

Využití strojového učení pro odhad křivek přežití

Bc. Tomáš Trdla <tomas.trdla@tul.cz>, Ing. Karel Paleček, Ph.D.

ABSTRACT

The thesis implements selected algorithms from artificial intelligence area for an estimation of survival curves. Each model is tested with usage of real data from database of transplantations on UHKT. Quality of all models is then compared with Cox Proportional Hazard Model by multiple different statistical methods. Results demonstrate different quality at investigated survival times and although AI methods achieve more accurate results the theoretical ensemble model is introduced as further possible solution and research.

ÚVOD

Cílem této práce je prozkoumání schopností některých vybraných algoritmů umělé inteligence v oblasti odhadu křivek přežití. Pro experimentální účely byly poskytnuty reálné vzorky dat z oblasti zdravotnictví, konkrétně z databáze transplantací na Ústavu hematologie a krevní transfuze v Praze.

Implementované modely jsou porovnávány s nejčastěji používanou metodou zvanou Coxův model proporcionálních rizik se snahou docílit kvalitnějších výsledků a zároveň zachovat jeho možnost zjištění vlivu vstupních proměnných na výsledek.

METODIKA

CoxPH model [4] je definován rovnicí 1, kde se částečně riziková funkce nachází v exponentu, b_i jsou jednotlivé tzv. regresní kovariáty z jejichž hodnoty lze zjistit i vliv vstupní proměnné (např. věk, váha, pohlaví, ...) na riziko, že se u subjektu projeví událost zájmu (např. úmrtí). Tyto kovariáty jsou lineární povahy a proto práce implementuje několik vybraných algoritmů, které odstraňují tento nedostatek a umožňují tak zmapování komplexních nelineárních vazeb v učicích datech.

Hlavní myšlenka spočívá v nahrazení této rizikové funkce neuronovou sítí [3] a pokročilými metodami stromových struktur, kdy v obou případech je třeba využít speciálních metod pro výpočet gradientní ztrátové funkce (tzv. loss function).

Mimo uvedené byly otestovány i metody Random Survival Forest [1] a AutoML (cílem bylo využít *state-of-art* algoritmy pro výběr a parametrizaci modelů).

Práce dále implementuje tzv. SHAP hodnoty [2], které doplňují možnost zjištění vlivu vstupních proměnných na výsledek analýzy přežití, jako tomu je u CoxPH modelu, i pro AI modely.

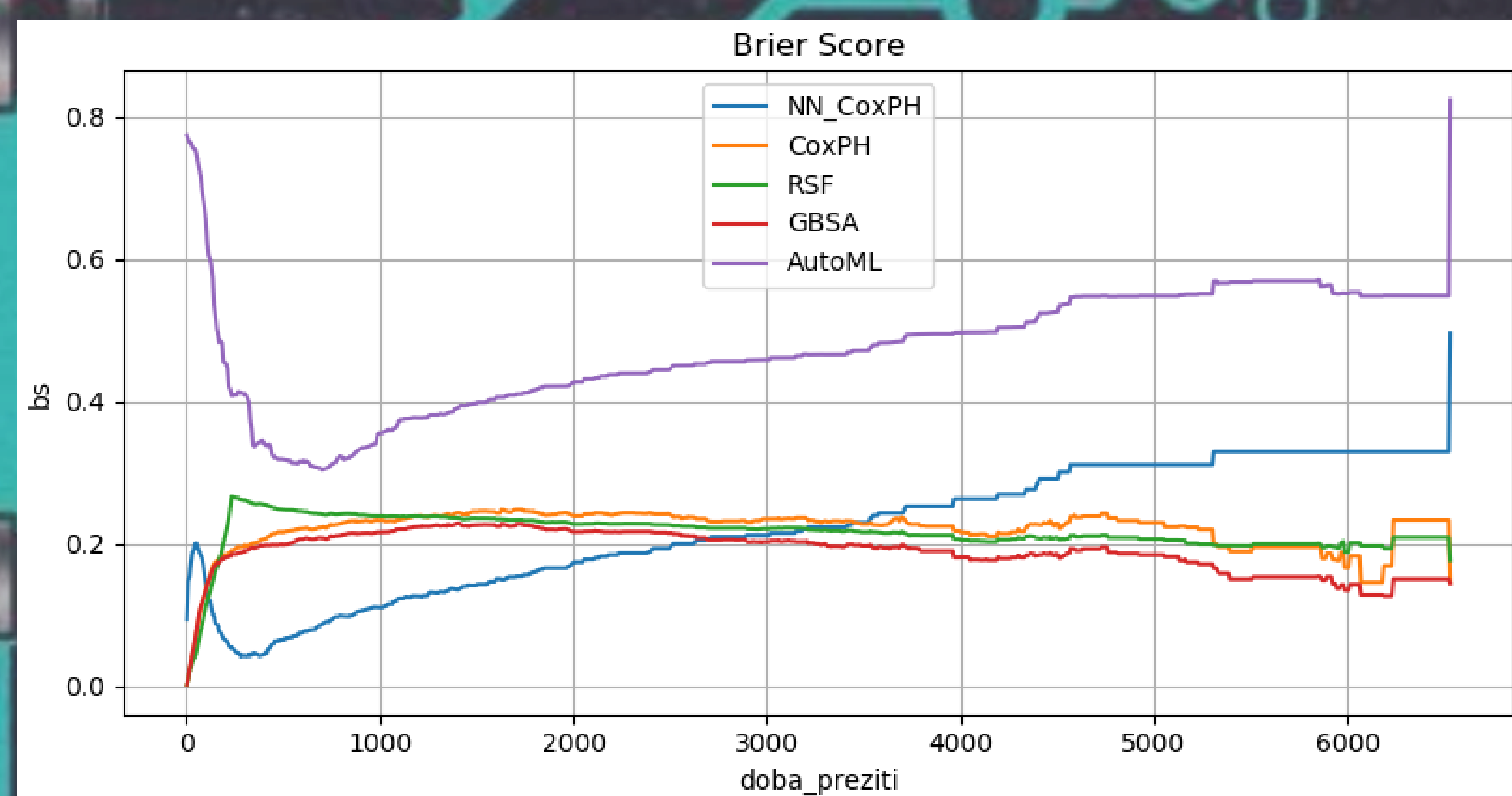
Modely jsou testovány s využitím křížové validace (K-Fold), aby byla zajištěna vyšší statistická významnost, než při testování pouze jedné náhodně vybrané sady dat z dostupného datasetu.

$$\lambda(t|x) = \lambda_0(t) \cdot \exp\left(\sum_{i=1}^n b_i x_i\right)$$

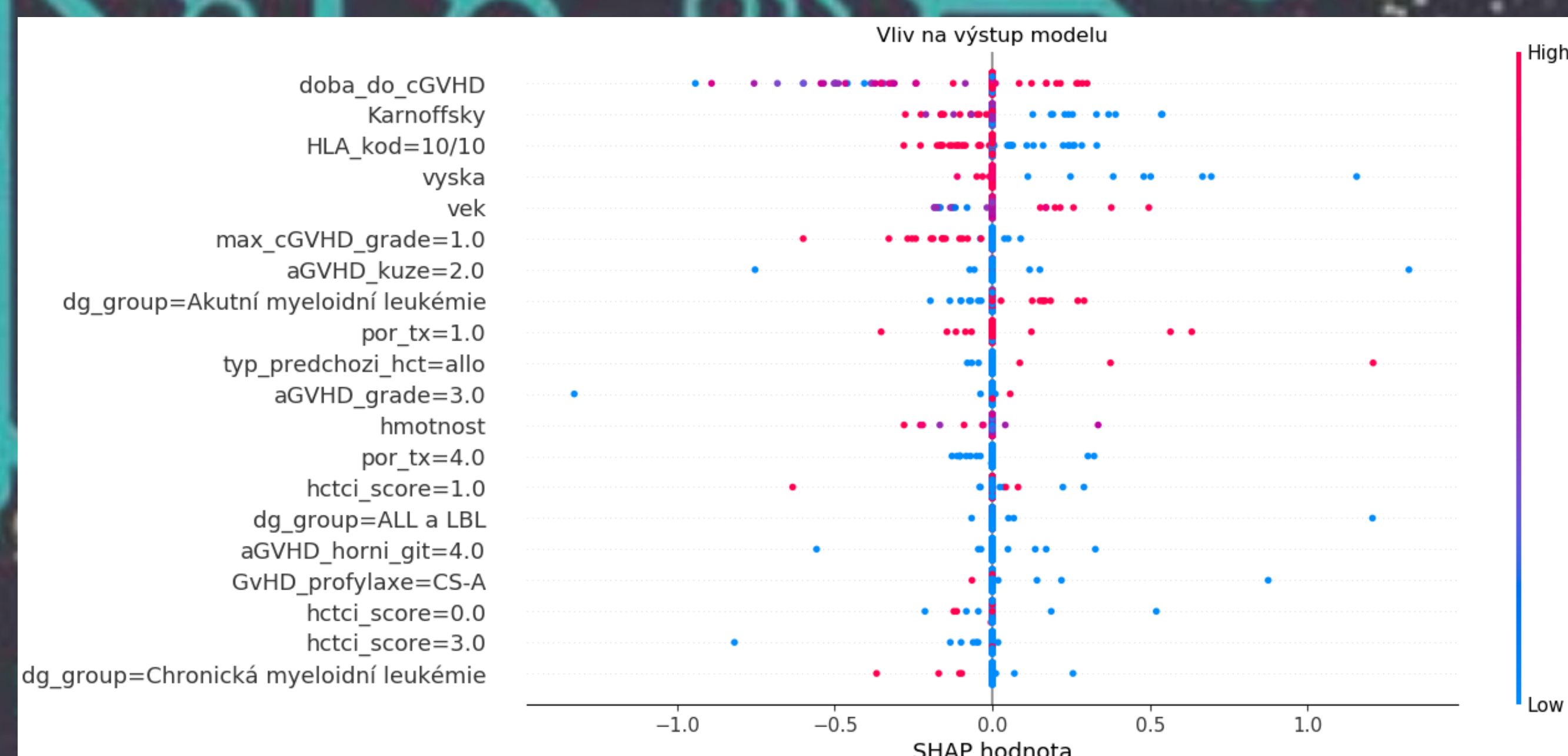
Rovnice 1 – Riziková funkce Coxova modelu proporcionálních rizik

VÝSLEDKY A ZÁVĚR

Porovnáním několika cílů (úmrtí pacienta, rozvinutí relapsu) včetně vlivu velikosti učicího datasetu se došlo k závěru, že nelze striktně říci, že jeden typ modelu je lepší než druhý, neboť pořadí pomyslného žebříčku je proměnné s učicími daty, cílem, ale i jejich velikostí. Avšak ve většině případů docílí AI modely lepších výsledků než standardní CoxPH. Proto byl diskutován návrh ensemble modelu. Ten však nebylo možné realizovat z důvodu omezeného rozsahu práce a může tedy být předmětem další studie v oblasti umělé inteligence a křivek přežití.



Obrázek 1 – Ukázka porovnání výsledků jedné z hodnotících metod



Obrázek 2 – Vliv vstupních proměnných GBSA modelu na riziko

REFERENCE

- [1] ISHWARAN, Hemant, Udaya B. KOGALUR, Eugene H. BLACKSTONE a Michael S. LAUER. Random survival forests. The Annals of Applied Statistics [online]. 2008, 2(3), 841-860 [cit. 2020-05-18]. DOI: 10.1214/08-AOAS169. ISSN 1932-6157. Dostupné z: <http://projecteuclid.org/euclid.aoas/1223908043>
- [2] LUNDBERG, Scott M. a Su-In LEE. A Unified Approach to Interpreting Model Predictions. Arxiv [online]. [cit. 2020-05-18]. Dostupné z: <https://arxiv.org/pdf/1705.07874.pdf>
- [3] KATZMAN, Jared L., Uri SHAHAM, Alexander CLONINGER, Jonathan BATES, Tingting JIANG a Yuval KLUGER. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Medical Research Methodology [online]. 2018, 18(1) [cit. 2020-05-18]. DOI: 10.1186/s12874-018-0482-1. ISSN 1471-2288. Dostupné z: <https://bmcmmedresmethodol.biomedcentral.com/articles/10.1186/s12874-018-0482-1>
- [4] PRENTICE, Ross L. Introduction to Cox (1972) Regression Models and Life-Tables. Breakthroughs in Statistics [online]. New York, NY: Springer New York, 1992, 1992, , 519-526 [cit. 2020-05-20]. Springer Series in Statistics. DOI: 10.1007/978-1-4612-4380-9_36. ISBN 978-0-387-94039-7. Dostupné z: http://link.springer.com/10.1007/978-1-4612-4380-9_36