Q5-Smoke Status Recognition

This Task can be basically regarded as a binary classification problem. To have a more accurate result, Ensembling Machine Learning methods are always the first choice. Thus, I use the XGBOOST algorithm to tackle this binary classification problem.

### Data Manipulation and Analysis

Firstly, I do some data manipulation and analysis to have a deep understanding of our dataset. The total training sample counts 159256, and there are some samples containing the missing value.

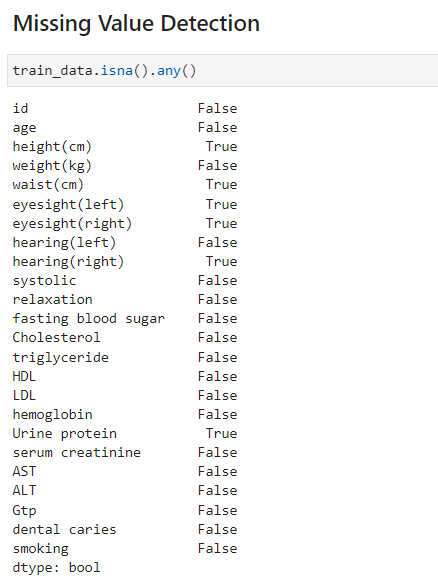


Figure 1 Missing Value Detection

We can also know the label distribution on the training data

A screenshot of a computer

Description automatically generated

Figure 2 Lable Distribution

The label distribution is very close to Evenly distributed among 0 and 1

### XGBoost

Before applying xgboost to classify the data, I compute the variance of each variable to see their impact on the prediction

A screenshot of a computer

Description automatically generated

Figure 3 Variance

Except id, The triglyceride variate largely among all the training sample. In addition, The ability of hearing is close for all the participants.

#### Grid Search for best parameters

I simply use the grid search method to search for the best parameters of xgboost classifier with 5-fold cross validation

* Learning rate: {0.1, 0.05, 0.01, 0.2}
* Number of estimators: {100, 200, 300, 400}
* Max Depth: {3, 5, 7}

And I also try some features selection and engineering to increase the quality of the training set.

Firstly, I try to remove the features with loweset variance.

|  |  |
| --- | --- |
|  |  |

Figure 4 Feature Importance and ROC Curve(remove features with low variance)

Average AUC on Validation dataset is 0.8674.

Secondly, I also try not to drop the missing value in training samples, which means I retain all the participants in my training sample.

|  |  |
| --- | --- |
|  |  |

Figure 5 Feature Importance and ROC Curve(without Dropna)

Lastly, I drop all the missing value in my dataset and apply the xgboost algorithm.

|  |  |
| --- | --- |
|  |  |

Figure 6 Feature Importance and ROC Curve(Dropna)

This time, the AUC on validation dataset is 0.87002.

Finally, I use xgboost with the best paramters to infer on test dataset and save the results.