mathematics

MDPI

Article

Sensitivity of Survival Analysis Metrics

Iulii Vasilev *, Mikhail Petrovskiy * and Igor Mashechkin *

Computer Science Department, Lomonosov Moscow State University, Vorobjovy Gory, 119899 Moscow, Russia
* Correspondence: iuliivasilev@gmail.com (I.V.); michael@cs.msu.su (M.P.); mash@cs.msu.su (I.M.)

Abstract: Survival analysis models allow for predicting the probability of an event over time. The specificity of the survival analysis data includes the distribution of events over time and the proportion of classes. Late events are often rare and do not correspond to the main distribution and strongly affect the quality of the models and quality assessment. In this paper, we identify four cases of excessive sensitivity of survival analysis metrics and propose methods to overcome them. To set the equality of observation impacts, we adjust the weights of events based on target time and censoring indicator. According to the sensitivity of metrics, *AUPRC* (area under Precision-Recall curve) is best suited for assessing the quality of survival models, and other metrics are used as loss functions. To evaluate the influence of the loss function, the *Bagging* model uses ones to select the size and hyperparameters of the ensemble. The experimental study included eight real medical datasets. The proposed modifications of *IBS* (Integrated Brier Score) improved the quality of *Bagging* compared to the classical loss functions. In addition, in seven out of eight datasets, the *Bagging* with new loss functions outperforms the existing models of the scikit-survival library.

Keywords: machine learning; survival analysis; Kaplan–Meier estimator; recursive partitioning; model averaging

MSC: 62N02

check for updates

Citation: Vasilev, I.; Petrovskiy, M.; Mashechkin, I. Sensitivity of Survival Analysis Metrics. *Mathematics* **2023**, *11*, 4246. https://doi.org/10.3390/math11204246

Academic Editor: Yiqiang Zhan

Received: 27 August 2023
Revised: 22 September 2023
Accepted: 27 September 2023
Published: 11 October 2023

1. Introduction

Models of event forecasting are important for describing the causes and effects of various phenomena and processes. Survival analysis is a set of methods for estimating the probability of occurrence of an event in time. The definition of an event varies depending on the application area.

For example, in the case of healthcare, the event is a fatal outcome, relapse, or recovery of the patient. Clinical data and anamnesis is used to predict disease history and prescribe treatment with minimal risks of negative consequences. In the case of reliability analysis, the event is a failure of the equipment. The history of the event allows for balancing the load or replacing individual components at an early stage to prevent the failure of the entire system.

Papers [1–3] emphasize the main features of survival analysis. The usage of incomplete data leads to the appearance of censored observations with an unknown time of the event (for example, the event does not occur before the end of the study). In this paper, we consider only right-censoring cases with a fixed beginning time and expect the event to the right. In addition, the distribution of events in time is important. There are several split criteria [4] with high sensitivity to early events and [5–7] late events. However, the models assume a constant sensitivity of events and do not take into account the event distribution of the source data.

Finally, the proportion of censored and terminal classes is the cause of bias in forecasting. The most popular approach to overcome the imbalance is data balancing (increasing the minor class or decreasing the dominant class). However, balancing leads to a bias of a priori probabilities of classes.

The influence of the highlighted characteristics applies not only to predictive models but also to quality metrics. In classic machine learning, metrics allow evaluation of the quality of models and are used as a loss function to solve optimization problems during model fitting. Finally, metrics influence the selection of the optimal model and its hyperparameters. Metrics of survival analysis cover the set of predicted values: point estimates (probability and time of occurrence of events) and integral estimates (history of occurrence of an event, cumulative risk in time).

The purpose of this work is to analyze the sensitivity of survival analysis metrics. We propose the classification of the sensitivity biases and modifications of metrics to overcome them. In addition, we explore the relationship between the loss function and the quality of the *Bagging* model. The *Bagging* model uses a loss function to select the size and hyperparameters of the ensemble. According to the analytical and experimental results, we define the best metrics as quality score and loss function. A critical analysis of the advantages and disadvantages of existing metrics allows us to assess the quality of models reliably and motivate researchers to conduct additional testing of the stability of metrics to data characteristics.

The paper is organized as follows: Section 2 presents an overview of survival analysis models and describes the motivation for choosing quality metrics for each type of predicted values. Section 4 describes the characteristics of open medical datasets, the presence of early and late events, and class imbalance. Section 3 describes the main steps of the sensitivity study. Section 5 classifies the causes of the excessive sensitivity of metrics and proposes relevant examples for their detection, such as the higher significance of particular events, metric changes over time, the significance of the time scale, and the influence of the imbalance of classes. In addition, we propose modified metrics to improve sustainability and select a reliable quality metric. Section 6 provides an experimental study of the relationship between modified loss functions and the quality of models. The result is the selection of the best loss functions for the expansion of the *Bagging* ensemble. Section 7 presents the main results of the work and directions for further research.

2. Background

2.1. Problem Statement

Let X denote a random vector of variables, T a non-negative random variable of event time, C a non-negative random variable of censoring time, and δ an event occurrence indicator. In this case, the target time, y_i, of the event is:

$$y_i = \begin{cases} T_i, & \text{if } \delta_i = 1, \\ C_i, & \text{if } \delta_i = 0. \end{cases}$$

Thus, the task is reduced to analyzing triplets (X_i, y_i, δ_i) for each observation, i, where y_i and δ_i are the target variables. Using variable X_j, the goal is to predict the true time, T_j, which is hidden for censored observations.

To predict the history of the occurrence of an event, the task of survival analysis is reduced to three functions [2]. The survival function determines the probability of non-occurrence of an event after a certain time:

$$S(t) = P(T > t),$$

where t is the observation time and T is a random variable of the event time.

The death density function determines the probability of an event occurring at a specific time, $t \in \mathbb{R}$:

$$f(t) = (1 - S(t))'$$

The hazard function or the conditional failure rate represents the probability of an event at a particular time, t, given that the event did not occur earlier:

$$h(t) = \frac{f(t)}{S(t)}.$$

Depending on the task, the survival analysis functions can be formulated in a continuous and discrete form [2,8]. The continuous-time problem uses the entire time scale, and the observations X_i correspond to the source event time, T_i. Then, the survival analysis functions have the following form:

$$\begin{aligned} S(t) &= P(T > t), \\ f(t) &= -\frac{d}{dt} S(t), \\ h(t) &= -\frac{d}{dt} [\log S(t)]. \end{aligned}$$

Usually, continuous-time models have strict assumptions about the time distribution and the differentiability of the survival function on the timeline.

In a discrete-time problem, the original timeline is split into n specified time intervals (bins). Let τ denote an ascending ordered set of time points, then the time points, T, of the sample are mapped to the set τ. Despite that the discrete-time problem requires fewer assumptions, it also leads to information loss due to time discretization. In addition, the number of time intervals is a hyperparameter that strongly affects the accuracy and computational complexity of the model. Finally, the discrete-time problem imposes a serious functional limitation on the allowable time points for forecasting. Further, we consider only continuous-time models.

2.2. Statistical Models

2.2.1. Kaplan–Meier Estimator

The most popular nonparametric estimation of the survival function is the Kaplan–Meier method [9]. Denote the set of source times of the event as $\{t_i\}$. For each time point, t_i, there are the number of remaining observations, N_i, and the number of events, O_i, that occurred at time t_i. Then, the survival function at the moment of t is a cumulative product of survival probability for each previous time:

$$S(t) = \prod_{t_i \leq t} \left(1 - \frac{O_i}{N_i}\right).$$

The expected lifetime is t_i, so that $S(t_i) = 0.5$. In practice, it also needs to predict the survival function before the first occurrence of the event and after the occurrence of the last event. There are two ways to expand Kaplan–Meier estimation. In the first case (henceforth, KM), the survival function is equal to the last cumulative product after all events. However, if the latest observation is censored, the survival function does not reach zero. In the second case (henceforth, KM10), the survival function assigns 0 after all events and 1 before the first event.

2.2.2. Cox Proportional Hazard

Nonparametric methods do not take into account the relationship between the signs of observations and target variables. At the same time, parametric methods assume a theoretical relationship and determine the significance of features based on their impact on the forecast.

The Cox Proportional Hazards (henceforth, CoxPH) [10] assumes that all observations have the same form of the hazard function and differ by a positive coefficient of proportionality.

$$h(t \mid X_i) = h_0(t) \cdot e^{X_i \beta}, \tag{1}$$

where $h_0(t)$ is the base hazard function, X_i is the vector of features, and β is the vector of linear model coefficients. The importance of features (hazard ratio) calculates as e^{β}.

The survival function of CoxPH consists of basic survival function, $S_0(t)$ (usually is the Breslow estimator [11]), that is shifted with the weights, β:

$$S(t \mid X_i) = S_0(t)^{\exp(X_i\beta)}. \quad (2)$$

However, the method has several significant disadvantages:

- The ratio of two hazard functions does not change over time;
- The significance of features does not depend on time. In real clinical practice, the influence of risk factors may vary over time. For example, the patient is at risk after surgery, but after rehabilitation is more stable;
- The weights of the model define a linear combination of the data features;
- CoxPH does not support categorical features and missing values.

2.3. Metrics

2.3.1. Concordance Index

The Concordance index (CI) [12] is a ratio of correctly ordered pairs relative to the event time, which can be calculated as follows:

$$CI = \frac{\sum_{i,j} I(T_j < T_i) \cdot I(\eta_j < \eta_i)}{\sum_{i,j} I(T_j < T_i)},$$

where T_k is the true time of occurrence of the event, $I(\cdot)$ is the indicator function, and η_k is the time expected by the model. The best metric value is 1 (correct ordering), the worst value is 0 (opposite order), and the value 0.5 reflects the randomness of the model response.

The CI metric uses only point forecasts and does not evaluate survival function as a whole. In addition, the value of CI does not change with the shift of the survival functions, although the predicted time itself may be extremely biased compared to the true one.

2.3.2. Integrated AUC

An alternative ranking metric is the integrated area under the curve ($IAUC$) [13,14], which extends the calculation of the ROC curve and area under the curve (AUC) to multi-class or temporary cases.

Denote the cumulative hazard as $H(t)$. There is the following relationship between cumulative hazard and hazard function:

$$H(t \mid X) = \int_0^t h(\tau \mid X) d\tau.$$

For each event time, there are two sets of observations with early and late events. The $\widehat{AUC}(t)$ metric measures the weighted proportion of right-ordered pairs (by cumulative hazard) of observations from different sets (occurred events should have a higher cumulative hazard at time t) and can be calculated as follows:

$$\widehat{AUC}(t) = \frac{\sum_i \sum_j I(y_j > t) \cdot I((y_i \le t) \cdot \delta_i) \cdot w_i \cdot I(\hat{H}(t \mid X_j) \le \hat{H}(t \mid X_i))}{(\sum_j I(y_j > t))(\sum_i I((y_i \le t) \cdot \delta_i) \cdot w_i)},$$

where $\hat{H}(t \mid X_i)$ is the cumulative hazard estimation of X_i in time t and w_i is the inverse probability of censoring in time t_i ($w_i = 1/G(t_i)$, where $G(t)$ is the Kaplan–Meier model with censoring event).

Integrated AUC aggregates $\widehat{AUC}(t)$ over time with minimal value, t_{min}, maximum value, t_{max}, and general survival function, $\hat{S}(t)$, as the Kaplan–Meier model.

$$IAUC(t_{min}, t_{max}) = \frac{1}{\hat{S}(t_{min}) - \hat{S}(t_{max})} \int_{t_{min}}^{t_{max}} \widehat{AUC}(t) d\hat{S}(t). \tag{3}$$

2.3.3. Likelihood

In general, the likelihood function (henceforth, LL) is the joint distribution of the sample, considered as a function of a parameter. Denote the describing parameter of the predictive model as θ. Therefore, the likelihood function of the sample $\{X', T'\}$ is $LL(\theta \mid X', T') = \prod_i P_\theta(T_i \mid X_i)$.

Based on the survival function, $\hat{S}$, and the hazard function, $\hat{h}$, paper [15] denotes the full likelihood as $FL(\hat{S}, \hat{h} \mid X, T) = \prod_i \hat{h}(T_i|X_i)^{\delta_i} \cdot \hat{S}(T_i|X_i)$. To solve an optimization problem, the logarithmic form is:

$$\log FL(\hat{S}, \hat{h} \mid X, T) = \sum_i \delta_i \cdot \log\Big(\hat{h}(T_i|X_i)\Big) + \sum_i \log\big(\hat{S}(T_i|X_i)\big).$$

In addition, paper [15] uses only the hazard function, $\hat{h}$, and denotes the partial likelihood as $PL(\hat{h} \mid X, T) = \prod_i \frac{\hat{h}(T_i|X_i)}{\sum_{T_j \geq T_i} \hat{h}(T_i|X_j)}$. In addition, the corresponding logarithmic form is:

$$\log PL(\hat{h} \mid X, T) = \sum_i \log\Big(\hat{h}(T_i|X_i)\Big) - \log\left(\sum_{T_j \geq T_i} \hat{h}(T_i|X_j)\right).$$

2.3.4. Kullback–Leibler Divergence

Kullback–Leibler Divergence (henceforth, KL) [16] measures the distance between continuous probability distributions P and Q. On the set $X \subseteq \mathbb{R}^k$, denote the density functions of the distributions P and Q as $p(X)$ and $q(X)$, accordingly. Then, KL divergence is $KL(P||Q) = \int_0^\infty p(x) \log\left(\frac{p(x)}{q(x)}\right) dx$.

In survival analysis, paper [17] proposes a modification of KL divergence (henceforth, KLS). Let $G_n(t)$ denote the estimation of the survival function on n observations. Based on the family of two-parameter Weibull survival functions, the true function is $F(t) = \exp\left[-\left(\frac{t}{\sigma}\right)^m\right]$, where $t \geq 0$, and $m, \sigma > 0$.

$$KLS(G_n||F) = \int_0^\infty G_n(t) \cdot \log\left(\frac{G_n(t)}{F(t)}\right) - [G_n(t) - F(t)]dt.$$

The main disadvantage of KLS is the usage of the nonparametric $G_n(t)$ to estimate proximity to the theoretical Weibull function, $F(t)$. In such a case, KLS does not take into account individual features, X_i, and a censoring indicator.

2.3.5. Integrated Brier Score

Integrated Brier score (IBS) [18] is based on the squared deviation of the predicted survival function from the true one. The true survival function equals 1 before the event time and 0 after. Denote the number of observations as N and the probability of censoring as the Kaplan–Meier model, $G(t) = P(C > t)$, with the censoring event. To assess the quality of the forecast at the time t, the Brier score is:

$$BS(t) = \frac{1}{N} \sum_i \begin{cases} \frac{(0 - S(t|X_i))^2}{G(T_i)}, & \text{if } T_i \leq t, \delta_i = 1, \\ \frac{(1 - S(t|X_i))^2}{G(t)}, & \text{if } T_i > t, \\ 0, & \text{if } T_i = t, \delta_i = 0, \end{cases} \tag{4}$$

where $S(t \mid X_i)$ is the predicted survival function at time, t for observation X_i with the true time T_i. Then, for a fixed moment, t, and an observation, X_i, if the event occurs before time t, we expect a low survival probability (close to 0). Otherwise, if the event occurs after the moment t, we expect a high survival probability (close to 1).

To score censored data, $BS(t)$ (4) uses the δ_i indicator and the probability of censoring. Squared deviations are weighted on the inverse probability: $\frac{1}{G(T_i)}$ if the event occurs before time t, and $\frac{1}{G(t)}$ if the event occurs after time t. Censored observations before time t are not taken into account. Integrated Brier score aggregates, $BS(t)$, over time:

$$IBS = \frac{1}{t_{max}} \int_0^{t_{max}} BS(t)dt. \tag{5}$$

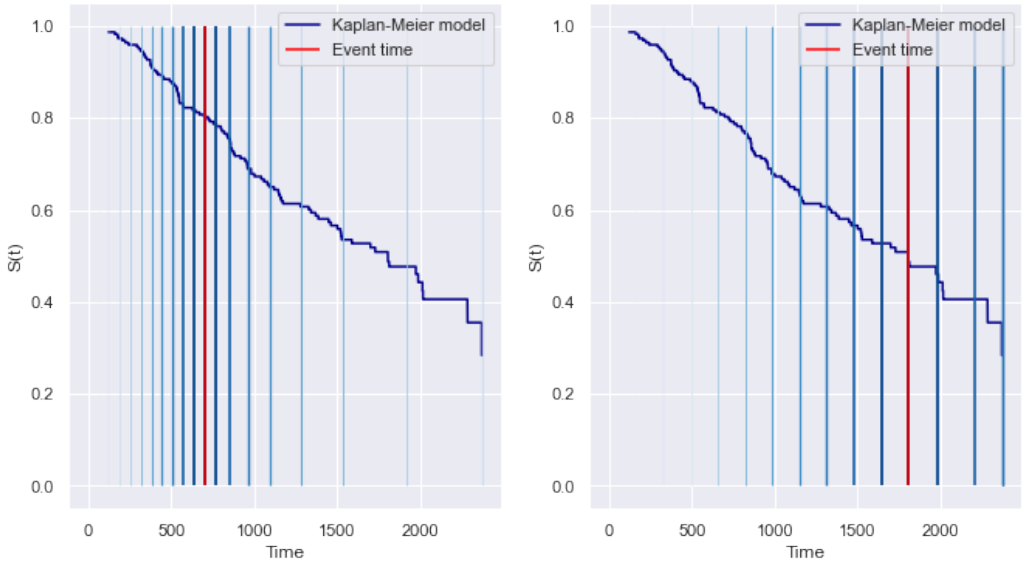
2.3.6. AUPRC

The Survival -AUPRC (henceforth, $AUPRC$) metric [19] measures the concentration of the distribution mass around the true time of the event. The idea is similar to the metric "area under Precision-Recall curve", but $AUPRC$ compares distributions for one observation.

For a terminal event with true time T_i and features X_i, the metric is an average difference between early and late values of the survival function $\hat{S}(t)$ at different intervals (each $\varphi \in [0, 1]$ determines the interval $[T_i \cdot \varphi, T_i/\varphi]$):

$$AUPRC_{\delta_i=1}(\hat{S}, T_i, X_i) = \int_0^1 \left[\hat{S}(T_i \cdot \varphi \mid X_i) - \hat{S}(T_i/\varphi \mid X_i)\right] d\varphi.$$

Figure 1 shows an example of calculating $AUPRC$ for two terminal events of the $GBSG$ dataset with the true times $T_i = 698$ (left figure) and $T_i = 1807$ (right figure). The vertical red line highlights the true time moment, and blue vertical lines correspond to the φ level and reflect the compared values $\hat{S}(T_i \cdot \varphi)$ and $\hat{S}(T_i/\varphi)$.



(**a**) (**b**)

Figure 1. Example of the $AUPRC$ (area under Precision-Recall curve) metric for two terminal observations of the $GBSG$ dataset with the event times $T_i = 698$ (**a**) and $T_i = 1807$ (**b**). The vertical red line highlights the true time moment, and the blue vertical lines correspond to the φ level. $AUPRC$ provides a symmetrical contribution of early and late intervals and covers the entire timeline. (**a**) True time is 698 ($AUPRC$ = 0.38). (**b**) True time is 1807 ($AUPRC$ = 0.42).

For censored observation with time T_i and features X_i, $AUPRC$ proposes zero survival probability after the event:

$$AUPRC_{\delta_i=0}(\hat{S}, T_i, X_i) = \int_0^1 \hat{S}(T_i \cdot \varphi \mid X_i) d\varphi.$$

The best value of the metric is 1 when the survival function is a step-function that equals 1 before the event and 0 after. The smallest value is 0 if the survival function is a constant (in terminal events for any constant function and in censored observations only for constant 0). However, paper [19] presents only partial metric values and does not describe a method to aggregate the $AUPRC$ for multiple observations. In this paper, we propose to aggregate values by mean:

$$AUPRC = \frac{1}{N} \sum_i AUPRC_{\delta_i}(\hat{S}(t), T_i, X_i). \tag{6}$$

2.3.7. Motivation for Choosing a Metric

In this section, we summarize the properties of all considered metrics. Firstly, each metric estimates one of the predicted values: the expected event time, T (CI), the survival function, $S(t)$ (KL, IBS, and $AUPRC$), and the hazard function, $h(t)$ ($IAUC$ and LL). For a comprehensive assessment of the quality of forecasting, it is necessary to choose a metric for each of the variables.

Based on the comparison approach, metrics can be classified into ranking (CI and $IAUC$ compare relative values) and regression (LL, KL, IBS, and $AUPRC$ compare absolute values) types. Based on the covering approach, metrics can be divided into point and integral types. Point metrics (CI and LL) use a single value to evaluate the quality. Integral metrics ($IAUC$, KL, IBS, and $AUPRC$) compare the predicted survival and hazard function with the target function. Thus, integral metrics allow us to score the quality of functions over time.

Finally, all metrics except KL have the ability to differ in the processing of censored observations from terminal events. In particular, the uncertain behavior after censoring time leads to the false comparison of forecasting.

Thus, we recommend several metrics to assess the following quantities:

1. CI metric for estimating the time of the event T. For censored observations, CI considers only pairs that the second event occurred before the moment of censoring.
2. $IAUC$ metric for estimating the hazard function, $h(t)$. Unlike the LL metric, $IAUC$ evaluates the overall survival function.
3. IBS and $AUPRC$ metrics for estimating the survival function, $S(t)$. Unlike KL, these metrics take into account censored observations. In addition, KL is limited by the nonparametric survival function, $S(t)$, which does not use the feature space of observations.

2.4. Machine Learning Models

Instead of classic parametric methods, tree-based approaches do not use strict theoretical assumptions (proportionality of hazards and definite distribution) and split the feature space into regions with a similar target variable.

Models applied in various tasks: classification, regression, and outlier detection. For each task, there is a criterion for splitting to calculate the proximity of the partitions by the values of the source features. In survival analysis, the most popular criteria are a statistical log-rank criterion (Section 2.4.1) and its modifications. Section 2.4.2 describes a method for constructing a binary survival tree.

To improve the quality of decision trees, there are ensemble algorithms. Decision trees are well suited as basic models, as they exactly describe the training sample. The random survival forest method (Section 2.4.3) is an ensemble of independent survival trees with

log-rank criteria. The survival *Bagging* (Section 2.4.4) method uses more confident trees with modified log-rank and works with missing and categorical data.

The gradient boosting method [20] is an ensemble of models where each next algorithm fixes the errors of the previous model. Gradient Boosting uses only one target variable, but survival analysis has two variables with different types: the time of the event and the event indicator. Section 2.4.5 describes the Gradient Boosting method in terms of survival analysis and its component-wise modification (Section 2.4.6).

2.4.1. Log-Rank Criterion

To measure the differences between two survival functions, the most widespread criterion is log-rank [3]. A higher value of log-rank statistics determines a greater difference between survival functions. The null hypothesis of the criterion assumes that survival functions from two samples are equal. However, papers [1,3] suggest poor sensitivity of log-rank to the real data with early events. The log-rank criterion does not assume a relationship between the censoring indicator and the forecast, and the significance of events is the same at the early and late stages of the study.

There are many ways to increase the sensitivity of the criterion. For example, several weighted criteria [1,3,4] have a high significance of the contribution of early events. Criteria define the weights of log-rank statistics by the following schemes:

1. Wilcoxon weights are the number of remaining observations at the time.
2. Peto-peto weights are the value of the parent survival function at the time.
3. Tarone-ware weights are the square root of the number of observations at the time. Tarone-Ware is the "golden mean" among weighted criteria [1].

Fleming–Harrington [7] is a flexible criterion sensitive to a certain type of events. The criterion is based on the family of statistics $\{G^{\rho,\gamma} \mid \rho \geq 0, \gamma \geq 0\}$ with weights $\hat{S}(\tau)^{\rho} \cdot (1 - \hat{S}(\tau))^{\gamma}$ for each time τ. In particular, the criterion $G^{0,0}$ equals the log-rank (sensitivity to proportional risks), $G^{1,0}$ equals the peto-peto (sensitivity to early events), and $G^{0,1}$ is sensitive to late events [5]. According to article [7], weighted log-rank statistics are distorted in the case of strong censorship, and the reliability of weighted statistics seriously decreases when censorship increases. In addition, local sensitivity enhancement does not apply to all data, and it needs to use different directions for assessing the proximity of survival functions.

A comprehensive study [21] noticed that MaxCombo [5] is a universal criterion for assessing the proximity of survival functions (it shows the best results for 18 datasets). The MaxCombo criterion combines several weighted log-rank criteria to provide sensitivity for early and late events simultaneously. In particular, MaxCombo is defined as the maximum of the criteria $G^{1,0}$ and $G^{0,1}$.

2.4.2. Survival Tree

A tree-like survival algorithm (ST) [22] recursively divides the sample into groups with different survival functions. The tree starts from the root node with full data. Using a log-rank criterion, the root node is divided into two child nodes, which are also divided. The process is repeated recursively for each subsequent node.

In the case of a binary tree, the splitting approach considers all possible intermediate values for each feature from the space X. For each intermediate point, the statistic value is calculated by target features T and δ from two branches of the partition. The best partition has the maximum value of statistics among all possible pairs of partitioning. For an observation with a feature vector x, the forecast of the survival function is Kaplan–Meier estimation for the associated leaf (end node) by the x feature.

The advantage of the method is a strong interpretation. Each leaf has a set of rules passing from the root to the leaf. Thus, if the depth of the tree is not too great, the expert can analyze the set of rules for consistency and correctness.

Nevertheless, the method has significant drawbacks. Firstly, a classical survival tree works only on filled data. Secondly, without any restrictions on the number of observations

in a node, the survival tree has an addiction to overfitting. Finally, a tree needs a sufficient amount of data to reach a high quality. In the case of limited data, the decision tree model is often used as a basic "weak" model in ensembles.

2.4.3. Random Survival Forest

Random Forest is the most widely used method of ensembling. For survival analysis, Random Survival Forest (RSF) [23] is an ensemble of independent survival trees [22], aggregating their forecasts. The construction algorithm is:

1. Build N bootstrap samples (with replacement) from the source sample. Each bootstrap subsample excludes approximately 37% of the data, which is called out-of-bag (OOB);
2. Build a survival tree for each bootstrap sample [22]. Finding the best partition uses only P features at each node. The best partition maximizes the difference between child nodes;
3. Each survival tree is built until bootstrap sampling is exhausted. In other words, there are no restrictions on the depth and number of observations for trees.

The constructed ensemble evaluates the error of the model with OOB_i (with $i = 1 \dots N$). For each observation with the feature vector x, the forecast is the average forecast over trees so that $x \in OOB_i$. Similarly, the forecast of the survival function is the average value for all tree forecasts in the ensemble for all time points. Averaging forecasts of trees allows us to improve quality and avoid overfitting.

2.4.4. Survival Bagging

In previous work [24], we developed a new model of survival tree that uses weighted log-rank criteria and has a high sensitivity to data characteristics. In particular, the model supports the following criteria: Peto-peto, Wilcoxon, and Tarone-ware. To predict the survival function, the model uses the following nonparametric estimators for each leaf: the Kaplan–Meier model (KM from the Section 2.2.1), expanded Kaplan–Meier model (KM10 from the Section 2.2.1).

The model handles categorical features using the weight of evidence (WOE) mapping method. To handle missing values, a set of observations with a missing feature are placed in turn in each branch of the partition. The final branch has the greatest statistical value. To reduce the computational complexity for continuous features, we use quantiles as splitting points (the number of quantiles is a hyperparameter).

In addition, we developed a Bagging model (*Bagging*) as an ensemble of proposed survival trees. To select the optimal size of the ensemble, for each iteration we evaluated the OOB error with the following quality metrics (loss functions): CI, LL, and IBS. Similar to RSF, the forecast of the survival function is the average forecast of the survival trees relative to each point in time.

Previously, we have not considered universal criteria and criteria with increased sensitivity to late events. In addition, it is necessary to study the impact of data characteristics on loss functions. Excessive sensitivity of metrics to early or late events influences the optimal size of the ensemble. In addition , it is necessary to investigate the stability of metrics to class imbalance.

2.4.5. Gradient Boosting Survival Analysis

The Gradient Boosting model [20] is based on iterative learning of each new decision tree on the errors of the previous one. The purpose of the algorithm is to minimize the loss function by the iterative counting of a gradient of prediction error. The forecast of gradient boosting is a weighted sum of all tree forecasts in the ensemble. The main advantages of the model are simplicity, versatility, flexibility to modifications, and high generalizing ability. Although the approach provides high-quality forecasting, the interpretability of the model is poor.

There are many difficulties of gradient boosting in survival analysis. In particular, Gradient Boosting uses only one target variable, but survival analysis has two variables with different types (the time of the event and the censoring indicator).

However, paper [25] proposes the Gradient Boosting Survival Analysis (GBSA) model with an expanded scope of applicability of classical Gradient Boosting to survival analysis tasks. GBSA predicts the probability of the event $g(X) : \mathbb{R}^{N_f} \to [0,1]$ (where N_f is the number of features) as the linear combination, $X^T\beta$, from CoxPH. An ensemble of regression decision trees is used to describe $g(X)$. As a loss function, GBSA uses a modification of the likelihood function, taking into account the assumption of proportional hazards:

$$loss = -\sum_i \delta_i \cdot \left(g(X_i) - \log\left(\sum_{t_j \geq t_i} e^{g(X_j)} \right) \right).$$

Based on the forecast of $g(X)$, the forecasts of functions $S(t \mid X), h(t \mid X)$ are:

$$h(t \mid X) = h_0(t) \cdot e^g(X),$$
$$S(t \mid X) = S_0(t)^{\exp(g(X))},$$

where $h_0(t)$ and $S_0(t)$ are the cumulative hazard and survival function from the CoxPH model.

Thus, GBSA reduces survival tasks to the regression problem of predicting the point value of $g(X)$. To predict the survival and hazard function, GBSA extends the point value using a nonparametric model and inherits the limitations of the CoxPH assumption.

2.4.6. Component-Wise Gradient Boosting

Article [26] provides an overview of the existing methods of Component-wise Gradient Boosting (CWGBSA). Instead of basic algorithms fitting on the entire feature space, X, component-wise boosting uses only one variable for fitting. This approach is also called likelihood-based boosting since the ensemble maximizes the overall probability at each iteration by choosing the underlying algorithm that leads to the greatest increase in probability.

In survival analysis, the goal of CWGBSA is to optimize the loss function with respect to linear coefficients, β, of the CoxPH model. The gradient of the loss functions is calculated by the weights, β. In this case, each loss function should be expressed in terms of the weights of the CoxPH model (which additionally requires a hazard proportionality condition).

The model response is $g(x, \beta) = x^T\beta$. Function $S(t \mid X), h(t \mid X)$ constructs according to formulas (1) and (2) of the CoxPH model. According to study [26], CWGBSA with $loss = BrierScore$ outperforms alternative approaches, where the base learner is the least squares method. Thus, the survival and hazard forecasts of CWGBSA are based on a nonparametric model and a point estimation of the probability of an event. To obtain point estimation, an ensemble of linear models iteratively finds optimal coefficients, β, of the CoxPH model. In that case, the model inherits the limitations of the CoxPH assumption.

2.5. Summary

Based on the review of the metrics of the survival analysis, we emphasize the following conclusions. To assess the predicted event time, we recommend using the concordance index, which equals the ratio of correctly ordered pairs of events in time. To assess the predicted survival and hazard functions, we consider point and integral metrics. Point metrics use a single value of the forecast and reduce the survival task to a classical regression problem, taking into account the censoring flag. Integral metrics compare the predicted and theoretical functions for all time points. In Section 2.3.7, we recommend using metrics $IAUC$, IBS, and $AUPRC$ as the most promising. In addition, we have not found studies on the sensitivity of metrics to data characteristics.

Based on the review of survival analysis models, we emphasize the following conclusions. The Kaplan–Meier method is applied to the continuous time problem and has

no restrictions on the time distribution. However, the KM method does not describe the relationship between the feature space and target variables. The classic Cox proportional hazards model has a strict assumption that leads to several disadvantages: independence of the significance of features in time, linearity of the model, inapplicability to categorical features, and missing values.

The survival tree model uses the log-rank criterion in order to find the best partition. However, the log-rank criterion has poor sensitivity to the characteristics of real data [1,3]. Ensembles of independent (Random Forest) and dependent (GBSA, CWGBSA) models are used to increase the accuracy of forecasting survival trees. The limitations of the Random Forest model follow from the base survival models. The gradient boosting model solves a classical regression problem by extending the point estimate using CoxPH. The most popular metrics of the boosting model are modification of likelihood with the assumption of proportional hazards (GBSA) and Brier score (CWGBSA).

3. Methodology

The formulated conclusions from Section 2 are the justification for the directions of further research. Figure 2 presents the scheme of the main steps of the sensitivity study. First, we consider the most popular open-source datasets and highlight the following characteristics of the data (Section 4): the concentration of events over time (early and late events) and class imbalance. The excessive sensitivity of models or metrics can negatively affect the quality of models and resistance to new data.

With high sensitivity to early events, the penalty for an error at early events exceeds the late event penalty. Thus, to optimize the loss function, it is profitable to understate the survival function to minimize the late error of an early event. In addition, an increased concentration of early events tends to decrease a survival function close to constant 0. With high sensitivity to late events, models tend to overstate the survival function. With a high concentration of late events, the survival function tends to be constant 1.

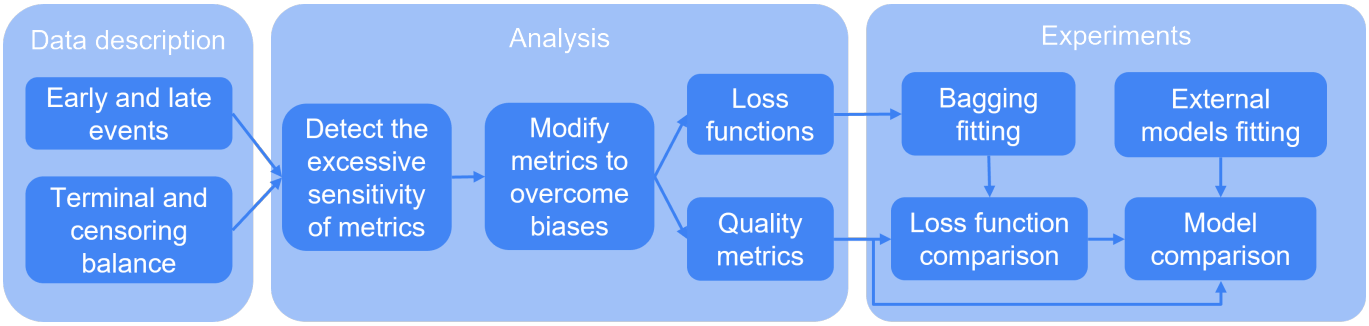


Figure 2. Scheme of the steps of the proposed algorithm. The first step is to highlight the characteristics of events for datasets. The second step is to investigate the sensitivity of metrics to these characteristics and check for equality between events. In the case of inequality, we modify the metrics. Metrics are compared with each other and divided into groups: quality metrics and loss functions. Finally, we evaluate the impact of modified metrics on the quality and compare models.

Later, we analyze existing metrics to detect the excessive sensitivity to described characteristics (Section 5). As we show below, the following cases distort the true quality of the forecast: the different significance of the contribution of events, increasing the contribution of late time, the influence of time scale, and a small contribution of terminal events in the case of an imbalance of censored observations. We present examples of the biases and propose modifications of metrics. To summarize the results of the analysis, we compare the metrics to each other based on their properties and resistance to biases. The most stable metrics are used to score the quality of the models, and others are used as the loss function.

At the experiments stage (Section 6), we use the previously proposed Survival *Bagging* (Section 2.4.4) and the following external models: Cox Proportional Hazard (Section 2.2.2),

Survival Tree (Section 2.4.2), Random Survival Forest (Section 2.4.3), Gradient Boosting Survival Analysis (Section 2.4.5), Component-wise Gradient Boosting (Section 2.4.6). Only the *Bagging* model is able to use the loss function in model fitting. To compare the influence of loss functions, *Bagging* considers them as hyperparameters. Based on experimental results, we provide a model comparison and a loss function comparison.

In addition, we visualize how modified metrics affect the choice of the best model. In the case of the leaf model hyperparameter, modified metrics allow for achieving better quality for the expanded Kaplan–Meier model (KM10).

4. Real Data Description

This section describes the most popular datasets in survival analysis. To provide a comprehensive analytical and experimental study, we consider eight datasets with different properties and characteristics. The SurvSet [27] library allows us to obtain files of the following datasets. In addition, we notice source links for each dataset.

The German Breast Cancer Study Group (*GBSG*) dataset, collected from 1984 to 1989, was presented in [28]. The dataset event is cancer relapse. The dataset contains 686 observations and 8 features according to anamnesis, tumor description, and treatment strategy. There are three categorical features: htreat, menostat, and tumgrad. The dataset does not contain missing values. During the study, 387 patients were censored.

The Cohort study on the breast cancer dataset from the Netherlands (*rott*2) was presented in [29]. The dataset event is the relapse of cancer. The dataset contains 2982 observations and 11 features according to anamnesis, tumor description, and treatment strategy. There are six categorical features: meno, tsize, grade, hormone, chemo, and recent. The dataset does not contain any missing values. During the study, 1710 patients were censored.

The Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (*support*2) [30] dataset describes incurable patients using life support devices. The dataset event is death. The dataset contains 9105 observations and 35 features according to anamnesis, the class of the patient's disease, the severity of physiological abnormalities, and concomitant diseases. There are 11 categorical features: sex, dzgroup, dzclass, num_co, race, diabetes, dementia, ca, dnr, sfdm2, income. In addition, 21 features contain missing values, where the maximum number of missing values is 5641 for the ADL feature. During the study, 2904 patients were censored.

The *WUHAN* dataset, collected from January 10 to 18 February 2020, was presented in [31]. The dataset event is the patient's discharge. The dataset contains 375 observations and 76 features according to anamnesis and the results of clinical studies during treatment. The feature space is formed from the minimum, maximum, and average indicators of the patient's clinical trials. All features contain missing values, where the maximum number of missing values is 173 for the indicators of antithrombin and fibrin breakdown products. During the study, 174 patients were censored.

The Primary Biliary Cirrhosis (*PBC*) dataset, collected from 1974 to 1984, was presented in [32]. The dataset event is death. The dataset contains 276 observations and 17 features according to anamnesis, cirrhosis status, treatment strategy, and clinical indicators. There are five categorical features: trt, sex, ascites, hepato, and spiders. In addition, 12 features of the dataset contain missing values (in particular, treatment strategies and clinical indicators), where the maximum number of missing values is 134 for the cholesterol index and 136 for the triglyceride index. During the study, 263 patients were censored.

The Second Manifestations of ARTerial Disease (*SMARTO*) [33] dataset is a sample from a study of patients hospitalized with clinically manifest atherosclerotic vascular disease or pronounced hazard factors for atherosclerosis. The dataset event is death. The dataset contains 3873 observations and 26 features of anamnesis, clinical indicators, and markers of atherosclerosis. There are nine categorical features: sex, diabetes, cerebral, aaa, periph, stenosis, albumin, smoking, and alcohol. In addition, 16 features contain missing values, where the maximum number of missing values is 1499 for diastolic by hand and 1498 for systolic by hand. During the study, 3413 patients were censored.

The AIDS Clinical Trials Group Study (*actg*) [34] dataset is a sample from a study comparing two sets of drugs in HIV-infected patients. The dataset event is the diagnosis of AIDS or death. The dataset contains 1151 observations and 11 features according to anamnesis, clinical indicators, and treatment strategy. There are seven categorical features: tx, txgrp, strat2, sex, raceth, ivdrug, hemophil. The dataset does not contain any missing values. During the study, 1055 patients were censored.

The Assay of Serum Free Light Chain (*flchain*) [35] dataset is a sample of Olmsted County residents from a study of the relationship between serum-free light chains (FLC) and mortality. The dataset event is the death of a patient. The dataset contains 7874 observations and 11 features according to anamnesis, clinical blood analysis, and the presence of monoclonal gammopathy. There are four categorical features: sex, chapter, sample_yr, and mgus. Only one feature (creatinine) contains 1350 missing values. During the study, 5705 patients were censored.

Figures 3 and 4 show the event time density for each of the considered datasets. Brief statistics for all datasets are shown in Table 1. The column "N" contains the number of observations, and the column "feat" contains the number of features. The percentage of terminal events is presented in column "Event". The number of features with possible missing values is presented in column "NaN (feat)".

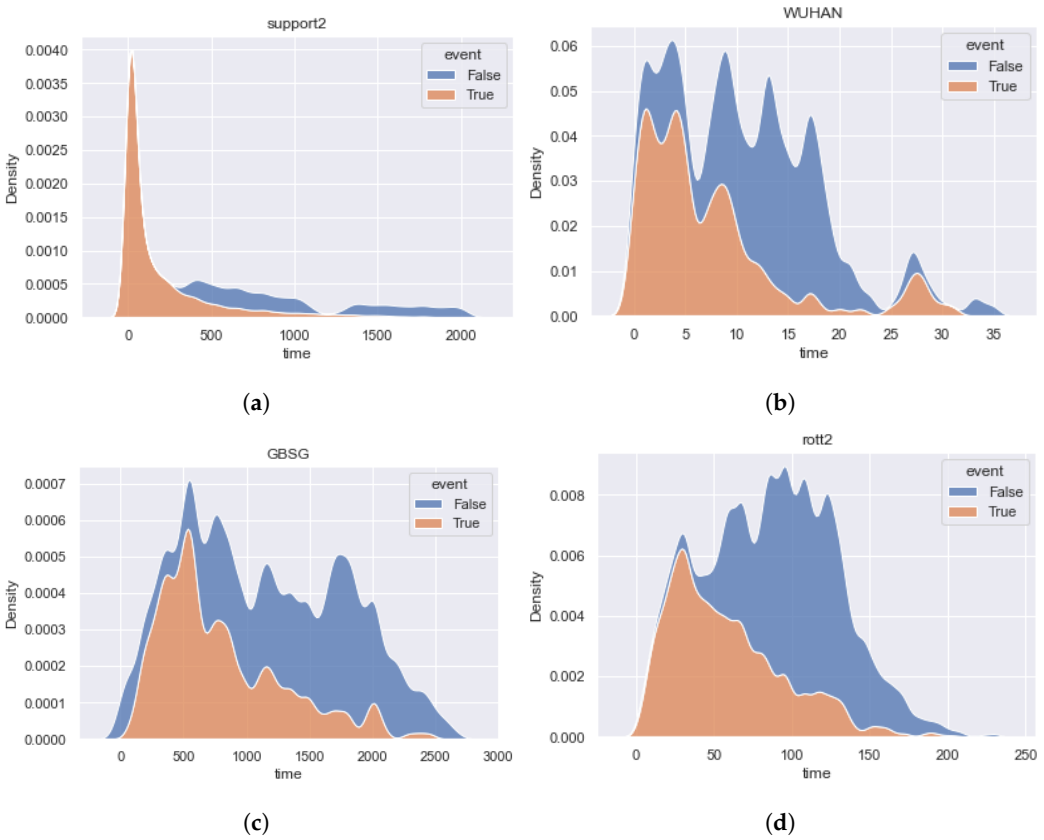


Figure 3. The time densities of terminal events and censored observations of the datasets *support*2, *WUHAN*, *GBSG*, and *rott*2. Datasets contain predominant early events. (**a**) *support*2 [30]. (**b**) *WUHAN* [31]. (**c**) *GBSG* [28]. (**d**) *rott*2 [29].

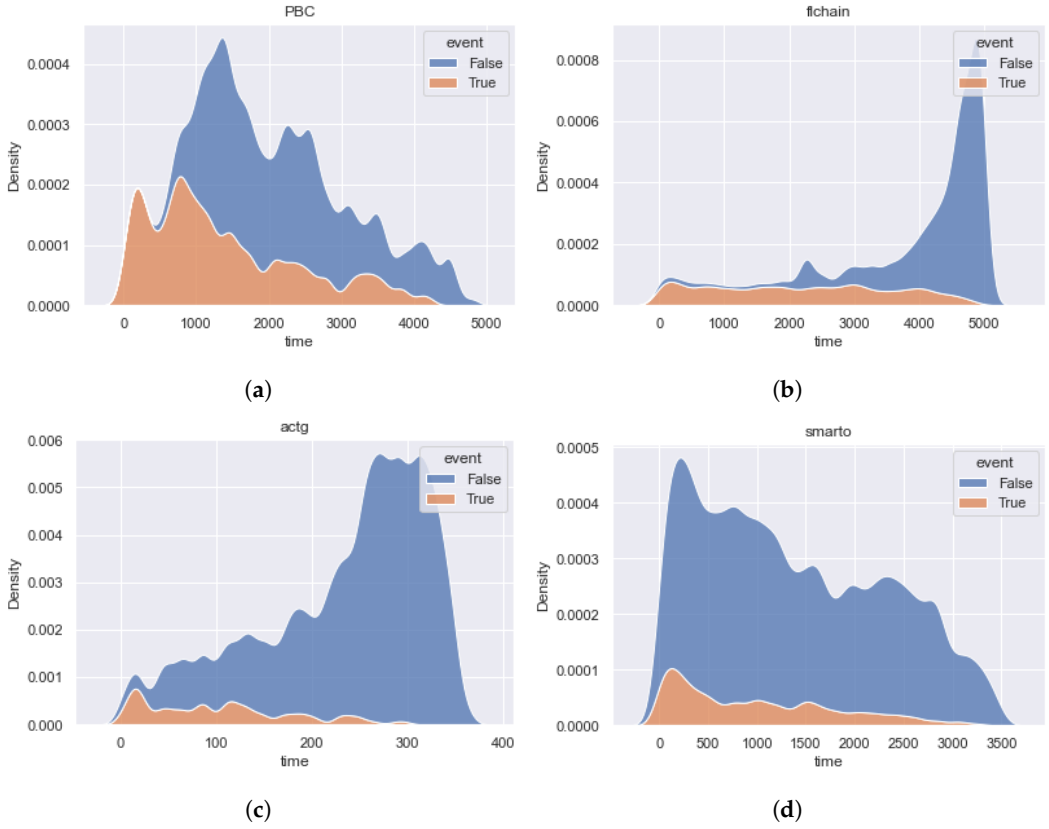


Figure 4. The time densities of terminal events and censored observations of the datasets *PBC*, *flchain*, *actg*, and *SMARTO*. The distributions of terminal and censored observations are very different for *flchain* and *actg* data. Datasets do not contain the predominant interval of events. (**a**) *PBC* [32]. (**b**) *flchain* [35]. (**c**) *actg* [34]. (**d**) *SMARTO* [33].

Table 1. Description of considered datasets. The datasets follow in descending order the percentage of terminal events.

Name	N	Feat	Cens, Event	Event (%)	NaN (Feat)
*support*2 [30]	9105	35	(2904, 6201)	0.681	21
WUHAN [31]	375	224	(201, 174)	0.464	222
GBSG [28]	686	8	(387, 299)	0.436	0
*rott*2 [29]	2982	11	(1710, 1272)	0.427	0
PBC [32]	418	17	(257, 161)	0.385	12
flchain [35]	7874	10	(5705, 2169)	0.275	1
SMARTO [33]	3873	26	(3413, 460)	0.119	16
actg [34]	1151	11	(1055, 96)	0.083	0

Based on the described open datasets and figures of the time density, we highlight the following characteristics of the datasets:

1. Early events have the highest significance for the *support*2 dataset because the data describes the incurable patients using life support devices. The imbalance of event classes is biased toward terminal events.
2. The *WUHAN*, *GBSG*, *rott*2, and *PBC* datasets are balanced relative to the event classes, and the early and middle events have the greatest importance.

3. The $flchain$, $SMARTO$, and $actg$ datasets have a high imbalance of censored events and are uniform for the importance of event time contribution. It is important to note that for $flchain$ and $actg$ datasets, the distribution of censoring time differs from the distribution of event time in the direction of increasing the importance of late events. The shapes of the $SMARTO$ dataset density functions are close.

5. Analysis of Biases in the Sensitivity of Metrics

In Section 2.3.7, we have described the primary motivation for choosing existing metrics to estimate survival forecasts. For a comprehensive study of the sensitivity of quality metrics, in this section, we consider the following characteristics and biases of chosen metrics:

1. The significance of the contribution of partial events (Section 5.1). The metric may have an implicit relationship between the contribution of events and its true time, which affects the reliability of the estimation. For example, the increased impact of late events is not suitable for data with dominant early events.
2. Dependence of integral metrics in time (Section 5.2). Integrated metric values may have a latent dependence on the timeline. In this case, the increased importance of a certain time period is not suitable for different data.
3. The influence of time variable in the integral metric (Section 5.3). The integrated variable directly affects aggregation over time and can lead to distortion of the significance of a certain period of time.
4. Resistance to the imbalance of censored observations (Section 5.4). The dominance of censoring can lead to false overstatement or understatement of the metric.

Later, we will reveal these biases for integrated metrics of survival analysis. To overcome the excessive sensitivity, we will adjust the weighting scheme of observation contributions. Finally, we will select the best metric for quality assessment and use other metrics as loss functions of the *Bagging* model. Comparing loss functions, we will assess the impact of biased and unbiased metrics on model quality.

5.1. The Significance of the Contribution of Partial Events

In this section, we investigate the metrics IBS, $IAUC$, and $AUPRC$ for the presence of dependence on the contribution of partial events. To check the dependence, we visualize the metric values for each observation relative to the event time. To ensure equal conditions for early and late events, we consider the constant forecast of the survival function as $S(t) = 0.5$. Using constant forecasts as one or zero, the quality of the events would be different. The main requirement for verification is the ability to represent the metric as an aggregation of values for each observation.

5.1.1. IBS

We present an alternative form of IBS (5). In particular, we transfer the integration operation for each iteration of summation. Since the sum and the constant N do not depend on the observation time, the formula has the following form:

$$IBS = \frac{1}{N}\sum_i \frac{1}{t_{max}} \int_0^{t_{max}} \left(\begin{cases} \frac{(0-S(t,x_i))^2}{G(T_i)}, & \text{if } T_i \le t, \delta_i = 1, \\ \frac{(1-S(t,x_i))^2}{G(t)}, & \text{if } T_i > t, \\ 0, & \text{if } T_i = t, \delta_i = 0, \end{cases} \right) dt.$$

Reveal the value of the integral for each of the conditions:

$$IBS = \frac{1}{N}\sum_i \frac{1}{t_{max}} \left(\int_0^{T_i} \frac{(1-S(t,x_i))^2}{G(t)} dt + \int_{T_i}^{t_{max}} \delta_i \cdot \frac{(0-S(t,x_i))^2}{G(T_i)} dt \right).$$

Then, IBS can be represented as the sum of partial IBS^i for each observation, i, with the corresponding event time, T_i, features, X_i, and the censoring indicator, δ_i:

$$IBS^i = \frac{1}{t_{max}} \left(\int_0^{T_i} \frac{(1 - S(t, x_i))^2}{G(t)} dt + \int_{T_i}^{t_{max}} \delta_i \cdot \frac{(0 - S(t, x_i))^2}{G(T_i)} dt \right), i = 1, 2, ..., N.$$

An alternative form of the IBS metric is the average of the partial IBS^i for each of the sample observations. The final form of IBS allows us to calculate the metric value for each observation:

$$IBS = \frac{1}{N} \sum_i IBS^i. \tag{7}$$

The first biased sensitivity of IBS is a growth of the values relative to the event time. For example, we consider $S(t) = 0.5$ as a constant survival function for each observation. The left side of Figure 5 shows the values of IBS for each observation of $GBSG$ dataset. The x-axis corresponds to the event time and the y-axis corresponds to the value of the partial IBS^i. The color of the dots determines the type of event: blue dots define censored events and orange dots define terminal events. Based on the left-hand figure, we notice that IBS increases during the growth of the event time. Consequently, late observations make a greater contribution in the metric (7).

Bias appears due to the monotonically increasing weight scheme, $1/G(t)$. The weight of the contribution depends on the time before the event occurrence and after the event is equal to the constant $1/G(T_i)$. Consequently, increasing the event time, the weights of deviations and their aggregation also increase. The logic is contrary to other studies, according to which weighted log-rank criteria are highly sensitive to early events. In addition, terminal and censored observations determine two curves, so the censored curve is located below. This effect is due to the censored IBS containing only up deviations to the moment of censorship.

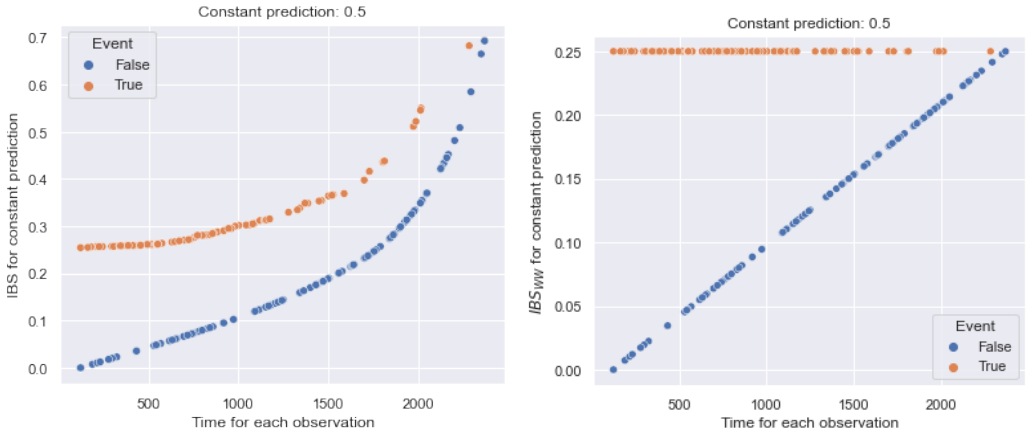
To overcome this disadvantage of the original IBS, we propose the IBS_{WW} metric without a weighting scheme (assuming $G(t) = 1$). Thus, the value of the $BS_{WW}(t)$ metric has the following form:

$$BS_{WW}(t) = \frac{1}{N} \sum_i \begin{cases} (0 - S(t, x_i))^2, & \text{if } T_i \leq t, \delta_i = 1, \\ (1 - S(t, x_i))^2, & \text{if } T_i > t, \\ 0, & \text{if } T_i = t, \delta_i = 0, \end{cases} \tag{8}$$

$$IBS_{WW} = \frac{1}{t_{max}} \int_0^{t_{max}} BS_{WW}(t) dt. \tag{9}$$

The right side of Figure 5 shows the values of IBS_{WW} for each observation of $GBSG$. The figure notations are equivalent to the left side of Figure 5. According to the figure, the metric value for terminal events equals the constant 0.25 and eliminates an increasing dependence on the true time. We see persistent linear dependence for censored observations because of the sum of early deviations, where the highest value is 0.25 (in the case of the latest censoring observation).

Thus, the inverse $G(t)$ weighting scheme gives excessive sensitivity to late observations, even in the case of a constant forecast. With the transition from the weighting scheme to the constant contribution of deviations, the increasing dependence disappears. An alternative disadvantage of the $1/G(t)$ weights is the dependence of the metric on the estimation of the general survival function for censored observations, $G(t)$. Consequently, the estimation of the global survival function changes with the growth of data, leading to a bias of the previous values of the metric.



(**a**) (**b**)

Figure 5. Example of the relationship between partial IBS^i (Integrated Brier Score) and IBS^i_{WW} values for each observation of the *GBSG* dataset relative to the event time. IBS increases depending on the event time and has a greater contribution to late observations. IBS_{WW} is constant over time (for terminal events) and determines equal contributions. (**a**) IBS. (**b**) IBS_{WW}.

5.1.2. IAUC

The $IAUC$ metric (3) does not have a representation in terms of partial $IAUC^i$, which depends on the forecast for only one object, i. The reason for this is the ranking type of metric. In particular, the construction of a partial score uses only one pair of the true and predicted hazard functions, but $IAUC$ is based on a comparison of different hazard functions for pairs of observations.

5.1.3. AUPRC

The $AUPRC$ metric (6) already has a partial form. Figure 6 shows an example of the dependence of $AUPRC$ on the true observation time for the constant forecast, $S(t) = 0.5$, of *GBSG*. All notations are the same as those in Figure 5. Note that the minimum value for censored observations is 0.5.

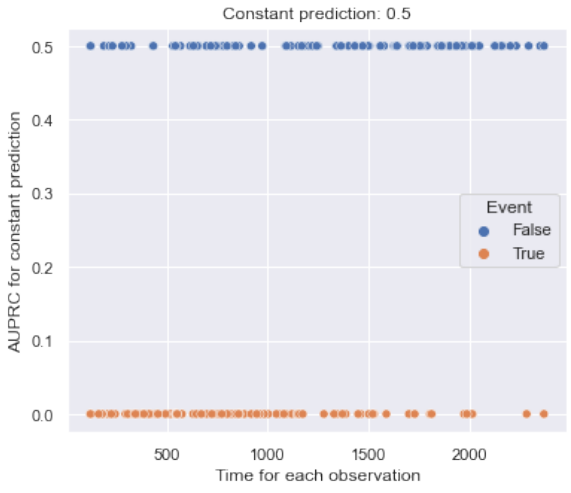


Figure 6. Example of the relationship between $AUPRC$ values and the event time for each observation of the *GBSG* dataset. $AUPRC$ is constant over time and determines equal contributions.

According to the figure, the metric value for terminal events equals the constant 0.0 and does not determine an increasing dependence on the true time. For censored observations, the metric equals the constant 0.5 and also does not assume increasing dependence. Thus, the $AUPRC$ metric does not have false sensitivity to partial observations.

5.2. Dependence of Integral Metrics in Time

In this section, we check the dependence of the time-based components $AUC(t)$, $BS(t)$, and $AUPRC(\varphi)$ over time. Later, integrated metrics will use the components to aggregate the scores for each moment of the timeline.

5.2.1. IBS

Calculating time-dependent metrics, $BS(t)$ (4) and $BS_{WW}(t)$ (8), averages the deviations of all observations for each time, t (the total number of observations is N).

As noted earlier, we consider the constant survival function, $S(t) = 0.5$, in time. Figure 7 shows the trend of quadratic deviations over time. The x-axis corresponds to the bins of the timeline, and the y-axis corresponds to the time-based values. The blue line refers to the $BS(t)$ values. The highest values of the metric are reached in the time interval from 1500 to 2000. Consequently, early observations (occurring before the 1000th moment) have a smaller contribution to the integral value of IBS. The reason for excessive sensitivity is the weight, $1/G(t)$, which increases the contribution of deviations of late events.

The orange line refers to $BS_{WW}(t)$ and monotonically decreases in time. Therefore, early observations have a greater contribution to the integral value of IBS. According to the $BS_{WW}(t)$ (8), the contribution equals 0 after the censoring time. In general, the deviation after censoring is indefinite (due to the absence of the true time of the event). At the averaging stage, the zero deviation makes a false contribution, assuming the high quality of forecasting, $S(t)$.

To overcome the problems of the $BS(t)$ and $BS_{WW}(t)$ metrics, we propose a $BS_{RM}(t)$ metric with controlled averaging of observed events by time, t. In this case, the constant of the total number of events, N, is replaced by the variable $N(t) = N_{event} + N_{cens}(t) = N_{event} + \sum_{i:\delta_i=0} I(T_i > t)$. Therefore, the following modification does not take into account the contribution of observations after the moment of censoring:

$$BS_{RM}(t) = \frac{1}{N(t)} \sum_i \begin{cases} (0 - S(t, x_i))^2, & \text{if } T_i \leq t, \delta_i = 1, \\ (1 - S(t, x_i))^2, & \text{if } T_i > t, \\ 0, & \text{if } T_i = t, \delta_i = 0, \end{cases} \tag{10}$$

$$IBS_{RM} = \frac{1}{t_{max}} \int_0^{t_{max}} BS_{RM}(t) dt. \tag{11}$$

In Figure 7, the green line relates to the $BS_{RM}(t)$ (10). Thus, the metric does not have false sensitivity, such as the contribution of each time is 0.25. It is important to note that the study (Section 5.1) of partial IBS^i assumes $N = N(t) = 1$ (for one observation). Consequently, this characteristic of time sensitivity affects only a set of observations. In the case of partial observations $IBS^i_{RM} = IBS^i_{WW}$.

Thus, the weighting scheme, $1/G(t)$, and the averaging approach of all deviations lead to excessive sensitivity to late and early events, respectively. Using a modification of controlled averaging, false sensitivity disappears.

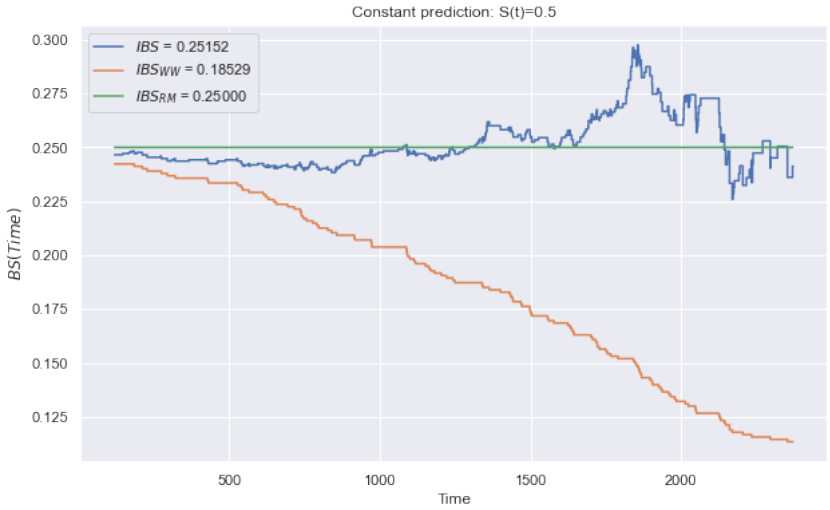


Figure 7. Example of $BS(t)$ trend over time for a constant forecast of $S(t) = 0.5$. The blue line relates to IBS, the orange line to IBS_{WW}, and the green line to IBS_{RM}. Metrics IBS and IBS_{WW} change over time and have false sensitivity. We propose IBS_{RM}, which determines equal contributions.

5.2.2. IAUC

Similarly, we consider the behavior of the $AUC(t)$ metric over time (Figure 8). Evaluating the cumulative hazard function, we use the following transformation: $H(t) = -log(S(t))$. For a constant function, $S(t) = 0.5$, we set $H(t) = -log(0.5) = log(2) = 0.693$. Based on the figure, we conclude that the metric does not have false sensitivity in time.

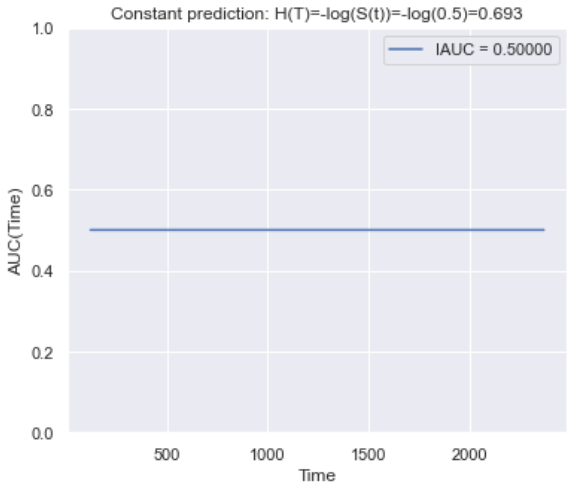


Figure 8. Example of the trend of $AUC(t)$ over time for a constant forecast of $H(t) = -log(S(t)) = -log(0.5)$. It is seen that $AUC(t)$ determines an equal contribution in time.

5.2.3. AUPRC

Figure 9 shows an example of changing the $AUPRC(\varphi)$ metric in time for a constant function, $S(t) = 0.5$, of $GBSG$. Based on the figure, the $AUPRC(\varphi)$ metric does not have false sensitivity in time.

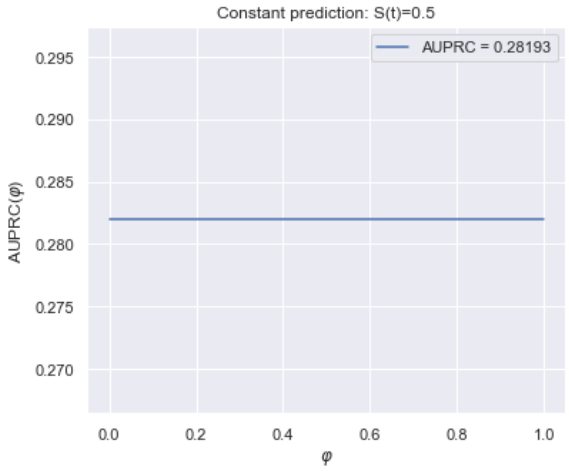


Figure 9. Example of $AUPRC(\varphi)$ trend over time for a constant forecast of $S(t) = 0.5$ on $GBSG$ dataset. It is seen that $AUPRC$ determines equal contributions.

5.3. The Influence of the Integration Variable

In Section 5.2, we considered the sensitivity of the time-based values relative to the time, t. Further, before obtaining the integral values, it is necessary to investigate the sensitivity of the variable of integration in time.

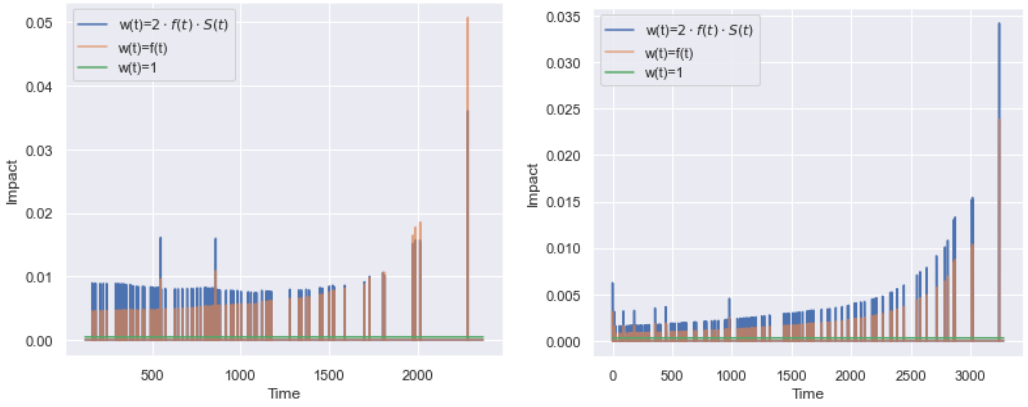
For integrated metrics (IBS, $IAUC$, $AUPRC$), we aggregate quality for all time points by calculating the integral on a timeline. In practice, a timeline can be set by the user to predict the function at a certain time. In this paper, the timeline is the set of times between the occurrence of the first and last event of the training sample (hereafter, we denote the set of bins as $\{t_i\} : t_{min} \leq T \leq t_{max}$).

Integrating with this set of bins leads to an equal contribution of $dt = 1$ in time. In terms of the IBS metrics (IBS, IBS_{WW}, IBS_{RM}), the integral is calculated directly from the time, t, and has an equal contribution each time. Similarly, integrating over the variable φ, $AUPRC$ determines the equal contribution.

For the metric $IAUC$ (Section 2.3.2), there are several defining sets of differentials to calculate the integral. In particular, papers [13,14,36,37] present weight schemes with the general formula $IAUC = \frac{1}{\int w(t)dt} \int AUC(t) \cdot w(t)dt$. At the same time, papers [14,36,37] assume weights as the density function, $w(t) = \hat{f}(t)$. Hence, $w(t)dt = \hat{f}(t)dt = -d\hat{S}(t)$, where $\hat{S}(t)$ is the KM estimation (Section 2.2.1). Paper [13] considers the weight scheme as $w(t) = 2 \cdot \hat{f}(t) \cdot \hat{S}(t)$. Hence, $w(t)dt = 2 \cdot \hat{f}(t) \cdot \hat{S}(t)dt = -2 \cdot \hat{S}(t)d\hat{S}(t) = -d\hat{S}^2(t)$, where $\hat{S}(t)$ is the KM estimation.

Figure 10 shows a behavior of different weights in time, $w(t)$, for the $GBSG$ and $SMARTO$ datasets. The green line relates to weighing with an equal contribution (used in IBS). The blue line relates to the weighted scheme $w(t) = 2 \cdot \hat{f}(t) \cdot \hat{S}(t)$, and the orange line to scheme $w(t) = \hat{f}(t)$ (used in $IAUC$).

Both $IAUC$ weight schemes have similar behavior and have two drawbacks. Firstly, some of the time points do not affect the value of the integral metric. The presence of a zero contribution is due to the equality of the survival functions for the start, t_1, and the end, t_2 (so that $t_2 \geq t_1$), points of the interval $[t_1, t_2]$ so $S(t_1) - S(t_2) = 0$. According to the KM estimation, it happens if there are no observed events in the interval $[t_1, t_2]$. In this case, censoring observations in the range $[t_1, t_2]$ changes the $AUC(t)$ value, which does not affect the integral metric, $IAUC$.



(**a**) (**b**)

Figure 10. Example of the dependence of the contribution of $w(t)$ on time for *GBSG* and *SMARTO* datasets. There are three weight schemes: $w(t) = 1$, $w(t) = \hat{f}(t)$, and $w(t) = 2 \cdot \hat{f}(t) \cdot \hat{S}(t)$. Only a constant weighting scheme provides equal contributions of bins. Other schemes increase over time and have false sensitivity to later bins. (**a**) *GBSG*. (**b**) *SMARTO*.

The second disadvantage is the increased significance of $AUC(t)$ for late time points. According to the dataset densities (Figures 3 and 4), the datasets *GBSG* and *SMARTO* contain predominant early events, and most of the events have happened by the time $t = 2000$. However, Figure 10 shows the higher importance of ranking quality for late time points. To overcome the disadvantages of $IAUC$, we recommend using the unit weight scheme $w(t) = 1$.

Thus, the metrics IBS and $AUPRC$ have equal contributions over time (without false sensitivity). At the same time, both weight schemes of $IAUC$ lead to the excessive contributions of late times and ignore $AUC(t)$ values at the moments of non-occurrence of terminal events (they do not take into account the fact of censoring).

5.4. The Impact of the Imbalance

In practice, real datasets have a different ratio of censored and terminal events. For a terminal event, the best survival function is a threshold function that equals 1 before the event and 0 after. For censored observation, the quality of the survival function is certain only before the moment of censoring T_i.

5.4.1. IBS

According to IBS (5), the best survival function for censored observation equals 1 until the moment of censoring and has an arbitrary value after. Therefore, for censored observations, the smallest value of $IBS = 0$ is reached by a constant forecast, $S(t) = 1$. The deviation before and after moment T_i are considered for terminal events. Therefore, for the same forecast, $S(t)$, it is true that $IBS_{\delta=1}(S(t)) > IBS_{\delta=0}(S(t))$.

To demonstrate the sensitivity of IBS to class imbalance, we present an alternative form of IBS (5). In particular, we represent the metric $BS(t)$ as the sum of deviations for each type of event, $BS(t) = BS_{\delta=1}(t) + BS_{\delta=0}(t)$, where $BS_{\delta=1}(t)$ is the proportion of deviations of terminal events and $BS_{\delta=0}(t)$ is the proportion of deviations of censored observations:

$$BS_{\delta=1}(t \mid N) = \frac{1}{N} \sum_{i:\delta_i=1} \begin{cases} \frac{(0-S(t,x_i))^2}{G(T_i)}, & \text{if } T_i \le t, \\ \frac{(1-S(t,x_i))^2}{G(t)}, & \text{if } T_i > t, \end{cases} \tag{12}$$

$$BS_{\delta=0}(t \mid N) = \frac{1}{N} \sum_{i:\delta_i=0} \begin{cases} \frac{(1-S(t,x_i))^2}{G(t)}, & \text{if } T_i > t, \\ 0, & \text{if } T_i \le t. \end{cases} \tag{13}$$

Therefore, IBS has the following form:

$$IBS = \frac{1}{t_{max}} \int_0^{t_{max}} BS_{\delta=1}(t \mid N) + BS_{\delta=0}(t \mid N)dt. \tag{14}$$

For data with the domination of censored observations, $N_{\delta=1} \ll N_{\delta=0}$, the optimal forecast, $S(t)$, is shifting to 1 and providing a smaller error for the dominant censored observations.

The left side of Figure 11 shows a change of $BS(t)$ for the $SMARTO$ dataset with a high imbalance of classes (12% of terminal events and 88% of censored observations). The dotted and dashed lines define $BS(t)$ values only for censored and terminal events, respectively. The solid line defines the total value of $BS(t)$ (sum of values). Based on the figure, there is an understating of the significance of the error of terminal events, and the curve, $BS(t)$, is close to $BS_{\delta=0}(t)$. Consequently, deviations of censored observations have a higher impact on the IBS metric. In this case, changing the deviations for terminal events does not lead to significant changes to the IBS metric.

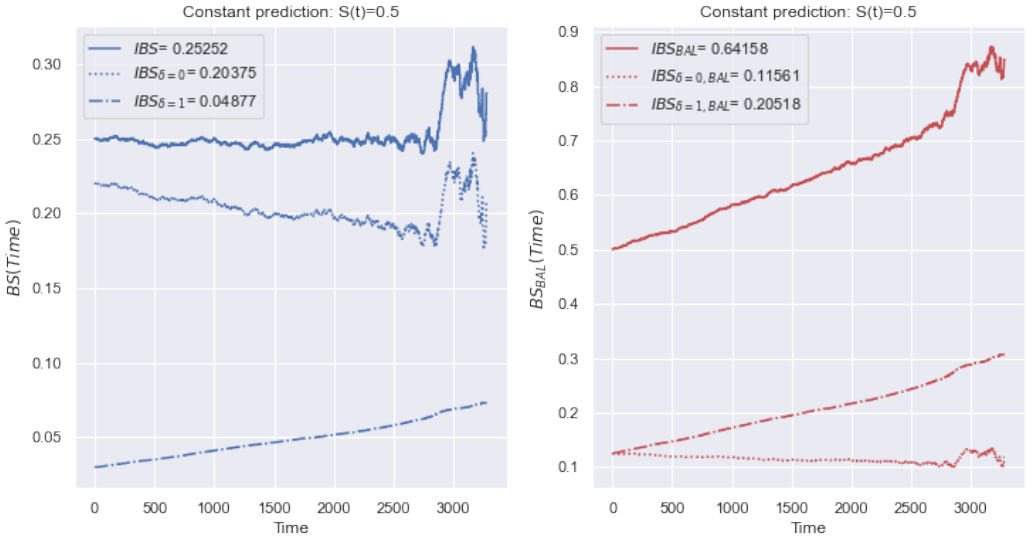


Figure 11. Example of the behavior of IBS and IBS_{BAL} metrics over time with a constant forecast of $S(t) = 0.5$ for the $SMARTO$ dataset. Due to the imbalance, there is an understating of the significance of the error of terminal events, and the curve, $BS(t)$, is close to $BS_{\delta=0}(t)$. Blue lines present IBS values (event, censoring, and total) with the source proportionality of classes. Red lines present IBS_{BAL} values (event, censoring, and total) with an equal ratio of classes.

In addition, we consider an example for other data. The left side of Figure 12 shows a change of $BS(t)$ for the $GBSG$ dataset, with a small class imbalance (44% of terminal events, 56% of censored observations). The notation is equivalent to Figure 11. There is no dominance of the contribution of a certain type of event. In this case, deviations for both types of event have a significant impact on the value of the IBS metric.

To overcome the disadvantage of the different contributions of event types, we propose establishing an equal contribution of the deviations of censored and terminal events to the value of $BS(t)$. In (12), we replace the total number of events, N, by the number of terminal events, $N_{\delta=1}$. Similarly, in (13), we replace N with the number of censored events, $N_{\delta=0}$. Thus, the value of the metric $BS_{BAL}(t)$ is defined as a balanced average relative to two types of event, and the metric IBS_{BAL} is defined similarly to (5):

$$BS_{BAL}(t) = \frac{1}{2}(BS_{\delta=1}(t \mid N_{\delta=1}) + BS_{\delta=0}(t \mid N_{\delta=0})),$$

$$IBS_{BAL} = \frac{1}{t_{max}} \int_0^{t_{max}} BS_{BAL}(t)dt. \quad (15)$$

Note that the metrics $BS_{\delta=1}(t \mid N_{\delta=1})$, $BS_{\delta=0}(t \mid N_{\delta=0})$, and $BS_{BAL}(t)$ can be defined for the previously described modifications, $BS_{WW}(t)$ and $BS_{RM}(t)$, similarly.

The right side of Figure 11 shows a change of the $BS_{BAL}(t)$ metric for the *SMARTO* dataset. Compared to the metric $BS(t)$ in the left-hand figure, the contribution of deviations of the terminal and censored events equally affects the values of $BS_{BAL}(t)$. The right side of Figure 12 shows a change of the $BS_{BAL}(t)$ metric for the *GBSG* dataset. The values of the metrics are close, and the contribution of deviations of terminal and censored events significantly affects the values of $BS(t)$ and $BS_{BAL}(t)$.

Thus, with an imbalance of a certain type of observation, the average deviation shifts towards the prevailing class. Using a balanced modification of *IBS*, we overcome excessive sensitivity to the prevailing class.

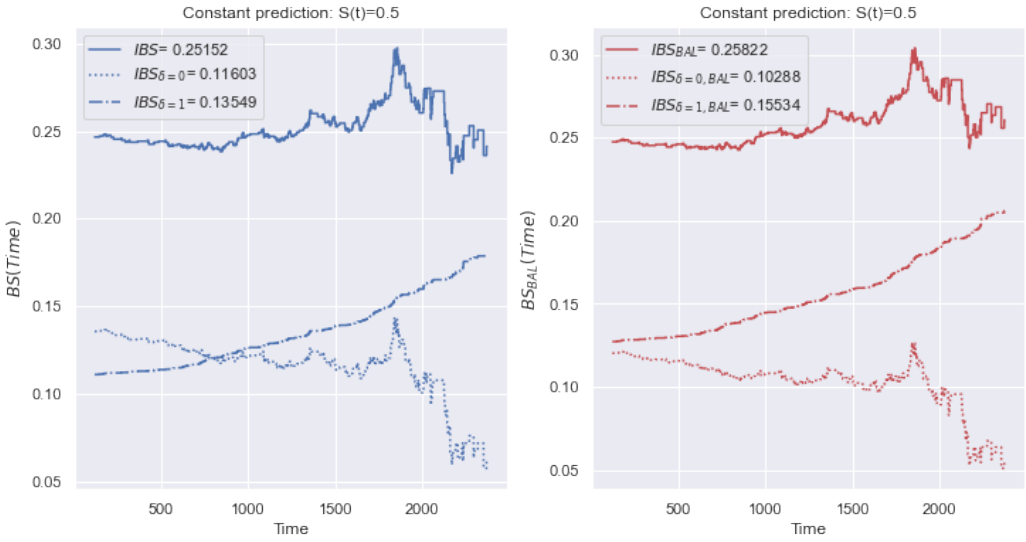


Figure 12. Example of the behavior of *IBS* and IBS_{BAL} metrics over time with a constant forecast of $S(t) = 0.5$ for the *GBSG* dataset. The right figure shows the proposed IBS_{BAL} metric, which is resistant to class imbalance. The metric values are close since the *GBSG* dataset is balanced relative to the event types.

5.4.2. AUPRC

Similarly, we conduct reasoning for the *AUPRC* metric. The left side of Figure 13 shows a change of $AUPRC(t)$ for the *SMARTO* dataset. Unlike the *IBS* metric, the final metric is the average of two values. The other notifications repeat Figure 11. There is an understating of the significance of the terminal events, and the curve $AUPRC(t)$ is close to $AUPRC_{\delta=0}(t)$. Consequently, deviations of censored observations contribute more to the metric *AUPRC*.

The left side of Figure 14 shows a change of $AUPRC(t)$ for the *GBSG* dataset. Unlike the *IBS* metric, the final metric is the average of two values. The other notifications repeat Figure 12. There is no dominance of a certain class of events.

Therefore, to increase the stability of *AUPRC*, we propose the following balanced modification (similar to the (15) metric):

$$AUPRC_{BAL}(\varphi) = \frac{1}{2}(AUPRC_{\delta=1}(\varphi \mid N_{\delta=1}) + AUPRC_{\delta=0}(\varphi \mid N_{\delta=0})),$$

$$AUPRC_{BAL} = \int_0^1 AUPRC_{BAL}(\varphi) d\varphi. \tag{16}$$

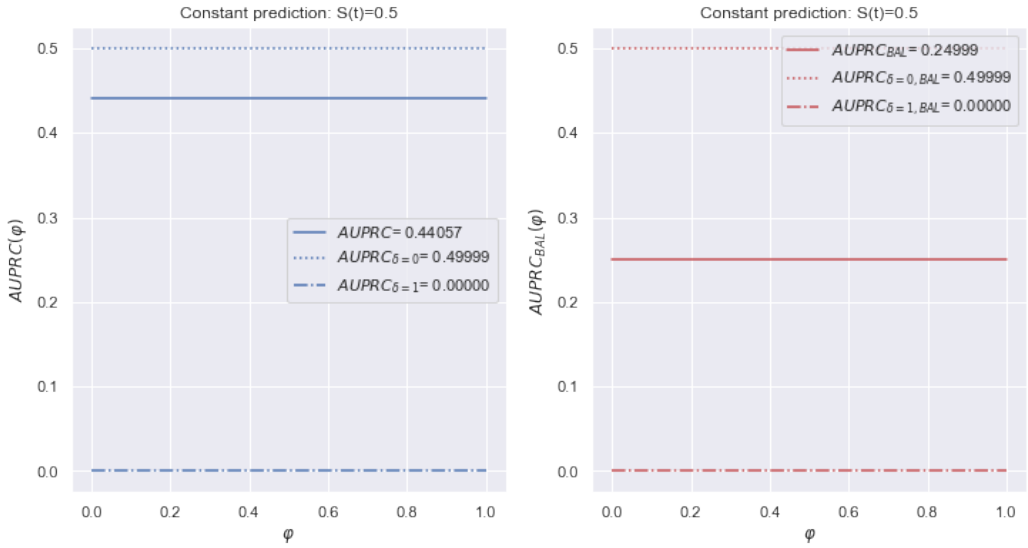


Figure 13. Example of a change of $AUPRC$ metric over time for $SMARTO$ dataset with a constant forecast of $S(t) = 0.5$. The left figure shows the total value of the $AUPRC$ metric is shifted toward the dominant class of censored events. The modification $AUPRC_{BAL}$ determines an equal ratio of the contributions.

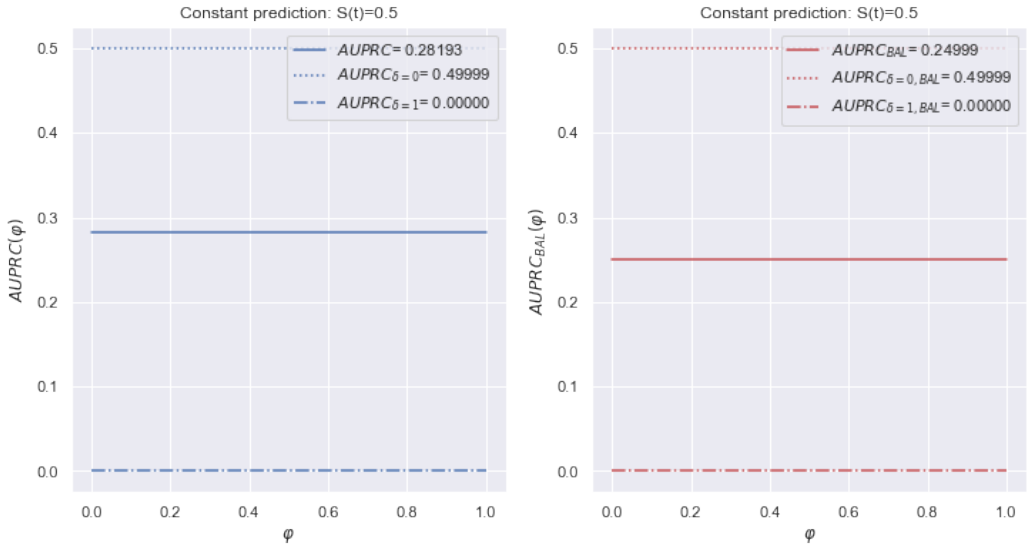


Figure 14. Example of a change of $AUPRC(t)$ over time for a $GBSG$ dataset with a constant forecast of $S(t) = 0.5$. $AUPRC_{\delta=0}$ and $AUPRC_{\delta=1}$ are close since $GBSG$ is balanced relative to the event types. In addition, $AUPRC_{BAL}$ determines an equal ratio of the contributions.

5.4.3. IAUC

Among the discussed metrics (Section 2.3.7), $IAUC$ is resistant to imbalance. In papers [38,39], the authors advise using AUC in the case of unbalanced data. In addition, papers [40,41] noticed that AUC is invariant for a priori probabilities of classes. In survival analysis, $AUC(t)$ and the integral value $IAUC$ have similar properties.

5.5. Summary

In this section, we studied the sensitivity of existing integral metrics of survival analysis to score the hazard and the survival function. In the course of the study, we

considered four different cases of metric excessive sensitivity involving false sensitivity to early events, late events, and class imbalance. To overcome the revealed biases, we have proposed several modifications to existing metrics.

Based on the conducted research (Sections 5.1–5.4), there is a summary, Table 2, that shows the properties of all integral metrics and their modifications. In addition, the summary table shows which metrics use the general survival function $G(t)$. In practice, the calculation and usage of $G(t)$ requires additional data and computing resources. Also, the estimation of the survival function changes with the growth of data, leading to a distortion of the previous values of the metric.

Table 2. The sensitivity properties of integral metrics. The column names are the numbers of the sections with studied excessive sensitivity: Section 5.1—the significance of the contribution of partial events, Section 5.2—the dependence of integral metrics in time, Section 5.3—the influence of the integration variable, Section 5.4—the impact of the imbalance. The column "$G(T)$" reflects the dependence of the metric on the theoretical general survival function. The cells indicate the presence of certain bias:"-"—stable, "+"—excessive, "?"—no information. The best metrics are grayed out.

Metric Name	Section 5.1	Section 5.2	Section 5.3	Section 5.4	G(T)
IBS	+	+	-	+	+
IBS_{WW}	-	+	-	+	-
IBS_{RM}	-	-	-	+	-
IBS_{BAL}	+	+	-	-	+
$IBS_{WW,BAL}$	-	+	-	-	-
$IBS_{RM,BAL}$	-	-	-	-	-
$IAUC$ $w(t) = 2 \cdot f(t) \cdot S(t)$	?	-	+	- ?	+
$IAUC$ $w(t) = f(t)$	?	-	+	- ?	+
$IAUC$ $w(t) = 1$	?	-	-	- ?	+
$AUPRC$	-	-	-	+	-
$AUPRC_{BAL}$	-	-	-	-	-

A better metric should have fewer biases. We highlight the stability of metrics to the considered bias as "-". Unknown verification or lack of information are highlighted as "?". For example, for the $IAUC$ family, we cannot check the contributions of partial observations because there is no representation of the metric in partial form. In addition, the stability of the $IAUC$ family to data imbalance is taken from papers [40,41], but this has not been tested in practice. The best metric from each family is grayed out.

Thus, the most stable metrics are $IBS_{RM,BAL}$, $IAUC(w(t) = 1)$, and $AUPRC_{BAL}$. The unknown behavior of $IAUC(w(t) = 1)$ to imbalance and partial events leads to less visibility and reliability compared with other families. Comparing the metrics $IBS_{RM,BAL}$ and $AUPRC_{BAL}$, we note an important property of the later. Estimating the probability of $P(T_i/\varphi > T > T_i \cdot \varphi)$, we evaluate the quality of the survival function before and after the event occurrence. Therefore, the early (before the event) and late (after the event) intervals have equal contributions to the $AUPRC$ metric. In the case of IBS, the early and late intervals have a different contribution proportional to the length of the time interval.

In addition, when calculating the integral, $AUPRC$ avoids additional false sensitivity. In particular, the distance between the bins does not affect the integral and avoids increasing the contribution of rare late events. In the case of IBS, the late outliers lead to additional late bins with an increased contribution of late events to the integral metric. Finally, paper [19] notes the stability of $AUPRC$ to calibration. In particular, saving the forecast form but changing the offset of predicted probabilities, IBS metrics give an unreliable assessment.

Thus, according to the results of the conducted research, the most stable and reliable metric for assessing the quality of the survival function is $AUPRC$.

6. Experiments

Based on the results of the conducted studies of the sensitivity of metrics (Section 5), we conclude that $AUPRC$ has the least false sensitivity. In addition, the proposed modifications of the IBS metrics allow us to achieve a comparable level of reliability.

Each noticed metric can be used as a loss function in machine learning models. The choice of the loss function for survival analysis models has a significant role. Firstly, during the model fitting, the loss function helps us to select the optimal size of the tree ensemble. Secondly, to deploy the model, it is necessary to determine a set of hyperparameters that achieve the best quality of the model. In the future, we plan to embed the loss function into boosting ensembles to minimize the error for each observation.

This section contains an experimental study of the impact of the loss function on the model quality. The $AUPRC$ metric and its modifications determine the model quality, but CI [42], LL [43,44], and IBS [24,26] and its modifications are used as the loss function. The study aims to select an optimal loss function (from the reviewed metrics in Section 5) and detect the impact of modifications of metrics to the model quality. The study of loss functions performs on the proposed survival bootstrap ensemble (Section 2.4.4).

The second goal of the experimental study is to evaluate and analyze the quality of existing and proposed models on the $AUPRC$ metric. At the time of this paper, we have not found any open studies about evaluating the models on $AUPRC$.

6.1. Experimental Setup

The experiment setup is divided into three stages (Figure 15). Initially, we process feature space and target variables of each dataset (time before the event, a censoring indicator). In the first stage, the source data is split into a training and a test sample (66% and 33%, respectively) with stratification by the censoring indicator.

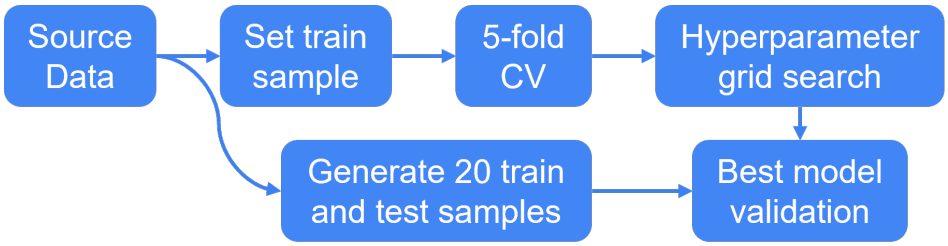


Figure 15. Scheme of experimental setup. There are three primary steps: train and test splitting, cross-validation (CV) grid search, and multi-sample validation. The pipeline is performed for each survival model (CoxPH, ST, RSF, GBSA, CWGBSA, *Bagging*), with a corresponding grid of hyperparameters.

In the second stage, using a training sample, we conduct a 5-fold cross-validation [45] according to a given grid of hyperparameters. During cross-validation, we divide the initial sample into five non-overlapping parts, four of which are used to train the model, and one part is used for model testing and for calculating metrics. Thus, there are five iterations of training/testing of the model, where each part becomes a test sample once. The resulting metric for cross-validation is the average value of the metric for all iterations. Using a predefined quality metric, we search the best hyperparameters for each model by grid.

In the third stage, we generate training and test data from the source 20 times (with 66% and 33% sizes, respectively). According to the best hyperparameters of each model (selected during cross-validation), we fit models on training data and apply them to test ones. The final quality of the model is the average quality for 20 test samples.

In the experimental study, we use the implementations of the CoxPH Survival Analysis (CoxPH), Survival Tree (ST), Random Survival Forest (RSF), Component-wise Gradient Boosting (CWGBSA), and Gradient Boosting Survival Analysis (GBSA) from the open

library *scikit-survival*. The scikit-survival [25] package is written in the Python programming language and allows one to build classical survival analysis models. According to the overview [25] of the completeness of existing libraries (scikit-survival, lifelines, statsmodels, pycox), scikit-survival contains the widest functionality. The implementation of *Bagging* was presented in Section 2.4.4 of this paper.

Table 3 contains the grid of hyperparameters for each model. The hyperparameters of the CoxPH model are the regularization parameter for ridge regression penalty (regularization penalty), and the method to handle tied event times (ties). The hyperparameters of the ST model control the tree growth by the depth of the tree (max depth), the number of splitting features (max features), the algorithm of best split choosing (split strategy), and the size of nodes (min sample leaf). The hyperparameters of the RSF model are the size of the ensemble (num estimators) and the tree growth control parameters. The hyperparameters of the GBSA model are the num estimators, the coefficient of the contribution of each tree (learning late), and the tree growth control parameters. The hyperparameters of the CWGBSA model are the num estimators, the learning rate, the fraction of samples to fit the individual base models (subsample), and the percentage of dropped base models during the fitting (dropout rate).

Table 3. The grid of hyperparameters of predictive models.

Model Name	Hyperparameter	Grid
CoxPH Survival Analysis	regularization penalty	0.1, 0.01, 0.001
	ties	breslow, efron
Survival Tree	split strategy	best, random
	max depth	from 10 to 30 step 5
	min sample leaf	from 1 to 20 step 1
	max features	sqrt, log2, none
Random Survival Forest	num estimators	from 10 to 100 step 10
	max depth	from 10 to 30 step 5
	min sample leaf	from 1 to 20 step 1
	max features	sqrt, log2, none
Component-wise Gradient Boosting	num estimators	from 10 to 100 step 10
	learning rate	from 0.01 to 0.5 step 0.01
	subsample	from 0.5 to 1.0 step 0.1
	dropout rate	from 0.0 to 0.5 step 0.1
Gradient Boosting SA	num estimators	from 10 to 100 step 10
	max depth	from 10 to 30 step 5
	min sample leaf	from 1 to 20 step 1
	max features	sqrt, log2, none
	learning rate	from 0.01 to 0.5 step 0.01
Bagging	bootstrap sample size	from 0.5 to 1.0 step 0.1
	num estimators	from 10 to 50 step 10
	max depth	from 10 to 30 step 5
	min sample leaf	0.05, 0.001
	max features	0.3, sqrt
	leaf model	KM, KM10 (Section 2.2.1)
	criterion	maxcombo, peto, tarone-ware, wilcoxon, logrank

The hyperparameters of *Bagging* are the num estimators, the fraction of samples to fit base trees (bootstrap sample size), the estimation of $S(t)$ and $h(t)$ for each leaf (leaf model), weighted statistic for split choosing (criterion), and the tree growth control parameters.

6.2. Results

In this section, we present detailed experimental results for 8 datasets: $support2$, $WUHAN$, $GBSG$, $rott2$, PBC, $flchain$, $SMARTO$, and $actg$. The purpose of the experiment is to evaluate the relationship between the loss function and the quality of model prediction. For each dataset, we evaluate the quality of existing and proposed methods using the metrics $AUPRC$, $AUPRC_{\delta=1}$ (metric is based on terminal events), and $AUPRC_{BAL}$ (metric is based on equal contributions of censured and terminal observations).

Figures 16–23 present the obtained results for each dataset. Each figure contains information about the quality of the existing models and *Bagging* for five loss functions: LL (likelihood), CI (concordance index), IBS (integrated Brier score), IBS_{WW} (9) (integrated Brier score without weighting), IBS_{RM} (11) (integrated Brier score with controlled averaging of remained events). The x-axis corresponds to method name and the y-axis corresponds to the quality of the method. The proposed models ($BSTR(IBS_{WW})$ and $BSTR(IBS_{RM})$) are marked in bold. For each method, we have created a boxplot with the distribution of metrics on 20 test samples. In addition, the medians of the values are marked by the gray line.

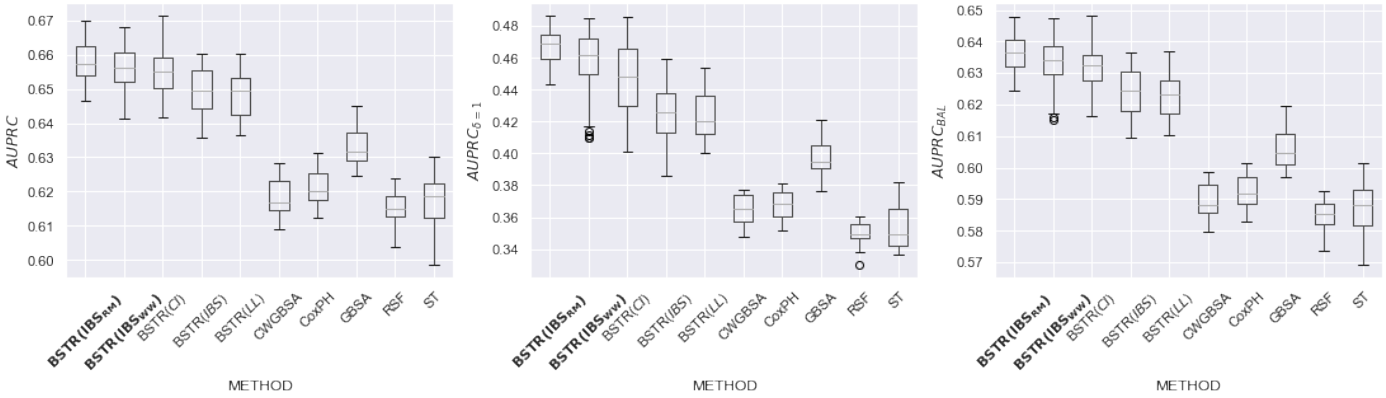


Figure 16. $AUPRC$ comparison for $GBSG$ dataset.

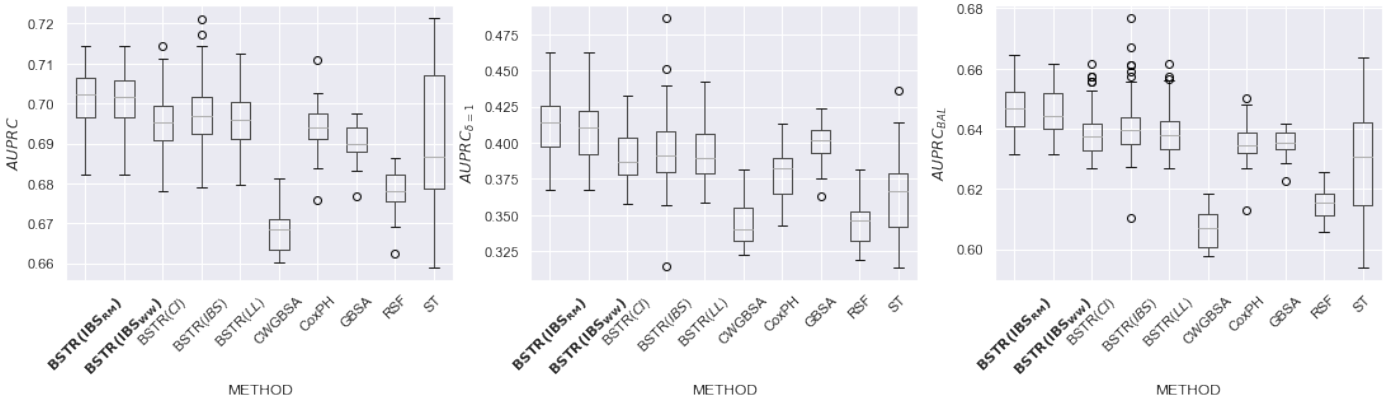


Figure 17. $AUPRC$ comparison for PBC dataset.

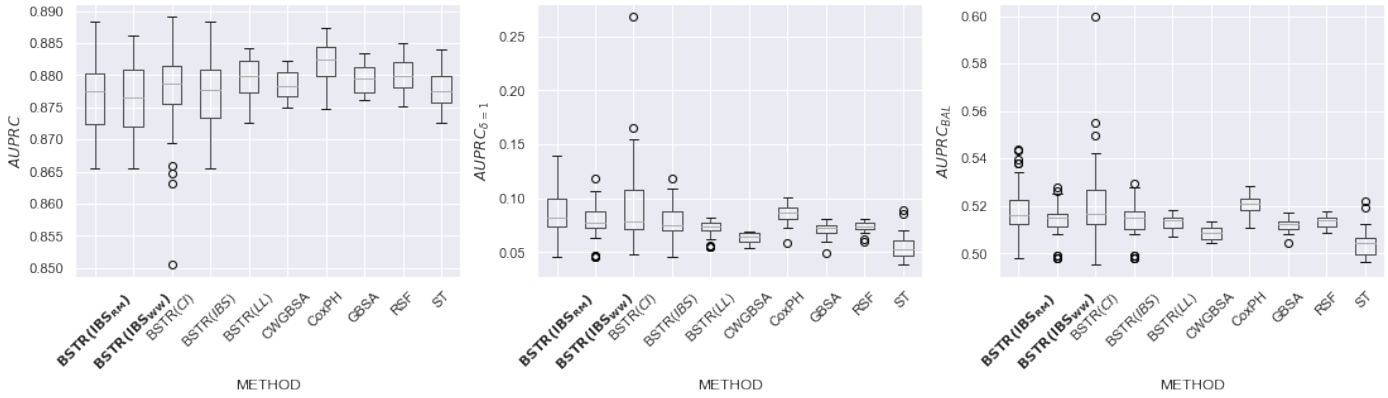


Figure 18. *AUPRC* comparison for *actg* dataset.

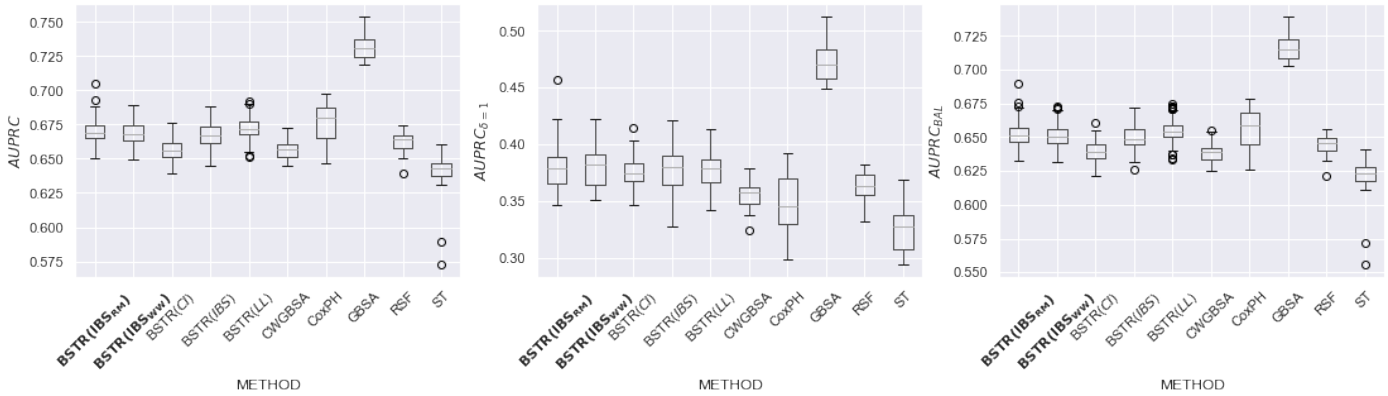


Figure 19. *AUPRC* comparison for *WUHAN* dataset.

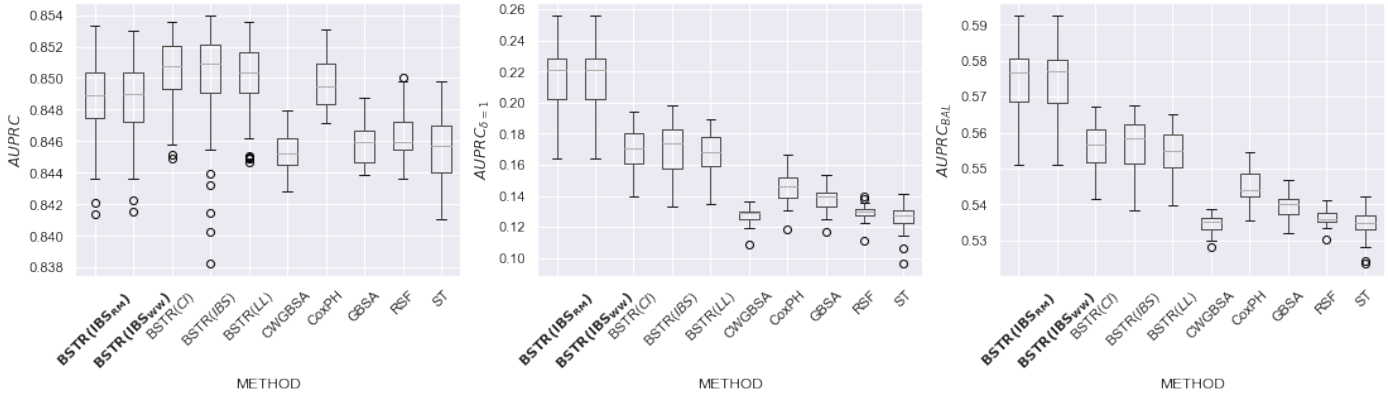


Figure 20. *AUPRC* comparison for *SMARTO* dataset.

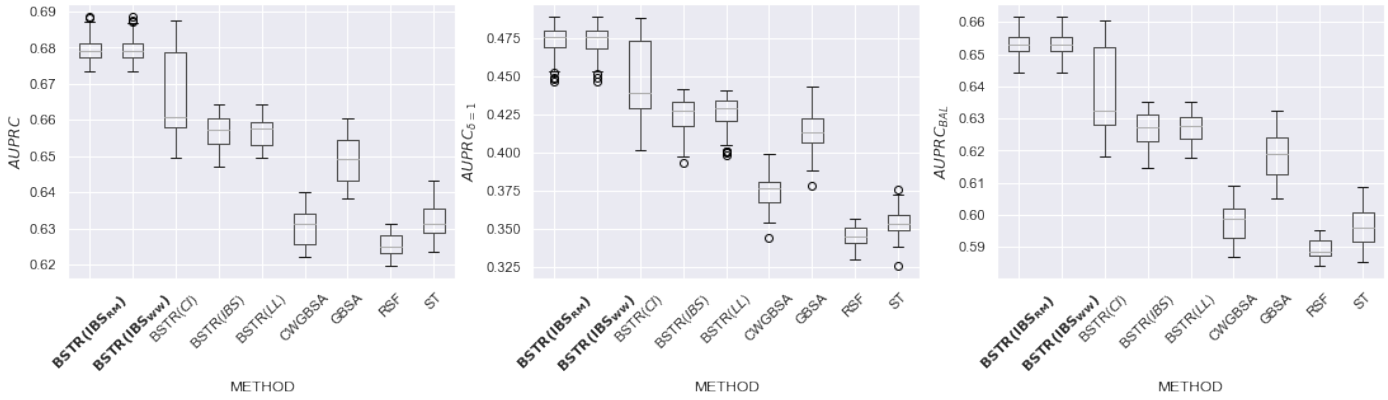


Figure 21. *AUPRC* comparison for *rott*2 dataset.

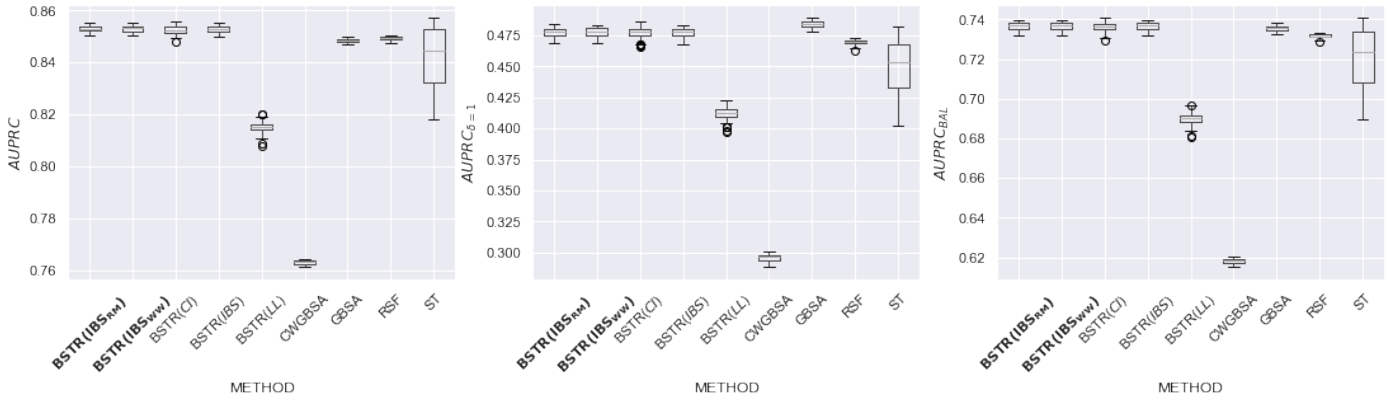


Figure 22. *AUPRC* comparison for $flchain$ dataset.

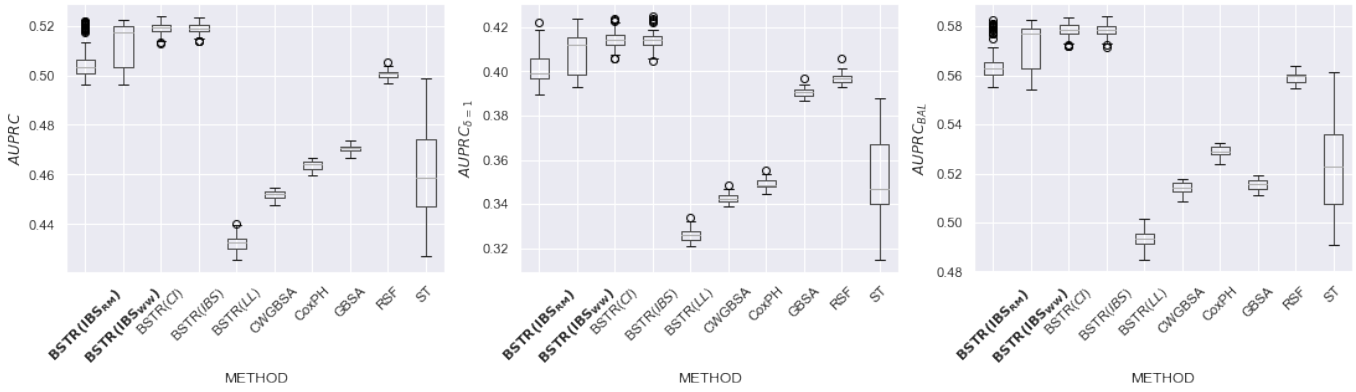


Figure 23. *AUPRC* comparison for $support2$ dataset.

Table 4 contains summary results for all datasets. According to the dataset and metric, each cell contains a list of the best loss functions, or the value "no winner" if all loss functions have achieved similar quality. The row "Total" determines the recommended loss function for each metric. The best loss function for all metrics is IBS_{RM}. The second place in all metrics is taken by IBS_{WW}. The third place in the $AUPRC$ metric is taken by CI. Thus, the IBS modifications lead to better quality compared to the original metric and alternative non-integrated metrics. In addition, for seven out of eight datasets, *Bagging* outperforms the existing scikit-survival models.

Table 4. Summary table of the best loss function for each dataset and metric (descending order of the datasets by the proportion of terminal events). The row "Total" defines the best loss function for each metric. The best loss function for all metrics is IBS_{RM}.

DATASET	$AUPRC$	$AUPRC_{\delta=1}$	$AUPRC_{BAL}$
*support*2	CI, IBS	CI, IBS	CI, IBS
WUHAN	IBS_{RM}, LL	IBS_{WW}, IBS_{RM}	IBS_{RM}, LL
GBSG	IBS_{WW}, IBS_{RM}	IBS_{RM}	IBS_{WW}, IBS_{RM}
*rott*2	IBS_{WW}, IBS_{RM}	IBS_{WW}, IBS_{RM}	IBS_{WW}, IBS_{RM}
PBC	IBS_{WW}, IBS_{RM}	IBS_{RM}	IBS_{WW}, IBS_{RM}
flchain	no winner	no winner	no winner
SMARTO	CI, IBS	IBS_{WW}, IBS_{RM}	IBS_{WW}, IBS_{RM}
actg	CI, LL	IBS_{RM}, CI	IBS_{RM}, CI
Total	IBS_{RM}	IBS_{RM}	IBS_{RM}

According to the results of an experimental study, the proposed modification of IBS_{RM} (used as a loss function for an ensemble of independent survival trees) showed an increase in quality for $AUPRC$, $AUPRC_{\delta=1}$, and $AUPRC_{BAL}$. Loss functions with an equal impact of events allow us to build models that are more resistant to imbalance. Thus, we should apply the proposed modifications of metrics to build a high-quality ensemble of survival trees. In further research, we plan to use IBS_{RM} for boosting model construction.

6.3. Discussion of Hyperparameters

In the experimental study, we showed the positive impact of the proposed modifications of IBS on the model quality. In this section, we use these modifications to analyze the hyperparameters of predictive models. For example, we have proposed analyzing the influence of two types of Kaplan–Meier leaf models (KM and KM10) that we defined in Section 2.2.1. Recall that KM10 is an extension of the standard Kaplan–Meier model with a zero value after the moment of the occurrence of the last event.

To analyze the hyperparameters, we use the obtained table in the second stage of the experiments. In particular, we compare the value of the loss function for each set of parameters. Based on the previously described algorithm, the result of this stage is the best set of hyperparameters for each model. We consider the relationship between the values of loss functions (IBS, IBS_{WW}, IBS_{RM}) and quality metrics ($AUPRC$, $AUPRC_{\delta=1}$, $AUPRC_{BAL}$). In plots, we note the values for each set of hyperparameters and stratify them relative to the type of leaf model. To increase visibility, we visualize an estimate of spatial density based on kernel functions.

Figure 24 shows nine plots of pairwise dependence of three quality metrics and three loss functions for *GBSG* datasets. The vertical axis corresponds to the quality of the cross-validation model. The horizontal axis corresponds to the values of the model's loss function on cross-validation. To choose the best hyperparameters, we consider the minimum value of IBS and the best quality related to the largest $AUPRC$. Hyperparameters with KM10 are highlighted in orange, with KM in blue.

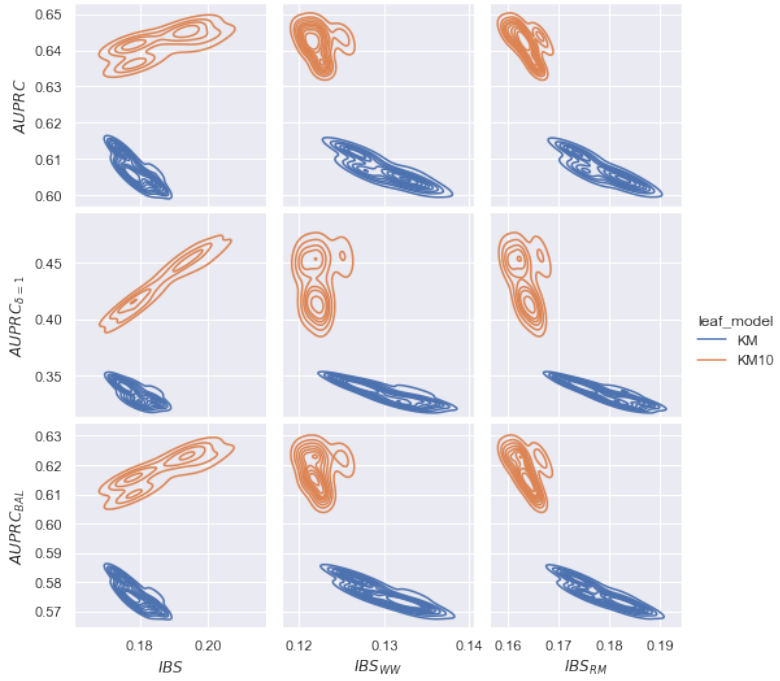


Figure 24. Example of the relationship between the model quality ($AUPRC$, $AUPRC_{\delta=1}$, $AUPRC_{BAL}$) at the cross-validation stage ($GBSG$ dataset) and loss functions (IBS, IBS_{WW}, IBS_{RM}). Based on the quality of *Bagging* for each hyperparameter, we visualize a kernel density estimation. The KM10 leaf model provides better quality than KM and corrects the linear relationship in the case of the modified IBS metrics. In addition, the best quality is reached for modified IBS metrics.

Based on Figure 24, we notice the following conclusions. Firstly, for all plots, the density of KM10 hyperparameters is higher than those for KM. Therefore, using the KM10 leaf model, $AUPRC$ increases. Secondly, minimizing the IBS metric (left plots), there is a linear relationship between IBS and $AUPRC$ for KM10. This dependence is contrary to the best quality of IBS and $AUPRC$. Therefore, by choosing hyperparameters with the lowest value of IBS, the quality of $AUPRC$ is also lower. Note that we observe the inverse linear relationship for the KM model ($AUPRC$ increasing for decreasing IBS).

However, minimizing IBS_{WW} and IBS_{RM} metrics (central and right plots), we observe an inverse linear relationship between the loss function and the quality metric for the KM10 hyperparameter. Thus, modifications of IBS restore the correct relationship for the KM10 parameter. Finally, minimizing IBS (left plots), there is an unstable choice between the KM10 and KM parameters, although the KM10 class leads to a significant improvement in quality. In the case of IBS modifications, the KM10 class shifts to the left and leads to a stable minimal loss value. Thus, modified metrics allows us to detect a class of hyperparameters that leads to an improvement of $AUPRC$.

7. Conclusions

Survival analysis data have several specific characteristics, such as rare early and late events and the proportion of classes. The rare late events often do not correspond to the general distribution of time and contribute to the bias of forecasts. In this paper, we have researched the excessive sensitivity of survival analysis metrics to data features. We have determined four cases of increased sensitivity: the higher significance of partial events, the growth of integral metrics in time, the impact of time bins, and the influence of the imbalance of censored observations.

IBS and $IAUC$ metrics increase the contribution of rare late events, which leads to a distortion of the quality assessment. $AUPRC$, IBS, and $IAUC$ metrics are unstable due to class imbalance. To set the equality of observation impacts, we adjust the weighting schemes of the event contribution and propose a controlled averaging approach. In particular,

IBS_{RM} has equal contributions of times and partial events. In addition, IBS_{BAL}, $IAUC_{BAL}$, and $AUPRC_{BAL}$ modifications provides equal contributions of censured and terminal observations. Based on the analytical study, we recommend $AUPRC$ to evaluate the prediction of the survival function.

The experimental study included eight datasets of real medical data. The goal of the study was to assess the impact of the loss function (LL, CI, IBS, and the proposed modifications) on the quality (according to the metrics $AUPRC, AUPRC_{\delta=1}, AUPRC_{BAL}$) of the *Bagging* ensemble of independent survival trees. The *Bagging* uses a loss function to select the size and hyperparameters of the ensemble. According to the experimental results, IBS_{RM} shows an increase in quality compared to the original metric and alternative non-integrated metrics. Loss functions with an equal impact of events allow us to build models that are more resistant to imbalance. In addition, IBS_{RM} allows us to detect a class of hyperparameters (with leaf model as extended Kaplan–Meier) that leads to an improvement of $AUPRC$. Finally, for seven datasets, the *Bagging* model outperforms the existing models of the scikit-survival library.

In further research, we plan to apply the proposed modifications of metrics to build a boosting ensemble of survival trees. The usage of stable metrics will prevent overfitting and bias in the model. In addition, we plan to study the quality of the proposed approaches on real datasets from alternative applications of survival analysis.

Author Contributions: Conceptualization, M.P. and I.M.; methodology, I.V. and M.P.; software, I.V.; validation, I.V.; formal analysis, I.V. and M.P.; investigation, M.P. and I.M.; resources, M.P. and I.M.; data curation, I.V.; writing—original draft preparation, I.V.; writing—review and editing, M.P. and I.M.; visualization, I.V.; supervision, M.P.; project administration, I.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The authors confirm that the data supporting the findings of this study are available within the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kleinbaum, D.; Klein, M. *Survival Analysis: A Self-Learning Text*, 3rd ed.; Statistics for Biology and Health; Springer: New York, NY, USA, 2016.
2. Wang, P.; Li, Y.; Reddy, C.K. Machine learning for survival analysis: A survey. *ACM Comput. Surv. (CSUR)* **2019**, *51*, 1–36. [CrossRef]
3. Lee, S.H. Weighted Log-Rank Statistics for Accelerated Failure Time Model. *Stats* **2021**, *4*, 348–358. [CrossRef]
4. Karadeniz, P.G.; Ercan, I. Examining tests for comparing survival curves with right censored data. *Stat Transit* **2017**, *18*, 311–28.
5. Lee, S.H. On the versatility of the combination of the weighted log-rank statistics. *Comput. Stat. Data Anal.* **2007**, *51*, 6557–6564. [CrossRef]
6. Brendel, M.; Janssen, A.; Mayer, C.D.; Pauly, M. Weighted logrank permutation tests for randomly right censored life science data. *Scand. J. Stat.* **2014**, *41*, 742–761. [CrossRef]
7. Hasegawa, T. Group sequential monitoring based on the weighted log-rank test statistic with the Fleming–Harrington class of weights in cancer vaccine studies. *Pharm. Stat.* **2016**, *15*, 412–419. [CrossRef]
8. Kvamme, H.; Borgan, Ø. Continuous and discrete-time survival prediction with neural networks. *Lifetime Data Anal.* **2021**, *27*, 710–736. [CrossRef] [PubMed]
9. Etikan, I.; Abubakar, S.; Alkassim, R. The Kaplan-Meier estimate in survival analysis. *Biom. Biostat. Int. J.* **2017**, *5*, 00128. [CrossRef]
10. Andersen, P.K. Fifty years with the Cox proportional hazards regression model. *J. Indian Inst. Sci.* **2022**, *102*, 1135–1144. [CrossRef]
11. Lin, D. On the Breslow estimator. *Lifetime Data Anal.* **2007**, *13*, 471–480. [CrossRef] [PubMed]
12. Longato, E.; Vettoretti, M.; Di Camillo, B. A practical perspective on the concordance index for the evaluation and selection of prognostic time-to-event models. *J. Biomed. Inform.* **2020**, *108*, 103496. [CrossRef] [PubMed]
13. Heagerty, P.J.; Zheng, Y. Survival model predictive accuracy and ROC curves. *Biometrics* **2005**, *61*, 92–105. [CrossRef]
14. Lambert, J.; Chevret, S. Summary measure of discrimination in survival models based on cumulative/dynamic time-dependent ROC curves. *Stat. Methods Med. Res.* **2016**, *25*, 2088–2102. [CrossRef]
15. Kvamme, H.; Borgan, Ø.; Scheel, I. Time-to-event prediction with neural networks and Cox regression. *arXiv* **2019**, arXiv:1907.00825.

16. Clim, A.; Zota, R.D.; TinicĂ, G. The Kullback-Leibler divergence used in machine learning algorithms for health care applications and hypertension prediction: A literature review. *Procedia Comput. Sci.* **2018**, *141*, 448–453. [CrossRef]
17. Yari, G.; Mirhabibi, A.; Saghafi, A. Estimation of the Weibull parameters by Kullback-Leibler divergence of Survival functions. *Appl. Math. Inf. Sci* **2013**, *7*, 187–192. [CrossRef]
18. Haider, H.; Hoehn, B.; Davis, S.; Greiner, R. Effective Ways to Build and Evaluate Individual Survival Distributions. *J. Mach. Learn. Res.* **2020**, *21*, 3289–3351.
19. Avati, A.; Duan, T.; Zhou, S.; Jung, K.; Shah, N.H.; Ng, A.Y. Countdown regression: sharp and calibrated survival predictions. In Proceedings of the Uncertainty in Artificial Intelligence, Virtual, 3–6 August 2020; pp. 145–155.
20. Hastie, T.; Tibshirani, R.; Friedman, J.H.; Friedman, J.H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: Berlin/Heidelberg, Germany, 2009; Volume 2.
21. Dormuth, I.; Liu, T.; Xu, J.; Yu, M.; Pauly, M.; Ditzhaus, M. Which test for crossing survival curves? A user's guideline. *BMC Med. Res. Methodol.* **2022**, *22*, 34. [CrossRef]
22. Bou-Hamad, I.; Larocque, D.; Ben-Ameur, H. A review of survival trees. *Statist. Surv.* **2011**, *5*, 44–71. [CrossRef]
23. Ishwaran, H.; Kogalur, U.B.; Blackstone, E.H.; Lauer, M.S. Random survival forests. *Ann. Appl. Stat.* **2008**, *2*, 841–860. [CrossRef]
24. Vasilev, I.; Petrovskiy, M.; Mashechkin, I.V. Survival Analysis Algorithms based on Decision Trees with Weighted Log-rank Criteria. In Proceedings of the 11th International Conference on Pattern Recognition Applications and Methods, Virtual, 3–5 February 2022; pp. 132–140.
25. Pölsterl, S. scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn. *J. Mach. Learn. Res.* **2020**, *21*, 1–6.
26. Nguyen, N.P. *Gradient Boosting for Survival Analysis with Applications in Oncology*; University of South Florida: Tampa, FL, USA, 2019.
27. Drysdale, E. SurvSet: An open-source time-to-event dataset repository. *arXiv* **2022**, arXiv:2203.03094.
28. Schumacher, M. Rauschecker for the german breast cancer study group, randomized 2× 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive lbreast cancer patients. *J. Clin. Oncol.* **1994**, *12*, 2086–2093. [CrossRef]
29. Royston, P.; Lambert, P.C. *Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model*; Stata Press: College Station, TX, USA, 2011; Volume 347.
30. Knaus, W.A.; Harrell, F.E.; Lynn, J.; Goldman, L.; Phillips, R.S.; Connors, A.F.; Dawson, N.V.; Fulkerson, W.J.; Califf, R.M.; Desbiens, N.; et al. The SUPPORT prognostic model: Objective estimates of survival for seriously ill hospitalized adults. *Ann. Intern. Med.* **1995**, *122*, 191–203. [CrossRef]
31. Yan, L.; Zhang, H.T.; Goncalves, J.; Xiao, Y.; Wang, M.; Guo, Y.; Sun, C.; Tang, X.; Jing, L.; Zhang, M.; et al. An interpretable mortality prediction model for COVID-19 patients. *Nat. Mach. Intell.* **2020**, *2*, 283–288. [CrossRef]
32. Kaplan, M.M. Primary biliary cirrhosis. *N. Engl. J. Med.* **1996**, *335*, 1570–1580. [CrossRef]
33. Simons, P.C.G.; Algra, A.; Van De Laak, M.; Grobbee, D.; Van Der Graaf, Y. Second manifestations of ARTerial disease (SMART) study: rationale and design. *Eur. J. Epidemiol.* **1999**, *15*, 773–781. [CrossRef]
34. Hammer, S.M.; Squires, K.E.; Hughes, M.D.; Grimes, J.M.; Demeter, L.M.; Currier, J.S.; Eron, J.J., Jr.; Feinberg, J.E.; Balfour, H.H., Jr.; Deyton, L.R.; et al. A controlled trial of two nucleoside analogues plus indinavir in persons with human immunodeficiency virus infection and CD4 cell counts of 200 per cubic millimeter or less. *N. Engl. J. Med.* **1997**, *337*, 725–733. [CrossRef]
35. Kyle, R.A.; Therneau, T.M.; Rajkumar, S.V.; Larson, D.R.; Plevak, M.F.; Offord, J.R.; Dispenzieri, A.; Katzmann, J.A.; Melton, L.J., III. Prevalence of monoclonal gammopathy of undetermined significance. *N. Engl. J. Med.* **2006**, *354*, 1362–1369. [CrossRef]
36. Hung, H.; Chiang, C.T. Estimation methods for time-dependent AUC models with survival data. *Can. J. Stat.* **2010**, *38*, 8–26. [CrossRef]
37. Uno, H.; Cai, T.; Tian, L.; Wei, L.J. Evaluating prediction rules for t-year survivors with censored regression models. *J. Am. Stat. Assoc.* **2007**, *102*, 527–537. [CrossRef]
38. Chawla, N.V. Data mining for imbalanced datasets: An overview. In *Data Mining and Knowledge Discovery Handbook*; Springer: New York, NY, USA, 2010; pp. 875–886.
39. He, H.; Ma, Y. *Imbalanced Learning: Foundations, Algorithms, and Applications*; Wiley-IEEE Press: Hoboken, NJ, USA, 2013.
40. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [CrossRef]
41. Tong, L.I.; Chang, Y.C.; Lin, S.H. Determining the optimal re-sampling strategy for a classification model with imbalanced data using design of experiments and response surface methodologies. *Expert Syst. Appl.* **2011**, *38*, 4222–4227. [CrossRef]
42. Chen, Y.; Jia, Z.; Mercola, D.; Xie, X. A gradient boosting algorithm for survival analysis via direct optimization of concordance index. *Comput. Math. Methods Med.* **2013**, *2013*, 873595. [CrossRef] [PubMed]
43. Binder, H.; Binder, M.H. Package 'CoxBoost'. 2015. Available online: https://cran.r-hub.io/web/packages/CoxBoost/CoxBoost.pdf (accessed on 20 August 2023).
44. Bai, M.; Zheng, Y.; Shen, Y. Gradient boosting survival tree with applications in credit scoring. *J. Oper. Res. Soc.* **2022**, *73*, 39–55. [CrossRef]
45. Refaeilzadeh, P.; Tang, L.; Liu, H. Cross-validation. *Encycl. Database Syst.* **2009**, *5*, 532–538.