Project 2: Breast Cancer Survival Prediction

Abstract

This project endeavors to leverage a dataset comprising information about breast cancer patients from a prospective study, with the primary objective of constructing predictive models for estimating the survival months and survival status of these patients. During our exploratory data analysis, noteworthy insights emerged: the non-normal distribution of variables and the interrelationships among them suggest a departure from the assumptions inherent in linear models. In response to these findings, we opted to develop two nonlinear models to enhance predictive accuracy—a survival analysis model for survival months and a logistic model for survival status. As a result, both models exhibited satisfactory predictive power, underscoring their efficacy in forecasting outcomes. Furthermore, the results provide valuable indications that some indicators significantly influence the breast cancer survival risk. Lastly, both models exhibited a different ability in the prediction ability between different races, which may lead to new models built on race-stratified data.

Introduction

In this breast cancer survival prediction project, our focus is on utilizing a dataset encompassing variables such as age, tumor characteristics, and patient demographics to predict the risk of death and identify significant factors influencing survival. With an emphasis on model performance and fairness evaluation, we also aim to address potential disparities between racial groups. The report will provide a comprehensive analysis, including data exploration, model selection, and model diagnostics.

Methods

1. Exploratory Data Analysis (EDA)

For our EDA, we commenced by providing fundamental numeric descriptive statistics (Table 1) for the original data. Visual representations, such as density plots (Figure 1) and histograms (Figure 2), were employed to illustrate variable distributions. The identification of outliers was conducted through boxplots (Figure 3). Additionally, the interrelationships between variables were elucidated via a correlation plot (Figure 4), offering a comprehensive overview of the dataset's key characteristics.

2. Survival Analysis

Survival analysis, pivotal in medical research for studying time-to-event data, employs both univariate and multivariate analyses. The Kaplan-Meier (KM) estimator, widely used for survival function approximation, facilitates univariate analysis, especially in datasets with non-informative right censoring. Simultaneously, the Cox proportional hazards (Cox_PH) model examines covariate impact on hazard rates, offering insights into factors influencing time to the event of interest while upholding proportional hazards assumptions.

Kaplan-Meier Method

In our analysis, we employed the Kaplan-Meier (KM) method to estimate survival functions, assessing the impact of population groups within predictor variables on survival time. To discern differentials in survival time based on predictor variables, both binary and continuous variables were categorized into discrete population groups. Non-weighted log-rank tests were then conducted to identify statistically significant disparities between survival functions, considering both two-sample and four-sample log-rank tests for binary and quartile-grouped continuous variables, respectively. The log-rank test statistic, following a chi-square distribution, determined the significance of observed versus expected events at each time interval, guiding our inclusion of variables in subsequent multivariate analyses. The significance level was set at $\alpha = 0.05$.

Cox-PH Method

In our analysis, the Cox-PH model served as the key tool for investigating the association between patients' survival time and various predictor variables. The Cox model was chosen because it is widely used when working with clinical data due to its applicability to a wide variety of studies. To streamline the model, we employed the Stepwise method for variable selection, followed by an exhaustive method to identify interaction terms.

The inclusion of interaction terms was assessed by comparing the performance of models with and without them. This comparative analysis allowed us to discern the necessity of interaction terms in

enhancing the model's predictive capabilities. Unlike the KM survival curve, the Cox model fit curve provides predicted survival outcomes after accounting for other variables.

3. Logistic Model

Logistic regression is a statistical model that predicts the probability of a binary outcome by transforming a linear combination of predictor variables through the logistic function, ensuring outputs fall within the range of 0 to 1.

Predictors Examination

We classified our predictors into categorical and continuous variables. For categorical ones, we assessed their distribution through histograms (Figure 2), conducted chi-square tests, and created correlation plots (Figure 4). Continuous variables were examined using density plots (Figure 1), rank-sum tests, and t-tests for median comparisons between status-stratified groups. Our findings led to strategic decisions: 1) exclusion of the 4^{th} level of *grade* due to anaplastic mode, 2) removal of *differentiate*, *x6th_stage*, and *estrogen_status* to address multicollinearity, 3) elimination of *survival_months*, and 4) retention of *regional_node_examined* for subsequent Walt test.

Furthermore, we explored interaction terms exhaustively, revealing two significant pairs: *a_stage* with *regional_node_positive*, and *reginol_node_examined* with *regional_node_positive*. These interactions were incorporated into our model selection process.

Model Variables Selection and Diagnosis

During this stage, we employed a comprehensive approach for variable selection in our logistic models. Utilizing the Stepwise Method, Random Forest (Figure 5), and LASSO, we gauged the importance of variables and ensured non-multicollinearity. The models generated through these methods underwent scrutiny via Nomogram, Calibration Curve, and Hosmer and Lemeshow Goodness-of-fit Test for comparison. The ultimate variables chosen for our logistic model were determined based on accuracy and kappa value, validated through a 5-fold cross-validation (Table 2).

Result

1. EDA

Upon examination of density plots (Figure 1) and histogram plot (Figure 2), a non-normal distribution was noted in the *regional_node_positive* variable, prompting consideration for log transformation. Boxplots (Figure 3) revealed numerous outliers in three numeric variables (excluding *age*), indicating a right-skewed distribution pattern. The correlation plot (Figure 4) highlighted significant correlations between *differentiate* and *grade, t-stage* and *6th-stage*, as well as *estrogen* and *progesterone status*. These findings suggest the potential need for nonlinear models in our analysis.

2. Survival Analysis

KM Method

The Kaplan-Meier survival curves (Figure 6) reveal distinct patterns across various predictors. *Progesterone status* and *age* exhibit significant differences in survival time, with positive progesterone and the age group 61-69 standing out. *Tumor size*, *regional_node_positivity*, *race* (particularly for Black individuals), *marital status* (especially for the separated group), *t-stage*, *n-stage* (particularly N3), *6th-stage* (for widows), *differentiate* (for undifferentiated individuals), *grade* (4^{th} level), and distant *a-stage* display notable variations in survival probabilities. The *Estrogen status* group also demonstrates significant differences, particularly for negative Estrogen Status. These findings underscore the diverse impact of predictor variables on survival outcomes, providing valuable insights for further analysis.

Cox-PH method

Following the Stepwise method, the survival model included *age, race, T stage, N stage, differentiate, estrogen status, progesterone status, regional node examined*, and *regional node positive*. Leveraging a combined approach of Exhaustive and Stepwise methods, eight interaction terms were identified and incorporated into the Cox-PH model. Model comparison based on AIC favored the model with interaction terms, signifying superior performance. Consequently, this refined model was deemed the final choice for survival analysis using the Cox-PH method (Figure 7).

Subsequently, we assessed model performance across different racial groups—Black, White, and other races (Figure 8). Notably, variations in the prediction of survival probability over time were observed among these groups, suggesting that race might play a crucial role in predicting survival time.

3. Logistic Model

In evaluating the models derived from different methods, the Stepwise Method demonstrated superior results in accuracy, kappa, nomogram score, and H-L value (Table 2). Therefore, the variables used in the final model include *age, race, t_stage, n_stage, grade, a_stage, progesterone_status, regional_node_examined, reginol_node_positive,* and an interaction term between *a_stage* and *reginol_node_positive*. The model revealed insightful associations: 1) a one-year increase in age corresponds to a 0.02-fold rise in the risk of death; 2) for Black individuals, the probability of death is 0.58 times higher than that for White individuals; 3) in the interaction term, a unit increase in regional node positive, given distant and regional a_stage, results in a 0.68-fold higher probability of death.

For validation and real-world applicability, the model was executed with race-stratified data Prediction accuracy was 86.3% for White individuals and 83.6% for other races, indicating a slight difference of 3%, deemed acceptable. To refine accuracy, a suggested approach involves separate modeling for each racial group, offering a tailored solution to enhance predictive performance.

Conclusion/Discussion

In summary, our analysis of breast cancer survival prediction involved EDA, survival analysis using KM and Cox-PH methods, and logistic modeling. EDA highlighted non-linearity, guiding us toward advanced models. KM provided nuanced insights into survival disparities, while Cox-PH unveiled a refined model with interaction terms, showcasing superior predictive performance. The logistic model, optimized through the Stepwise Method, demonstrated superior accuracy, and highlighted key interactions. Notably, both survival and logistic models indicated varying performances among different racial groups, suggesting the need for separate modeling to enhance predictive accuracy. In conclusion, our findings underscore the intricate dynamics of breast cancer survival, emphasizing the importance of nuanced, inclusive modeling for robust prognostication.

Contribution

Yunshen Bai – Cox-PH method

Yangyang Chen – EDA and KM method

Kindle Zhang – EDA and logistic regression

Jingyi Xu – Interpretation and report writing

Appendix

Figures

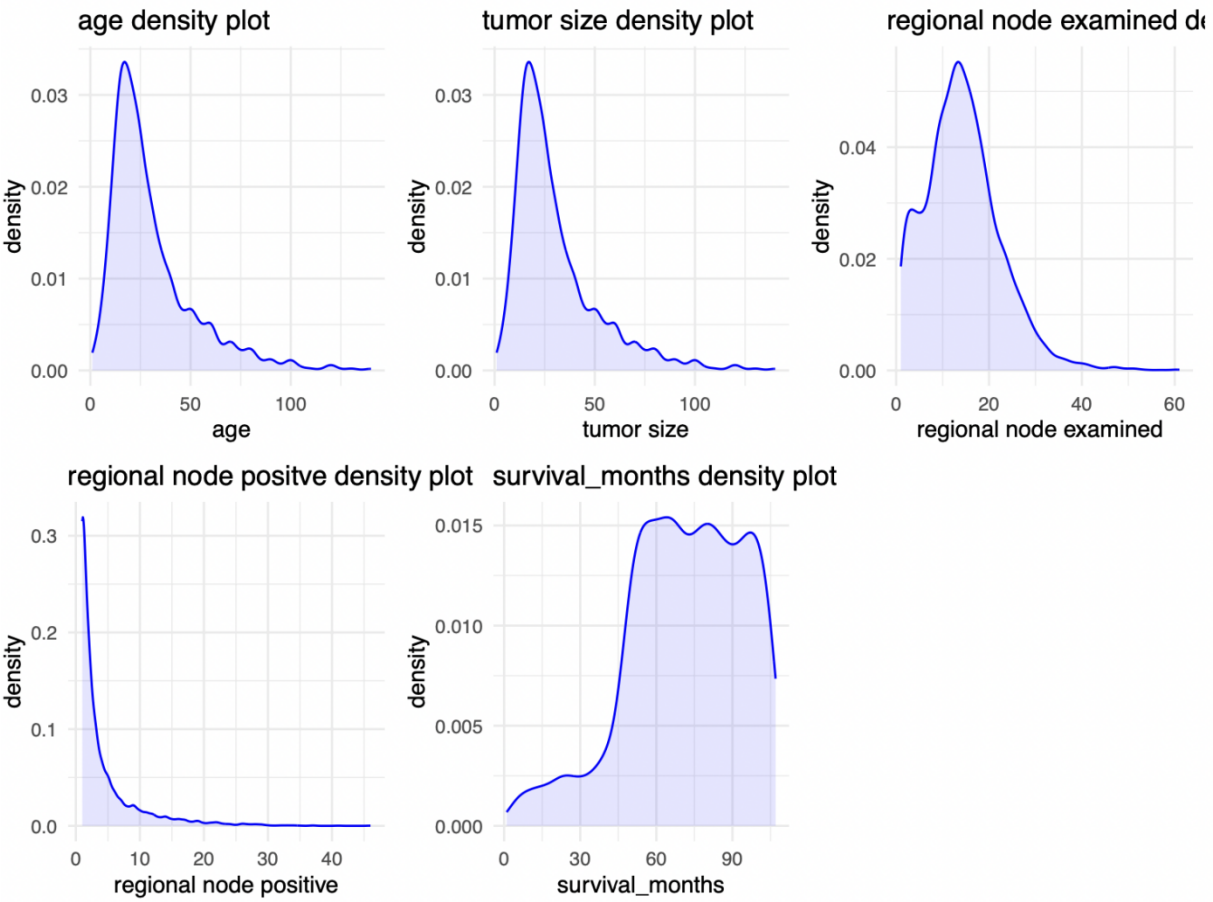

Figure 1 Density plots for all numeric variable

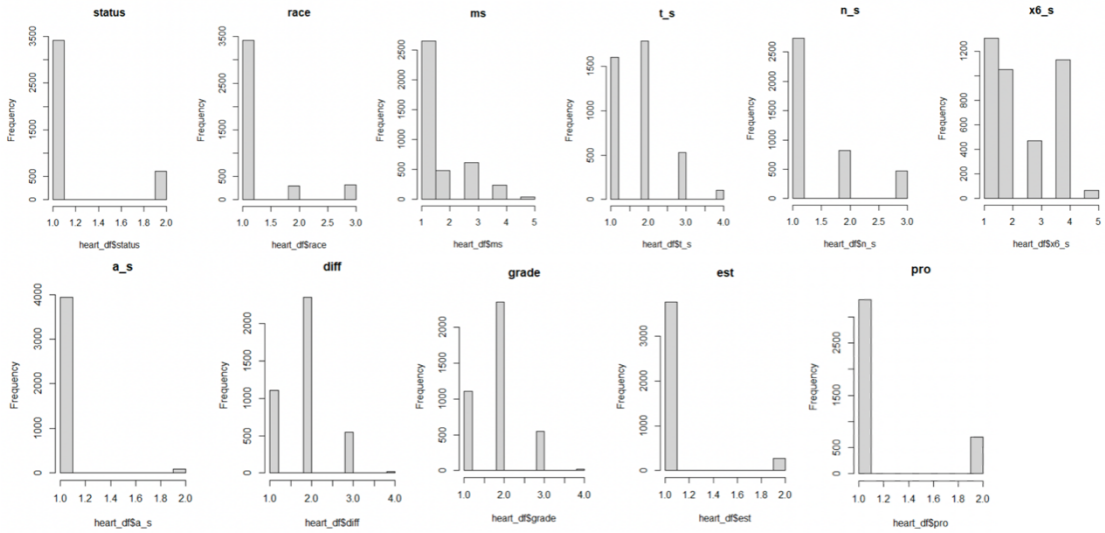

Figure 2 Histogram plots for categorical variables.

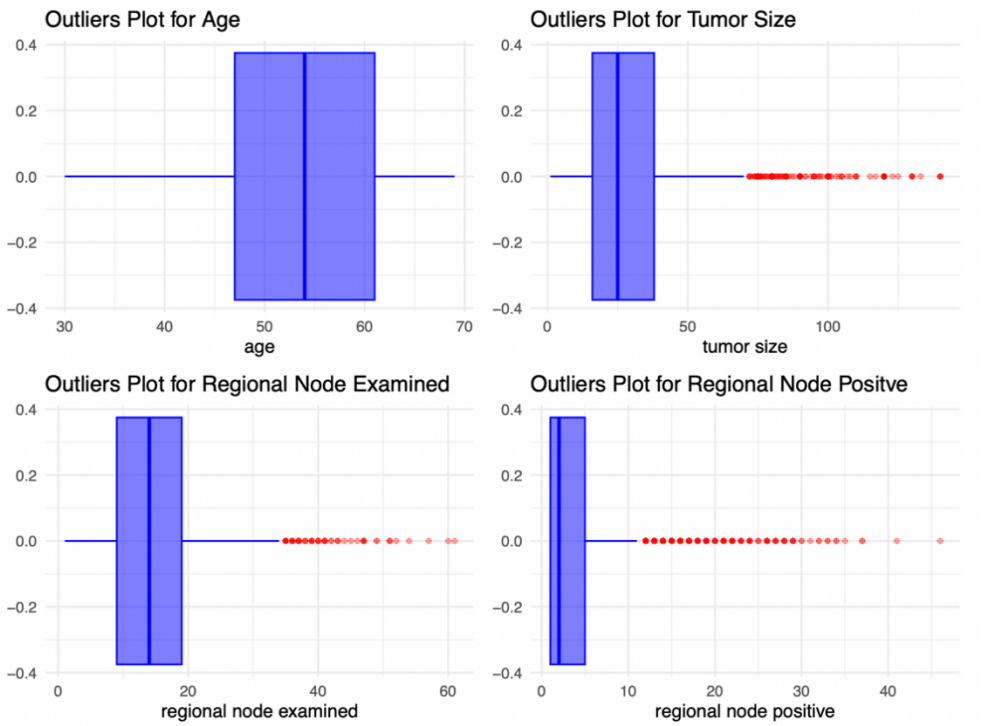

Figure 3 Boxplots for continuous variables to identify outliers.

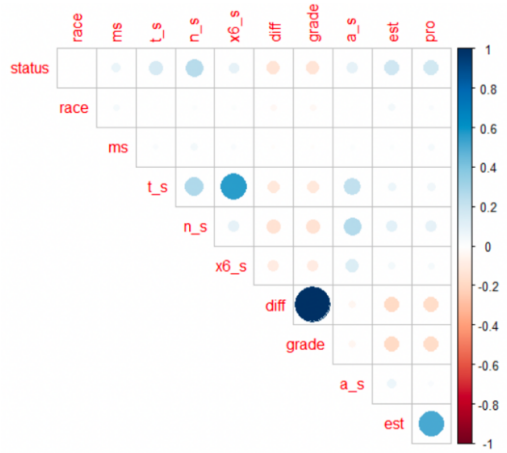

Figure 4 Correlation plots for categorical variables.

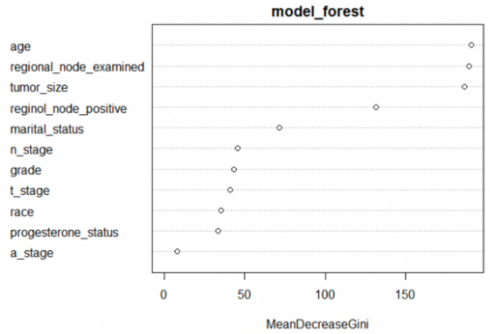

***Figure 5 Feature importance for Random Forest**. According to the picture, the first 5 most important variables with a mean decreased gini over 50 were selected.*

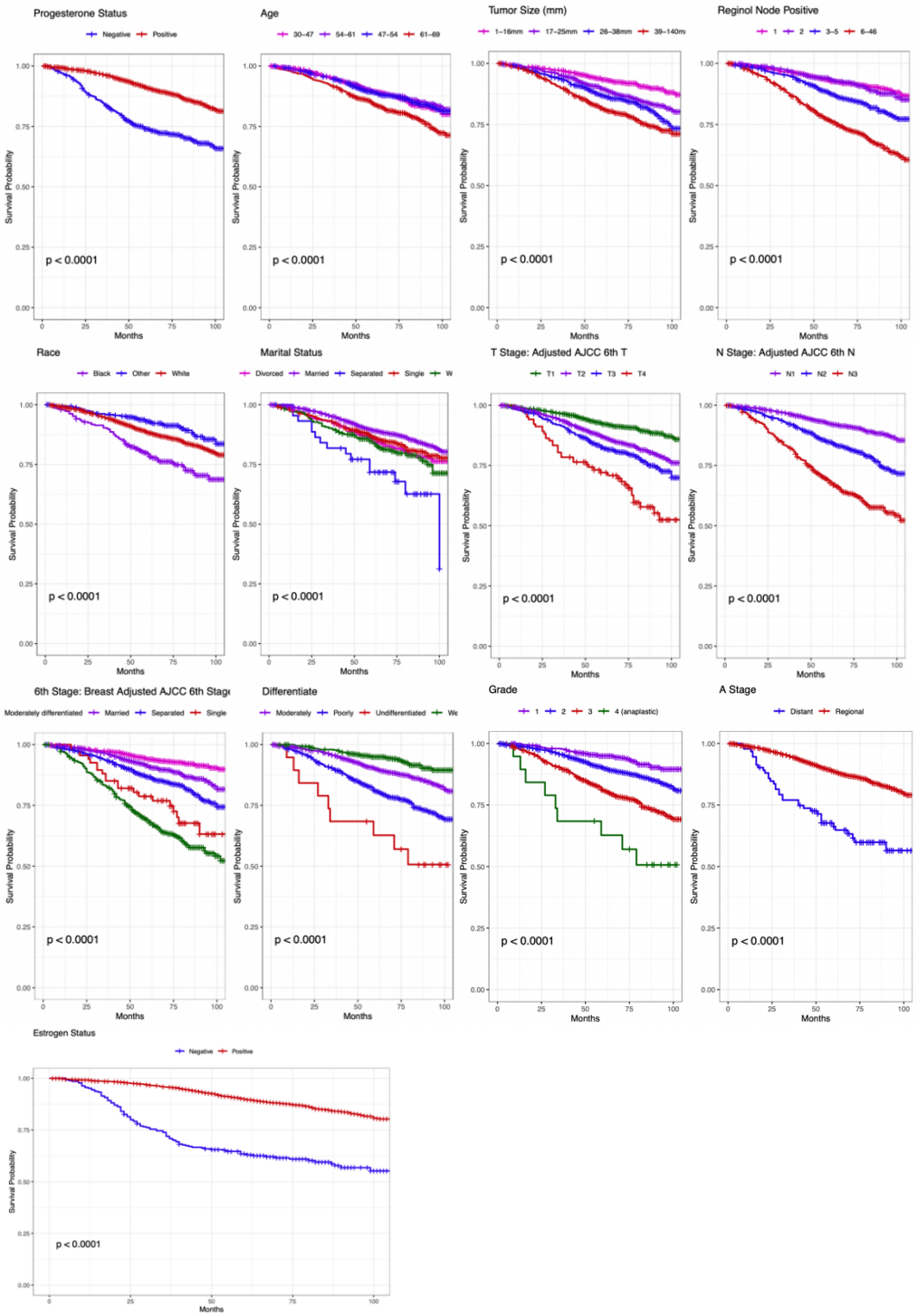

Figure 6 Kaplan-Meier survival curves of variables with a significant log-rank test result. *Progesterone status and age exhibit significant differences in survival time, with positive progesterone and the age group 61-69 standing out. Tumor size, regional_node_positivity, race (particularly for Black individuals), marital status (especially for the separated group), t-stage, n-stage (particularly N3), 6th-stage (for widows), differentiate (for undifferentiated individuals), grade (4th level), and distant a-stage display notable variations in survival probabilities. The Estrogen status group also demonstrates significant differences, particularly for negative Estrogen Status.*

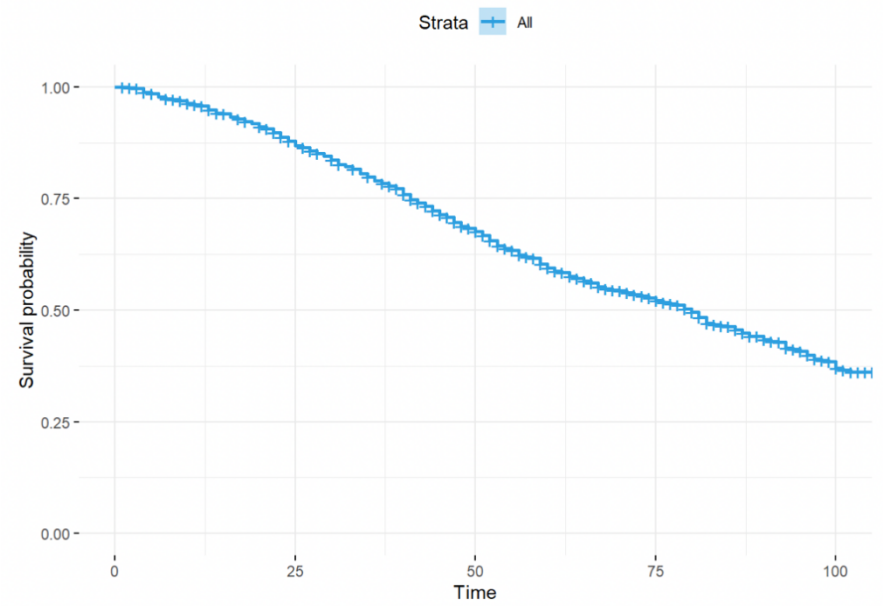

Figure 7 Survival time - survival probability plot by Cox-PH method.

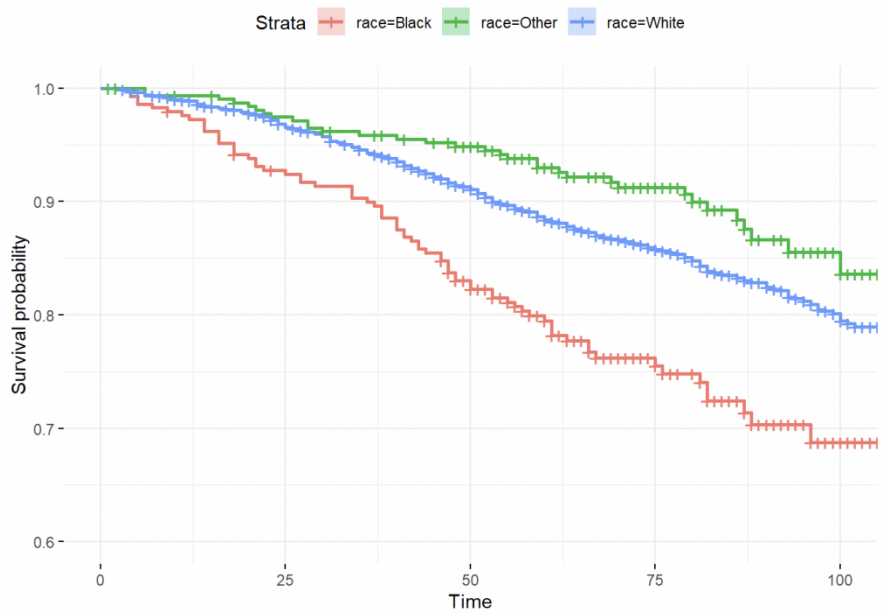

Figure 8 Survival time - survival probability stratified by races.

Tables

age	tumor_size	examined_node	positive_node	survival_month	status
Min. :30.00	Min. : 1.00	Min. : 1.00	Min. : 1.000	Min. : 1.0	Min. :0.0000
1st Qu.:47.00	1st Qu.: 16.00	1st Qu.: 9.00	1st Qu.: 1.000	1st Qu.: 56.0	1st Qu.:0.0000
Median :54.00	Median : 25.00	Median :14.00	Median : 2.000	Median : 73.0	Median :0.0000
Mean :53.97	Mean : 30.47	Mean :14.36	Mean : 4.158	Mean : 71.3	Mean :0.1531
3rd Qu.:61.00	3rd Qu.: 38.00	3rd Qu.:19.00	3rd Qu.: 5.000	3rd Qu.: 90.0	3rd Qu.:0.0000
Max. :69.00	Max. :140.00	Max. :61.00	Max. :46.000	Max. :107.0	Max. :1.0000

Table 1 Descriptive statistics of numeric data.

model	cro_val_accuracy	cro_val_kappa	nomo_plot	hl_pvalue
model1	0.8566796	0.1874696	7.6%	0.6940802
model2	0.8524347	0.1152586	6.3%	0.2399465
model3	0.8521874	0.1191757	11.8%	0.7272005

Table 2 Model comparison in logistic model selection. *Model 1: Stepwise method. Method 2: Random Forest. Model 3: LASSO*