A Simulation-based comparison of the predictive accuracy of the random survival forest and the lasso-regularized Cox Model in Survival Analysis

An Annotated Bibliography

Willem Van Der Merwe (2914429)
University of Witwatersrand

March 24, 2024

References

[1] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society*, vol. vol. 34, pp. 187–220, 1972. [Online]. Available: http://www.jstor.org/stable/2985181

Aim: Introduction of the foundational Cox proportional hazards model for analyzing time-to-event data, providing proof and application for survival analysis and the handling of censored data.

Style/Type: journal article, theoretical

Cross references: The Cox model's foundational aspects of survival analysis can be extended, comparable or complemented by methods introduced in Tibshirani (1996) with the Lasso for feature selection, Ishwaran et al. (2008) through Random Survival Forests as a machine learning approach, and utilized in the ISD models discussed by Haider et al. (2018) for individualized survival predictions.

Summary: The paper introduces a statistical model, which acts as an extension to prior work formalized as the Kaplan-Meier estimator, by exploring time-to-event data (life tables). The major benefit, speaks to the concept of censored data, which is a known concept in survival analysis, there is missing information within the data, specifically, event occurrence without observation on a continuous time scale. The proposition consists of covariates, known as attributes regarding a unit in a distribution of data, which is associated with a coefficient β scaling the impact of said covariates, this product is then bound by the baseline hazard $h_0(t)$.

$$h(t|X) = h_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n)$$

Hazard in other words is the estimated conditional probabilities, in line with the observed conditional frequencies of events or simply the risk of event occurrence at a specific time. An assumption of the Cox model is the proportional hazards assumption,

suggesting that the hazard ratios for different covariates remain constant over time we see this for two events observations,

$$\frac{h(t|X_1)}{h(t|X_2)} = \frac{h_0(t)\exp(\beta^T X_1)}{h_0(t)\exp(\beta^T X_2)} = \frac{\exp(\beta^T X_1)}{\exp(\beta^T X_2)} = \exp(\beta^T (X_1 - X_2))$$

This is an operational Assumption and a limitation as this is not always true for survival data. The model can handle censoring, by adjusting the likelihood function for observations where event occurrence did not happen in a particular continuous time frame, and by maximizing the likelihood of all observed events, it is possible to estimate the coefficients that could work the best under the Cox formulation.

$$L(\beta) = \prod_{i:\delta_i=1} \frac{\exp(\beta^T X_i)}{\sum_{j \in R(t_i)} \exp(\beta^T X_j)}$$

[2] H. Haider, B. Hoehn, S. Davis, and R. Greiner, "Effective ways to build and evaluate individual survival distributions," *Journal of Machine Learning Research (JMLR)*, vol. 21, no. 2020, pp. 18–772, 11 2018.

Aim: Exploration of the possible advantages and methodologies for modeling individual survival distributions as opposed to population-level approaches, introducing tools for personalizing survival predictions.

Style/Type: journal article, empirical

Cross references: The ISD paper builds upon the foundational work of Cox (1972) by applying survival analysis on the individual level and could leverage advancements in regularization and feature selection methods like those introduced by Tibshirani (1996) to improve model performance. It also aligns with the machine learning approaches to survival analysis discussed by Ishwaran et al. (2008), showcasing a practical application of combining traditional statistical models with modern computational techniques for individualized predictions.

Summary: This paper is a detailed exploration of the differences and benefits of modelling individual survival distributions as compared to the current models that act on a general population. It is provided that standard methods fall short in assessing effects for an individual and suggest the use of Individual Survival Distribution (ISD) models. The key outline is that the paper presents five classes of tools, One Value individual risk models ($[R, 1, V_i]$), which is in line with the Cox model which provides risk scores for each patient, single time group risk predictors ($[R, 1_{t^*}, g]$), such as prognostic scales, assigning patients to risk groups, single time individual probabilistic predictors ($[R, 1_{t^*}, g]$), like the Gail Model, which gives the probability at a specific time point for an individual, Group survival distributions ($[P, \infty, g]$), like Kaplan-Meier curves, provides population level probabilities over time, and lastly individual survival distribution models($[P, \infty, i]$), which are a group of models outlined in the paper for evaluation, models like cox extensions are used which generates probability curves per patient across all time. For evaluation, the authors highlight concordance, which compares predicted risk scores with outcomes, L1-loss which yields the average absolute difference between predicted and outcome

survival times, 1-calibration which checks specific predicted time point probabilities are well calibrated to actual survival rates, Integrated Brier score (IBS), which indicates calibration and discrimination measures. The authors propose an evaluation method called D-calibration, a method that aims to indicate if probability estimates are meaningful across survival curves. The method operates over subsets (D_Θ) across the entire data distribution.

$$\frac{|D_\Theta([a,b])|}{|D|} = b - a$$

During model comparison the authors point out that there is an important difference between the datasets in order to make a fair evaluation, and the four datasets which have the most complete data are labelled as NICE datasets. The other datasets are processed to include mean data imputation methods. The study provides empirical results after applying the above groups with relevant methods among the groups being, Cox Kalbfleisch-Prentice (COX-KP), accelerated time failure model (ATF), Random survival forests (RSF), multi-task linear regression (MLTR), Kaplen-Meier (KM), elstic net Cox (COXEN-KP) which show the MLTR method outperforms most models across the different measures and demonstrates the concept of D-calibration, proving that some of the methods demonstrate cross-distribution calibration. The authors round off with key considerations for selecting the appropriate model, based on the proposed groups, which stems from data complexity factors and censoring contained within datasets. Finally it is also important to notice when ISD are most useful across calibration and discrimination scenarios.

[3] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random survival forests," *Annals of Applied Statistics*, vol. 2, no. 3, pp. 841–860, 2008. [Online]. Available: DOI 10.1214/08-AOAS169

Aim: Introduction of the foundational Random Survival Forests machine learning approach for survival analysis, capable of managing right-censored data and estimating survival functions.

Style/Type: journal article, theoretical

Cross references: Ishwaran's method, Random Survival Forests presents an alternative to the Cox model (Cox, 1972) by incorporating machine learning into survival analysis, a methodological advancement that shares the goal of the ISD paper (Haider et al., 2018) in enhancing predictive performance in survival data.

Summary: Random survival forests are an extension of random forests, which have the ability to handle right-censored data and aim to estimate the appropriate survival function. Consisting of an ensemble of trees, which are grown from a bootstrap sample, and each node of underlying trees, consists of covariates. Per node splitting criteria are conditional to survival time and censoring, whereby node "impurity" is determined by the survival differences. Methods like log-rank, conservation of events splitting rule, and random log rank are used. Terminal nodes are the result of saturated splitting criteria, with each endpoint having d-dimensional covariates of the individuals encapsulated. A key component of the model is the conservation of events principle, which is used to define a type of predicted outcome, namely

ensemble mortality,

$$\hat{M}^*_{e,i}(t) = \sum_{j=1}^{n} H^*_e(t_j|X_i)$$

which is derived from the cumulative hazard function (CHF) using the Nelson-Aalen estimator. All terminal nodes share the estimated hazard function.

$$\hat{H}_h(t) = \sum_{t_i,h \leq t} \frac{d_i, h}{n_i, h}$$

Another key concept is the out-of-bag (OOB) samples which act as a validation subset. The OOB error is calculated on the ensemble survival function with regard to the observed data using metrics like concordance.

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

Prediction error metrics, like the concordance index which calculates the permissible pairs per node and OOB prediction error, are used for accuracy metrics. Variable Importance (VIMP) is assessed by looking at each predictor variable in the sample and assessing the impact on prediction error, increases in error indicating importance. The paper puts forward an approach to deal with missing data, outlining the shortcomings of prior methods like replacing missing values with distribution medians, and for categorical data replacing with most frequent occurrences. The method is called adaptive tree imputation and relies on the OOB data set to determine missing data, for both continuous or integer values. A Key benefit of the model is within its ability to capture survival functions for an individual in the distribution by estimating its survival function across all trees where the individual is captured in terminal nodes. Another benefit is that the model is well suited for high dimensional data because of the random subset selection process, which helps mitigate overfitting. Due to the permutative nature of the ensemble bound to the brevity of the underlying data distribution, the model is computationally demanding, and although the model can yield variable information, it might be difficult to interpret the final resulting model

[4] G. Kantidakis, E. Biganzoli, H. Putter, and M. Fiocco, "A simulation study to compare the predictive performance of survival neural networks with cox models for clinical trial data," *Computational and Mathematical Methods in Medicine*, vol. 2021, p. 2160322, 2021, publisher: Hindawi. [Online]. Available: https://doi.org/10.1155/2021/2160322

Aim: Comparison of predictive performance of Survival Neural Networks (SNNs) with traditional Cox models in the context of simulated data extracted from clinical trial data, focusing on survival outcomes.

Style/Type: journal article, empirical

Cross references: This paper compares survival neural networks (SNNs) with the Cox model, thereby indirectly interacting with the model and methods thereof, which is introduced by Cox (1972), furthermore the regularization technique by Tibshirani (1996), could be relevant for enhancing SNNs. The comparison of

SNNs with traditional models mirrors the ISD paper's exploration of innovative versus traditional approaches for individual survival predictions.

Summary: The paper shows a comparison between machine learning methods, the author's term survival neural networks (SNNs) compared to the Cox proportional hazards model, using clinical trial data for survival outcomes. The models are formulated subject to the European Osteosarcoma intergroup trial data, which is used as the foundation for the synthetic data generation that would ultimately be used for simulation training. The original dataset contains various instances of censoring, and the authors approach this issue, by segmenting the datasets into samples with degrees of censoring present, after, data imputation techniques such as the inverse probability censoring weighting method, which is used that is based on calibration procedures outlined in the paper, to ensure the synthetic data retains the statistical properties of the original clinical data. The authors go on throughout the methodology to explain the architecture of the neural network architectures used, namely partial logistic artificial neural networks(PLANN) such as underlying components like the error functions, activation functions employed, and how the methods are applied for simulation by looking at fitting the underlying dataset to the model. The basic model is shown by,

$$\lambda(x_i, \alpha_l, w) = f\left[w'_{0k} + \sum_{h=1}^{H} w'_{hk} g_h \left(w_{0h} + w_{1h}\alpha_l + \sum_{j=1}^{p} w_{(j+1)h} x_{ij}\right)\right]$$

where j = 1, 2, ..., J are the nodes in the input layer, h = 1, 2, ..., H are the nodes in the hidden layer, and w are the weights of the network. The training is performed with training sets and validation sets, using cross-validation and hyperparameter tuning. Furthermore, evaluation techniques are explored, specifically looking at factors like discrimination (C-index) by using the average time dependant nonlinear prognostic index,

$$\theta(x_i, \alpha_l, w) = \log\left[\frac{\lambda(x_i, \alpha_l, w)}{1 - \lambda(x_i, \alpha_l, w)}\right],$$

$$\theta(x_i, w) = \frac{1}{L}\sum_{l=1}^{L} \theta(x_i, \alpha_l, w),$$

accuracy (Brier) interpreted in continuous form as the integrated Brier score for prediction error over the total period,

$$\text{Brier}(y, \hat{S}(t_0|x)) = (y - \hat{S}(t_0|x))^2,$$

$$\text{Err}_{\text{score}}(\hat{S}, t_0) = -\frac{1}{n}\sum_i \{d_i = 1\{t_i > t_0\}\} \cdot \text{score}(1\{t_i > t_0\}, \hat{S}(t_0|x_i)) \cdot \frac{1}{\hat{C}(min(t_i - t_0)|x_i)}$$

And lastly Miscalibration (mean squared error) for censored groups. The results indicated comparable predictive performance but highlighted a lack of accuracy for calibration measures with SNNs. The authors point out that although machine learning techniques are attractive for survival analysis scenarios because of the ability to model interactions and nonlinearities and no assumption approach, the robustness of

the Cox model, regarding ease of implementation as well as interpretability of covariates makes it formidable in situations where limited sample sizes and variables are available. The paper ties in nicely with the other literature in this document, where the authors indicated the need for clear and better implementation of calibration metrics specifically with machine learning models, and caution against indiscriminate application of these models.

[5] H. Smith, M. Sweeting, T. Morris, and M. J. Crowther, "A scoping methodological review of simulation studies comparing statistical and machine learning approaches to risk prediction for time-to-event data," *Diagnostic and Prognostic Research*, 2022. [Online]. Available: https://doi.org/10.1186/s41512-022-00124-y

Aim: Methodological review of simulation studies that compare statistical and machine learning approaches for risk prediction in time-to-event data, which highlights trends, gaps, and standards.

Style/Type: journal article, empirical

Cross references: Smith's methodological review underscores the importance of comparing statistical and machine learning approaches, referencing foundational works like Cox's (1972) introduction of the proportional hazards model, Tibshirani's (1996) Lasso regression, Ishwaran et al.'s (2008) Random Survival Forests, and Kantidakis et al.'s (2021) analysis on Survival Neural Networks. This review aligns with the objectives of Haider et al. (2018), who investigate these methodologies within the context of individual survival distributions.

Summary: The paper explores the Pubmed database, regarding simulation studies on time-to-event data, where both statistical and machine learning approaches for risk prediction are used. The first part entails the mechanism the authors utilized to derive a target set of 10 papers most relevant to the criterion for selection. The comparative findings are broken into categories for analysis, where components are evaluated within several sections. The study looks at the data-generating methods (DGM) used in the models, and the coverage these mechanisms obtain in terms of sample sizes, covariate attributes, censoring, and failure distributions to name a few. Then covariates and failure time analysis within the studies, are compared across data distribution families and covariate relationships and effects and operational model assumptions. Next percentage censoring is compared across underlying distribution families and training subsets splits with regards to DGM methods used. The studies are categorized into three sets noting which statistical and machine learning models the studies used namely statistical methods, which include Cox proportional hazards and variants thereof, hybrid models like the superlearner method and Mahalanobis K nearest neighbour, and machine learning methods like random survival forests and support vector machines. Finally, the studies are compared across estimands like survival function, hazard functions, and linear predictors to name a few, and performance measures, which entail common methods like the C-index, mean square prognostic error, etc. The findings of the paper point to key considerations, when conducting research simulation studies, of the nature of the underlying papers. Authors highlight poor reporting standards across studies,

as well as missing information on the implementation of DGM's where it is noted that the method performance is directly dependent on the DGM. Furthermore, it is pointed out that there is a level of bias towards machine learning methods used. The authors suggest that researchers consider comprehensive evaluation, and the use of standards, pointing to two sources of interest, and highlight the need for simulation studies that perform method comparisons independent of novel method development, assessing both discrimination and calibration, report variations in performance measures and consider fairness for comparing underlying methods.

[6] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society.*, vol. vol. 58, pp. 267–88, 1996. [Online]. Available: https://www.jstor.org/stable/2346178

Aim: Introduction of the Lasso method for regression analysis, aiming to improve model interpretability and prediction accuracy by incorporating feature selection through regularization.

Style/Type: journal article, theoretical

Cross references: This work directly complements Cox (1972) by offering a method (Lasso) that can enhance the Cox model's predictive accuracy and interpretability. Furthermore, the ISD paper by Haider et al. (2018) implicitly benefits from such regularization techniques to handle high-dimensional data efficiently as it uses statistical regression models.

Summary: The paper presents proof for a new method that aims to solve two objective issues with Ordinary Least Squares (OLS) estimates. It improves prediction accuracy and model interpretability for scenarios where the number of predictors is large or when there is collinearity among predictors. The method minimizes the residual sum of squares in relation to the sum of absolute values of the coefficients being less than a constant.

$$\hat{\beta}^{lasso} = \arg\min_{\beta} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} X_{ij}\beta_j)^2 \right\}$$

bound by

$$\sum_{j=1}^{p} |\beta_j| \leq t.$$

The benefit of the model is that on a continuous scale, the method has the ability to set some coefficients exactly to zero, which excludes feature contributions and thereby performs variable selection, which yields improved interpretability. In the evaluation section of the paper, lasso is compared with subset selection, ridge regression, and the non-negative Garotte method, which shows competitive, prediction accuracy in various scenarios characterized by the number and magnitude of effect sizes of the predictors. Specifically, it performs well in settings with a moderate number of moderate-sized effects and when there's a large number of small effects, indicating its versatility across different regression contexts.