Ann. Data. Sci.
DOI 10.1007/s40745-017-0105-4

Classification in Non-linear Survival Models Using Cox Regression and Decision Tree

Reza Mokarram[1] · **Mehdi Emadi**[1]

Received: 27 December 2016 / Revised: 18 February 2017 / Accepted: 28 February 2017

Abstract Classification is the most important issues that have gained much attention in various fields such as health and medicine. Especially in survival models, classification represents a main objective and it is also one of the main purposes in data mining. Among data mining methods used for classification, implementation of the decision tree due to its simplicity and understandable and accurate results, has gained much attention and popularity. In this paper, first we generate the observations by using Monte-Carlo simulation from hazard model with the three degrees of complexity in different levels of censorship 0 to 70%. Then the accuracy of classification in the Cox and the decision tree models is compared for the number of samples 1000, 5000 and 10,000 by area under the ROC curve(AUC) and the ROC-test.

Keywords Proportional hazard · Cox model · Decision tree · Classification · Non-linear survival model · Non-Parametric model

1 Introduction

A main topic of scientific research in the field of data science is classification [4]. Especially in medical and health area, classification of survival continuous data or failure times is very important. Two major characteristics of survival continuous data are censoring and violation of normal assumption for ordinary least squares regressions [7]. These two characteristics of time variables explain why straightforward logistic and multiple regression techniques cannot be used.

✉ Mehdi Emadi
Emadi@math.um.ac.ir

[1] Department of Statistics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, Mashhad, Iran

Published online: 12 July 2017

Different parametric and semiparametric models for survival data, such as proportional hazards and accelerated life time models, are based on the assumptions which may be untenable in some situations. Hazard functions may not be proportional and determining the baseline distribution for the survival time may be hard to justify. When classification is required, nonparametric models such as decision tree are proposed that have fewer presuppositions and produce good results. The use of decision trees is one of the most promising and popular approaches for classification and it is a very common technique in data mining [2,3]. According to many researchers, decision trees are popular due to their simplicity and transparency. Decision trees are self-explanatory and there is no need to be expert in order to follow a certain decision tree. Moreover decision trees are widely used in areas such as text mining, information extraction, machine learning, and pattern recognition [2].
In this paper, simulated data by Monte-Carlo method, covariates and survival times, from three models of hazard with different rates of censoring were used to predict the outcome using decision trees and Cox's regression models. Then the predictions of the models were compared with the area under Receive Operating Characteristics curve (AUC) and an appropriate statistical test for AUC.

2 Methods

2.1 Decision Trees (Non-Parametric Model)

In the science of machine learning, much attention is given to the research based on the classification.The goal of classification is to choose the most appropriate class of default classes available to a case. The main goal of the these researches is to develop the methods which minimize an error rate.
Considerable progress has been achieved toward this goal in recent decades. Several methods to allocate a class are now available. Particularly, the use of decision tree is one of the most favorable and popular methods to classify tree [3]. A decision tree determines the classification and explains the reasons for selecting a certain case. This form of the classification approach is of particular importance in areas such as medical and health. The ability to present a simple and understandable branch structure is another reason to use a decision tree. This capability has shown the decision tree to be a more acceptable method compared with the other classifier such as neural network and support vector machine [2,5].
Simplicity and comprehensibility of a decision tree are reduced as nodes and tree depth increase and in this case, it is not profitable to use structural graphic. A decision tree is a classifier which acts on a given space as recursive discriminant. It includes a node in root without any input, and other nodes in the branches with only one entrance. A node with at least one output as the internal node and other nodes without the output are considered as leaves or external nodes. In decision tree, the sample space is divided into at least two subspace by each internal node. The nodes are selected from the entries based on a specific property. For non-discrete characteristics, the parts refers to pre-specified ranges. Each leaf represents the class that has the highest relation with the target value. It may also show a probability vector representing the probability of

belonging to the target feature. Training examples are gradually classified from root to the leaves according to corresponding tests. The process of forming a decision tree can be briefly stated as follows [5]:

1. In the event all the cases in S are tagged with identical class, giving back a leaf labeled with this class.
2. Select some tests (not necessarily statistical one) such as T with respect to some criteria which include two or more mutually exclusive results $\{P_1, P_2, \ldots, P_r\}$.
3. Split S into disjoint subsets $S_1, S_2, \ldots, S_r$ so that S_i includes those cases getting outcome P_i with the test T, for $i = 1, 2, \ldots, r$.
4. Repeat this tree-construction process recursively in every subsets $S_1, S_2, \ldots, S_r$ and suppose the decision trees that are given back by these recursive process named $S_{ij}, i = 1, 2, \ldots, r, j = 1, 2, \ldots, r_i$.
5. A decision tree is returned with a node tagged S_0 as the root and the trees $S_{ij}, i = 1, 2, \ldots, r, j = 1, 2, \ldots, r_i$ with subtrees under of each node.

2.1.1 Growing Decision Tree

Ordinary supervised learning involves a given training set aiming at formation of an explanation which would be useful for predicting past unobserved samples. Various definitions of the training set have been proposed. Often, it is defined as a bag instance of a specific schematic design. Bag instance contains a set of records or instances probably containing duplicates. Every individual tuple can be described with the aid of the values pertaining to its vector of attribute. The attributes and their domains are described by the bag schema.
Often, the assumption is that, based on a random pattern, the training set tuples are independently created in accordance with a number of unchanging and unidentified shared probability distribution. For example, assuming a training set of $S = \{\langle X_1, y_1\rangle, \langle X_2, y_2\rangle, \ldots, \langle X_n, y_n\rangle\}$ with input attributes set of $A = \{a_1, a_2, \ldots, a_p\}$ in relation with covariate X and y nominal target attribute from an F fixed distribution across the labeled instance space, the objective can be an optimum classifier that generates minimum generalization error. As for the generalization error, it means the misclassification rate in concerning the distribution F. Considering the nominal attributes, the generalization error can be obtained by the following phrase [2]:

$$\epsilon\left(DT(S), F\right) = \sum F\left(x, y\right) * L\left\langle y, DT\left(S\right)\left(x\right)\right\rangle$$

where $L\left\langle y, DT\left(S\right)\left(x\right)\right\rangle$ is the zero-one loss function which can be defined as below clause:

$$L\left\langle y, DT\left(S\right)\left(x\right)\right\rangle = \begin{cases} 0 & if\ y = DT\left(S\right)\left(x\right) \\ 1 & if\ y \neq DT\left(S\right)\left(x\right) \end{cases}$$

Regarding to the sum operator over a training set, a decision tree inducer is represented by the notation DT and the $DT(S)$ is indicative of a classification tree, induced by operationalizing DT on the training set of S. If the attributes are numeric, then the sum operator is replaced by the integration operator.

2.1.2 Decision Trees and Estimation of Probability

The inducer product, i.e., the classifier is useful for classifying the unobserved tuples through either expressly allocating it to a determined class or preparing a probability vector indicating the conditional probability of an assumed instance to be classified in any of the individual classes (probabilistic classifier). The inducers capable of constructing probabilistic classifiers are called "the probabilistic inducers". As an outstanding feature, estimating the conditional probability of $\hat{P}_{DT(S)}\left(y = c_j | a_k = x_{k,q}, k = 1, 2 \ldots, p\right)$ of a given observation $\left\langle X_q, y_q \right\rangle$ would be possible in decision trees. Approximating the probability in classification trees is performed individually for each leaf through class frequency calculation considering the training instances belonging to the desired leaf [2]. The frequency use entails the drawback of estimating the probability which could be problematic in cases, a class will never occur in a given leaf. Obviously, the probability will be zero in such cases.

2.1.3 Algorithmic Outline for Decision Trees

Most of the algorithms of recursive partitioning type can be regarded as the product of particular cases of a non-complex two-stage algorithm: In the first stage, the observations are partitioned under the effect of univariate splits in a recursive manner and in the second stage a constant model is fitted in individual cells of the resulting partition. As the most prevalent applications of such algorithms, CART (Breiman, Friedman, Olshen and Stone, 1984) and C4.5 (Quinlan 1993) carry out an elaborative search concerning all the probable splits; there by a maximization of an information measure of node impurity is performed and the covariate indicating the best split is selected [3]. Still, the approach suffers from two major problems, i.e., overfitting and selection bias in relation with covariates involving too many probable splits.
A recent approach for developing trees is based on the conditional inferences (Hothorn, Hornik and Zeileis 2006). In the present paper the authors try to implement an integrated framework embedding recursive binary partitioning feature in the well-adjusted permutation tests theory hypothesized by Strasser and Weber (1999). The statistics that measure the relationship between the responses and covariates, when conditionally distributed, function as the basis for an unbiased selection among covariates that are measured against different scales. Furthermore, several testing processes need to be employed to ensure that there would be no major relationship between the individual covariates, and the response can be described and the recursion must be halted. Using non-negative integer valued case weights $W = (w_1, \ldots, w_n)$ can be used to formulate a generic algorithm for recursive binary partitioning for any given learning sample. Each tree node is shown and represented by a vector consisting of case weights having one and zero elements, in case the relevant observations are the node elements and the otherwise respectively. In the following, the procedure recursive binary partitioning is given [1]:

1. Regarding to the case weights W, one should test the global null hypothesis of independence among the p covariate and the response. We should halt in case the

above hypothesis is acceptable. Otherwise, we can select the covariate Xj with strongest association to Y .

2. Select a $\Theta \subset \chi_p$ to split χ_p into two disjoint sets Θ and $\chi_p \backslash \Theta$ in which the p-dimensional covariate vector $X = \left(X_1, \ldots, X_p\right)$ is obtained from a sample space $\chi = \chi_1 \times \ldots \times \chi_p$. The case weights W_l and W_r specify two subgroups with elements $w_{l,i} = I\left(X_{ji} \in \Theta\right)$ and $w_{r,i} = I\left(X_{ji} \notin \Theta\right)$ where $i = 1, \ldots, n$, $j = 1, \ldots, p$ (I(.) denotes the indicator function).

Afterwards you should repeat steps 1 and 2 recursively with modified case weights W_l and W_r respectively. Step 1 of the generic algorithm is regarded as an independence problem because we need to decide if there is any information on the response variable which might be covered by any covariate [1,2].

The segregation of selection of variables and implementing steps 1 and 2 of the procedure is essential for the construct of the tree structure which is interpretable, meanwhile avoiding systemic trend toward covariates with many probable splits or missing values. Moreover, a stopping criterion which is statistically motivated and intuitive can be employed: stopping at the time when the global null hypothesis of independence among the response and the p covariates is impossible to be rejected at a pre-specified nominal level α. The algorithm induces the partition $B_1, B_2, \ldots, B_r$ of the covariate space X, in which an individual cell $B \in \{B_1, B_2 \ldots, B_r\}$ is accompanied with a vector of case weights. In the nodes characterized by case weight W, the overall hypothesis of independence can be formulated by considering the p partial hypothesis $\mathrm{H}_0^j{:}F\left(Y|X_j\right) = F\left(Y\right), j = 1, \ldots, p$ and the global null hypothesis of $\mathrm{H}_0 = \bigcap_{j=1}^{p} \mathrm{H}_0^j$. Where the rejection of H_0 at the pre-determined level α is impossible, we should stop the recursion. If we can reject the global hypothesis, we should measure the relationship (association) between Y and the covariates $X_j, \quad j = 1, \ldots, p$, through p-values that indicate the deviation from partial hypotheses H_0^j [1]. In this paper, for the implementation of this process in the R (version 3.1.3)software, we use "ctree" command from "party" package.

2.2 Cox Regression Model (Semi-Parametric Model)

The survival time T associated with an event can be a continuous, non-negative random variable having survival function:

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

The related hazard function $h(t)$, demonstrates the probability density of an event occurring around time t, considering that it doesn't occur prior to time t.

The hazard function is

$$\begin{aligned} h(t) &= \lim_{\triangle t \to 0^+} \frac{P(t < T < t + \triangle t \mid T > t)}{\triangle t} \\ &= \frac{f(t)}{S(t)} \end{aligned}$$

where, $f(t)$ is survival density function for T. The cox regression, generally known as the proportional hazards regression (PH), is demonstrated by:

$$h_x(t, \boldsymbol{\beta}) = h_0(t) exp(\boldsymbol{\beta}' X) \tag{1}$$

where $\boldsymbol{\beta}$ (p-dimensional parameter) is the regression coefficient vector, $h_0(t)$ is the baseline hazard which is related to time and an unspecified function (it is the property that makes the Cox model a semiparametric model). Here $X = \left(x_1, \ldots, x_p\right)$ is a vector of x_i elements that are predictors or covariates [6].

$$\begin{aligned} \frac{h_{x_i}(t, \boldsymbol{\beta})}{h_{x_j}(t, \boldsymbol{\beta})} &= \frac{h_0(t)\, exp\left(\boldsymbol{\beta}' x_i\right)}{h_0(t)\, exp\left(\boldsymbol{\beta}' x_j\right)} \\ &= exp\left\{\boldsymbol{\beta}'(x_i - x_j)\right\} \end{aligned}$$

Which is independent of the survival time. The Likelihood function will be given by

$$\begin{aligned} L(\boldsymbol{\beta}; \boldsymbol{x}) &= \prod_{i=1}^{n} h_{x_i}(t_i, \boldsymbol{\beta})^{\delta_i} S_{x_i}(t_i, \boldsymbol{\beta}) \\ &= \prod_{i=1}^{n} (h_0(t_i) exp(\boldsymbol{\beta}' x_i))^{\delta_i} S_0(t_i)^{exp(\boldsymbol{\beta}' x_i)} \end{aligned}$$

where δ_i is the binary variable, $\delta_i = 1$ if a case is observed and $\delta_i = 0$ if censored. Considering that the baseline distribution is usually unknown, this likelihood function cannot be used to obtain estimates for $\boldsymbol{\beta}$ [7]. Cox (1972) suggested to maximize the so-named partial likelihood function. In order to introduce this concept we will denote the observed ordered event times by $t_{(i)}, \quad i = 1, 2, \ldots, d$

$$t_{(1)} < t_{(2)} < \ldots < t_{(d)}, \qquad d = \sum_{i=1}^{n} \delta_i.$$

The covariate linked to the case observed to failure time in t_i can be through by $x_{(i)}$. The partial likelihood proposed by Cox, is given by [6]

$$\prod_{i=1}^{d} \frac{exp(\boldsymbol{\beta}' x_{(i)})}{\sum_{j \in R_i} exp(\boldsymbol{\beta}' x_j)}.$$

By maximizing this partial likelihood through utilizing the Newton-Raphson technique, we could match the Cox regression and estimate parameter $\boldsymbol{\beta}$.
In this paper, for the implementation of this process in the R software, we use "coxph" command from "survival" package [9].

Table 1 Results of simulation data

ID	Age	Sex	Social welfare(SW)	Postoperative Care-1(PC1)	Postoperative Care-2(PC2)
1	49	Male	Mw	22	27
2	36	Female	Lw	18	23
3	55	Female	Lw	28	36
4	57	Male	Uw	21	31
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
1000	45	Female	Mw	29	45

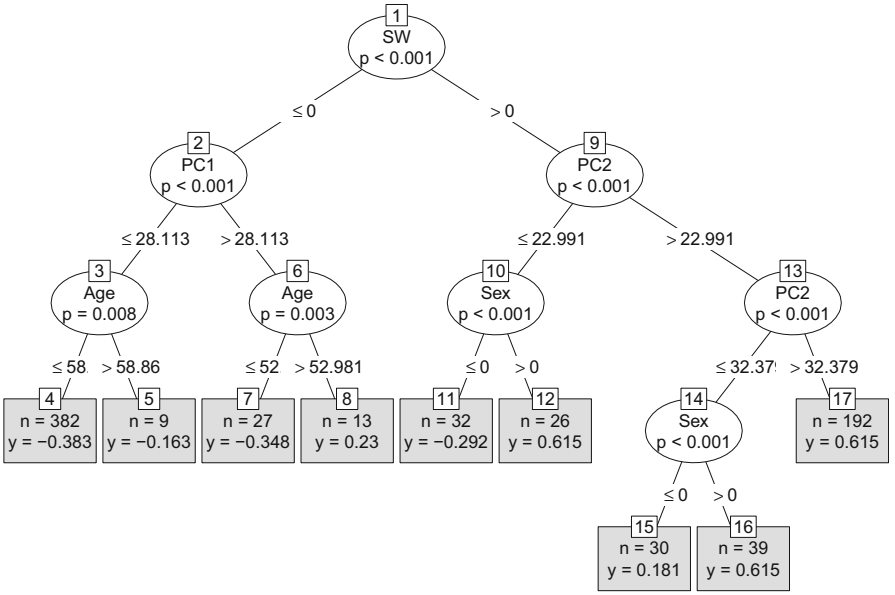


Fig. 1 Tree-structured survival model for the simulated data based on conditional inference

3 Simulation

Consider the training set in Table 1 containing Monte-Carlo simulated data about patients. Each patient is characterized by six attributes: Age, Sex, Social welfare, Censor, Postoperative Care-1 and Postoperative Care-2 describe time care after surgery for patient. The goal is to induce a classifier with the highest accuracy in predicting health recovery. In order to examine the precision of classification in the models DT and Cox, a collection of N = 1000 samples in the medical and health including variables age from the normal distribution and sex and social welfare class from the binomial

Table 2 Results of simulation for exponential hazard model with N = 1000

Hazard	Censored rate (%)	COX-AUC	COX-AUC CI	TREE-AUC	TREE-AUC CI	COX-TREE P-value
Model1	0	0.915	0.031	0.966	0.032	0.028
	10	0.916	0.032	0.798	0.036	0.009
	20	0.914	0.034	0.687	0.041	0.007
	30	0.912	0.037	0.762	0.038	0.013
	40	0.904	0.038	0.760	0.039	0.002
	50	0.903	0.028	0.766	0.030	0.012
	60	0.905	0.030	0.637	0.036	0.004
	70	0.905	0.033	0.473	0.048	0.000
Model2	0	0.895	0.038	0.986	0.032	0.008
	10	0.896	0.035	0.981	0.036	0.009
	20	0.894	0.043	0.947	0.041	0.004
	30	0.882	0.037	0.962	0.038	0.014
	40	0.874	0.048	0.960	0.039	0.012
	50	0.863	0.028	0.866	0.030	0.032
	60	0.850	0.030	0.877	0.036	0.000
	70	0.835	0.043	0.838	0.048	0.000
Model3	0	0.769	0.058	0.969	0.002	0.008
	10	0.766	0.053	0.961	0.003	0.003
	20	0.754	0.043	0.967	0.009	0.000
	30	0.742	0.038	0.962	0.008	0.000
	40	0.755	0.048	0.967	0.004	0.000
	50	0.763	0.033	0.866	0.010	0.000
	60	0.750	0.040	0.857	0.016	0.000
	70	0.732	0.045	0.834	0.018	0.000

distribution and the period of care after operation in two stages hospitalization and outside hospital from the uniform distribution based on Monte-Carlo were simulated.

Then three different models for the hazard ratio (HR) in different complexity levels including a simple linear model, a model with double interaction and a model including at least double interaction were assumed as follow:

$$\text{Model1: } \exp\{.25 * \text{Sex} + .5 * \text{SW} - .5 * \text{Age} + .5 * \text{PC1} + .25 * \text{PC2}\}$$

$$\text{Model2: } \exp\{.25 * \text{Sex} + .5 * \text{SW} - .5 * \text{Age} + .5 * \text{PC1} + .25 * \text{PC2} + .5 * (\text{Sex} * \text{PC1} + \text{PC2} * \text{SW2} - \text{PC1} * \text{Age} + \text{PC2} * \text{Age})\}$$

$$\text{Model3: } \exp\{.25 * \text{Sex} + .5 * \text{SW} - .5 * \text{Age} + .5 * \text{PC1} + .25 * \text{PC2} + .5 * (\text{Sex} * \text{PC1} + \text{PC2} * \text{SW2} - \text{Sex} * \text{PC1} * \text{Age} + \text{SW} * \text{PC2} * \text{Age}) + .25 * \text{PC1} * \text{PC2} * \text{Sex} * \text{SW} * \text{Age}\}$$

In the next step survival times from the exponential distribution of these models were simulated [8]. Next, the survival times greater than kth percentile were taken as censored observation in k percent level. The averages of the censor rates in eight levels

Table 3 Results of simulation for exponential hazard model with N = 5000

Hazard	Censored rate (%)	COX-AUC	COX-AUC CI	TREE-AUC	TREE-AUC CI	COX-TREE P-value
Model1	0	0.911	0.011	0.976	0.032	0.028
	10	0.912	0.012	0.798	0.036	0.019
	20	0.914	0.014	0.757	0.041	0.017
	30	0.915	0.017	0.742	0.038	0.013
	40	0.914	0.018	0.760	0.039	0.012
	50	0.913	0.017	0.766	0.030	0.014
	60	0.915	0.010	0.622	0.036	0.000
	70	0.911	0.013	0.485	0.048	0.000
Model2	0	0.905	0.018	0.996	0.013	0.008
	10	0.906	0.015	0.991	0.016	0.002
	20	0.891	0.014	0.994	0.011	0.004
	30	0.882	0.011	0.972	0.013	0.003
	40	0.872	0.014	0.960	0.019	0.002
	50	0.863	0.018	0.963	0.013	0.008
	60	0.850	0.013	0.887	0.011	0.010
	70	0.825	0.019	0.838	0.018	0.035
Model3	0	0.759	0.028	0.996	0.006	0.000
	10	0.766	0.023	0.991	0.004	0.000
	20	0.744	0.026	0.994	0.007	0.000
	30	0.742	0.028	0.968	0.004	0.000
	40	0.755	0.035	0.977	0.008	0.000
	50	0.743	0.023	0.956	0.010	0.000
	60	0.730	0.024	0.897	0.013	0.000
	70	0.722	0.045	0.864	0.011	0.000

from 0 to 70% were considered. In order to learn the models 75% of the samples was used and the remaining was put aside for the model test. Survival times were divided into two groups based on the median. The Cox and The DT classifiers were examined after learning by criteria Area Under Curve (AUC) and the ROC-test [10,11]. Figure 1 shows the structure of the conditional decision tree for simulated data. Whole the above procedure was repeated 200 times and the average of criteria AUC, confidence interval (CI) for AUC and P-value of ROC-test were considered. When there are many samples to examine the results, were considered as N2 = 5000 and N3 = 10, 000, and all the procedure was repeated. The results are presented in Tables 2, 3 and 4.

4 Results and Conclusion

According to this simulation study, when data are censored, for the hazard linear function, the Cox regression model works better than nonparametric model DT especially

Table 4 Results of simulation for exponential hazard model with N = 10, 000

Hazard	Censored rate (%)	COX-AUC	COX-AUC CI	TREE-AUC	TREE-AUC CI	COX-TREE P-value
Model1	0	0.915	0.010	0.978	0.012	0.007
	10	0.916	0.015	0.798	0.016	0.009
	20	0.914	0.014	0.737	0.016	0.010
	30	0.913	0.018	0.602	0.018	0.003
	40	0.911	0.018	0.764	0.019	0.002
	50	0.916	0.015	0.756	0.016	0.014
	60	0.910	0.014	0.622	0.015	0.000
	70	0.911	0.013	0.435	0.018	0.000
Model2	0	0.911	0.013	0.997	0.005	0.000
	10	0.908	0.012	0.991	0.006	0.000
	20	0.891	0.010	0.994	0.011	0.000
	30	0.886	0.011	0.973	0.003	0.003
	40	0.872	0.019	0.966	0.009	0.002
	50	0.866	0.018	0.964	0.011	0.001
	60	0.850	0.016	0.885	0.018	0.012
	70	0.835	0.013	0.846	0.022	0.024
Model3	0	0.765	0.018	0.998	0.011	0.000
	10	0.756	0.013	0.995	0.014	0.000
	20	0.759	0.016	0.991	0.007	0.002
	30	0.752	0.018	0.988	0.004	0.004
	40	0.753	0.015	0.973	0.003	0.000
	50	0.749	0.013	0.916	0.005	0.000
	60	0.730	0.014	0.897	0.002	0.000
	70	0.716	0.015	0.872	0.014	0.000

in the censor level more than 50%. However when data are not censored, it is observed that the DT model works better than the COX model.
An examination of AUC criterion in the non-linear hazard model shows that a higher precision of the DT model results compared with the Cox model in all censorship levels. Change of censor levels has no remarkable effect on AUC criterion in the Cox model. In addition, it is noted that in the non-linear hazard, with an increase complexity in hazard model, AUC criterion in the DT is almost constant and significantly more than the amount for the Cox model. The significance of these differences are confirmed by the ROC test. These results to be established for all number of samples and they can be summarized in terms of HR as follows:

1. If HR is linear with respect to covariates, when the data were censored, Cox regression model performs better than DT non-parametric model, especially in the censor levels more than 50% the different performance of the models is remarkable. If the data are not censorship, the DT model is outperformed the Cox model.

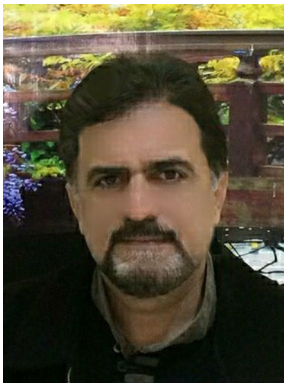
2. If the HR is non-linear with respect to variables and includes double interactions, both models perform efficiently but the DT model offers better results compared to the Cox model. If the data are not censorship the DT model is outperformed the Cox model.
3. In the case HR quantity is non-linear with respect to variables include minimum double interactions, the DT model significantly better results than the Cox model in all levels of censorship.

It is noted that increase in the censor levels, especially for the censor levels over 50%, decreases the performance of the DT model. But this increase does not have a remarkable effect on the Cox model. In addition, with the increase in complexity of the HR form, using the DT non-parametric model is suggested as the preferred approach, but in the case having simple linear form, application of the Cox semiparametric model is recommended. The increase in the number of samples in all the three models of HR, has no significant effect on the both classifier Cox and DT.

Acknowledgements The authors feel much obligated for the time spent to review this paper by highly informed the referees.

References

1. Hothron T, Hornik K, Zeileis A, ctree: Conditional Inference Trees. https://cran.r-project.org/web/packages
2. Rokach L, Maimon O (2007) Data mining with decision trees, series in machine perception and artificial intelligence, vol. 69. World Scientific Publishing Co. Pte. Ltd, Singapore
3. Kotsiantis SB (2013) Decision trees: a recent overview. Artif Intell Rev 39(4):261–283
4. Batagelj V, Bock H-H, Ferligoj A, Ziberna A (2006) Data science and classification, ISBN-10 3–540-34415-2. Springer, Berlin
5. Almuallim H, Kaneda S, Akiba Y (2001) Development and applications of decision trees. In: Proceedings in the expert systems-the technology of knowledge management and decision making for the 21st century six-volume set, pp 53–77
6. Lee ET, Wang JW (2003) Statistical methods for survival data analysis, Wiley series in probability and statistics, 3rd edn. Wiley-Interscience, Hoboken
7. Kleinbaum DG, Klein M (2012) Survival analysis a self learning text, 3rd edn. Springer, New York
8. Bender R, Augustin T, Blettner M (2005) Generating survival times to simulate cox proportional hazards models, Statistics Med 24:1713–1723
9. Fox J, Weisberg S (2011) An R companion to applied regression, 2nd edn. Sage Publications
10. Robin X, Turk N, Hainard A, Tiberti N, Lisacek F, Sanchez J, Muller M (2011) pROC: an open-source package for R to analyze and compare ROC curves. BMC Bioinform 12:77
11. Fawcett T (2006) An introduction to ROC analysis. Pattern Recognit Lett 27:861–874



Reza Mokarram I am a Ph.D. student in statistics from Ferdowsi University of Mashhad and I teach in Islamic Azad University. My favorite topics and studies are in data mining field.

Mehdi Emadi I am master of university in university Ferdowsi of Mashhad. My degree is associate Professor. I work in university from 15 years ago.