Mfolozi Dlamini

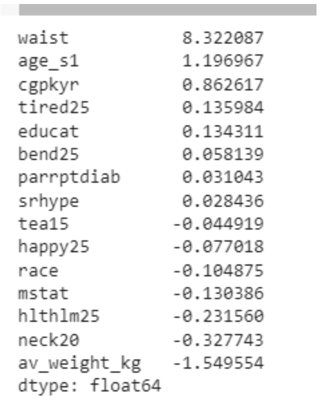
BSAN 6070 CA05 Logistic Regression

Dr Arin Brahma

April 04, 2022

Part 3: Explanation and Conclusion of Performance Outputs

1) Features of Importance in descending order



```
waist          8.322087
age_s1         1.196967
cgpkyr         0.862617
tired25        0.135984
educat         0.134311
bend25         0.058139
parrptdiab     0.031043
srhype         0.028436
tea15         -0.044919
happy25       -0.077018
race          -0.104875
mstat         -0.130386
hlthlm25      -0.231560
neck20        -0.327743
av_weight_kg  -1.549554
dtype: float64
```

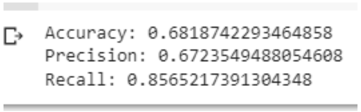
It is clear from our input that the top 5 features of importance for predicting the risk to cardiovascular disease (CVD) are:

- **Waist (Waist measurement in centimeters)**
- **age_s1 (Age at time of study in years, based on start date of Sleep Heart Health Study Visit One (SHHS1) Polysomnography (PSG) recording. Values equal to 90 indicate an age of 90 or greater.)**
- **cgpkyr (Cigarette pack-years),**
- **tired25 (Quality of Life (QOL) (Sleep Heart Health Study Visit One (SHHS1)): Felt tired),**
- **educat (Education level of participant)**

2) When performing the data pre-processing, we discovered that 'waist' and 'hip' are highly correlated. Therefore, it made sense to remove one of these variables from the model. 'hip' was removed from the Logistic Regression model.

3)

a) Performance evaluation of model

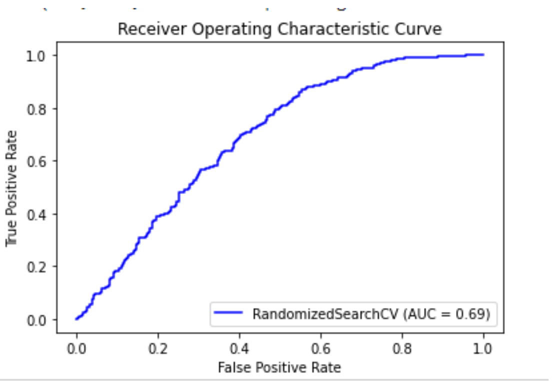


```
Accuracy: 0.6818742293464858
Precision: 0.6723549488054608
Recall: 0.8565217391304348
```

From the above report we note the following performance metrics:

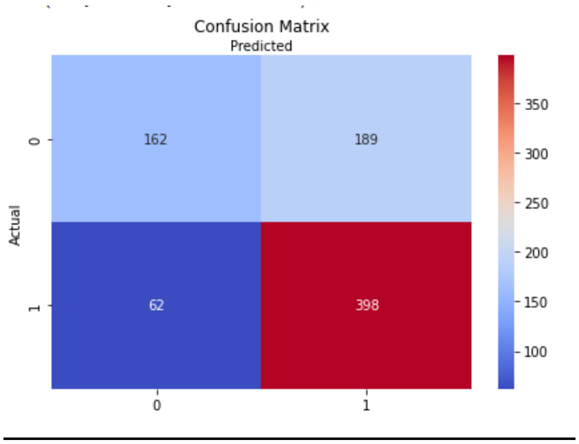
- **Accuracy: 0.68**
- **Precision: 0.67**
- **Recall: 0.86**

b) ROC Curve



a. Area Under Curve: 0.69

c) Confusion Matrix



The model returned the highest recall of 0.83 using newton-cg method when fitting the logistic regression. The highest value of AUC was also obtained after the 'hip' feature was removed and the readjustment of the penalty.

Additionally, the confusion matrix of the model shows that the model does a good job at predicting TP but the accuracy of 0.69 is still very worrying because it is not as high as we would want it to be. Ideally an accuracy measure of 0.85 or higher would've been more plausible. It is important to note that the prediction is through randomized search and therefore will not be a perfect prediction.

In conclusion: Although the logistic regression model is a good classifier it might not be the best option for this dataset given the model performance when predicting Negative. We could try implore other methods that we have learnt in this class.