Comparative Analysis of Machine Learning Models for Survival Analysis: Empirical Study and Performance Assessment

Key words and phrases. survival analysis, survival predictive models, survival machine learning

Abstract:
Survival analysis is used in healthcare, but also in economics, finance, and marketing as well as sickness analysis and monitoring. Machine learning algorithms can provide accurate survival analysis predictions. This paper examines popular machine learning methods for predictive survival analysis to help healthcare professionals and researchers choose models for their datasets. The research applies four machine learning models to SimulatedONE, SimulatedTWO, and NKI Breast Cancer Data. The Cindex measures surviving machine learning model prediction capability. The DeepSurv model beats other models in most circumstances, showing well-calibrated projected probabilities and effective event occurrence and non-occurrence discrimination.

1. Introduction

Survival analysis is crucial for predicting patients' time-to-event outcomes and supporting healthcare practitioners in making the best treatment decisions (Ping, W., et al. 2017), not just in illness analysis or monitoring procedures but also in assisting in the quantitative and qualitative improvement of preventative medicine (lifestyle interventions, vaccine efficacy, screening programs, among others).

Although in recent years, survival analysis (time to event) is beginning to be used in economics, finance, operations, marketing, credit risk, among other disciplines (see Chava and Jarrow, 2004; Carling, Pan, Ariyan, Narayan, and Truini, 2007; Leong, Nguyen, Meredith, et al., 2008; and Leonardis and Rocci, 2009). In the last years, machine learning algorithms have matured into outstanding instruments for survival analysis, providing precise and trustworthy forecasts (Balan and Putter, 2020; Balan, 2018; Alyass, Turcotte, and Meyre, 2015).

Considering the relevance of the use of machine learning models, the primary goal of this study is to highlight the most widely used machine learning approaches for predictive survival analysis. This article provides a thorough examination of machine learning models for survival analysis, which may aid healthcare practitioners and researchers in selecting the best model for their datasets.

2. Case Studies

We present the performance of 4 machine learning models on three different datasets: SimulatedONE, SimulatedTWO, and NKI Breast Cancer Data (Lum, P., Singh, G., Lehman, A. et al, 2013). The SimulatedONE is a data frame with 2000 observations generated using *coxed* R library. It contains 10 variables, with 30% of censored data. For SimulatedTWO the same library has been used, with the same characteristics. The only change that was introduced was that 80% of censored data was considered. The NKI Breast Cancer Data includes survival data of 272 breast cancer patients[1] with 1570 columns, which is the most extensive dataset we analyzed.

Our empirical analysis provides insights into the performance of different machine learning models for survival analysis and their suitability for various types of datasets. For the evaluation and comparison of models we have used the Cindex, which allows a comparative evaluation of the predictive capacity of different survival machine learning models. The evidenced results imply that the DeepSurv model performed the best in two cases, with a Cindex of 0.84512 in the first test. This indicates that the model's predicted probabilities of the event occurring are well-calibrated and the model can effectively differentiate between patients who will experience the event and those who will not and the time when it happens.

[1] NKI Breast Cancer Data: https://data.world/deviramanan2016/nki-breast-cancer-data

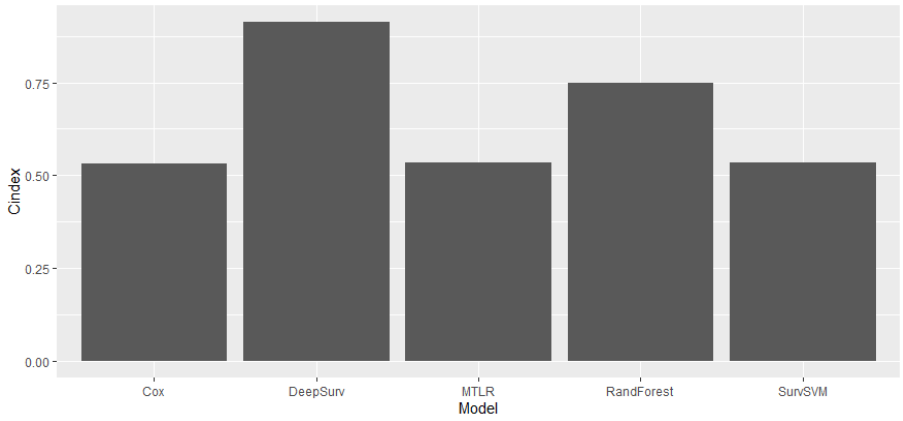

Figure 1: results from different machine learning models on SimulatedONE dataset.

The other models, including Cox, MTLR, Random Forest, and Kernel SVM, all performed reasonably well but not as well as DeepSurv. The differences in performance could be due to various factors, such as differences in the model's architecture, the type of loss function used, or the hyperparameters chosen during training. Something happens when we change the percentage of censored data. We have noticed that the performance of the models, with the same characteristics (same hyperparameters) have a different absolute and relative performance. It is evident what happens when a dataset similar characteristic, but the only differential is the percentage of censored data.

As we noted, in the case of SumulatedONE and SimulatedTWO the relative performance of the DeepSurv algorithm has decreased notably. We can conclude that this algorithm was affected by the volume of censored data, in line with what we have presented in the previous analysis.

It is good to note at this moment that the standard parameters have been used to develop each of the models. It would be desirable, in future investigations, to deepen the analysis that certain changes could generate in some of the models hyperparameters.

3. Conclusions and Next Steps.

To summarize, although machine learning algorithms have made tremendous progress in resolving some of the shortcomings of classic survival analysis approaches, there is still room for advancement. Overfitting, interpretability, data amount and quality, and temporal effects in variables are among shortcomings shared by all machine learning models. Although ensemble machine learning systems have been created to mitigate these flaws, the most complicated models remain hyperparameter sensitive.

Furthermore, since access to longitudinal or supplemental data is restricted, many assumptions cannot be checked ex post, posing a challenge to the validity of the findings. Nonetheless, by changing parameters as they learn from fresh data, machine learning algorithms provide huge productivity increases.

To address these constraints, future research should concentrate on building more resilient machine learning algorithms that can manage vast amounts of data while maintaining accuracy over time, as well as enhancing access to complementary and longitudinal data. Furthermore, the problem of interpretability must be addressed, ensuring that models can be understood and confirmed by domain experts.