





BART-Survival: A Bayesian machine learning approach to survival analyses in Python

Jacob Tiegs ^{1,2}, Julia Raykin ¹, and Ilia Rochlin ¹

¹ Inform and Disseminate Division, Office of Public Health Data, Surveillance, and Technology, Centers for Disease Control and Prevention, Atlanta, Georgia, United States of America ² Metas Solutions, Atlanta, Georgia, United States of America  Corresponding author

DOI: [10.21105/joss.07213](https://doi.org/10.21105/joss.07213)

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [AHM Mahfuzur Rahman](#)

Reviewers:

- [@turgeonmaxime](#)
- [@rich2355](#)

Submitted: 12 August 2024

Published: 28 January 2025

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).

Summary

BART-Survival is a Python package that allows time-to-event (survival) analyses in discrete-time using the non-parametric machine learning algorithm, Bayesian Additive Regression Trees (BART). BART-Survival combines the performance of the BART algorithm with the complementary data and model formatting required to complete the survival analyses. The library contains a convenient application programming interface (API) that allows a simple approach when using the library for survival analyses, while maintaining capabilities for added complexity when desired. The package is intended for analysts exploring use of flexible non-parametric alternatives to traditional (semi-)parametric survival analyses.

Statement of need

Survival analyses are a cornerstone of public health and clinical research in such diverse fields as cancer, cardiovascular disease, and infectious diseases ([Altman & Bland, 1998](#); [Bradburn et al., 2003](#)). Traditional parametric and semi-parametric statistical methods, such as the Cox proportional hazards model, are commonly employed for survival analyses ([Cox, 1972](#)). However, these methods have several limitations, particularly when applied to complex data. One major issue is the need for restrictive assumptions, such as proportional hazards and predefined functional forms, which may not hold true in complex, real-world healthcare data ([Harrell, 2015](#); [Ishwaran et al., 2008](#)). Additionally, these methods often struggle with high-dimensional datasets, leading to problems with over-fitting, multi-collinearity, and dealing with complex interactions ([Ishwaran et al., 2008](#); [Joffe et al., 2013](#)).

More recently, non-parametric machine learning approaches have been introduced to address the limitations of the traditional methods ([Harrell, 2015](#); [Ishwaran et al., 2008](#)). BART is a one such approach that has demonstrated exceptional performance across a variety of analytic settings, typically outperforming the traditional methods in terms of predictive accuracy. BART's performance is linked to its ability to flexibly model complex non-linear and variable interactions within the data, while being inherently regularized to reduce issues of over-fitting. The BART method is fully non-parametric and can adaptively model data complexities without prior knowledge or specification of a particular functional form. Finally, the method is generally accepted as being a user-friendly machine learning approach, as it typically requires minimal hyperparameter tuning and the outcomes can be easier to interpret than those produced by other similar methods ([Chipman et al., 2010](#); [R. Sparapani et al., 2021](#); [R. A. Sparapani et al., 2016](#)).

Currently, the only BART survival algorithm readily available exists as part of the BART R package, which contains a library of various BART-based approaches in addition to a BART survival analysis application ([R. Sparapani et al., 2021](#); [R. A. Sparapani et al., 2016](#)). The

BART-Survival package described here combines the survival analysis approach outlined in the BART R package with the foundational Python-based probabilistic programming language library, PyMC (Abril-Pla et al., 2023), and the accompanying BART algorithm from the PyMC-BART library (Quiroga et al., 2023). Our aim in developing BART-Survival is to provide accessibility to the BART survival algorithm within the Python programming language. This contribution is beneficial for analysts when Python is the preferred programming language, the analytic workflow is Python-based, or when the R language is unavailable for analyses.

The need for a complete BART-Survival python package is given by the simple fact that the BART survival algorithm is non-trivial to implement. Both the required data transformations and the internal model definition requires precise implementations to ensure generation of accurate survival models. Our BART-Survival library provides accessibility to these precise methods while removing the technical barriers that would limit user adoption of the BART survival approach.

More specifically, the BART-Survival library abstracts away the complexities of generating the proper training and inference datasets, which are conceptually complex and prone to being specified incorrectly if implemented from scratch. Similarly, the BART-Survival library provides a pre-specified internal Bayesian model using the PyMC probabilistic programming language. This pre-specified model is primarily accessed through the BART-Survival API removing the requirement for users to have more than a cursory knowledge of the PyMC or PyMC-BART libraries. Since the BART-Survival package is intended for students and professional in the public health and clinical fields, it is expected that users of the BART-Survival library will not have extensive programming expertise, adding to the need for a fully self-contained and accessible approach.

In summary, the BART-Survival package provides a simple and accessible approach to implementing the BART survival algorithm. The provided approach can be beneficial for users who are looking for non-parametric alternatives to traditional (semi-)parametric survival analysis. The BART survival algorithm can be especially useful in large, complex healthcare data, where machine learning methods can demonstrate improved performance over the traditional methods.

Conclusion

BART-Survival provides the computational methods required for completing non-parametric discrete-time survival analysis. This approach can have several advantages over alternative survival methods. These advantages include capabilities to incorporate non-linear and interaction effects into the model, naturally ability to regularize the model (which reduces the risk of over-fitting) and of being robust to issues of multi-collinearity. The BART-Survival approach is especially useful when the assumptions of alternative survival methods are violated.

Our BART-Survival algorithm has been tested in a rigorous simulation study, with additional applications to real-world data. While the manuscript for this work is currently under development, the results indicate similar performance as the the R-based BART survival method across settings of varied complexity. Both methods demonstrate the previously describe advantages over other survival approaches (such as Cox Proportional Hazard Models) when the relationships within the data becomes more complex or assumptions of the these other models are violated. A comparison of the R-based method and our BART-Survival algorithm is included in the [examples folder of the Github repository](#).

Our library provides a convenient API for completing discrete-time survival analysis, along with the functionality to customize the methodology as needed. The associated API documentation can be found [here](#), along with the associated Github repository [BART-Survival](#). An extended review of the methods is additionally provided within the documentation and can be found [here](#).

Acknowledgements

We thank Oscar Rincón-Guevara for helpful suggestions and review. We also thank Tegan Boehmer, Sachin Agnihotri, and Matt Ritchey for supporting the project throughout its development.

References

- Abril-Pla, O., Andreani, V., Carroll, C., Dong, L., Fonnesebeck, C. J., Kochurov, M., Kumar, R., Lao, J., Luhmann, C. C., Martin, O. A., Osthege, M., Vieira, R., Wiecki, T., & Zinkov, R. (2023). PyMC: A modern, and comprehensive probabilistic programming framework in Python. *PeerJ Computer Science*, 9, e1516. <https://doi.org/10.7717/peerj-cs.1516>
- Altman, D. G., & Bland, J. M. (1998). Statistics Notes: Time to event (survival) data. *BMJ*, 317(7156), 468–469. <https://doi.org/10.1136/bmj.317.7156.468>
- Bradburn, M. J., Clark, T. G., Love, S. B., & Altman, D. G. (2003). Survival Analysis Part II: Multivariate data analysis – an introduction to concepts and methods. *British Journal of Cancer*, 89(3), 431–436. <https://doi.org/10.1038/sj.bjc.6601119>
- Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1). <https://doi.org/10.1214/09-AOAS285>
- Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 34(2), 187–202. <https://doi.org/10.1111/j.2517-6161.1972.tb00899.x>
- Harrell, F. E. (2015). *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Springer International Publishing. <https://doi.org/10.1007/978-3-319-19425-7>
- Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, 2(3). <https://doi.org/10.1214/08-AOAS169>
- Joffe, E., Coombes, K. R., Qiu, Y. H., Yoo, S. Y., Zhang, N., Bernstam, E. V., & Kornblau, S. M. (2013). Survival Prediction In High Dimensional Datasets – Comparative Evaluation Of Lasso Regularization and Random Survival Forests. *Blood*, 122(21), 1728–1728. <https://doi.org/10.1182/blood.V122.21.1728.1728>
- Quiroga, M., Garay, P. G., Alonso, J. M., Loyola, J. M., & Martin, O. A. (2023). *Bayesian additive regression trees for probabilistic programming* (No. arXiv:2206.03619). arXiv. <https://doi.org/10.48550/arXiv.2206.03619>
- Sparapani, R. A., Logan, B. R., McCulloch, R. E., & Laud, P. W. (2016). Nonparametric survival analysis using Bayesian Additive Regression Trees (BART). *Statistics in Medicine*, 35(16), 2741–2753. <https://doi.org/10.1002/sim.6893>
- Sparapani, R., Spanbauer, C., & McCulloch, R. (2021). Nonparametric Machine Learning and Efficient Computation with Bayesian Additive Regression Trees: The **BART** R Package. *Journal of Statistical Software*, 97(1). <https://doi.org/10.18637/jss.v097.i01>