Q5. Smoke Status Recognition

Preprocessing

After printing the information of df, we can see that there are many missing values. Next, we will fill them in.

Firstly, for the missing values of hearing and vision, due to the symmetry of the human body, we will fill in the missing data on the other side.

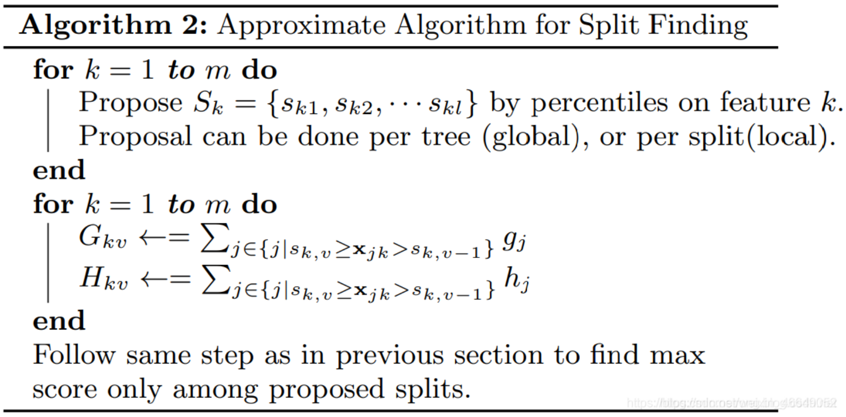
The most reliable method for professional indicators related to medicine is to fill in the mean value.

Finally, we roughly estimate the three closely related variables of height, weight, and waist circumference using some recognized calculation methods.

Modeling

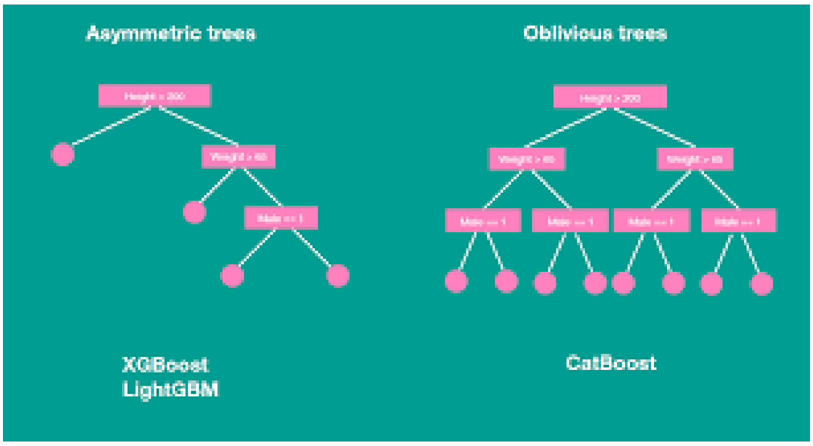
Now we use processed data to train the model. For this task, I chose two models, xgboost and catboost, for training.

XGB is an implementation method of boosting algorithm, mainly aimed at reducing bias, that is, reducing model errors. Therefore, it adopts multiple base learners, each of which is relatively simple to avoid overfitting.



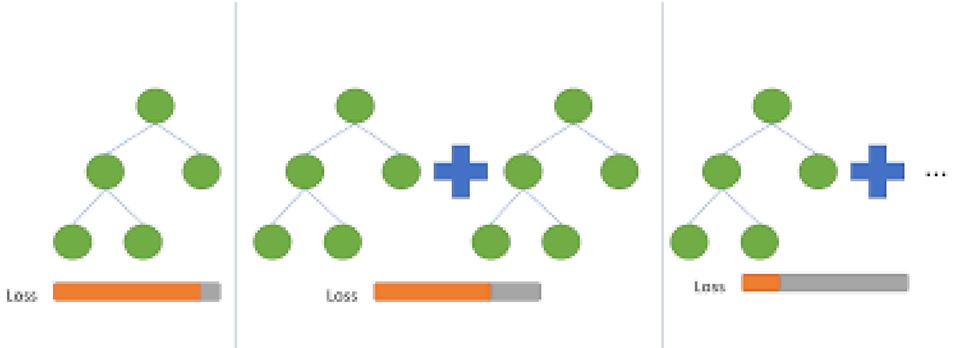
```
Algorithm 2: Approximate Algorithm for Split Finding
for $k = 1$ to $m$ do
    Propose $S_k = \{s_{k1}, s_{k2}, \cdots s_{kl}\}$ by percentiles on feature $k$.
    Proposal can be done per tree (global), or per split(local).
end
for $k = 1$ to $m$ do
    $G_{kv} \leftarrow= \sum_{j \in \{j | s_{k,v} \geq \mathbf{x}_{jk} > s_{k,v-1}\}} g_j$
    $H_{kv} \leftarrow= \sum_{j \in \{j | s_{k,v} \geq \mathbf{x}_{jk} > s_{k,v-1}\}} h_j$
end
Follow same step as in previous section to find max
score only among proposed splits.
```

CatBoost is a GBDT framework based on symmetric decision trees as the base learner, which has fewer parameters, supports categorical variables, and high accuracy. Its main pain point is to efficiently and reasonably process categorical features, which can be seen from its name. CatBoost is composed of Categorical and Boosting. In addition, CatBoost also solves the problems of gradient bias and prediction shift, thereby reducing overfitting and improving the accuracy and generalization ability of the algorithm.



Training

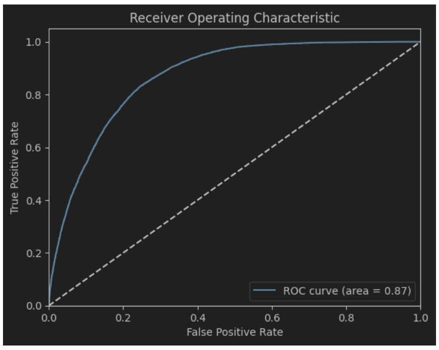
Training using two models and then subtract the average of their predicted probabilities. It has been proven that the cooperation between the two models is better than that of a single model.



We evaluate the models on val set. Then we calculate the AUC score.

AUC: 0.8686464662867673

Draw ROC curve:



Output the results

Doing these on test set and save the results as csv.

Q5_output.csv