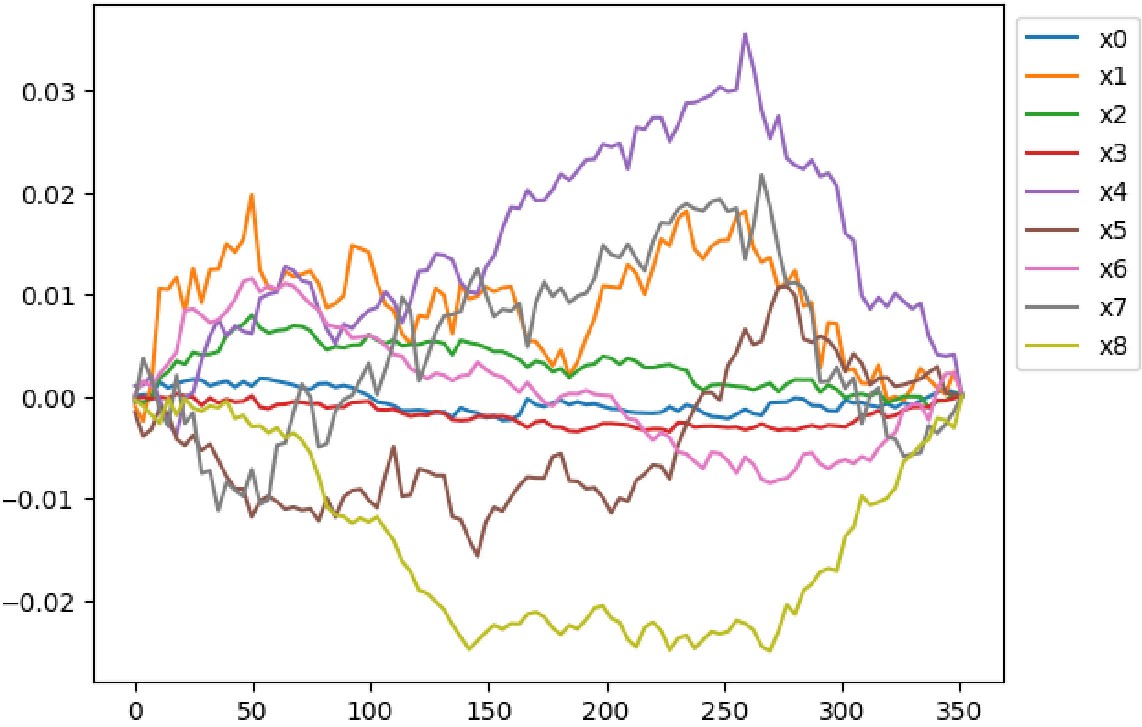

Figure 1: SurvSHAP(T) explenation of DeepHit on Metabric dataset

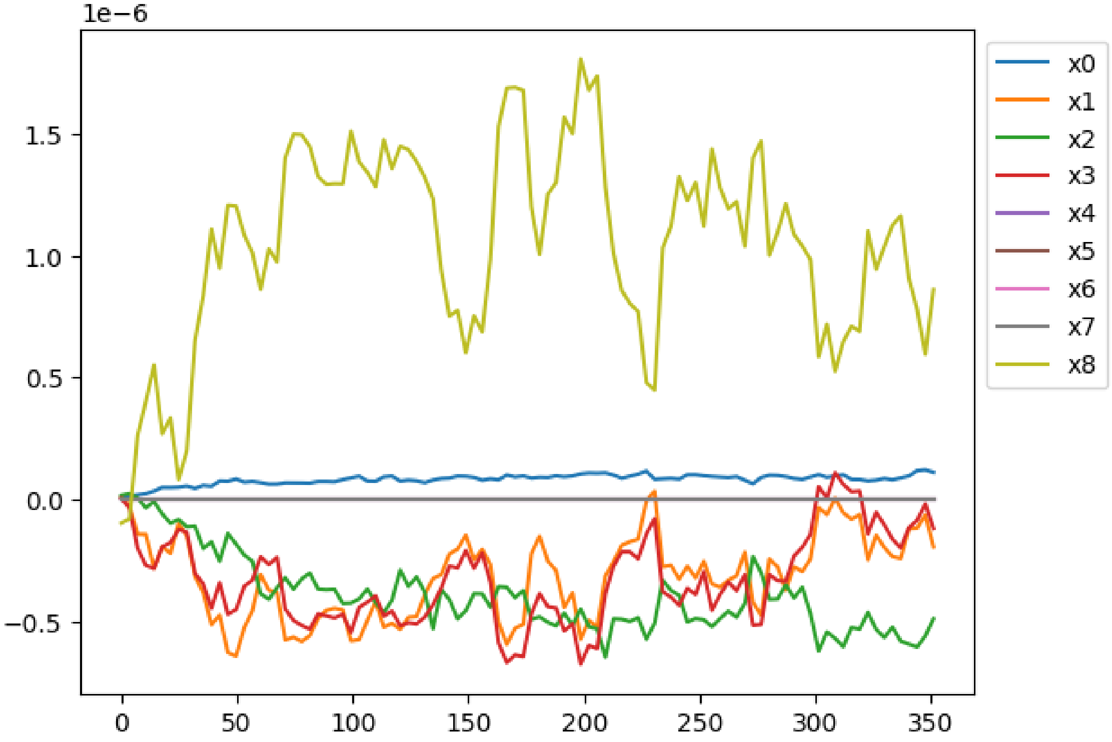

Figure 2: Integrated Grandients explenation of DeepHit on Metabric dataset

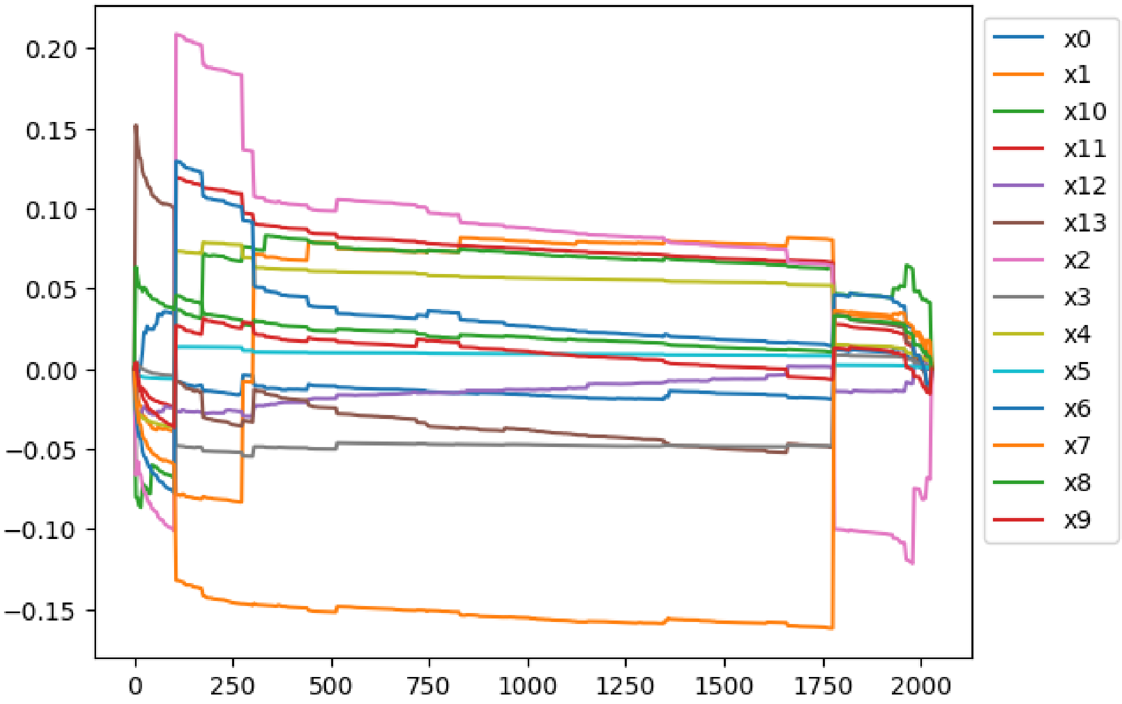

Figure 3: SurvSHAP(T) explenation of DeepHit on Support Experiment dataset

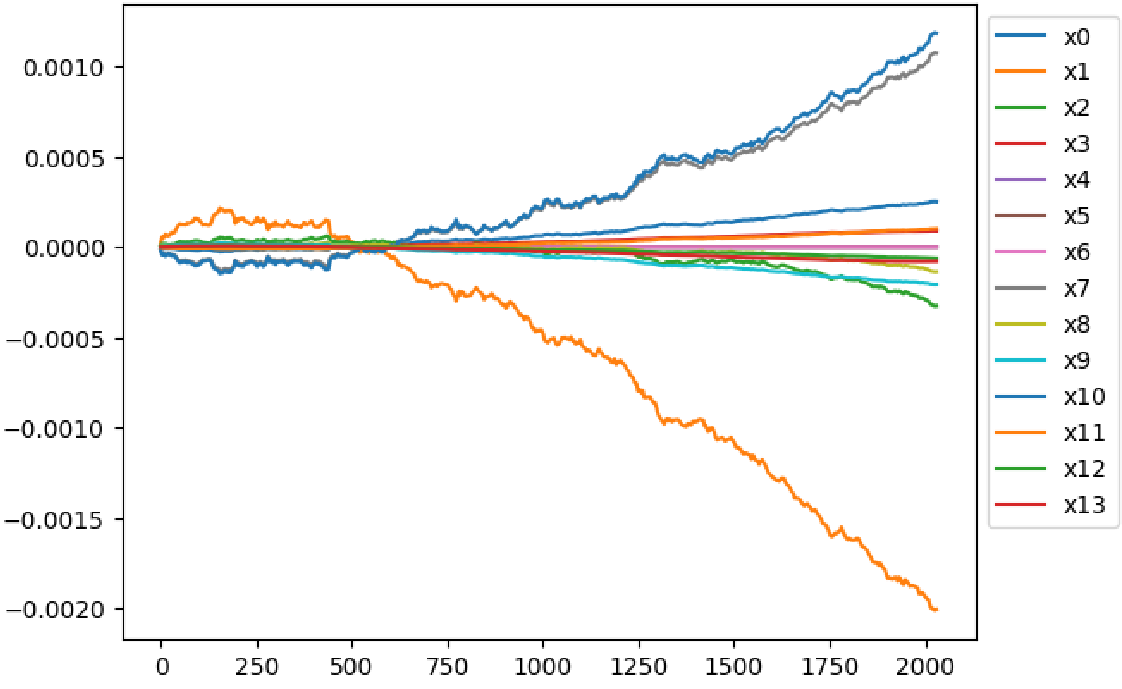

Figure 4: Deeplift explenation of DeepHit on Support Experiment dataset

References

Baniecki Biecek Krzyziński, Spytek. Time-dependent explanations of machine learning survival models. *Knowledge-Based Systems*, 262(110234), 2023.

Zame W. Yoon J. Van Der Schaar M. Lee, C. A deep learning approach to survival analysis with competing risks. *In Proceedings of the AAAI conference on artificial intelligence*, 32 (1), 2018.

Greenside P. Kundaje A. Shrikumar, A. Learning important features through propagating activation differences. *International conference on machine learning*, pages 3145–3153, 2017.

Taly A. Yan Q. Sundararajan, M. Axiomatic attribution for deep networks. *International conference on machine learning*, pages 3319–3328, 2017.