Predicting Lung Cancer Outcomes: Leveraging Machine Learning Techniques For Detection And Prognosis

Sunset Coral

Ho Inn Jong Jereme, Ian Oon, Poon Hiu Ching, Sean Gastinov Soeandi, Tan Lu Yan

MH4510 Data Mining
AY2024/25 Semester 1, 15 November 2024

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Definition

- Lung cancer is a leading cause of cancer-related deaths globally
- Often undetectable until advanced stages

Objectives

- Use machine learning models to predict lung cancer risk based on survey data
- Generate probabilistic outputs for early detection without formal diagnosis
- Enables informed decision-making for patients and healthcare providers to prioritise treatment levels

Overview

- Widely used on Kaggle, chosen for its simplicity in deploying YES/NO survey questions to patients
- Data consists of 15 features related to lung cancer symptoms

EDA

- Significant class imbalance (87% positive for lung cancer), balanced by gender
- Focus on ages 55+, with highest diagnoses in the 60-65 range
- Key features: Alcohol consumption, allergies, and anxiety; low multicollinearity

Implemented Models for Testing

- Gradient Boosting Algorithms
- Deep Forest
- Logistic Regression
- Artificial Neural Network
- Stacking Ensemble Method

LASSO

- Loss Function $L_L(\beta) = -\sum_i (y^i log p(x^i) + (1 - y^i) log(1 - p(x^i))) + \sum_{j=1}^{p} |\beta_j|$
- Adds a penalty term to penalize the coefficients' absolute values

Categorical Boosting (CatBoost) Process and Key Features

- Loss Function $L(f(x), y) = \sum_i w_i \cdot l(f(x_i), y_i) + J(f)$
- BinarisedTargetMeanValue, Symmetric Tree Structure, Ordered Boosting, `scale_pos_weight` parameter, and Variable Importance.

Results and Conclusion

Model	Training B.A.	Testing B.A.	Difference	Overfitting
Extreme Gradient Boosting	87.52%	75.18%	12.62%	Mild
Adaptive Boosting	89.11%	70.27%	18.84%	Severe
Categorical Boosting	93.95%	82.40%	11.55%	Mild
Deep Forest	98%	79.68%	18.33%	Severe
Logistic Regression	89.11%	78.14%	10.97%	Mild
LASSO	88.79%	80.09%	8.70%	Mild
Elastic Net	87.51%	78.79%	8.72%	Mild
Artificial Neural Network	$\sim 88\%$	$\sim 78\%$	$\sim 10\%$	Mild

Table: Model Performance Comparison

Conclusion

- CatBoost has the highest test balanced accuracy, while LASSO best minimised overfitting
- Significant Variables **(CatBoost)**: Swallowing difficulty, alcohol consumption, age
- Significant Variables **(LASSO)**: Swallowing difficulty, fatigue, chronic disease
- 0.1 threshold: Minimize false negatives (undetected cases)
- Balanced accuracy: Address dataset imbalance