Final Report for Summer Semester

by

Roungu Ahmmad

PhD Candidate of Biostatistics and DataScience

Submitted to

Yunxi Zhang, Ph.D.

Assistant Professor

Course Director: BDS 791 - Special Topics

Department of Data Science

John D Bower School of Population Health

The University Mississippi Medical Center MS 39216 USA

November 19, 2022

Abstract

Michigan has the highest death rate among the states in the United States. Detroit is one of the areas in Michigan with the highest mortality rate for men and women. There are many factors that influence this death rate, but we are looking at breast cancer outcomes among male respondents. There is a very low likelihood of death, but the researchers have taken the death risk into account. Due to different diagnosis systems and a variety of factors affecting this event, predicting the survivability of men suffering from breast cancer is a complex process. The Surveillance, Epidemiology, and End Results Program (SEER) is a large institute with a large number of respondents and records of breast cancer, but we only have a minimal number of male breast cancer patients responding.

Missing data values are very common because of the complexity of the data structures and results. In statistical methods, we use multiple imputation and regularization methods to infer missing values. In machine learning methods, the miss forest algorithm is used to estimate missing values. Furthermore, step-wise method and lasso regression model have been used in survival scenario for variable selection, while variable hunting and random forest methods have been used for machine learning approaches. The survivability of male breast cancer patients is predicted using Cox model and machine learning methods. AUC, prediction errors, Brier score, and OOB survival rate score are used to compare which methods prediction more accurate.

According to machine learning methods, age is the single greatest predictor of mortality among male breast cancer patients. By the Cox model, for single-year increase of age, approximately 5% chance

increase of death (HR: 0.05, SD = 0.006). Similarly, for T1-4a stage cancer patients have approximately 30% lower chance of death compared with T0-X stages. Total tumor around the prime site is negatively related to a survival probability of male breast cancer patients. Considering all the comparisons of OOB survival, OOB Brier score, prediction error, AUC, machine learning methods provided better prediction compare with the Cox Proportional model.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Background

There are different types of breast cancer depending on which cells in the breast cause cancer symptoms. Breast cancer is the most common cancer among women in the United States, except for skin cancer, according to the CDC report. As noted in Kopans (2008) over the last decade, while overall rates of breast cancer in US women have not changed dramatically, rates have increased dramatically across racial and age groups in Detroit, Michigan. As the American Cancer Society reports every year in the U.S., about 245 thousand cases of breast cancer are diagnosed and out of this number, around 41 hundred women die every year from breast cancer yet a few data are recorded for male respondents. According to some researchers, breast cancer is part of the death rate for males in the United States.

Various studies have shown that the risk factors for breast cancer are a combination of factors, but the main factors are getting older and having a family history of breast cancer. The majority of breast cancers are found in women over 50 years old. According to the Detroit Atla report, about 11% of all new cases of breast cancer in the United States are found in younger patients, yet very few studies have examined male breast cancer events and consequences.

This study aims to consider over 26 potential factors that affect breast

cancer and to identify the most important factors for events and predict survivability using machine learning methods and Cox regression models. A very low proportion of respondents from the data set and an extremely low proportion of outcomes are observed in this research. Furthermore, a lot of missing data have been found in this data set which are highly important for predictions the breast cancer. Therefore, we are going to imputation of this missing data by various traditional and and machine learning methods and then analysis the data by traditional cox model, machine learning model. Finally we compare the outcomes, which model give us better estimate by considering the concordance index and prediction errors.

1.2 Methods

In the study of biological systems, inference and prediction are two major objectives. As a result of inference, formalize the understanding of how the system behaves or test a hypothesis. It is possible to determine the best course of action (e.g., treatment choice) without understanding the underlying mechanisms. In a typical research project, both inference and prediction are important-we want to know how biological processes work and what will happen next.

Several methods from statistics and machine learning (ML) can, in theory, be used for both prediction and inference. However, statistical methods have a long-standing focus on inference, which is achieved by creating and

fitting a project-specific probability model. The model allows us to compute a quantitative measure of confidence that a discovered relationship describes a genuine effect that is unlikely to be caused by noise. Furthermore, if enough data are available, we can explicitly verify assumptions (e.g., equal variance) and refine the specified model as needed. On the other hand, ML focuses on prediction using general-purpose algorithms to find patterns in often complex and large dataset. They are effective even when the data come from systems that are not carefully controlled and when there are complex nonlinear relationships among them. However, since ML solutions lack an explicit model, they can be difficult to tie directly to biological knowledge, despite the prediction results being convincing.

The computational tractability of classical statistics and machine learning varies as the number of variables per subject increases. Statistical models of the past were designed for data with a few dozen input variables and sample sizes that would be considered small to moderate today. By adding the models to the system, the unobserved aspects are filled in. However, as the number of input variables and possible associations increases, the model that captures those relationships becomes more complex. Thus, statistical inferences become less precise and the line between statistical and ML approaches becomes blurred. In this study we are focusing on both statistical methods and machine learning methods for imputation of missing values, variables selection and finally, comparison the performance of them.

2 Statistical Method

Traditionally, the Cox proportional hazards model has been the most widely used technique for analysing censored data. In case we have lot of missing values at first we need to impute the missing values. Imputation is the process of substituting values for missing data in statistics. Imputation is used to avoid pitfalls that can arise when deleted cases that have missing values are analyzed since missing data can create problems with analysis. As a result, when one or more values are missing for a case, most statistical packages default to discarding that case entirely, which may introduce bias or affect the representatives of the results. By substituting an estimated value for missing data based on other available information, imputation preserves all cases. There are a number of theories that scientists have embraced to account for missing data, but they all introduce bias into the analysis. Once missing values have been imputed, the data set can be analyzed using standard methods for complete data.

2.1 Imputations

The absence of data introduces considerable bias, makes handling data more difficult, and impairs interpretation of the findings. When using Imputation, we can avoid pitfalls that can occur when deleted cases with missing values are analyzed as missing data can cause problems during analysis. An imputation preserves all cases by substituting estimated values for missing data

based on other information. For missing data, there are several theories, but all of them introduce some bias into analysis. The data set can then be analyzed using standard methods for complete data after missing values have been imputed. For this study, we used the **mice** r package for multiple imputation and created 10 different data sets. By using Rubin methods, we finally pooled these results.

2.2 Variables Selections

Variable selection is a critical component of research and further investigation. The advantage of a model-based approach is that the importance calculation is more closely related to the model performance and may be able to incorporate the correlation structure between the predictors. All the predictor of the dataset are not very relevant for fitting the model. So we are using step-wise methods and regularization methods for variable selections and finally fitting the cox proportional hazard model.

2.2.1 Stepwise Methods for Variable Selections

Step wise regression (or step wise selection) involves iteratively adding and removing variables in the predictive model in order to find the subset of variables in the data set that result in the most accurate model, that is, a model that lowers prediction error. There are three types of step wise regression. 1. With forward selection, the model starts with no predictors, adds the most significant predictors iteratively, and stops when the improvement is no

longer statistically significant. 2. Backward selection (or backward elimination), which starts with all predictors in the model (full model), iteratively removes the least significant predictors, and stops when all predictors are statistically significant. and finally, A step wise selection (or sequential replacement) is a combination of forward and backward selections. Starting with no predictors, you add the most significant ones sequentially (such as forward selection). Remove any variables that no longer improve the model fit (such as backward selection) after adding each new variable. In survival data, we used iterative approach and continue the process unit the final model and the previous one have equal c-index. In the final stage of the iteration we find that the maximum c-index is 0.747, and this methods provide the given bellow list of predictors.

```
*** Stepwise Final Model (in.lr.test: sle = 0.05;
out.lr.test: sls = 0.25;
variable selection restrict in vif = 999):
Call:
coxph(formula = Surv(time, status) ~ trt_prime_site + age +
    no_lymph_nodes + tumor_size + tumor_ext_area +
    tumer_away_fr_primsite + Tumor_1_mrkr,
    data = data, method = "efron")

  n= 926, number of events= 635

                           coef exp(coef)  se(coef)      z Pr(>|z|)
trt_prime_site         -0.094003  0.910280  0.010691 -8.792  < 2e-16 ***
age                     0.039977  1.040787  0.003593 11.125  < 2e-16 ***
no_lymph_nodes          0.092868  1.097317  0.009966  9.318  < 2e-16 ***
tumor_size              0.016674  1.016814  0.003293  5.064 4.11e-07 ***
tumor_ext_area         -0.046524  0.954542  0.013417 -3.467 0.000525 ***
tumer_away_fr_primsite  0.025091  1.025408  0.008923  2.812 0.004925 **
Tumor_1_mrkr           -0.057414  0.944203  0.028307 -2.028 0.042529 *
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

                       exp(coef) exp(-coef) lower .95 upper .95
trt_prime_site            0.9103     1.0986    0.8914    0.9296
age                       1.0408     0.9608    1.0335    1.0481
no_lymph_nodes            1.0973     0.9113    1.0761    1.1190
tumor_size                1.0168     0.9835    1.0103    1.0234
tumor_ext_area            0.9545     1.0476    0.9298    0.9800
tumer_away_fr_primsite    1.0254     0.9752    1.0076    1.0435
Tumor_1_mrkr              0.9442     1.0591    0.8932    0.9981

Concordance= 0.747  (se = 0.01 )
Likelihood ratio test= 393  on 7 df,   p=<2e-16
Wald test            = 431.8  on 7 df,   p=<2e-16
Score (logrank) test = 466  on 7 df,   p=<2e-16
```

```
--------------- Variance Inflating Factor (VIF) ---------------
Multicollinearity Problem: Variance Inflating Factor (VIF)
is bigger than 10 (Continuous Variable)
or is bigger than 2.5 (Categorical Variable)

trt_prime_site     age          no_lymph_nodes              tumor_size
 1.135419    1.133260                1.419760                1.336365
tumor_ext_area        tumer_away_fr_primsite            Tumor_1_mrkr
1.287382                 1.287085                           1.018227
```

2.2.2 Penalized Methods for Variable Selections

Partial least squares regression involves replacing given predictor variables with smaller sets of linear combinations of the predictor variables. It has the advantage of finding directions that are aligned with either the signal of the data or the response, but the disadvantage is it can be difficult to interpret the new predictors. These techniques work directly with the predictors, and lead to models that are easier to interpret. Having arbitrarily large coefficients has no effect on minimizing SSE.

If we include a penalty related to the size of the coefficients, then minimizing this new cost function will keep the coefficients relatively small. In an ideal world, it would force some of them to be zero, or close enough to zero that it could be rounded to zero and removed from the model. It's still not a huge difference, but the large coefficients are now closer to the same size. In ridge regressions with highly correlated predictors, the coefficients associated with the highly correlated predictors tend to have about the same size. In other words, ridge regression tends to spread out the weight of the direction across all of the predictors. from the below table we see that some predictors shown NA of betas minimum information criterion standard error and the statistically insignificant.

Table 1: Lasso Regression Regression method for selecting importance predictors

	beta0	gamma	se.gamma	LB	UB	z.stat	p.value	beta.MIC	se.beta.MIC
race	0.000	0.000	0.045	-0.089	0.089	0.000	1.000	0.000	NA
marital	0.000	0.000	0.044	-0.087	0.087	0.000	1.000	0.000	NA
CStage	0.000	0.000	0.085	-0.167	0.167	0.000	1.000	0.000	NA
seer cstage	0.000	-0.067	0.048	-0.160	0.027	-1.395	0.163	-0.066	0.046
trt prime site	0.000	-0.339	0.067	-0.470	-0.207	-5.059	0.000	-0.339	0.047
trt surgy other region	0.000	0.000	0.038	-0.075	0.075	0.000	1.000	0.000	NA
surgery primesite	0.000	0.000	0.083	-0.162	0.162	0.000	1.000	0.000	NA
tumor size	0.000	0.253	0.057	0.142	0.364	4.463	0.000	0.253	0.047
tumor ext area	0.000	-0.141	0.052	-0.242	-0.039	-2.708	0.007	-0.141	0.048
no lymph nodes	0.000	0.441	0.057	0.329	0.552	7.752	0.000	0.441	0.049
Reg Node Eval	0.000	0.000	0.057	-0.111	0.111	0.000	1.000	0.000	NA
Reg nodes pos	0.000	-0.073	0.074	-0.217	0.071	-0.995	0.320	-0.073	0.052
tumer away fr primsite	0.000	0.134	0.053	0.030	0.238	2.520	0.012	0.134	0.049
Tumor 1 mrkr	0.000	0.000	0.118	-0.231	0.231	0.000	1.000	0.000	NA
Tumor 2 mrkr	0.000	-0.077	0.118	-0.309	0.155	-0.650	0.516	-0.077	0.043
Total tumors	0.000	-0.057	0.045	-0.144	0.031	-1.270	0.204	-0.055	0.044
borderline tumors	0.000	0.000	0.042	-0.081	0.082	0.000	1.000	0.000	NA
age	0.000	0.525	0.047	0.433	0.617	11.229	0.000	0.525	0.046

So, from the table we can say that those insignificant predictors is not important for prediction and further analysis. Similarly for the given below plot we see that same predictors shown not important considering the error bar of lambda and betas.

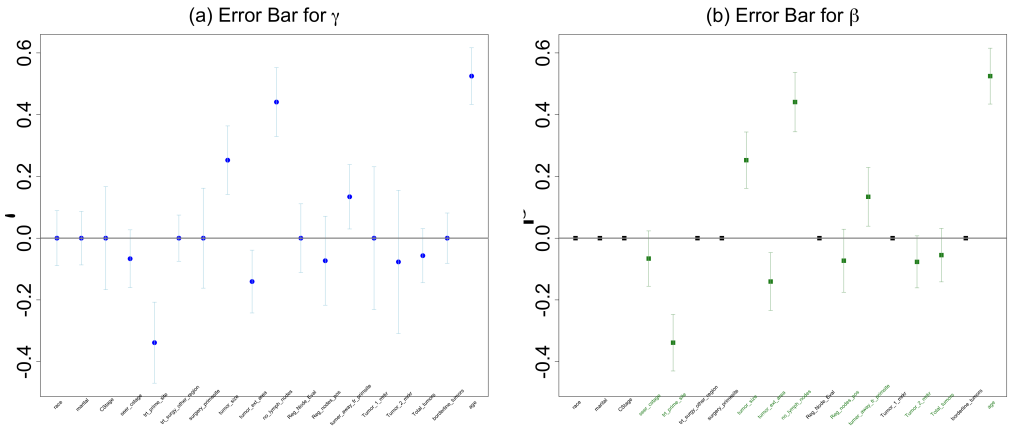

Figure 1: Penalized method for variables selections with 95% CI for the parameter lambda and betas

2.2.3 Predictions

According to Vanneschi et al. (2011) imputed data set give better prognosis and accuracy and reliability estimates. For excluding the missing data, fitting cox model then we get the following outcomes considering the important predictors

As mentioned above, we imputed the missing values by multiple imputation methods and create 10 data sets. Fitting cox model of these 10 dataset and show that all the outcomes is approximately similar. From the below

Table 2: Fitting Cox model with 95% CI by excluding the missing data

Variable	HazardRatio	CI.95	p-value
race	0.97	[0.79;1.18]	0.72579
CStage	1.03	[0.96;1.11]	0.34472
trt prime site	0.92	[0.89;0.95]	<0.001
trt surgy other region	1.17	[0.72;1.92]	0.52135
surgery primesite	1.2	[0.71;2.04]	0.49815
tumor size	1.02	[1.01;1.02]	<0.001
tumor ext area	0.96	[0.94;0.99]	0.01576
no lymph nodes	1.1	[1.08;1.13]	<0.001
Reg Node Eval	0.97	[0.85;1.10]	0.58837
Reg nodes pos	0.97	[0.94;1.00]	0.05731
tumer away fr primsite	1.03	[1.01;1.05]	0.00385
Tumor 1 mrkr	0.97	[0.84;1.12]	0.67206
Tumor 2 mrkr	0.98	[0.85;1.13]	0.77279
age	1.04	[1.03;1.05]	<0.001

table shown the estimated coefficients and corresponding standard error for each datasets.

Table 3: Multiple Imputation of missing data and create 10 dataset and fitting the cox model for each dataset shown the coefficient estimates and corresponding standard errors

	Dataset 1		Dataset 2		Dataset 3		Dataset 4		Dataset 5	
term	estimate	std.error	estimate	std.error	estimate	std.error	estimate	std.error	estimate	std.error
race	-0.0356	0.1015	-0.1199	0.0962	-0.2862	0.0960	-0.0452	0.0951	-0.2427	0.0985
CStage	0.0333	0.0352	0.0807	0.0379	0.3952	0.0518	0.0868	0.0346	0.1516	0.0376
trt prime site	-0.0860	0.0165	-0.1105	0.0170	-0.1389	0.0159	-0.0594	0.0166	-0.1121	0.0150
trt surgy other region	0.1605	0.2504	0.4276	0.3565	1.4369	0.4289	1.7728	0.3091	0.6508	0.3295
surgery primesite	0.1833	0.2705	0.3360	0.2595	1.6301	0.2635	0.4287	0.2153	0.8956	0.2194
tumor size	0.0158	0.0041	0.0110	0.0037	-0.0235	0.0041	0.0236	0.0039	0.0040	0.0040
tumor ext area	-0.0367	0.0152	-0.0376	0.0119	-0.0092	0.0144	0.0073	0.0138	-0.0141	0.0126
no lymph nodes	0.0981	0.0125	0.0761	0.0124	0.1241	0.0130	0.0497	0.0130	0.0928	0.0122
Reg Node Eval	-0.0356	0.0658	0.0338	0.0648	-0.0978	0.0677	-0.3065	0.0654	-0.0704	0.0642
Reg nodes pos	-0.0307	0.0162	-0.0340	0.0154	-0.1546	0.0199	-0.0564	0.0168	-0.0621	0.0147
tumer away fr primsite	0.0286	0.0099	0.0273	0.0094	0.0375	0.0094	0.0105	0.0094	0.0352	0.0096
Tumor 1 mrkr	-0.0320	0.0755	0.1934	0.1154	0.8712	0.1327	0.2124	0.0913	0.0820	0.0881
Tumor 2 mrkr	-0.0207	0.0718	-0.1964	0.1123	-0.9757	0.1300	-0.2439	0.0910	-0.0791	0.0827
age	0.0415	0.0037	0.0435	0.0038	0.0482	0.0037	0.0335	0.0037	0.0441	0.0038
	Dataset 6		Dataset 7		Dataset 8		Dataset 9		Dataset 10	
race	-0.4073	0.1007	-0.1665	0.0983	-0.0449	0.0991	-0.2193	0.0991	0.2909	0.1009
CStage	0.1050	0.0314	0.1514	0.0314	0.0582	0.0359	0.2283	0.0349	0.0868	0.0356
trt prime site	-0.1216	0.0157	-0.0255	0.0162	-0.1068	0.0160	-0.1105	0.0171	-0.1093	0.0160
trt surgy other region	0.9263	0.3351	0.3814	0.3018	0.3658	0.3506	1.2112	0.5928	1.2542	0.4272
surgery primesite	0.3367	0.2221	0.0771	0.2102	0.3759	0.2288	1.1122	0.2372	0.6153	0.2424
tumor size	-0.0062	0.0038	0.0115	0.0039	0.0125	0.0036	0.0009	0.0035	0.0267	0.0038
tumor ext area	-0.0254	0.0114	-0.0361	0.0123	-0.0481	0.0124	-0.0502	0.0160	-0.0732	0.0136
no lymph nodes	0.1112	0.0127	0.0359	0.0115	0.0860	0.0114	0.0642	0.0125	0.1209	0.0131
Reg Node Eval	0.1114	0.0606	-0.3732	0.0608	-0.0270	0.0651	-0.1086	0.0653	-0.1828	0.0604
Reg nodes pos	-0.0433	0.0154	-0.0417	0.0133	-0.0417	0.0157	-0.0568	0.0148	-0.0639	0.0171
tumer away fr primsite	0.0141	0.0097	0.0341	0.0100	0.0252	0.0097	0.0476	0.0086	-0.0072	0.0097
Tumor 1 mrkr	0.0185	0.0952	0.2215	0.0823	-0.1323	0.0669	0.3188	0.1330	0.0325	0.0787
Tumor 2 mrkr	-0.1693	0.0911	-0.2766	0.0789	0.1037	0.0606	-0.3564	0.1338	-0.1376	0.0788
age	0.0434	0.0037	0.0389	0.0038	0.0395	0.0036	0.0413	0.0036	0.0421	0.0037

By using Rubin Strine Methods, pooling these outcome and finally fitting the cox model, we find that the Hazard ratio of missing data and completed dataset are different. All the predictors have same directional output the magnitude is different.

Table 4: By using Rubin Stine Methods pooling the ten output of multiple imputation in a single output with Hazard ratio and 95% Confidence Interval

term	estimate	HaazardRatio	95% CI		p.value
race	-0.128	0.881	0.705	1.099	0.575
CStage	0.138	1.148	1.021	1.291	0.272
trt prime site	-0.098	0.907	0.873	0.942	0.025
trt surgy other region	0.859	2.361	1.192	4.674	0.226
surgery primesite	0.599	1.821	1.043	3.178	0.303
tumor size	0.008	1.008	0.992	1.024	0.645
tumor ext area	-0.032	0.968	0.942	0.995	0.262
no lymph nodes	0.086	1.09	1.054	1.127	0.027
Reg Node Eval	-0.106	0.901	0.761	1.064	0.542
Reg nodes pos	-0.059	0.943	0.906	0.982	0.178
tumer away fr primsite	0.025	1.026	1.006	1.045	0.206
Tumor 1 mrkr	0.179	1.196	0.878	1.628	0.576
Tumor 2 mrkr	-0.235	0.79	0.574	1.089	0.482
age	0.042	1.042	1.037	1.048	0.000

From the above analysis we can say that, age is a significant predictors for breast cancer. Similarly, With increasing the number of lymphs nodes, the risk of death for breast cancer patients significantly increased. Finally, treatment of breast cancer is significant factor for death of breast cancer patients.

3 Machine Learning Methods

Several literature found on machine learning approaches usually used in breast cancer detection. Very recently Jain and Kumar (2020), applied machine learning approaches for prediction the breast cancer and founded 97.89% accurately predict the breast cancer. Agarap (2018) used Machine learning method on Wisconsin Diagnostic dataset for detection of breast cancer and predict breast cancer events. Some advanced machine leaning method such as support vector network and artificial neural network methods used for cancer diagnosis and prognosis (Sharma et al. (2017) , Amrane et al. (2018) and Sharma et al. (2018)). Gayathri and Sumathi (2016) provided a comparative study of relevance vector machine with various machine learning methods for detection of breast cancer.

According to Obermeyer and Emanuel (2016) machine learning methods offers an alternative approach to standard prediction modeling that may address current limitations of survival methods and improve the accuracy of breast cancer prediction tools. Vanneschi et al. (2011) mentioned that this methods has been used in models related to cancer prognosis and survival and produced better accuracy and reliability estimates. Still today, very few studies applied machine methods for personalized breast cancer survival prediction or compared the predictive accuracy and reliability with models commonly used in clinic practice, but Heidari et al. (2018) mentioned that, machine learning approaches provided more accurate prediction

on short term breast cancer risk.

3.1 Imputations

Implemented in R in the missForest() package, it is a random forest imputation algorithm for missing data. MissForest initially imputes all missing data using the mean/mode, then for each variable with missing values, it fits a random forest on the observed part and then predicts the missing part. Training and predicting are repeated in an iterative process until a stopping criterion is reached, or a maximum number of user-specified iterations is reached.

This study using miss forest forest package of random forest methods for imputation the missing values and compared the accuracy of imputations. At nearly every level of missingness in every dataset, MissForest outperformed these other algorithms, in some cases reducing the imputation error by more than 50%. Moreover, it is easier to use than all of these alternatives and does not require tuning (because random forests are so effective at default parameters).

3.2 Variables Selections

Variable importance evaluation can be separated into two groups, those that use the model information and those that do not. The advantage of using a model-based approach is that is more closely tied to the model performance

and that it may be able to incorporate the correlation structure between the predictors into the importance calculation. Regardless of how the importance is calculated for most classification models, each predictor will have a separate variable importance for each class (the exceptions are classification trees, bagged trees and boosted trees). All measures of importance are scaled to have a maximum value of 100, unless the scale argument.

In random survival forest for each tree, the prediction accuracy on the out-of-bag portion of the data is recorded. Then the same is done after permuting each predictor variable. The difference between the two accuracy's are then averaged over all trees, and normalized by the standard error. For regression, the MSE is computed on the out-of-bag data for each tree, and then the same computed after permuting a variable. The differences are averaged and normalized by the standard error. If the standard error is equal to 0 for a variable, the division is not done.

There are several approaches have for measuring the importance score. Gini score is more popular for measuring the importance which give the means Gini score produced by overall trees. Gini Permutation importance measure the mean decrease in classification accuracy after permuting over all trees. Let t be the tree size then the Variable importance score will be

$$VI^{(t)}(x_j) = \frac{\sum_{i \epsilon \hat{\sigma}^{(t)}} I\left(y_i = \hat{y}_i^{(t)}\right)}{|\hat{\sigma}^{(t)}|} - \frac{\sum_{i \epsilon \hat{\sigma}^{(t)}} I\left(y_i = \hat{y}_{i,\pi_j}^{(t)}\right)}{|\hat{\sigma}^{(t)}|} \tag{1}$$

Where, $\hat{y}_i^(t) = f^t(x_i)$ = predicted class before permuting

$\hat{y}^{(t)}_{i,\pi_j} = f^t(x_{i,\pi_j})$ = predicted class after permuting of X_j

And $X_{i,\pi_j} = (x_{i,1}, x_{i,2}, \ldots, x_{i,j-1}, X_{\pi_j(i),j}, \ldots, X_{\pi_j(p),j}$

Note that, $VI^{(t)}(X_j) = 0$ by definition if X_j is not in the tree t.

By missforest, the minimum depth and corresponding relative frequency of the predictors is given bellow,

Table 5: Random Forest Model for variable selection by Minimum depth and relative frequency

Predictors	depth	rel.freq
tumor size	1.2898483	60
CStage	1.304	40
tumor away from prime site	1.516	40
Tumor marker 1	1.321	40
race	1.43	20
marital	1.326	20
trt surgary other region	2.156	20
surgery primesite	1.42116183	20
Number of lymph nodes	1.25518672	20
Reg Node Eval	1.0605428	20
Tumor marker 2	1.564	20
Total tumors	1.32157676	20
borderline tumors	2.25621974	20
seer cstage	NA	0
trt prime site	NA	0
tumor ext area	NA	0
Reg Node Eval	NA	0
age	NA	0

From the above table we see that tumor size have more relative frequencies but the depth is not minimum, on the other hand Regional node evaluation have lower minimum depth but the relative frequncies is not high. Some other predictors shown the relative frequencies is zeros and minimum depth

is not applicable.

For machine learning model, the family used survival and variable selection method is chosen minimum depth. The most effective criterion is selected medium values importance. In the model of random survival forest, number of tree is selected 1000 and number of each predictors splits 10. The node size is chosen 20. This methods suggested the predictors by considering 2.09% out of bag survival rate errors.

minimal depth variable selection ...

---

family : surv

var. selection : Minimal Depth

conservativeness : medium

x-weighting used? : TRUE

dimension : 18

sample size : 926

ntree : 1000

nsplit : 10

mtry : 5

nodesize : 20

refitted forest : FALSE

model size : 10

depth threshold : 6.1616

PE (true OOB) : 2.099

In random forest model, the variable importance score is called and base on this, this method suggest some important predictors. From the given bellow table we see that borderline number of tumors have most important score but the minimum depth is higher.

Table 6: Variable selection by machine learning methods with considering minimum depth and vimp

predictors	Minimum Depth	Variable imprtannce
race	9.861	13.348
marital	11.878	13.648
CStage	3.941	9.525
seer cstage	4.643	11.683
trt prime site	1.804	7.787
trt surgy other region	9.672	13.347
surgery primesite	7.578	13.322
tumor size	4.236	8.702
tumor ext area	5.014	8.934
no lymph nodes	3.656	8.319
Reg Node Eval	8.219	12.308
Reg nodes pos	5.436	9.278
tumer away fr primsite	4.044	9.649
Tumor 1 mrkr	6.206	10.637
Tumor 2 mrkr	6.887	11.566
Total tumors	5.773	9.673
borderline tumors	13.462	13.853
age	2.395	7.248

Similarly the relative importance score we can find the most importance predictors. From the below table we see that borderline tumors have more importance.

So the list of top variable suggesting by machine learning methods by considering least square methods distance is given bellow, which is approxi-

predictors	Relative importance by percentile
race	0.70708371
marital	0.8396174
CStage	0.27366726
seer cstage	0.33310243
trt prime site	0.11582087
trt surgy other region	0.68508523
surgery primesite	0.52220753
tumor size	0.29412606
tumor ext area	0.35829901
no lymph nodes	0.25041578
Reg Node Eval	0.59309487
Reg nodes pos	0.39227935
tumer away fr primsite	0.28018048
Tumor 1 mrkr	0.44230231
Tumor 2 mrkr	0.48912131
Total tumors	0.41091607
borderline tumors	0.93351379
age	0.14473263

mately same as lasso and ridge regression methods provided.

```
trt_prime_site
age
no_lymph_nodes
CStage
tumer_away_fr_primsite
tumor_size
seer_cstage
tumor_ext_area
Reg_nodes_pos
Total_tumors
```

3.3 Predictions

Classification and regression trees (CART) have proved a valuable resource in modelling complex and non-linear effects for categorical and continuous outcomes, which has motivated a range of attempts to adapt tree methodology to right-censored survival outcomes. Ishwaran et al. (2008) proposed a method for analysis survival data that very helpful for fitting and implementation. Any tree method relies on the recursive binary partitioning of a given features space into increasingly smaller regions, containing observations with similar responses, until a certain stopping criterion is reached. The regions are referred to as nodes, and the final regions created on termination of the growth of a tree are known as terminal nodes, or leafs, while the initial node is commonly referred to as the root node. There are two necessary features to any tree algorithm: a node splitting rule, which informs how a partition is best split and a stopping rule, according to Zhou and McArdle (2015), which provides a criterion for termination the growth of a tree . For a given feature space with x possible variables, and c possible split values, variable x* and split point c* are chosen in a way that maximizes the difference between the two daughter nodes.

While a tree has great potential for modelling non-linear and interactive effects, as will be detailed later, the very way a tree is grown makes it highly variable and very unstable. Each split of a node is dependent on previous partitioning [Strobl et al. (2009)]. Consider a situation in which several samples are taken from a larger dataset and used to grow individual trees.

Even a slight variation in the data may result in a selection of a different variable x* or splitting point c* for one of the early nodes, or even the root node, giving rise to vastly different trees. The sensitivity of a single tree to minor training data variations is likely to result in poor generalization to new data. Breiman (1996) mentioned that, a way to combat the over fitting reduce the variability - is by introducing a measure of randomness in the way of bootstrapping. In this procedure individual trees are grown for multiple bootstrap samples drawn from the data, and subsequently aggregated over, producing a single ensemble tree. This procedure of bootstrap aggregation, is commonly known as bagging.

The survival difference can be quantified using the log-rank statistic or the log-rank score statistic. The log-rank test has been shown to be a valid test for splitting survival trees given both proportional and non-proportional hazards. For a split on a continuous predictor defined as $x < c$ and $x > c$ let $t_1, \ldots, t_m$ be the event times in the parent node h. Let $d_{k,l}$ and $y_{k,l}$ be the number of events and subjects at risk at time k in the left daughter node, respectively, and let $d_{k,r}$ and $Y_{k,r}$ be the same for the right daughter node, then $Y_{k,l}$ will be all the subject i at risk time k with covariates values $x_i \leq c$ and $Y_{k,r}$ the same for subjects with $x_i \geq c$.

$$\begin{aligned} Y_{k,l} &= \#T_i \geq t_k, x_i \leq c \\ Y_{k,r} &= \#T_i \geq t_k, x_i \geq c \end{aligned} \tag{2}$$

Let Y_k and d_k be the total number of subjects at risk time k, and total event at time k, then the log-rank statistic for the split point on the variable x is

$$L(c,x) = \frac{\sum_{k=1}^{m}\left(d_{k,l} - Y_{k,l}\frac{d_k}{Y_K}\right)}{\sqrt{\sum_{k=1}^{m}\frac{Y_{k,l}}{Y_k}\left(1-\frac{Y_{k,l}}{Y_k}\right)\left(\frac{Y_k-d_k}{Y_k-1}\right)d_k}} \tag{3}$$

The large the value of $|L(c,x)|$ the greater difference between the survival curves and the greater the node separation. The objective is to find a best variable x* with an optimal split c* such that $|L(c*,x8)| \geq |L(c,x)|$ for all variables x and split point c. An alternative splitting rule is given by the log-rank score statistic, which differs from the log-rank statistic in the assumption that the variable x is ordered. The particular splitting rule used, the number of variables considered at each split along with the number of split points must all be considered when growing a tree.

The survival time, and censoring status for individuals $i = 1, \ldots, n$ can then be written as $(T_{1,h}, \delta_{1,h}), \ldots, (T_{n(h),h}, \delta_{n(h),h})$ where, $\delta_{i,h} = 0$ denotes a subject right censored at time $T_{i,h}$ otherwise the event occurs. Define the n(h) distinct event times as $t_{1,h}, t_{2,h}, \ldots, t_{n(h),h}$. At time, $t_{1,h}$, $d_{1,h}$ is the number of death and $Y_{1,h}$ is the subjects at risk, then, Cumulative hazard function (CHF) for a terminal event can be estimated as

$$\hat{H}_h(t) = \sum_{t_{l,h} \leq t}^{n} \frac{d_{l,h}}{Y_{l,h}} \tag{4}$$

So, At time, $t_{1,h}$, $d_{1,h}$ is the number of death and $Y_{1,h}$ is the subjects at risk,

then, survival function for a terminal event can be estimated as

$$S_h(t) = e^{-\hat{H}_h(t)} \tag{5}$$

To obtain the ensemble cumulative hazard function, the average over the survival trees is taken. Defining the hazard for a tree grown from a bootstrap sample b as $H_b^*(t|x)$, then the bootstrap ensemble $H_e^*(t|x_i)$ is given by

$$H_e^*(t|x_i) = \frac{1}{B}\sum_{b=1}^{B} H_b^*(t|x) \tag{6}$$

Where, B is the number of survival trees. For a subject i with covariates vectors x_i all survival trees are used. This approach gives a CHF estimate which, given a large enough number of trees, is comparable to one that would be obtained with a leave-one-out cross-validation. Such a CHF estimate for a subject i is valid, as it has been obtained from prediction using only those trees in which i was not included as an observation. Then, the indicators, $I_{i,b}$ is used to select the correct trees and will equal to 1 if the observation in OOB and 0 if it lies in-bag. Then the survival probability will be

$$S_e(t) = exp\left(-(H_e^{**}(t|x_i)\right) = exp\left\{-\frac{\sum_{b=1}^{B} I_{i,b} H_b^*(t|x)}{\sum_{b=1}^{B} I_{i,b}}\right\} \tag{7}$$

In a random survival forest context, the prognostic index is replaced by a predicted outcome: the ensemble mortality. When choosing the latter option, the ensemble mortality scores for the subjects of each test set are obtained

by dropping test observations down the trees of a survival random forest grown from the training set, and calculating their ensemble CHF and from the ensemble CHF the ensemble mortality. Ensemble mortality is based on the conservation of events principle. The conservation of events for a given terminal node $h \epsilon T$ in a given tree can be written as

$$\sum_{i=1}^{n(h)} \hat{H}_h(T_{i,h}) = \sum_{i=1}^{n(h)} \delta_{i,h} \tag{8}$$

For each terminal nodes $h \epsilon T$, which shows that the total number of deaths is conserved within h. Summing the estimated CHF over all observed survival times over all terminal nodes h amounts to the total number of deaths.

$$\sum_{i=1}^{n} H(T_i|x_i) = \sum_{h \epsilon T} \sum_{i=1}^{n(h)} \hat{H}_h(T_{i,h}) = \sum_{h \epsilon T} \sum_{i=1}^{n(h)} \delta_{i,h} = \sum_{i=1}^{n} \delta_i \tag{9}$$

From the below plot we see that age is the most important predictor and for less than 41 years have higher survival rate compare with greater than 41 years. Similarly, treatment on prime site is coded as resection breast cancer patients have higher survival compare with autopsy treatment.

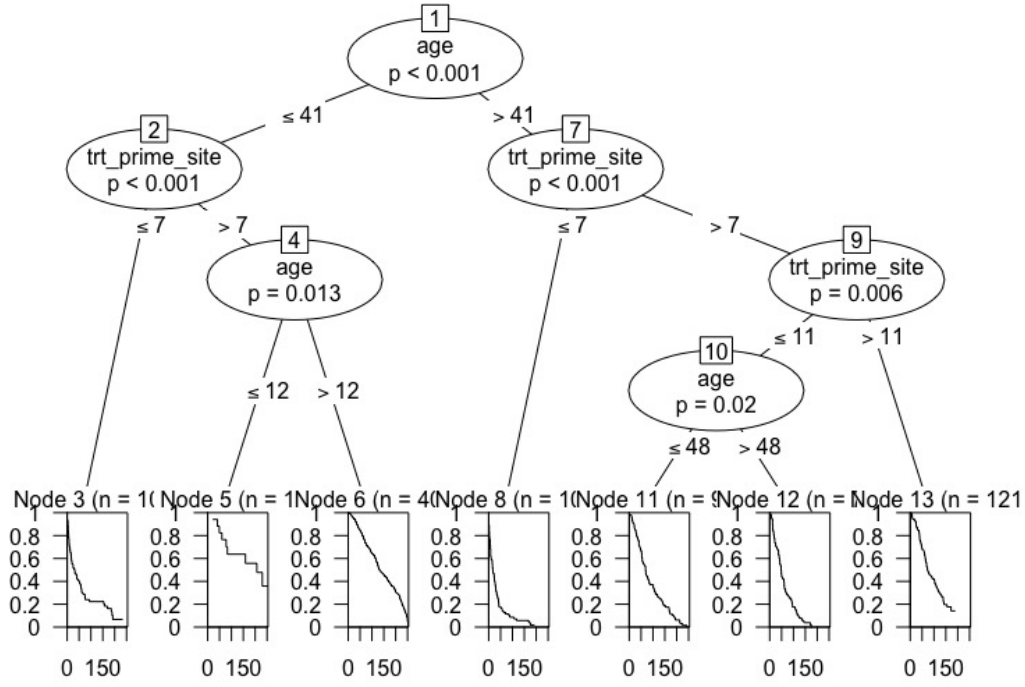

Figure 2: Overall predicted survival tree for random survival forest methods

4 Comparison and Validations

For comparison the prediction performance of different model different methods are available. Usually prediction errors, mean square errors, AIC, BIC, sensitivity and specificity analysis, ROC curve, AUC and so on measurements are used. For comparison of random survival forest model and cox model, OOB survival, OOB mortality curve, AUC, Brier score and predictors errors are more popular. From the below table we see that the Concordance index for cox proportional hazard model and random survival forest model. From this table we see that C-index is higher for Random survival models compare with Cox PH model over time time points. We also see that, from the figure 3 random forest model have higher C index over time than Cox model.

Table 7: Concordance Index over time for Cox proportional hazard model and random survival forest model

time	C-Index for Cox Model	C-Index for RSF Model
1	0.821466091	0.905029094
11	0.74304913	0.780378847
55	0.736865321	0.749881641
70	0.785570961	0.866022203
77	0.766200981	0.822651596
103	0.739070831	0.777688579
104	0.739305766	0.776382946
123	0.735961566	0.734800307
142	0.732378452	0.679890675
198	0.745957568	0.791398863

The reference is the no predictors model and consider the time and censoring indicator are are only associated each others. From the figure 4 we

also see that, cox model have higher prediction error compare with random survival forest over the time points. So, for our dataset we can conclude that the machine learning approaches give better prediction with lower prediction errors and higher Concordance.

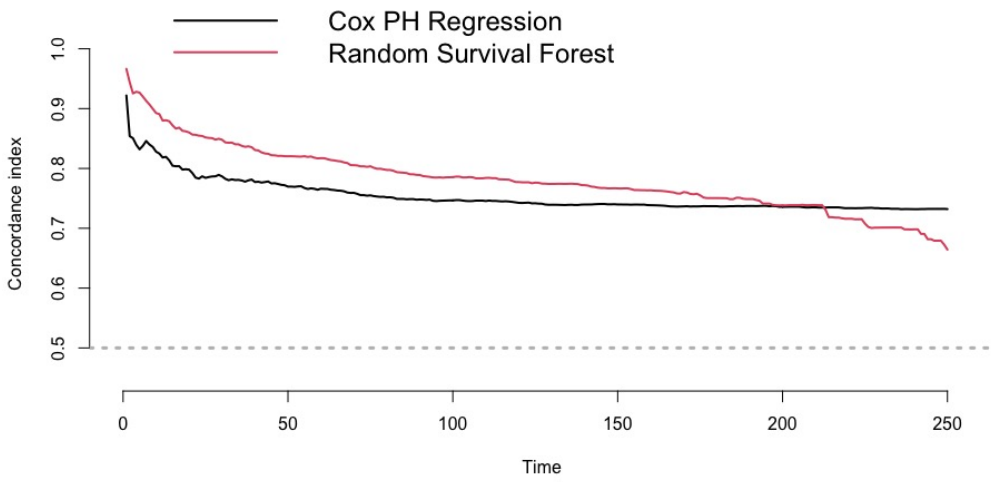

Figure 3: Concordance Index of the random survival forest and cox model over times

From the below table 8 we see that the predictor errors of random survival model lower compare with null model and cox model. The integrated Brier score also lower for random survival model shown in table below.

Table 8: Prediction errors over time for Null model, Cox proportional hazard model and random survival forest model with number of risk

time	n.risk	Reference	Cox PH Regression	Random Survival Forest
0	926	0.000	0.000	0.000
63.5	436	0.244	0.176	0.130
127	210	0.227	0.181	0.131
190.5	90	0.165	0.134	0.092
254	0	0.001	0.004	0.009

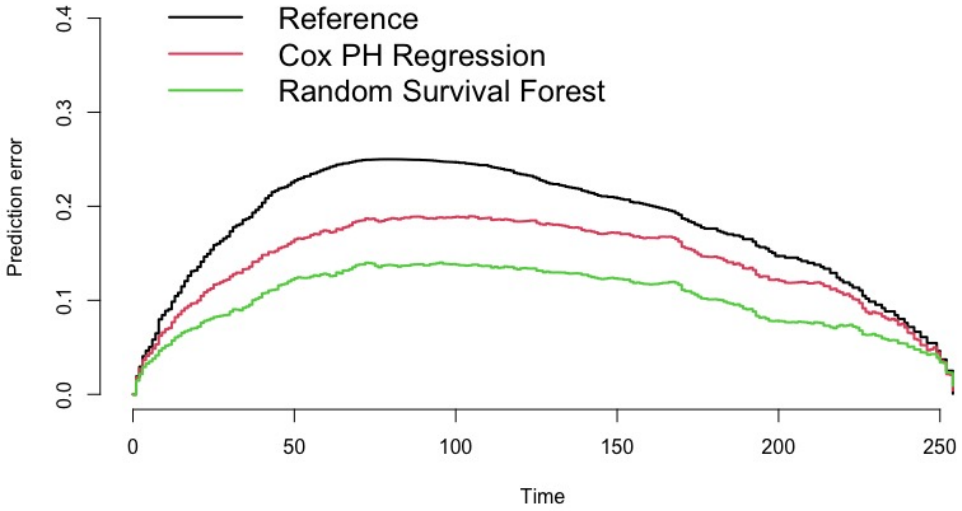

Figure 4: Prediction error for Cox model, Random survival forest model

Table 9: Integrated Brier Score for Null model, Cox proportional hazard model and Random survival forest model

Integrated Brier score (crps):	IBS [0;time=254)
Reference	0.179
Cox PH Regression	0.141
Random Survival Forest	0.101

5 Conclusions

Overall, considering machine learning approaches may be more powerful tools for predicting the survivability of male breast cancer patients. So, the main purpose in this study was considering regression tree methods for survival

data and predicted the survivability of male breast cancer in Detroit Michigan. By the random survival forest methods, we find that age is the most potential factors for death of male breast cancer patients. Primary tumor diameter is an another important features that, with increasing of diameter, the change of survivability decreasing.

The concordance index over time for random survival forest model is lower compare with cox model and the predictor errors of this model is lower over times compare with cox model. Integrated brier score is higher for cox model compare with random survival forest model. So, we can says that, machine learning model is fitted better for our data set.

References

Agarap, A. F. M. (2018). On breast cancer detection: an application of machine learning algorithms on the wisconsin diagnostic dataset. In *Proceedings of the 2nd international conference on machine learning and soft computing*, pages 5–9.

Amrane, M., Oukid, S., Gagaoua, I., and Ensari, T. (2018). Breast cancer classification using machine learning. In *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*, pages 1–4. IEEE.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140.

Gayathri, B. and Sumathi, C. (2016). Comparative study of relevance vector machine with various machine learning techniques used for detecting

breast cancer. In *2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, pages 1–5. IEEE.

Heidari, M., Khuzani, A. Z., Hollingsworth, A. B., Danala, G., Mirniaharikandehei, S., Qiu, Y., Liu, H., and Zheng, B. (2018). Prediction of breast cancer risk using a machine learning approach embedded with a locality preserving projection algorithm. *Physics in Medicine & Biology*, 63(3):035020.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., Lauer, M. S., et al. (2008). Random survival forests. *Annals of Applied Statistics*, 2(3):841–860.

Jain, S. and Kumar, P. (2020). Prediction of breast cancer using machine learning. *Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science)*, 13(5):901–908.

Kopans, D. B. (2008). Basic physics and doubts about relationship between mammographically determined tissue density and breast cancer risk. *Radiology*, 246(2):348–353.

Obermeyer, Z. and Emanuel, E. J. (2016). Predicting the future—big data, machine learning, and clinical medicine. *The New England journal of medicine*, 375(13):1216.

Sharma, A., Kulshrestha, S., and Daniel, S. (2017). Machine learning approaches for breast cancer diagnosis and prognosis. In *2017 International conference on soft computing and its engineering applications (icSoftComp)*, pages 1–5. IEEE.

Sharma, S., Aggarwal, A., and Choudhury, T. (2018). Breast cancer detection using machine learning algorithms. In *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, pages 114–118. IEEE.

Strobl, C., Malley, J., and Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological methods*, 14(4):323.

Vanneschi, L., Farinaccio, A., Mauri, G., Antoniotti, M., Provero, P., and Giacobini, M. (2011). A comparison of machine learning techniques for survival prediction in breast cancer. *BioData mining*, 4(1):1–13.

Zhou, Y. and McArdle, J. J. (2015). Rationale and applications of survival tree and survival ensemble methods. *Psychometrika*, 80(3):811–833.

... The End ...