BSAN6070

Machine Learning

CA05-A Logistic Regression

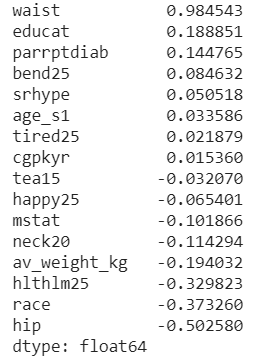
Vincent Chen

Model choosing

For choosing the best hyperparameters in the logistic model, I was grouping all the possible models by penalty methods, because different penalty methods could affect the other parameters significantly. With the help of ‘for’ loop to list performance evaluations for each combination of parameters, I used the sum of AUC and accuracy score as the criteria to choose the best model, which is:

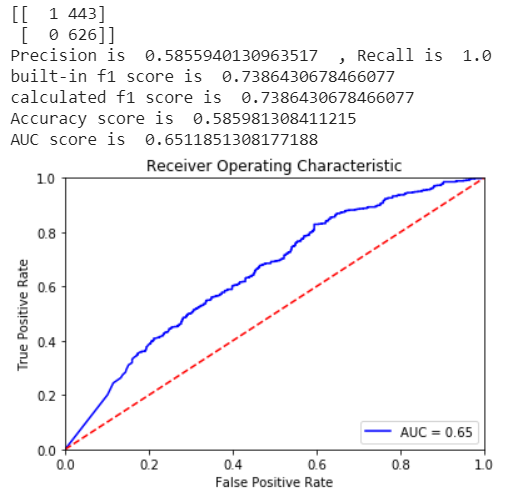
#penalty='l2', C=0.06951927961775606, solver='newton-cg'

Feature importance



Waist seemed to be the best predicter for CVD risks, followed by education and parrptdiab, which is Parent Cohort reported Diabetes Status, one sort of History of Diabetes, according to sleepdata.org.

Model Performance



Precision is too low, which means the model is very easy to predict that someone has the CVD risk, while he/she might not. Recall rate is very high, meaning very few persons with CVD risk would not be detected. AUC score and accuracy score are about 0.6 and 0.65, not too high, due to low precision score.