Study on Lung Cancer Dataset for Both Male and Female Patients

- BY **SREYA BHATTACHARYA**

(UNDER THE GUIDANCE OF **PROF. DITHI BHATTACHARYA**)

Introduction

- Lung cancer is a type of cancer that originates in the lungs.
- It is one of the most common and deadliest forms of cancer worldwide.
- Lung Cancer Statistics :
  - According to WHO there is an estimated 2.21 million of new cases of lung cancer every year.
  - According to National Cancer Institute the rate of new cases of lung and bronchus cancer is 50.0 per 100,000 men and women per year. The death rate is 35.0 per 100,000 men and women per year
  - As per Lung Cancer Research Foundation, an estimated 238,340 People will be diagnosed with lung cancer in 2023 in the US.

Data Description

- The dataset is collected from Kaggle website. (https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer)
- This data consists :

  Total no. of attributes: 16 & No .of instances: 276
- Here I divided my dataset into two parts i.e. - Male (142 observations) and Female (134 observations.

Objective

- The main objective of this project is to see if the Multiple Logistic Regression model fits the collected data set.
- We'll check how accurately the model fits our data for both Male and Female.
- We'll compare the accuracy of the fitting in Male and Female data.

Methodology

- Dataset is splitted into two parts for both Male and Female.
- Train Data consisted 70% of the real Dataset & Test Data consisted 30% of the real Dataset.
- The multiple logistic regression model is fitted on the train data.
- We'll check the following values for both Male and Female Data :

1. P-Value
2. McFadden's R^2 Value & VIF values
3. Confusion Matrix
4. Sensitivity & Specificity
5. Accuracy

- Lastly the ROC Curve is plotted.

Calculation & Analysis

- R-Studio is used for all calculations.
- Based on those calculations the data is analyzed.
- Same procedure is used for both cases.

Train Data and Test Data

After splitting the data into Train Data and Test data we get

- For Female Data,
  - No. of rows in Train Data = 94 &
  - No. of rows in Test Data = 40
- For Male Data,
  - No. of rows in Train Data = 100 &
  - No. of rows in Test Data = 42

P-Value & McFadden's R^2 Value

- After calculating the P-Values we see
- For Female *Data Anxiety, Peer Pressure and Allergy* are important covariates.
- For Male Data *Yellow Fingers, Fatigue and Allergy* are important covariates.
- With these Important Factors we'll fit the Logistic Regression Model.

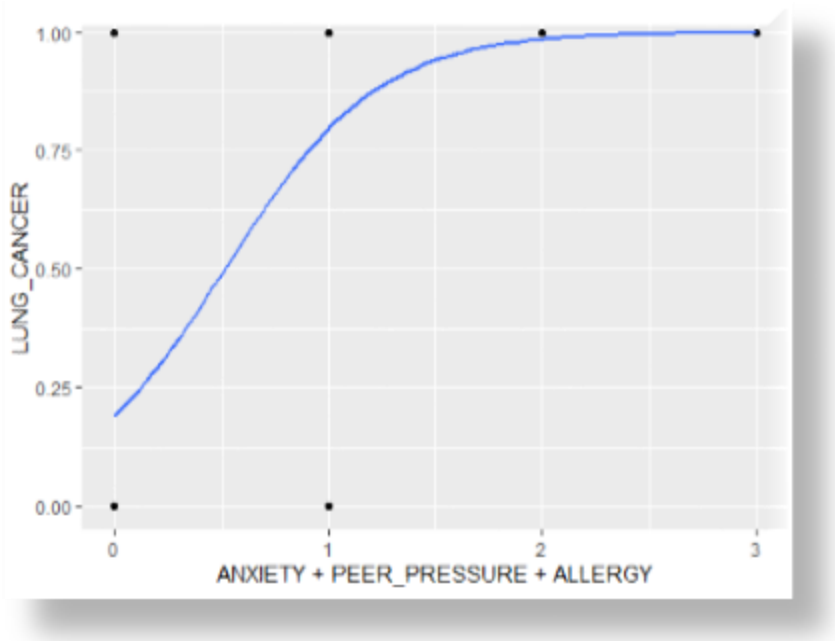
- For Female *Data* McFadden's R^2 value is 0.5815674.
- For Male *Data* McFadden's R^2 value is 0.39953.

VIF Values

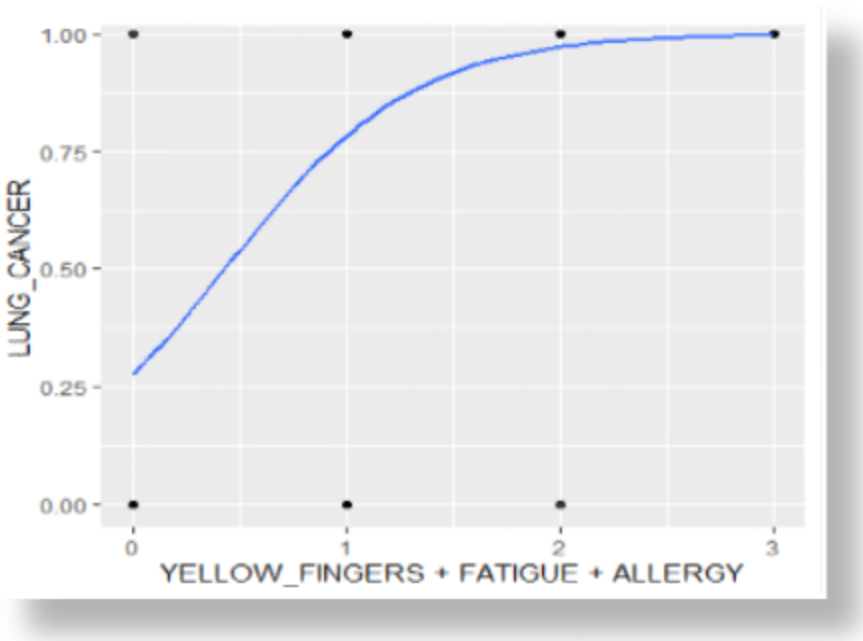
- For Female Data VIF Values of the important factors are,
  - Anxiety : 1.095288
  - Peer Pressure : 1.101561
  - Allergy : 1.136863
- For Male Data VIF Values of the important factors are,
  - Yellow Fingers : 1.314700
  - Fatigue : 1.219443
  - Allergy : 1.186989
- Multicollinearity is not an issue in our Model

Plotting the Data

Female Data



Male Data



Model Diagnosis

- We'll analyze how well our fitted Model performs on the Test Data.

Confusion Matrix

Female Data

	0	1
False	8	3
True	5	78

Male Data

	0	1
False	4	1
True	8	87

Sensitivity, Specificity & Accuracy

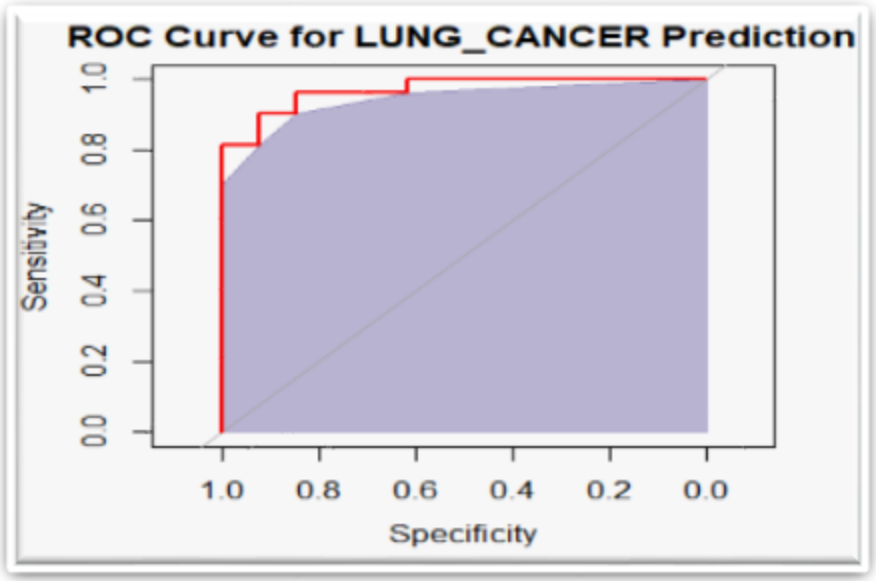
- For Female Data, Sensitivity Value is 0.61539.
- For Male Data, Sensitivity Value is 0.5.

- For Female Data, Specificity Value is 0.9397690.
- For Male Data, Specificity Value is 0.945652173.

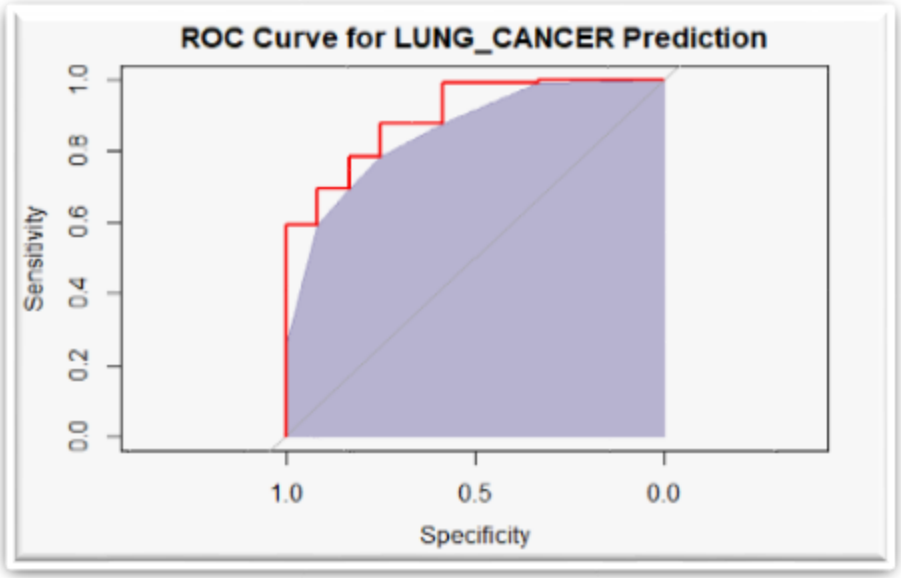
- The Accuracy for our Female Data is 0.9148.
- The Accuracy for our Female Data is 0.91

ROC Curve

Female Data



Male Data



Area under Curve

- For Female Data Area under Curve is 0.9435.
- For Male Data Area under Curve is 0.8532.

Conclusion

- From this report we can conclude that Lung Cancer Data of Female Patients will fit better in our model than Male Patient,

  i.e. Female Data will fit **94 % accurately** & Male Data will fit **85.32 % accurately** in our model.

Reference

- Fundamental of Statistics by Gun Gupta Dasgupta.
- An Introduction to Categorical Data Analysis by Alan Agresti.
- Logistic Regression A Self-Learning Text by David G. Kleinbaum, Mitchel Klein.

Acknowledgement

- I would like to express my special thanks of gratitude to my teacher and mentor of this project Prof. Dithi Bhattacharya for her constant support and guidance.
- I also thank to all teachers of statistics department who gave me knowledge which helped me to complete this project.
- I also want to thank University of Calcutta for giving me this opportunity.

Thank you