Case Study – Breast Cancer Data Set

Shruthi Bashetti (4217619)  
Date 10-7-2022

**Executive Summary**

The objective of this report is to answer several questions that may be surmised from the Breast Cancer dataset that was provided as a data set for analysis. Regression analysis was used to determine factors like if the age has impact on cancer tumor size.

**Definition**

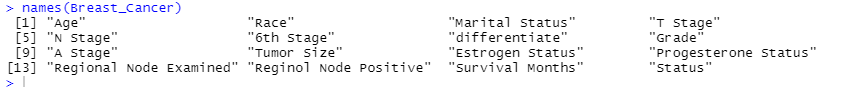
The purpose of this report is to analyze the “Breast Cancer” data set that was provided as a case study on Machine Learning. The data set contains information collected about age, race, marital status, tumor size, and different stages like T stage, N stage, 6th stage. This dataset also has information about estrogen, progesterone stages, survival months and if the person is alive or dead by the time this data was captured. This information will be used to analyze and determine if the age, race are affecting tumor size survival months of a affected.

**Preparation**

In order to gain a better understanding of the data, R and RStudio were used to perform various statistical analysis on the data set. EDA analysis gives brief information about dataset like first 6 rows of dataset, last 6 rows of dataset, variable names using functions like names(), head(), tail(), nrow(), ncol(), str().

**EDA Analysis**

Names() function gives all variables in dataset as shown in below.



This function gives all variables in dataset. In this dataset it has 16 variables which has character and numeric types of data.

Str() function gives in detail information about internal information of R object.

Text, letter

Description automatically generated

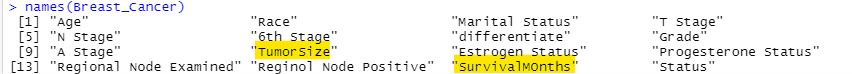
Using the dim() function in R, I was able to see that there are a total of 8048 observations and 16 columns. To gain further insight and understanding of the data I have used summary() which summarize various model fitting functions.

A screenshot of a computer

Description automatically generated with low confidence

For better functionality aspects I have renamed few variable names by removing space between words like changing “Survival Months” and “Tumor Size” to “SurvivalMonths”, “TumorSize” respectively using colnames() as shown below.





To understand further in detail, I have used histograms (Fig.1) to know if which age group has affected more to breast cancer. By looking at histogram which I got, ladies between 45-65 are affecting more with cancer than any other age group.

Chart, bar chart, histogram

Description automatically generated

**Figure.1**

I created plots using R to look at the how the size of tumor is behaving with respect to the age in different races of people like white, black and others. The purpose of this was to see if there were any possible trends that could correlate the tumor size and age. The following represents the histogram created Age Vs Tumor size.

Chart, bar chart, histogram

Description automatically generated

**Figure 2**

According to Figure 2, Black and other women are having higher tumor size than White women. Reasons for having high tumor size in black and others could be different based on the food habits, exposure to sun and also based on estrogen and progesterone levels in body.



Based on the dataset average age where women affected with cancer is 53. This we can find using mean() function as shown below.

Chart

Description automatically generated with medium confidence

**Figure 3**

Chart, scatter chart

Description automatically generated

**Figure 4**

According to Fig 3, Fig 4, the dataset has more of white women than black and other category of women and cancer victims whose tumor size is below 50 has higher survival months than others.

**Analysis**

Since my preliminary finding suggested that there is a correlation among survival months of a victim to the size of tumor, age and race I wanted to first create a validation set as was prescribed by Jared P Lander (2017). I decided to do a regression analysis these variables to verify if age, race, tumor size has significant impact on survival months of victims.

I would like to see if Estrogen and Progesterone has impact on Tumor Size which are actual reasons for one to get cancer. So, I have created 2 models see if Tumor Size is affected by Estrogen and Progesterone and age, race, tumor size has significant impact on survival months of victims

Below mentioned model is created using a regression formula for Tumor Size and the following statistical information was generated as a summary of that model fit (Table 2).

Call:

lm(formula = TumorSize ~ EstrogenStatus + ProgesteroneStatus,

data = Breast\_Cancer)

Residuals:

Min 1Q Median 3Q Max

-33.473 -14.776 -5.776 7.224 110.224

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 35.4734 0.9109 38.943 < 2e-16 \*\*\*

EstrogenStatusPositive -2.7224 1.0953 -2.486 0.013 \*

ProgesteroneStatusPositive -2.9754 0.7225 -4.118 3.85e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.06 on 8045 degrees of freedom

Multiple R-squared: 0.005647, Adjusted R-squared: 0.0054

F-statistic: 22.84 on 2 and 8045 DF, p-value: 1.281e-10

**Table 1**

According to table 1, it clearly says that Tumor size is affected by Estrogen and Progesterone levels as p value is lesser than 0.05, The next model is to see if age, race, tumor size has significant impact on survival months.

Call:

lm(formula = SurvivalMOnths ~ Age + TumorSize + Race, data = Breast\_Cancer)

Residuals:

Min 1Q Median 3Q Max

-71.726 -15.006 1.508 18.456 42.709

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 71.82840 1.83581 39.126 < 2e-16 \*\*\*

Age -0.04357 0.02857 -1.525 0.127

TumorSize -0.09586 0.01207 -7.943 2.24e-15 \*\*\*

RaceOther 6.60559 1.30596 5.058 4.33e-07 \*\*\*

RaceWhite 4.97186 0.98553 5.045 4.64e-07 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.79 on 8043 degrees of freedom

Multiple R-squared: 0.01152, Adjusted R-squared: 0.01103

F-statistic: 23.43 on 4 and 8043 DF, p-value: < 2.2e-16

**Table 2**

Based on the statistical information in Table 2 we can see that there is significant correlation among survival months to age, race, tumor size as a since p < 2.2e-16.

**Conclusion**

Based on the evidence presented by the models created to see if that Tumor size is affected by Estrogen and Progesterone and Age, Race, Tumor size has significant impact on survival months of victims, it’s clear that Tumor size, Survival Months are having dependencies on mentioned variables. Age and Race also have significant impact on tumor. Reasons for having high tumor size in black and others could be different based on the food habits, exposure to sun and also based on estrogen and progesterone levels in body.

# **References**

Lander, J. P. (2017). *R for everyone : advanced analytics and graphics*. Addison-Wesley.