Build Classification Model to Help Predict the Survival Results of Breast Cancer Patients

Zirong Huang, Zhuoqi Cai, and Shuhao Wang
Department of Data Science, University of Southern California
DSCI 550: Data Science at Scale
Dr. Seon Ho Kim
December 4, 2022

1. Motivation and Background

1.1 Project idea & Problem Definition

This project topic is not only important to understand the main factors that influence the high possibility of death from breast cancer, but also to raise people's awareness about breast cancer. Increasing breast cancer cases have forced people all over the world to raise awareness about the significance of identifying breast cancer and the urgency of taking action. This project is aimed to answer: 1) Which classification model, logistic, KNN, or AdaBoost, is the best model when it comes to predicting the survival result (i.e. patient has survived or dead from the breast cancer) of breast cancer patients? 2) What are the most influential variable(s) when predicting the outcome of breast cancer patients?

1.2 Project Plan & Methods

The three classification methods that we choose are Logistic Regression, KNN, and AdaBoost. We are going to compare the performance of those three models using the ROC and AUC. Brownless (2021) suggested using ROC plot and AUC as the evaluation metrics for imbalanced data if both classes are important, and we follow the suggestion and use ROC plot and AUC to determine the best performance model. In addition, we also constructed baseline models- models directly trained from the "try" dataset (i.e. original distributed dataset without downsampling) to illustrate how downsampling an imbalanced dataset can improve the overall model performance. Baseline models follow the same training procedures as the downsampled model, the only difference is the dataset that those models are trained from. For getting the most influential variable, we are going to use the variable importance plot from Adaboost to accomplish this task.

1.3 Description of the Dataset

This database of breast cancer patients was acquired from the SEER Program of the NCI's November 2017 update, which offers details on population-based cancer statistics. The dataset included female patients who had been diagnosed between 2006 and 2010 with breast cancer. In the end, 4024 patients were included after the exclusion of patients with uncertain tumor sizes, studied regional LNs, positive regional LNs, and patients whose survival months were less than one month. Since the dataset is provided by the National Cancer Institute (NCI) (the federal government's principal agency for cancer research and training), and the dataset does

not have missing values or duplicated data, it has high reliability to be considered valid and valued data to analyze.

2. Basic Analysis

Based on the boxplot of Regional Node Positive and Tumor size, we can easily find that the distributions of Alive and Dead are right skewed with the median of Dead being significantly higher than the median of Alive, but there are many outliers in survival. From these graphs, we can guess that the number of Regional Node Positives and Tumor size are proportional to the probability of death.

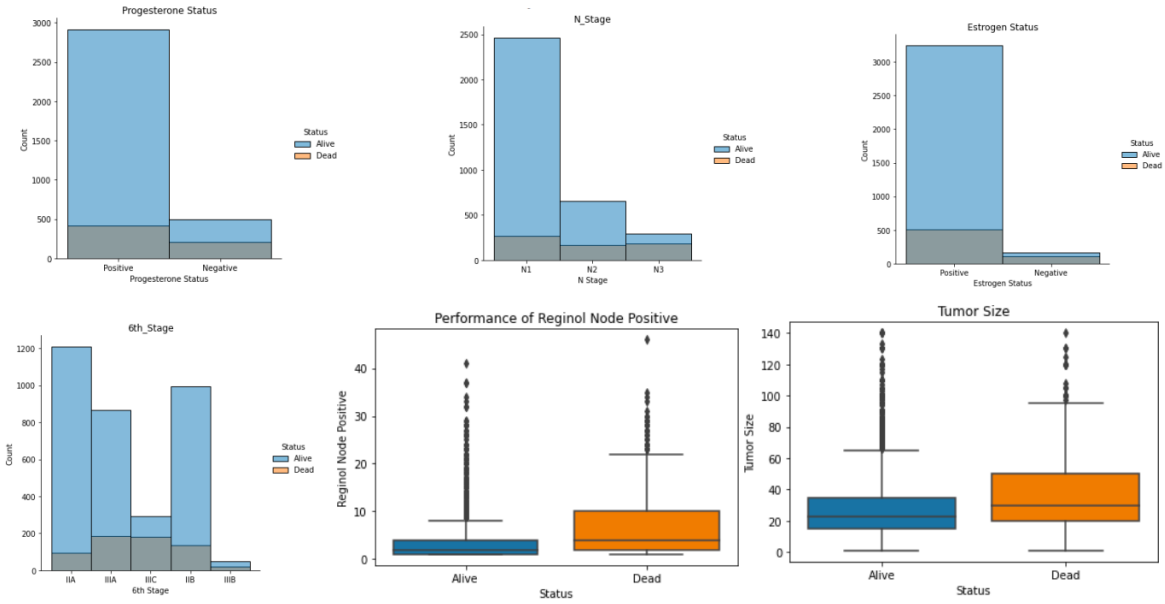

Figure 1 : boxplot and displot of potential factors

By the graph of Progesterone Status, we can observe that the number of positive patients is much higher than the number of negative patients but the death rate of negative patients is around 50%. In the N_Stage, the number of patients in N1 is much more than in other stages and the death rate in N3 is the highest. In the observation of Estrogen Status, the number of positive patients is higher than the number of negative patients but the death rate of negative patients is much higher. By the graph of 6th_Stage, we can notice that the most alive number is in the IIA and the IIIC has the highest death rate. In this case, we will explore these factors in more depth by Logistic Regression, KNN, and AdaBoost.

3. Deal with Imbalanced Data

3.1 Original Data

We first check the dataset and find out the data is imbalanced. The dataset has about 85% record has Status_Dead = 0 and only 15 % of the dataset has Status_Dead = 1 (0 - the patient has survived, 1- the patient has died).

3.2 Stratified Sampling

We separated the dataset into two parts using stratified sampling. One part is for training our model (the “try” set), and the other part is for testing our model (the “ultimate” set). After stratified sampling, “try” set has 0 - 2896 records and 1 - 524 records. “Ultimate” set has 0 - 512 records and 1 - 92 records. (0 - the patient has survived, 1- the patient has dead): The reason that we did stratified sampling is because we want to make sure that we get accurate test results.

3.3 Downsampling

To better deal with unbalanced data and avoiding our models overfitting on the majority data(0 - the patient has survived), we decided to perform random downsampling on our “try” dataset. After downsampling, 0 - 524 records and 1 - 524 records. We will use the “res” dataset for model construction.

4. Logistics Regression

4.1 Bivariate Analysis

An efficient method for selecting non-redundant features is the Bivariate Analysis, which is a method based on correlation that analyzes the relationship between pairs of elements. In calculating correlations between variables, we can observe some high correlated features (values greater than 0.9). Hence, we eliminate 'Moderately differentiated', 'Poorly differentiated', 'N Stage_N3','Undifferentiated', and 'Well differentiated'.

4.2 Select Proper Features

We use Recursive Feature Elimination, which is based on the idea to repeatedly construct a model and choose either the best or worst performing feature, setting the feature aside and then repeating the process with the rest of the features. After getting all the features, we implement them in the model. By removing features whose p-value > 0.05, we finally decide 6 features shown in Figure 3.

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Race_Black	0.6517	0.2371	2.7492	0.0060	0.1871	1.1163
N Stage_N1	-0.7436	0.1046	-7.1067	0.0000	-0.9487	-0.5386
6th Stage_IIIC	0.8844	0.1855	4.7668	0.0000	0.5208	1.2481
Grade_1	-0.6189	0.2288	-2.7051	0.0068	-1.0672	-0.1705
Grade_3	0.3894	0.1467	2.6537	0.0080	0.1018	0.6769
Progesterone Status_Negative	0.6998	0.1596	4.3833	0.0000	0.3869	1.0127

Figure 3 Logit Result

4.3 Logistic Regression Performance

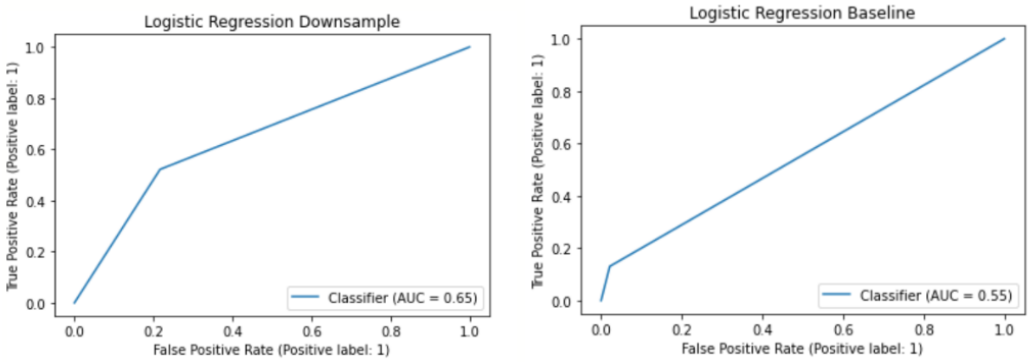

Figure 4 AUC based on 'ultimate' test set (left: from downsampled logistic model, right: from baseline logistic model)

In Figure 4, downsample helps improve the AUC from 0.55 to 0.65 (about 18% increase). When the Progesterone Status is Negative, the 6th Stage reaches to Stage_IIIC, and the Grade reaches to Grade 3, the patients are more likely to die.

5. KNN

5.1 Choose Best k

To choose the best K value, we compared the AUC that yielded from different K (ranged 1 to 30) and chose k=23 as our final KNN model since it yields the highest AUC for the downsampled KNN model.

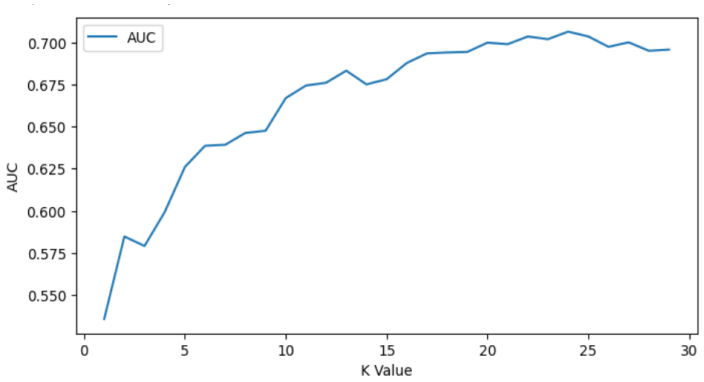

Figure 5 AUC resulted from different K-value KNN downsampled models

5.2 KNN Performance

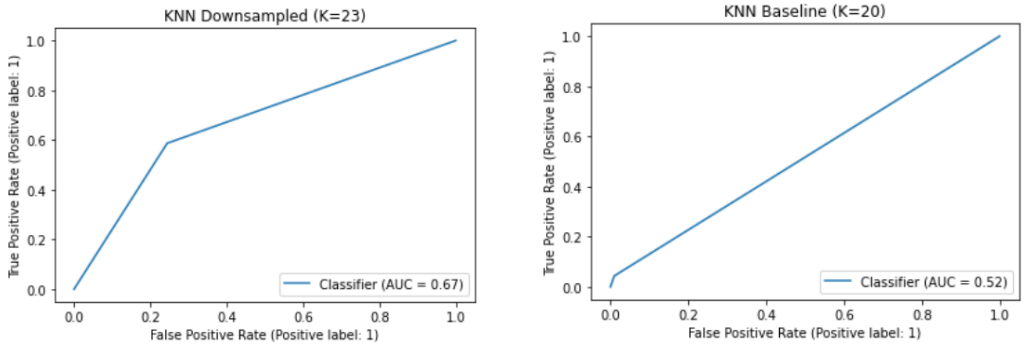

Figure 6 AUC based on 'ultimate' test set (left: from downsampled KNN model, right: from baseline KNN model)

By downsampling the dataset, AUC has increased from 0.52 to 0.67, which is a 28% increase.

6. AdaBoost

6.1 Choose Best n_estimator

Following the similar procedure, to choose best n_estimator value, we compared the AUC that yielded from different n estimator(ranged 1 to 30) and chose n_estimator=23 as our final AdaBoost model since it yields the highest AUC for the downsampled AdaBoost model.

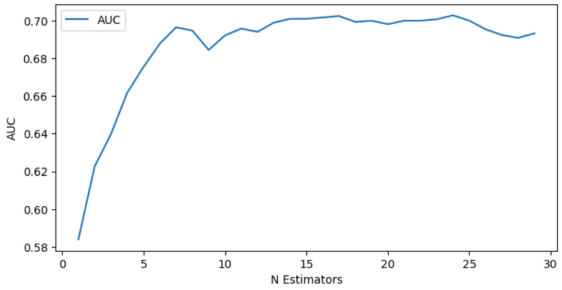

Figure 7 AUC resulted from different n_estimator AdaBoost downsampled models

6.2 AdaBoost Performance

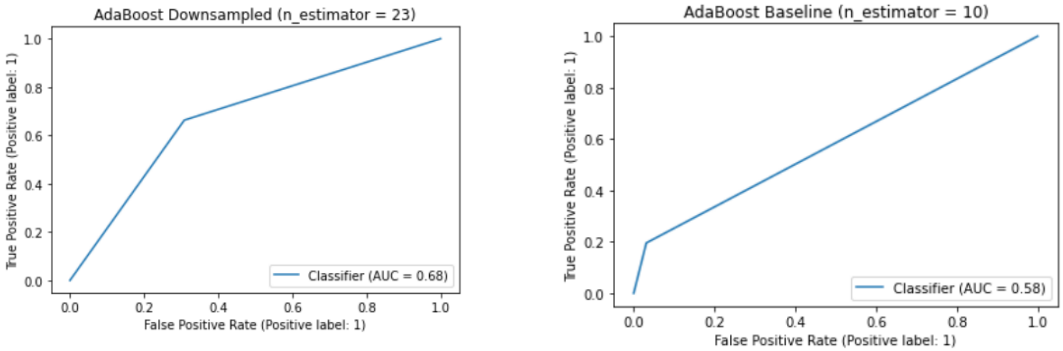

Figure 8 AUC based on 'ultimate' test set (left: from downsampled AdaBoost model, right: from baseline AdaBoost model)

By downsampling the dataset, AUC has increased from 0.58 to 0.68, which is a 17% increase.

6.3 Influential Variables

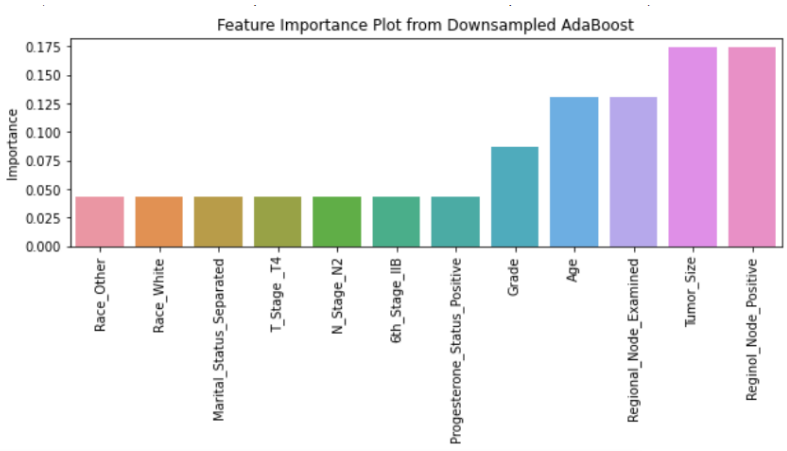

Figure 9 Feature importance plot generated from downsampled AdaBoost model

From the plot above, it can be seen that Regional Node Positive, Tumor Size have a strong influence on the survival outcome of breast cancer patients.

7. Conclusion

In conclusion, the AdaBoost model from a downsampled dataset is the best classification model since it yields the highest AUC (0.68). Downsampling did help improve model performance; average AUC has increased 21% by downsampling the unbalanced dataset. For all downsampled ROC curves, we can observed that when the threshold is increasing, the baseline model acts similar to a random classifier, which cannot find out the difference between two survival outcomes. In comparison, the downsampled models performed better than baseline models, especially AdaBoost.

The most influential variables when it comes to predicting the survival outcome of breast cancer patients are Regional Node Positive and Tumor Size. Unfortunately, due to the limit of AdaBoost, we are not able to quantify the impact of those two variables. However, according to the box-plots from basic analysis, we noticed that the average number of tumor size and positive regional nodes are higher than patient survived from breast cancer. We think these are the two potential directions for further analysis on factors affecting breast cancer patients' survival outcomes, especially for experts in causal inference that can draw casual relationship between those two variables and breast cancer. We hope this project can increase people's awareness of breast cancer and is willing to learn more about breast cancer.

References

Breast cancer: Breast cancer information & overview. American Cancer Society. (n.d.). Retrieved November 8, 2022, from https://www.cancer.org/cancer/breast-cancer.html

Brownlee, J. (2021, April 30). Tour of evaluation metrics for imbalanced classification. Machine Learning Mastery. Retrieved November 8, 2022, from https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/

Namdari, R. (2022, August 8). Breast cancer. Kaggle. Retrieved November 8, 2022, from https://www.kaggle.com/datasets/reihanenamdari/breast-cancer