Evaluating time-dependent explanations of neural networks for survival analysis

Jakub Krajewski JK406798@STUDENTS.MIMUW.EDU.PL
Maciej Wojtala MW406587@STUDENTS.MIMUW.EDU.PL
Stanisław Frejlak S.FREJLAK@STUDENT.UW.EDU.PL
University of Warsaw, Poland

Abstract

Explainability of survival models has been recently addressed by Krzyziński et al.. The proposed explanation, SurvSHAP(t), is a model-agnostic method meeting the local accuracy property. Its downside is high computational complexity. Gradient-based explanations, such as Integrated Gradients (Sundararajan et al.), have much lower complexity but they are not crafted specifically for time-dependent data. In this work, we show that explanations of deep survival models provided by Integrated Gradients do not resemble those of SurvSHAP so the former method cannot be treated as a computationally cheaper alternative to the latter.

1. Introduction

The main goal of our work is to compare the Integrated Gradients method and the SurvSHAP method. SurvSHAP is a model-agnostic method designed to explain survival models and Integrated Gradients is a general method used to explain neural network models. We evaluate both models on a neural network trained to solve the censored survival analysis problem and analyse the results.

1.1 Survival analysis

We consider the censored survival analysis problem. In this framework an observation i is described by a triple (x_i, y_i, σ_i), where $x_i \in \mathbb{R}^n$ is the variables vector, σ_i is the censoring indicator (i.e. the observation is censored for $\sigma_i = 0$ and revealed for $\sigma_i = 1$), y_i is either revealed survival time T_i for $\sigma_i = 1$, or censored survival time C_i for $\sigma_i = 0$ (Krzyziński et al.). The main goal of the survival analysis is predicting the survival time T_i from the variables vector x_i. Usually the output is a function of time instead of a single time moment. Using the censored framework is modelling data gaps and makes predicting the survival time more difficult. The most fundamental function considered in the survival analysis is the survival function $S(t) = P(T > t)$. Another important object is the hazard function $h(t) = -\frac{dS}{dt} \ln(S(t))$. The most frequently used approach to the survival analysis is the Cox Proportional Hazards models (Cox). Another solutions in this area are Random Survival Forest (Ishwaran et al.), Cox-nnet (Ching T.), DeepSurv (Katzman J.L.), Cox-Time (Kvamme et al.).

1.2 Explanation methods

1.2.1 SurvSHAP

SurvSHAP (Krzyziński et al.) is a model-agnostic method designed to explain survival models with functional output. The novelty of this method is creating explanations that are time-dependent. In its core SurvSHAP is based on the Shapley value calculated on conditional expected values of the survival function. SurvSHAP explanations satisfy the local accuracy property.

1.2.2 Integrated Gradients

Integrated Gradients (Sundararajan et al.) is an explanation method designed to explain neural networks, in particular it can be used to explain neural networks trained to perform the survival analysis. The method creates explanations that satisfy two fundamental axioms - Sensitivity and Implementation Invariance. Explanations in the Integrated Gradients method for an input and a baseline are obtained by integrating the gradient of explained function over a path from the baseline to the input point. The method utilizes the fact that neural networks are differentiable and the calculus theorem stating that (under mild assumptions) the integral over a path between two points does not depend on the choice of the path.

1.3 Evaluation of explanation methods

Evaluation of explanation methods is a complex and active research area. One of the difficulties is the lack of one ground truth. We can therefore judge the quality of explanations from various perspectives. Two popular quantitative metrics for evaluating performance of explanation methods are described in Bhatt et al. and Yeh et al.. Faithfulness Correlation measures consistency of attributions with the decision-making process by comparing a random subset of the attributions with model's output. Average Sensitivity assesses robustness of the explanation by measuring how much it changes when small variations are made to the input. The result is obtained using Monte Carlo sampling-based approximation. Another approach might be to use a special dataset with known, straightforward relations between certain features and target value. Having a model that is well fitted, we can assume these correspondences are also captured. We can then make sure that the chosen explanation method is able to detect the dependency.

2. Methodology

In this work, we have evaluated effectiveness of two metrics for explaining the workings of a neural network model used in Survival Analysis: SurvSHAP and Integrated Gradients. To achieve this, we fitted the model to both artificial and medical datasets and compared computation times and quality of explanations generated by both methods.

2.1 Neural network

DeepHit (Lee et al.) is a neural network architecture designed to perform the survival analysis. The model consists of shared sub-network and some numbers of heads called

cause specific sub-networks. Each of them consists of fully-connected layers. Output of the network are obtained via softmax and are modelling the joint probability of the first hitting time (we assume that time horizon is finite). The network is trained using the loss consisting of two terms - the first one is responsible for optimising the log-likelihood of the estimated joint probability and the second one is an additional ranking loss used to incorporate the cause-specifics effects. Additionally, we use the Brier Score as an auxiliary metric of fittness quality. Both the loss function and the Brier Score indicated behaviour significantly better than the chance level, which made explaining the model feasible.

2.2 Datasets

For the first experiment, we have used synthetic dataset consisting of N = 1000 observations, generated using the method proposed by Crowther and Lambert. The coefficients were chosen to highlight differences between variables. Only x_1 has a time-dependent effect. Variables x_2, x_3 and x_4 present constant influence, while x_5 can be treated as insignificant random noise. In order to evaluate the effectiveness of the explanation methods in real-world applications, we also used a dataset of 299 patients with heart failure presented in Ahmad et al.. The dataset contains various factors that may have affected the patient's mortality, divided into continuous and categorical variables.

2.3 Evaluating explanations

For the first (synthetic) dataset, our goal was to determine if the methods were able to effectively identify the relationships between variables and their corresponding outputs. Specifically, we wanted to observe the attribution of variable x_1 varying over time, as well as the constant attribution of x_2, x_3, and x_4, and the minimal attribution of x_5. To accomplish this, we prepared aggregated plots showing change in time of mean and standard deviation of attribution for each variable over the entire dataset. To measure quality of explanations quantitatively, we have utilized Average Sensitivity and Faithfulness Correlation metrics from the quantus package in Python. Unfortunately, the available implementation only covers some models, which do not include SurvSHAP(t). Therefore, we have used explanations of an untrained model as a baseline for evaluating Integrated Gradients.

3. Experimental results

3.1 Assessing performance of methods

3.1.1 Experiment on artificial data

In this part we have trained the DeepHitSingle model based on a simple neural network with one hidden layer of size 5. The model was relatively stable during training. We achieved validation loss of 0.79 and an Integrated Brier Score of 0.18. These metric values indicate that the model was well-fitted to the data and thus, we could proceed to generate explanations. We applied the SurvSHAP explanation method and observed results similar to those reported by Krzyziński et al.. The generated plots clearly showed the time-dependent effect of variable x_1. Additionally, the other variables also behaved consistently with their attribution to the prediction. However, when using the Integrated Gradients explanation

method, the results were not fully satisfactory. We were unable to observe a time-dependent effect for variable x_1 and the explanations for the other variables were not sufficient. It's important to note that Faithfulness Correlation and Average Sensitivity plots presented generally better results on the trained model. However, the differences were not large and especially only marginal for Faithfulness Correlation. Exemplary plots for the results achieved in this section are shown in Figure 1 in Appendix.

3.1.2 Experiment on medical data

In the second experiment, we faced more challenges while fitting the model to the data. We used have found the same neural network architecture to work in this case. Despite the increased difficulty, we were able to achieve a validation loss of 0.54 and an Integrated Brier Score of 0.228. These metrics indicate that the model was able to capture the underlying relationships in the data. When it came to explaining the model's predictions, we found that the task was more complex due to the increased number of variables. In this case, there was a lack of a "ground truth" to compare the explanations to. However, the results for SurvSHAP were again consistent with what has been previously described in Krzyziński et al.. The generated plots showed sensible results and highlighted the time-dependent effect for some variables. On the other hand, the results for Integrated Gradients seemed random and did not show any clear correspondences when aggregated over the whole dataset. The Average Sensitivity and Faithfulness Correlation were comparable to the previous case, and again slightly better for the trained model. Figure 2 in Appendix shows an example of the results obtained in this case. Overall, the results suggest that Integrated Gradient produce explanations better than random guess, but not satisfactory.

3.2 Comparison of computation time

In terms of computation time, the results were similar to expected. Integrated Gradients method was significantly faster than SurvSHAP. However, the computation time for both methods was generally satisfactory and not a major limiting factor. A detailed comparison of computation times for both methods can be found in the appendix.

4. Conclusion

In conclusion, we have compared two methods of explanation for survival analysis models: Integrated Gradients, which is specific for neural networks, and SurvSHAP, which is model-agnostic but constructed specifically for survival analysis. We have found that while the computation time for Integrated Gradients was significantly shorter, the results were worse than those obtained with SurvSHAP. Additionally, the computation time for SurvSHAP was not found to be a limiting factor. Therefore, for the time being, we recommend the use of SurvSHAP as the method of choice for explaining survival analysis models. However, it should be noted that with further improvements to the neural network explanation methods or additional transformations of the model output, it may be possible to shorten computation time without loss of accuracy.

References

Tanvir Ahmad, Assia Munir, Sajjad Bhatti, and Muhammad Aftab. Survival analysis of heart failure patients: A case study. *PLoS ONE 2017*, abs/12(7). URL https://doi.org/10.1371/journal.pone.0181001.

Umang Bhatt, Adrian Weller, and José M. F. Moura. Evaluating and aggregating feature-based model explanations. *CoRR 2020*, abs/2005.00631. URL https://arxiv.org/abs/2005.00631.

Garmire LX. Ching T., Zhu X. Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS Comput Biol. 2018 Apr 10.*

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological) 34, no. 2 (1972): 187–220. http://www.jstor.org/stable/2985181.*

Michael Crowther and Paul C. Lambert. Simulating biologically plausible complex survival data. *Statistics in Medicine 2013*, abs/10. URL https://doi.org/10.1002/sim.5823.

Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. *The Annals of Applied Statistics 2008*, 2(3):841 – 860. doi: 10.1214/08-AOAS169. URL https://doi.org/10.1214/08-AOAS169.

Cloninger A. et al. Katzman J.L., Shaham U. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Med Res Methodol 18, 24 (2018).* doi: https://doi.org/10.1186/s12874-018-0482-1.

Mateusz Krzyziński, Mikołaj Spytek, Hubert Baniecki, and Przemysław Biecek. Survshap(t): Time-dependent explanations of machine learning survival models. URL https://arxiv.org/abs/2208.11080.

Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-event prediction with neural networks and cox regression. *Journal of Machine Learning Research 2019*, 20(129):1–30. URL http://jmlr.org/papers/v20/18-424.html.

Changhee Lee, William Zame, Jinsung Yoon, and Mihaela van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. *Proceedings of the AAAI Conference on Artificial Intelligence 2018*, 32(1), Apr. . doi: 10.1609/aaai.v32i1.11842. URL https://ojs.aaai.org/index.php/AAAI/article/view/11842.

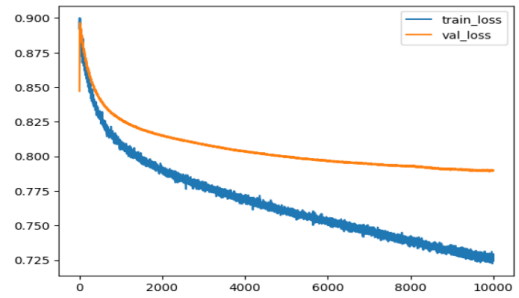
Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *CoRR 2017*, abs/1703.01365. URL http://arxiv.org/abs/1703.01365.

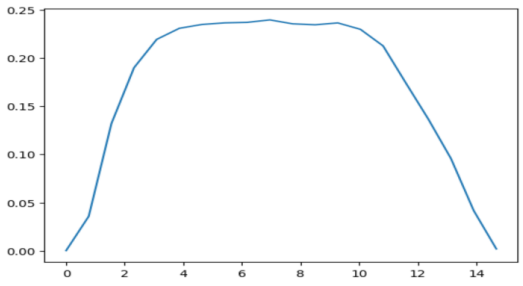
Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Suggala, and David I. Inouye. On the (in)fidelity and sensitivity for explanations 2019. URL https://arxiv.org/abs/1901.09392.

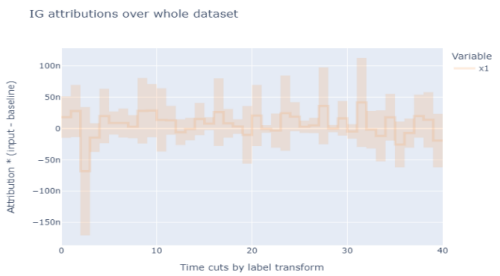
Appendix

Results for artificial dataset

Figure 1: Comparison for artificial dataset. For calculations over the entire dataset mean and standard deviation is depicted on the plots.

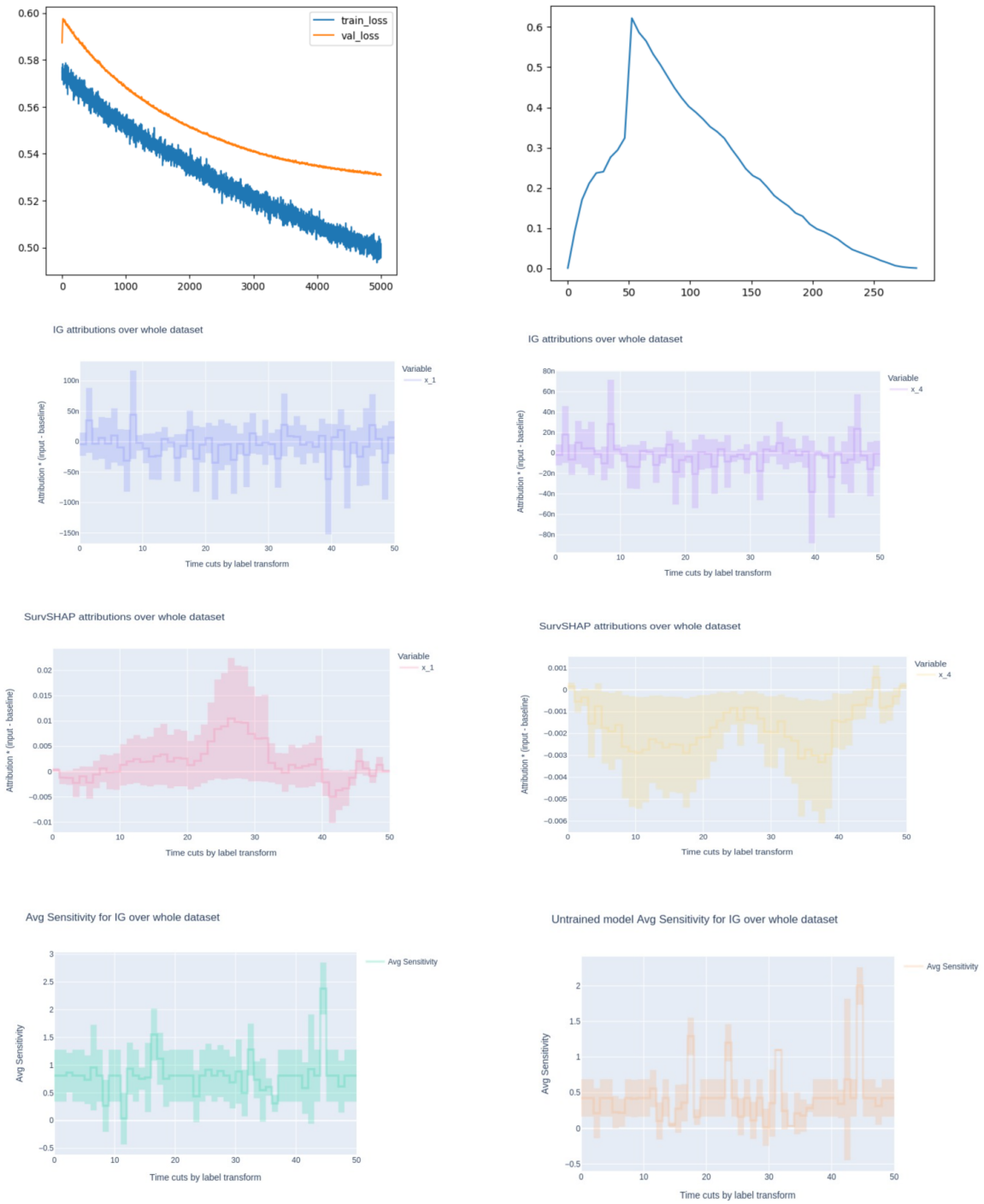






Results for medical dataset

Figure 2: Results for medical dataset. For calculations over the entire dataset mean and standard deviation is depicted on the plots.



Comparison of computation time

Table 1: Time of computation for compared methods on both datasets

	artificial data	medical data
SurvSHAP(t)	102 ms ± 7.82 ms	2.71 s ± 13.6 ms
Integrated Gradients	670 ns ± 16.2 ns	680 ns ± 15.6 ns