Survival_tree

Runze Cui

2023-11-27

Introduction:

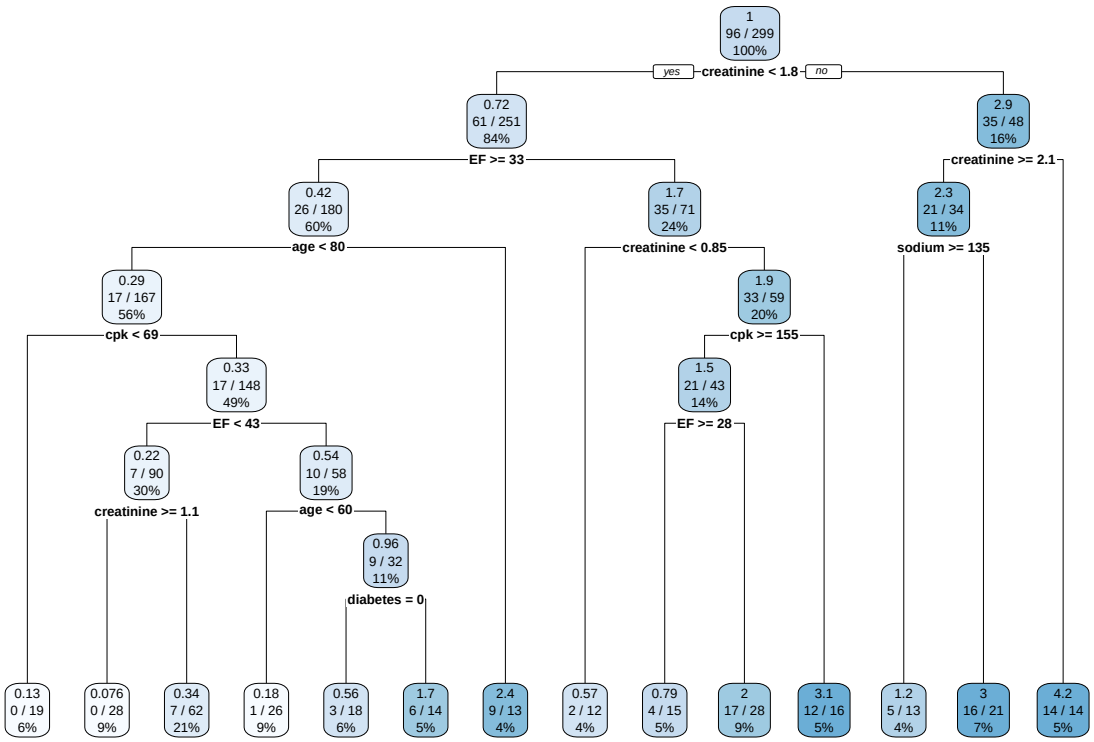
Survival trees are a specialized form of decision trees that are used for analyzing time-to-event data, commonly known as survival analysis. Unlike standard regression trees that predict a numeric outcome, survival trees incorporate censored data and use it to predict the time until an event occurs. Censored data refers to incomplete information about the survival time, which is a common aspect in survival analysis.

Survival trees partition data by the values of predictor variables. They create branches that lead to nodes representing subgroups with similar survival characteristics. This recursive partitioning continues until it reaches a stopping criterion, typically a minimum node size or a maximum tree depth.

```
surv_obj <- Surv(time = dat$time, event = dat$event)
surv_tree <- rpart(surv_obj ~ gender + smoking + diabetes + bp + anaemia + age + EF + sodium + creatini
                   data = dat, method = "exp")
```

Code:

The R code provided fits an exponential survival tree using the rpart package. The model is specified with a range of predictor variables, which include gender, smoking status, diabetes, blood pressure status, presence of anaemia, age, ejection fraction (EF), sodium levels, creatinine levels, platelet count, and creatine phosphokinase (cpk) levels. The method="exp" indicates that an exponential model is used for the survival tree.



Explanations and Conclusions of the Survival Tree Plot:

The survival tree plot visually represents the model fitted on the dataset. The tree is composed of nodes and branches:

- Each internal node represents a binary decision based on one of the predictor variables.
- The branches represent the outcome of the decision, leading to two child nodes.
- Each leaf node (also called a terminal node) provides survival information for the observations in that subgroup.

In the provided plot:

- The root node splits on the creatinine variable, dividing the dataset into two groups based on a creatinine level of 1.8.
- Subsequent splits are made based on other important variables such as EF, age, sodium, and so on.
- Each node displays the survival probability, the number of subjects in the node, the percentage of subjects that have the event, and the predictor variable used for the split.
- The terminal nodes provide the estimated survival distribution for the subjects in that node.
- The exponential survival tree helps identify the most significant predictors of survival time and how these variables interact to define risk subgroups within the population.

Interpreting the Results:

The interpretation of the survival tree involves understanding how the splits correspond to the survival probabilities. For instance, patients with creatinine levels above or below 1.8 have different survival profiles, and this distinction is further refined by subsequent splits in the tree.

It is important to note that survival trees can handle complex interactions between variables and can reveal non-linear relationships. They are a valuable tool for exploratory analysis and can be used to inform more complex survival models.

Note on Model Validity:

It's crucial to validate the survival tree using techniques like cross-validation or an independent dataset to ensure its predictive performance. Overfitting can occur with decision trees, so pruning and validation steps are often necessary to ensure the model's generalizability.

In summary, the survival tree provides a visual and quantitative method for understanding the factors that influence survival time. It allows clinicians and researchers to identify high-risk subgroups and factors that may contribute to patient outcomes, informing decisions for personalized treatment strategies.