Imperial College London

23 September 2021

A Survey of Machine Learning Methodologies for Survival Analysis

MSc Health Data Analytics and Machine Learning

Author: Camilla Callierotti
Supervisor: Dr Abhijit Dasgupta

Imperial College London

Outline

Imperial College London

Aims

Imperial College London

Aims

- This project uses the Cox PH model as the benchmark and surveys three other machine learning models:
  1. Random Survival Forest (RSF)
  2. eXtreme Gradient Boosting (XGBoost)
  3. neural network based methods (DeepSurv and DeepHit)
- The models are compared to determine:
  1. extent of improvement of **prediction performance**
  2. **interpretability** via variable selection measures
  3. robustness to **censoring**

"There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown."
- Leo Breiman, Statistical Modeling: The Two Cultures

Imperial College London

Background

Imperial College London

Survival data

We will focus on **right-censored** survival data, i.e. characterised by incomplete observations of the nature $[y, +\infty)$. We assume that that censoring is **independent** of survival time i.e. the chance that an individual's data is censored is not affected by their risk of experiencing the event.

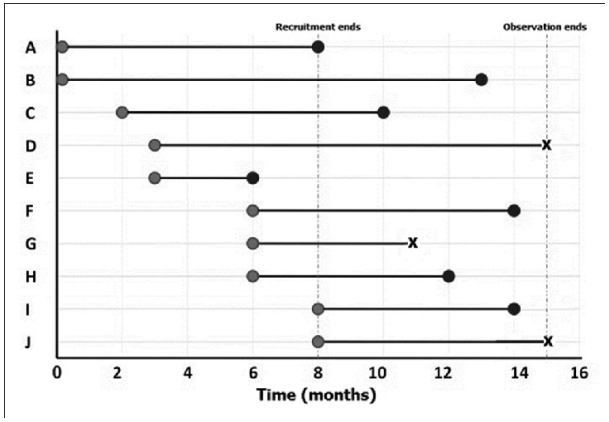

Figure 1: Patient timelines in survival analysis [5]

Descriptive methods

To describe right-censored survival data we use its cumulative distribution function for the random variable survival time T:

$$F(t) = Pr(T \leq t)$$

We use the **survival function** $S(t)$ to better describe survival data by the chance of surviving at least for a particular length of time:

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

We use the **hazard function** $h(t)$ to describe the *instantaneous* risk of an individual at a time t, given the event has not yet occurred until just before time t:

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

We use the **cumulative hazard** to model the hazard against covariates:

$$\begin{aligned} H(t) &= \int_0^t h(s)\, ds \\ &= -\ln S(t) \end{aligned}$$

Nonparametric methods

Give an overall view of the patterns in the data, either marginally or stratified by some predictors of interest.
The survival function can be estimated nonparametrically from observed data via the **Kaplan-Meier estimator**:

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

The hazard function can be estimated nonparametrically from observed data via the textbfNelson-Aalen estimator:

$$H(t) = \sum_{t_i \leq t} \frac{d_i}{n_i}.$$

Regression models for survival data

Regression models for survival analysis are specified by a hazard function $h(t)$ built of two components which accomplish two separate goals [6]:

1. error component: describes the **underlying distribution** of survival time,
2. systematic component: describes how the distribution **changes** as a function of the covariates

Semiparametric models

- Treat the baseline hazard $\lambda_0(t)$ as a nuisance parameter
- Estimates parameters defining ratio of **relative risk** between groups characterised by different covariates (not absolute risk predictions)

Cox regression

Cox (1972) [2] defined a semiparametric hazard function as

$$\lambda(t, x, \beta) = \lambda_0(t) e^{x\beta}$$

thus the hazard ratio between two values of the covariates can be written as

$$HR(t, x_1, x_0) = e^{\beta(x_1 - x_0)} = \frac{e^{\beta x_1}}{e^{\beta x_0}}$$

Fully parametric models

- The underlying distribution of the survival time random variable is defined upon previous knowledge of the data.
- Hence, can make absolute risk predictions

Accelerated Failure Time models

Common distributions of survival time [6]:

Parametric model	Hazard function	Survival function
Exponential	γ	$exp(-\gamma t)$
Weibull	$\gamma\alpha t^{\alpha-1}$	$exp(-\gamma t^{\alpha})$
Gompertz	$\gamma exp(\alpha t)$	$exp(-(\gamma/\alpha)(e^{\alpha t}-1)$
Log-logistic	$\alpha\gamma^{\alpha} t^{\alpha-1}/(1+(\gamma t)^{\alpha})$	$1/(1+(\gamma t)^{\alpha})$

Example: Weibull regression:

$$ln(T) = \beta_0 + \beta_1 x + \sigma * \epsilon^*$$

Machine learning methods

We use machine learning for survival analysis to [1, 10]:

1. Relax assumptions like linearity that are made in the systematic component
2. Allow for data-driven complexity and interactions
3. Make individual and overall predictions about survival times
4. Obtain a slightly biased but more precise predictive model (low bias, low variance models)

ML methods: Random Survival Forest

Random Survival Forest (RSF) is a specific random forest algorithm for survival data developed by Ishwaran (2008) [7], composed of:

- Base learners= binary survival trees with their associated Nelson-Aalen cumulative hazard function
- Ensemble= cumulative hazard function formed by the average of each terminal node's Nelson-Aalen cumulative hazard function

ML methods: Random Survival Forest

Advantages:

- Can learn from lots of features and incorporate complex interactions automatically
- Randomisation at two levels:
  1. Each tree is grown from a different bootstrap sample
  2. A tree is grown by splitting nodes based on a random subset of features

$\rightarrow$ Reduce the effects of overfitting and allow the forest to capture multiple aspects of the data within the training process [8]

ML methods: eXtreme Gradient Boosting

Also uses decision trees as the building blocks which are added in a step-wise fashion such that each successive tree is fit utilising the residuals of the previous tree as the response rather than the survival outcome [9].
Advantages:

- Complementary trees
- Step-wise addition of learners

$\rightarrow$ Improves prediction performance by progressively reducing bias

Imperial College London

ML methods: neural networks

Objective: learn the relationship between the patient's features and the their outcome, utilising the relationship's complexities and non-linearities, without any previous knowledge provided by the data analyst [10, 3]

$\rightarrow$ distribution of survival times is learned by the network
$\rightarrow$ risk networks consist of feed-forwards neural networks outputting the **risk of failure**

Prediction performance metrics

- Standard regression error metrics for continuous outcomes are unfit to assess model performance in survival analysis due to the presence of censored observations
- Harrell's concordance index (c-index) is a rank order statistic defined as the ratio of concordant pairs to total comparable pairs [11, 13]

Harrell's c-index

- Comparable pairs:
  - two uncensored observations, or
  - one censored observation and one uncensored observation whose event time is smaller than the censored observation's censoring time
- the pair (i, j) is concordant if $t_i > t_j$ and $S(t_i) > S(t_j)$
- the pair is discordant if $t_i > t_j$ and $S(t_i) < S(t_j)$

Interpretability measures

P-values versus variable importance

- The variable importance measures (used for RSF and XGBoost) are machine learning alternatives to p-values
- The p-value indicates significance of the variable *within the model itself*
- Variable importance measures how much each feature contributes to the prediction compared to random noise

Imperial College London

Interpretability: Methods

Cox regression:

- Penalised regression
- Optimises the objective function subject to a penalty constraint on the parameter

Random Survival Forest:

- Permutation VIMP
- The variable's 37% OOB data is permuted, the new covariates are passed through the RSF, and OOB error is recalculated

XGBoost:

- SHAP values
- Describe non-linearities in dependency plots

Imperial College London

Methods

Datasets

- Primary Biliary Cirrhosis (PBC) dataset [4]:
  - Randomized placebo controlled trial of the drug D-penicillamine on individuals affected by PBC
  - 424 total cases, 312 of which are complete cases and 112 of which were not randomised, 6 of which were lost to follow-up $\rightarrow$ **312** complete cases used
  - Included 17 covariates including: age, sex, treatment, and variables specific to the disease such as presence of edema, presence of ascites, serum bilirubin, and more
  - In R in the `randomForestSRC` package
- Simulated data
  - Simulated five datasets of 1000 observations characterised by increasing censoring rates (1%, 25%, 50%, 75%, 99%)
  - In R with the `sim.survdata` function of the `coxed` package,
  - In Python with the `SimStudyLinearPH` function of the `pycox.simulations` package

Imperial College London

Software used

All computations were carried out on R version 4.0.3 and Python version 2.7.16. All code is found at the GitHub repository: https://github.com/camicallierotti/imperial-summer-project.

Package	Version
dplyr	1.0.7
rms	6.2
tidyverse	1.3.0
pec	2020.11.17
survival	3.2-12
coxed	0.3.3
glmnet	4.1-2
randomForestSRC	2.12.0
xgboost	1.4.1.1
survXgboost	1.0
ggplot2	3.3.5

Table 1: R packages used

Package	Version
pandas	1.1.3
numpy	1.19.2
sklearn	0.23.2
matplotlib	3.3.2
pycox	0.2.2
pytorch	1.7.1
torchtuples	0.2.0

Table 2: Python packages used

Imperial College London

Results

Imperial College London

Prediction performance

We split the `pbc` data into a training and test set, in a 4:1 ratio. We trained each model on the training set and evaluated predictive performance using Harrell's c-index on the test set.

Model	Cox PH	RSF	XGBoost	DeepHit	DeepSurv
c-index	0.835	0.878	0.901	0.635	0.576

Table 3: Harrell's c-index from the PBC test data using various algorithms

We made 10-year survival probabilities and full survival curve predictions for three test set individuals: youngest, median, oldest.

Patient ID	Age (days)	Cox	RSF	XGBoost	DeepSurv	DeepHit
1	10550	92.1	73.9	99.9	42.7	22.2
2	16941	17.8	18.8	< 0.0001	19.6	24.7
3	25006	11.1	26.8	1.49	24.3	24.7

Table 4: Predicted 10-year survival probabilities for three test set data to all models evaluated.

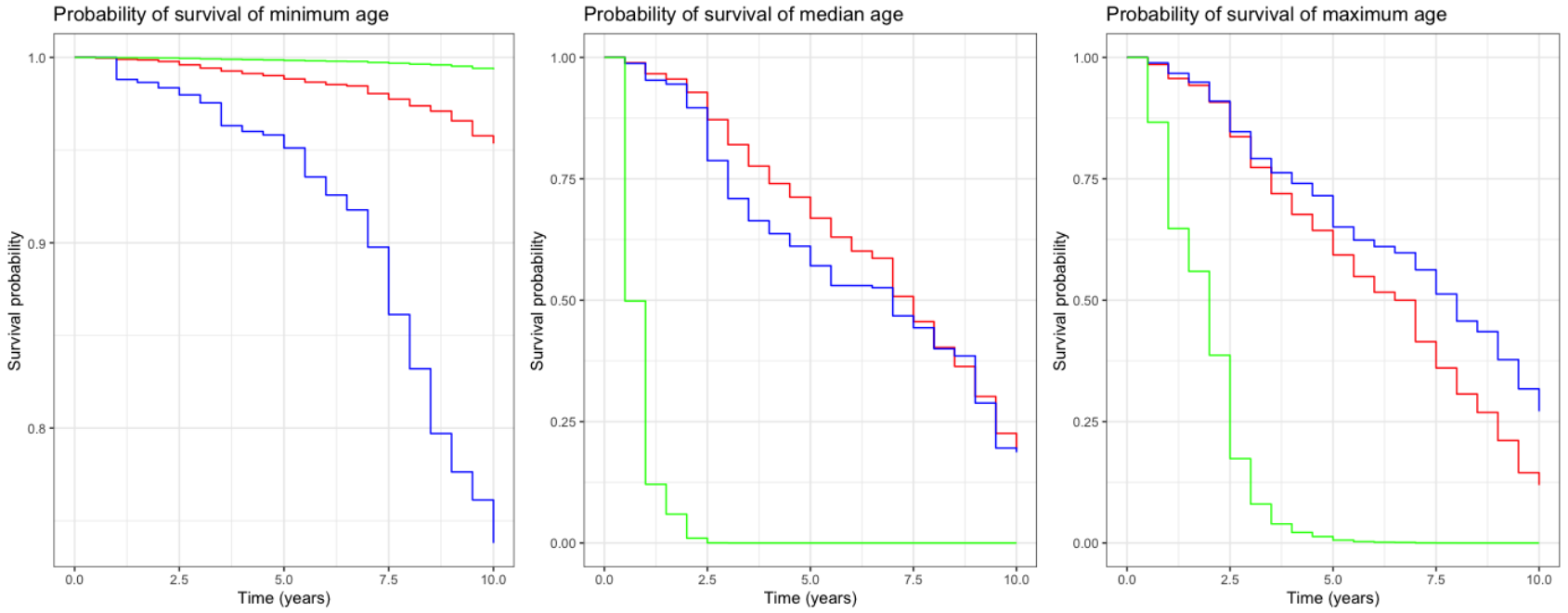

Figure 2: Full Kaplan-Meier survival curves for three individuals of increasing age predicted by Cox model (red), RSF (blue), and XGBoost (green).

- RSF **underestimates** survival at lower ages (patient 1) and **overestimates** survival at higher ages (patient 3) compared to Cox
- XGBoost makes **extreme predictions** on either outcome compared to Cox: almost complete chance of survival after 10 years (patient 1) and an almost complete chance of death after 10 years (patient 3)

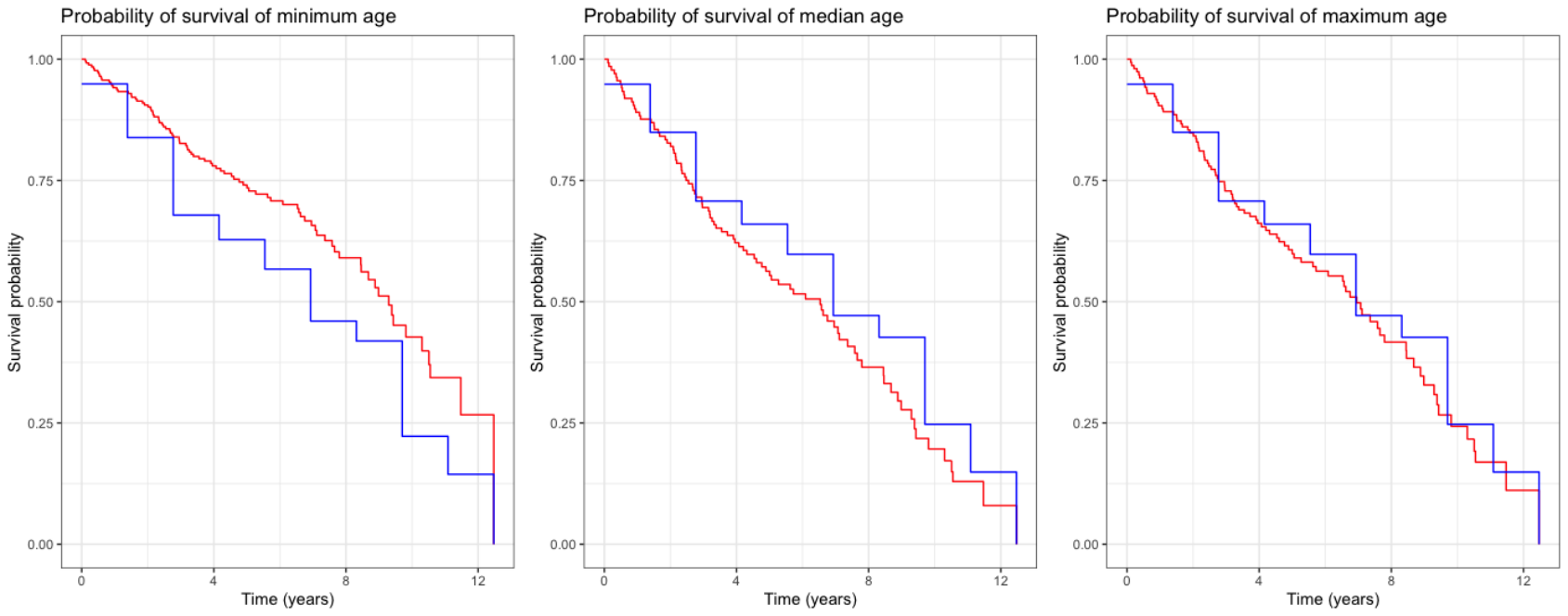

Figure 3: Full Kaplan-Meier survival curves for three individuals of increasing age predicted by DeepSurv (red) and DeepHit (blue).

- DeepSurv and DeepHit produce very **similar** curves for all individuals
- Predicted survival curves are nearly **linear**, incoherent with the known survival pattern for human life

Variable importance

$\rightarrow$ The gold standard for statistical significance are p-values (from Cox regression), hence all variable importance measures will be compared to this benchmark.

Variable	Cox β	Cox p-value	Lasso Cox β
treatment	1.08e-03	0.997	0.0
age	9.58e-05	0.007 **	4.4e-06
sex	-3.05e-02	0.929	0.0
ascites	-8.45e-02	0.848	1.7e-01
hepatom	7.19e-02	0.786	0.0
spiders	-1.98e-01	0.491	0.0
edema	1.58	0.0002 ***	4.5e-01
bili	7.41e-02	0.019 *	7.9e-02
chol	9.94e-04	0.076	0.0
albumin	-6.36e-01	0.068	-3.2e-01
copper	4.14e-03	0.004 **	2.1e-03
alk	-2.17e-05	0.627	0.0
sgot	1.46e-03	0.528	0.0
trig	-1.26e-03	0.414	0.0
platelet	2.13e-04	0.883	0.0
prothrombin	1.84e-01	0.147	3.4e-02
stage	5.20e-01	0.007 **	1.5e-01

Table 5: Full Cox regression variable effect sizes and p-values, and Cox regression with Lasso regularisation variable effect sizes.

- **Good consistency** between variables significant by p-value and variables selected by **Lasso penalty**

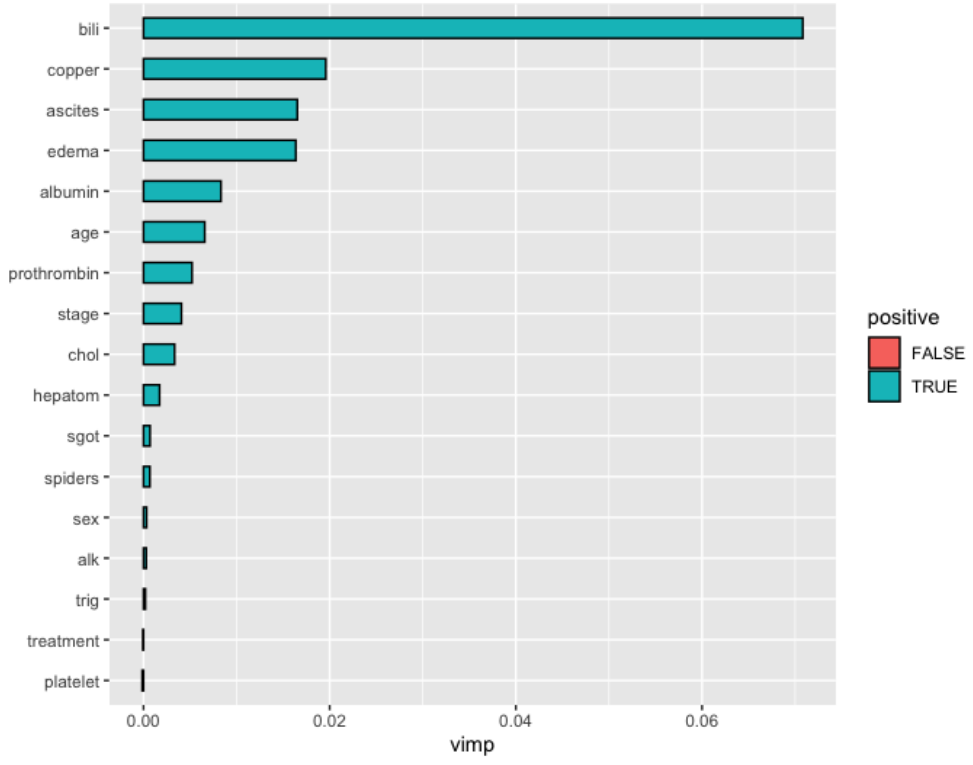

Figure 4: Variable importance (VIMP) plot obtained via the permutation method in RSF.

- Serum bilirubin (bili) is the most important variable in survival prediction (contributing to c-index by 7%); however was not the most significant variable ($p<0.05$, Lasso β=7.9e-02)
- Presence of edema which was the most significant variable, however is not the most important (VIMP<0.02)

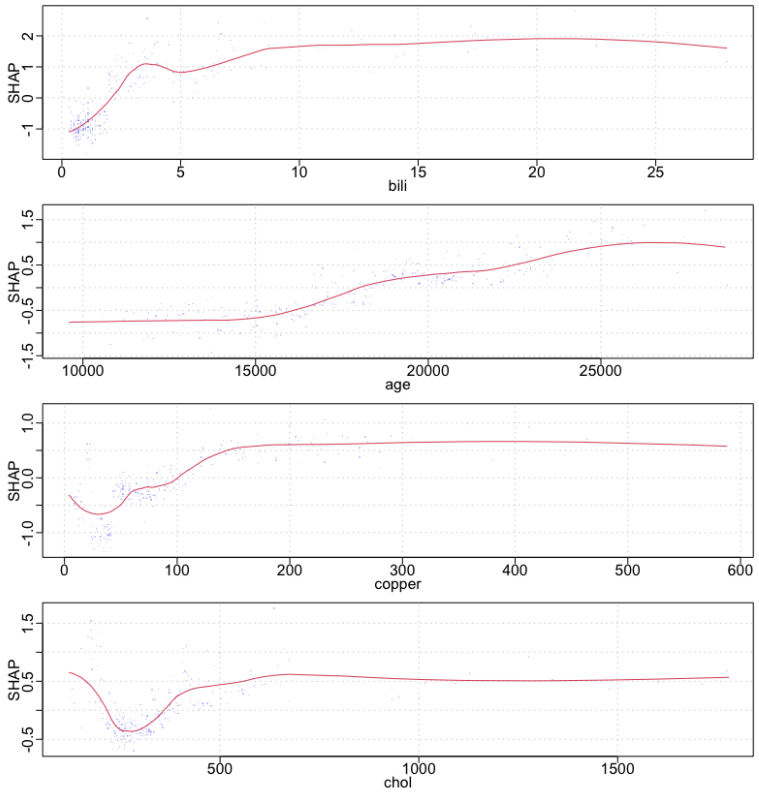

Figure 5: SHAP dependency plots for four most important variables in prediction of survival according to XGBoost.

$\rightarrow$ SHAP values have the added advantage of being **non-linear** with respect to the variable

- Serum bilirubin (bili) positively affects probability of death consistently at values greater than 2.5 mg/dl
- Age positively affects probability of death from ages greater than 48

Imperial College London

Censoring rate sensitivity analysis

Censoring rate	1%	25%	50%	75%	99%
True HR	1.001	0.971	0.899	0.797	0.964
Cox HR	1.004	0.975	0.912	0.819	0.988
95% CI	0.992-1.017	0.961-0.989	0.896-0.928	0.799-0.840	0.874-1.117
HR diff.	0.003	-0.010	0.013	0.022	0.035
CI width	0.025	0.028	0.032	0.041	0.243

Table 6: Censoring rate sensitivity analysis for Cox regression

With increasing censoring rates, Cox regression:

- increasingly overestimates the hazard ratio compared, and
- widens confidence intervals

Censoring rate	1%	25%	50%	75%	99%
Cox PH	0.494	0.581	0.533	0.564	0.604
RSF	0.504	0.553	0.505	0.493	0.854
XGBoost	0.511	0.483	0.535	0.524	0.684
DeepSurv	0.991	0.990	0.990	0.984	0.986

Table 7: Censoring rate sensitivity analysis for Cox PH, RSF, XGBoost, and DeepSurv prediction performance

With increasing censoring rates, all models produce a **higher** c-index (better performance?)
$\rightarrow$ Attributed to a property of the c-index calculation.

Imperial College London

Discussion

Imperial College London

Prediction performance

The different models made *widely different* survival predictions, e.g. recall a 68-year-old female individual with stage 1 primary biliary chirrhosis has the following predicted 10-year survival probabilities:

Patient ID	Cox	RSF	XGBoost	DeepSurv	DeepHit
1	92.1	73.9	99.9	42.7	22.2

The models ranked in the following order in terms of c-index of prediction:

1. **XGBoost**
   - Coherent with results from the literature [1]
   - Ensemble algorithm: learns slowly, prevents overfitting, corrects residual error
   - Recall that predicted Kaplan-Meier curves were extreme: clinical implications
2. **Random Survival Forest**
   - Comparable performance to Cox regression
   - Slightly over- and under- estimated predictions at age range extremes [12]
3. **Cox regression** (benchmark)
   - The literature shows that this classical model outperforms complex models when data is linear [10]
4. **Neural networks**
   - Likely due to overfitting (i.e., the PBC dataset is small and the models learned the noise)

$\rightarrow$ No one-size-fits-all model

Imperial College London

Variable importance

- Statistical methods give p-values which are **highly interpretable**, while machine learning methods perform better but are considered **"black boxes"**
- Results show that p-values, VIMP, and SHAP values are largely overlapping
- SHAP values give added benefit of explaining non-linearities

Imperial College London

Robustness to censoring

- Results show higher c-indices at higher censoring rates
- Misleading: more censoring $\rightarrow$ less complete cases used in calculation $\rightarrow$ bias introduced

$\rightarrow$ Importance of accounting for censoring rate in survival analysis

Imperial College London

Conclusions and future work

- Random Survival Forest was the optimal solution between the extremes
- PBC dataset is a **classical clinical dataset** in terms of features and observation size
- **Growing dimensionality** of data collected in clinical research in recent years thanks to high throughput methods

Imperial College London

Appendix I: RSF Algorithm

The Random Survival Forest algorithm

1. Draw B boostrap samples 63% of the original data, on average
2. Grow a survival tree recursively for each bootstrap sample
3. Reach the constraint that a node should have no less than $d_0 > 0$ unique deaths, indicating that this is the terminal node.
4. Calculate the Nelson-Aalen estimate of the cumulative hazard function for each tree

$$H(t) = \sum_{t_{l,h} \leq t} \frac{d_{l,h}}{Y_{l,h}}$$

5. Calculate the ensemble CHF by averaging over B survival trees

$$H_e^{**}(t|x_i) = \frac{\sum_{b=1}^{B} I_{i,b} \hat{H}_b(t|x_i)}{\sum_{b=1}^{B} I_{i,b}}$$

Appendix II: XGBoost Algorithm

The XGBoost algorithm

1. Set $\hat{f}(x) = 0$ and $r_i = y_i$ for all i in the training set
2. For each bootstrap sample $b = 1, 2, ..., B$:
   1. Fit a tree to that bootstrap sample $\hat{f}^b$ with d splits to the training data (X, r), where the features are the covariates X and the response are the residuals of the previous tree r
   2. Update $\hat{f}$ by adding in a shrunken version of the new tree

$$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x)$$

   3. Update the residuals

$$r_i \leftarrow r_i - \lambda \hat{f}^b(x_i)$$

3. This results in a large number of decision trees $\hat{f}^1, ..., \hat{f}^B$, whose sum produces the output of the boosted model

$$\hat{f}(x) = \sum^{B} \lambda \hat{f}^b(x)$$

Imperial College London

References I

Yifei Chen, Zhenyu Jia, Dan Mercola, and Xiaohui Xie.
A Gradient Boosting Algorithm for Survival Analysis via Direct Optimization of Concordance Index.
Computational and Mathematical Methods in Medicine, pages 1–8, 2016.

D. R. Cox.
Regression Models and Life-Tables.
Journal of the Royal Statistical Society: Series B (Methodological), 34(2):187–202, January 1972.

David Faraggi and Richard Simon.
A neural network model for survival data.
Statistics in Medicine, 14(1):73–82, January 1995.

References II


Thomas R. Fleming and David P. Harrington.
Counting Processes and Survival Analysis.
John Wiley & Sons, Inc, 1991.


Avijit Hazra and Nithya Gogtay.
Biostatistics Series Module 9: Survival Analysis.
Indian journal of dermatology, 62:251–257, May 2017.


David W. Hosmer, Stanley Lemeshow, and Susanne May.
Applied Survival Analysis: Regression Modeling of Time-to-Event Data.
Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., 2008.

References III


Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer.
Random survival forests.
The Annals of Applied Statistics, 2(3), September 2008.


Hemant Ishwaran, Udaya B. Kogalur, Xi Chen, and Andy J. Minn.
Random survival forests for high-dimensional data: Random Survival Forests for High-Dimensional Data.
Statistical Analysis and Data Mining: The ASA Data Science Journal, 4(1):115–132, February 2011.


Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani.

An Introduction to Statistical Learning, volume 103 of *Springer Texts in Statistics*.
Springer New York, New York, NY, 2013.

References IV


Jared L. Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger.
DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network.
BMC Medical Research Methodology, 18(1):24, December 2018.


Andreas Mayr and Matthias Schmid.
Boosting the Concordance Index for Survival Data – A Unified Framework To Derive and Evaluate Biomarker Combinations.
PLoS ONE, 9(1):e84483, January 2014.


Ulla B. Mogensen, Hemant Ishwaran, and Thomas A. Gerds.
Evaluating Random Forests for Survival Analysis Using Prediction Error Curves.
Journal of Statistical Software, 50(11), 2012.

References V

Hajime Uno, Tianxi Cai, Michael J. Pencina, Ralph B. D'Agostino, and L. J. Wei.
On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data.
Statistics in Medicine, 30(10):1105–1117, May 2011.