






Research article

DOI: <https://doi.org/10.18721/JCSTCS.17302>

UDC 004.85



INTERPRETATION METHODS FOR MACHINE LEARNING MODELS IN THE FRAMEWORK OF SURVIVAL ANALYSIS WITH CENSORED DATA: A BRIEF OVERVIEW

L.V. Utkin  , *A.V. Konstantinov* , *D.Yu. Eremenko* ,
V.S. Zaborovsky, *V.A. Muliukha* 

Peter the Great St. Petersburg Polytechnic University,
St. Petersburg, Russian Federation

 lev.utkin@gmail.com

Abstract. Methods of interpretation, or explanation, of predictions are an integral part of modern black-box machine learning models. They have become widespread due to the need for the user to understand what the machine learning model is predicting. This is especially important for survival analysis models, as they are used in medicine, system reliability, safety, and also have features that make them difficult to explain and interpret. The paper discusses the main methods for interpreting survival models that deal with censored data and determine the characteristics of the time until a certain event. A feature of such models is that their predictions are presented not as a point value, but as a probabilistic function of time, for example, a survival function or a risk function. This requires the development of special interpretation methods. The most well-known methods SurvLIME, SurvLIME-KS, SurvNAM and SurvBeX, SurvSHAP(t) are considered, which are based on the use of LIME and SHAP interpretation methods, the Cox model and its modifications, as well as the Beran estimator.

Keywords: machine learning, survival model, explainable artificial intelligence, censored data, Cox model, Beran estimator

Acknowledgements: The research was partially financially supported by the Ministry of Science and Higher Education of the Russian Federation within the framework of the state assignment “Development and research of machine learning models for solving fundamental problems of artificial intelligence in the fuel and energy complex” (FSEG-2024-0027).

Citation: Utkin L.V., Konstantinov A.V., Eremenko D.Yu., et al. Interpretation methods for machine learning models in the framework of survival analysis with censored data: a brief overview. Computing, Telecommunications and Control, 2024, Vol. 17, No. 3, Pp. 22–31. DOI: 10.18721/JCSTCS.17302





Научная статья

DOI: <https://doi.org/10.18721/JCSTCS.17302>

УДК 004.85



МЕТОДЫ ИНТЕРПРЕТАЦИИ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ В РАМКАХ АНАЛИЗА ВЫЖИВАЕМОСТИ ПРИ ЦЕНЗУРИРОВАННЫХ ДАННЫХ: КРАТКИЙ ОБЗОР

Л.В. Уткин , А.В. Константинов , Д.Ю. Еременко ,
В.С. Заборовский, В.А. Мулюха 

Санкт-Петербургский политехнический университет Петра Великого,
Санкт-Петербург, Российская Федерация

✉ lev.utkin@gmail.com

Аннотация. Методы интерпретации, или объяснения, предсказаний являются неотъемлемой частью современных моделей машинного обучения типа «черный ящик». Они получили широкое распространение, что обусловлено необходимостью понимания пользователем того, что предсказывает модель машинного обучения. Это особенно относится к моделям анализа выживаемости, так как они используются в медицине, надежности, безопасности, а также имеют особенности, которые усложняют их объяснение и интерпретацию. В работе рассматриваются основные методы интерпретации моделей выживаемости, которые оперируют с цензурированными данными и определяют характеристики времени до определенного события. Особенностью таких моделей является то, что их предсказания представляются не в виде некоторого точечного значения, а в виде вероятностной функции времени, например, функции выживаемости или функции риска. Это требует необходимости разработки специальных методов интерпретации. Рассмотрены наиболее известные методы SurvLIME, SurvLIME-KS, SurvNAM и SurvBeX, SurvSHAP(t), которые основаны на использовании методов интерпретации предсказаний LIME и SHAP, модели Кокса и ее модификации, а также оценки Берана.

Ключевые слова: машинное обучение, модель выживаемости, объяснимый искусственный интеллект, цензурированные данные, модель Кокса, оценка Берана

Финансирование: Исследование выполнено при частичной финансовой поддержке Министерства науки и высшего образования Российской Федерации в рамках государственного задания «Разработка и исследование моделей машинного обучения для решения фундаментальных задач искусственного интеллекта в топливно-энергетическом комплексе» (FSEG-2024-0027).

Для цитирования: Utkin L.V., Konstantinov A.V., Eremenko D.Yu., et al. Interpretation methods for machine learning models in the framework of survival analysis with censored data: a brief overview // Computing, Telecommunications and Control. 2024. Т. 17, № 3. С. 22–31. DOI: 10.18721/JCSTCS.17302

Introduction

The increasing importance of machine learning models, particularly deep learning models, and their widespread use in various applications has led to the problem of prediction explanation and interpretation. The development and implementation of intelligent systems based on machine learning models for solving various application tasks is currently one of the most rapidly growing areas of the artificial intelligence (AI) applications, and it leads to the problem of explaining or interpreting predictions provided by the models. This problem stems from the fact that, despite the importance of the AI applications in many real tasks, there are several obstacles to further implementation of AI especially in such areas as medicine, system reliability, etc. because the corresponding machine learning models are often perceived as “black

boxes” meaning that the inner workings of these models are often completely unknown. As a result, it is difficult to understand and explain, why the models provide a particular prediction for a particular input instance. The importance of the explanation problem prompts the development of corresponding additional models aimed at explaining and interpreting the obtained predictions. The explanation of a model prediction means to find features of the explained instance that most influence the prediction. In other words, the meta-model should suggest, which features of the explained instance cause the corresponding prediction.

It should be noted that many explanation methods have been proposed recently, which are discussed in several review articles [1–4]. If we consider explanation methods in terms of simultaneously explained instance numbers, all methods can be divided into two large groups. The first group consists of *local* methods, which try to explain predictions obtained for a single test instance or for a small set of instances. The second group contains the *global* methods which explain predictions of the entire dataset on average. The first group of methods is more important, because many applications require to explain a certain instance, for example, a doctor prefers to understand a diagnosis, which is predicted by a machine learning model for a certain patient, but not for all the patients in a hospital.

Most methods of the local explanation are based on training a special meta-model, which is self-explainable, and it approximates the black-box model prediction function at a point, which corresponds to the explained instance. One of the ideas behind several explanation methods is to approximate with a linear model, because the coefficients of the linear model can be interpreted as quantitative measures of the feature importance. Following this idea, the well-known explanation model, called LIME (Local Interpretable Model-Agnostic Explanation), has been proposed by Ribeiro et al. [5]. According to LIME, a linear approximation of a non-linear prediction function of the black-box model at an instance is built. It is carried out by generating synthetic instances around the explained instance with such weights that each weight depends on a distance between the explained instance and the generated synthetic instance. Another well-known explanation method is so-called SHAP method (SHapley Additive exPlanations) proposed in [6, 7]. SHAP is based on the game-theoretic Shapley values, which can be regarded as the contributions of features into the black-box model prediction. Applications of SHAP meet two important difficulties. First, its complexity rapidly increases with the number of features. Second, it uses subsets of features as inputs for the black-box model, which must be added by some values of removed features that are not strongly defined.

To improve the linear explanation models and to overcome weakness of the linear approximations, more complex explanation models have been proposed. They are based on sums of the feature shape functions, which form the Generalized Additive Model (GAM) introduced by Hastie and Tibshirani in [8]. The GAM motivated to develop several interesting explanation models, including the Neural Additive Model (NAM) introduced by Agarwal et al. in [9], a weighted sum of separate gradient boosting machines (GBMs) presented in [10].

It is important to point out that the aforementioned models and their modifications have been developed to deal with various types of data. However, there is a class of datasets, which consider times to the events of interest.

Machine learning models, trained on data characterizing the time to occurrence of certain events of some objects depending on the structure of these objects, are becoming increasingly widespread [11]. This is due to their use in a variety of areas, for example, in the system reliability, when events of system failure are considered, in medicine, when the event is the recovery or death of a patient. One of the crucial features of many models is that events associated with some objects may not be observed, but only the last moment of observation is recorded, assuming that the event will occur in the future, but we do not know when. Such data are called censored, and they contain significantly less information about the object than uncensored data, for which the time of the event is known. However, censored data can also be used in machine learning models called survival models.

One of the well-known survival models is the Cox proportional hazard model [12]. According to the model, the covariates (features) of an object are linearly connected. On the one hand, this feature can be viewed as a limitation of the model since the relationship between features may be significantly nonlinear in some cases. To account for various relationships between features, a large number of survival models have been developed recently, for example, random survival forests, deep survival neural networks, modifications of the support vector machine and others [11]. At the same time, each model is a black box, that is, only inputs and the corresponding outputs (predictions) are known, but it is not known how a prediction is obtained, which features of the object influence the prediction of the model. However, an important feature of many survival models is that their predictions are presented in the form of the survival function (SF) or the cumulative hazard function (CHF). This fact significantly complicates the solution of the explanation problem and requires special approaches to its solution.

We provide a brief overview of the most important explanation methods within survival models.

Concepts of survival analysis

Let us consider the training set D consisting of n triplets $(\mathbf{x}_i, T_i, \delta_i)$, $i = 1, \dots, n$, where each triplet characterizes an object, $\mathbf{x}_i \in \mathbf{R}^m$ is the feature vector; T_i is the event time of the i -th object; δ_i is the indicator of the event, $\delta_i = 1$, if the event is observed (uncensored observation), $\delta_i = 0$, if the event is not observed (censored observation). We aim to estimate the event time T on the basis of D for a new object having features \mathbf{x} .

Important concepts in survival analysis are SFs and CHFs [11]. The SF $S(t|\mathbf{x})$ is defined as the probability of surviving the object \mathbf{x} up to time t . Another concept is the CHF $H(t|\mathbf{x})$, which is expressed through the SF as $H(t|\mathbf{x}) = -\ln(S(t|\mathbf{x}))$.

One of the base survival models, which can be regarded as the basis for several explanation methods, is the well-known Cox proportional hazards model [12]. According to the Cox model, the conditional CHF is determined as follows [12]:

$$H(t|\mathbf{x}, \mathbf{b}) = H_0(t) \cdot \exp(\mathbf{b}^T \mathbf{x}),$$

where $H_0(t)$ is the baseline CHF, which can be estimated by using the Nelson–Aalen or Kaplan–Meier estimators; \mathbf{b} is the vector of the model parameters.

It can be seen from the expression for the Cox CHF that the linear relationship assumption between covariates and the log-risk of an event is accepted in the model. This is a very important property of the Cox model, which allows us to approximate an arbitrary CHF produced as an output of the black-box survival model by the Cox model and, therefore, to construct methods explaining survival models.

The Cox model is popular in many real tasks. However, its linear assumption significantly restricts its application, because many real survival datasets violate this assumption. GAM incorporated into the Cox model instead of the linear expression partially relaxes this assumption. This representation will be used in one of the explanation methods. Another shortcoming of the Cox model is that it does not take into account the positional relationship of feature vectors. This difficulty can be resolved by using the Beran estimator [14]. According to the Beran estimator, a SF can be estimated as follows:

$$S_B(t|\mathbf{x}) = \prod_{T_i \leq t} \left\{ 1 - \frac{W(\mathbf{x}, \mathbf{x}_i)}{1 - \sum_{j=1}^{i-1} W(\mathbf{x}, \mathbf{x}_j)} \right\}^{\delta_i}.$$

Here the weight $W(\mathbf{x}, \mathbf{x}_i)$ characterizes how the vectors \mathbf{x}, \mathbf{x}_i are close to each other. It should be noted that the kernel function measures how similar any two feature vectors are. Therefore, the weights

are nothing else, but the normalized kernels. For example, if the kernel is Gaussian, then the weight is defined through the softmax operation: $W(\mathbf{x}, \mathbf{x}_i) = \text{softmax}\left(\frac{\|\mathbf{x} - \mathbf{x}_i\|^2}{\tau}\right)$, where τ is a hyperparameter. It is interesting to note that the Kaplan–Meier estimator can be viewed as a special case of the Beran estimator, when all weights $W(\mathbf{x}, \mathbf{x}_i)$ are identical and equal to $1/n$, where n is the number of instances in the dataset.

Explanation formal problem statements

Formally, the explanation problem is solved by means of training a meta-model or an explanation model that approximates the explainable black-box model in the vicinity of the example being explained and belongs to a set of “simple” models that are interpretable (linear models, decision trees). The explainable black-box model implements the function $f: \mathbf{R}^m \rightarrow \mathbf{R}^d$, for example, in classification $f(\mathbf{x})$ is the probability (or indicator) that the feature vector \mathbf{x} belongs to a certain class. A meta-model is a model $g \in G$, where G is a class of interpretable models, which is a solution to the optimization problem:

$$\min_{g \in G} \{L(f, g, \theta) + \Omega(g)\},$$

where $L(f, g, \theta)$ is the measure of how poorly g approximates f ; θ is the parameter vector; $\Omega(g)$ is the regularization term.

One of the most popular interpreted functions is the linear function. This is due to the fact that the coefficients of a linear function precisely characterize the influence of each feature on the value of the function. In fact, the local explanation problem comes down to approximating the function $f(\mathbf{x})$ of the black-box model by a linear function $g(\mathbf{x})$ at instance \mathbf{x} . The well-known LIME method [5] and its modifications are based on this idea. In LIME, to construct the approximating function $g(\mathbf{x})$, N instances (vectors $\mathbf{z}_i \in \mathbf{R}^m$) are generated in the vicinity of the instance \mathbf{x} . Using the black-box model, the prediction $y_i = f(\mathbf{z}_i)$ is computed for each generated point \mathbf{z}_i and a new dataset of N points (\mathbf{z}_i, y_i) is formed. The resulting dataset is used to construct the linear function $g(\mathbf{x})$. LIME is used to explain the classification and regression models. However, its application to survival models meets some difficulties, since, firstly, survival models deal with censored data, for which the construction of a regression model differs from standard models. Secondly, the output of the survival model, that is y , is the SF rather than the point value, which also complicates the task of explanation, since the entire SF has to be interpreted.

Another explanation method is SHAP [6, 7]. According to this method, the i -th feature average contribution is estimated by the Shapley value:

$$\phi_i(f) = \phi_i = \sum_{S \subseteq N/\{i\}} \frac{|S|!(m-|S|-1)!}{m!} [f(S \cup \{i\}) - f(S)],$$

where $f(S)$ is a prediction of the black-box model under condition that a subset S of features is used as the corresponding input; m is the number of all features.

A controversial question in SHAP is how to represent or to fill features from the subset $\{1, \dots, m\}/S$ to apply the black-box model. There are several approaches to partially solve this [14]. However, every approach has disadvantages and cannot be used in all cases.

Explanation methods in survival analysis

Due to the mentioned peculiarities of the survival machine learning models, the known explanation methods like LIME and SHAP to explain the corresponding predictions cannot be directly used. Therefore, every explanation method is based on applying a trick, which allows us to adapt it to LIME or SHAP.

The main idea behind the first group of explanation methods, called SurvLIME, SurvLIME-KS, SurvLIME-Inf, is to use the Cox model to approximate the black-box model in a local area around the explained object \mathbf{x} . This idea stems from the fact that the Cox model assumes the linear relationship $\mathbf{b}^T \mathbf{x}$ of the object features or covariates. An important peculiarity of the Cox model is that functions of features and the time are separated. Hence, coefficients \mathbf{b} in the Cox model can be regarded as measures how the features impact on predictions. However, we do not approximate a point prediction, but rather functions, such as CHF. According to the SurvLIME method [15], synthetic instances (feature vectors) \mathbf{z}_i are randomly generated around the explainable example, and for each synthetic vector the CHF is calculated using the black-box model. Therefore, we propose to consider the distance between logarithms of the CHFs $H(t|\mathbf{z}_i)$ predicted by the black-box model and CHFs $H^{Cox}(t|\mathbf{z}_i)$ computed by using the Cox model. The distance is determined as follows:

$$d(\mathbf{z}_i) = \int_0^\infty (\ln H(t|\mathbf{z}_i) - \ln H^{Cox}(t|\mathbf{z}_i))^2 dt.$$

Let $t_1 \leq t_2 \leq \dots \leq t_n$ be the distinct event times obtained from the set of training times $\{T_1, \dots, T_n\}$, where $t_1 = \min_{k=1, \dots, n} T_k$, $t_n = \max_{k=1, \dots, n} T_k$. Substituting the corresponding expressions for CHF into the above expression and using the fact that the CHF are stepwise functions due to the finite number of the observed event times, we obtain after simplification:

$$d(\mathbf{z}_i) = \sum_{j=0}^n (\ln H_j(\mathbf{z}_i) - \ln H_{0,j}(\mathbf{z}_i) - \mathbf{b}^T \mathbf{z}_i)^2 \tau_j,$$

where $\tau_j = t_{j+1} - t_j$; $H(\mathbf{z}_i)$ is the CHF $H(t|\mathbf{z}_i)$ in the interval $[t_j, t_{j+1}]$; $H_{0,j}(\mathbf{z}_i)$ is the baseline CHF of the Cox model in the same interval of time.

Since the CHF in the Cox model is a function of unknown coefficients \mathbf{b} , the optimization problem for calculating the coefficients is determined by the weighted average distance between the CHF of the black-box model and the Cox model over all generated points \mathbf{z}_i , $i = 1, \dots, N$, taking into account weights w_i that are determined by the distances between point \mathbf{x} and each point \mathbf{z}_i . This implies that the loss function is of the form:

$$L(\mathbf{b}) = \min_{\mathbf{b}} \sum_{i=0}^N w_i d(\mathbf{z}_i).$$

It is interesting to point out that the obtained optimization problem is convex, therefore, its solution does not meet any difficulties.

The general scheme of the method is shown in Fig. 1. If the distance between risk functions is calculated based on the quadratic norm L_2 , then the optimization problem is reduced to quadratic programming, which makes it possible to find a solution (vector \mathbf{b}) quite simply. For norms L_1 and L_∞ (SurvLIME-Inf [16]) it is shown that the corresponding optimization problems are reduced to linear.

To ensure the robustness of the explanation mode, the Kolmogorov–Smirnov bounds for the SF were introduced in [17]. The proposed method, called SurvLIME-KS, is an extension of the SurvLIME method and uses the results of this method, but under the assumption that instead of a single SF, a set of SFs is used. As a result, a maximin optimization problem for calculating the optimal vector \mathbf{b} is constructed, where the maximum is determined over all SFs within the Kolmogorov–Smirnov bounds. Despite the seeming complexity of the optimization problem, it reduces to a finite set of quadratic or linear optimization problems whose solutions do not meet any difficulties.

SurvLIME provides explanation of the survival black-box model predictions. However, it is interesting to explain why predicted results of a survival model are uncertain or to answer the question: Which

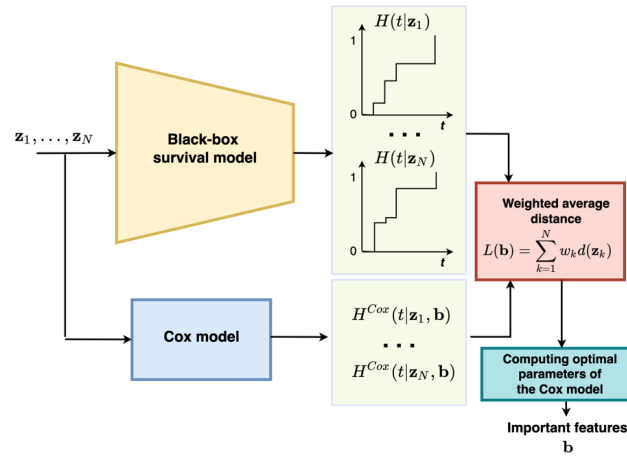


Fig. 1. SurvLIME structure [16]

features of an explained instance lead the black-box model prediction to be uncertain? Following the above question, a method for uncertainty interpretation of the black-box survival model predictions called UncSurvEx (Uncertainty Survival Explanation) has been proposed in [18]. The method like SurvLIME applies the Cox model to approximate the survival model. It is assumed that the most uncertain density function of the prediction is uniform at times T_1, \dots, T_n . This is similar to the multi-class classification, where the prediction is the class probability distribution. The most uncertain prediction is when all probabilities are identical, and we cannot make decision about a class of an instance. According to UncSurvEx, an “certainty” measure $c(\mathbf{z}_i)$ as the distance d between the actual SF $S(t|\mathbf{z}_i)$ and the most uncertain “uniform” SF $S^u(t|\mathbf{z}_i)$ is determined. In the same way, the “certainty” measure $c^{Cox}(\mathbf{z}_i, \mathbf{b})$ is determined as the distance between the Cox SF $S^{Cox}(t|\mathbf{z}_i, \mathbf{b})$ and the most uncertain SF $S^u(t|\mathbf{z}_i)$. The weighted difference between $c(\mathbf{z}_i)$ and $c^{Cox}(\mathbf{z}_i, \mathbf{b})$ is minimized to get the optimal values \mathbf{b} , which show the most important features from the prediction uncertainty point of view.

Another interesting explanation method is a generalization of the NAM method [10] to the case of censored data, called SurvNAM [19]. The idea behind the method is similar to the SurvLIME method, but unlike the linear combination $\mathbf{b}^T \mathbf{x}$ of attributes adopted in SurvLIME in accordance with the Cox model, this combination in SurvNAM is replaced by a set of neural networks that calculates the functions $g_i(x_i)$ of features. The modified Cox model is of the form:

$$H(t|\mathbf{x}, \mathbf{g}) = H_0(t) \cdot \exp(g_1(x_1) + \dots + g_m(x_m)),$$

where $\mathbf{g} = (g_1(x_1), \dots, g_m(x_m))$ is the vector of functions.

Neural networks implementing functions $g_i(x_i)$ are trained according to a loss function defined by the average distance between CHF of the black-box model and the Cox model over all generated synthetic points. As a result, we obtain functions $g_i(x_i)$, called the shape functions. In order to explain a prediction by using the shape functions, the rate of change of each function has to be estimated. It characterizes the impact of the corresponding feature on the prediction. The rate of change can be simply estimated by computing the variance of $g_i(x_i)$.

The Cox model, even with the functions implemented by the neural network, requires the calculation of the baseline CHF or SF that are independent of features. A more powerful survival model is the Beran estimator. Therefore, a method called SurvBeX (Survival Beran eXplanation) [20] was developed that uses the Beran estimator instead of the Cox model. The main idea of the method is that in the Beran estimator the weight is defined as follows:

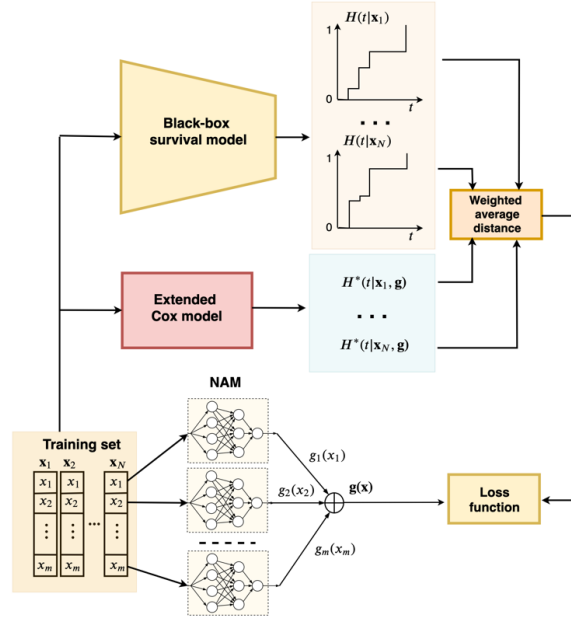


Fig. 2. SurvNAM structure [21]

$$W(\mathbf{x}, \mathbf{x}_i, \mathbf{b}) = \text{softmax}\left(\left\| \mathbf{b} \odot (\mathbf{x} - \mathbf{x}_i) \right\|^2\right),$$

where vector \mathbf{b} characterizes the influence of features on predictions, \odot is the scalar product.

In general, SurvBeX uses the algorithm of the SurvLIME method with the generation of synthetic instances near the object being explained, but with the difference that the Beran estimator is used instead of the Cox model. In this case, we obtain a more complex optimization problem. However, numerous numerical experiments show that it can be solved quite simply by existing methods and provide significantly better interpretation results.

SurvBeX is a flexible method, which can be modified in several ways. First, we can change the kernel function. For example, if we replace the Gaussian kernel, used in the method, with the triangle kernel, then only a local area of instances around \mathbf{x} will be included into consideration. Second, the set of training parameters can also be changed. If the number of training instances is rather large, then additional parameters may lead to a more accurate approximation of the black-box model prediction function.

So far, the explanation methods based on survival modifications of LIME and NAM have been considered. However, there is an explanation method, which extends SHAP to a case of survival analysis. The method is called SurvSHAP(t) [21], and it is based on SHAP with solid theoretical foundations and a broad adoption among machine learning practitioners. According to SurvSHAP(t), the Shapley values (contributions of the d -th feature) $\phi_d(\mathbf{x}, t)$ are assigned for the explained instance \mathbf{x} at a time moment t . Since the predicted SF of the black-box model is a step-wise function with changes at times $t_1 \leq t_1 \leq \dots \leq t_n$ due to the finite number of observations, then values $\phi_d(\mathbf{x}, t)$ are determined at all times, i.e., $\phi_d(\mathbf{x}, t_1), \dots, \phi_d(\mathbf{x}, t_n)$. Values $\phi_d(\mathbf{x}, t)$ are computed for every feature d and every time t_i in the standard way by taking points of the predicted SF $S(t|\mathbf{x})$ at the point t_i as a point-valued prediction of the black-box model for the explained instance \mathbf{x} . The obtained time-dependent values $\phi_d(\mathbf{x}, t)$ are aggregated to calculate the overall importance $\psi_d(\mathbf{x})$ of the d -th feature as follows:

$$\psi_d(\mathbf{x}) = \int_0^{t_n} |\phi_d(\mathbf{x}, t)| dt.$$

In contrast to SurvLIME, SurvNAM, and SurvBeX, where the meta-model is based on an assumption of the approximating model (the Cox model, the Beran estimator), an important advantage of SurvSHAP(t) is that it does not require to make any assumption about the meta-model. On the other hand, SurvSHAP(t) as a SHAP-based method inherits all problems of using the SHAP method, which include the problem of representing subsets of features as inputs and the problem of the complexity of computing the Shapley values.

Conclusion

We have briefly considered only the main explanation methods used in recent survival analysis. Due to the importance of survival analysis in many applications, including medicine, system reliability, safety, and predictive analytics, new survival explanation models are being developed and will be proposed in the near future. There are several areas of survival analysis, for which explanation methods have not been developed. For example, there are no effective explanation methods for competing risks. The development of such methods is an active area for further research. The same can be said for specific applications of survival analysis. For example, the development of explanation methods for models dealing with multi-modal data is another area for further research. New explanation methods that improve existing methods (SurvLIME, SurvNAM, SurvSHAP(t), etc.) can also be considered as areas for further research.

REFERENCES

1. **Adadi A., Berrada M.** Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 2018, Vol. 6, pp. 52138–52160. DOI: 10.1109/ACCESS.2018.2870052
2. **Burkart N., Huber M.F.** A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 2021, Vol. 70, pp. 245–317. DOI: 10.48550/arXiv.2011.07876
3. **Carvalho D.V., Pereira E.M., Cardoso J.S.** Machine learning interpretability: A survey on methods and metrics. *Electronics*, 2019, Vol. 8, no. 8, article no. 832. DOI: 10.3390/electronics8080832
4. **Guidotti R.** Evaluating local explanation methods on ground truth. *Artificial Intelligence*, 2021, Vol. 291, article no. 103428. DOI: 10.1016/j.artint.2020.103428
5. **Ribeiro M.T., Singh S., Guestrin C.** “Why should I trust you?”: Explaining the predictions of any classifier. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144. DOI: 10.1145/2939672.2939778
6. **Lundberg S.M., Lee S.-I.** A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 2017, pp. 4765–4774. DOI: 10.48550/arXiv.1705.07874
7. **Štrumbelj E., Kononenko I.** An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research*, 2010, Vol. 11, pp. 1–18. DOI: 10.5555/1756006.1756007
8. **Hastie T.J., Tibshirani R.J.** *Generalized additive models* (Chapman & Hall/CRC Monographs on Statistics and Applied Probability), 1st ed. London: Chapman & Hall/CRC press, 1990. 352 p.
9. **Agarwal R., Melnick L., Frosst N., Zhang X., Lengerich B., Caruana R., Hinton G.** Neural additive models: Interpretable machine learning with neural nets. 35th Conference on Neural Information Processing Systems (NeurIPS 2021), 2021, Vol. 34, pp. 4699–4711. DOI: 10.48550/arXiv.2004.13912
10. **Konstantinov A.V., Utkin L.V.** Interpretable machine learning with an ensemble of gradient boosting machines. *Knowledge-Based Systems*. 2021, Vol. 222, article no. 106993. DOI: 10.1016/j.knosys.2021.106993
11. **Wang P., Li Y., Reddy C.K.** Machine learning for survival analysis: A survey. *ACM Computing Surveys*. 2019, Vol. 51, pp. 1–36. DOI: 10.48550/arXiv.1708.04649
12. **Cox D.R.** Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1972, Vol. 34, no. 2, pp. 187–202. DOI: 10.1111/j.2517-6161.1972.tb00899.x
13. **Beran R.** *Nonparametric regression with randomly censored survival data*, 1981.

14. **Covert I., Lee S.-I.** Improving KernelSHAP: Practical Shapley Value Estimation via Linear Regression. International Conference on Artificial Intelligence and Statistics, 2021, pp. 3457–3465. DOI: 10.48550/arXiv.2012.01536
15. **Kovalev M.S., Utkin L.V., Kasimov E.M.** SurvLIME: A method for explaining machine learning survival models. Knowledge-Based Systems. 2020, Vol. 203, article no. 106164. DOI: 10.1016/j.knosys.2020.106164
16. **Utkin L.V., Kovalev M.S., Kasimov E.M.** An explanation method for black-box machine learning survival models using the Chebyshev distance. Artificial Intelligence and Natural Language (AINL 2020). 2020, Vol. 1292, pp. 62–74. DOI: 10.1007/978-3-030-59082-6_5
17. **Kovalev M.S., Utkin L.V.** A robust algorithm for explaining unreliable machine learning survival models using the Kolmogorov–Smirnov bounds. Neural Networks. 2020, Vol. 132, pp. 1–18. DOI: 10.1016/j.neunet.2020.08.007
18. **Utkin L.V., Zaborovsky V.S., Kovalev M.S., Konstantinov A.V., Politayeva N.A., Lukashin A.A.** Uncertainty interpretation of the machine learning survival model predictions. IEEE Access. 2021, Vol. 9, pp. 120158–120175. DOI: 10.1109/ACCESS.2021.3108341
19. **Utkin L.V., Satyukov E.D., Konstantinov A.V.** SurvNAM: The machine learning survival model explanation. Neural Networks. 2022, Vol. 147, pp. 81–102. DOI: 10.1016/j.neunet.2021.12.015
20. **Utkin L.V., Eremenko D.Y., Konstantinov A.V.** SurvBeX: An explanation method of the machine learning survival models based on the Beran estimator. 2023. DOI: 10.48550/arXiv.2308.03730
21. **Krzyżniński M., Spytek M., Baniecki H., Biecek P.** SurvSHAP(t): Time-dependent explanations of machine learning survival models. Knowledge-Based Systems, 2023, Vol. 262, article no. 110234. DOI: 10.1016/j.knosys.2022.110234

INFORMATION ABOUT AUTHORS / СВЕДЕНИЯ ОБ АВТОРАХ

Utkin Lev V.

Уткин Лев Владимирович

E-mail: lev.utkin@gmail.com

ORCID: <https://orcid.org/0000-0002-5637-1420>

Konstantinov Andrei V.

Константинов Андрей Владимирович

E-mail: andrue.konst@gmail.com

ORCID: <https://orcid.org/0000-0002-1542-6480>

Eremenko Danila Y.

Еременко Данила Юрьевич

E-mail: eremenko.dyu@edu.spbstu.ru

ORCID: <https://orcid.org/0000-0002-1115-7543>

Zaborovsky Vladimir S.

Заборовский Владимир Сергеевич

E-mail: vlad2tu@yandex.ru

Muliukha Vladimir A.

Мулюха Владимир Александрович

E-mail: vladimir.muliukha@spbstu.ru

ORCID: <https://orcid.org/0000-0002-3583-7324>

Submitted: 10.06.2024; Approved: 26.08.2024; Accepted: 18.09.2024.

Поступила: 10.06.2024; Одобрена: 26.08.2024; Принята: 18.09.2024.