Comparison between methods of explaining deep survival analysis models

Jakub Krajewski, Stanisław Frejlak, Maciej Wojtala

Deep survival analysis models are a powerful tool for analyzing data and predicting outcomes related to time-to-event data, such as the length of time a patient will survive after a medical treatment. However, these models can be complex and difficult to interpret, which can make it challenging to understand the underlying mechanisms and factors that contribute to the predictions made by the model. In this project, we will compare two different methods for explaining survival analysis models based on deep neural networks: SurvSHAP(t), which is known for its accuracy but can be slow to compute, and methods specifically designed for explaining neural networks. We will evaluate the strengths and weaknesses of each approach. Hopefully, we would like to show that model-specific methods give similar results, but are much faster to compute.

We have decided to work with the DeepHitSingle model, because it is designed for predicting events of one type (in contrast to for example DeepHit). For neural explanation methods, we have chosen DeepLift and Integrated Gradients.

We have run the code for training/explaining in two data scenarios. So far numerical results are limited (.csv files are uploaded along with this report). However, we have overcome the main technical issues and understood the problem. Therefore, we expect to get to the next results faster.

What has been done:

- Installing and getting familiar with captum, pycox, survshap packages, solving compatibility/hardware issues
- Analysis and understanding of the problem of explaining survival analysis models
- Multiple attempts to run SurvSHAP(t) on artificial data prepared by authors (we have encountered problems due to long time of computation and hanging program). Finally we have found out that it's probably the same issue as mentioned here: link to github
- Successful training of DeepHitSingle model on medical data provided by authors of SurvSHAP(t)
- Running methods for neural networks (DeepLift, Integrated Gradients) on the trained neural model
- Comparing first results of mean attributions/mean convergence deltas

Next steps:

- Explaining DeepHitSingle results by SurvSHAP(t) on medical data
- Comparing results of the three explanation methods (time/accuracy tradeoff)
- Training DeepHitSingle on artificial data and explaining it aith all the three methods
- Comparing results on artificial data and making final conclusions

Difficulties:

- Compatibility/technical issues (several different libraries with distinct dependencies)
- Long computation time (we need to wait for the results to compare, computation may hang/stop)
- Concise documentation of packages (we need more time to understand the codebase)