PhD NKI Breast Cancer Survival Analysis

Diego Vallarino

2023-03-08

Contents

Introduction

For the development of this document, the database was used: NKI Breast Cancer Data. This base has the following characteristics: 272 rows (data.train = 190 and data.test =82), 1570 columns. Gene-expression-only network. Patient, treatment, and survival meta data.

Each node has similar patients. Flares (left) denote subpopulations. Estrogen expression (low = top flare, high = bottom flare) distinguishes them. Bottom flare patients survive 100%. Top flare survival is low at the tip (red) and excellent at the base (circled).

Understanding the ringed group of excellent survivors, who had genetic markers of bad survivors (low ESR1 levels, which is a predictive predictor of poor breast cancer outcomes), might help reduce breast cancer death rates. The Outcome Column (Event Death, binary - 0,1) was used as a Data Lens to swiftly analyze why this group survived (which we term Supervised vs Unsupervised analyses).

PNAS and Nature papers: http://www.nature.com/articles/srep01236 The dataset is here: https://data.world/deviramanan2016/nki-breast-cancer-data

Descriptive

```
## [1] FALSE
```

```
## # A tibble: 6 x 12
##     age status  time chemo hormonal amputa~1 histt~2  diam posno~3 grade angio~4
##   <dbl>  <dbl> <dbl> <dbl>    <dbl>    <dbl>   <dbl> <dbl>   <dbl> <dbl>   <dbl>
## 1    43      0 14.8      0        0        1       1    25       0     2       3
## 2    48      0 14.3      0        0        0       1    20       0     3       3
## 3    38      0  6.64     0        0        0       1    15       0     2       1
## 4    50      0  7.75     0        1        0       1    15       1     2       3
## 5    38      0  6.32     0        0        1       1    15       0     2       2
## 6    42      0  2.74     1        0        1       1    10       1     1       1
## # ... with 1 more variable: lymphinfil <dbl>, and abbreviated variable names
## #   1: amputation, 2: histtype, 3: posnodes, 4: angioinv
```

```
## tibble [272 x 12] (S3: tbl_df/tbl/data.frame)
##  $ age       : num [1:272] 43 48 38 50 38 42 50 43 47 39 ...
##  $ status    : num [1:272] 0 0 0 0 0 0 0 0 0 1 ...
##  $ time      : num [1:272] 14.82 14.26 6.64 7.75 6.32 ...
##  $ chemo     : num [1:272] 0 0 0 0 0 1 1 1 1 0 ...
##  $ hormonal  : num [1:272] 0 0 0 1 0 0 1 0 0 0 ...
##  $ amputation: num [1:272] 1 0 0 0 1 1 0 0 0 0 ...
##  $ histtype  : num [1:272] 1 1 1 1 1 1 1 1 1 1 ...
##  $ diam      : num [1:272] 25 20 15 15 15 10 25 15 18 17 ...
##  $ posnodes  : num [1:272] 0 0 0 1 0 1 1 3 1 0 ...
##  $ grade     : num [1:272] 2 3 2 2 2 1 1 2 3 3 ...
##  $ angioinv  : num [1:272] 3 3 1 3 2 1 1 2 1 1 ...
##  $ lymphinfil: num [1:272] 1 1 1 1 1 1 1 1 2 1 ...
```

Data Management

```r
set.seed(123)
data.train <- sample_frac(df1, 0.7)
train_index <- as.numeric(rownames(data.train))
data.test <- df1 [-train_index, ]

surv_obj = Surv(data.test$time, data.test$status)
```

Traditional Models

Modelo de Tobit - parametric survival model

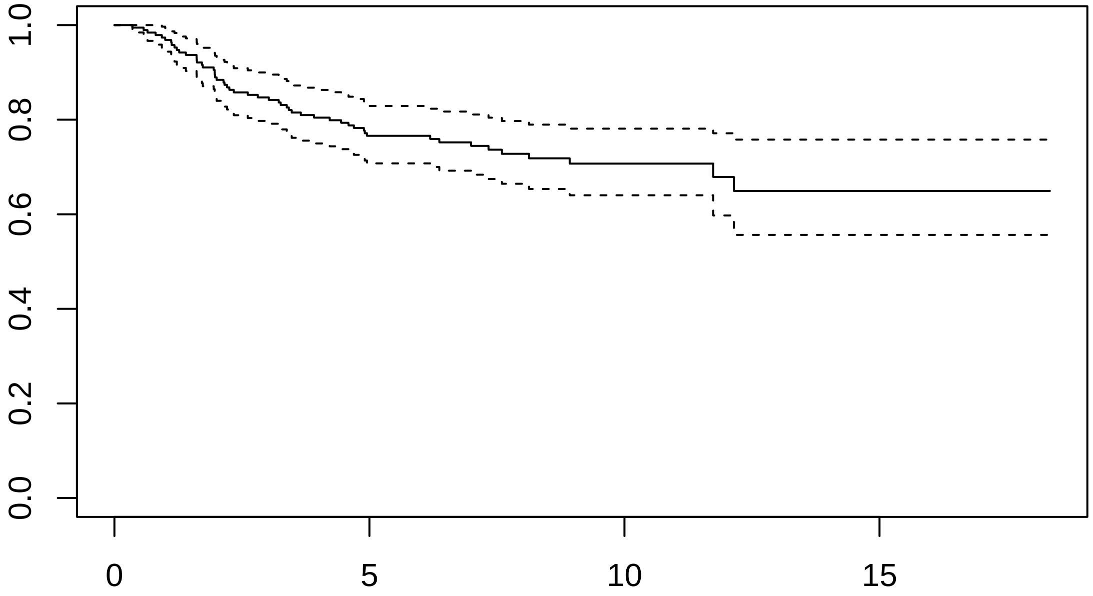
```
##
## Call:
## tobit(formula = time ~ age + chemo + hormonal + amputation +
##     histtype + diam + posnodes + grade + angioinv + lymphinfil,
##     data = data.train)
##
## Observations:
##          Total  Left-censored     Uncensored Right-censored
##            190              0            190              0
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.16857    3.02109   1.711   0.0871 .
## age          0.11646    0.05832   1.997   0.0458 *
## chemo        1.37681    0.68013   2.024   0.0429 *
## hormonal    -1.85659    0.87826  -2.114   0.0345 *
## amputation   0.18770    0.60368   0.311   0.7559
## histtype    -0.57305    0.45514  -1.259   0.2080
## diam        -0.02652    0.03644  -0.728   0.4669
## posnodes    -0.12107    0.15936  -0.760   0.4474
## grade       -0.82212    0.43236  -1.902   0.0572 .
## angioinv    -0.53795    0.35056  -1.535   0.1249
## lymphinfil   0.60313    0.56529   1.067   0.2860
## Log(scale)   1.35346    0.05130  26.384   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Scale: 3.871
##
## Gaussian distribution
## Number of Newton-Raphson Iterations: 3
## Log-likelihood: -526.8 on 12 Df
## Wald-statistic:  20.1 on 10 Df, p-value: 0.02832
```

Modelo Regression

```
##
## Call:
## survreg(formula = Surv(time, status) ~ age + chemo + hormonal +
##     amputation + histtype + diam + posnodes + grade + angioinv +
##     lymphinfil, data = data.train)
##               Value Std. Error     z       p
## (Intercept)  3.1064     1.6023  1.94  0.0525
## age          0.0751     0.0299  2.51  0.0120
## chemo        0.9435     0.4184  2.26  0.0241
## hormonal    -0.1701     0.6065 -0.28  0.7791
## amputation   0.1351     0.3358  0.40  0.6873
## histtype    -0.3497     0.2299 -1.52  0.1283
## diam        -0.0409     0.0167 -2.45  0.0143
## posnodes    -0.1252     0.0723 -1.73  0.0833
## grade       -1.2248     0.2962 -4.13 3.6e-05
## angioinv    -0.0521     0.1821 -0.29  0.7749
## lymphinfil   0.8894     0.3041  2.92  0.0034
## Log(scale)   0.0884     0.1164  0.76  0.4477
##
## Scale= 1.09
##
## Weibull distribution
## Loglik(model)= -199.5   Loglik(intercept only)= -226
##  Chisq= 53.07 on 10 degrees of freedom, p= 7.2e-08
## Number of Newton-Raphson Iterations: 6
## n= 190
```

Kaplan-Meier Model - non parametric survival model

```
## Call: survfit(formula = Surv(time, status) ~ 1, data = data.train)
##
##        n events median 0.95LCL 0.95UCL
## [1,] 190     53     NA      NA      NA
```



Cox models - semi parametric survival model

```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  fit4$residuals
## W = 0.88825, p-value = 1.043e-10
```

```
## Analysis of Deviance Table
##  Cox model: response is Surv(time, status)
## Terms added sequentially (first to last)
##
##             loglik   Chisq Df Pr(>|Chi|)
## NULL       -262.80
## age        -260.34  4.9302  1   0.026392 *
## chemo      -258.71  3.2621  1   0.070897 .
## hormonal   -258.52  0.3825  1   0.536272
## amputation -258.03  0.9775  1   0.322820
## histtype   -258.02  0.0077  1   0.930152
## diam       -252.18 11.6971  1   0.000626 ***
## posnodes   -250.78  2.7803  1   0.095430 .
## grade      -242.08 17.4162  1  3.003e-05 ***
## angioinv   -241.25  1.6524  1   0.198631
## lymphinfil -234.92 12.6629  1   0.000373 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
## Call:
## coxph(formula = Surv(time, status) ~ ., data = data.train, x = TRUE)
##
##   n= 190, number of events= 53
##
##                 coef exp(coef)  se(coef)      z Pr(>|z|)
## age        -0.071170  0.931303  0.027128 -2.623 0.008704 **
## chemo      -0.794107  0.451985  0.378796 -2.096 0.036047 *
## hormonal    0.050192  1.051473  0.558201  0.090 0.928352
## amputation -0.080302  0.922837  0.304051 -0.264 0.791697
## histtype    0.384261  1.468529  0.214089  1.795 0.072675 .
## diam        0.041178  1.042037  0.015214  2.707 0.006799 **
## posnodes    0.108560  1.114672  0.065710  1.652 0.098510 .
## grade       1.203876  3.333010  0.252353  4.771 1.84e-06 ***
## angioinv    0.001297  1.001298  0.167899  0.008 0.993838
## lymphinfil -0.905513  0.404334  0.273939 -3.306 0.000948 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
##            exp(coef) exp(-coef) lower .95 upper .95
## age           0.9313     1.0738    0.8831    0.9822
## chemo         0.4520     2.2125    0.2151    0.9496
## hormonal      1.0515     0.9510    0.3521    3.1401
## amputation    0.9228     1.0836    0.5085    1.6747
## histtype      1.4685     0.6810    0.9653    2.2342
## diam          1.0420     0.9597    1.0114    1.0736
## posnodes      1.1147     0.8971    0.9800    1.2679
## grade         3.3330     0.3000    2.0325    5.4656
## angioinv      1.0013     0.9987    0.7205    1.3915
## lymphinfil    0.4043     2.4732    0.2364    0.6917
```

```
##
## Concordance= 0.783  (se = 0.031 )
## Likelihood ratio test= 55.77  on 10 df,   p=2e-08
## Wald test            = 49.9  on 10 df,   p=3e-07
## Score (logrank) test = 57.64  on 10 df,   p=1e-08
```

Machine Learning Models

MTLR Model - machine learning model

```
##
## Call:  mtlr(formula = Surv(time, status) ~ ., data = data.train, nintervals = 9)
##
## Time points:
##  [1]  1.95  3.39  5.19  5.86  6.77  7.49  8.47  9.58 11.12 13.10
##
##
## Weights:
##         Bias     age   chemo  hormonal amputation  histtype   diam  posnodes
## 1.95   0.151 -0.0185 -0.0269 -0.008012   -0.00447  0.021674 0.0343 -0.003925
## 3.39   0.201 -0.0179 -0.0239 -0.013447    0.00848  0.015081 0.0298 -0.000627
## 5.19   6.676 -0.0258 -0.0311  0.000527    0.01316  0.004127 0.0386  0.015005
## 5.86  -5.226 -0.0257 -0.0311  0.000640    0.01314  0.004127 0.0385  0.015023
## 6.77  -0.198 -0.0256 -0.0383 -0.002194    0.01397  0.001726 0.0361  0.008159
## 7.49  -0.226 -0.0323 -0.0346 -0.004778    0.01422 -0.000737 0.0265  0.003560
## 8.47   0.804 -0.0255 -0.0311 -0.006915    0.01330  0.006110 0.0302  0.018384
## 9.58   2.377 -0.0212 -0.0242 -0.007970    0.00786  0.005206 0.0313  0.017527
## 11.12 -4.324 -0.0212 -0.0241 -0.007955    0.00791  0.005217 0.0313  0.017513
## 13.1  -2.317 -0.0185 -0.0205 -0.007617   -0.00226  0.004424 0.0260  0.012821
##        grade  angioinv lymphinfil
## 1.95  0.0350 -0.000261   -0.01038
## 3.39  0.0425  0.014005   -0.01104
## 5.19  0.0575  0.022705   -0.00802
## 5.86  0.0574  0.022690   -0.00803
## 6.77  0.0655  0.026050   -0.01313
## 7.49  0.0671  0.017725   -0.01773
## 8.47  0.0558  0.016029   -0.02152
## 9.58  0.0602  0.012248   -0.02356
## 11.12 0.0602  0.012304   -0.02350
## 13.1  0.0517  0.005247   -0.02553
```

Random Forest

```
##                          Sample size: 190
##                     Number of deaths: 53
##                      Number of trees: 500
##            Forest terminal node size: 15
##        Average no. of terminal nodes: 7.49
## No. of variables tried at each split: 4
##               Total no. of variables: 10
##        Resampling used to grow trees: swor
##     Resample size used to grow trees: 120
##                             Analysis: RSF
##                               Family: surv
##                       Splitting rule: logrank *random*
##        Number of random split points: 10
##                           (OOB) CRPS: 0.15663457
##    (OOB) Requested performance error: 0.32179584
```

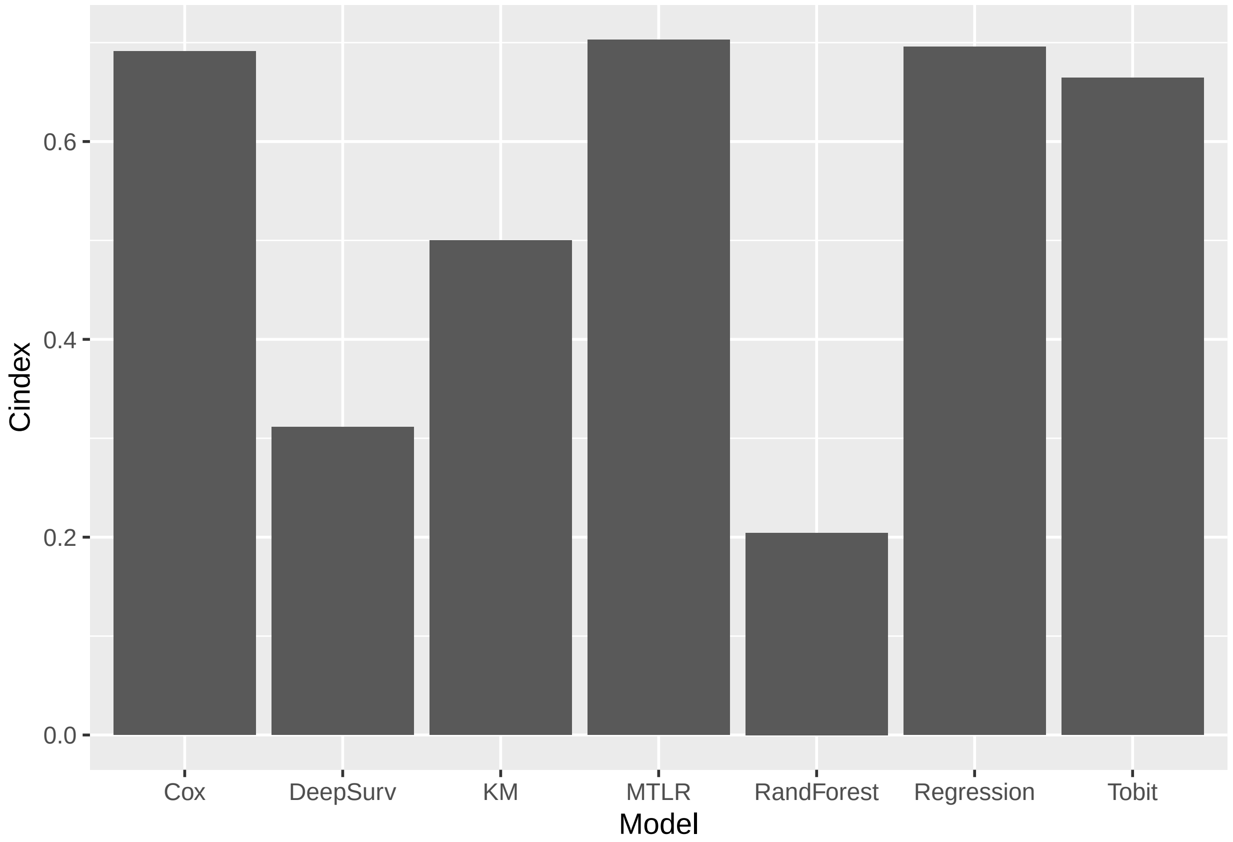
DeepSurv Model

```
##
##  DeepSurv Neural Network
##
## Call:
##   deepsurv(data = data.train, frac = 0.3, activation = "relu",      num_nodes = c(4L, 8L, 4L, 2L), di
##
## Response:
##   Surv(time, status)
## Features:
##   {age, chemo, hormonal, amputation, histtype, diam, posnodes, grade, angioinv, lymphinfil}
```

Models Ranking List

```
## [[1]]
##     Cindex      Model
## 1 0.664537      Tobit
## 2 0.695847 Regression
## 3 0.500000         KM
## 4 0.691374        Cox
## 5 0.702875       MTLR
## 6 0.204473 RandForest
## 7 0.311502   DeepSurv
```

Models Ranking Chart



Conclusion

Why MTLR outperforms neural networks and forest random survival?

Without understanding the data collection and model, answering this question is challenging. I can provide some generic explanations: Data size: Neural Networks and Random Forests are more complex than Multiple Linear Regression (MTLR). They need additional data to fit appropriately. MTLR may be more accurate with less data than neural networks or random forests.

Hyperparameter Tuning: The hyperparameters of a machine learning model determine its performance. The neural network or random forest may not be accurate if not properly configured. The MTLR features fewer hyperparameters, making tweaking simpler. Sensitivity to data distribution: Different models may respond differently to input data distribution. The random forest or neural network design may not operate well with the dataset's data distribution. MTLR data distribution is more robust.

Survival statistics: Survival statistics might favor particular models. MTLR may be appropriate if your data show a linear association between entrance characteristics and survival. However, a neural network or random forest may function better if there are intricate survival-feature correlations or key nonlinear characteristics.

In survival issues, neural networks and random forest models may underperform multiple linear regression models for several reasons. The data set and model determine the reasoning.