Explainable AI for survival analysis: a median-SHAP approach

Anonymous Authors[1]

Abstract

With the adoption of machine learning into routine clinical practice comes the need for Explainable AI methods tailored to medical applications. Shapley values have sparked wide interest for locally explaining models. Here, we demonstrate their interpretation strongly depends on both the summary statistic and the estimator for it, which in turn define what we identify as an 'anchor point'. We show that the convention of using a mean anchor point may generate misleading interpretations for survival analysis and introduce median-SHAP, a method for explaining black-box models predicting individual survival times.

1. Introduction

Recent years have seen a rapid growth in the Explainable AI (XAI) literature. Given the great variety of explanation models and the diversity of interpretations they can generate (Sundararajan & Najmi, 2020), decision-makers are increasingly demanding that XAI methods be designed for a single application. Contrasting with one-size-fits all solutions, such tailored XAI techniques would better fulfill the precise requirements of the task they're meant for (Arrieta et al., 2020). Here, we introduce a specific method for explaining black-box survival analysis models, which is based on Shapley values.

Shapley values (SV) have become a gold standard for local model explainability. They compute feature attributions by quantifying the change in model output when dropping certain feature values and sampling them from a reference distribution. The choice of a reference probability distribution on feature values has been the subject of multiple debates (Aas et al., 2019; Janzing et al., 2020; Frye et al., 2020). In contrast, we argue that a key challenge of SV arises from the convention of both using (i) the expectation as a summary statistic for the distribution of model outcomes and (ii) the sample mean to estimate it, and that this question has been overlooked. Most importantly, we show these parameters jointly induce a mean 'anchor point' against which comparisons are made, and how such an anchor point isn't adapted the time-to-event models. Ultimately, we demonstrate that using the original formulation of Shapley values can generate misleading interpretations when the distribution of the value function is right-skewed, which is the case for survival analysis models as some individuals may not be experiencing the event throughout the study (Rao & Schoenfeld, 2007; Ying et al., 1995).

To resolve this problem, we introduce median-SHAP, which uses the median as the summary statistic and for estimation. We demonstrate the benefits of this approach for explaining survival analysis models that output predicted survival times from a set of features. Note that such models differ from inferential methods (e.g. the Cox proportional hazards model (Lin & Wei, 1989)) for understanding the drivers of survival. For point prediction models, the median predicted survival time is more commonly reported than the mean.

To the best of our knowledge, our method is the first additive feature attribution method tailored to survival analysis. In contrast, SurvLIME (Utkin et al., 2020; Kovalev et al., 2020) is a local linear approximation based on the Cox proportional hazards model.

Contributions The gaps in the existing literature motivate the following contributions of this paper:

1. We shed a new light on the discussion around SV by interrogating the notions of summary statistics and mean estimators within Shapley values and the subsequent 'anchor points' they induce. We identify the anchor point as the instance our observation of interest is compared against, and explain why it is essential for the interpretation of SV.
2. We show how using a mean anchor point can generate misleading explanations of survival analysis models
3. We introduce median-SHAP, an explanation model specific to survival analysis. Our method is based on observational SV and uses the median as a summary statistic. We experimentally show that our method has improved interpretability and robustness compared with the original SV.

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

2. Shapley Values

We consider a model $f : \mathbb{R}^m \to \mathbb{R}^l$ and aim to explain the prediction of an instance $x \in \mathbb{R}^m$, given only black-box access to the model. Shapley values quantify the contribution of input features $\{1, \dots, m\}$ to the prediction of a complex model $f : \mathbb{R}^m \to \mathbb{R}^l$ at an instance x as follows:

$$f(x) = \phi_0^f + \sum_{i=1}^{M} \phi_i^f(x) \tag{1}$$

where $\phi_j^f(x)$ is the Shapley value of feature j to $f(x)$ and $\phi_0^f = \mathbb{E}[f(X)]$ is the prediction averaged over the observed data distribution. The attribution of a feature j is computed from the difference in value function v_f comparing when the feature j is equal to the value x_j with when it is removed from the coalition. We denote this difference $\Delta v_f(S, j, x) = v_f(S \cup \{j\}, x) - v_f(S, x)$. When a feature is included in the coalition its value is set to the observed instance value x_S. When a feature is not in the coalition its value is sampled from a reference distribution $r(X^* \mid x_S)$.

$$v_f(S) = \mathbb{E}_{r(X^* \mid x_S)}[f(x_S, X^*_{\overline{S}})]$$

for $\overline{S} := \{1, \dots, m\} \backslash S$ with $(x_S, x_{\overline{S}})$ denoting the concatenation of its two arguments. To account for the dependence with other features, one takes the difference in value function v averaged over all possible coalitions S of features excluding feature j. Ultimately, a binomial weight $\frac{|S|!(m-|S|-1)!}{m!}$ is added to recover the original Shapley values which account for all possible orderings. All in all, the Shapley values of feature j is defined as follows:

$$\begin{aligned} \phi_j^f(x) &= \frac{1}{m} \sum_{\substack{|S| \in \\ \{0, \dots, m-1\}}} \left[\frac{1}{\binom{m-1}{|S|}} \sum_{\substack{S \subseteq \{1, \dots, m\}/j \\ \text{with } |S| = j}} \Delta v_f(S, j, x) \right] \\ &= \mathop{\mathbb{E}}_{S} \left[\mathop{\mathbb{E}}_{r(X^* \mid x_S)} [f(x_S, X^*_{\overline{S}})] \right] \end{aligned}$$

Further, note that the two expectations can be taken in whichever order. We refer to the *expectation as a summary statistic* to describe the fact that the distribution of model outcomes (whether over coalitions or reference points) is summarised using the central location.

The choice of a reference distribution for SV has been subject to recent debates (Aas et al., 2019; Janzing et al., 2020; Merrick & Taly, 2020). Interventional SV (Lundberg & Lee, 2017; Janzing et al., 2020) define $r(X^* \mid x_S) := p(X^*)$ where p denotes the marginal data distribution. Observational Shapley values (Aas et al., 2019) set the reference distribution equal to the conditional distribution given x_S $r(X^* \mid x_S) := p(X^* | X^*_S = x_S)$. Observational SV are often described as 'true to the data' because they do not break the correlations between features. Moreover, this makes them robust to adversarial attacks, as their computation does not involve evaluations on out-of-distribution instances i.e. concatenated inputs which were built by sampling from a marginal distribution.

At this stage, it is important to note that, in reality, SV assess the importance of a feature *value*, and not a feature. In other words, the central question of SV is 'By how much did the feature j shift the local model's prediction $f(x)$ away (in any direction) from the average prediction ϕ_0^f, once feature j was set to x_j?'. By implementing Equation 1 – also known as the *Efficiency* property – we are able to derive bespoken drift from the SV, e.g. 'mode shift'. Consequently, as feature attributions result from comparisons between $f(x)$ and ϕ_0^f, our instance x is being compared with a synthetic input for which the model prediction is equal to $\phi_0^f(x)$. We call this synthetic input the *anchor point* in the following. Such comparisons are not readily interpretable as the anchor point is not an actual realisation of the population.

Finally, given that Shapley values are Monte Carlo approximated in practice, the expectation over the reference distribution is estimated by the mean over a finite set of L observations $\{x^*_{i,\overline{S}}\}_{i=1}^{L}$ from the reference distribution, henceforth called 'reference points' $r(X^*_S \mid x_S)$,

$$\begin{aligned} \widehat{\phi}(j) = \frac{1}{m} \sum_{S \subseteq \{1, \dots, m\}/j} [\frac{1}{\binom{m-1}{|S|}} (\frac{1}{L} \sum_{i=1}^{L} f(x_{S \cup j}, x^*_{i, \overline{S}/j}) \\ - \frac{1}{L} \sum_{i=1}^{L} f(x_S, x^*_{i, \overline{S} \cup j}))]. \end{aligned}$$

We will refer to this as using the *sample mean as an estimator* of the expectation, or central location.

3. Summary statistics, estimators, and anchor points

In the following, we differentiate between the drawbacks of the expectation as a summary statistic, the sample mean over reference points as its estimator and the previously defined an anchor point.

3.1. Drawbacks of the expectation as a summary statistic

When averaging the difference in value function Δv_f over all coalitions, the original Shapley formulation summarises the distribution of the model outcomes of concatenated inputs by taking their expectation. First, taking the expectation to summarise central location is problematic when the distribution is skewed. Second, the value of the expectation – and not its relative position within the sample distribution – can lead to misleading interpretations. For instance, if the single reference values are centred around zero, this may give the impression that the corresponding feature has a negligible impact on the model outcome. For instance, consider a model

$$f(x) = 1000x_1 + x_2$$

at observation $x = (0, 1)$ and where marginally $x_1 \sim \mathrm{Normal}(0, 1)$ and $x_2 \sim \mathrm{Normal}(0, 1)$. Adding x_1 to the empty coalition does not, on average, make a difference in terms of model outcome as on the one hand we marginalise over x_1 which is centered on 0, and on the other hand we specify $x_1 = 0$. Similarly, model outcomes are be equal in expectation for the coalitions $\{x_2\}$ and $\{x_1, x_2\}$. (Fryer et al., 2021) argue this is why Shapley values should not be used for feature selection. We then have Shapley values:

$$\phi^f_{X_1}(x_1, x_2) = 0 \qquad \phi^f_{X_2}(x_1, x_2) = 1$$

which can be misleading. Challenges can also arise from taking the expectation over feature coalitions, as shown by Merrick et.al (2020). To solve this problem, the authors propose to cluster the single reference Shapley values and return the cluster means. Nonetheless, multidimensional clustering algorithms used on one-dimensional data have several shortcomings, including the fact that two clusters might be of considerably different size. Instead we remark that since we are interested in summarising a distribution, we should thus use summary statistics that are robust to distribution shifts, such as quantiles.

3.2. Drawbacks of the mean estimator

Using the sample mean over reference points as an estimator of central location is challenging when in the presence of heavy-tailed data distributions and outliers. Consider, for instance a black-box model

$$f(x) = x + \epsilon$$

where ϵ is Cauchy noise. Here, the Shapley value at x should be ∞ in theory. In practical applications, an adversary might return extreme values for unlikely observations, i.e. $f(x) = -x - 1000 \cdot \mathbb{I}(x > 0.99), x \sim \mathrm{Uniform}[0, 1]$. The extreme values with small probability push the mean change in model outcome of including the feature to the positive at any observation with x outside of $[0.99, 1]$. Such an effect is all the more likely to happen for marginal Shapley values, as the model estimator can suffer under high variance.

3.3. Drawbacks of a global mean 'anchor point'

Taking the expectation as a summary statistic and the sample as its estimator together imply that comparisons are made w.r.t. the global mean, which acts as an anchor point. This has negative impacts on the interpretability of the resulting attributions as the average model outcome over reference points does not link to any single particular individual. Put differently, in a clinical example we would compute the respective SV for a patient comparing them with an 'artificial' patient whose model outcome would be equal to the global mean ϕ_0. Further, when this distribution is marginal, the anchor point might not relate to a plausible instance. This poses a problem for clinical management, as patients may want to understand the influence of modifiable features (e.g. smoking, physical activity) to lengthen their survival.

4. median-SHAP for predictions of median survival times

Our proposed method is a local explanation model based on *observational* Shapley values tailored to black-box models that predict a median survival time. median-SHAP takes the median of the expected change in model outcome for a coalition S, that is

$$\widehat{\phi}(j)^{med} = \frac{1}{m} \sum_{S \subseteq \{1,\dots,m\}/j} [\frac{1}{\binom{m-1}{|S|}}$$
$$(med(\{f(x_{S\cup j}, x^*_{i,\overline{S}/j})\}_{i=1}^{L}) - med(\{f(x_S, x^*_{i,\overline{S}})\}_{i=1}^{L})]$$

Here, like in the original Shapley, each ϕ_j may be interpreted as the contribution of feature value x_j to the shift $f(x) - \phi_0$. However in median-SHAP ϕ_0 is the median prediction from f in the reference population. Therefore comparing $f(x)$ with ϕ_0 implies we're comparing our individual x with the median *predicted* individual i.e. the individual with median prediction in the cohort. Ultimately, with median-SHAP the anchor point is the median individual, and thus an observed data point. Using a conditional reference distribution further ensures that our black-box isn't evaluated off the data manifold i.e. concatenated instances aren't out of distribution. Contrastively, using a marginal reference distribution would jeopardise the interpretation by breaking the correlation and defining an out-of-distribution anchor point. Ultimately, using the median over references makes our estimation robust to outliers and skewed distribution compared to using the sample mean.

Given the clinical focus of our targeted application, we chose observational Shapley values to preserve the robustness to adversarial attacks and thus maintain the high safety standards needed for medical purposes. Further, note that in the case of a binary classification problems (e.g. having a survival time above/below a threshold) and assuming that the data set $\{f(x_{S\cup j}, x^*_{i,\overline{S}/j})\}_{i=1}^{L}$ is balanced, the median model outcome will be a realisation at the decision boundary. Such an approach is easier (but also more restricted) than a linear search for the decision boundary with e.g. b perturbations used for LIME (White & Garcez, 2019). In contrast to $l1$ and $l2$ regularisation, additivity to $f(x)$ of the Shapley values is preserved with median-SHAP. Moreover, while Shapley values lose their intuition when regularisation is added, the median Shapley values can be interpreted as the expected change in median model outcome when a feature is added.

5. Experiments

We demonstrate the benefits of using median-SHAP by doing the following experiment on two datasets. We use the entire set as a training set, and a reference distribution.

1. Training a regression model f on (X, Y) where X is a set of features and Y is a median survival time. We denote $med = \text{Median}[f(X)]$.

2. Using SHAP and median-SHAP to generate feature attributions for the outcome of a given individual x .

3. 'Re-labeling' the outcome of all individuals in the dataset to create a classification problem, such that: $Y' = 1$ if $f(x) > med$ and $Y' = 0$ otherwise.

4. Training a classification model g on (X, Y').

5. Using SHAP to compute feature attributions explanations for the classification of x by our model g

We repeat this experiment for randomly sampled individuals and report the relative differences (i) between median-SHAP for f and SHAP for g (ii) between SHAP for f and SHAP for g. This process may attest if median-SHAP captures which are the most important features for predicting how an individual *ranks* within the cohort, w.r.t. a model's predictions. In other words, if a feature is important for model g it means that it is important to predict how our individual ranks w.r.t. the median individual. Therefore, if median-SHAP for f computes feature attributions similar to the Shapley values for g, this confirms our method captures how the instance of interest compares with the anchor point, or median individual here. Details regarding the experiments and the datasets are shown in the Supplements B.

5.1. The Worcester Heart Attack Study

The Worcester Heart Attack Study (Whas) is a longitudinal study on on acute myocardial infarction which looks at patient features at presentation (Goldberg et al., 2000). A Random Survival Forest (Ishwaran et al., 2008) – which is a classification tree method for the analysis of right-censored survival data – was trained on the entire set (N=500) to predict time of death, which occurred for 215 patients (43.0%).

Here, feature attributions should be interpreted comparing each instance of interest with the median patient, who is a 59 year old individual with a BMI of 29.3 kg/m^2, a heart rate of 117 bpm and who has experienced congestive heart failure. Table 1 shows the results of our experiment on the Whas dataset.

Shapley type	Age	BMI	Chf	Heart Rate
SHAP on f vs SHAP on g	7.84	2.1	-0.54	6.51
median-SHAP on f vs SHAP on g	1.87	-0.1	-1.03	1.22

Table 1. Mean difference of scaled feature attributions for the Whas data comparing (i) mean Shapley values on f vs mean Shapley values on the classification model g, (ii) median-SHAP on f vs mean Shapley values on the classification model g.

5.2. The breast cancer survival data

The breast cancer survival data (Desmedt et al., 2007) is a longitudinal study which explores the time dependence between a genetic prognostic signature and the occurrence of distant metastases. A Random Survival Forest was trained on 198 individuals to predict the time of potential metastases, which occurred for 51 patients (25.8%). Resulting feature attributions are described in table 2, and should be interpreted as comparisons with the median individual who has the following features: `X202240 = 6.96`, `X214806 = 7.32`, `X219724_s = 6.33` and `X203306_s = 10.51` and has a predicted median survival time of 51 time units.

For both experiments, feature attributions computed using median-SHAP are closer to the ones for the classification task than the original mean Shapley values. This further demonstrates that median-SHAP is able to capture which features are predominant when predicting how our individual compares with the rest of the cohort, and more specifically here how it compares with the median individual. This further illustrates the applicability of our method in this context. Ultimately, note that the median is more appropriate as an estimator for such examples with a small sample size, due to its robustness to outliers.

Shapley type	X202240	X219724_s	X214806	X203306_s
SHAP on f vs SHAP on g	0.04	-3.32	-2.54	10.51
median-SHAP on f vs SHAP on g	<0.001	-1.41	-3.57	-1.28

Table 2. Mean difference of scaled feature attributions for the breast cancer survival data comparing (i) mean Shapley values on f vs mean Shapley values on the classification model g, (ii) median-SHAP on f vs mean Shapley values on the classification model g.

6. Concluding remarks

We introduce median-SHAP, a specific explanation method for survival models. In addition to the robustness of the median estimator w.r.t. outliers and its higher likelihood for skewed distributions, median-SHAP offers increased interpretability as comparisons are made with an actual individual from the reference population. Ultimately, we advocate that median-SHAP has improved clinical applicability as it allows physicians to illustrate their recommendations using person-centric explanations.

References

Aas, K., Jullum, M., and Løland, A. Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *arXiv preprint arXiv:1903.10464*, 2019.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

Desmedt, C., Piette, F., Loi, S., Wang, Y., Lallemand, F., Haibe-Kains, B., Viale, G., Delorenzi, M., Zhang, Y., d'Assignies, M. S., et al. Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the transbig multicenter independent validation series. *Clinical cancer research*, 13(11):3207–3214, 2007.

Frye, C., de Mijolla, D., Cowton, L., Stanley, M., and Feige, I. Shapley-based explainability on the data manifold. *arXiv preprint arXiv:2006.01272*, 2020.

Fryer, D., Strümke, I., and Nguyen, H. Shapley values for feature selection: The good, the bad, and the axioms. *arXiv preprint arXiv:2102.10936*, 2021.

Goldberg, R. J., Yarzebski, J., Lessard, D., and Gore, J. M. Decade-long trends and factors associated with time to hospital presentation in patients with acute myocardial infarction: the worcester heart attack study. *Archives of Internal Medicine*, 160(21):3217–3223, 2000.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. Random survival forests. *The annals of applied statistics*, 2(3):841–860, 2008.

Janzing, D., Minorics, L., and Blöbaum, P. Feature relevance quantification in explainable ai: A causal problem. In *International Conference on Artificial Intelligence and Statistics*, pp. 2907–2916. PMLR, 2020.

Kovalev, M. S., Utkin, L. V., and Kasimov, E. M. Survlime: A method for explaining machine learning survival models. *Knowledge-Based Systems*, 203:106164, 2020.

Lin, D. Y. and Wei, L.-J. The robust inference for the cox proportional hazards model. *Journal of the American statistical Association*, 84(408):1074–1078, 1989.

Lundberg, S. and Lee, S.-I. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.

Merrick, L. and Taly, A. The explanation game: Explaining machine learning models using shapley values. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pp. 17–38. Springer, 2020.

Rao, S. R. and Schoenfeld, D. A. Survival methods. *Circulation*, 115(1):109–113, 2007.

Sundararajan, M. and Najmi, A. The many shapley values for model explanation. In *International Conference on Machine Learning*, pp. 9269–9278. PMLR, 2020.

Utkin, L. V., Kovalev, M. S., and Kasimov, E. M. Survlime-inf: A simplified modification of survlime for explanation of machine learning survival models. *arXiv preprint arXiv:2005.02387*, 2020.

White, A. and Garcez, A. d. Measurable counterfactual local explanations for any classifier. *arXiv preprint arXiv:1908.03020*, 2019.

Ying, Z., Jung, S.-H., and Wei, L.-J. Survival analysis with median regression models. *Journal of the American Statistical Association*, 90(429):178–184, 1995.