Theoretical Machine Learning - DASC 5420

Cancer Survivability Prediction Using Machine Learning Techniques

Thai Pham - T00727094
Bardelosa, John Joshua - T00728432

April 18, 2024

Abstract

Cancer, a complex disease influenced by various factors such as type, stage, and individual responses to treatment, poses a significant global health challenge. With an alarming projection of over 35 million new cases by 2050, early diagnosis and effective treatment are paramount. While advancements have been made, factors like lifestyle, timely intervention, and ongoing support play pivotal roles in patient outcomes. This study delves into predicting cancer survivability, utilizing data from Kaggle focusing on cancer patients' outcomes. Through exploratory data analysis and a range of prediction models including logistic regression, LASSO regression, ridge regression, support vector machine, random forest, and gradient boosting machine, we aim to identify influential predictors of patient survival. Among the key findings, time of recurrence emerged as a critical factor affecting survivability, followed by the presence of the ESR1 gene. Models like random forest and gradient boosting machine showcased superior performance, underscoring the importance of robust prediction methodologies. Our study highlights the significance of early detection, treatment optimization, and the need for tailored interventions based on individual patient characteristics. By understanding the intricate interplay of variables impacting survivability, healthcare professionals can make informed decisions, ultimately improving patient outcomes and quality of life.

1 Introduction

Cancer is a serious disease that can affect different parts of the body and the severity depends on several factors including the type of cancer, its stage at diagnosis, and individual factors like response to treatment. Over 35 million new cancer cases are predicted in 2050, a 77% increase from the estimated 20 million cases in 2022 [1]. Although considered a complex disease, cancer can be treated if diagnosed early on. However, other types of cancer progress quicker than others, making it difficult to cure as the damage it does to a human's body develops faster than the body's response to the treatment. Worldwide, an estimated 19.3 million new cancer cases (18.1 million excluding nonmelanoma skin cancer) and almost 10.0 million cancer deaths (9.9 million excluding nonmelanoma skin cancer) occurred in 2020 [2]. Breast cancer is the most commonly diagnosed cancer, with an estimated 11.7% of the 19.3 million followed by lung cancer with 11.4%. Lung cancer, on the other hand, is the leading cause of cancer death with an estimated 1.8 million deaths in 2020. Overall, cancer poses the highest clinical, social, and economic burden in terms of cause-specific Disability-Adjusted Life Years (DALYs) among all human diseases [3].

While there have been advancements in research and treatment, other factors such as early detection, lifestyle, timely treatments, and ongoing support are pivotal in a cancer patient's survivability. Survival analysis is a field in medical prognosis that deals with the application of various methods to historical data to predict the survival of a particular patient suffering from a disease over a particular time period [4] Predicting Breast Cancer Survivability. With the increased utilization of computer-driven automated tools, extensive medical data is becoming more accessible, allowing more enhanced prediction models for survivability rates. Accuracy has been regarded as a primary measurement for the performance evaluation of the models, but stability which indicates the robustness of the performance to model parameter variation also becomes essential [5].

The main purpose of this research is to build a model that can predict cancer survivability and identify which predictors mainly affect the death of a cancer patient. Several methods were used as prediction models to extract metrics such as accuracy and precision. The most influential factors were identified through these models and cross-validation techniques were used to compare the results. This research can help cancer patients and health professionals make efficient decisions for treating cancer that can help slow down the progression of the disease, or even better, cure cancer.

2 Data

The data for this research is Predicting Cancer Diagnosis [6] from Kaggle. Although the title of the data set suggests that cancer diagnosis will be predicted, the study focuses on cancer patients' survivability, assuming everyone in the dataset already has cancer. This dataset has one response variable and 13 predictors and contains 217 observations. The definition of the predictors in the medical context is as follows:

1. Patient: This likely represents a unique identifier for each patient in a dataset.
2. Age: The age of the patient.
3. EventDeath: Indicates whether the patient experienced death (e.g., binary variable: 1 for death, 0 for survival).
4. TimeRecurrence: Time to recurrence of the disease.
5. Chemo: Indicates whether the patient received chemotherapy (e.g., binary variable: 1 for received chemotherapy, 0 for not received).
6. Hormonal: Indicates whether the patient received hormonal therapy.
7. Amputation: Indicates whether amputation was performed as part of treatment.

8. HistType: Histological type or subtype of the disease.
9. Diam: Diameter of the tumor.
10. PosNodes: Number of positive lymph nodes.
11. Grade: Grade or stage of the disease.
12. AngioInv: Presence of angioinvasion.
13. LymphInfil: Presence of lymphatic infiltration.
14. ESR1: Likely refers to the presence or expression of the estrogen receptor 1 gene.

3 Method

3.1 EDA and Data Preprocessing

Given that the response variable yields binary outcomes, our study naturally aligns itself with a classification problem. To commence model construction, we employed Exploratory Data Analysis (EDA) to meticulously scrutinize the dataset. This pivotal step facilitated the handling of missing values, categorical variables, and data visualization, thereby fostering precision in our analysis and enhancing the interpretability of our findings. Firstly, recognizing the limited relevance of the "Patient" column, which solely contained patient ID information devoid of any substantive value to our model, we judiciously removed it from our dataset. Next, we embarked on a transformative journey with the variable "age." Rather than treating age as a continuous variable, we sought to enhance its utility by segmenting it into meaningful age groups. By leveraging the minimum and maximum age values, we deftly categorized individuals into age cohorts, thereby imbuing our analysis with a nuanced understanding of age-related dynamics. Subsequently, we ingeniously engineered dummy variables for each age group, fortifying our model with granular insights into age-specific patterns.

Turning our attention to categorical variables such as "Histtype," "Angioinv," "Grade," and "Lymphinfil," we adopted a strategic approach to enhance their utility. Recognizing the inherent limitations of categorical data in traditional modeling frameworks, we artfully transformed these variables into binary counterparts. This transformative endeavor not only bolstered the accuracy of our model fitting process but also endowed us with a richer interpretative lens through which to glean insights from our results.

3.2 Procedure

Before proceeding with model fitting, we partitioned the data into training and validation sets using an 80-20 split ratio. Subsequently, we trained our models on the training set and evaluated their performance on the test set using appropriate metrics such as Accuracy, Recall, Precision, F1-Score, and AUC-ROC. Additionally, we assessed the significance of variables through their corresponding p-values. These parameters collectively informed our decision-making process regarding the efficacy of models fitted using various methodologies.

- **Logistic Regression and Logistic Regression using weighted class**: We utilize the train function from the caret package to conduct logistic regression model training with repeated 5-fold cross-validation on the training dataset. The best model index is determined by selecting the model with the highest accuracy from cross-validation results. Subsequently, the model is employed to predict the probabilities of eventdeath = 1 on the test set. These probabilities are then transformed into predicted class labels (0 or 1) using a threshold of 0.5. Evaluation of the model's performance on the test set is conducted by calculating the confusion matrix.

Next, we perform logistic regression model training considering class weights to handle imbalanced data. We calculate the class weights based on the ratio of samples in each class, assigning higher weights to the minority class to balance the training process. Subsequently, we proceed with the same steps as logistic regression without class weights, including model evaluation and performance assessment.

- **LASSO Regression and LASSO Regression using weighted class**: We utilize Lasso regression with 5-fold cross-validation to determine the optimal regularization parameter λ This parameter is selected based on the minimum mean cross-validated error. Once the optimal λ is identified, we train the Lasso model using the entire training dataset and make predictions of eventdeath = 1 on the test set using metrics similar to those mentioned earlier. Furthermore, we examine the coefficients derived from the Lasso model to pinpoint significant variables, focusing on those with non-zero coefficients.

  In the case of Lasso regression with weighted class, we compute class weights by considering the sample distribution across each class. This involves assigning higher weights to the minority class to ensure balance during the training phase. Subsequently, we follow the same procedures outlined above to evaluate the model's performance and identify important variables.

- **Ridge Regression and Ridge Regression using weighted class**: Ridge regression and Ridge regression with weighted class are conducted in a manner analogous to Lasso as described earlier, by modifying the alpha parameter within the glmnet function provided in R.

- **Linear Support Vector Machine**: We conduct a grid search to fine-tune parameters in a linear support vector machine (SVM) model, employing 5-fold cross-validation. Initially, a logistic regression formula is established, and a grid of tuning parameters (C values) for the SVM model is configured. Subsequently, an empty dataframe is created to record the outcomes of the grid search. Subsequently, each C value in the tuning grid is sequentially processed. During each iteration, a linear SVM model is trained using the designated C value, with 5-fold cross-validation executed, repeating the process 5 times.

- **Non-linear Support Vector Machine**: We perform a grid search to optimize parameters in a non-linear support vector machine (SVM) model utilizing a radial basis function kernel. Each row of the tuning grid is iterated over, training a nonlinear SVM model with a radial basis function kernel using 5-fold cross-validation and 5 repetitions for each combination of C and sigma values.

- **Random Forest**: We formulate a random forest model by amalgamating the target variable and predictors. Subsequently, we conduct model training using 5-fold cross-validation on the training set, with repetition performed 5 times. We assess the model's performance using similar evaluation metrics as previously described, and additionally, we generate a confusion matrix to delve deeper into the model's performance.

- **Gradient Boosting Machine**: We conduct a similar procedure as with Random Forest, but instead utilizing the gbm method within the built-in caret package in R.

3.3 Document link

- GitHub repositor link: `https://github.com/Jena-Thaipham/DASC-5420---Final-project/tree/0764501d5b5a4c599adbb985cce6a9c48fd3a9c2`

- Data link: `https://www.kaggle.com/competitions/mubravo/data?select=kaggle_to_students.csv`

4 Results and discussion

During the EDA step, it was seen that there were no missing values, therefore, imputation was not necessary. The average age of all the cancer patients is 43.83 years old with the oldest patient being 53 years old and the 26 years old as the youngest. This suggests that cancer is more common in middle-aged persons, however, it can still that having cancer does not automatically lead to death.

In the correlation heatmap, Figure 1, the grade or the stage (Grade) of the disease has the highest correlation to the event that a person dies with a 34% correlation. Followed by the diameter of the node (Diam) and the presence of angioinvasion (AngioInv) with 20% and 18% correlation.

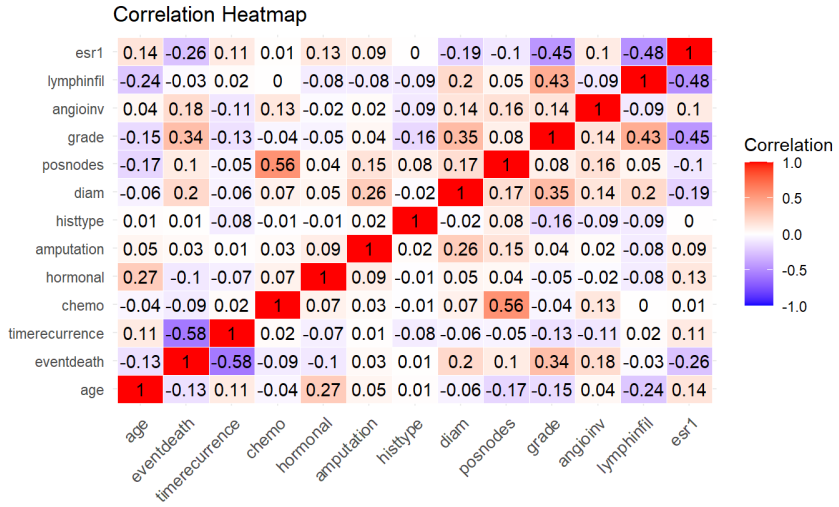

Figure 1: Correlation Heatmap

After the repeated 5-fold CV, a weighted class was added to Logistic Regression, LASSO Regression, and Ridge Regression to address imbalanced data. However, the weighted class yielded lower results than its original counterpart, therefore, the comparison will be made using the original models without weighted class. As seen in Table 1, GBM and Random Forest have the highest accuracy with 90.48% and 88.1% respectively. For the precision, LASSO has the highest precision with 90% followed by linear SVM with 88.09%. The precision of non-linear SVM is 100% but will be disregarded as getting 100% is theoretically possible but unrealistic in practice. This suggests that the model has a range of parameters that result in no false positives. Logistic Regression and GBM both resulted in 90.91% in the recall metrics.

To check for the overall performance of the models, the trade-offs of all the metrics, including F1-score and AUC - ROC, were considered. Logistic Regression has an overall good performance even though its precision is one of the lowest. As for LASSO, it has relatively high metrics but a lower recall which is caused by the regularization. Similar to LASSO, Ridge also indicates a good performance but a lower recall as well due to the penalty. For the SVM, only the linear SVM is considered to have a good overall performance while the non-linear has showcased extreme values. As for Random Forest and GBM, both models showcased a high overall performance making them the top two models with the best metrics.

	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Logistic regression	0.833	0.625	0.909	0.7407	0.8504
Logistic regression (weighted)	0.714	0.476	0.909	0.625	0.828
LASSO	0.861	0.900	0.643	0.75	0.828
LASSO (weighted)	0.767	0.625	0.714	0.667	0.815
Ridge	0.884	0.857	0.600	0.706	0.809
Ridge (weighted)	0.767	0.625	0.714	0.667	0.857
Linear Support Vector Machine	0.881	0.875	0.636	0.737	0.802
Non-linear Support Vector Machine	0.786	1.000	0.182	0.308	0.591
Random Forest	0.881	0.750	0.818	0.308	0.591
Gradient Boosting Machine	0.905	0.769	0.909	0.833	0.965

Table 1: **Model Performance Comparison**

Overall, the Random Forest model and the GBM model have the best performance and both resulted in TimeRecurrence as the most important variable followed by ESR1. As seen in the correlation heatmap above, these variables have a negative correlation to death, however, Random Forest and GBM capture non-linear relationships between the predictors and target variable. This means that even though a variable has a negative correlation, it could still have a significant impact on the response variable when considered in combination with other variables in a non-linear model.

5 Conclusion

In the top two models, the most important variable was the time recurrence of cancer (TimeRecurrence). This means that a person who has had cancer develops a secondary primary cancer has a lesser chance of surviving. Secondary primary cancer may be of the same cancer type or a different type developed in another part of the body. Recurrence not only increases the risk of death but might also lead to changes in exposure status [7]. This means that patients who experienced a recurrence of their disease later in time after the initial diagnosis or treatment had a better prognosis compared to patients who experienced a recurrence earlier. Advanced stage at primary diagnosis, distant metastases, adjuvant treatment for locoregional recurrence, and systemic treatment for distant recurrence were associated with increased mortality after late recurrence [8]. This means that the grade or stage of the cancer, prior to the secondary primary cancer, also plays a crucial role. Aside from the grade, the presence of ESR1 or the Estrogen Receptor 1 gene is the second most important variable in both models.

ESR1 is commonly found in various tissues and organs and this protein can influence cancer development and progression. Approximately 70% of breast cancers are estrogen receptor (ER)-positive, and many of these patients are effectively cured of their disease [9]. This also supports the claim as to why breast cancer is the leading diagnosis of cancer but not the leading cause of cancer death. One reason for this is there have been hormonal therapies used in the treatment of hormone receptor-positive breast cancer, however, there were some cases that caused mutations, resulting in cancer having attributes resistant to the therapy. Despite effective hormonal and targeted therapies, half of these patients will relapse or progress to incurable metastatic disease [9].

Based on the significant variables from the GBM and Random Forest models, regardless of what age an individual is, cancer is just as deadly at any other age. The factors that affect survivability the most are time of recurrence, presence of ESR1, number of positive lymph nodes, presence of lymphatic infiltration,

and grade of cancer. This implies that a person who has had breast cancer has a greater chance of recovery with the help of treatments and therapy, however, it also has a high chance of recurring again. A secondary primary cancer has a higher risk of incurability if it recurs at an early period after recovery. An even greater risk awaits the cancer patient if the stage of the cancer is already at a high level, there are several positive lymph nodes and the presence of lymphatic infiltration.

Cancer is a disease that continues to evolve and mutate which resulted in some cancers still incurable to this date. It can be noted that there are people who recovered from this disease but the resurgence of it increases the risk of mortality, especially if the cancer relapses shortly after recovery. Although there is no guaranteed way to prevent cancer, adopting certain lifestyle choices and behaviors can significantly lower the risk of developing the disease. It is also best to consult with a health professional if symptoms of cancer persist as early detection can help avoid further progression of the disease.

References

[1] Global cancer burden growing, amidst mounting need for services

[2] Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries

[3] Current Cancer Epidemiology

[4] Predicting breast cancer survivability: a comparison of three data mining methods

[5] Robust predictive model for evaluating breast cancer survivability

[6] Predicting Cancer Diagnosis

[7] Threats to Validity of Nonrandomized Studies of Postdiagnosis Exposures on Cancer Recurrence and Survival

[8] Mortality After Late Breast Cancer Recurrence in Denmark

[9] ESR1 mutations in breast cancer