Krissy Wong

BSAN 6070

4/1/2022

CA05 – Logistic Regression

Part 3: Evaluate the performance of your model (including ROC Curve), explain the performance and

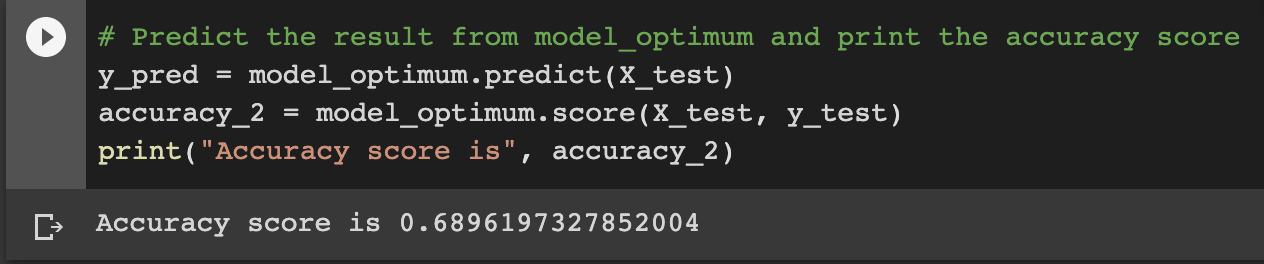
draw a meaningful conclusion. (deliverable: Performance outputs in Notebook, explanation and

conclusion in Word/PDF document)

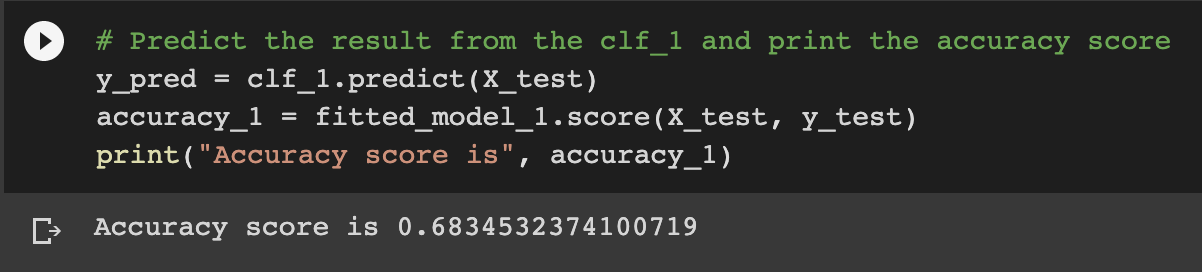
**Model Choosing**:

I chose the model with a penalty of “l2” and a C value of “0.8884981901414992”. I tested three different models by tuning the hyperparameters and adjusting the C values. The accuracy score using “l2” as the penalty and “0.8884981901414992” as the C value yielded the highest accuracy compared to the other two models. Although there is a slight increase in accuracy when the penalty and C value are adjusted for, the accuracy does not significantly differ from Models 1 and 2. Accuracy remained the same for Models 1 and 2.

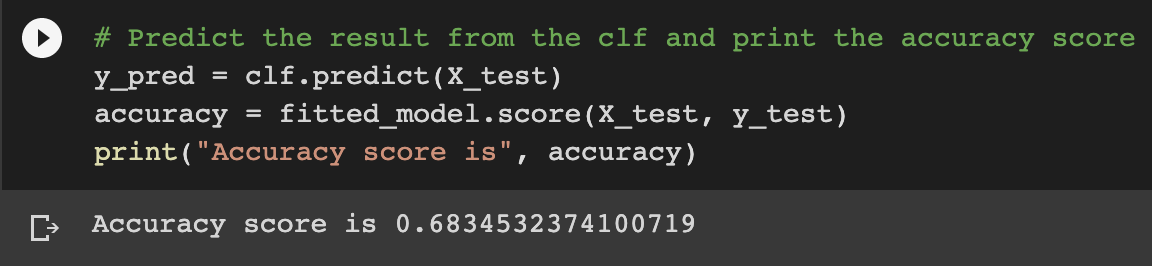
Chosen Model/ Best Model



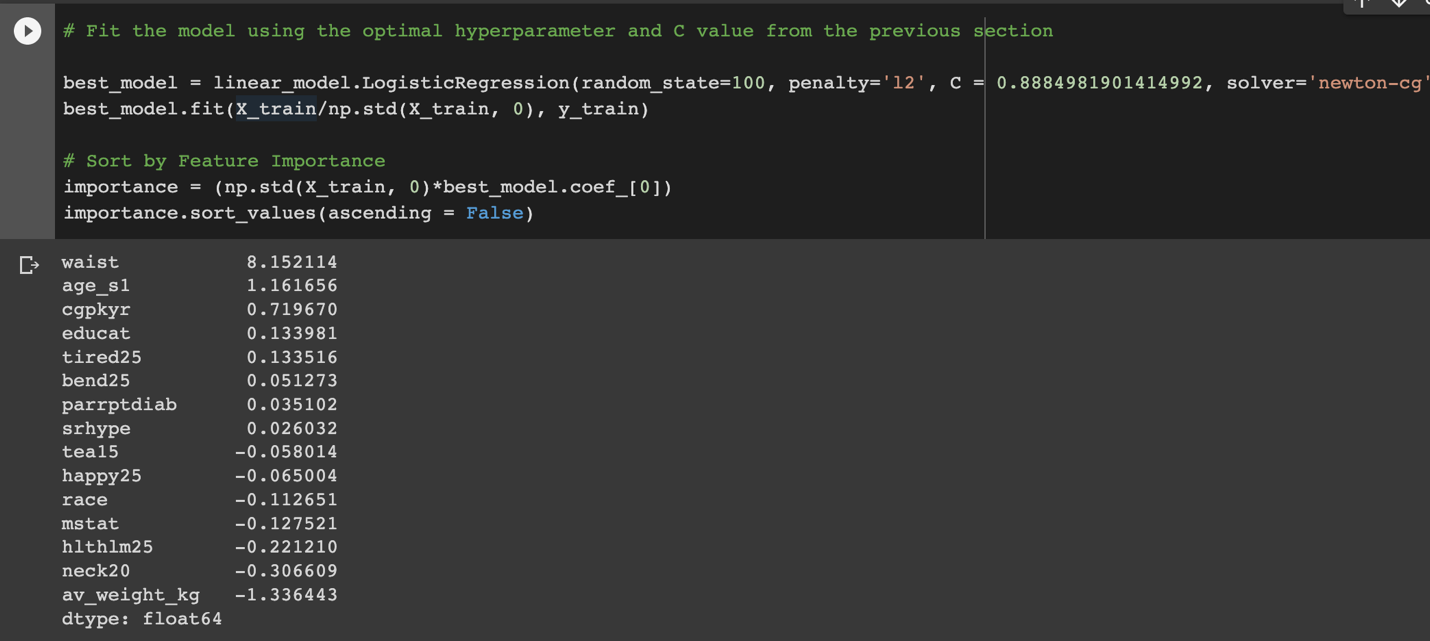
Model 1 with C = 1e40



Model 2 with adjustments only penalty and not C value.



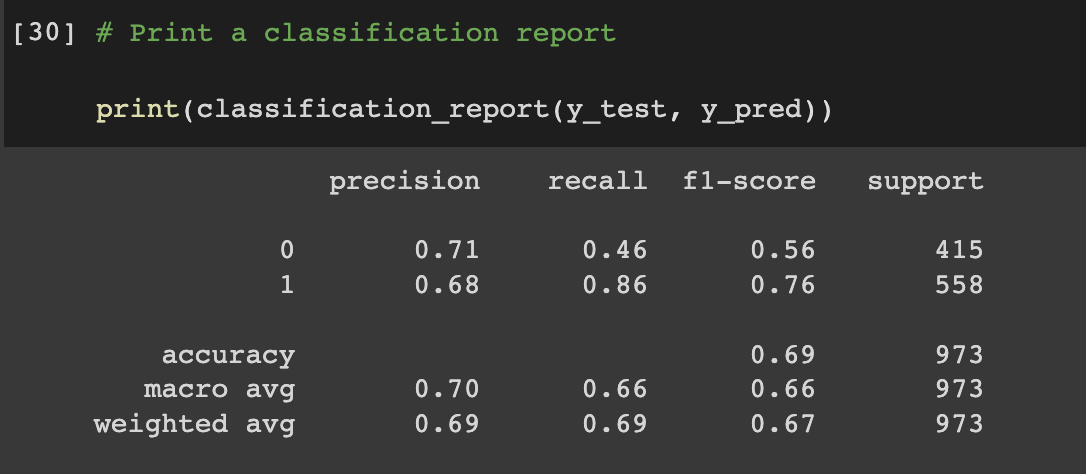
**Feature Importance**

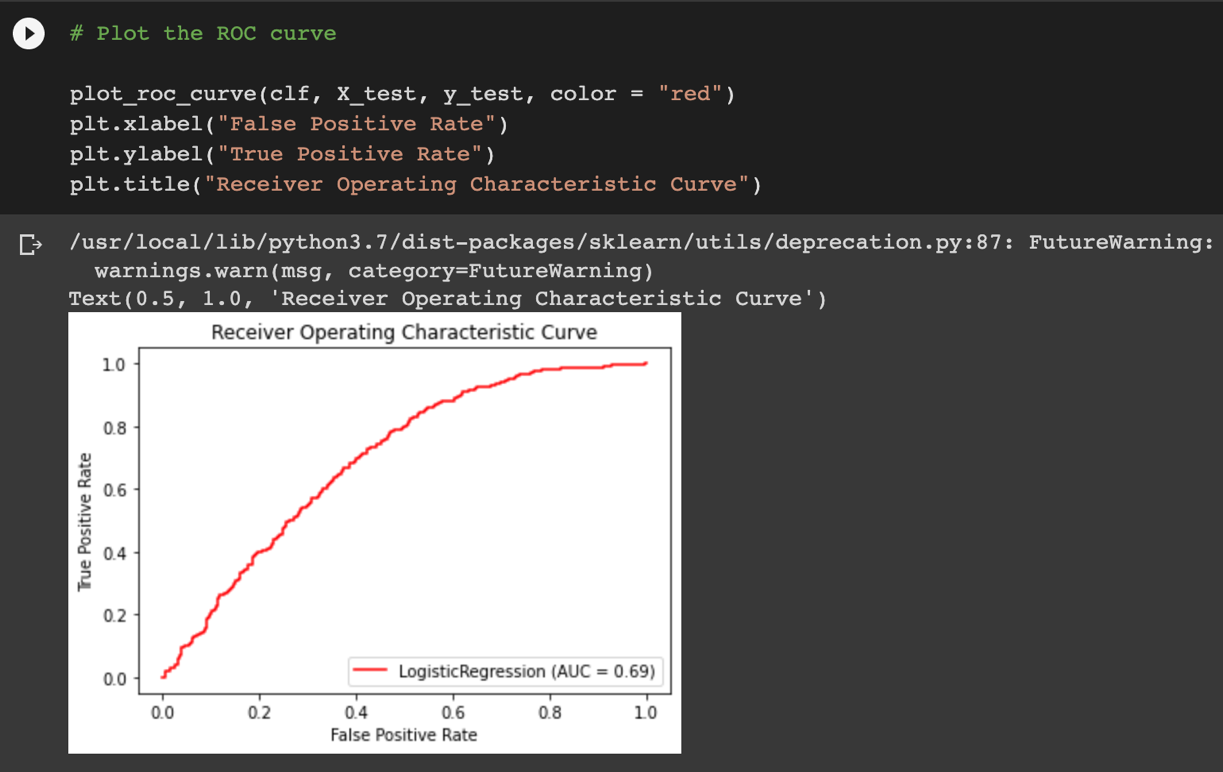
****

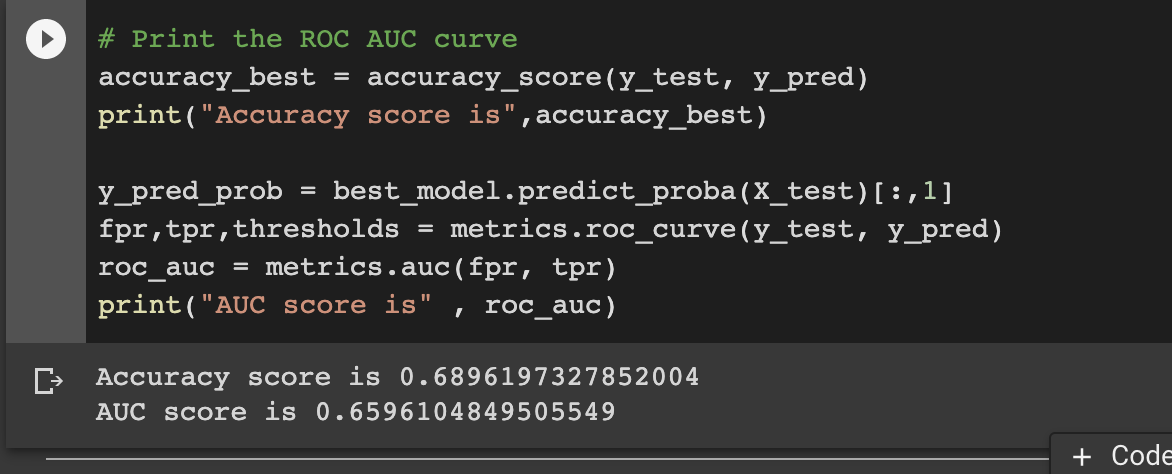
Waist seems to be the best predictor for CVD risks, followed by age\_s1 and cgpkyr, which is cigarette pack-years, according to sleepdata.org.

**Model Performance**

The precision for risk (1) is lower than the precision for no risk (0) which indicates that the model has a higher chance for falsely predicting that a person has CVD. Recall is significantly higher for risk than no risk, meaning that fewer people with CVD risk would not be detected.

****

****

****

AUC and accuracy scores are 0.689 and 0.659, respectively, which are roughly in the midpoint due to the lower precision scores for risk and no risk.