ML model for Shared Decision-Making Tool for CRC screening

Daiane M. Seibert and Karen Feyen
Thomas More University of Applied Sciences, Belgium

ORINT
tOwaRds Informed dEcisions iN colorecTal cancer screening

Dataset

The PLCO dataset, Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial dataset.

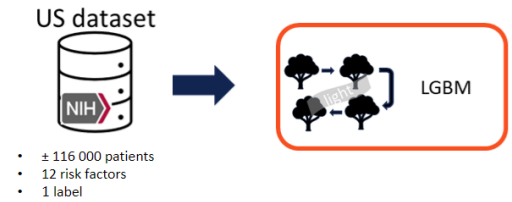
Enrolment period: 1993 - 2001
Data size: ~154,000
Screening period: 1993 - 2009
Data Collection: +500 risk factors. 6 questionnaires

Features

- Sex,
- Age,
- Height,
- Weight,
- BMI,
- Hypertension,
- Heart problems,
- Diabetes,
- Smoke history,
- Smoke quantity and
- Alcohol drink history.

Exception made directly in the family history feature.*

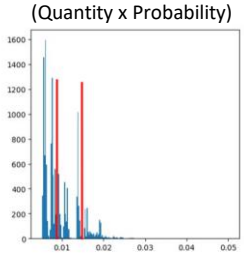
Model



- ± 116 000 patients
- 12 risk factors
- 1 label

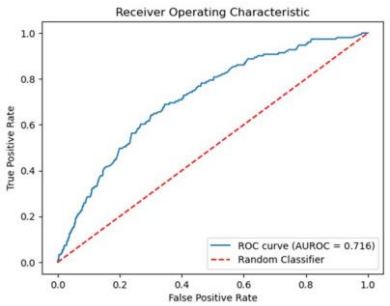
We trained a Light Gradient Boosting Machine (LGBM) Regressor model with PLCO dataset. With ~116000 patient data, between these ~1700 were positive cases.

Model prediction (Test set)
(Quantity x Probability)



Model's Performance

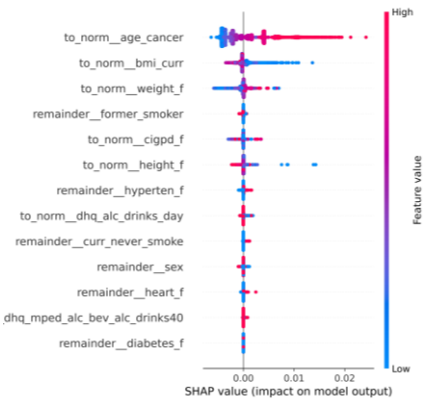
AUROC of the model is 0.716.



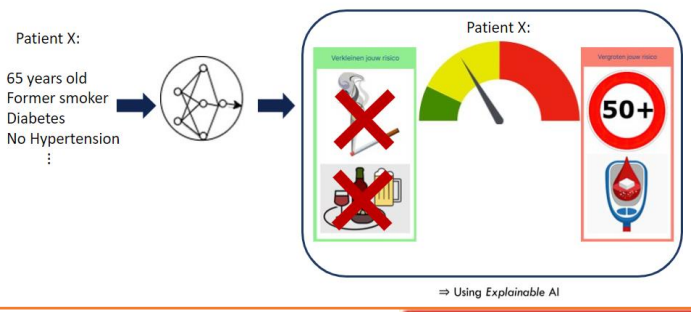
The Receiver Operating Characteristic (ROC) curve is a **graphical representation of the true positive rate (sensitivity) plotted against the false positive rate (1 - specificity)** as the discrimination threshold of the model varies. The area under this ROC curve, the AUROC, is a single scalar value that summarizes the overall performance of the model across various threshold settings.

SHAP

SHAP values quantify the impact of each feature on a specific prediction compared to a baseline prediction. They show how much each feature pushes the prediction up (positive impact) or down (negative impact).



Inference Example



Kom op tegen Kanker
DOMUS MEDICA
vito
Universiteit Antwerpen
RESEARCH @THOMAS MORE