Technische Universität Braunschweig

Hannover Medical School

BRICS
Braunschweig Integrated Centre of Systems Biology

PLRI
PETER L. REICHERTZ INSTITUTE FOR MEDICAL INFORMATICS

Hackathon: Breast cancer survival time analysis

August 29, 2024

Daniel Dehncke, Lisa-Marie Bente

Division Data Science in Biomedicine
Peter L. Reichertz Institute for Medical Informatics
of TU Braunschweig and Hannover Medical School
daniel.dehncke@plri.de, lisa-marie.bente@plri.de, www.plri.de

Content

- Breast Cancer
- Survival time analysis
- Tasks
  - **Challenge 1**: Survival Prediction
  - **Challenge 2**: Cox Regression
- Dataset
- Wrap-up

Breast Cancer

- type of cancer that starts in the cells of the breast.
- most common cancer among women worldwide.
- risk factors can be age, family history, genetic mutations, hormonal factors, and lifestyle.
- Treatment options include surgery, radiation therapy, chemotherapy and hormone therapy.

Survival Time Analysis

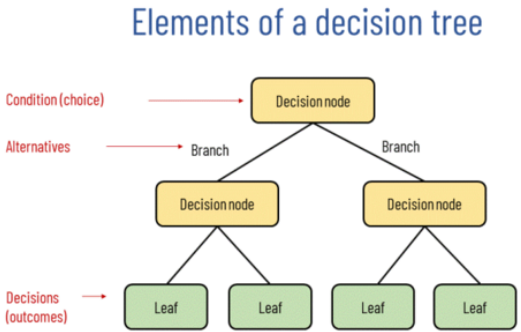
- used to analyze the time until an event of interest occurs.
- Concepts:
  - **Censoring**: Data points where the event of interest has not occurred by the end of the study.
  - **Survival Function**: Probability of surviving past a certain time point.
  - **Hazard Function**: Instantaneous rate of occurrence of the event.
- **Cox Proportional Hazards Model**:
  - Parametric method for analyzing the association between covariates and survival time.

Challenges

- **Challenge 1: Survival Prediction**
  - Task: Predicting survival outcomes of breast cancer patients.
  - Objective: Develop models to estimate survival time or predict survival status based on patient characteristics.
- **Challenge 2: Cox Regression**
  - Train a Cox Regression model for breast cancer survival.
  - Assess the impact of various covariates (e.g., age, tumor stage, treatment types) on survival time.

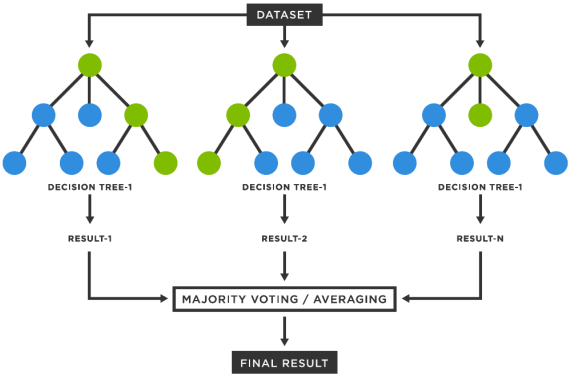
Decision Trees

- Non-parametric, supervised
- Classification, regression
- Splitting criteria (e.g. Gini impurity, variance reduction)
- **highly interpretable**



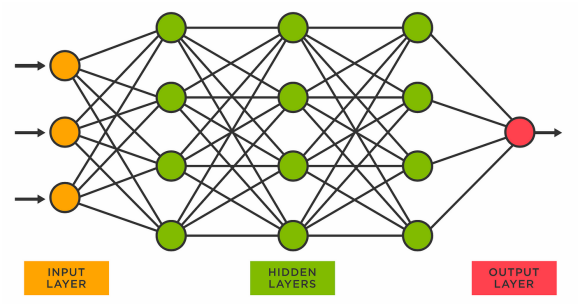
Random Forest

- Ensemble of decision trees
- Bagging: build multiple subsets of original dataset and use to train trees
- Random subspace: use random subset of features for each tree
- Voting/averaging to get final output
- Advantages: less prone to overfitting, better performance



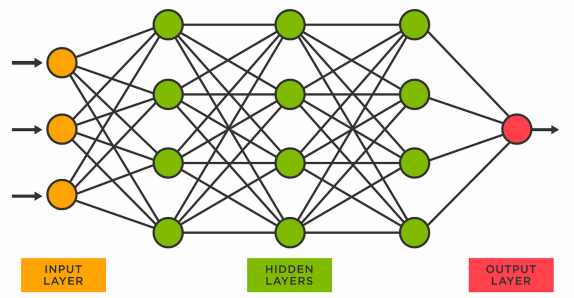
Neural Networks

- 3 layer types: input, hidden, output
  (and a whole lot of other types)
- Black box: what happens in hidden layers?
  $\rightarrow$ bad interpretability



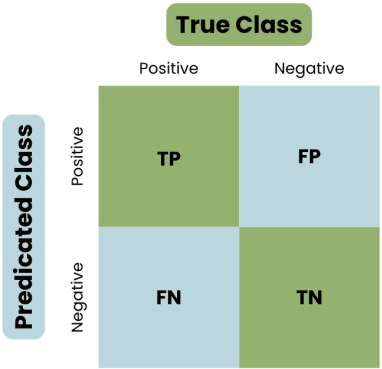
Neural Networks

- Nodes have weight
- Optimizer updates weights of nodes
- Goal: minimize loss function $\rightarrow$ good performance



How to measure performance?
(for classification problems)

- **Accuracy**: number of correct predictions among all predictions
- Confusion matrix
- Precision: $\frac{TP}{TP+FP}$
- Recall: $\frac{TP}{TP+FN}$
- F1 score: $\frac{2}{\frac{1}{precision}+\frac{1}{recall}}$
- AUROC



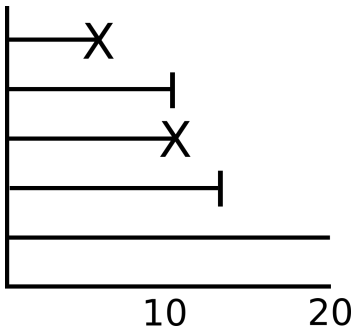
Time-to-Event

- Often interested in the time until something happens.
  - How long will a cancer patient survive?
  - What's the probability of survival over 12 months?
  - When will I finish my PhD?
- The "sth" is typically called **event**.
- Almost always, data to investigate these questions is **censored**.

Censored Data

- **right censored**: don't know what happened **after** specific point in time
- **left censored**: don't know what happened **before** specific point in time
- **interval censored**: don't know **when exactly** the event happened

ID	Time	Event
1	5	1
2	10	0
3	10	1
4	13	0
5	20	0

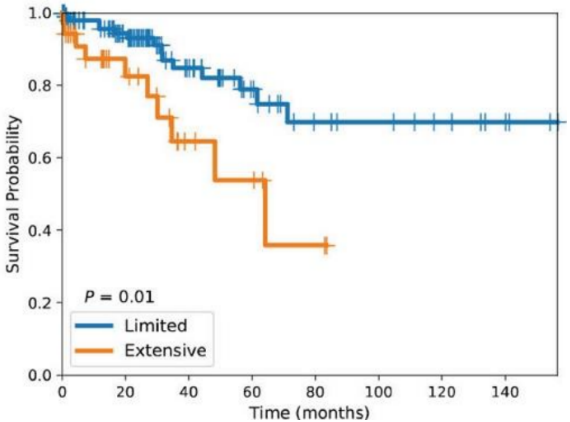


- **Survival function**: $S(t) = P(T > t)$, with T being a continuous random variable, ususally time, and S(t) denoting the proportion of oberservations without event, yet.
- **Kaplan-Meier Estimate**:

$$\hat{S}(t) = \prod_{t' \leq t} (1 - \frac{f_{t'}}{r_{t'}})$$

with $f_{t'}$ and $r_{t'}$ being # individuals failing or at risk at t', respectively.

Kaplan-Meier Curves and Estimates



Stögbauer *et al.*, submitted

Cox Regression

- Association between the survival time of patients and predictor variables.
- Indicating how much a unit change in the covariate affects the hazard of the event.

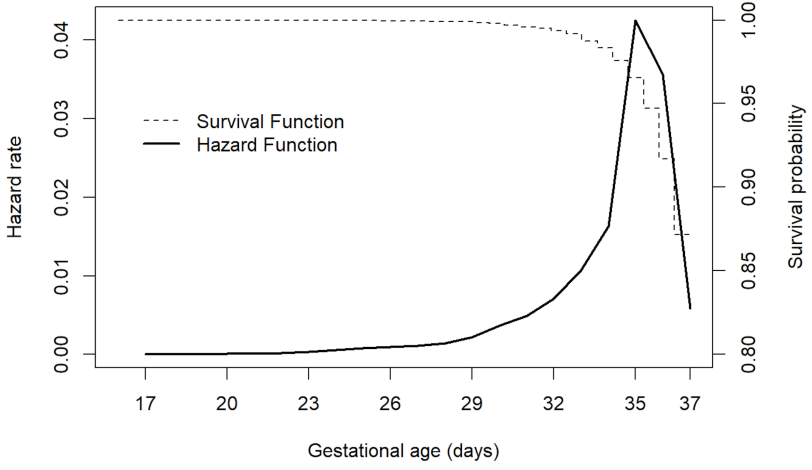
Hazard function

- **Hazard function**: $h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$
- **Cox regression** (a.k.a. *proportional hazards regression*) model has the form

$$h_z(t) = h_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p) \tag{1}$$

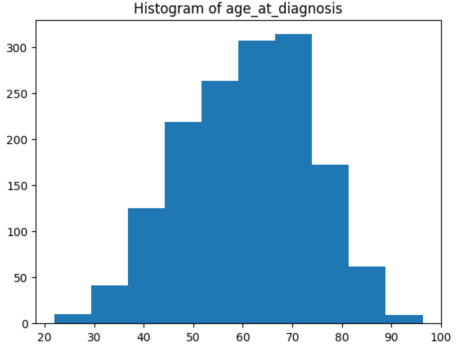
- Coefficients determined by maximum likelihood approach
- Hazard and survival function determine each other.

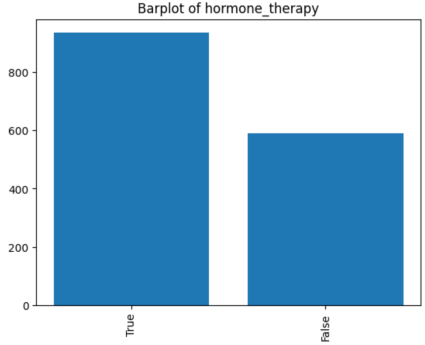
Survival function vs Hazard function



Dataset

- Contains 1522 patients
- Survival rate: 32.6 % of the patients died from cancer.
- Patients that died from other causes are considered as survivors.





Dataset

- clone the repository at: `https://rb.gy/1b09c3`
- In the files x_train.csv and y_train.csv, you can find the data we are going to work with.
- data contains 22 features, like age at diagnosis, tumor_stage or if they received chemotherapy.
- death_from_cancer gives the survival status.
- **Challenge 1:** In the file x_test.csv, you find test data without a label. Predict the label, and send it us. We will evaluate your predictions and elect the winner

Wrap-up

- The winner is presenting their analysis, with an open discussion.

References & Additional Reading

- Some material adopted from
  - https://www2.stat.duke.edu/courses/Spring21/sta102.001/slides/lecture-20.pdf
  - https://socialsciences.mcmaster.ca/jfox/Courses/soc761/survival-analysis.pdf

Result submission

Name your .csv file: $< name > .csv$
e.g.: lisa.csv

Send results to **lisa-marie.bente@plri.de**

Results

Rank	Name	Accuracy [%]
1		
2		
3		
4		
5		
6		
7	Benchmark	75.00

Results

Rank	Name	Accuracy [%]
1		
2		
3		
4		
5		
6	Sadat	77.43
7	Benchmark	75.00

Results

Rank	Name	Accuracy [%]
1		
2		
3		
4		
5	tim (chatGPT)	79.53
6	Sadat	77.43
7	Benchmark	75.00

Results

Rank	Name	Accuracy [%]
1		
2		
3		
4	Tegegne	80.32
5	tim (chatGPT)	79.53
6	Sadat	77.43
7	Benchmark	75.00

Results

Rank	Name	Accuracy [%]
1		
2		
3	Anna	81.10
4	Tegegne	80.32
5	Tim (chatGPT)	79.53
6	Sadat	77.43
7	Benchmark	75.00

Results

Rank	Name	Accuracy [%]
1		
2	Corinna	82.15
3	Anna	81.10
4	Tegegne	80.32
5	Tim (chatGPT)	79.53
6	Sadat	77.43
7	Benchmark	75.00

Results

Rank	Name	Accuracy [%]
1	Luca	85.04
2	Corinna	82.15
3	Anna	81.10
4	Tegegne	80.32
5	Tim (chatGPT)	79.53
6	Sadat	77.43
7	Benchmark	75.00