A Comparison Study of Random Survival Forest and Cox Proportional Hazards for Predicting the Survival Risk of 76-gene prognostic signature from breast cancer

Project Report submitted to the

UNIVERSITY OF MADRAS

in partial fulfilment of the requirement
For the award of the degree of

MASTER OF SCIENCE
IN
STATISTICS

by

KRITHIKA DEVI C (32820006)

Under the guidance of

Dr. M. RAMADURAI
Assistant Professor

DEPARTMENT OF STATISTICS
UNIVERSITY OF MADRAS
CHENNAI-600005

JUNE 2022

DEPARTMENT OF STATISTICS
UNIVERSITY OF MADRAS
CHENNAI-600005

CERTIFICATE

This is to certify that the Project Report entitled **“A Comparison Study of Random Survival Forest and Cox Proportional Hazards for Predicting the Survival Risk of 76-gene prognostic signature from breast cancer”** submitted in partial fulfilment of the requirement of the award of degree of **MASTER OF SCIENCE** in **STATISTICS** is a bonafide record of work done by **KRITHIKA DEVI C (32820006)** under the guidance of **Dr. M. RAMADURAI** during the academic year of 2020-2022, in the Department of Statistics , University of Madras, Chennai – 600 005.

Dr. M. RAMADURAI
Associate Professor and Head (I/c)
Department of Statistics,
University of Madras

Dr. M. R. SINDHUMOL
Assistant professor
Department of Statistics,
University of madras

Place: Chennai

Date: 17.06.2022

ACKNOWLEDGEMENT

We are profoundly grateful and sincerely thankful to our guide **Dr. M. RAMADURAI,** Assistant Professor, Department of statistics, University of Madras for his unstinted support and valuable guidance throughout this project and encouraging us in times of discouragement and failures for the successful completion of this project.

We express our gratitude to **Dr. M. R. SINDHUMOL,** Associate Professor and Head (I/c), University of Madras, for providing us with all necessary facilities.

We wish to thanks **Dr. S. SURESH,** Assistant Professor, Department of statistics, University of Madras, for his assistance during the time of our project work.

We thank all the non-teaching staff of Department of Statistics for their timely help during our course.

We would also like to thank all the guest faculties and non-teaching staffs for their help and support.

JUNE 2022

KRITHIKADEVI C

CONTENTS

S.No.	INDEX	PAGE No.
	CHAPTER 1	
1.1	Introduction	1
1.2	Basic Concepts of Survival Analysis	2
1.3	Kaplan –Meir Estimate	4
	CHAPTER 2	
2.1	Cox Proportional Hazard Model	5
2.2	Cox Partial Likelihood	6
2.3	Random Survival Forest	7
2.4	RSF Log-rank Splitting rule	9
2.5	Ensemble CHF and Survival Function	10
2.6	Variable Importance in RSF	11
2.7	Prediction Performance	11
2.8	Evaluation Metric	13
	CHAPTER 3	
3.1	Overview of Cancer	14
3.2	Genetic factors to Cancer	15
3.3	Breast Cancer	16
3.4	Estrogen Receptor	19
3.5	Use of Genomic Data	21
	CHAPTER 4	
4.1	Data Description	22
4.2	Brief Overview of Data	23
4.3	Analysis of Breast Cancer Data	25
4.4	Comparison of CPH and RSF on Predictive ability	41
	CHAPTER 5	
5.1	Conclusion	43
	REFERENCES	44
	APPENDIX	45

CHAPTER 1

1.1 INTRODUCTION

This study provides an overview of statistical methods for the analysis of high-dimensional data with time-to-event endpoint analysis. The identification of appropriate covariates and the adequate manipulation of non-linearity and high dimensionality are highly critical to investigate the effect of several important covariates on the event times and the accurate survival prediction. In survival analysis many different regression modelling strategies can be applied to predict the risk of future events. Often, however, the default choice of analysis relies on Cox regression modelling due to its convenience.

The Cox proportional hazards (CPH) model seems to be the first natural step since it is the most commonly used method in modelling event times with the covariates. Its popularity is rooted in its simple interpretation, semi-parametric nature, that is, survival time models that have a fully parametric regression structure but leave their dependence on time unspecified. The CPH model, however, may result in biased parameter estimates if the covariates fail to follow the proportional hazards assumption. Moreover, the assumption is often violated due to the presence of complex relationships in the data structure.

Extensions of the Random Survival Forest (RSF) approach to survival analysis provide an alternative way to build a risk prediction model. This approach is the use of classification and regression trees (CART) that gives a more detailed study of the effects of covariates on the survival distributions and allows for accurate prediction. A RSF bypasses the need to impose parametric or semi-parametric constraints on the underlying distributions and provides a way to automatically deal with higher-order terms in variable that fit a linear combination of trees. The intended approaches of CPH and RSF were compared in terms of prognostic factor (risk factor) detections and prediction performance.

Several measures can be used to assess the resulting probabilistic risk predictions. Most popular is the rank statistics like the concordance index which equals the area under the ROC curve (AUC) for binary responses. For event time outcome in survival analysis these measures can be estimated pointwise over time, where pioneering work was done by Heagerty, Lumley, and Pepe (2000) for time-dependent ROC analysis. Performance curves are obtained by combining time pointwise measures. In this article we concentrate on prediction error curves that are time dependent estimates of AUC.

In this thesis, the CPH and RSF models are compared and contrasted using a benchmark breast cancer dataset that has 76-gene prognostic signature able to predict distant metastases in lymph node-negative (N-) patients. This contains about three-fourth of women who are 40 or older when they are diagnosed with an invasive breast cancer. The expression level of 76 genes has an estrogen receptor which helps in the progression of breast cancer and also has the report of tumor size and tumor grade on the survival of breast cancer.

1.2 Basic Concepts of Survival Analysis

1.2.1 The survival and hazard functions

The survival function S(x) is the probability that an individual survives to time x and defined as S(x) = P (X > x). A non-parametric method to estimate S(x) was provided by Kaplan and Meier (1958). The product-limit estimator, as an estimator of S(x), is defined as

$$\hat{s}(\mathrm{t})=\begin{cases} 1 & , if\ t < t_1 \\ \prod_{t_i \le t}\left[1-\frac{d_i}{n_i}\right] & , \ if\ t \ge t_1 \end{cases} \qquad (2.1)$$

where $d_i$and n_i are respectively the number of events and the number of risk at time point, $i = 1,2,\ \ldots, n$. The Kaplan-Meier estimator is a stepwise function, with jumps at event times. It takes into account censoring and assumes this to be independent of the survival.

The variance of the product-limit estimator is given by the Greenwood's formula

$$\hat{V}\left[\hat{S}(t)\right] = \hat{S}(\mathrm{t})^2 \sum_{i \le t} \frac{d_i}{n_i(n_i-d_i)} \qquad (2.2)$$

The steepness of the survival function is determined by the hazard function. Let T be the random variable representing time until the event of interest occurs. The distribution of T can be characterized by the hazard function, that is, the risk of an individual is the probability who is alive at t experiences the event in the next instant of time period δt given by

$$h(t) = \lim_{\delta t \to 0}\left(\frac{\Pr(t<T<(t+\delta t)|T \ge t)}{\delta t}\right) \qquad (2.3)$$

When hazards are large, many events occur and the survival curve will quickly decrease. The cumulative hazard function H(t) is defined as

$$H(t) = \int_0^t h(t)dt = \ -\ln[S(t)] \qquad (2.4)$$

Since the cumulative hazard H(t) is related to the survival function, it may be estimated from the product-limit estimator as follows

$$\widehat{H}(t) = -\ln[\hat{S}(t)] \qquad (2.5)$$

Alternatively, the Nelson-Aalen estimator can be used

$$\widehat{H}(t) = \begin{cases} 0, & if\ t < t_1 \\ \sum_{t_i \leq t} \frac{d_i}{n_i}, & if\ \ t \geq t_1 \end{cases} \qquad (2.6)$$

1.2.2 Basic Quantities including covariates:

The main goal in many applications is to determine the influence of a (set of) covariates in the survival, for example to compare treatments or establish prognostic factors. In general, this is performed using regression models on the hazard function. For each subject $i \epsilon\ (i = 1,\ \ldots,\ n)$ in the study, assume the observed data of patient i consists of the tuple $(t_i,\ \delta_i)$, the vector of covariates $x_i = (x_{i1},\ \ldots, x_{ip})' \in R^p$, and the subgroup membership $s_i \in \{1, \ldots, S\}$ with S the number of subgroups in the complete data set. The observations , where $t_i = \min(T_i,\ C_i)$ denotes the observed time of patient i, with T_i the event time and C_i the censoring time and let $\delta_i = \mathbb{1}\ (T_i \leq C_i)$ be a binary indicator that indicates whether a patient experienced an event (δ_i =1) or was (right-) censored ($\delta_i = 0$). The covariates may consist of clinical variables such as age, sex, tumor grade or tumor size and a potentially large set of genomic variables.

1.3 Kaplan –Meier Estimate:

The Kaplan –Meier survival curve is defined as the probability of surviving in a given length of time while considering time in many small intervals. The Kaplan –Meier estimate is also called as “product limit estimate”.

Kaplan-Meier estimate is one of the best options to be used to measure the fraction of subjects living for a certain amount of time after treatment. In clinical trials or community trials, the effect of an intervention is assessed by measuring the number of subjects survived or saved after that intervention over a period of time. The time starting from a defined point to the occurrence of a given event, for example death is called as survival time and the analysis of group data as survival analysis.

This can be affected by subjects under study that are uncooperative and refused to be remained in the study or when some of the subjects may not experience the event or death before the end of the study, although they would have experienced or died if observation continued, or we lose touch with them midway in the study. We label these situations as censored observations. The Kaplan-Meier estimate is the simplest way of computing the survival over time in spite of all these difficulties associated with subjects or situations. The survival curve can be created assuming various situations. It involves computing of probabilities of occurrence of event at a certain point of time and multiplying these successive probabilities by any earlier computed probabilities to get the final estimate. This can be calculated for two groups of subjects and also their statistical difference in the survivals. This can be used in Ayurveda research when they are comparing two drugs and looking for survival of subjects.

CHAPTER 2

2.1 Cox Proportional Hazard Model

The most popular regression model used in survival analysis is the Cox proportional hazards model developed by Cox (1972) investigates the covariate effects on survival time which takes into account the effect of censored observations. When subjects have not experienced the event of interest at the end of the study, the exact survival times of these subjects are unknown and these are called censored observations. It models the hazard rate $h(t|x_i)$of an individual at time t and consists of terms, the non-parametric baseline hazard rate $h_0(t)$ and a parametric form of the covariate effects.

$$h(t|x_i) = h_0(t)\, exp(\beta' x_i) = h_0(t) \exp(\textstyle\sum_{j=1}^{p} \beta_j x_{ij}) \qquad (2.1)$$

where t denote time until the event of interest, x_i is a row vector of covariates for individual i of dimension p and $\beta = (\beta_1, \beta_2 \dots, \beta_p)'$, the unknown parameter vector that represents the strength of influence of the covariates on the hazard rate. In matrix notation, $\beta' x_i$ is the linear component of the model which is also known as the 'risk score' or 'prognostic index' for the i^{th} individual.

The general proportional hazards model becomes

$$h(t|x_i) = \exp\left(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}\right) h_0(t)$$

$$h(t|x_i) = \exp\left(\textstyle\sum_{j=1}^{p} \beta_j x_{ij}\right) h_0(t) \qquad (2.2)$$

Thus, the linear model for the logarithm of the hazard ratio is

$$log\left\{\frac{h(t|x_i)}{h_0(t)}\right\} = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} = \sum_{j=1}^{p} \beta_j x_{ij}$$

The regression coefficients β_i are traditionally estimated via maximum likelihood based on the partial likelihood. Cox showed that the relevant partial likelihood function for the model in Equation (2.2) is given by

$$L(\beta) = \prod_{j=1}^{r} \frac{h(t_{(j)}|x_{(j)})}{\sum_{\ell \in R_j} h(t_{(j)}|x_\ell)} = \prod_{j=1}^{r} \frac{\exp(\sum_{i=1}^{p} \beta_i x_{(j)i})}{\sum_{\ell \in R_j} \exp(\sum_{i=1}^{p} \beta_i x_{\ell i})} \qquad (2.3)$$

2.2 Cox Partial Likelihood

The partial log-likelihood is defined as

$$l(\beta) = \ln(L(\beta)) = \sum_{j=1}^{r}\sum_{i=1}^{p}\beta_i x_{(j)i} - \sum_{j=1}^{r}\ln\left[\sum_{\ell\in R_j}\exp\left(\sum_{i=1}^{p}\beta_i x_{\ell i}\right)\right]$$

Here, we suppose data are available for n patients, amongst whom there are r distinct death times and $(n - r)$ right-censored survival times. We view the Risk model by ignoring $h_0(t)$ (as $h_0(t)$ is the same for all the patients). The r ordered death times are denoted by $t_{(1)} < t_{(2)} < \cdots < t_{(r)}$. $x_{(j)}$ is the vector of covariates for a patient who dies at the j^{th} ordered death time, $t_{(j)}$ and $R_j = \{\ell : t_\ell \geq t_{(j)}\}, j = 1,2,\ldots,r$, denote the set of patients who are "at risk" for failure at time to $t_{(j)}$, called the risk set. x_ℓ is vector of p covariates for the individual who is in the risk set.

Individuals for whom the survival times are censored do not contribute to the numerator of the log-likelihood function, but they do enter into the summation over the risk sets at death times that occur before a censored time.

In presence of tied observations, there are two popular methods suggested by Breslow (1974) and Efron (1977).

The partial likelihood function, suggested by Breslow is defined as

$$L_B(\beta) = \prod_{j=1}^{r}\frac{\exp(\beta'\, s_j)}{\left\{\sum_{\ell\in R_j}\exp(\beta'\, x_\ell)\right\}^{d_j}} \qquad (2.4)$$

Where, s_j is the vector of sums of each of the p covariates for those individuals who die at the j^{th} death time, $t_{(j)}, j = 1,2,\ldots,r$. If there are d_j deaths at $t_{(j)}$, the h^{th}element of s_j is S_{hj} = $\sum_{k=1}^{d_j} X_{hjk}$, where X_{hjk} is the value of the h^{th} covariate, $h = 1,2,\ldots,p$, for the r^{th} of d_j individuals, r = 1,2,... d_j who die at the j^{th} death time, j = ,1 2, ..., r.

When the number of ties is large, the partial likelihood proposed by Erfon is more precise

$$L_E(\beta) = \prod_{j=1}^{r}\frac{\exp(\beta'\, s_j)}{\prod_{g=1}^{d_j}\left\{\sum_{\ell\in R_j}\exp(\beta'\, x_\ell) - \frac{g-1}{d_j}\right\}^{d_j}} \qquad (2.5)$$

Coefficient vectors of the covariates are estimated using a maximum likelihood procedure and Maximum Likelihood estimates are obtained by maximizing a (partial) likelihood function. The problem of maximizing $l(\beta)$ can be solved numerically using a Newton-Raphson technique.

Since the Cox regression model relies on the PH assumption, it is very important to verify that covariates satisfy the assumption of proportionality. The assessment of PH assumption can be done by many numerical or graphical approaches. None of these approaches is known to be superior in finding out non-proportionality. Interpreting graphical plots can be arbitrary. The conclusions are highly dependent on the subjectivity of the researcher.

2.3 Random Survival Forest

In survival analysis many different regression modeling strategies can be applied to predict the risk of future events. Often, however, the default choice of analysis relies on Cox regression modeling due to its convenience. So, an extension of the random forest approach to survival analysis is useful in building a risk prediction model which is a non-parametric machine learning strategy. In this application, it is of interest to compare the predictive accuracies of Cox regression to random forest for building a risk prediction model. Several measures can be used to assess the resulting probabilistic risk predictions.

If a risk prediction model fits well over the training data used to build the model, and has good prediction accuracy (assessed using the training data), we would like to know if it continues to predict well over independent validation data and what that prediction accuracy is. Various data splitting algorithms have been proposed, based on bootstrap, to correctly estimate the prediction accuracy of a model in the typical situation where a single data set has to be used to build the prediction models and again to estimate the prediction performance. Early applications of random forests (RF) focused on regression and classification problems. Random survival forests (RSF) does not make the proportional hazards assumption and has the flexibility to model survivor curves that are of dissimilar shapes for contrasting groups of subjects. We applied both techniques to a number of publicly available datasets and compared their fits using prediction error curves and the concordance index. was introduced to extend RF to the setting of right-censored survival data. Implementation of RSF follows the same general principles as RF: (a) Survival trees are grown using bootstrapped data; (b) Random feature selection is used when splitting tree nodes; (c) Trees are generally grown deeply, and (d) The survival forest ensemble is calculated by averaging terminal node statistics (TNS).

The general strategy of random forest is as follows:

Step 1. Draw *B* bootstrap samples.

Step 2. Grow a survival tree based on the data of each of the bootstrap samples $b = 1, \dots, B$:

a) At each tree node select a subset of the predictor variables.

b) Among all binary splits defined by the predictor variables selected in a), find the best split into two subsets (the daughter nodes) according to a suitable criterion for right censored data, like the log-rank test.

c) Repeat a) - b) recursively on each daughter node until a stopping criterion is met.

Step 3. Aggregate information from the terminal nodes (nodes with no further split) from the *B* survival trees to obtain a risk prediction ensemble.

When **training**, each tree in a random forest learns from a random sample of the data points. The individual trees are constructed based on *bootstrap samples (sampling without replacement)*, which means that some samples will be used multiple times in a single tree and for the split selection at each node of a tree a random subset of $q^* \leq q$ covariates is chosen. When **testing**, predictions are made by averaging the predictions of each decision tree. This procedure of training each individual learner on different bootstrapped subsets of the data and then averaging the predictions is known as *bagging*, short for *bootstrap aggregating*.

The presence of censoring is a unique feature of survival data that complicates certain aspects of implementing RSF compared to RF for regression and classification. In right-censored survival data, the observed data is $(\boldsymbol{T}, \boldsymbol{\delta})$ where $\boldsymbol{T}$ is time and $\boldsymbol{\delta}$ is the censoring indicator. The observed time $\boldsymbol{T}$ is defined as the minimum of the true (potentially unobserved) survival event time $\boldsymbol{T^o}$ and the true (potentially unobserved) censoring time $\boldsymbol{C^o}$; thus $\boldsymbol{T} = \boldsymbol{min}(\boldsymbol{T^o}, \boldsymbol{C^o})$ and the actual event time might not be observed. The censoring indicator is defined as $\boldsymbol{\delta} = \mathbb{I}\{\boldsymbol{T^o} \leq \boldsymbol{C^o}\}$. Whe $\boldsymbol{\delta} = \boldsymbol{1},$ an event has occurred (i.e., death has occurred) and we observe the true event time, $\boldsymbol{T} = \boldsymbol{T^o}$. Otherwise when $\boldsymbol{\delta} = \boldsymbol{0},$ the observation is censored and we only observe the censoring time, $\boldsymbol{C} = \boldsymbol{C^o}$: thus we know that the subject has survived to time$\boldsymbol{C^o}$, but not when the subject actually dies.

We denote the data by $(\boldsymbol{T_1}, \boldsymbol{X_1}, \boldsymbol{\delta_1}), \dots, (\boldsymbol{T_n}, \boldsymbol{X_n}, \boldsymbol{\delta_n})$ where, $\boldsymbol{X_i}$ is the feature vector (covariate) for individual $\boldsymbol{i}$ and $\boldsymbol{T_i}$ and $\boldsymbol{\delta_i}$ are the observed time and censoring indicators for $\boldsymbol{i}$. RSF trees just like RF trees are grown using resampling (for example by using the bootstrap; the default is to use sampling without replacement). However for notational simplicity, we will sometimes ignore this distinction.

2.4 RSF Log-rank Splitting rule

The true event time being subject to censoring must be dealt with when growing a RSF tree. In particular, the splitting rule for growing the tree must specifically account for censoring. Thus, the goal is to split the tree node into left and right daughters with dissimilar event history (survival) behavior.

The default splitting rule used by the package is the log-rank test statistic and is specified by split rule="log rank". The log-rank test has traditionally been used for two-sample testing with survival data, but it can be used for survival splitting as a means for maximizing between-node survival differences.

To explain log-rank splitting, consider a specific tree node to be split. Without loss of generality let us assume this is the root node (top of the tree). For simplicity assume the data is not bootstrapped, thus the root node data is $(T_1, X_1, \delta_1), \ldots, (T_n, X_n, \delta_n)$. Let X denote a specific variable (i.e., one of the coordinates of the feature vector). A proposed split using X is of the form $X \leq c$ and $X > c$ (for simplicity we assume X is nominal) and splits the node into left and right daughters, $L = \{X_i \leq c\}$ and $R = \{X_i > c\}$ respectively. Let

$$t_1 < t_2 < \cdots < t_m$$

be the distinct death times and let $d_{j,L}$, $d_{j,R}$ and $Y_{j,L}$, $Y_{j,R}$ equal the number of deaths and individuals at risk at time t_j in daughter nodes L, R. At risk means the number of individuals in a daughter who are alive at time t_j, or who have an event (death) at time t_j:

$$Y_{j,L} = \#\{T_i \leq t_j, X_i \leq c\}, \qquad Y_{j,R} = \#\{T_i \geq t_j, X_i > c\}.$$

Define

$$Y_j = Y_{j,L} + Y_{j,R}, \qquad d_j = d_{j,L} + d_{j,R}.$$

The log-rank split-statistic value for the split is

$$L(X, c) = \frac{\sum_{j=1}^{m}(d_{j,L} - Y_{j,L}\frac{d_j}{Y_j})}{\sqrt{\sum_{j=1}^{m}\frac{Y_{j,L}}{Y_j}\left(1 - \frac{Y_{j,L}}{Y_j}\right)\left(\frac{Y_j - d_j}{Y_j - 1}\right)d_j}}$$

The value $|\boldsymbol{L}(\boldsymbol{X},\boldsymbol{c})|$ is a measure of node separation. The larger the value, the greater the survival difference between $\boldsymbol{L}$ and $\boldsymbol{R}$, and the better the split is. The best split is determined by finding the feature $\boldsymbol{X}^*$ and split-value $\boldsymbol{c}^*$ such that $|\boldsymbol{L}(\boldsymbol{X}^*,\boldsymbol{c}^*)| \geq |\boldsymbol{L}(\boldsymbol{X},\boldsymbol{c})|$ for all $\boldsymbol{X}$ and $\boldsymbol{c}$.

2.5 Ensemble CHF and Survival Function

Once the survival tree is grown, the ends of the tree are called the terminal nodes. The survival tree predictor is defined in terms of the predictor within each terminal node. Let $\boldsymbol{h}$ be a terminal node of the tree and let

$$\boldsymbol{t_{1,h}} < \boldsymbol{t_{2,h}} < \cdots < \boldsymbol{t_{m(h),h}}$$

be the unique death times in $\boldsymbol{h}$ and let $\boldsymbol{d_{j,h}}$ and $\boldsymbol{Y_{j,h}}$ equal the number of deaths and individuals at risk at time $\boldsymbol{t_{j,h}}$. The Cumulative hazard function (CHF) and survival functions for $\boldsymbol{h}$ are estimated using the bootstrapped Nelson-Aalen and Kaplan-Meier estimators,

$$H_h(t) = \sum_{\boldsymbol{t_{j,h \leq t}}} \frac{d_{j,h}}{Y_{j,h}}, \qquad S_h(t) = \prod_{\boldsymbol{t_{j,h \leq t}}} \left(1 - \frac{d_{j,h}}{Y_{j,h}}\right).$$

An in-bag (IB) estimator for CHF and survival function is,

$$H^{IB}(t|X) = H_h(t), \qquad S^{IB}(t|X) = S_h(t), \quad if\ X \in h.$$

To define the Out-of-bag (OOB) estimator, let $I_i \in \{0,1\}$ indicate whether case i is IB or OOB. Let $I_i = 1$ if and only if i is OOB. Drop i down the tree and let h denote i's terminal node. The OOB tree estimators for i is

$$H^{OOB}(t|Xi) = H_h(t), \quad S^{OOB}(t|Xi) = Sh(t),\ if\ Xi \in h\ and\ Ii = 1.$$

Observe these are NULL unless i is OOB for the tree.

The bootstrap ensemble CHF takes the average over the B survival trees such that

$$H_{RF}(t|Xi) = \frac{1}{B}\ H_b(t|Xi),$$

where $\boldsymbol{H_b(t|Xi)}$ indicates the CHF obtained from a tree grown on the bth bootstrap sample.

In a random forest, each of the trees is grown using an independent bootstrap sample from the set of training observations. One-third of observations are not used to construct a tree from the particular bootstrap sample. The remaining observations are referred to as out of bag (OOB) observations and are used to predict the ensemble CHF or ensemble survival function. The resulting OOB predicted ensemble CHF or ensemble survival function are used as valid test set predictions obtained from the random forest model.

2.6 Variable Importance in RSF

The most popular Variable Importance (VIMP) method uses a prediction error approach involving "noising-up" each variable in turn. VIMP for a variable x_v is the difference between prediction error when x_v is randomly permuted, compared to prediction error under the observed values. Since VIMP is the difference in OOB prediction error before and after permutation, a large VIMP value indicates that misspecification detracts from the predictive accuracy in the forest. VIMP close to zero indicates the variable contributes nothing to predictive accuracy, and negative values indicate the predictive accuracy improves when the variable is mis-specified. In the latter case, we assume noise is more informative than the true variable. As such, we ignore variables with negative and near zero values of VIMP, relying on large positive values to indicate that the predictive power of the forest is dependent on those variables. This VIMP is used to select the top most features for getting a better predictive results.

2.7 Prediction Performance

In order to assess the prediction performance of a fitted model, it is important to validate it on independent test data. If the same (training) data are used for learning and evaluating a model, the estimated prediction error will generally be too optimistic and underestimate the true prediction error. This means that the model performs well on the training data but worse on independent test data. The more complex a model becomes, the more training data are used for learning, which makes the model more specific to the training data but less generalizable. This is called overfitting.

When no independent test data are available for validation, resampling can be applied. Resampling is the random (repeated) partitioning of the entire available data into training and test sets. The model is fitted on the training set and predictions are made on the test set.

2.7.1 C-Index

C-Index is the measure of prediction performance for time-to-event data are presented. They are particularly important for the comparison of two or more survival models with regard to their predictive accuracy. Another important criterion for model evaluation besides prediction performance is the stability of variable selection. A desirable property is that the set of selected covariates remains stable across different resampling data sets.

Prediction performance of all Cox models is evaluated by Harrell's C- (concordance), implemented in the R package. The C-index is a measure of predictive discrimination and defined as the proportion of all usable pairs of patients with concordant predicted and observed survival times.

The C-index measures predictive information derived from a set of predictor variables in a model. In predicting the time until death, C is calculated by considering all possible pairs of patients, at least one of whom has died. If the predicted survival time is larger for the patient who lived longer, the predictions for that pair are said to be concordant with the outcomes. If one patient died and the other is known to have survived at least to the survival time of the first, the second patient is assumed to outlive the first. When predicted survivals are identical for a patient pair, rather than 1 is added to the count of concordant pairs in the numerator of C. In this case, one is still added to the denominator of C (such patient pairs are still considered usable). A patient pair is unusable if both patients died at the same time, or if one died and the other is still alive but has not been followed long enough to determine whether she will outlive the one who died.

Let t_i, t_{i^*} be the observed survival times of patients i and i^*, and the corresponding risk scores are $\hat{r}(x_i) = \hat{\beta}' x_i$, $\hat{r}(x_{i^*}) = \hat{\beta}' x_{i^*}$. A pair $(i,\ i^*)$ is considered concordant if $t_i \overset{<}{\underset{>}{-}} t_{i^*} \Leftrightarrow \hat{r}(x_i) \overset{>}{\underset{<}{-}} \hat{r}(x_{i^*})$. The C-index is defined as,

$$CI = \frac{1}{n_c} \sum_{\{i:\delta_i=1\}} \sum_{\{i^*:t_i^*>t_i\}} \left(\mathbb{I}\big(\hat{r}(x_{i^*}) < \hat{r}(x_i)\big) + \frac{1}{2} \mathbb{I}\big(\hat{r}(x_{i^*}) = \hat{r}(x_i)\big) \right)$$

where n_c is the number of comparable pairs $(i,\ i^*)$ that standardizes CI to [0, 1]. A patient pair is considered unusable, if both patients die at the same time, or both patients are censored, or if one is censored before the other one dies. CI ≈ 1 stands for a very good prediction and values around 0.5 suggest a random prediction.

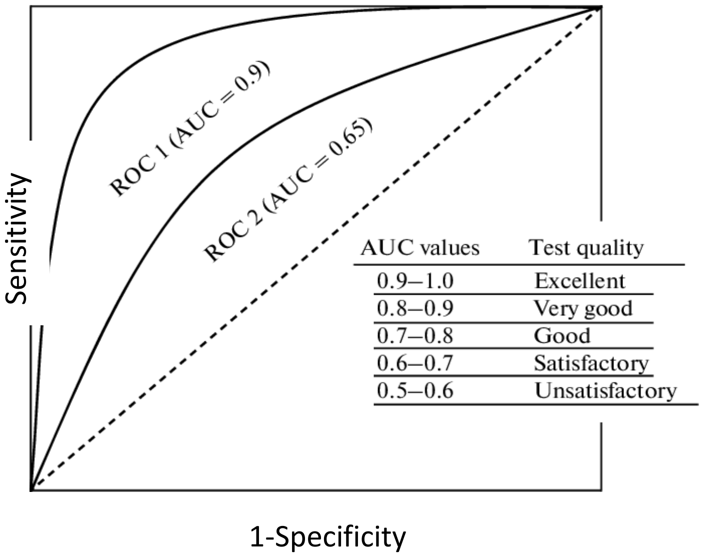
C-index is equal to area under Roc Curve that ranges from 0.5 to 1. Overall Concordance score is actually going to represent the proportion of pairs of individuals who experience the event.

The probability error rate is $PE = 1 - C$. Note that $0 \leq PE \leq 1$ and that $PE = 0.5$ corresponds to a procedure doing no better than random guessing, whereas $PE = 0$ indicates perfect prediction.

2.8 Evaluation Metric

The Receiver Operator Characteristic (ROC) curve is most commonly used to visualize the performance of binary classification problems. It is a probability curve that plots the True Positive Rate (*TPR*) against False Positive Rate (*FPR*) at various threshold values and essentially separates the 'signal' from the 'noise'.

The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.



AUC values	Test quality
0.9–1.0	Excellent
0.8–0.9	Very good
0.7–0.8	Good
0.6–0.7	Satisfactory
0.5–0.6	Unsatisfactory

Figure 2.1: ROC Curve and AUC Statistic measurement quality

$TPR = Sensitivity = \frac{TP}{TP+FN}$; tells us what proportion of the positive class got correctly classified.

$\mathrm{F}PR = 1 - Specificity = \frac{FP}{FP+TN}$; tells us what proportion of the negative class got incorrectly classified

CHAPTER 3

3.1 Over View of Cancer

Cancer causes cells to divide uncontrollably. This can result in tumors, damage to the immune system, and other impairment that can be fatal. It describes the disease that results when cellular changes cause the uncontrolled growth and division of cells. Some types of cancer cause rapid cell growth, while others cause cells to grow and divide at a slower rate.

A cell receives instructions to die so that the body can replace it with a newer cell that functions better. Cancerous cells lack the components that instruct them to stop dividing and to die. As a result, they build up in the body, using oxygen and nutrients that would usually nourish other cells. Cancerous cells can form tumors, impair the immune system and cause other changes that prevent the body from functioning regularly. Cancerous cells may appear in one area, then spread via the lymph nodes. These are clusters of immune cells located throughout the body.

3.1.1 Cancer Therapy

Doctors usually prescribe treatments based on the type of cancer, its stage at diagnosis, and the person's overall health.

Below are examples of approaches to cancer treatment:

- Chemotherapy aims to kill cancerous cells with medications that target rapidly dividing cells. The drugs can also help shrink tumors, but the side effects can be severe.
- Hormone therapy involves taking medications that change how certain hormones work or interfere with the body's ability to produce them. When hormones play a significant role, as with prostate and breast cancers, this is a common approach.
- Immunotherapy uses medications and other treatments to boost the immune system and encourage it to fight cancerous cells. Two examples of these treatments are checkpoint inhibitors and adoptive cell transfer.
- Precision medicine, or personalized medicine, is a newer, developing approach. It involves using genetic testing to determine the best treatments for a person's

particular presentation of cancer. However, Researchers have yet to show that it can effectively treat all types of cancer.

- Radiation therapy uses high-dose radiation to kill cancerous cells. Also, a doctor may recommend using radiation to shrink a tumor before surgery or reduce tumor-related symptoms.
- Stem cell transplant can be especially beneficial for people with blood-related cancers, such as leukemia or lymphoma. It involves removing cells, such as red or white blood cells, that chemotherapy or radiation has destroyed. Lab technicians then strengthen the cells and put them back into the body.
- Surgery is often a part of a treatment plan when a person has a cancerous tumor. Also, a surgeon may remove lymph nodes to reduce or prevent the disease's spread.
- Targeted therapies perform functions within cancerous cells to prevent them from multiplying. They can also boost the immune system. Two examples of these therapies are small-molecule drugs and monoclonal antibodies.

Doctors will often employ more than one type of treatment to maximize effectiveness.

3.2 Genetic factors to Cancer

Genetic factors can contribute to the development of cancer. A person's genetic code tells their cells when to divide and expire. Changes in the genes can lead to faulty instructions, and cancer can result.

Genes also influence the cells production of proteins, and proteins carry many of the instructions for cellular growth and division. Some genes change proteins that would usually repair damaged cells. This can lead to cancer. If a parent has these genes, they may pass on the altered instructions to their offspring. Some genetic changes occur after birth, and factors such as smoking and sun exposure can increase the risk. Other changes that can result in cancer take place in the chemical signals that determine how the body deploys, or "expresses" specific genes. Finally, a person can inherit a predisposition for a type of cancer. A doctor may refer to this as having a hereditary cancer syndrome. Inherited genetic mutations significantly contribute to the development of 5–10 percent trusted Source of cancer cases. Innovative research has fueled the development of new medications and treatment technologies.

3.3 Breast Cancer

Breast cancer is a cancer that starts in breast tissue. It happens when cells in the breast change and grow out of control. The cells usually form a tumor. Sometimes the cancer does not spread any further. If the cancer spreads outside the breast, the cancer is called “invasive”. It may just spread to nearby tissues and lymph nodes. Or the cancer may metastasize (spread to other parts of the body) through the lymph system or the blood. Breast cancer is the second most common type of cancer in women in the United States. Rarely, it can also affect men. There are different types of breast cancer. The types are based on which breast cells turn into cancer. The types include: Ductal carcinoma, which begins in the cells of the ducts. This is the most common type. Lobular carcinoma, which begins in the lobules. It is more often found in both breasts than other types of breast cancer. Inflammatory breast cancer, in which cancer cells block lymph vessels in the skin of the breast. The breast becomes warm, red, and swollen. This is a rare type. Paget's disease of the breast, which is a cancer involving the skin of the nipple. It usually also affects the darker skin around the nipple. It is also rare.

In some cases, cancerous cells can invade surrounding breast tissue. In these cases, the condition is known as invasive breast cancer. Sometimes, tumors spread to other parts of the body. If breast cancer spreads, cancerous cells most often appear in the bones, liver, lungs, or brain. Tumors that begin at one site and then spread to other areas of the body are called metastatic cancers. A small percentage of all breast cancers cluster in families. These cancers are described as hereditary and are associated with inherited gene mutations. Hereditary breast cancers tend to develop earlier in life than non-inherited (sporadic) cases, and new (primary) tumors are more likely to develop in both breasts.

3.3.1 Stages of breast cancer

Breast cancer can be classified into different stages depending on several factors, such as tumor size, lymph node involvement, and metastasis. This method of classification is known as ***TNM system*** and the stages are defined as follows:

Stage I: are small tumors located in the breast, without infiltration to lymph nodes.

Stage II: the tumor size is between 2 and 5 cm, and there may or may not be lymph node involvement.

Stage III: the size of the tumor can be greater than 5 cm and if there is involvement of lymph nodes.

Stage IV: is when metastasis occurs, regardless of tumor size and whether or not lymph nodes are involved.

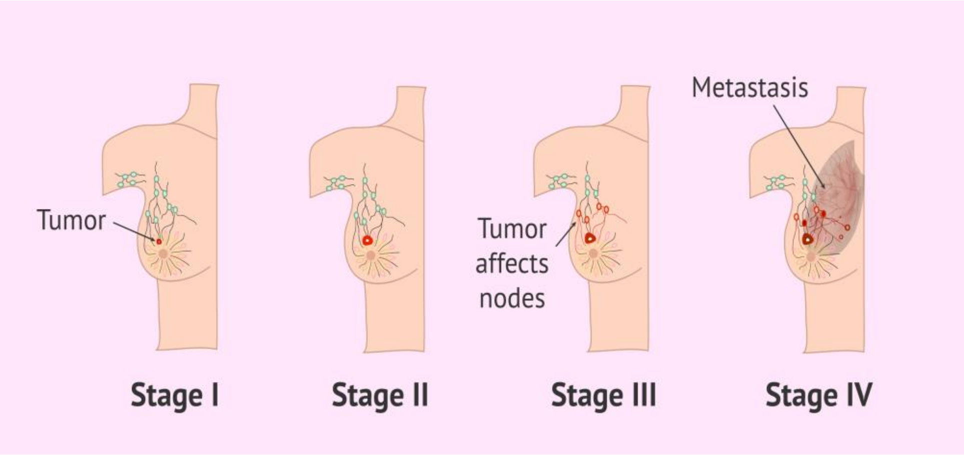


Figure 3.1: Stages of Breast Cancer

3.3.2 Causes of Breast Cancer

Breast cancer happens when there are changes in the genetic material (DNA). Often, the exact cause of these genetic changes is unknown. But sometimes these genetic changes are inherited, meaning that you are born with them. There are also certain genetic changes that can raise your risk of breast cancer, including changes called BRCA1 (BReast CAncer gene 1) and BRCA2 (BReast CAncer gene 2) . These genes produce proteins that help repair damaged DNA. Everyone has two copies of each of these genes—one copy inherited from each parent. *BRCA1* and *BRCA2* are sometimes called tumor suppressor genes because when they have certain changes, called harmful (or pathogenic) variants (or mutations), cancer can develop. These two changes also raise your risk of ovarian and other cancers. Besides genetics, your lifestyle and the environment can affect your risk of breast cancer.

3.3.3 Risk and Symptoms of Breast Cancer

The factors which raise your risk of breast cancer include: Older age History of breast cancer or benign (non-cancer) breast disease Inherited risk of breast cancer, including having BRCA1 and BRCA2 gene changes Dense breast tissue A reproductive history that leads to more exposure to the estrogen hormone, including Menstruating at an early age Being at an older age when you first gave birth or never having given birth Starting menopause at a later age Taking hormone therapy for symptoms of menopause Radiation therapy to the breast or chest Obesity Drinking alcohol.

A new lump or thickening in or near the breast or in the armpit A change in the size or shape of the breast A dimple or puckering in the skin of the breast. It may look like the skin of an orange. A nipple turned inward into the breast Nipple discharge other than breast milk. The discharge might happen suddenly, be bloody, or happen in only one breast. Scaly, red, or swollen skin in the nipple area or the breast Pain in any area of the breast. Having one or more of these symptoms does not mean that a person definitely has breast cancer.

3.3.4 Diagnosis of Breast Cancer

Health care provider may use many tools to diagnose breast cancer and figure out which type you have: A physical exam, including a clinical breast exam (CBE). This involves checking for any lumps or anything else that seems unusual with the breasts and armpits. A medical history Imaging tests, such as a mammogram, an ultrasound, or an MRI Breast biopsy Blood chemistry tests, which measure different substances in the blood, including electrolytes, fats, proteins, glucose (sugar), and enzymes. Some of the specific blood chemistry tests include a basic metabolic panel (BMP), a comprehensive metabolic panel (CMP), and an electrolyte panel. If these tests show that you have breast cancer, you will have tests which study the cancer cells. These tests help your provider decide which treatment would be best for you. The tests may include: Genetic tests for genetic changes such as BRCA and TP53 HER2 test. HER2 is a protein involved with cell growth. It is on the outside of all breast cells. If your breast cancer cells have more HER2 than normal, they can grow more quickly and spread to other parts of the body.

An estrogen and progesterone receptor test. This test measures the amount of estrogen and progesterone receptors in cancer tissue. If there are more receptors than normal, the cancer is called estrogen and/or progesterone receptor positive. This type of breast cancer may grow more quickly. Another step is staging the cancer. Staging involves doing tests to find out whether the cancer has spread within the breast or to other parts of the body. The tests may include other diagnostic imaging tests and a sentinel lymph node biopsy. This biopsy is done to see whether the cancer has spread to the lymph nodes. Treatments for breast cancer include: Surgery such as A mastectomy, which removes the whole breast A lumpectomy to remove the cancer and some normal tissue around it, but not the breast itself Radiation Therapy, Chemotherapy, Hormone therapy, which blocks cancer cells from getting the hormones they need to grow Targeted therapy, which uses drugs or other substances that attack specific cancer cells with less harm to normal cells Immunotherapy.

3.3.5 Cancer Stat Facts: Female Breast Cancer

Estimated New Cases in 2022	287,850
% of All New Cancer Cases	15.0%

Estimated Deaths in 2022	43,250
% of All Cancer Deaths	7.1%

3.4 Estrogen receptor

3.4.1 Estrogen receptor - Positive and Negative Breast Cancer

Breast cancer is a major concern worldwide and is responsible for one of the highest causes of death. Determination of estrogen receptor (ER) status on invasive carcinomas (cancer) prior to therapeutic procedures has become a standard practice in the management of breast cancer and approximately 60–65% of primary breast cancers are ER-positive. ER has also proven to be a successful target for the treatment of ER-positive breast carcinomas; the effectiveness of antiestrogens such as tamoxifen and raloxifene is well known.

Better-differentiated tumors are likely to be ER-positive and these ER-positive tumors have relatively better prognosis. Conversely, ER-negative tumors are more likely to be of higher histological grade, and the patients to have a decreased overall survival depending on age and lymph node status. ER receptor status of breast cancers in postmenopausal women is also associated with survival; a higher recurrence rate is observed in ER-negative group. Although most of the ER-negative tumors are presumed to be poorly differentiated, a significant proportion of a small subset of invasive cancers (adenoid cystic carcinoma, secretory carcinoma) are ER-negative. These tumors have an excellent prognosis with minimal regional recurrence. On the other hand, not all poorly differentiated, ER-negative tumors behave poorly. Medullary and atypical medullary cancers are reported in some series to have a relatively better prognosis than expected. Some ER-negative tumors also show a higher BRCA 1 germline mutation. All these features point towards the heterogeneous nature of an ER-negative subgroup of invasive breast cancers. The purpose of this study was to determine the characteristics of ER-negative breast cancers through analysis of several morphological features and to correlate these features and their immunophenotypical profile with other prognostic variables and clinicopathological data to better understand their biological behavior.

ER-positive breast cancer is the most common type of breast cancer diagnosed today. When breast cancer cells test positive for estrogen receptors, it's called estrogen receptor positive (ER-positive) breast cancer. It means that estrogen is fueling the growth of the cancer. It's one of several important characteristics of breast cancer that help determine the best treatment options. About 67 to 80 percentTrusted Source of breast cancers in women and 90 percent of breast cancers in men are ER-positive, per the National Cancer Institute.

Estrogen and progesterone are the two hormones associated with breast cancer. If the cancer has either or both receptors, it's also known as hormone-positive or HR-positive breast cancer. Breast cancers that test negative for both hormone receptors are HR-negative. When doctor suspects ER-positive breast cancer, it is diagnosed by a biopsy to test for cancerous cells. If there is cancer, your doctor will also test the cells for characteristics that include what receptors, if any, are present on the surface of the cancer cells.

Hormone receptors are proteins located in and around breast cells. When the corresponding hormone binds to a receptor, it tells the cells how to grow and divide. In the case of breast cancer, these receptors allow abnormal cells to grow out of control, which results in a tumor.

3.5 Use of Genomic Data

Genomic data science applies statistics and computer science to the genome. The goal is to understand, analyze, and interpret information from genome sequences.

Gene expression profiling has been used extensively in biological research and has resulted in significant advances in the understanding of the molecular mechanisms of complex disorders, including cancer, heart disease, and metabolic disorders.

The hope is that by analyzing patterns of gene expression (e.g. profiling) scientists will be able to better understand the molecular etiology of multi-factorial disorders such as obesity, diabetes, heart disease, or cancer.

CHAPTER 4

4.1 Data Description

The data has been gathered from the GEO Gene Expression Omnibus from National Centre for Biotechnology Information (National Library of Medicine) which advances science and health by providing access to biomedical and genomic information. The data that was accessed is the strong Time Dependence of the 76-Gene Prognostic Signature that was able to predict distant metastases in lymph node-negative (N-) breast cancer patients. The data contains the expression levels of 76 genes, age, estrogen receptor status (er), tumor size, grade, days and status for 198 individuals. The objective is to predict the time to distant metastasis and probabilistic risk predictions. The Study data consists of 198 patient observations with 82 features. Of these 82 features, 2 features such as 'days' and 'status' were survival dependent variable.

Predictor Variables	Levels/ IQR	Variable name	Total (*n = 198)*
Continuous:			
76-gene expression ID ref	IQR	X200726_at, X200965_s_at, X201068_s_at, X201091_s_at, X201288_at, X201368_at, X201663_s_at, X201664_at, X202239_at, X202240_at, X202418_at, X202687_s_at, X203391_at, X204014_at, X204015_s_at, X204073_s_at, X204218_at, X204540_at, X204631_at, X204740_at, X204768_s_at, X204888_s_at, X205034_at, X205848_at, X217019_at, X217102_at, X217404_s_at, X206295_at, X207118_s_at, X208180_s_at, X208683_at, X209500_x_at, X209524_at, X209835_x_at, X209862_s_at, X210028_s_at, X210314_x_at, X210593_at, X211040_x_at, X211382_s_at, X211762_s_at, X211779_x_at, X212014_x_at, X212567_s_at, X214806_at, X214915_at, X214919_s_at, X215510_at, X215633_x_at, X216010_x_at, X216103_at, X216693_x_at, X219724_s_at, X220886_at, X221028_s_at, X217471_at, X221241_s_at, X217767_at, X221344_at, X217771_at, X221634_at, X217815_at, X221816_s_at, X218430_s_at, X221882_s_at, X218478_s_at, X221916_at, X218533_s_at, X221928_at, X218782_s_at, X218883_s_at, X218914_at, X219340_s_at, X219510_at, X219588_s_at,	8.678166 [14.38958; 0.516942]

X203306_s_at

Age	IQR	age	46 [24; 60]
Tumor Size	IQR	size	2 [0.6; 5]
Period of days taken	IQR	days	4384 [125; 9108]
Survival Status	0/1	status	147/51
Factors:			
Status of Estrogen Receptor	positive/negative	er	134/64
Tumor Grade	poorly differentiated/ intermediate/ well differentiated/ unknown	grade	83/83/30/2

Table 4.1: Shown the interquartile range (IQR); median [Q1; Q3] for continuous variable. The counts are shown for each level in column 4 for factors.

4.2 Brief Overview of data

The entire dataset does not contain non-null values. It has a float data type for 76-gene expression and tumor size. Also, it has an integer data type for remaining covariates except estrogen receptor status and tumor grade which consists of object (Character) data type.

To perform the CPH and RSF model for the estimation of survival prediction, fixed random samples of size 198 were considered and 178 datasets were generated, for each of the models. The datasets were considered for training and evaluation (i.e., 178 for training set and 20 for testing set). Performances of the CPH and RSF methods were compared based on measures of accuracy derived from the test set of 20 samples, for both the model settings.

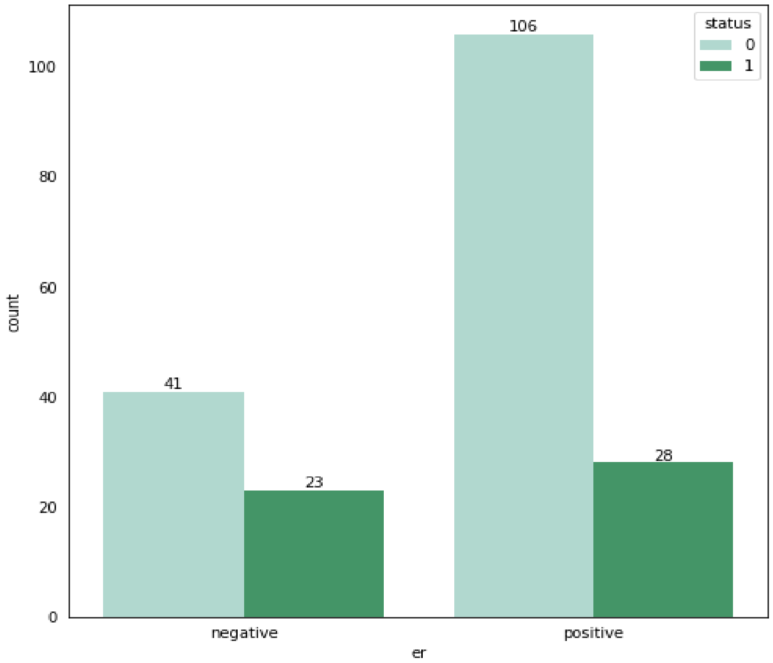


Figure 4.1: Distribution of Status corresponding to its estrogen receptor.

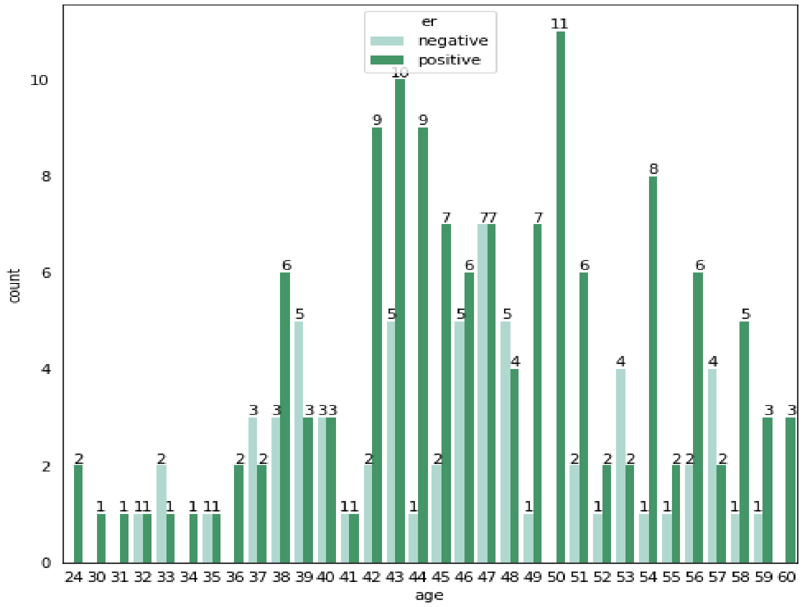


Figure 4.2: Distribution of estrogen positive and negative corresponding to its age.

4.3 ANALYSIS OF BREAST CANCER DATA

4.3.1 Estimation of Kaplan-Meier

The ***survival*** package is the cornerstone of the entire R survival analysis. Not only is the package itself rich in features, but the object created by the ***Surv()*** function, which contains failure time and censoring information, is the basic survival analysis data structure in R.

To begin our analysis, we use the formula ***Surv(days, status) ~ .*** and the ***survfit()*** function to produce the Kaplan-Meier estimates of the probability of survival over time. Using ***ggsurvplot()*** function to plot survival curves.

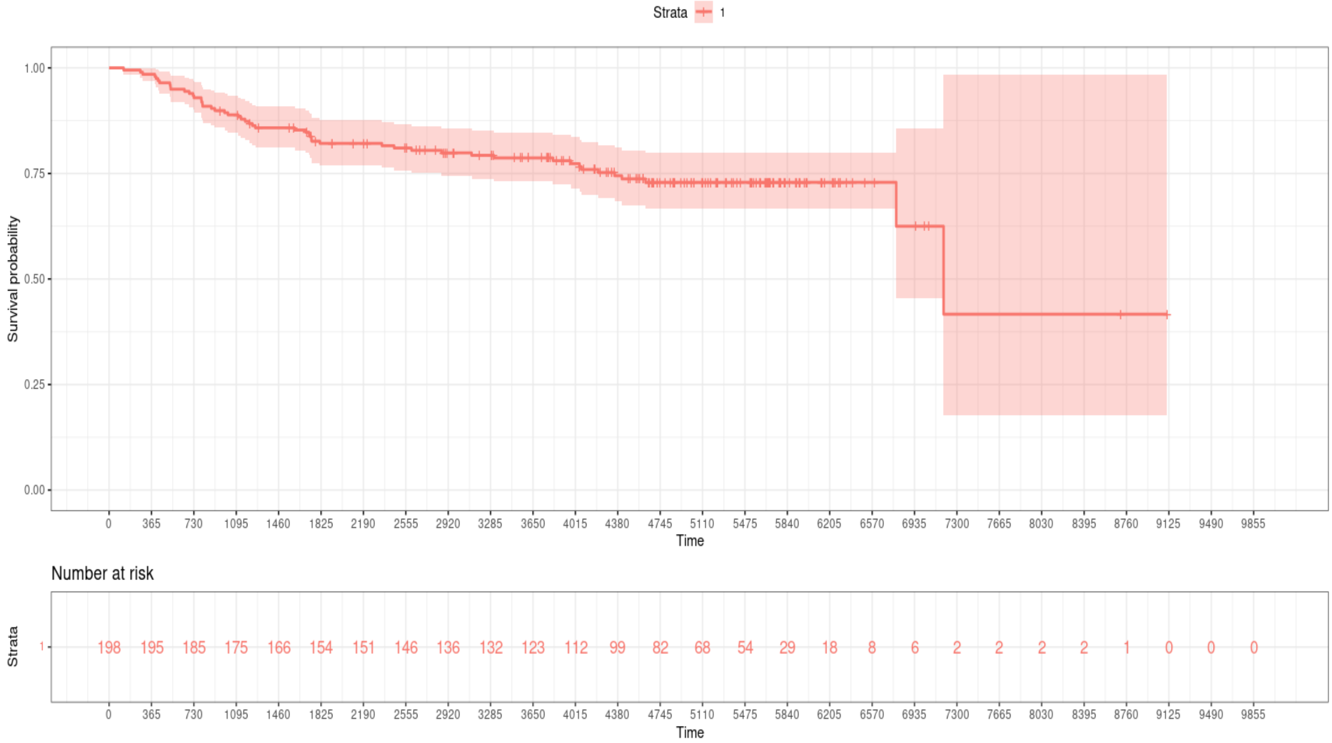


Figure 4.1: Shows the Probability of survival for the patient who is dead ('=1') and number of patients who are at risk for every period of 365 days.

4.3.2 Performance of CPH model

coxph() function, fits a Cox proportional hazards regression model. Here, an applied CPH method is ***'efron'***. The proportional hazards model is usually expressed in terms of a single survival time value for each person, with possible censoring. The following is the ***summary of coxph()*** for 178 training observations that makes use of all covariates in the dataset.

```
                          coef  exp(coef)   se(coef)        z     Pr(>|z|)
X200726_at          -2.433e+02 2.103e-106  3.791e+00  -64.187  < 2e-16 ***
X200965_s_at          1.606e+02   5.337e+69  1.814e+00   88.502  < 2e-16 ***
X201091_s_at        -3.616e+02 8.872e-158  2.979e+00 -121.396  < 2e-16 ***
X201288_at           2.283e+02  1.355e+99  2.293e+00   99.546  < 2e-16 ***
X201368_at           3.407e+01  6.273e+14  2.374e+00   14.355  < 2e-16 ***
X201663_s_at         2.457e+02 5.082e+106  1.593e+00  154.251  < 2e-16 ***
X201664_at           9.558e+01  3.219e+41  1.623e+00   58.897  < 2e-16 ***
X202239_at          -3.918e+02 6.729e-171  3.577e+00 -109.534  < 2e-16 ***
X202240_at           2.308e+02 1.641e+100  2.215e+00  104.169  < 2e-16 ***
X202418_at           3.005e+02 3.089e+130  3.330e+00   90.229  < 2e-16 ***
X202687_s_at        -1.482e+02  4.529e-65  1.435e+00 -103.279  < 2e-16 ***
X203306_s_at        -1.488e+02  2.410e-65  2.892e+00  -51.444  < 2e-16 ***
X203391_at          -4.436e+02 2.250e-193  2.356e+00 -188.296  < 2e-16 ***
X204014_at          -9.180e+01  1.361e-40  6.583e-01 -139.436  < 2e-16 ***
X204015_s_at         7.046e+01  3.972e+30  1.078e+00   65.354  < 2e-16 ***
X204073_s_at         2.171e+01  2.687e+09  1.018e+00   21.324  < 2e-16 ***
X204218_at          -1.038e+02  8.173e-46  4.105e+00  -25.292  < 2e-16 ***
X204540_at           6.423e+01  7.862e+27  7.302e-01   87.960  < 2e-16 ***
X204631_at          -2.962e+01  1.375e-13  1.541e+00  -19.217  < 2e-16 ***
X204740_at           8.687e+00  5.923e+03  2.518e+00    3.449 0.000562 ***
X204768_s_at        -7.534e+01  1.903e-33  2.848e+00  -26.458  < 2e-16 ***
X204888_s_at         9.862e+01  6.782e+42  1.640e+00   60.143  < 2e-16 ***
X205034_at          -4.718e+01  3.249e-21  1.945e+00  -24.257  < 2e-16 ***
X205848_at           1.115e+02  2.535e+48  1.137e+00   98.012  < 2e-16 ***
X206295_at          -6.763e+00  1.156e-03  1.705e+00   -3.967 7.27e-05 ***
X207118_s_at        -3.568e+01  3.183e-16  1.046e+00  -34.111  < 2e-16 ***
X208180_s_at         3.821e+01  3.927e+16  8.676e-01   44.041  < 2e-16 ***
X208683_at          -1.281e+02  2.345e-56  2.723e+00  -47.038  < 2e-16 ***
X209500_x_at         2.041e+02  4.328e+88  2.147e+00   95.067  < 2e-16 ***
X209524_at           3.430e+02 9.590e+148  2.213e+00  155.020  < 2e-16 ***
X209835_x_at        -1.122e+02  1.895e-49  1.510e+00  -74.272  < 2e-16 ***
X209862_s_at        -1.883e+01  6.643e-09  2.464e+00   -7.641 2.15e-14 ***
X210028_s_at         2.684e+02 3.566e+116  1.923e+00  139.573  < 2e-16 ***
X210314_x_at         3.931e+02 5.031e+170  1.979e+00  198.636  < 2e-16 ***
X210593_at          -8.507e+01  1.135e-37  1.651e+00  -51.535  < 2e-16 ***
X211040_x_at        -3.474e+02 1.312e-151  2.849e+00 -121.945  < 2e-16 ***
X211382_s_at        -1.832e+02  2.789e-80  2.533e+00  -72.326  < 2e-16 ***
X211762_s_at         1.929e+02  5.728e+83  2.291e+00   84.176  < 2e-16 ***
X211779_x_at         7.606e+01  1.073e+33  4.016e+00   18.938  < 2e-16 ***
X212014_x_at        -1.019e+01  3.746e-05  1.551e+00   -6.573 4.94e-11 ***
X212567_s_at         2.812e+02 1.275e+122  3.562e+00   78.927  < 2e-16 ***
X214806_at           1.731e+02  1.541e+75  1.378e+00  125.638  < 2e-16 ***
X214915_at           5.561e+01  1.416e+24  1.895e+00   29.349  < 2e-16 ***
X214919_s_at         7.433e+01  1.912e+32  3.022e+00   24.593  < 2e-16 ***
X215510_at          -8.436e+01  2.296e-37  1.060e+00  -79.554  < 2e-16 ***
X215633_x_at        -6.917e+01  9.084e-31  1.813e+00  -38.144  < 2e-16 ***
X216010_x_at         7.522e+00  1.848e+03  8.843e-01    8.507  < 2e-16 ***
X216103_at          -1.396e+01  8.674e-07  1.161e+00  -12.027  < 2e-16 ***
X216693_x_at        -2.412e+02 1.796e-105  2.120e+00 -113.771  < 2e-16 ***
X217019_at          -9.314e+01  3.547e-41  1.446e+00  -64.416  < 2e-16 ***
X217102_at           7.945e+01  3.199e+34  1.856e+00   42.801  < 2e-16 ***
X217404_s_at        -1.455e+01  4.780e-07  7.432e-01  -19.583  < 2e-16 ***
X217471_at          -3.395e+01  1.793e-15  1.464e+00  -23.189  < 2e-16 ***
X217767_at           2.812e+01  1.627e+12  1.430e+00   19.660  < 2e-16 ***
X217771_at          -2.098e+01  7.734e-10  1.520e+00  -13.799  < 2e-16 ***
X217815_at          -3.219e+02 1.517e-140  3.304e+00  -97.456  < 2e-16 ***
```

```
X218430_s_at          -1.387e+02   5.529e-61   1.805e+00  -76.881  < 2e-16 ***
X218478_s_at           2.274e+01   7.518e+09   2.848e+00    7.985 1.41e-15 ***
X218533_s_at          -4.961e+00   7.003e-03   2.746e+00   -1.807 0.070804 .
X218782_s_at          -7.209e+01   4.900e-32   1.549e+00  -46.531  < 2e-16 ***
X218883_s_at           1.350e+02   4.271e+58   2.045e+00   66.017  < 2e-16 ***
X218914_at             5.272e+01   7.874e+22   2.869e+00   18.378  < 2e-16 ***
X219340_s_at           1.222e+01   2.036e+05   1.878e+00    6.508 7.64e-11 ***
X219510_at             7.727e+00   2.268e+03   1.091e+00    7.084 1.40e-12 ***
X219588_s_at          -2.435e+02  1.736e-106   2.870e+00  -84.864  < 2e-16 ***
X219724_s_at          -1.770e+02   1.315e-77   1.449e+00 -122.182  < 2e-16 ***
X220886_at             8.119e+01   1.819e+35   1.393e+00   58.266  < 2e-16 ***
X221028_s_at           1.334e+02   8.967e+57   3.484e+00   38.302  < 2e-16 ***
X221241_s_at           8.687e+01   5.357e+37   1.306e+00   66.527  < 2e-16 ***
X221344_at            -9.489e+01   6.154e-42   2.524e+00  -37.594  < 2e-16 ***
X221634_at             1.698e+02   5.730e+73   2.656e+00   63.947  < 2e-16 ***
X221816_s_at          -1.568e+02   7.880e-69   2.703e+00  -58.014  < 2e-16 ***
X221882_s_at          -2.004e+02   9.445e-88   1.652e+00 -121.279  < 2e-16 ***
X221916_at            -7.075e+01   1.883e-31   1.434e+00  -49.350  < 2e-16 ***
X221928_at             8.315e+00   4.086e+03   1.133e+00    7.340 2.14e-13 ***
age                   -6.400e+00   1.661e-03   2.067e-01  -30.963  < 2e-16 ***
erpositive             2.404e+01   2.753e+10   3.253e+00    7.389 1.47e-13 ***
gradepoorly
differentiated        -1.438e+02   3.376e-63   3.580e+00  -40.178  < 2e-16 ***
gradeunkown           -2.165e+02   9.598e-95   1.145e+03   -0.189 0.850030
gradewell
differentiated        -3.288e+02  1.561e-143   5.482e+00  -59.977  < 2e-16 ***
size                  -5.766e-01   5.618e-01   1.896e+00   -0.304 0.760998
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                            exp(coef) exp(-coef)  lower .95  upper .95
X200726_at                 2.103e-106 4.755e+105 1.247e-109 3.546e-103
X200965_s_at                5.337e+69  1.874e-70  1.525e+68  1.868e+71
X201068_s_at               4.311e-141 2.320e+140 1.873e-144 9.923e-138
X201091_s_at               8.872e-158 1.127e+157 2.584e-160 3.045e-155
X201288_at                  1.355e+99 7.382e-100  1.514e+97 1.212e+101
X201368_at                  6.273e+14  1.594e-15  5.985e+12  6.576e+16
X201663_s_at               5.082e+106 1.968e-107 2.240e+105 1.153e+108
X201664_at                  3.219e+41  3.106e-42  1.338e+40  7.745e+42
X202239_at                 6.729e-171 1.486e+170 6.067e-174 7.464e-168
X202240_at                 1.641e+100 6.094e-101  2.136e+98 1.261e+102
X202418_at                 3.089e+130 3.238e-131 4.521e+127 2.110e+133
X202687_s_at                4.529e-65  2.208e+64  2.722e-66  7.535e-64
X203306_s_at                2.410e-65  4.149e+64  8.321e-68  6.981e-63
X203391_at                 2.250e-193 4.445e+192 2.222e-195 2.277e-191
X204014_at                  1.361e-40  7.348e+39  3.745e-41  4.945e-40
X204015_s_at                3.972e+30  2.517e-31  4.801e+29  3.286e+31
X204073_s_at                2.687e+09  3.722e-10  3.652e+08  1.976e+10
X204218_at                  8.173e-46  1.224e+45  2.620e-49  2.549e-42
X204540_at                  7.862e+27  1.272e-28  1.879e+27  3.289e+28
X204631_at                  1.375e-13  7.272e+12  6.708e-15  2.819e-12
X204740_at                  5.923e+03  1.688e-04  4.255e+01  8.245e+05
X204768_s_at                1.903e-33  5.256e+32  7.169e-36  5.050e-31
X204888_s_at                6.782e+42  1.475e-43  2.726e+41  1.687e+44
X205034_at                  3.249e-21  3.078e+20  7.182e-23  1.470e-19
X205848_at                  2.535e+48  3.945e-49  2.729e+47  2.354e+49
X206295_at                  1.156e-03  8.651e+02  4.092e-05  3.266e-02
X207118_s_at                3.183e-16  3.142e+15  4.096e-17  2.473e-15
```

```
X208180_s_at                3.927e+16  2.547e-17  7.171e+15  2.150e+17
X208683_at                  2.345e-56  4.265e+55  1.128e-58  4.876e-54
X209500_x_at                4.328e+88  2.311e-89  6.440e+86  2.908e+90
X209524_at                 9.590e+148 1.043e-149 1.254e+147 7.335e+150
X209835_x_at                1.895e-49  5.278e+48  9.813e-51  3.658e-48
X209862_s_at                6.643e-09  1.505e+08  5.307e-11  8.317e-07
X210028_s_at               3.566e+116 2.804e-117 8.231e+114 1.545e+118
X210314_x_at               5.031e+170 1.988e-171 1.041e+169 2.432e+172
X210593_at                  1.135e-37  8.812e+36  4.465e-39  2.884e-36
X211040_x_at               1.312e-151 7.622e+150 4.930e-154 3.491e-149
X211382_s_at                2.789e-80  3.585e+79  1.948e-82  3.993e-78
X211762_s_at                5.728e+83  1.746e-84  6.423e+81  5.107e+85
X211779_x_at                1.073e+33  9.322e-34  4.093e+29  2.811e+36
X212014_x_at                3.746e-05  2.669e+04  1.793e-06  7.826e-04
X212567_s_at               1.275e+122 7.846e-123 1.183e+119 1.373e+125
X214806_at                  1.541e+75  6.489e-76  1.035e+74  2.295e+76
X214915_at                  1.416e+24  7.060e-25  3.454e+22  5.808e+25
X214919_s_at                1.912e+32  5.231e-33  5.114e+29  7.147e+34
X215510_at                  2.296e-37  4.355e+36  2.873e-38  1.835e-36
X215633_x_at                9.084e-31  1.101e+30  2.598e-32  3.176e-29
X216010_x_at                1.848e+03  5.410e-04  3.267e+02  1.046e+04
X216103_at                  8.674e-07  1.153e+06  8.920e-08  8.434e-06
X216693_x_at               1.796e-105 5.567e+104 2.818e-107 1.145e-103
X217019_at                  3.547e-41  2.819e+40  2.085e-42  6.035e-40
X217102_at                  3.199e+34  3.126e-35  8.413e+32  1.216e+36
X217404_s_at                4.780e-07  2.092e+06  1.114e-07  2.051e-06
X217471_at                  1.793e-15  5.576e+14  1.017e-16  3.163e-14
X217767_at                  1.627e+12  6.146e-13  9.863e+10  2.684e+13
X217771_at                  7.734e-10  1.293e+09  3.929e-11  1.522e-08
X217815_at                 1.517e-140 6.592e+139 2.339e-143 9.838e-138
X218430_s_at                5.529e-61  1.809e+60  1.609e-62  1.900e-59
X218478_s_at                7.518e+09  1.330e-10  2.831e+07  1.997e+12
X218533_s_at                7.003e-03  1.428e+02  3.220e-05  1.523e+00
X218782_s_at                4.900e-32  2.041e+31  2.352e-33  1.021e-30
X218883_s_at                4.271e+58  2.341e-59  7.760e+56  2.351e+60
X218914_at                  7.874e+22  1.270e-23  2.847e+20  2.178e+25
X219340_s_at                2.036e+05  4.913e-06  5.126e+03  8.082e+06
X219510_at                  2.268e+03  4.408e-04  2.675e+02  1.924e+04
X219588_s_at               1.736e-106 5.761e+105 6.266e-109 4.810e-104
X219724_s_at                1.315e-77  7.606e+76  7.684e-79  2.250e-76
X220886_at                  1.819e+35  5.496e-36  1.185e+34  2.793e+36
X221028_s_at                8.967e+57  1.115e-58  9.708e+54  8.282e+60
X221241_s_at                5.357e+37  1.867e-38  4.144e+36  6.926e+38
X221344_at                  6.154e-42  1.625e+41  4.371e-44  8.663e-40
X221634_at                  5.730e+73  1.745e-74  3.144e+71  1.044e+76
X221816_s_at                7.880e-69  1.269e+68  3.942e-71  1.575e-66
X221882_s_at                9.445e-88  1.059e+87  3.705e-89  2.408e-86
X221916_at                  1.883e-31  5.311e+30  1.134e-32  3.127e-30
X221928_at                  4.086e+03  2.447e-04  4.436e+02  3.763e+04
age                         1.661e-03  6.021e+02  1.108e-03  2.491e-03
erpositive                  2.753e+10  3.633e-11  4.685e+07  1.617e+13
gradepoorly differentiated  3.376e-63  2.962e+62  3.027e-66  3.766e-60
gradeunkown                 9.598e-95  1.042e+94  0.000e+00        Inf
gradewell differentiated   1.561e-143 6.406e+142 3.363e-148 7.246e-139
size                        5.618e-01  1.780e+00  1.368e-02  2.307e+01

Likelihood ratio test= 442.9  on 82 df,   p=<2e-16
Wald test            = 469159  on 82 df,   p=<2e-16
```

```
Score (logrank) test = 153.2  on 82 df,   p=3e-06
```

After performing CPH model for training dataset, the performance of test data set is predicted by computing fitted values and regression terms for a model fitted by ***coxph()*** with ***type="lp"*** using ***predict()*** function. This is interpreted as the individual having a lower or higher risk for survival, depending on whether the value is below zero or above zero.

The linear predictors for 20 test observation is,

```
[1]  -847.152253   340.971951  -524.196853
[4] -1062.031255   460.435388   -39.030206
[7]  -935.374832 -2196.101652   273.136244
[10]    2.325584 -1168.855957    98.704610
[13] 1063.406382  -597.507267    90.553627
[16]  494.187446  1367.868203   349.651086
      [19] -109.332359  -770.892665
```

4.3.3 Perfomance of RSF model

This section is the implementation of the random survival forests and how to generate plots for inference. The main frame is structured as a tutorial for using the ***randomForestSRC*** package for building and post-processing random survival forest models and using the ***ggRandomForests*** package for understanding how the forest is constructed. We will build a random survival forest for the considered dataset.

The ***rfsrc()*** function fits the forest. It determines the type of forest by the response supplied in the formula argument. An argument applied to this function is,

- *mtry = 9*, number of variables to possibly split at each node is divided by 9 for regression.
- *ntree = 1000*, number of trees.
- *nsplit = 800*, non-negative integer specifying number of random splits for splitting a variable.
- *splitrule = "logrank"*, log-rank splitting rule
- *importance = TRUE*, is a method for computing VIMP.

Bootstap protocol helps to grow the trees. Type of bootstrap that was used here is ***by.root***, which is in effect of bootstrapping the data by ***sampling without replacement (swor; default)***.

We now grow a random forest for survival, by passing a survival ***Surv()*** object to the forest. The forest uses all variables in the trial (train) data set to generate the RSF survival model. As a consequence of the application ***rfsrc()***, the following outcome is achieved.

```
Sample size: 178
                    Number of deaths: 46
                     Number of trees: 1000
           Forest terminal node size: 15
       Average no. of terminal nodes: 15.396
No. of variables tried at each split: 9
              Total no. of variables: 80
       Resampling used to grow trees: swor
    Resample size used to grow trees: 112
                            Analysis: RSF
                              Family: surv
                      Splitting rule: logrank *random*
       Number of random split points: 800
                          (OOB) CRPS: 0.17462061
   (OOB) Requested performance error: 0.37202056
```

Using the function ***gg_error()*** of ***ggRandomForests***, can help to plot the results across the number of trees. Choosing the point (minimum number) where plot converges into a minimum. If it does not converge, trying it with a higher number of trees.

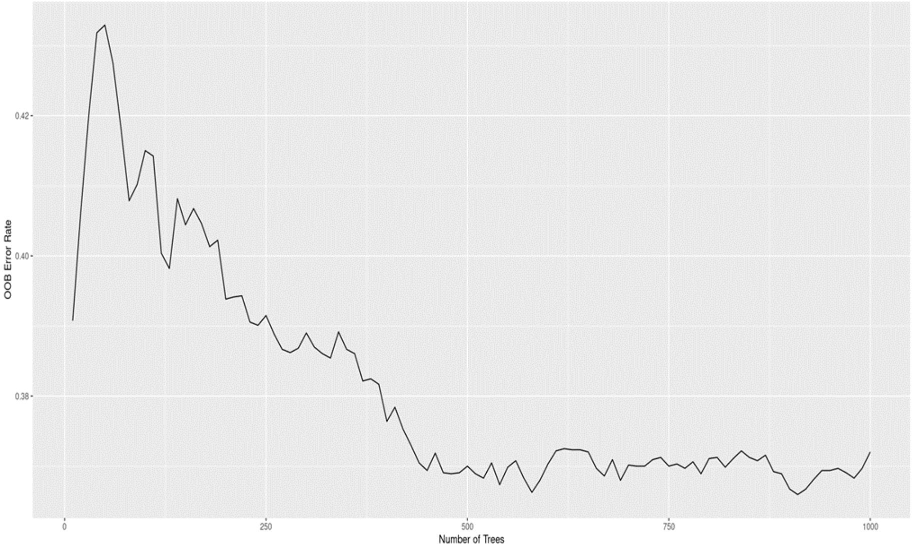


Figure 4.2: Shows the Number of Trees per error rate that is calculated for each observation by predicting the response over the set of trees which were not trained with that particular observation.

To Extract a single tree from a forest of 1000 number of trees, we use ***get.tree.rfsrc()***. This function extracts a single specified tree from a forest and converts the tree to a hierarchical structure.

Plotting the object will conveniently render the tree on the user's browser. Left tree splits are displayed. For continuous values, left split is displayed as an inequality with right split equal to the reversed inequality. For factors, split values are described in terms of the levels of the factor. In this case, the left daughter split is a set consisting of all levels that are assigned to the left daughter node and the right daughter split is the complement of this set. Terminal nodes are highlighted by colour and display the sample size and predicted value. By default, predicted value equals the tree predicted value and sample size are terminal node in-bag sample sizes. Figure 4.3, shows the single
tree for tree ID 1 and 1000

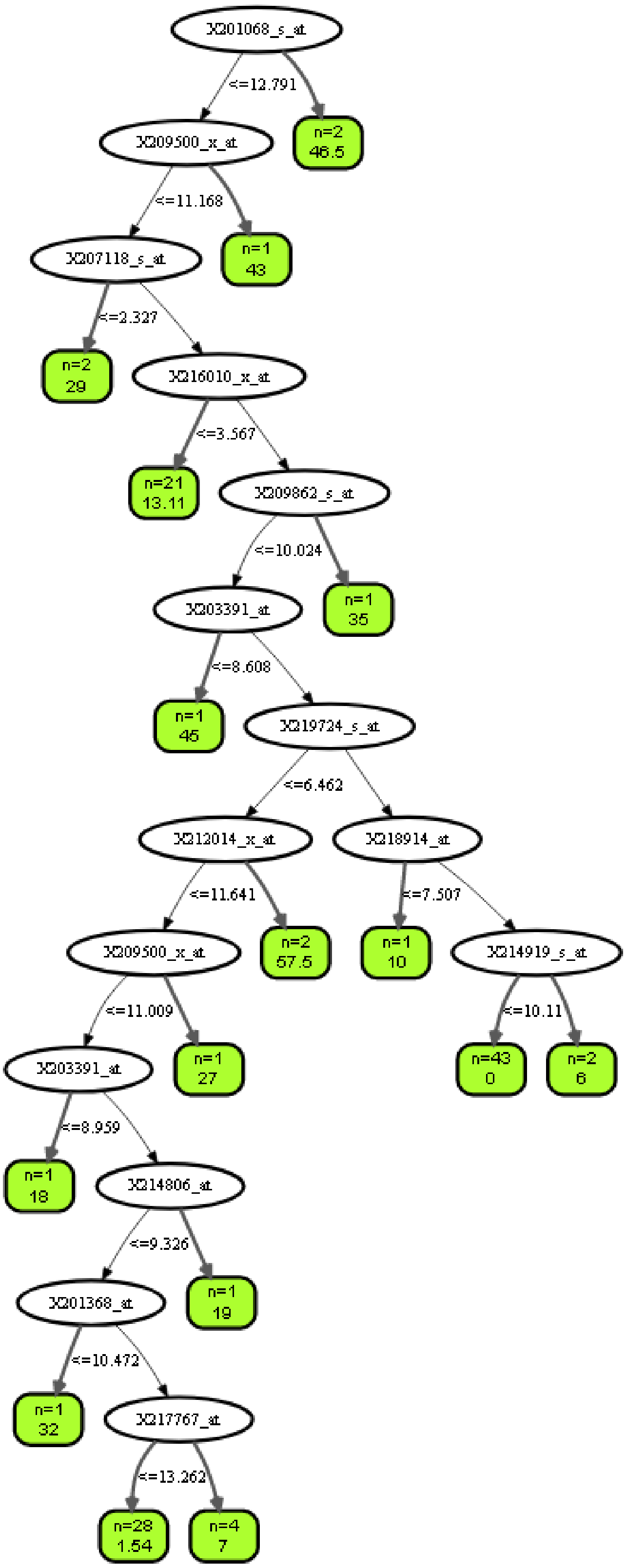


Figure 4.3: Single tree ID = 1

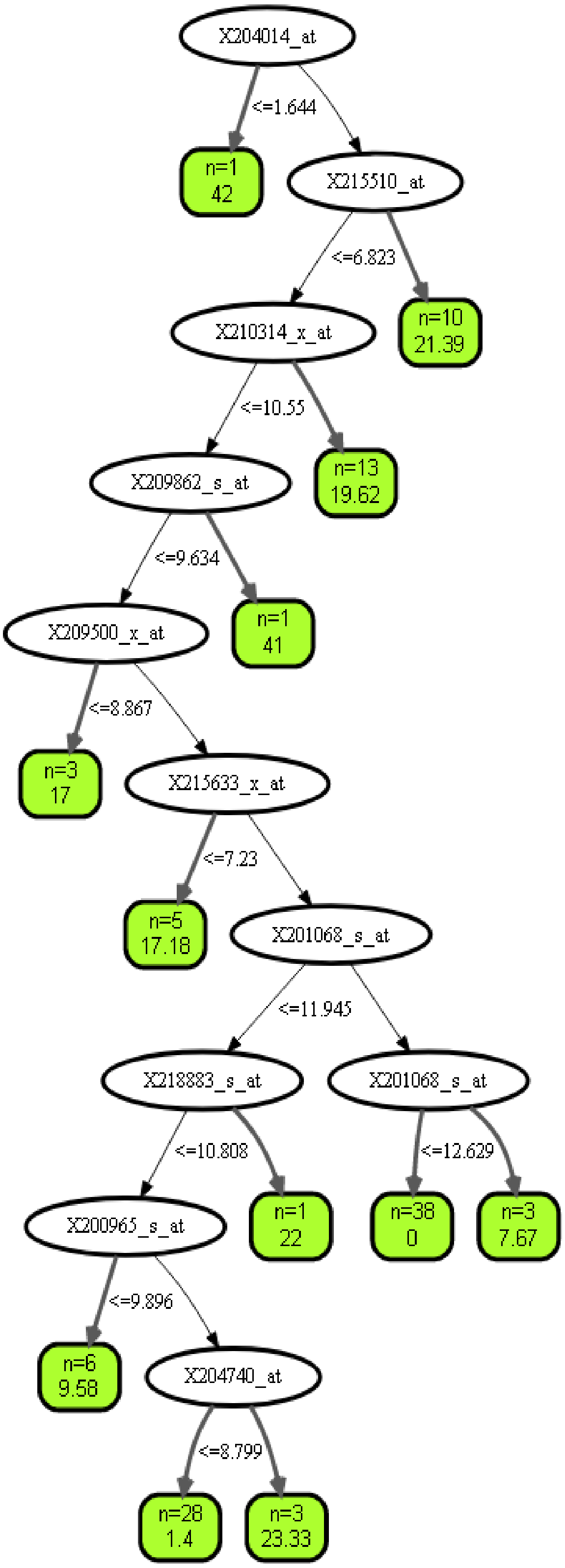


Figure 4.4: Single tree ID = 1000

4.3.4 Training-set predictions

We can now select a single patient in the training data set by using a function ***ggRFsrc ()*** of ***ggRandomForests*** and determine how positive or negative estrogen receptor status would affect the survival function. Figure 4.5, Shows where censored patients are colored green, and patients who have experienced the event (death) are colored in red.

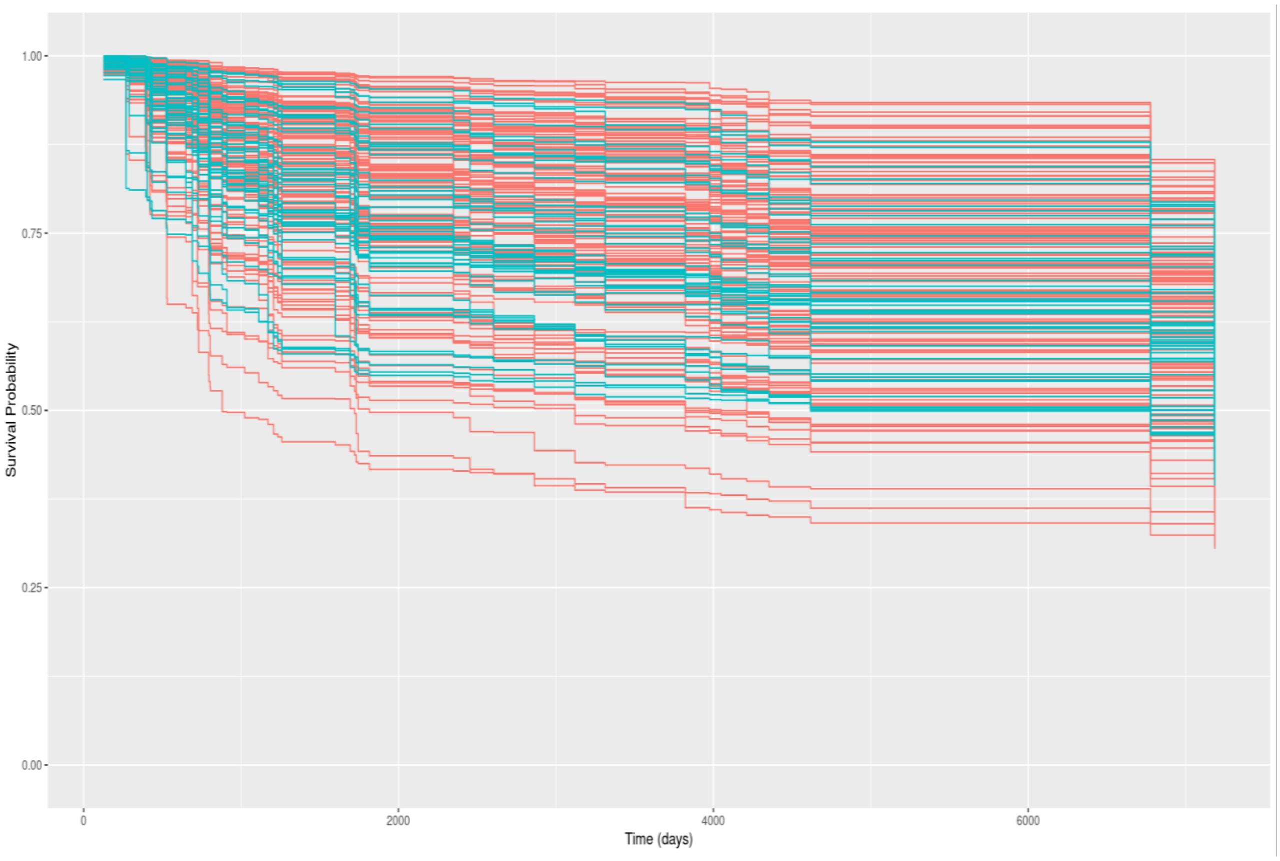


Figure 4.5: Survival Plot from RSF model for a single patient in the trial observations.

Interpretation of general survival properties from the plot above is difficult because of the number of curves displayed. To get more interpretable results, we plot a summary of the survival results. From figure 5, below we can observe that patients with positive estrogen receptor status tend to have a better prognosis than the negative estrogen receptor status.

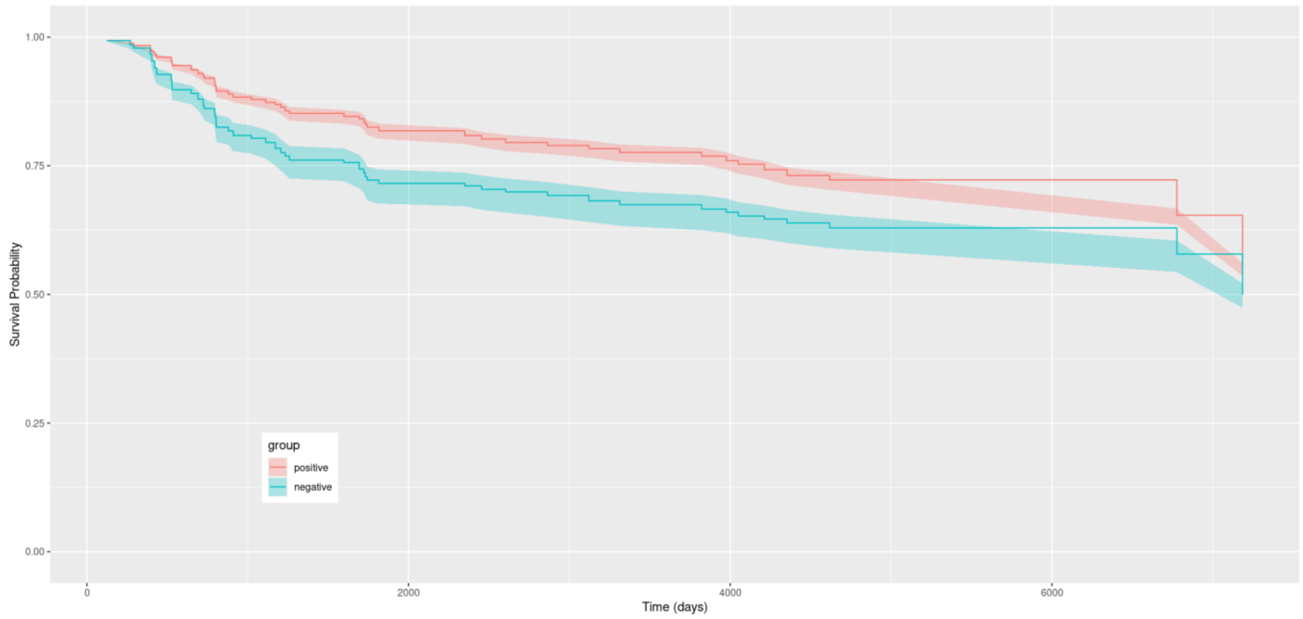


Figure 4.6: Median survival with a 95% shaded confidence band for the er-positive group in red and er-negative group in green.

4.3.5 Test-Set Predictions

The strength of adaptive tree imputation becomes clear when doing prediction on test set observations. If we want to predict survival for patients that did not participate in the trial using the model that was created, new data need to be applied. The function ***predict.rfsrc()*** takes the forest object and the test data set and returns a predicted survival using the same forest within the test data set. The following outcome is received when applying this function ***predict.rfsrc()***.

```
Sample size of test (predict) data: 20
                Number of grow trees: 1000
  Average no. of grow terminal nodes: 15.396
         Total no. of grow variables: 80
       Resampling used to grow trees: swor
    Resample size used to grow trees: 112
                            Analysis: RSF
                              Family: surv
                                CRPS: 0.26676909
         Requested performance error: 0.20253165
```

The function ***predictSurvProb ()*** is employed, to foresee the survival probability predictions on this test observation. To use this, the package ***pec ()*** (Prediction error curves) is required which evaluates the performance of risk prediction models in survival analysis. The following result is obtained when predicting survival probabilities for median day (4598.5) in each 20 observations of test dataset.

```
          [,1]
 [1,] 0.8751525
 [2,] 0.7047764
 [3,] 0.5044267
 [4,] 0.5348995
 [5,] 0.5192836
 [6,] 0.7338981
 [7,] 0.8144397
 [8,] 0.8245206
 [9,] 0.7773142
[10,] 0.6955194
[11,] 0.7480731
[12,] 0.6626270
[13,] 0.7432664
[14,] 0.8330090
[15,] 0.8257696
[16,] 0.7080668
[17,] 0.5253787
[18,] 0.7000726
[19,] 0.6365489
[20,] 0.7244519
```

Plotting various survival estimates on test set using ***plot.survival.rfsrc()***. This function produces the plots for forest estimated survival function for each individual (thick red line is overall ensemble survival; thick green line is Nelson-Aalen estimator). Also, it plots for mortality of each individual versus observed days (time). Points in blue correspond to events, black points are censored observations. Forest estimated cumulative hazard function (CHF) (displayed using black lines). Blue lines are the CHF from the estimated hazard function.

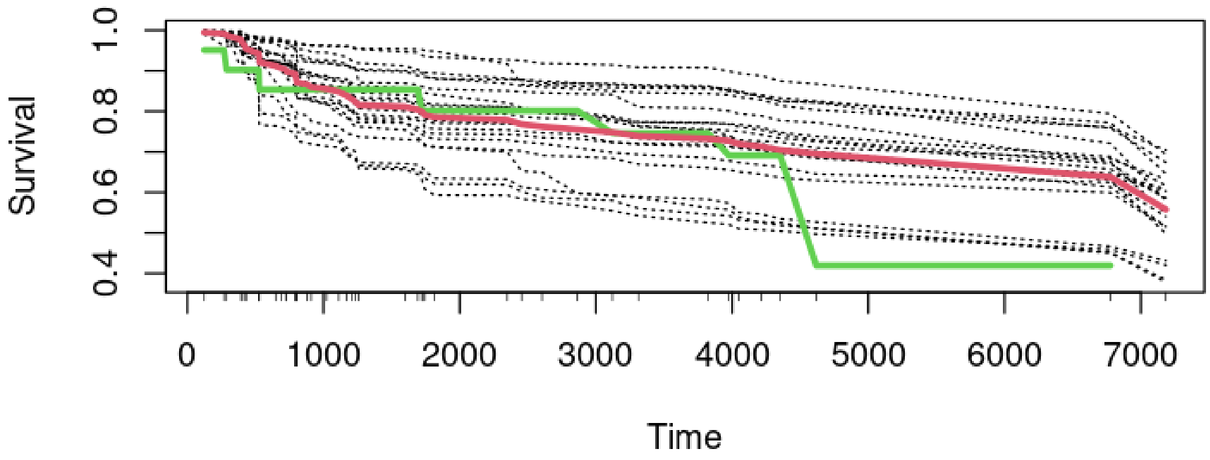


Figure 4.7: Estimated survival function on test observations.

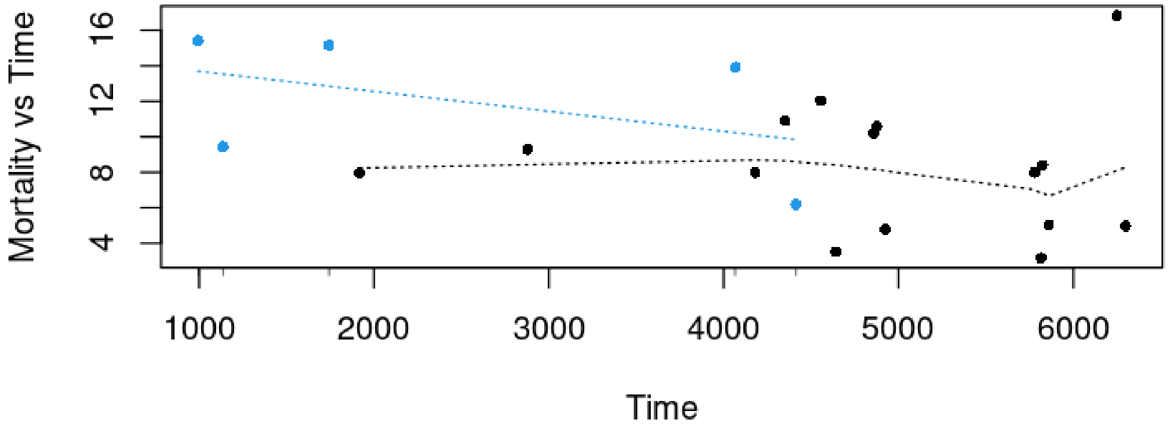


Figure 4.8: Estimated Mortality rate on test observations.

As obtained in training set, prediction survival probabilities can also be obtained for new data. Figure 4.10, shows that patients with positive estrogen receptor status tend to have a better prognosis than the negative estrogen receptor status as obtained in trail.

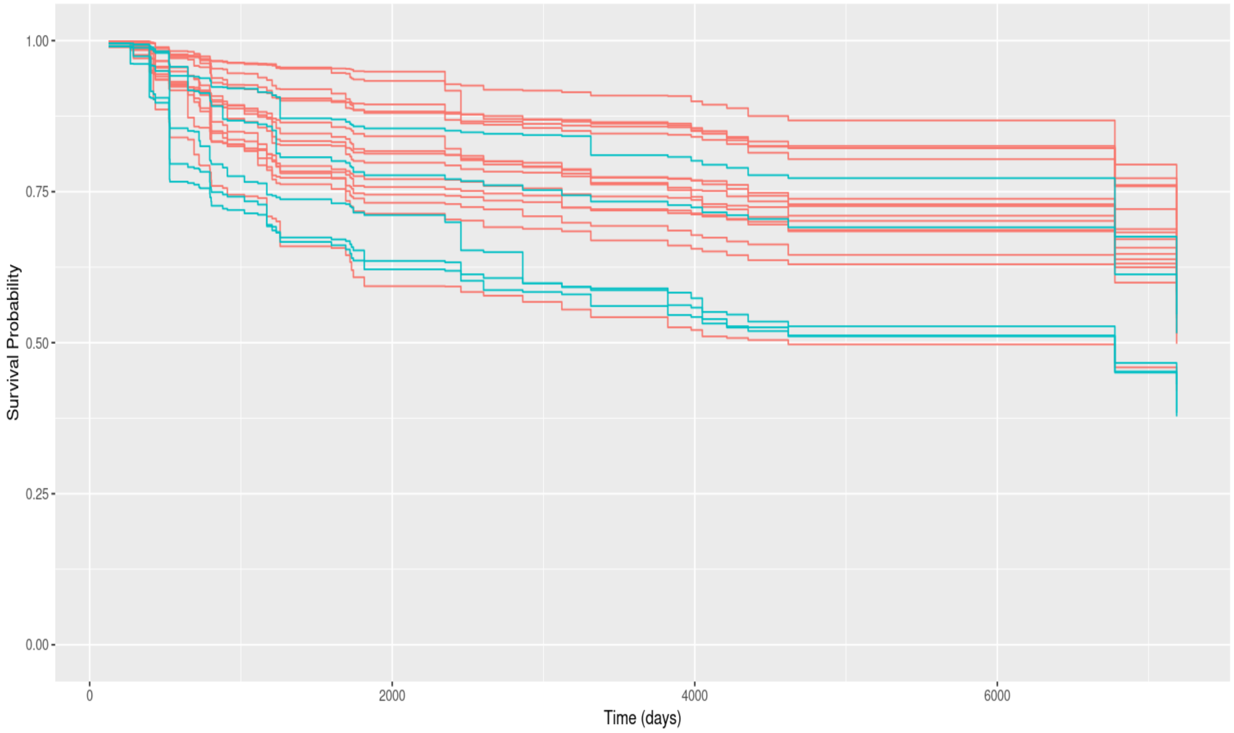


Figure 4.9: Predicted Survival Plot from RSF model for a single patient in the test observations.

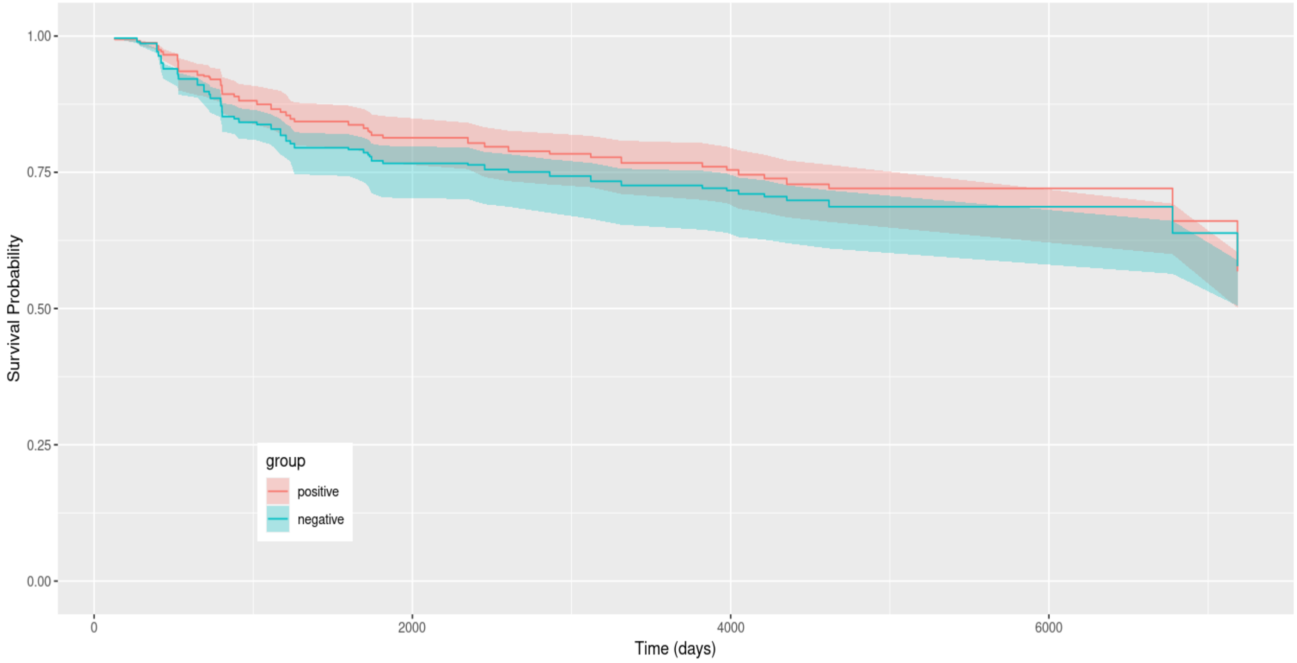


Figure 4.10: Median survival with a 95% shaded confidence band for the er-positive group in red and er-negative group in green in the test observations.

4.3.6 Variable Importance in RSF

An essential parameter is the number of trees. To compute the best number of trees, use ***importance=TRUE,*** so that unnecessary computing is not used, and set a sufficiently high number of trees.

Random forest uses all variables available in the data set to construct the response predictor. We can also demonstrate random forest variable selection techniques using Variable Importance (VIMP) to improve the prediction outcome of interest. Hence care should be taken while including variables into the model. The VIMP of trees for ***$importance*** object contains variables importance, in same order as in input dataset. We sort it out to show a ranking and use ***ggRandomForests ()*** to plot this object. This library uses ***ggplot2***, so that it can easily retouch these plots.

	Importance		Importance
X202240_at	0.0603	X214806_at	0.0479
X203391_at	0.0371	X218883_s_at	0.0287
X204014_at	0.0275	X215510_at	0.0267
X201091_s_at	0.0249	X204740_at	0.0222
X211779_x_at	0.0207	X209862_s_at	0.0193
X209500_x_at	0.0187	X218478_s_at	0.0151
X221028_s_at	0.0141	X214915_at	0.0116
X205848_at	0.0114	X201068_s_at	0.0114
X207118_s_at	0.0114	X204218_at	0.0112
X204631_at	0.0110	X211762_s_at	0.0101
X203306_s_at	0.0101	X220886_at	0.0101
X221916_at	0.0098	X208180_s_at	0.0097
X205034_at	0.0095	X216693_x_at	0.0091
X201368_at	0.0084	X212014_x_at	0.0066
X211040_x_at	0.0066	X215633_x_at	0.0065
size	0.0065	X219724_s_at	0.0058
X214919_s_at	0.0052	X219510_at	0.0051
X204015_s_at	0.0050	X221882_s_at	0.0048
X209835_x_at	0.0046	X217815_at	0.0045
X217019_at	0.0044	X217767_at	0.0036
X210028_s_at	0.0030	X202418_at	0.0024
X210314_x_at	0.0024	er	0.0021
X221816_s_at	0.0020	X201288_at	0.0020
X218533_s_at	0.0019	X200726_at	0.0018
X204540_at	0.0018	X201663_s_at	0.0017
X217404_s_at	0.0015	X216103_at	0.0014
X202687_s_at	0.0013	X209524_at	0.0012
X217471_at	0.0010	X202239_at	0.0009
X221241_s_at	0.0009	X204768_s_at	0.0009
X201664_at	0.0008	X217102_at	0.0005
age	0.0005	X219340_s_at	0.0004
X208683_at	0.0003	X219588_s_at	0.0001
X218782_s_at	0.0001	X200965_s_at	0.0000
X217771_at	-0.0003	X212567_s_at	-0.0004

X221344_at	-0.0004	X204073_s_at	-0.0005
X210593_at	-0.0006	grade	-0.0007
X216010_x_at	-0.0008	X221634_at	-0.0010
X218914_at	-0.0012	X211382_s_at	-0.0016
X221928_at	-0.0017	X206295_at	-0.0017
X204888_s_at	-0.0019	X218430_s_at	-0.0053

Table 4.2: Importance of Variable sorted from higher rank features

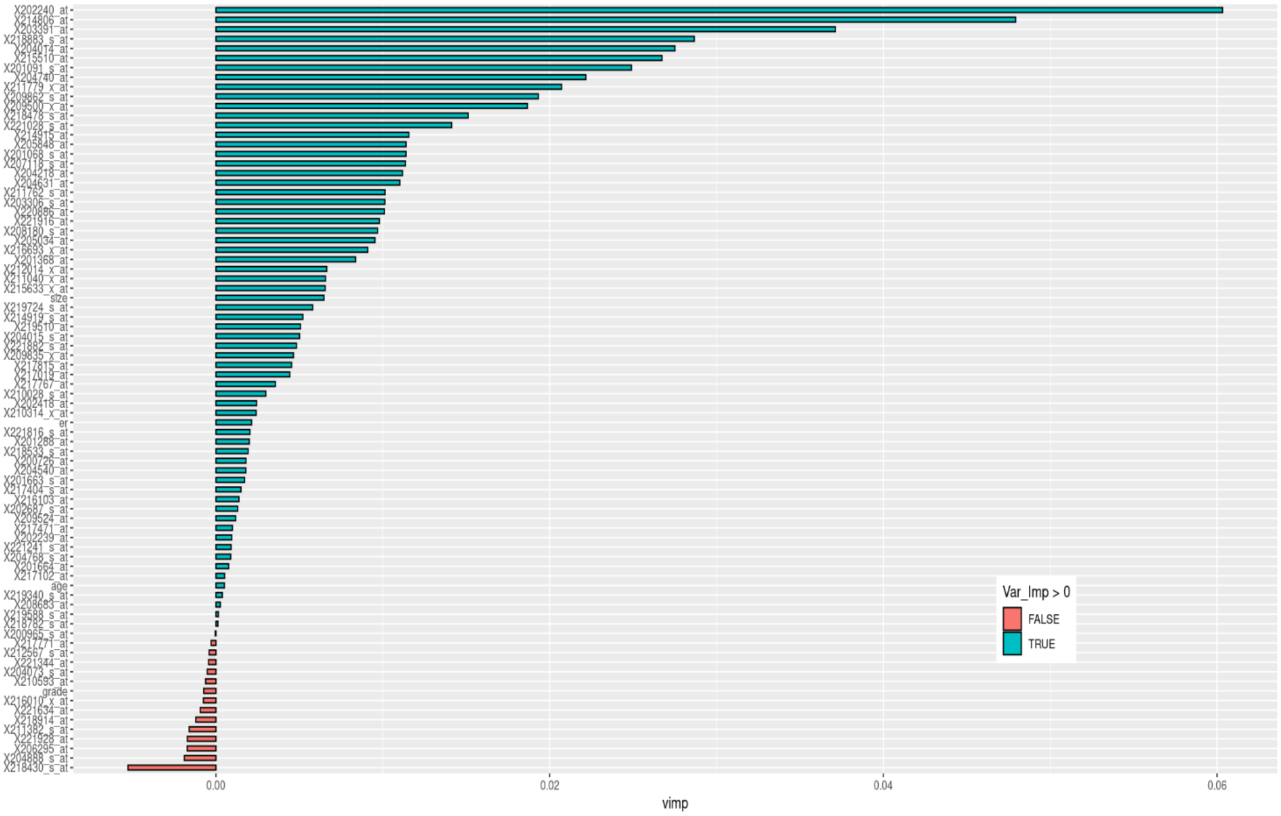


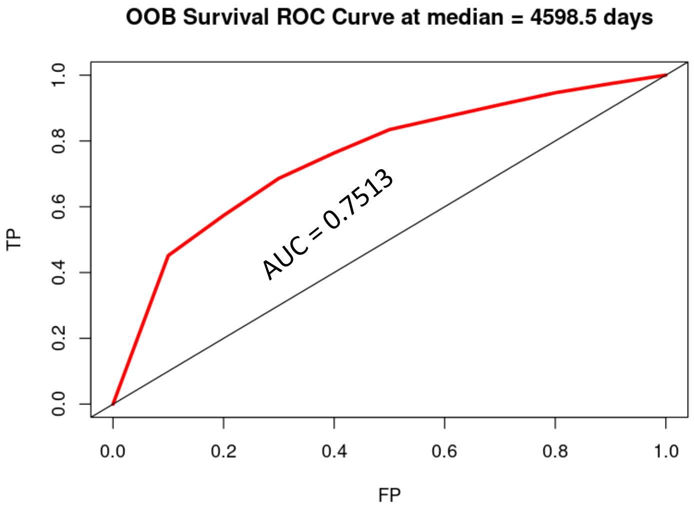
Figure 4.11: Variable importance ranking plot of applied random forest.

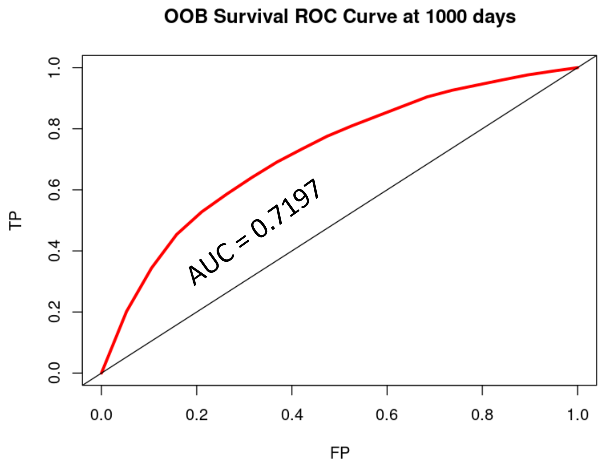
Once an idea of which variables are most interested in are known, we understand how these variables are related to the response. It gives an idea of the overall trend of a variable/response relation. Model Can be improved by selecting a large positive values of a variable relying on high predictive power of the forest.

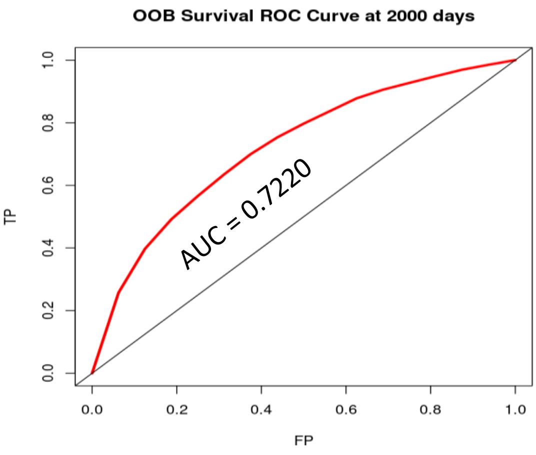
4.4 Comparison of CPH and RSF on Predictive ability

In CPH model, obtained linear predictor of the test model is used as a marker to plot ***SurvivalROC()*** helps to give the probability of AUC. It creates time-dependent ROC curve estimation from censored survival data using the Kaplan Meier (KM) or Nearest Neighbor Estimation (NNE) method. Survival probability obtained on test set is 0.7272, for death (=1) time.

In RSF model, we obtain error rate corresponding to the number of trees by using ***$err.rate*** and ***$ntree***. The C-Index is obtained by using ***rcorr.cens()*** function. This is equal to 1 - error rate. Here, ***risksetROC()*** library provides functions to compute a ROC curve and its associated Area Under Curve (AUC) in a time-dependent context. Let's see predictive ability of OOB samples (in testing set) at median time. We must be careful about method as, depending on assumptions, but let's assume we meet Cox's proportional hazards assumption. Figure 4, represents OOB Survival ROC Curve at median days, 1000 days, 2000 days, 3000 days, and 5000 days.


OOB Survival ROC Curve at median = 4598.5 days


OOB Survival ROC Curve at 1000 days


OOB Survival ROC Curve at 2000 days

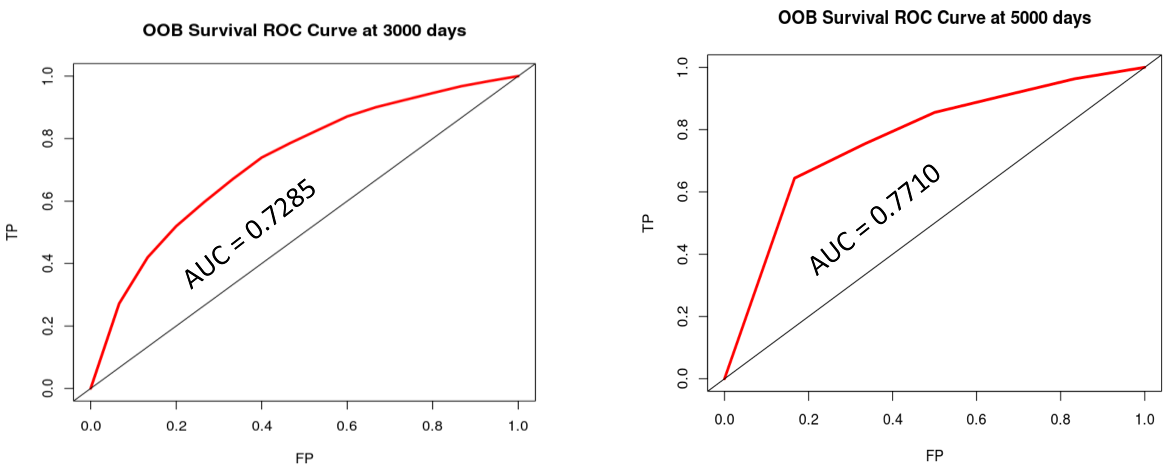


Figure 4.12: ROC Curves over median,1000, 2000, 3000, and 5000 days for RSF Model.

AUC on the new test dataset (20 observations) is an average computation of true positive Rates over all possible values of the false positive rate.

Model	Time (days)
	Median Time = 4598.5
CPH	0.7130
RSF	0.7513

Table 4.3: Performance on AUC for the test set

The Concordance-Index (C-Index) on the new test dataset (20 observations) is the computation of an agreement between an observed response and a predictor.

Model	C- Index
	Test
CPH	0.7215
RSF	0.7975

Table 4.4: Performance on risk prediction (C-Index) for the test set

Based on all these evaluation measures, it can be inferred that RSF method improves survival prediction accuracy when compared to CPH method. Thus, RSF model outperforms the CPH model.

CHAPTER 5

5.1 Conclusion

This study indicated a superior accuracy of RSF model as compared to CPH present better fit to predict survival probability when feeding into a new observation. Prognosis prediction plays a critical role in clinical and personal decision-making for cancer patients. There have been attempts to conduct traditional statistics and ML methodology to predict individual survival. CPH and RSF model are extensively used in application of cancer survival that generally refers to time-to-event censored data. Our primary motivation to compare the performance between CPH and RSF model in using the 76-gene expression breast cancer dataset in identifying a statistical model to predict overall survival effectively based on a set of covariates. Our results showed that RSF model present better fit to predict new dataset.

The main advantage of our study is that it approached progression prediction based on a time-to-event dataset. According to Harrell's concordance index, RSF based on approximate log-rank splitting rule determined major risk factors for survival which shows a slightly better performance than CPH approach. RSF is an attractive method when the goal is to do prediction. Its advantage is more apparent when the dimension of covariates is large, relationship between response and covariates are complex or when the proportional hazard assumption is at risk. Although RSF acts as a great alternative in analysing time-to-event data, it is worth nothing that interpretability is burdensome and that care must be taken when choosing and tuning the trees based on the form of available data.

REFERENCES

1) Fast Unified Random Forests with randomForestSRC, Random Survival Forest by Hemant Ishwaran, Michael S. Lauer, Eugene H. Blackstone, Min Lu, Udaya B. Kogalur, 2022-06-01. https://www.randomforestsrc.org/index.html

2) ggRandomForests: Exploring Random Forest Survival John Ehrlinger Microsoft arXiv:1612.08974v1 [stat.CO] 28 Dec 2016

3) https://aacrjournals.org/clincancerres/article/13/11/3207/193398/Strong-Time-Dependence-of-the-76-Gene-Prognostic

4) https://cran.r-project.org/web/packages/randomForestSRC/randomForestSRC.pdf

5) https://cran.r-project.org/web/packages/randomForestSRC/randomForestSRC.pdf

6) https://cran.r-project.org/web/packages/survival/survival.pdf

7) https://pedroconcejero.wordpress.com/2015/11/12/survival-random-forests-for-churn-prediction-3/

8) https://scikit-survival.readthedocs.io/en/stable/user_guide/random-survival-forest.html

9) https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE7390

10) Ikeda, K., Horie-Inoue, K. & Inoue, S. Identification of estrogen-responsive genes based on the DNA binding properties of estrogen receptors using high-throughput sequencing technology. *Acta Pharmacol Sin* 36, 24–31 (2015). https://doi.org/10.1038/aps.2014.123

11) Michael Ingrisch, Franziska Schöppe, Karolin Paprottka, Matthias Fabritius, Frederik F. Strobl, Enrico N. De Toni, Harun Ilhan, Andrei Todica, Marlies Michl and Philipp Marius Paprottka. Journal of Nuclear Medicine May 2018, 59 (5) 769-773; DOI: https://doi.org/10.2967/jnumed.117.200758

12) RANDOM SURVIVAL FORESTS, By Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone and Michael S. Lauer Cleveland Clinic, Columbia University, Cleveland Clinic and National Heart, Lung, and Blood Institute. DOI: 10.1214/08-AOAS169

13) **Shu Jiang, Prediction Based on Random Survival Forest. Am J Biomed Sci & Res. 2019 - 6(2). AJBSR.MS.ID.001005. DOI: 10.34297/AJBSR.2019.06.001005.**

14) Strobl, C., Boulesteix, AL., Kneib, T. *et al.* Conditional variable importance for random forests. *BMC Bioinformatics* 9, 307 (2008). https://doi.org/10.1186/1471-2105-9-307

15) Time-dependent ROC for Survival Prediction Models in R, by Kazuki Yoshida, R Project, May 22, 2018. https://datascienceplus.com/time-dependent-roc-for-survival-prediction-models-in-r/

16) **Zhou D. Prognostic factors and predictions of survival data using cox ph models and random survival forest approaches. *Biom Biostat Int J*. 2017;5(5):165-181. DOI: 10.15406/bbij.2017.05.00142**

APPENDIX

```
library(survival)
library(tidyverse)
library(caret)
library(randomForestSRC) # package for random forest
library(ggRandomForests) # package for random forest
library(ggplot2)
library(ggpubr)
library(survivalROC)
library(dplyr)
ddft<-read.csv(file.choose(),header = T)
ddft[sapply(ddft,is.character)]<-lapply(ddft[sapply(ddft, is.character)],as.factor)
str(ddft)
data.frame(sapply(ddft,class))
set.seed(2110)
train<-sample(1:198,178,replace = FALSE)
train
ddft.trial = ddft[train,]
ddft.test = ddft[-train,]
install.packages("survminer")
library(survminer)
#Kaplan-Meir
km <- survfit(Surv(days, status==1) ~ 1,
            data = ddft)
ggsurvplot(km,
  data = ddft.trial,
  size = 1,                # change line size
  risk.table = TRUE,       # Add risk table
  break.time.by=365,
  risk.table.col = "strata",# Risk table color by groups
```

```
 legend.labs ="1",    # Change legend labels
 risk.table.height = 0.25, # Useful to change when you have multiple groups
 ggtheme = theme_bw()      # Change ggplot2 theme
)
#CPH
cox_survival <- survival::coxph(Surv(days, status==1) ~ ., data = ddft.trial, method = 'efron')
summary(cox_survival)
ddft.test$lp <- predict(cox_survival,type="lp", newdata = ddft.test)
ddft.test$lp
summary(ddft.test$lp)
cind_cox_test <- survival::concordance(cox_survival, newdata = ddft.test)$concordance
round(cind_cox_test, 4)
roc<-survivalROC(Stime = ddft.test$days,
         status = ddft.test$status,
         marker = ddft.test$lp,
         predict.time = median(ddft.test$days),
         method="KM")
roc
# RSF
set.seed(999)
rfsrc.df = rfsrc(Surv(days, status) ~ ., # survival object
         data = ddft.trial, # data
         mtry=9,
         ntree = 250,
         nsplit = 50, # number of random splits to consider
         #na.action = "na.impute", # imputing missing data
         #tree.err = TRUE,
         #seed = -9,
         #nodesize = 150,
         #oob=TRUE,
                splitrule = "logrank",
                importance = TRUE) # variable importance
rfsrc.df
par(mar = c(2,2,2,2))
plot(rfsrc.df)
plot(gg_error(rfsrc.df))
dev.off()
vimp<-sort(rfsrc.df$importance, decreasing = T)
round(vimp,4)
plot(gg_vimp(rfsrc.df))+
  theme(legend.position = c(0.8, 0.2))+
  labs(fill = "Var_Imp > 0")
ggRFsrc = plot(gg_rfsrc(rfsrc.df))+
  theme(legend.position = "none" )+
  labs(y = "Survival Probability", x = "Time (days)")+
  coord_cartesian(ylim = c(-0.01, 1.01))
show(ggRFsrc)
library(devtools)
devtools::install_github('araastat/reprtree')
library(reprtree)
plot(get.tree.rfsrc(rfsrc.df,1))
plot(get.tree.rfsrc(rfsrc.df,1000))
plot(gg_rfsrc(rfsrc.df, by = "er"))+
  theme(legend.position = c(0.2, 0.2))+
  labs(y = "Survival Probability", x = "Time (days)")+
  coord_cartesian(ylim = c(-0.01, 1.01))
install.packages("pec")
library(pec)
rfsrc.df.test = predict.rfsrc(rfsrc.df, newdata = ddft.test)
rfsrc.df.test
```

```
test_survprob = predictSurvProb(rfsrc.df, newdata = ddft.test,times = median(ddft.test$days))

test_survprob

plot.survival.rfsrc(rfsrc.df)

plot(gg_rfsrc(rfsrc.df.test))+

  theme(legend.position = "none")+

   labs(y = "Survival Probability", x = "Time (days)")+

   coord_cartesian(ylim = c(-0.01, 1.01))

plot(gg_rfsrc(rfsrc.df.test, by = "er"))+

   theme(legend.position = c(0.2, 0.2))+

   labs(y = "Survival Probability", x = "Time (days)")+

   coord_cartesian(ylim = c(-0.01, 1.01))

library(rms)

library(risksetROC)

err.rate.rsf_test = rfsrc.df.test$err.rate[rfsrc.df.test$ntree]

err.rate.rsf_test

cind_rf_test<-rcorr.cens(-rfsrc.df.test$predicted,

                                        Surv(ddft.test$days, ddft.test$status))["C Index"]

cind_rf_test

w.ROC_test = risksetROC(Stime = ddft.test$days,

                                   status = ddft.test$status, marker = rfsrc.df.test$predicted,

                                predict.time = median(ddft.test$days), method = "Cox",

                                main = paste("OOB Survival ROC Curve at median = 4598.5 days"),

                              lwd = 3, col = "red" )

  w.ROC_test
```