final report

2023-12-19

Abstract

Introduction

Data and Methods

Descriptive Data

Data Cleaning and Preparation

The dataset utilized in this study was derived from a comprehensive breast cancer database. The initial step in data preparation involved the standardization of variable names to ensure consistency. In addition, we converted several categorical variables into factors with defined levels. Specifically, we recoded the survival status variable into a binary format with `Dead` as 1 and `Alive` as 0. Finally, a new variable `node_positive_prop` was created and calculated based on the ratio of `reginol_node_positive` to `regional_node_examined`. This variable represents the proportion of examined nodes that were found to be positive.

Variable and Model Selection Procedures

After preprocessing the dataset, we subdivided our dataset into categorical and numerical variables in order to have a general outline data patterns. In **Table 1**, we summarized the essential statistics of all the categorical variables, which includes variable names, number of missing values, unique and top counts.

For numerical variables, we employed boxplot visualizations to effectively represent their distribution patterns. As illustrated in **Figure 1**, these boxplots serve as a comprehensive visualization, which include potential outliers, important quartiles, and medians of all the numerical variables.

After gaining an initial understanding of the data trends, we employed both stepwise selection and regularization techniques like LASSO and Ridge Regression for selecting the most appropriate model.

Initially, a comprehensive model incorporating all available predictors was developed, targeting survival status as the response variable. The corresponding estimates, standard errors, and P-values are illustrated in details in **Table 2**. In addition, we conducted a series of diagnostic evaluations on this full model. Our first step in this process was to assess multicollinearity, the results of which are presented in the Variance Inflation Factor (VIF) table depicted in **Table 3**. After the implementation of stepwise selection methods, along with LASSO and Ridge Regression, we proceeded to evaluate the classification accuracy of all the models. This was achieved through the generation of Receiver Operating Characteristic (ROC) curves and the analysis of the Area Under the Curve (AUC) statistics. The summary of all AUC statistics are included in **Table 4**. In addition, all ROC curves arr illustrated with **Figure 2**, **Figure 3**, **Figure 4**, and **Figure 5**.

Assumption Checking

Result

Variable Selection

Model Selection

Conclusion

Appendix

Table

Table 1: Summary Statistics of Categorical Variables

Variable	Missing	Unique Counts	Top Counts
race	0	3	Whi: 3413, Oth: 320, Bla: 291
marital_status	0	5	Mar: 2643, Sin: 615, Div: 486, Wid: 235
t_stage	0	4	T2: 1786, T1: 1603, T3: 533, T4: 102
n_stage	0	3	N1: 2732, N2: 820, N3: 472
x6th_stage	0	5	IIA: 1305, IIB: 1130, III: 1050, III: 472
differentiate	0	4	Mod: 2351, Poo: 1111, Wel: 543, Und: 19
grade	0	4	2: 2351, 3: 1111, 1: 543, ana: 19
a_stage	0	2	Reg: 3932, Dis: 92
estrogen_status	0	2	Pos: 3755, Neg: 269
progesterone_status	0	2	Pos: 3326, Neg: 698
status	0	2	0: 3408, 1: 616
node_positive_prop	0	NA	NA

Table 2: Table 2: Full Model Summary

Term	Estimate	Standard Error	P Value
(Intercept)	-3.926	0.461	0.000
age	0.024	0.006	0.000
raceBlack	0.515	0.162	0.002
raceOther	-0.416	0.203	0.040
marital_statusMarried	-0.132	0.135	0.327
marital_statusDivorced	0.082	0.175	0.641
marital_statusSeparated	0.721	0.383	0.060
marital_statusWidowed	0.098	0.219	0.653
t_stageT2	0.279	0.195	0.153
t_stageT3	0.542	0.314	0.084
t_stageT4	0.949	0.450	0.035
n_stageN2	0.562	0.241	0.020
n_stageN3	0.586	0.305	0.055
x6th_stageIIB	0.216	0.232	0.352
x6th_stageIIIA	-0.101	0.295	0.733
x6th_stageIIIB	0.053	0.529	0.921
x6th_stageIIIC	NA	NA	NA
differentiatePoorly differentiated	0.391	0.105	0.000

Term	Estimate	Standard Error	P Value
differentiateUndifferentiated	1.364	0.535	0.011
differentiateWell differentiated	-0.533	0.184	0.004
grade2	NA	NA	NA
grade3	NA	NA	NA
gradeanaplastic; Grade IV	NA	NA	NA
a_stageRegional	-0.060	0.266	0.821
tumor_size	0.000	0.004	0.992
estrogen_statusNegative	0.737	0.178	0.000
progesterone_statusNegative	0.589	0.128	0.000
regional_node_examined	-0.021	0.011	0.053
reginol_node_positive	0.055	0.020	0.007
node_positive_prop	0.590	0.316	0.062

Table 3: Table 3: VIF for Full Model

Term	VIF	CI_low	CI_high	SE_factor	Tolerance	Tolerance_low	Tolerance_high
age	1.1	1.1	1.2	1.1	0.9	0.9	0.9
race	1.1	1.0	1.1	1.0	0.9	0.9	1.0
marital_status	1.1	1.1	1.2	1.1	0.9	0.8	0.9
t_stage	30.6	28.8	32.5	5.5	0.0	0.0	0.0
n_stage	31.8	30.0	33.8	5.6	0.0	0.0	0.0
x6th_stage	61.7	58.0	65.5	7.9	0.0	0.0	0.0
differentiate	1.1	1.1	1.2	1.1	0.9	0.9	0.9
a_stage	1.3	1.2	1.3	1.1	0.8	0.8	0.8
tumor_size	3.7	3.5	3.9	1.9	0.3	0.3	0.3
estrogen_status	1.5	1.4	1.5	1.2	0.7	0.6	0.7
progesterone_status	1.4	1.4	1.5	1.2	0.7	0.7	0.7
regional_node_examined	3.4	3.3	3.6	1.9	0.3	0.3	0.3
reginol_node_positive	7.3	6.9	7.8	2.7	0.1	0.1	0.1
node_positive_prop	4.4	4.2	4.7	2.1	0.2	0.2	0.2

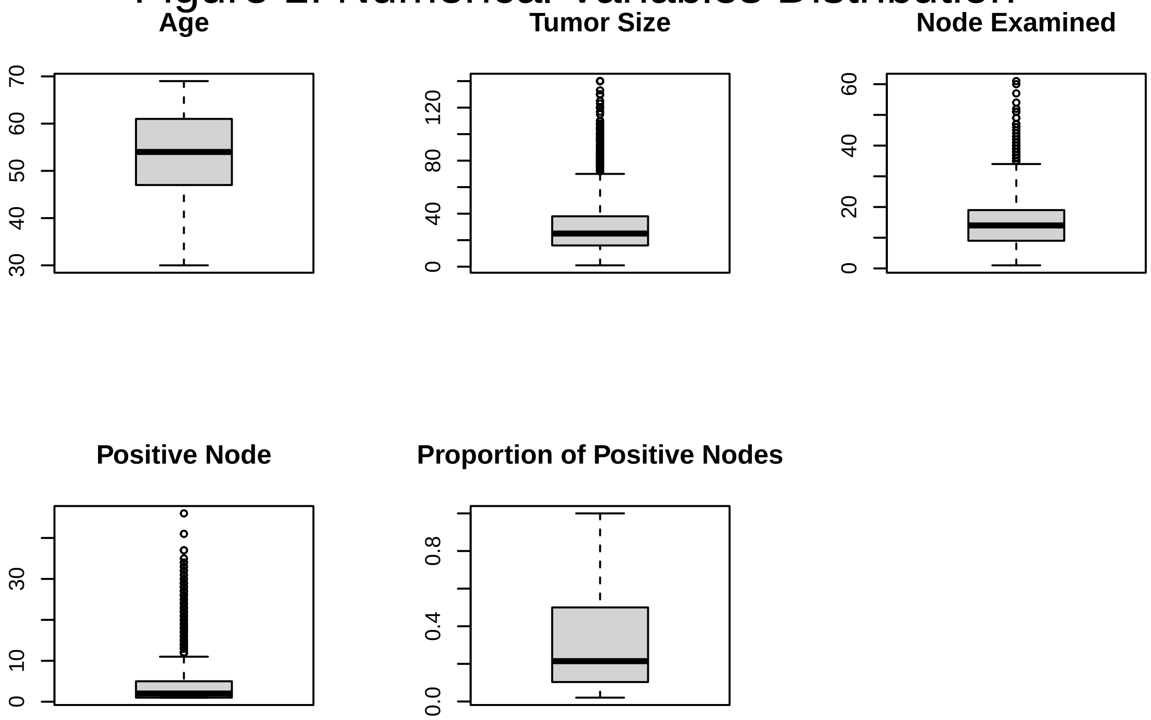
Table 4: Table 4: Backward Model Summary

Term	Estimate	Standard Error	P Value
(Intercept)	-4.043	0.364	0.000
age	0.024	0.005	0.000
raceBlack	0.571	0.159	0.000
raceOther	-0.436	0.202	0.031
t_stageT2	0.415	0.113	0.000
t_stageT3	0.537	0.149	0.000
t_stageT4	1.081	0.243	0.000
n_stageN2	0.359	0.133	0.007
n_stageN3	0.483	0.239	0.043
differentiatePoorly differentiated	0.390	0.105	0.000
differentiateUndifferentiated	1.343	0.527	0.011
differentiateWell differentiated	-0.514	0.183	0.005
estrogen_statusNegative	0.737	0.177	0.000
progesterone_statusNegative	0.598	0.127	0.000
regional_node_examined	-0.021	0.011	0.053

Term	Estimate	Standard Error	P Value
reginol_node_positive	0.056	0.020	0.005
node_positive_prop	0.603	0.314	0.054

Figure

Figure 1: Numerical Variables Distribution



Contribution