Predictive modelling. LUNG CANCER

Kulakova Tatiana, 5694626W

Temasek Polytechnic, School of Informatics & IT – SPECIALIST DIPLOMA IN BUSINESS ANALYTICS AY2023/2024

OCT SEMESTER (TERM A), DATA ANALYTICS FOR BUSINESS INSIGHTS (CBA1C09)

INTRODUCTION

The reason why I choose this data set from Kaggle is that having a question in mind is how it possible to predict lung cancer based on set of data. As my uncle had died from lung cancer being a smoker. What are the key factors (variables)? WHO defines **Lung cancer is the leading cause of cancer-related deaths worldwide** and **Smoking is the leading cause** of it (around **85% of all cases)**.

DATA EXPLORATION & PREPARATION

Sample Properties

Property	Value
Rows	309
Columns	17

Sample Statistics

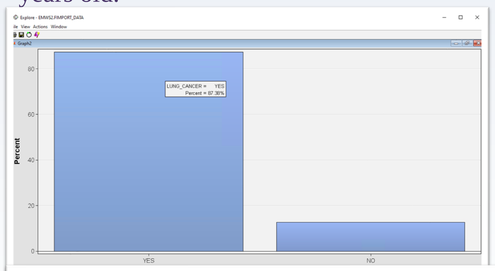
Obs #	Variable Name	Type	Percent Missing	Minimum	Maximum	Mean	Number of Levels	Mode Per...	Mode
1	GENDER	CLASS	0	.	.	.	2	52.42718	M
2	LUNG_CANCER	CLASS	0	.	.	.	2	87.37864	YES
3	PatientrID	CLASS	0	.	.	.	128+	0.775194	P001
4	AGE	VAR	0	21	87	62.67314	.	.	.
5	ALCOHOL_CONSUMING	VAR	0	1	2	1.556634	.	.	.
6	ALLERGY	VAR	0	1	2	1.556634	.	.	.
7	ANXIETY	VAR	0	1	2	1.498382	.	.	.
8	CHEST_PAIN	VAR	0	1	2	1.556634	.	.	.
9	CHRONIC_DISEASE	VAR	0	1	2	1.504854	.	.	.
10	COUGHING	VAR	0	1	2	1.579288	.	.	.
11	FATIGUE	VAR	0	1	2	1.673139	.	.	.
12	PEER_PRESSURE	VAR	0	1	2	1.501618	.	.	.
13	SHORTNESS_OF_BREATH	VAR	0	1	2	1.640777	.	.	.
14	SMOKING	VAR	0	1	2	1.563107	.	.	.
15	SWALLOWING_DIFFICULTY	VAR	0	1	2	1.469256	.	.	.
16	WHEEZING	VAR	0	1	2	1.556634	.	.	.
17	YELLOW_FINGERS	VAR	0	1	2	1.569579	.	.	.

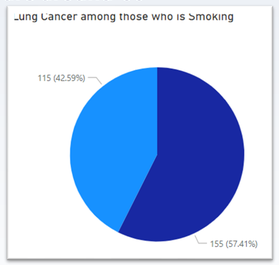
Attribute information:

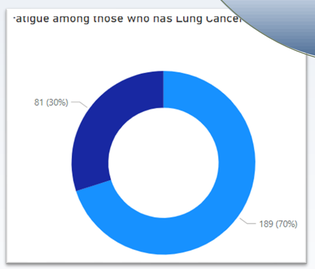
1. Gender: M(male), F(female)
2. Lung Cancer: YES, NO.
3. PatientID- unique ID
4. Age: Age of the patient
5. Other attributes: YES=2, NO=1.

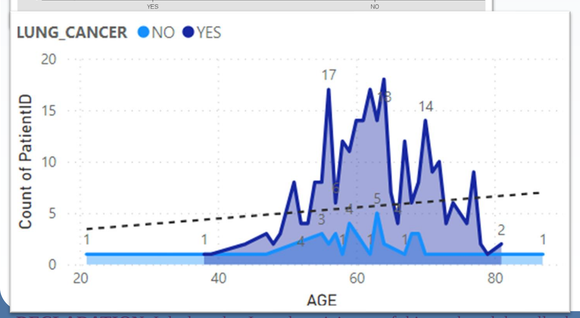
Data profiling shows that there are 309 rows, 17 columns, there is no missing values. Minimal age is 23, maximum age is 87, while Mean is 62.7 years in the dataset. Dataset is clean.

Following charts showing that 87.38% of cases have Lung Cancer, while 12.62% cases do not have it. Most of patients have shortness of breath and are smoking. Most of patients with lung cancer diagnosis are around 60 years old.







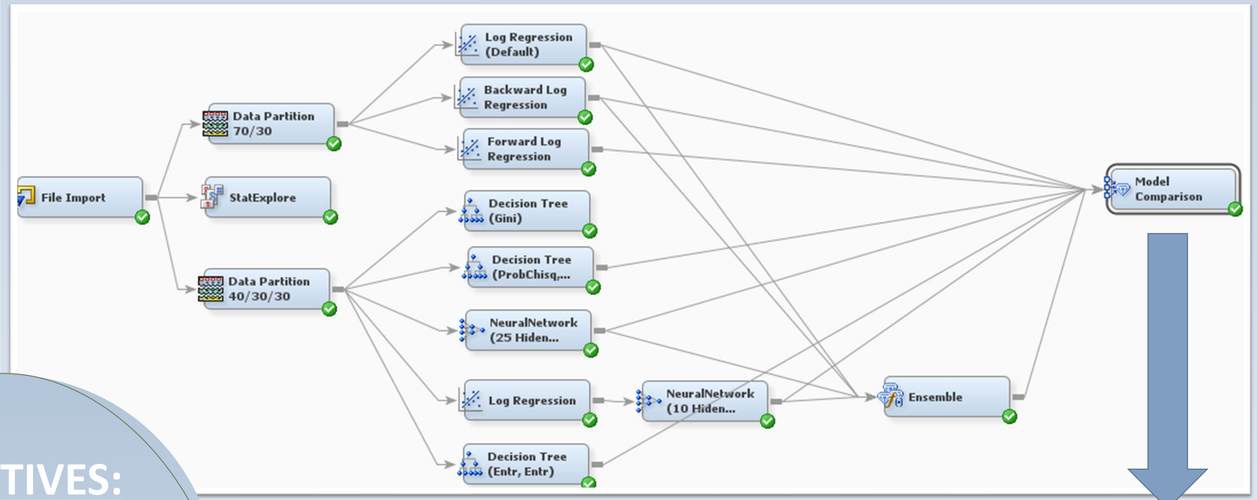


Score Code Variable	Sign	Absolute Coefficient
SWALLOWING_DIFFICULTY	+	3.819213597
FATIGUE	+	3.512904682
COUGHING	+	3.25497496
CHRONIC_DISEASE	+	2.873965683
PEER_PRESSURE	+	2.691626547
SMOKING	+	2.357114636
YELLOW_FINGERS	+	2.128970013
ALLERGY	+	1.923966883
ALCOHOL_CONSUMING	+	1.798545847

OBJECTIVES:

Performance of Predictive Modeling Task, Interpretation, Recommendation

WORKFLOW



The best performing model is Backward Logistic Regression
(Test: 0.08, Train: 0.05,
Selection statistic is Misclassification rate).

Model Description	Test: Misclassification Rate
Backward Log Regression	0.08421
Log Regression (Default)	0.09474
Ensemble	0.10638
NeuralNetwork (10 Hiden units)	0.10638
NeuralNetwork (25 Hiden units)	0.10638
Forward Log Regression	0.11579
Decision Tree (Entr, Entr)	0.11702
Decision Tree (ProbChisq,Entropy)	0.12766

Ensemble model is based on 4 best performing models. Ensemble model as 3d best result (Test: 0.106, Train: 0.05,Validation:0.04).
The slight difference in Misclassification rate indicates that Ensemble model is minor overfitting.

FINDINGS & RECOMENDATIONS

Doctors can use Backward Logistic Regression for predicting Lung Cancer as 1st level filter (Misclassification rate: 0.08, Accuracy: 0.92). Then Doctors can look at most urgent cases.

To make Model even more effective for predicting, the future improvement can be done towards **Over Sampling techniques** to handle Imbalanced Data. We train imbalanced dataset (Lung Cancer=Yes, 87.38%). As a result, the models try to learn only the majority class (Lung Cancer=Yes, 87.38%) and result in overfitting.

REFERENCES

1. Data Analytics For Business Insights Module course materials from Temasek Polytechnic
2. SAS Enterprise Miner Workstation 14.1 Help
3. WHO website https://www.who.int/news-room/fact-sheets/detail/lung-cancer
4. 7 Over Sampling techniques to handle Imbalanced Data, https://towardsdatascience.com/7-over-sampling-techniques-to-handle-imbalanced-data-ec51c8db349f