Title:

A novel interpretation method for explaining machine learning survival models

Abstract:

Machine learning models such as tree-based ensemble methods or neural networks have been adapted to handle survival data and have shown superior predictive performance compared to traditional statistical approaches. However, the lack of interpretability restricts the adoption of these machine learning models in survival analysis. In these lines, a novel interpretation method is proposed for explaining machine learning survival models. It extends the framework of the popular interpretation method LIME by applying the joint model to approximate the machine learning survival model at the local scale of a test example. The proposed method explains a machine learning survival model through the linear combination of covariates included in the joint model, such that coefficients of the covariates can be regarded as quantitative impacts on the prediction. Besides, by using the joint model, the proposed method has the advantage of handling the endogenous time-varying covariate, which is critical to survival analysis.