Final Resear****ch****

Course: CS555

Professor: Hong Pan

Researcher: Yidan Sun

Research Scenario

The data is called **Lung Cancer** from Kaggle recording the real-world dataset. Original data volume has amount of 309. The data is collected from the website online lung cancer prediction system .

Data include 15 variables in total, and most of the data is binary variables. In dataset value 2 represents “Yes”, value 1 represents “No”.

Variables explained as followings:

AGE: The age of the respondent.

GENDER: The gender of the respondent (Male or Female).

SMOKING: Whether the respondent is a smoker (Yes/No).

YELLOW\_FINGERS: Indicates if the respondent has yellow fingers (Yes/No), a common sign among smokers.

ANXIETY: Indicates if the respondent experiences anxiety (Yes/No).

PEER\_PRESSURE: Whether the respondent feels peer pressure to smoke (Yes/No).

CHRONIC DISEASE: Indicates if the respondent has any chronic diseases (Yes/No).

FATIGUE: Whether the respondent experiences fatigue (Yes/No).

ALLERGY: Whether the respondent has allergies (Yes/No).

WHEEZING: Indicates if the respondent experiences wheezing (Yes/No).

ALCOHOL\_CONSUMING: Whether the respondent consumes alcohol (Yes/No).

COUGHING: Whether the respondent experiences coughing (Yes/No).

SHORTNESS\_OF\_BREATH: Whether the respondent experiences shortness of breath (Yes/No).

SWALLOWING\_DIFFICULTY: Whether the respondent has difficulty swallowing (Yes/No).

CHEST\_PAIN: Whether the respondent experiences chest pain (Yes/No).

LUNG\_CANCER: The outcome variable indicating whether the respondent has been diagnosed with lung cancer (Yes/No).

My research question is: What are the significant risk factors for developing lung cancer, considering demographic characteristics, smoking behavior, and other health indicators?

Data Exploration

The dataset I used from Kaggle is cleaned, and so I converted GENDER and LUNG\_CANCER to factor.

First to understand data, I explore the data by making some data visualizations. The age distribution for respondents in this survey mostly gather around 50 to 70 with median (62) and mean (62.7) and hence the age distribution is roughly normal distributed, the only outlier we can tell from the histogram is the minimum age is 21, which is rare to see. It may indicate that lung cancer mostly happen in the late period of lifetime.

A graph of a number of people

Description automatically generated

A close up of numbers

Description automatically generated

The distribution of gender shows that among the lung cancer respondents, males can be a little higher than females, not giving too much difference.

A graph of a number of people

Description automatically generated with medium confidence

Statistical Approach

To investigate the research question, I used logistic regression model, a statistical method commonly used to model the relationship between a binary outcome variable (such as having or not having lung cancer) and one or more predictor variables (such as smoking status, age, gender, etc.). Logistic regression estimates the probability of the occurrence of an event (in this case, lung cancer) based on the values of the predictor variables.

Report Result

With default significance level of alpha=0.05, we can see exploratory variables like Smoking, Peer Pressure, peer pressure, chronic disease, fatigue, allergy, coughing, swallowing difficulty are statistically significant and will be contributed to the lung cancer.

Surprisingly, smoking variable isn’t the most significant contributor, top few smallest p-value, that are contributed significantly are peer pressure, chronic disease and fatigue. In contrast, age, gender, anxiety, alcohol consuming and some other health indicators were not found to be statistically significant in this model.

Residual deviance 91.909 on 293 degrees of freedom, it measures the goodness of fit of the model. A lower value indicates a better fit. The model has significantly improved the fit by including the predictors.

A screenshot of a computer

Description automatically generated

Conclusion and Limitations

From the analysis, I found that several factors significantly increase the likelihood of developing lung cancer: Smoking, Peer Pressure, Chronic Disease, Fatigue, Allergies, Coughing, and Swallowing Difficulty

In confirms that smoking and several health symptoms significantly raise the risk of lung cancer. People who smoke or have chronic diseases, experience fatigue, or have symptoms like coughing and difficulty swallowing should be particularly vigilant about lung cancer.

The potential limitation is that the sample size is of 309 and might be too small to make that conclusion, it might show bias and cannot fully represent the truth.

Additionally, the dataset might not capture all relevant risk factors (e.g., genetic factors, environmental exposures). Also, some predictors might be correlated with each other, which can complicate the interpretation of their individual effects on lung cancer risk.

Finally, since the logistic regression model is linear relationship between the log-odds of the outcome and the predictors, which may not perfectly capture complex biological interactions.

Data Reference:

https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer